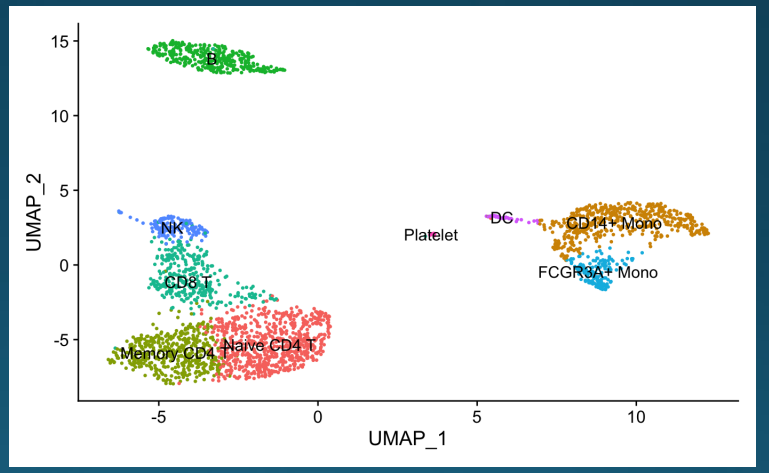
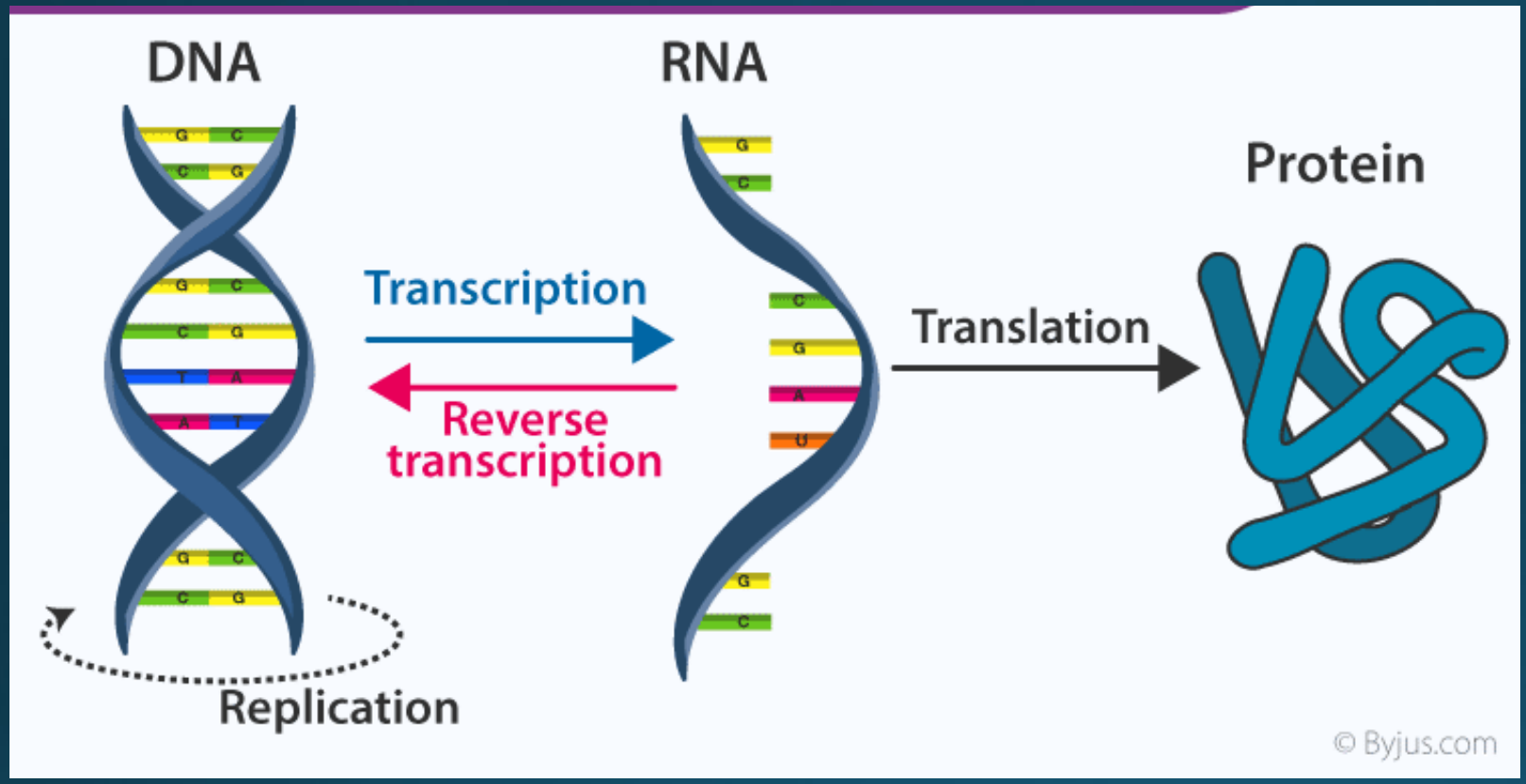
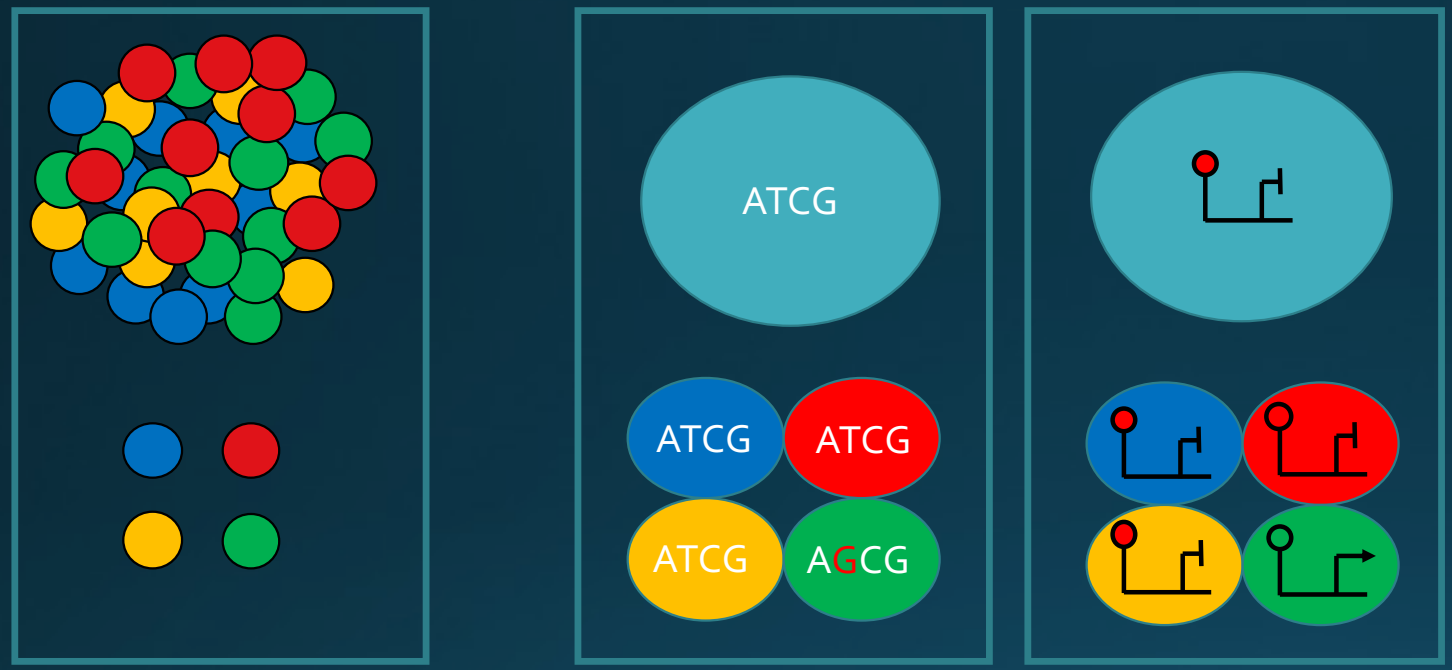
Analyzing the single cell RNA sequencing (sc-RNASeq) data and assigning cell type identity

