BT3041: **Analysis and Interpretation of Biological Data**

Term Project Report

Gene Expression Profile at Single Cell Level of Hippocampal Organoids

Anirudh Rao

Govinda K

Nithish Gokul

Sanyukta Gedam

Shankhanabha Ghosh

Shreeharsha Bhat

INDIAN INSTITUTE OF TECHNOLOGY MADRAS • सिद्धिर्भवति कर्मजा •

Contents

Abstract

Hippocampal organoids can be used to study human brain development and function, and to model diseases and screen drugs. It is important to ensure that the organoid resembles the *in vivo* hippocampus as closely as possible, in terms of cell types present and gene expression patterns. In this project, we analyse the results of a single cell RNA-Seq (scRNA-Seq) experiment conducted on a hippocampal organoid grown for 70 days. Using unsupervised learning and dimensionality reduction techniques, we uncover the cell types present in the organoid and the unique patterns in their transcriptomic signatures. We also compare the results of two clustering algorithms on the same dataset.

Keywords – scRNA-Seq, organoid, hippocampus, cell subtyping, clustering

Introduction

The dataset for this project was obtained from the Gene Expression Omnibus. It was published on April 20, 2024 by Yan Wu's group at the Southern University of Science and Technology (SUSTech) in Shenzhen, China, with the accession number GSE264363 [1]. It is yet to published in a research paper.

Hippocampal organoids were created using human induced pluripotent stem cells (IPSCs) with the help of various growth factors (Figure 1). After 70 days of development, the organoid was dissected into fragments to obtain single cells.

The gene expression profiles of individual cells in the sample were characterised by scRNA-Seq (Figure 2). Overall, 2,065,299 cells were sequenced and the expression patterns of 33,538 genes was found.

The data from the experiment was stored in 3 files – `barcodes`, `features`, `matrix`. The `barcodes` file includes information about the cells. The `features` file includes information about the genes. The `matrix` file describes the genes expressed in each cell (Figure 3).

The objective of this analysis is to uncover the cell types present in the organoid and the unique patterns in their transcriptomic signatures.

Methods and Results

All analysis was performed using libraries in Python or in R using the Seurat library [2] developed by Rahul Satija's group at New York University.

Distribution of gene expression patterns

We first analysed the distribution of the number of genes being expressed in the cells. We then analysed the distribution of the number of cells that expressed each of the genes.

Both the distributions showed a decaying behaviour as the number of genes / cells increased (Figure 4). We found that, on average, the cells in the sample express 31 genes, and on average, a given gene is expressed in 1,493 cells.

Quality control

To limit the size of the data being analysed and to minimise outliers, we dropped those cells that expressed fewer than 30 genes. We also dropped those genes that were expressed in fewer than 1,500 cells. This reduced the number of cells to be analysed to 83,262 and the number of genes to be analysed to 6,488. This was used to create the final gene expression matrix in the form of a “Seurat object”.

Variably expressed genes

To reliably identify cell types using unsupervised learning methods, it is important to identify those genes whose expression is highly varied across individual cells in the sample. These can be used as “features” for data analysis. When we analysed the variance in gene expression across all the genes, we found that the average variance in expression was 1.068. The number of genes with variance greater than or equal to the average variance was 1,473.

Using the Seurat library, we found the top 1,500 genes that had the most variance in their expression. We performed the same analysis with and without the quality control measures mentioned earlier (Figure 5).

When variably expressed genes are identified without performing quality control, collagen genes (COL3A1, COL1A2, COL1A1) are shown to be highly variable. This is unusual as collagen is considered a housekeeping gene. We found that these collagen genes were expressed in fewer than 1,493 cells, which is lower than average.

When quality control measures are used, the top 3 highly variable genes are found to be:

- FN1 – Fibronectin 1 – Involved in neuroprotection [3]
- TTR – Transthyretin – Involved in thyroxine and retinol transport, protects against fibril formation in neurons [4]
- PTGDS – Prostaglandin D2 synthase – Synthesises a neuromodulator in the central nervous system [5]

These give us more confidence that the organoid has developed neuronal cells.

Principal component analysis

To reduce the number of features being analysed, we performed linear dimensionality reduction using principal component analysis (PCA) on the 1,500 variable genes identified. First, we scaled the data such that the mean expression of each gene across cells is 0, and the variance across cells is 1. Then, we identified 20 principal components (PCs).

We then plotted the percentage explained variance of each of the 20 PCs against the cumulative percentage explained variance, which ranges from 0 to 100% (Figure 6). PC-1 explains slightly more than 20% of the variance. This gradually decreases as the rank of the PC increases from 1 to 20.

We found that the change in percentage explained variance is lesser than 0.1% when moving from PC-11 to PC-12. Thus, we decided to consider the first 11 PCs for further analysis. Plotting the data points using the first 2 PCs revealed a good spread with some clusters (Figure 7).

