CLUSTERING AND VISUALIZATION OF SINGLE-CELL RNA SEQ DATA OF PERIPHERAL BLOOD MONONUCLEAR CELLS (PMBCs) USING: SCANPY

B-573 Programming for Science Informatics
Final Project Report
By Anusha Bellapu

INTRODUCTION:

Ribonucleic acid (RNA) is formed from the Deoxyribonucleic acid (DNA) by transcription and is translated to proteins, which have many vital functions in the human body. RNA can alter the normal human body function during the assembly of amino acids or nucleotides in the process of its assembly or during its translation or improper post-translational changes before forming the functional protein.[1] Ribonucleic acid (RNA) transcription and translation and its stability affect the physiological and pathological state of the body. When there is gene overexpression which can be caused epigenetically or due to internal mutations can lead to uncontrolled cell growth and proliferation.[2,3] The human cell never remains in the same state even though it is derived from a common zygote. During differentiation to form an organ its function and expression change. The dysregulation or dysfunction in any of these cell types may lead to diseased states.[4] Bulk cell RNA sequencing is the measurement of the average gene expression of the sample population of cells. Single cell RNA sequencing (scRNA-seq) is the measurement of gene expression and genomic changes at the cellular level. With the advancements in the single cell sequencing techniques, scRNA-seq had a humungous effect on the research field. Bulk cell RNA sequencing could measure the differential gene expression between the samples but failed to analyze the expression at the cellular and tissue levels. That is where scRNA-seq has taken the front seat in analyzing the expression at the cellular and transcriptomic levels, especially in diseases where the cell states were unstable. [5] The single cell analysis also made it possible to study the sub-populations of cells, their transition in the fate of the cell during the disease progression, and their gene expression.[4] This helps in finding the subtypes of the diseases like cancer, monitoring the drug response, and helps us design the cell-targeted therapies.[6] PMBCs are any blood cells that are a vital part of the body's defense system and play a critical role in fighting infections. They are of many types like T-cells, B-cells, monocytes, and platelets. There are many reasons to study the PMBCs in the human body one main reason is to observe the drug toxicity to the new treatment in the body; drug toxicity can affect the PMBCs as they are the most common targets. They are also important in personalized medicine and pharmacogenomics to study differential drug responses. The study of PMBCs also gives a glance at the diseased and healthy people's gene expressions.[7]

OVERVIEW OF SCANPY:

SCANPY is a freely accessible python-based tool to analyze single-cell gene expression data. As the number of cells in transcriptomic datasets grows crossing the one million range, the frameworks like SEURAT [8], MONOCLE [9], SCDE/PAGODA[10], SCATER[11], SCRAN[12], and CELL RANGER[13], become increasingly complex to use and thus difficult to scale. This study demonstrates that SCANPY's multi-modal nature makes it well-suited for both batch processing and interactive exploration through the usage of advanced machine learning algorithms in python. The proposed framework integrates complex analysis workflows into a single dashboard that enables multidimensional data exploration, deep learning classification and prediction, and built-in visualization through Python scripts.

SCANPY combines the scalability and modularity functions in analysis comparable to established R-based frameworks. Preprocessing methods of SCANPY are comparable to SEURAT[8] and CELL RANGER[13], visualization methods to t-distributed stochastic neighbor embedding (tSNE) [14,15] diffusion maps [14,16,17]and graph-drawing[18,19,20] and, clustering like PHENOGRAPH [21,22,23]

Scanpy uses a data form called ANNDATA for the analysis. An ANNDATA object represents a data matrix with annotations. It gives access to machine-learning tools to easily extract information without occupying much memory.[24] It is similar to R's EXPRESSIONSET but supports sparse data and saves data on a disc in HDF5-based format, which is not dependent on the platform, framework, and language. This allows operating on an ANNDATA object without fully loading it into memory. Its graph of neighborhood relations among data points is better than the popular package SCIKIT-LEARN.[25]

SCANPY is an open-source graph visualization and analytics engine that helps users quickly extract insights from relational data. It provides as a convenient import/export tool, allowing users to easily share their graphs. Using built-in tools such as random walk-based metrics and distance matrix computation, SCANPY allows you to use a single dataset for multiple purposes without re-importing each time. This platform can be managed by a community and data can be shared amongst them as the transfer of data is easy and independent of the language used in the analysis.[26]