Nitin Mahajan
DSC 680





RNA-Sequencing

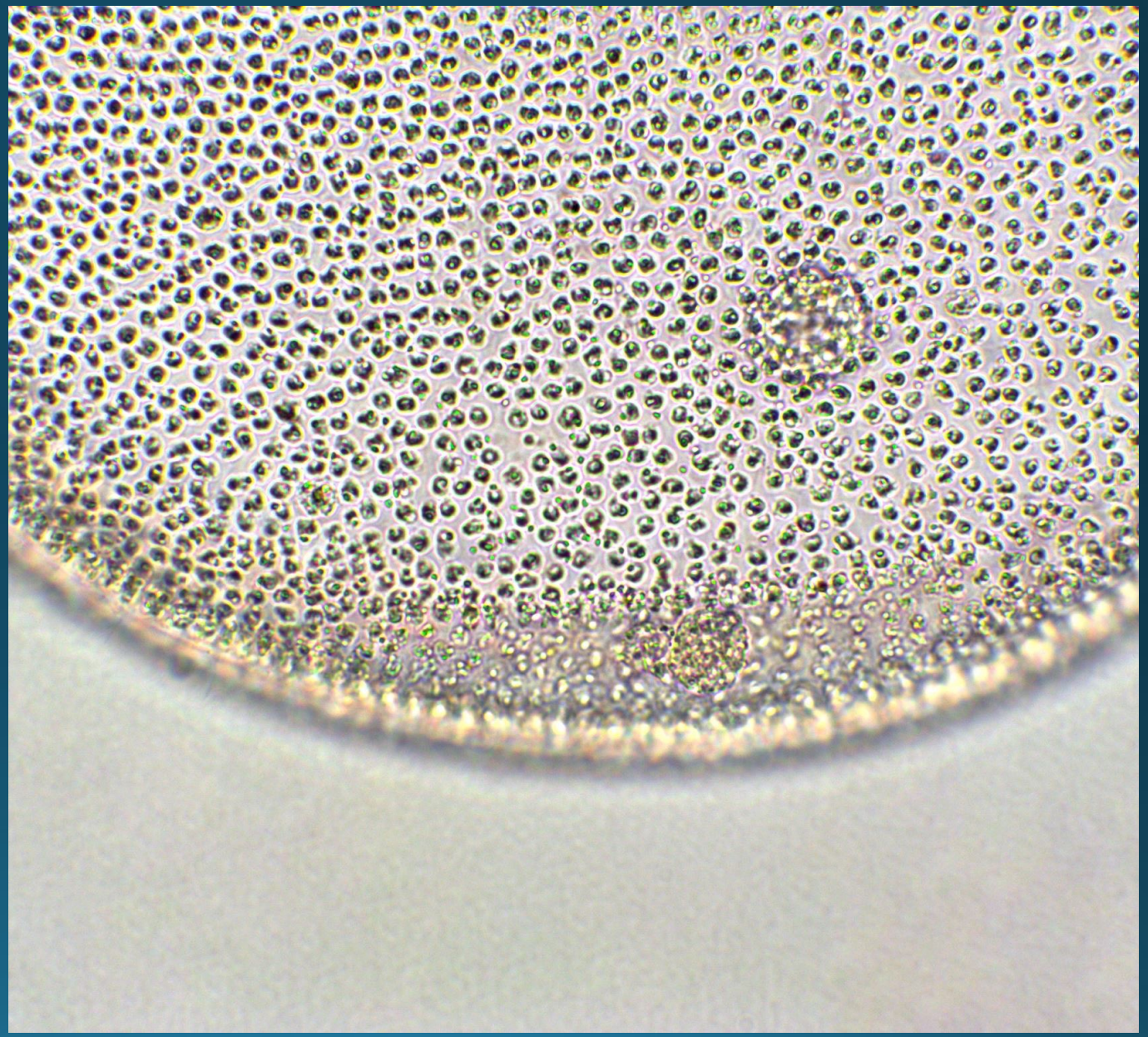


Bulk vs Single Cell

- ***What we see is not actually what we have***
  - Bulk studies averages the cells expression
  - Inappropriate for the rare population studies
  - Bulk Analysis misses rare single cells
  - Identification and analysis of rare cell types in heterogeneous cell population

Aim

Identification and marking the cell populations in mixed cell population

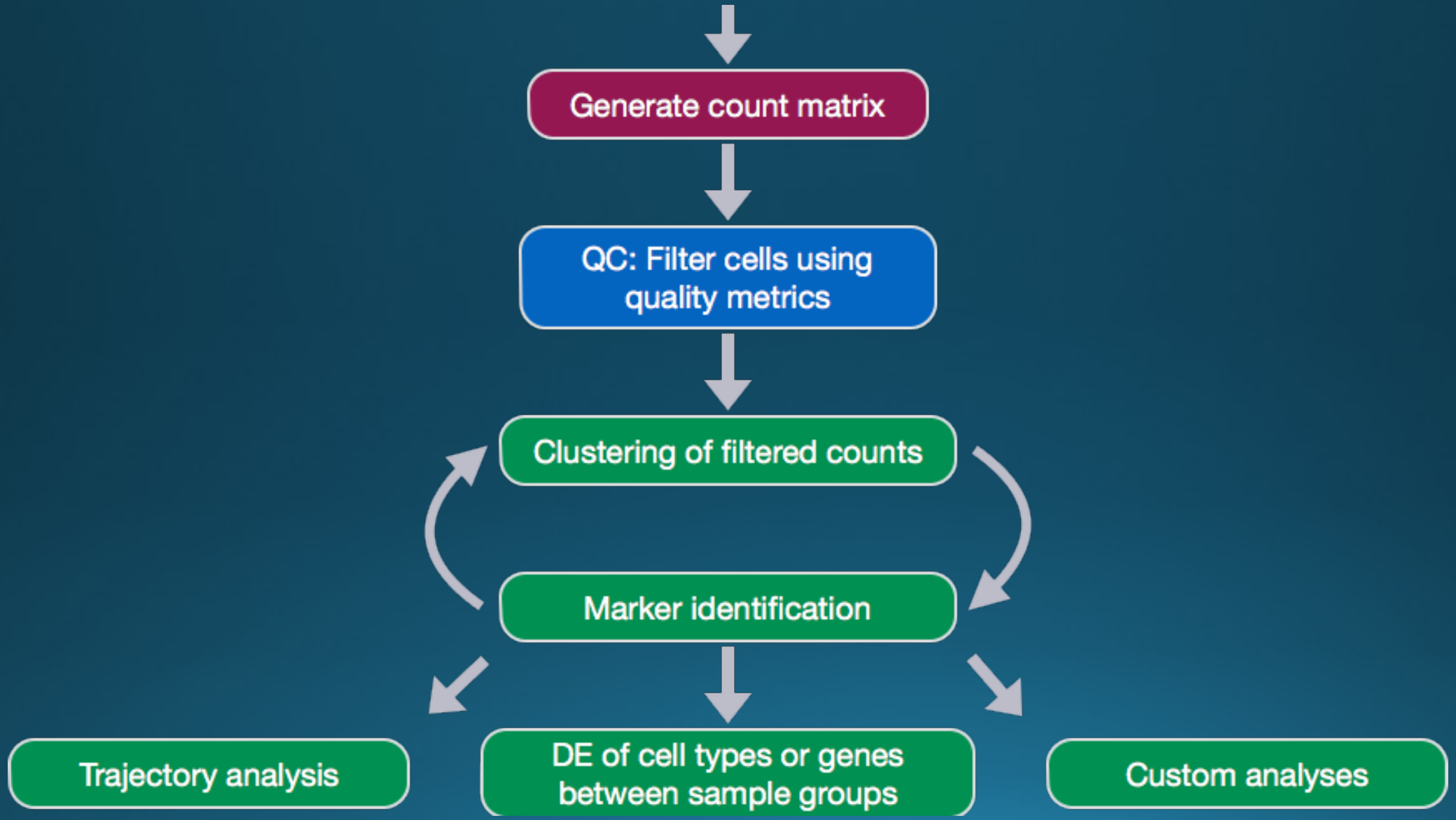


Data Source

- 10xGenomics - https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz

- Ethical Statement – There is no personal information is attached to this data set and the source of the data is duly acknowledged.