The top 3 contributing genes to PC-1 were found to be RBM25, NEAT1, and ANKRD11. RBM25 is a splicing factor found in a wide range of eukaryotic cells [6]. NEAT1 is a long non-coding RNA that has been shown to play a role in long-term memory in the hippocampus [5]. The top 3 contributing genes to PC-2 were found to be KCNQ1OT1, CELF4, and SSTR2. KCNQ1OT1 is a potassium ion channel with a crucial role in the central nervous system [8]. CELF4 is a well-studied RNA-binding protein that is expressed in the hippocampus [9]. The coefficients of the top genes that contribute to the first 2 PCs were then visualised (Figure 8).

Uniform Manifold Approximation and Projection

We then used the first 11 PCs to perform non-linear dimensionality reduction to 2 dimensions using Uniform Manifold Approximation and Projection (UMAP). The algorithm was able to identify 7 clusters (Figure 9). We found that two of these clusters (Cluster 0 and Cluster 4) had significant overlaps. The other clusters were relatively separated. Each cluster corresponds to a particular cell type.

We then varied the `n.neighbors` parameter of the algorithm from its default value of 30. We observed minor structural changes in the UMAP when the value of the parameter was changed to 15 (Figure 10) and 60 (Figure 11). However, the overall clustering was unaffected.

k-Means clustering

We also used classical unsupervised learning by performing k-means clustering on the first 11 PCs. We set the number of clusters to be 7 to match the number of clusters identified by UMAP. The algorithm was able to successfully identify 7 clusters with only some minor overlaps. This was visualised using the first 2 PCs (Figure 12).

The Pearson correlation and Spearman correlation between the clusters assigned by UMAP and those by k-means were found to be 0.35 and 0.43, respectively. This indicates a relatively poor agreement between the two methods.

Biomarker discovery and cell subtyping

The final part of our analysis involved the identification of differentially expressed genes (i.e., biomarkers) in the 7 cell types / clusters identified by UMAP using `n.neighbors` = 30. For each cluster, we identified genes that had the highest average fold change with respect to the other clusters. Essentially, we found genes that were significantly overexpressed (with p-value < 0.05) in a given cluster when compared to the other clusters. The in-built algorithm in Seurat uses the Wilcoxon rank sum test to identify such biomarkers. The results are visualised in Figure 13 and tabulated in Table 1.

Using the Human Protein Atlas [10], we attempted to identify the specific cell types present. We used the single cell type specificity information provided by the Atlas for each of the biomarkers identified. We were able to identify the following cell types – horizontal cells, Muller glia cells, astrocytes, bipolar cells, Schwann cells, microglial cells, and excitatory neurons. These are known neuronal cell types.

Discussion

By analysing scRNA-Seq data from a hippocampal organoid, we were able to identify the genes that were variably expressed across the different cells in the sample. Using this, we could reliably identify 7 different cell types that were present in the organoid, along with their associated biomarkers. The genes and cell types identified have important roles to play in hippocampal development and function.

Future Work

In this work, we have analysed 70 day organoids. There is another sample from 81 day organoids that can be analysed and compared with the results presented here. This can provide information about the relation between the duration of development and types of cells that have differentiated, as well as changes in their transcriptomes. A similar comparison can be made with the data from an *in vivo* hippocampus.

To appropriately decide the number of clusters present, silhouette score can be used as a metric. This was not possible in this project due to the lack of computational resources to compute the silhouette scores of 83,262 cells.

Acknowledgements

We would like to thank the course instructor Dr. Meiyappan Lakshmanan for his valuable inputs during this project and throughout the course. We would also like to thank the teaching assistants for their tutorial sessions and feedback on this project.

References

[1] GEO Accession viewer https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE264363

[2] Hao, Y., et al. (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology, 42(2), 293–304. https://doi.org/10.1038/s41587-023-01767-y

[3] Wang, J., et al. (2013). Neuroprotective role of fibronectin in neuron-glial extrasynaptic transmission. PubMed. https://doi.org/10.3969/j.issn.1673-5374.2013.04.010

[4] Li, X., et al. (2013). Mechanisms of transthyretin inhibition of β-Amyloid Aggregation in vitro. The Journal of Neuroscience, 33(50), 19423–19433. https://doi.org/10.1523/jneurosci.2561-13.2013

[5] Shimizu, T., et al. (1979). Prostaglandin D2, a neuromodulator. Proceedings of the National Academy of Sciences of the United States of America, 76(12), 6231–6234. https://doi.org/10.1073/pnas.76.12.6231

[6] Carlson, S. M., et al. (2017). RBM25 is a global splicing factor promoting inclusion of alternatively spliced exons and is itself regulated by lysine mono-methylation. Journal of Biological Chemistry 292(32), 13381–13390. https://doi.org/10.1074/jbc.m117.784371

[7] Butler, A. A., et al. (2019). Long noncoding RNA NEAT1 mediates neuronal histone methylation and age-related memory impairment. Science Signaling, 12(588). https://doi.org/10.1126/scisignal.aaw9277

[8] Alam, K. A., et al. Potassium channels in behavioral brain disorders. Molecular mechanisms and therapeutic potential: A narrative review. Neuroscience & Biobehavioral Reviews, 152, 105301. https://doi.org/10.1016/j.neubiorev.2023.105301

[9] Wagnon, J. L., et al. (2012). CELF4 Regulates Translation and Local Abundance of a Vast Set of mRNAs, Including Genes Associated with Regulation of Synaptic Function. PLOS Genetics, 8(11), e1003067. https://doi.org/10.1371/journal.pgen.1003067

[10] The Human Protein Atlas https://www.proteinatlas.org/

Annexure

All files and codes used for this project can be found here –

https://drive.google.com/drive/folders/1i8aaSUWss3FaolrJTfyVbbHn1R9qjVxj?usp=sharing .