METHODS:

1. *DATA COLLECTION:*

Dataset (with intronic reads) a single cell expression dataset by Cell Ranger 6.1.2 version is taken from the 10x genomics database. Human peripheral blood mononuclear cells (PBMCs) were obtained by 10x Genomics from AllCells from a healthy female donor aged 25-30 years.
Out of 16,000 cells (11,984 cells recovered) as described in the Chromium Single Cell 3' Reagent Kits User Guide (v3.1 Chemistry Dual Index) (CG000315 Rev C) using the Chromium X. The reads are sequenced on an Illumina NovaSeq 6000 to a read depth of approximately 40,000 mean reads per cell.

2. *PACKAGES AND TOOLS USED IN THE ANALYSIS:*

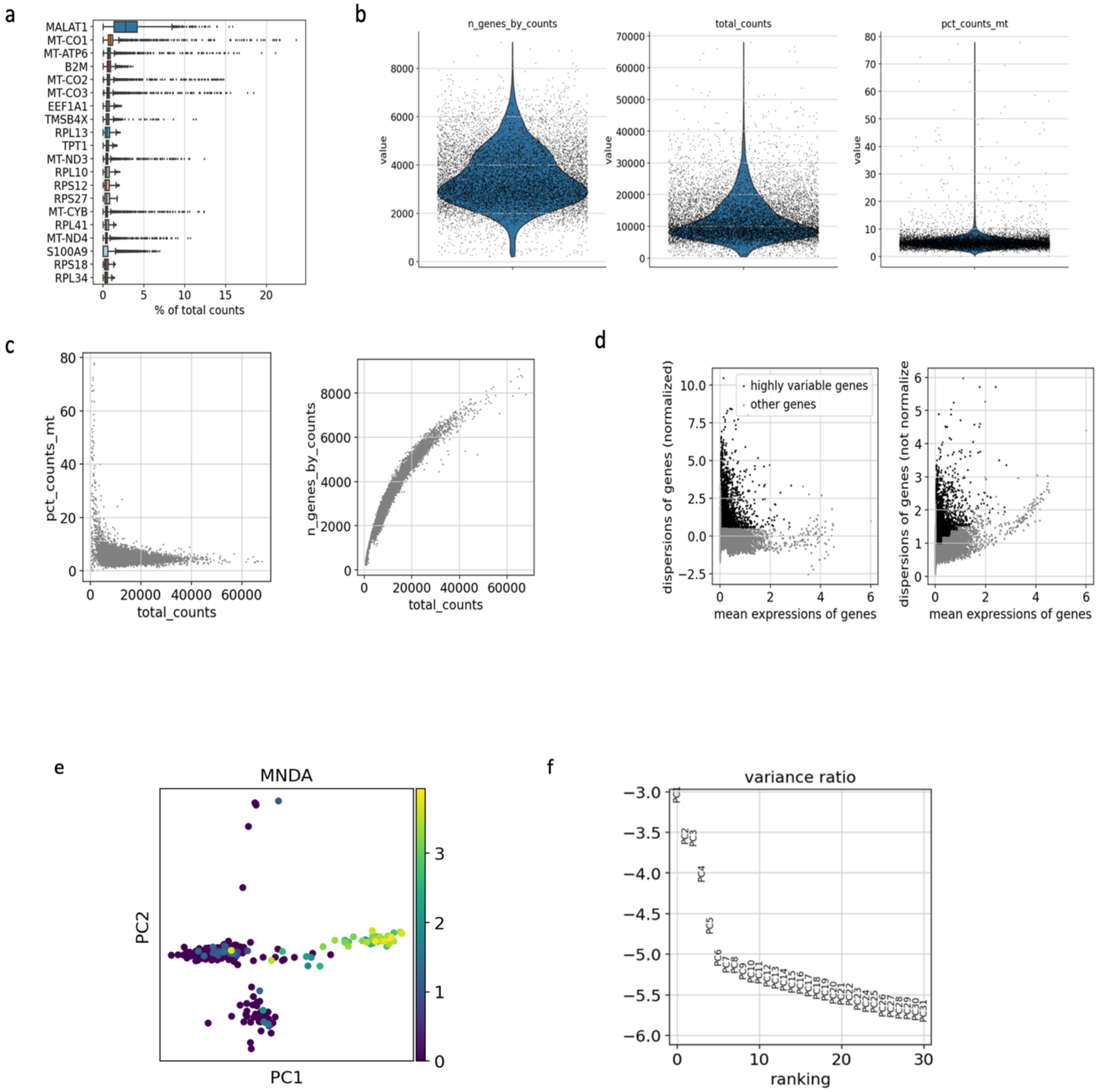
The scRNA-seq analysis was based on Python and performed in Google Collaboratory[27] environment. The packages such as SCANPY[26], NumPy[28], Pandas[29], Matplotlib[30], AnnDATA[31], leidnalg[32], Seaborn[33].