Sequencing Data

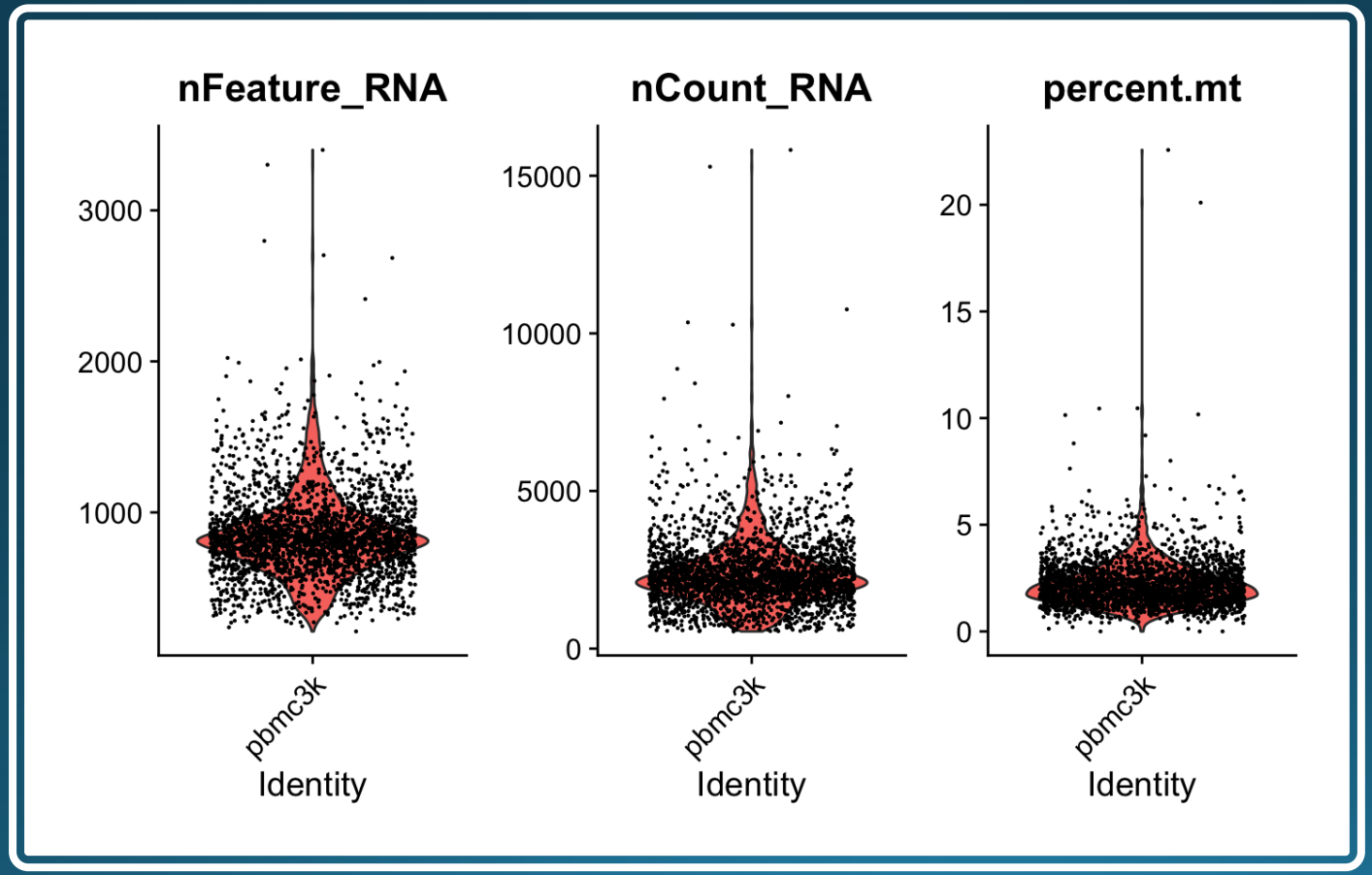


Exploratory Data Analysis

- 2,700 single cells
- Each cell is treated as row and the expression of different genes are treated as columns
- To get a quality data, we initially set a cut off, where we decided to take the cells (min = 3 cells) have the expression of minimum 200 features (genes)
- Count matrix