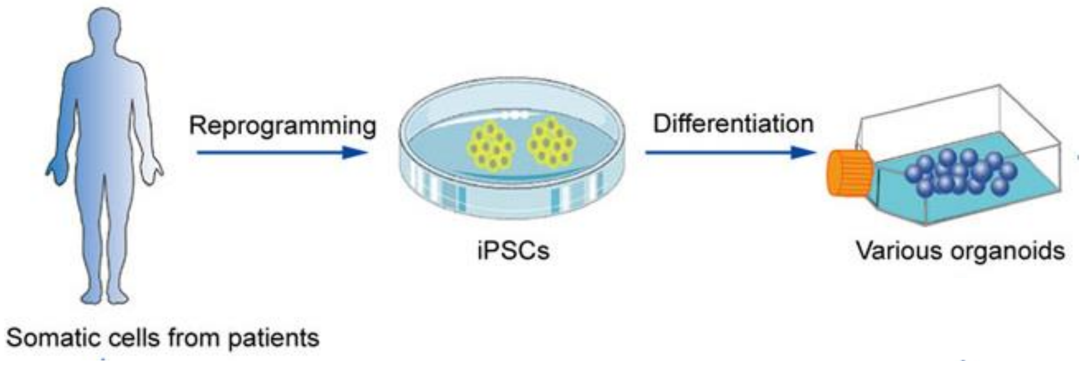


Figure 1: Creation of organoids from IPSCs

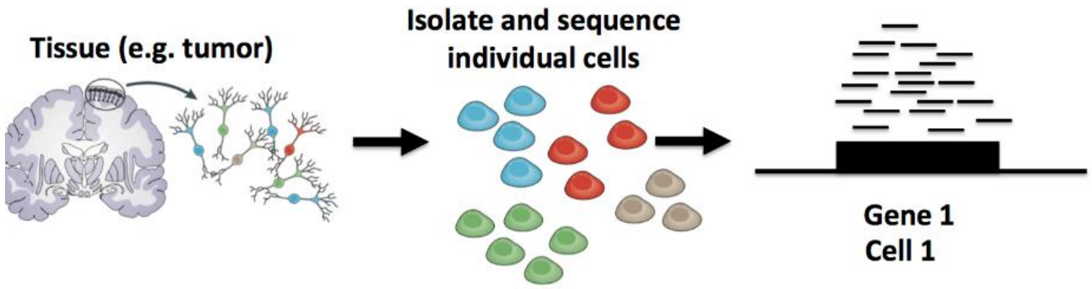
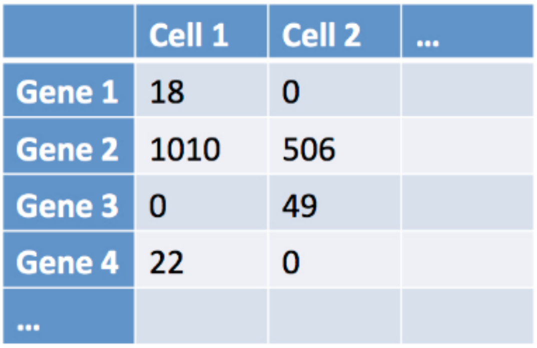


Figure 2: Principle of scRNA-Seq



	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

Figure 3: Schematic representation of the gene expression matrix

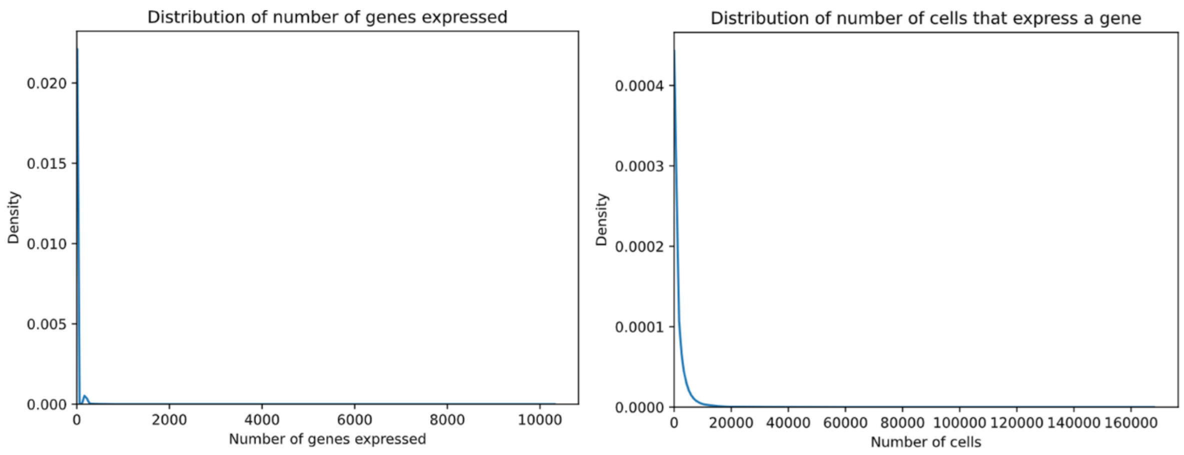


Figure 4: Distributions in expression in the 70 day organoids, shown as kernel density estimates