3. DATA ANALYSIS:

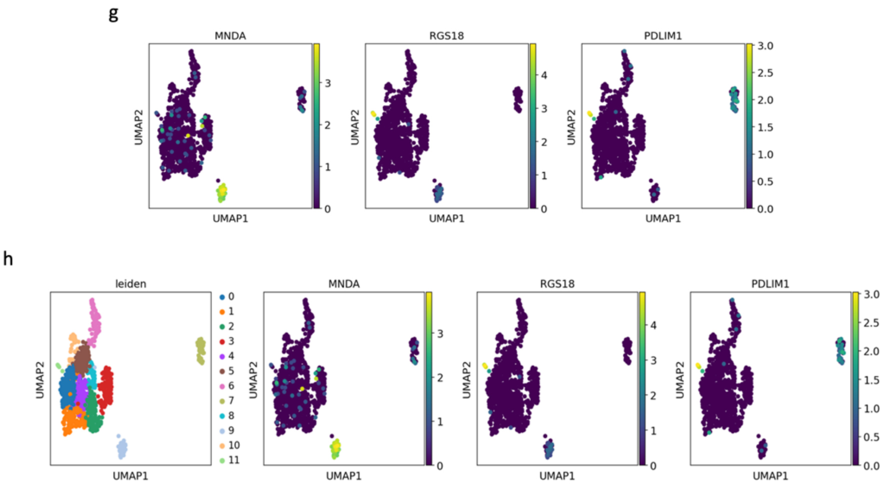
The single cell expression dataset has been analyzed using SCANPY using the above-mentioned packages. The analysis follows the below mentioned steps.
Data was loaded into the Linux terminal and unzipped files were retrieved and used for the analysis. **Data preprocessing** is the first step where the high fraction genes are identified, and the cells are filtered to remove the genes that are detected in less than 3 cells. This step also involves removing the mitochondrial genes and computing the number of genes after filtering into a count matrix. All the data is stored in the form of AnnData. Total count normalization is also done by making the 10,000 reds per cell so, that all the cells have uniform counts followed by log normalizing the data and finding the highly variable genes. The highly variable genes stored in the data matrix are detected by the PCA. **Principle component analysis** (PCA) is performed on the filtered data to reduce the dimensionality. PCA is a method to reduce the dimensionality of the data without the information loss by retaining the variability of the dataset. [34] Principal components (PCs) that are contributing to the variance of the dataset are identified and selected roughly to compute the neighborhood relations of cells.
The **neighborhood graph** was computed and embedded in two dimensions by uniform manifold approximation and projection (UMAP)[35] It is better than the tSNE. tSNE and UMAP are the analysis tools in single-cell transcriptomics used to visualize the scatterplots of the cells which belong to a similar cell type and are positioned closely. Both use the loss functions and combine similar points and separate the dissimilar points.[36] In **Clustering the neighborhood graph** Leiden clustering is used which directly clusters the graphs of the cells. The Leiden does move the nodes and partition the cluster and aggregation of the network based on the refined partition. It is the improved algorithm of Louvain algorithm to avoid badly connected communities.[37] As the last step statistical tests are performed to identify the highly differential gens called **marker genes**, using t-test, Wilcoxon rank-sum test and logistic regression is also used to introduce the multi-variate approach while the other two tests use the uni-variate approach. Using the marker genes and the literature the cell types are determined and the clusters are annotated using the cell types. We can use different visualization methods to represent the genes and clusters.