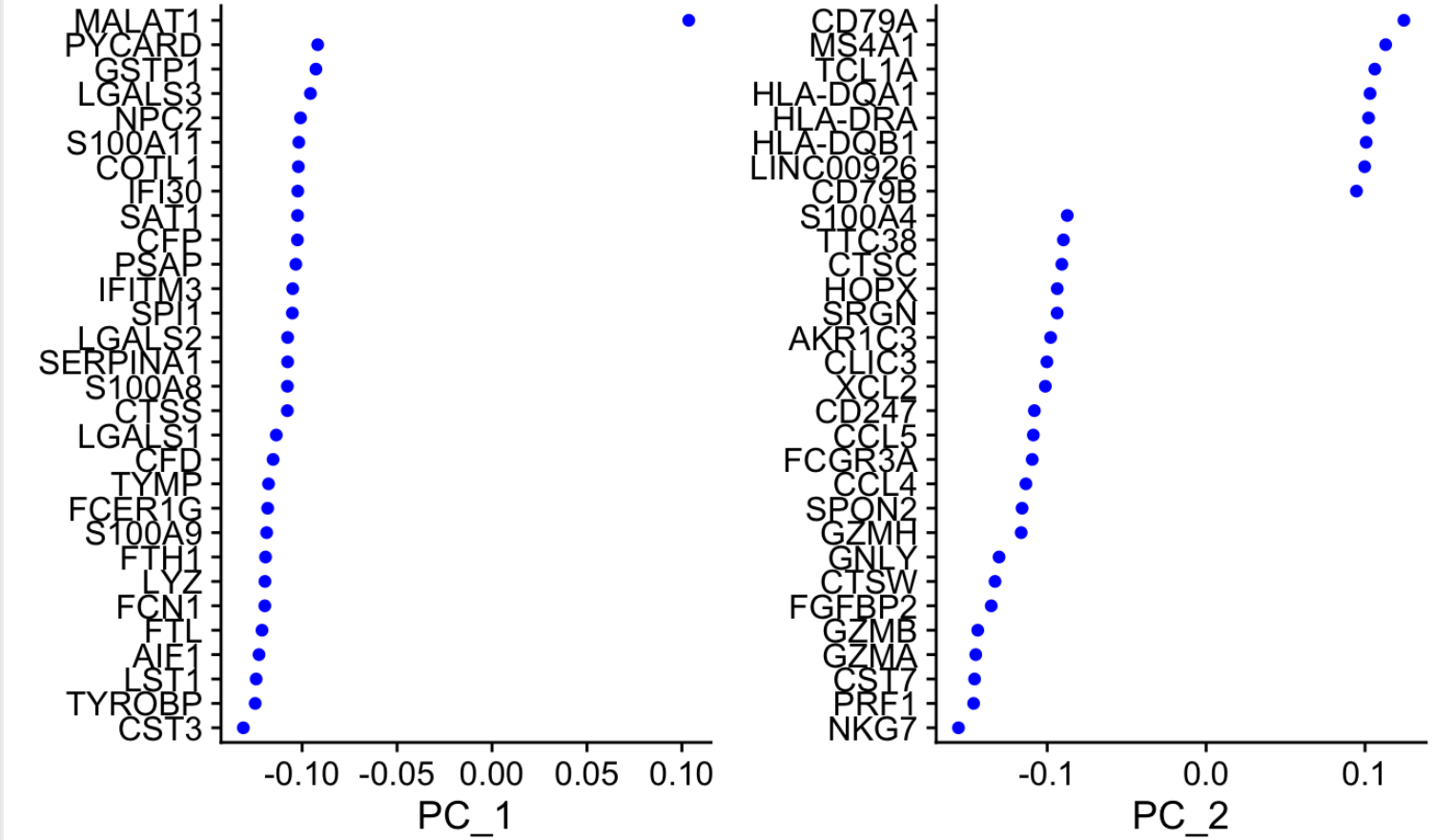
Visualization of data after QC metrics

- unique feature counts over 2,500 or less than 200 and
- >5% mitochondrial counts.



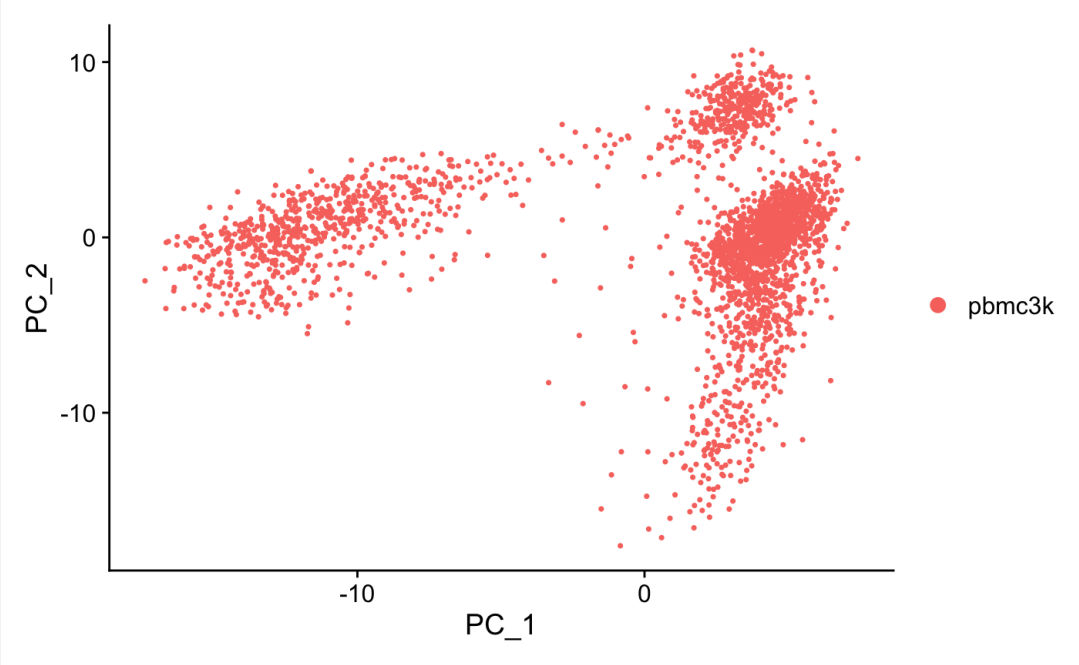
Feature Selection

- Identify a subset of features (e.g., genes) exhibiting high variability across cells,
- For each gene, the variance of standardized values across all cells was computed .
- This variance represents a measure of single-cell dispersion after controlling for mean expression, and we use it directly to rank the features.
- The *FindVaribaleFeatures()* function in the Seurat was used to do the feature selection.