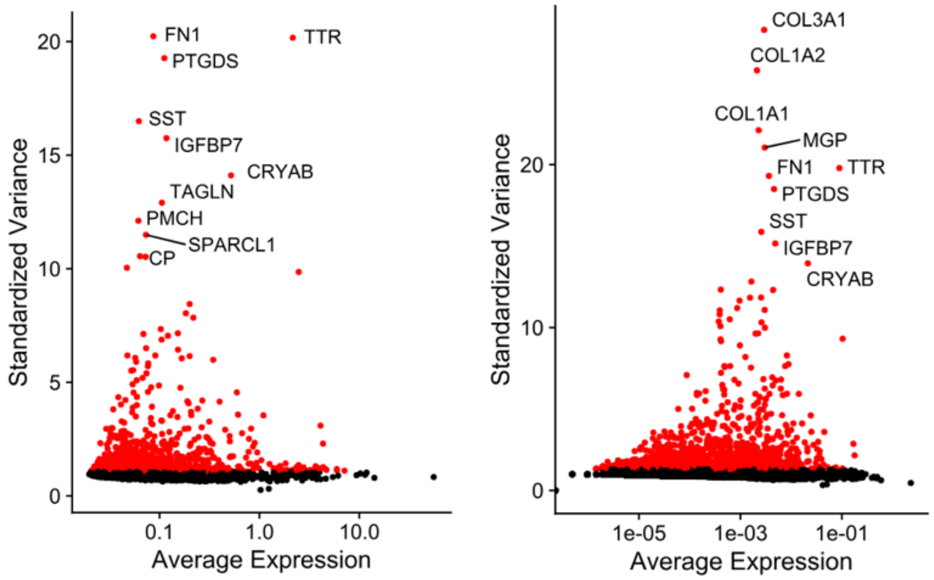


Figure 5: Variably expressed genes – with (left) and without (right) quality control