Results:

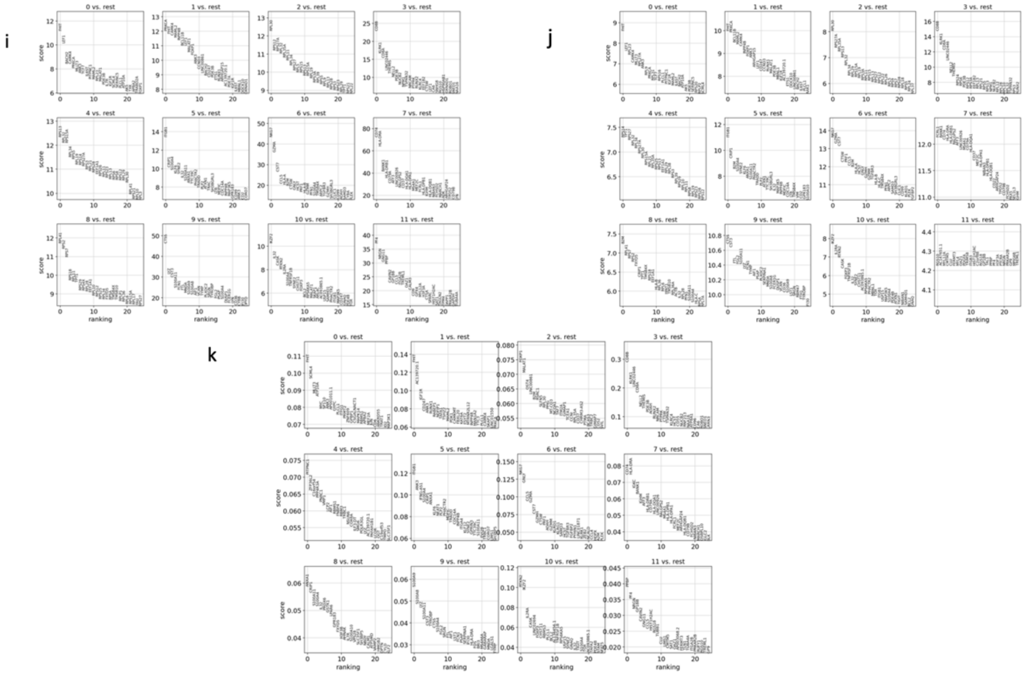
The analysis of this dataset resulted in an AnnData matrix with 11984 observations and 36601 genes. In the data preprocessing 32 cells were filtered out which have less than 200 genes and 9238 genes that are present in less than 3 cells are filtered out. After performing the data scaling further analysis was done with 964 observations and 3553 genes. The loadings of the three PCs are visualized to see the expressed gene for each PC and embeddings are visualized using UMAP. After the application of Leiden 11 clusters were found. Further analysis was done, and marker genes were identified as FHIT, FOXP1, CD8B, PITPNC1, ITGB1, NKG7, CD74, ANXA1, S100A9, RTKN2, PPBP.