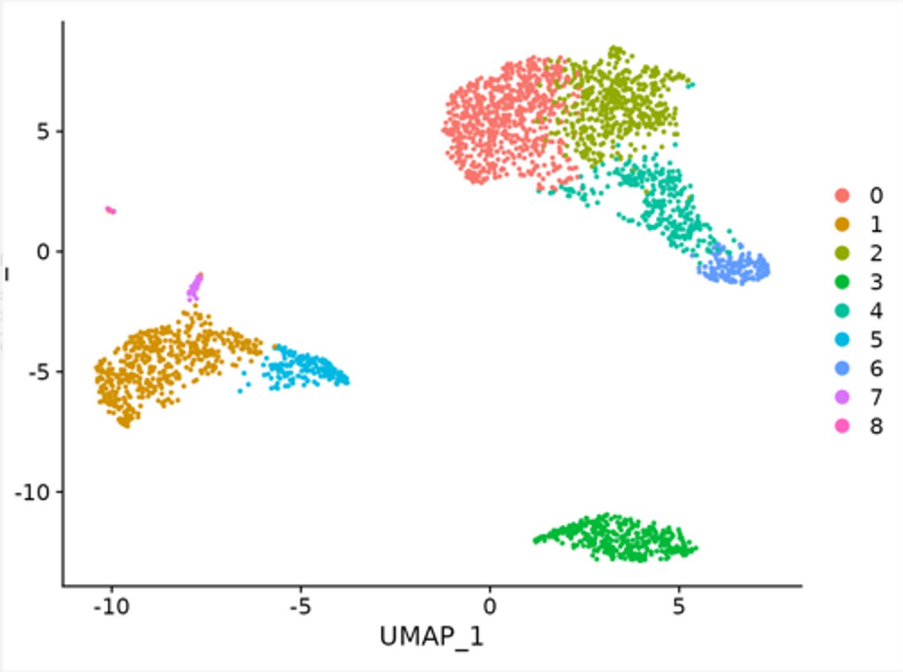


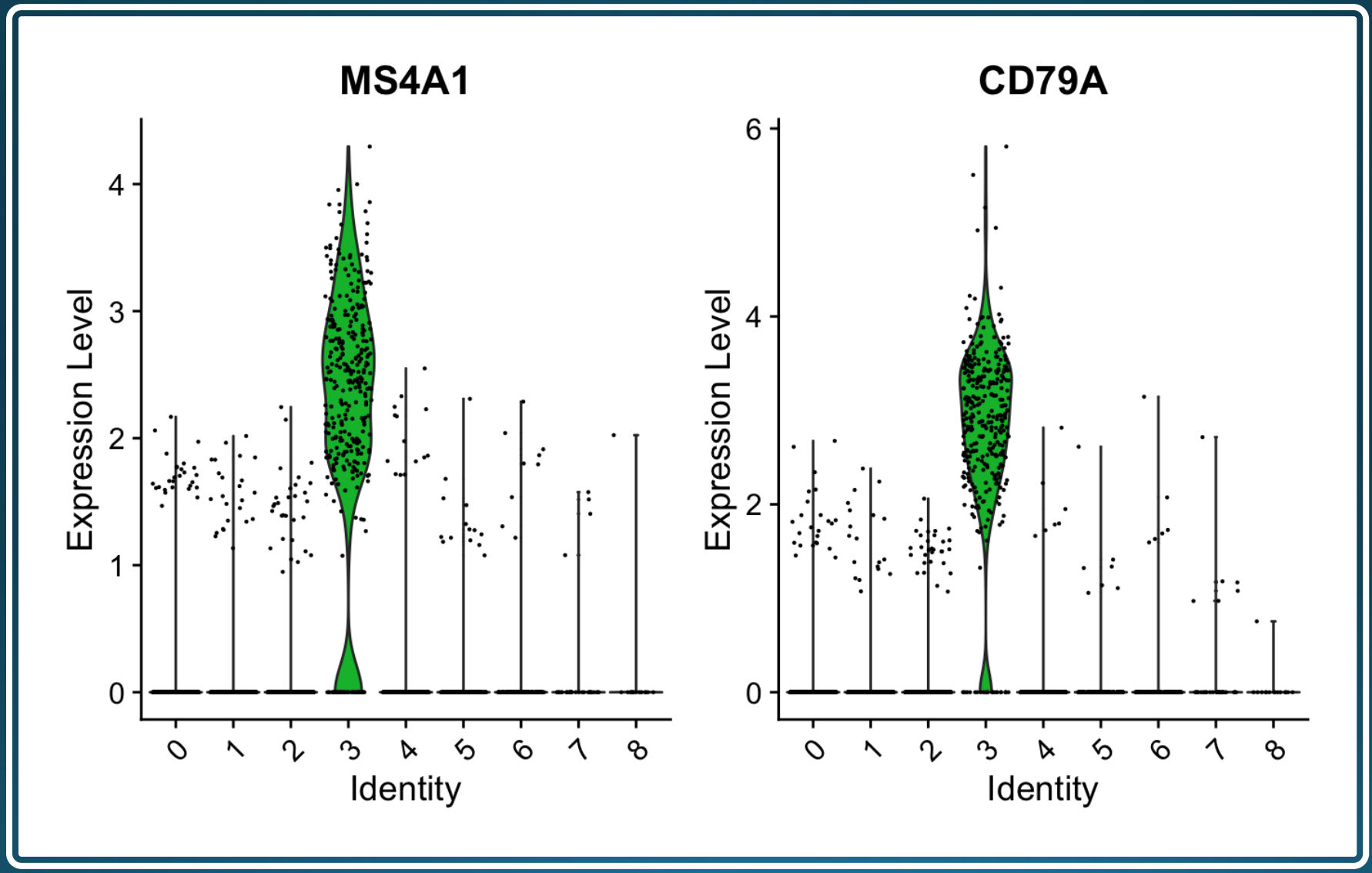
Dimension Reduction