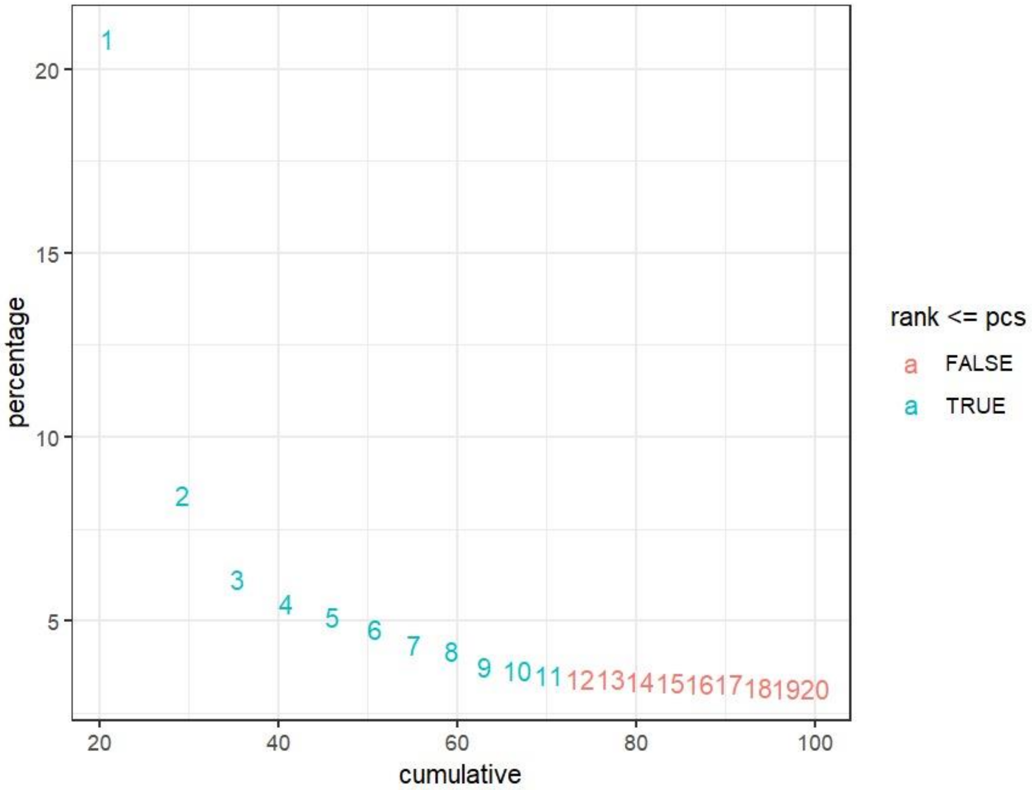


Figure 6: Percentage of explained variance of the 20 principal components

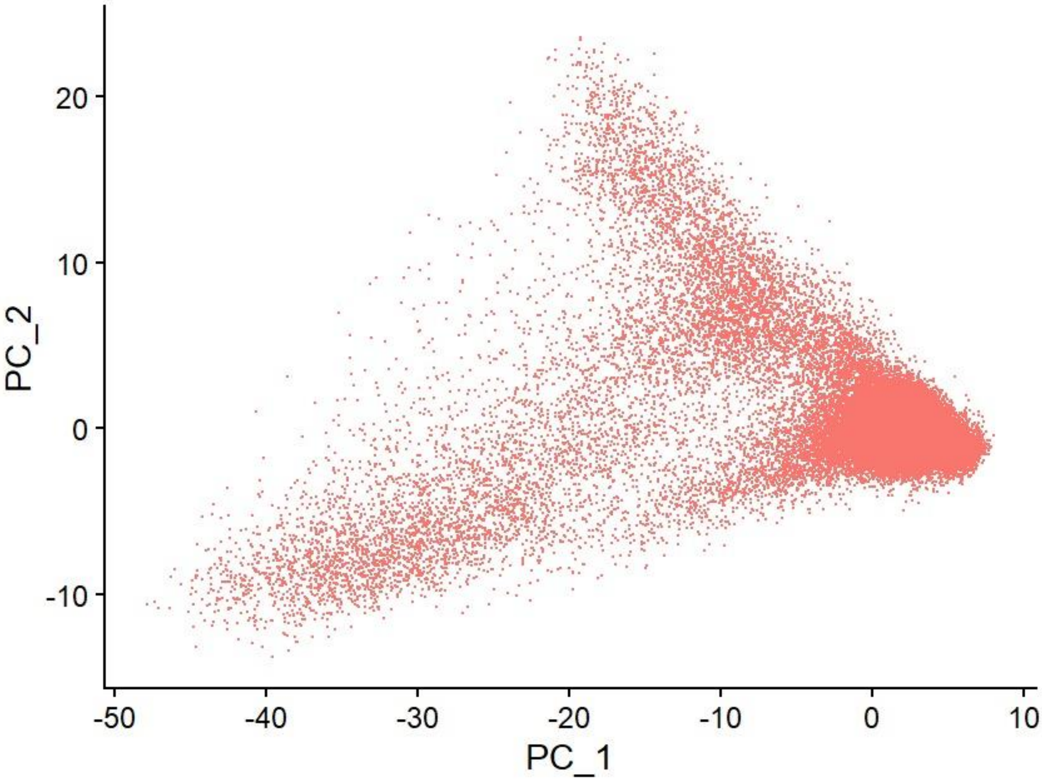


Figure 7: PCA plot of the first 2 principal components

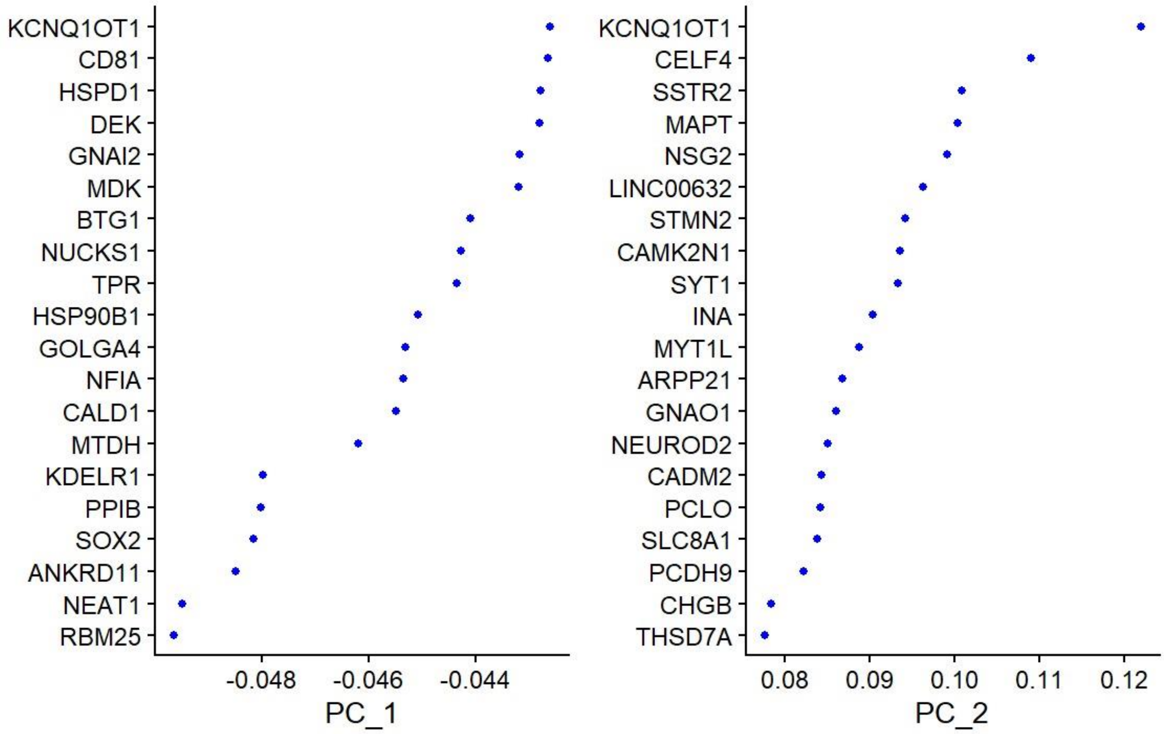


Figure 8: Loadings plot of genes contributing to the first 2 PCs

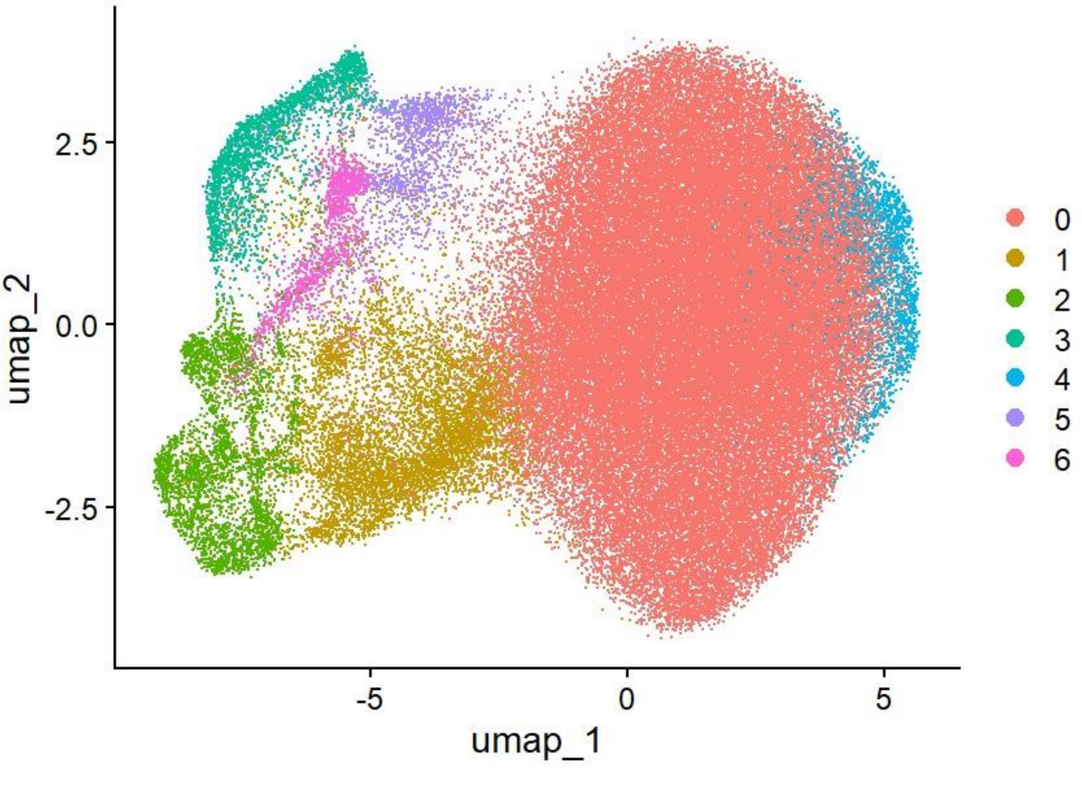


Figure 9: UMAP with `n.neighbors` = 30

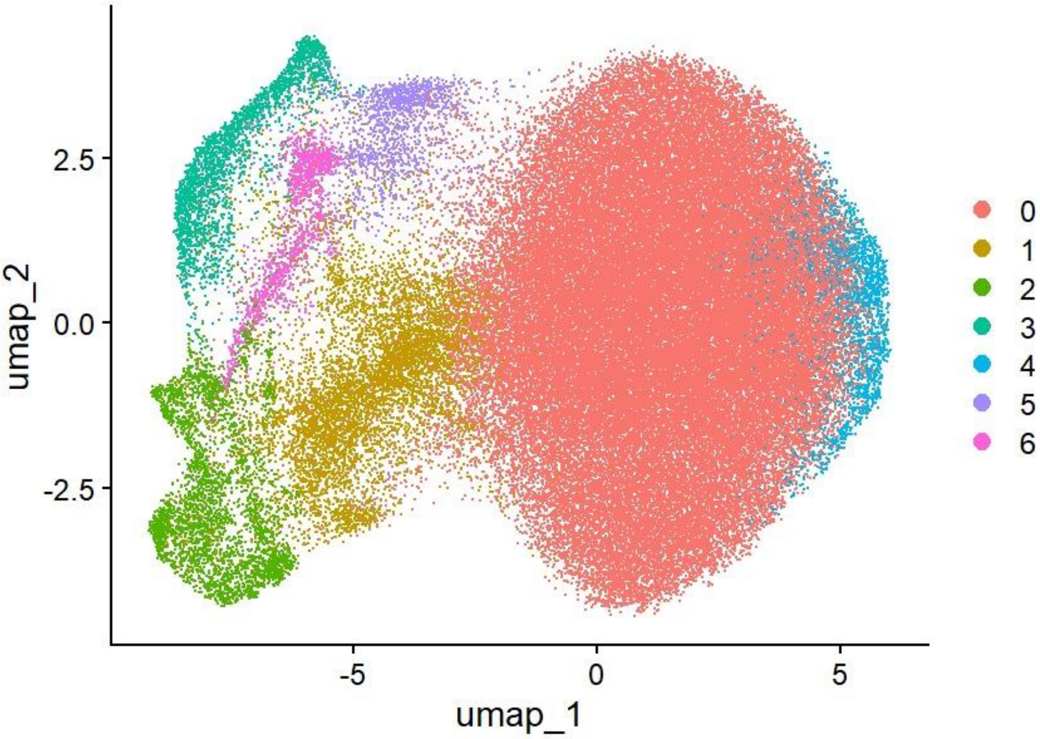


Figure 10: UMAP with `n.neighbors` = 15