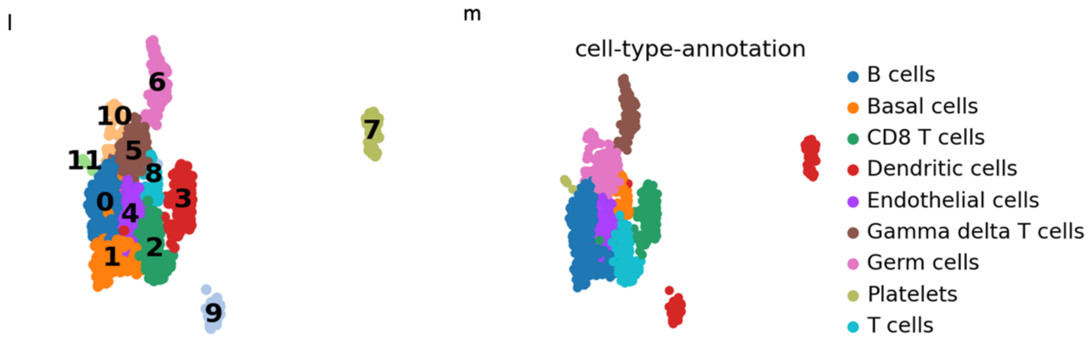
a. Figure (a) shows the genes that yield the highest fraction of counts in every single cell, across all cells. **b.** A violin plot of the number of genes expressed in the count matrix, total counts, and the percentage of the mitochondrial genes in the cell. **c.** Scatter plot showing the distribution of total gene counts in the cell's vs mitochondrial genes and genes by counts. **d.** The distribution of the highly variable genes before and after data normalization. **e.** Scatter plot of the PCA coordinates of the gene that is highly expressed in the PC1. **f.** The variance of each PC in the data set is visualized to Consider these for further analysis.



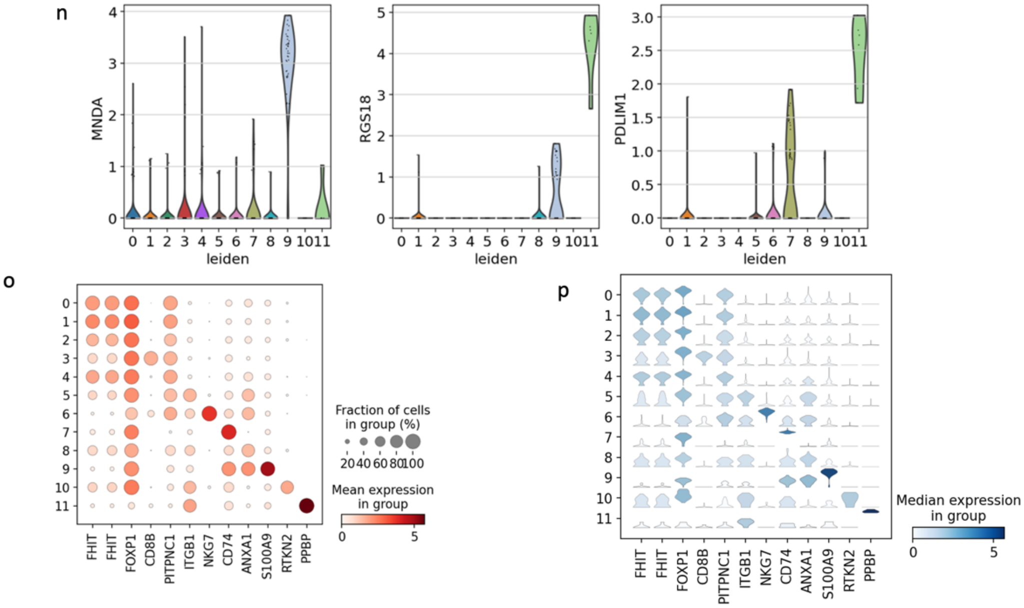
g. The graph embedding is visualized in two dimensions using UMAP of the highly expressed genes in the first three PCs. **h.** The graph embeddings of the scaled and normalized data.



i. Analysis of highly differential genes in all the 11 clusters using t-test **j.** Analysis of highly differential genes in all the 11 clusters using the Wilcoxon rank-sum (Mann-Whitney-U) test. **k.** Analysis of highly differential genes in all the 11 clusters using logistic regression.



l. 11 clusters resulted from the Leiden clustering depending on the marker genes **m.** The cell type annotation of the clusters is based on the cell type. The clusters were merged into 9 clusters as two of the genes belonged to the B-cells and germ cells



n. The violin plots of the genes MNDA, RGS18, and PDLIM1 in all 11 clusters. **o.** The dot plot shows the expression of the marker genes in each cluster. **p.** The visualization of the expression of each gene in the clusters in the form of violin plots.