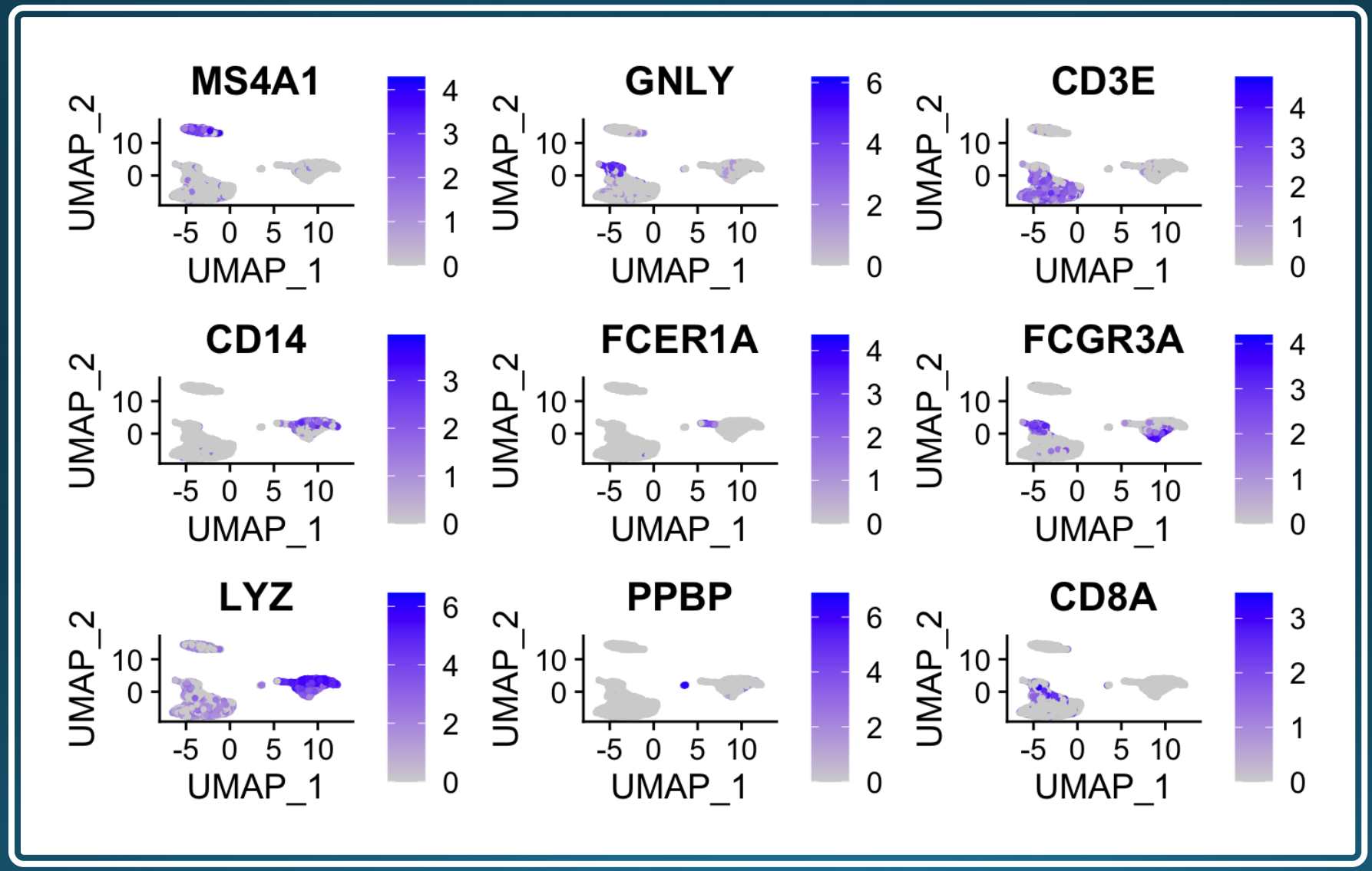
- Linear graph-based method (PCA)
- Preserve Global Representation