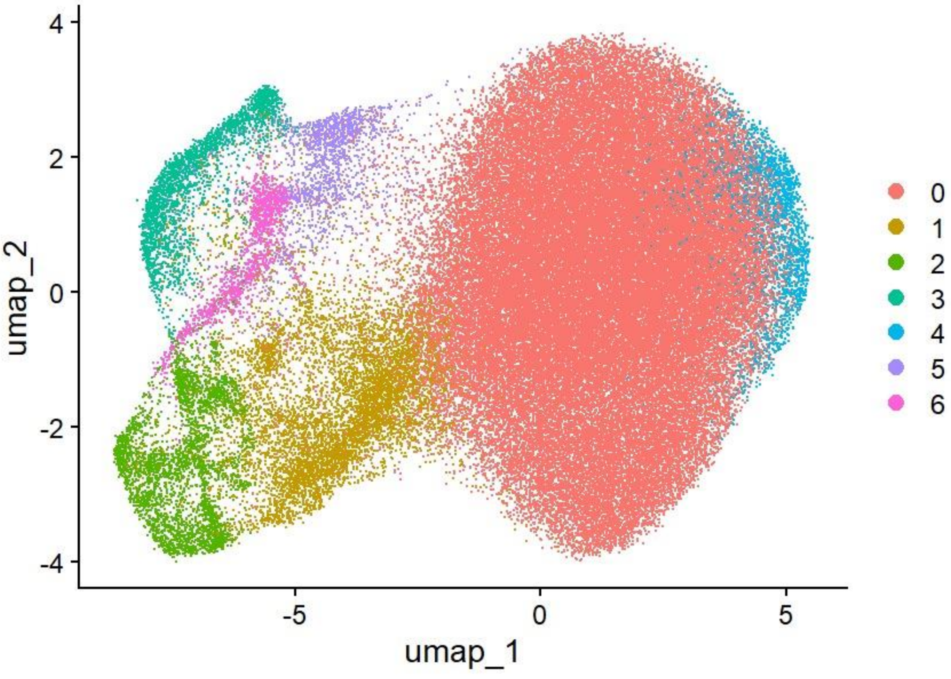


Figure 11: UMAP with `n.neighbors` = 60

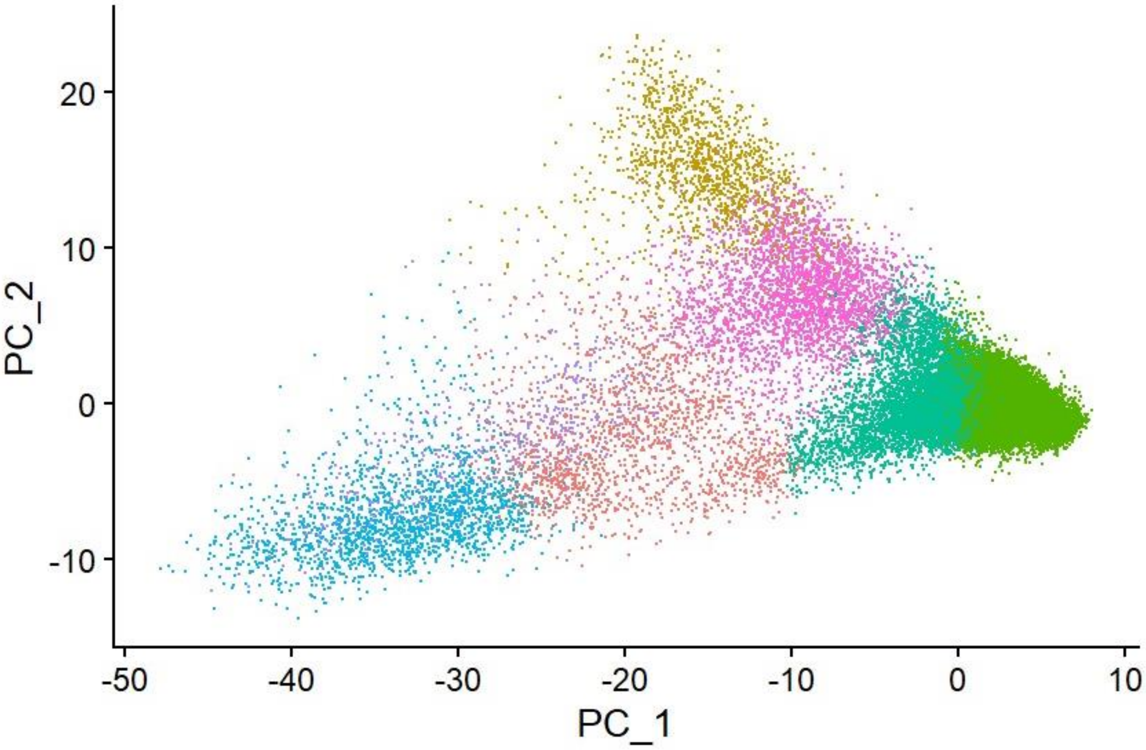


Figure 12: Clusters identified by k-means

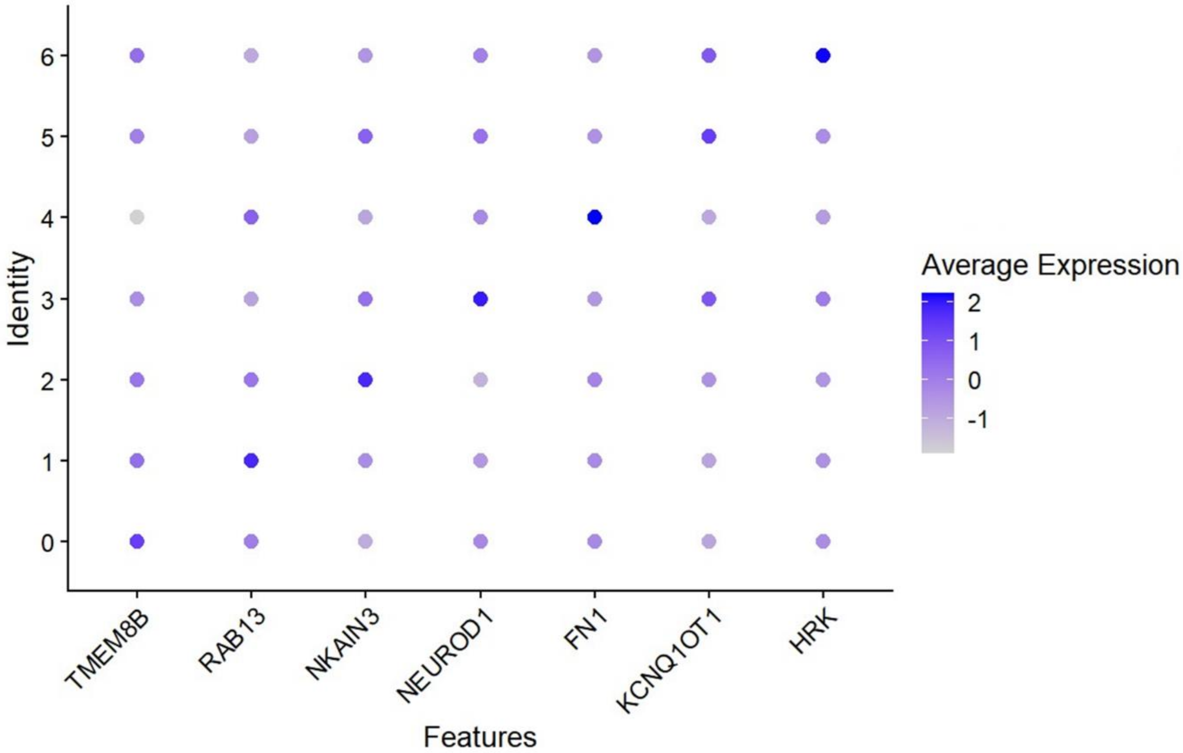


Figure 13: Biomarkers across the different cell types. A darker shade of blue implies that the expression of the gene is higher in that cluster.

Table 1

Cluster	Biomarker	Average log fold change	p-value	Possible cell type
0	TMEM8B	0.79	0	Horizontal cells
1	RAB13	1.73	7.64e-159	Muller glia cells
2	NKAIN3	1.66	0	Astrocytes
3	NEUROD1	2.11	0	Bipolar cells
4	FN1	6.12	1.24e-153	Schwann cells
5	KCNQ1OT1	2.68	0	Microglial cells
6	HRK	2.32	0	Excitatory neurons

Contributions

Name	Roll No.	Contribution
Anirudh Rao	BE21B004	Non-linear dimensionality reduction with UMAP, biomarker discovery, cell subtyping
Govinda K	BE21B016	Dataset identification, data preprocessing, quality control
Nithish Gokul	CH20B075	Dataset identification, data preprocessing, quality control
Sanyukta Gedam	BE21B035	Identification of variably expressed genes, principal component analysis, k-means clustering
Shankhanabha Ghosh	BE21B036	Identification of variably expressed genes, principal component analysis, k-means clustering
Shreeharsha Bhat	BE21B037	Non-linear dimensionality reduction with UMAP, biomarker discovery, cell subtyping

This is just a rough overview of the contributions of team members. However, it must be noted that all team members were involved in all aspects of the project.