Conclusion: scRNA-seq using SCANPY can be a promising advancement in single cell RNA sequencing transcriptomics and is documented to analyze the data and visualize more accurately compared to many other tools and frameworks available currently. It can make the scRNA data analysis simpler and user-friendly and can lead to study the disease like cancer, rare diseases, and immunological diseases, also improves safer drug discovery by helping study the drug toxicities

References:

1. Wang D, Farhana A. Biochemistry, RNA Structure. InStatPearls [Internet] 2022 May 8. *StatPearls Publishing*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK558999/
2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*. 2009 ;10(1):57-63.
3. Ozsolak F, Milos PM. RNA sequencing: advances, challenges, and opportunities. *Nature reviews genetics*. 2011;12(2):87-98.
4. Adil A, Kumar V, Jan AT, Asger M. Single-Cell transcriptomics: current methods and challenges in data acquisition and analysis. *Front Neurosci*. 2021; 15:591122.
5. De Windt LJ, Hegenbarth JC, Lezzoche G, Stoll M. Perspectives on bulk-tissue RNA sequencing and single-cell RNA sequencing for cardiac transcriptomics. *Frontiers in Molecular Medicine*.:2.
6. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*. 2022;12(3): e694.
7. Pourahmad J, Salimi A. Isolated human peripheral blood mononuclear cell (PBMC), a cost-effective tool for predicting immunosuppressive effects of drugs and xenobiotics. *Iranian journal of pharmaceutical research: IJPR*. 2015;14(4):979.
8. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015; 33:495–502.
9. Trapnell C, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014; 32:381–6.
10. Kharchenko PV, Silberstein L, Scadden DT, Bayesian approach to single-cell differential expression analysis. *Nat Methods*. 2014; 11:740–2.
11. McCarthy D, Wills Q, Campbell K. SCATER: single-cell analysis toolkit for gene expression data in R. *Bioinformatics*. 2017; 33:1179.
12. Lun A, McCarthy D, Marioni J. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with BIOCONDUCTOR. *F1000Research*. 2016; 5:2122.
13. Zheng GXY, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017; 8:14049.
14. Coifman RR, et al.Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci*. 2005; 102:7426–31.
15. Amir EAD, Davis KL, Tadmor MD, et al. VISNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol*. 2013; 31:545–52.
16. Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*. 2015; 31:2989–98.
17. Angerer P, et al. DESTINY: diffusion maps for large-scale single-cell data in R. *Bioinformatics*. 2015; 32:1241.
18. Reingold EM. Graph drawing by force-directed placement. *Softw Pract Exp*. 1991; 21:1129–64.
19. Csardi G, Nepusz T. The Igraph Software Package for Complex Network Research. *Interjournal Compl Syst*. 2006; 2006:1695.
20. Weinreb C, Wolock S, Klein A. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *bioRxiv*. 2017.
21. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech*. 2008; 2008: P10008.
22. Levine JH, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*. 2015; 162:184–97.
23. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics.* 2015; 31:1974–80.
24. Huber, W, et al. Orchestrating high-throughput genomic analysis with BIOCONDUCTOR. *Nat Methods*. 2015; 12:115–21.
25. Pedregosa F, et al. SCIKIT-LEARN: machine learning in Python. *J Mach Learn Res*. 2011; 12:2825–30.
26. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology.* 2018;19(1):1-5
27. M. Canesche, L. Bragança, O. P. V. Neto, J. A. Nacif and R. Ferreira, "Google Colab CAD4U: Hands-On Cloud Laboratories for Digital Design," 2021 *IEEE International Symposium on Circuits and Systems (ISCAS),* 2021, pp. 1-5.
28. Harris CR, Millman KJ, Van Der Walt SJ, et al. Array programming with NumPy. Nature. 2020 Sep;585(7825):357-62.

29. McKinney W, others. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference.2010. p. 51–6.
30. Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering. 2007; 9(03):90-5.
31. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. anndata: Annotated data. *bioRxiv*. 2021.
32. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports. 2019;9(1):1-2.
33. Waskom ML. Seaborn: statistical data visualization. Journal of Open-Source Software. 2021;6(60):3021.
34. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016 ;374(2065):20150202.
35. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* preprint arXiv:1802.03426. 2018.
36. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature biotechnology*. 2021; 39(2):156-7.
37. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports.* 2019;9(1):1-2.