- Non-linear, graph-based methods (tSNE/UMAP)
- Preserve Local Representation

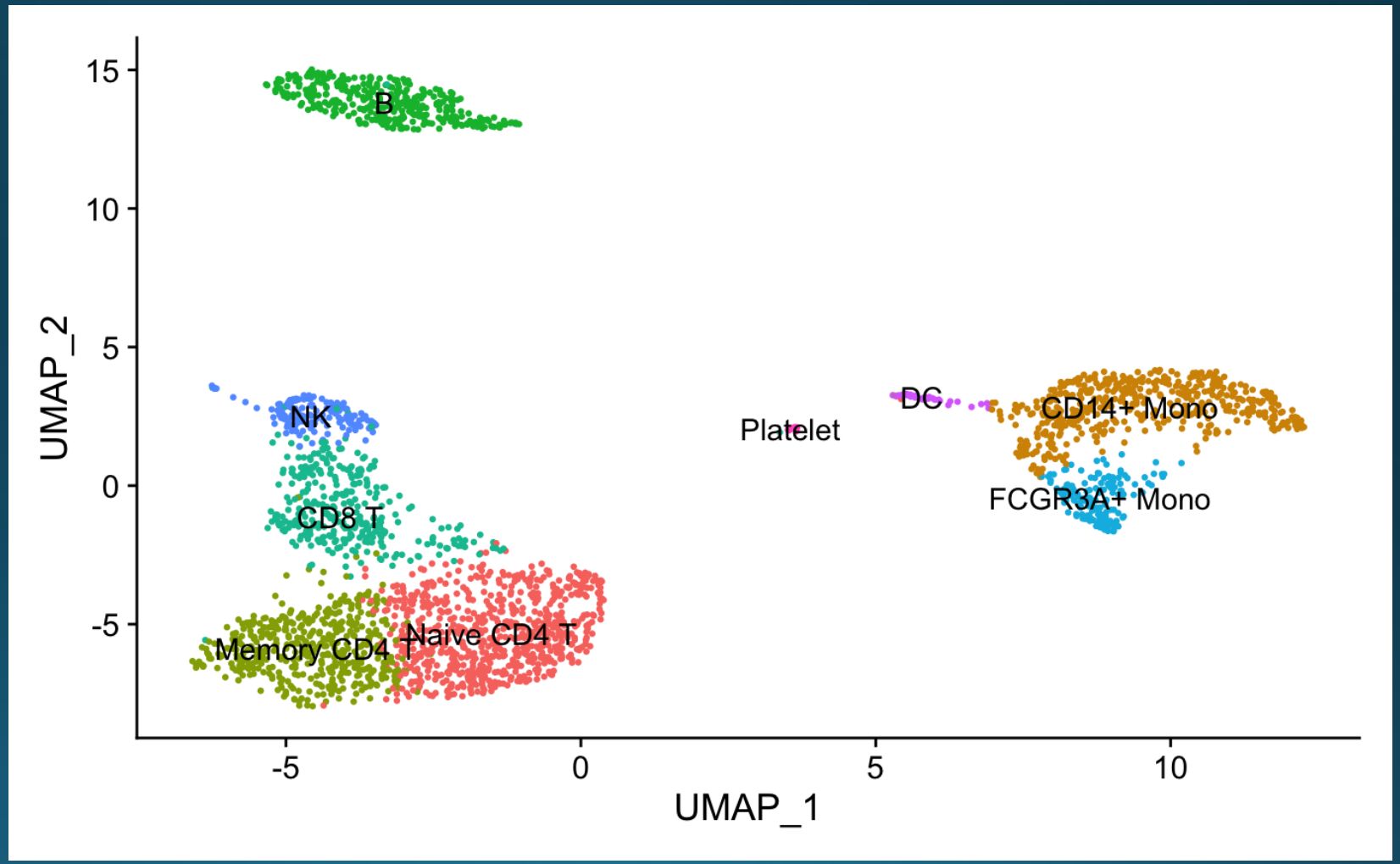



MS4A1
CD79A
Expression Level
Identity
0
1
2
3
4
5
6
7
8


MS4A1
GNLY
CD3E
CD14
FCER1A
FCGR3A
LYZ
PPBP
CD8A
UMAP_2
UMAP_1
10
0
-5
0
5
10
4
3
2
1
0
6
4
2
0
4
3
2
1
0
3
2
1
0
4
3
2
1
0
4
3
2
1
0
6
4
2
0
6
4
2
0
3
2
1
0

Finding differentially expressed features

Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet



Assumptions

- Minimum 3 cell expressing the more than 2000 features are normal
- A representation of of all humans, however

Limitations

- The present data set is from a normal healthy person, however, data from different diseases and personals would be interesting to explore and comparison.
- Such techniques are expensive and require special NGS facility, high efficiency computing facility.

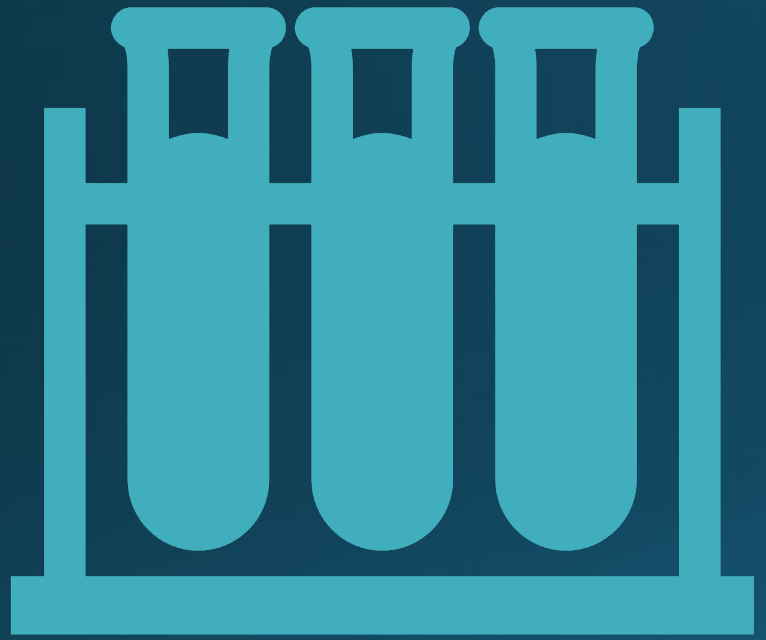
Challenges

- Selection of features and set the cut off for feature selection
- Deleting the mitochondrial genes
- We used the canonical genes to mark the cell populations, however, that would vary from person to person and from disease to disease, and therefore, we might need to tweak the cutoffs and cell specific markers

Future Use and Recommendations

- Comparison of normal to disease or another person sample would be interesting to explore.
- Identify the novel cell population in the different cells which will be helpful to understand the pathobiology and to generate the novel biomarkers for disease identification

Deployment



- The pipeline can be deployed to analyze normal healthy PBMC samples, however, with change in source of the sample would affect the characterization of the cell population. There will be change in the gene expression and therefore can be updated from time to time.
- Caution must be extrapolated if data is from another species.

Conclusion

Using feature selection method, dimensions reduction and cluster analysis, we identify the different cell type in the peripheral blood mononuclear cells (PBMCs). Based on prior knowledge of cell specific markers, different cell population have been successfully marked.

Appendix (Supporting information)

- Contributor, T. (2018, November 17). *dimensionality reduction*. WhatIs.com. Retrieved 22 October 2022, from https://www.techtarget.com/whatis/definition/dimensionality-reduction
- GeeksforGeeks. (2022a, January 14). *Difference between PCA VS t-SNE*. Retrieved 22 October 2022, from https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/
- GeeksforGeeks. (2022b, September 27). *Introduction to Dimensionality Reduction*. Retrieved 22 October 2022, from https://www.geeksforgeeks.org/dimensionality-reduction
- Haque, A. (2017, August 18). *A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications - Genome Medicine*. BioMed Central. Retrieved 22 October 2022, from https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-
- Li, X. (2021, November 15). *From bulk, single-cell to spatial RNA sequencing*. Nature. Retrieved 22 October 2022, from https://www.nature.com/articles/s41368-021-00146-0?error=cookies_not_supported&code=b6892f1d-ebeo-4cf7-ba24-5c74981e5e25
- *Seurat - Guided Clustering Tutorial*. (n.d.). Retrieved 22 October 2022, from https://satijalab.org/seurat/articles/pbmc3k_tutorial.html
- Stuart, T. (2019, June 13). *Comprehensive Integration of Single-Cell Data*. Cell. Retrieved 22 October 2022, from https://www.cell.com/cell/fulltext/S0092-8674(19)30559-8?_returnURL=https://linkinghub.elsevier.com/retrieve/pii/S0092867419305598?showall=true
- Wikipedia contributors. (2022, August 19). *Scree plot*. Wikipedia. Retrieved 22 October 2022, from https://en.wikipedia.org/wiki/Scree_plot