BMEN6003_HW3

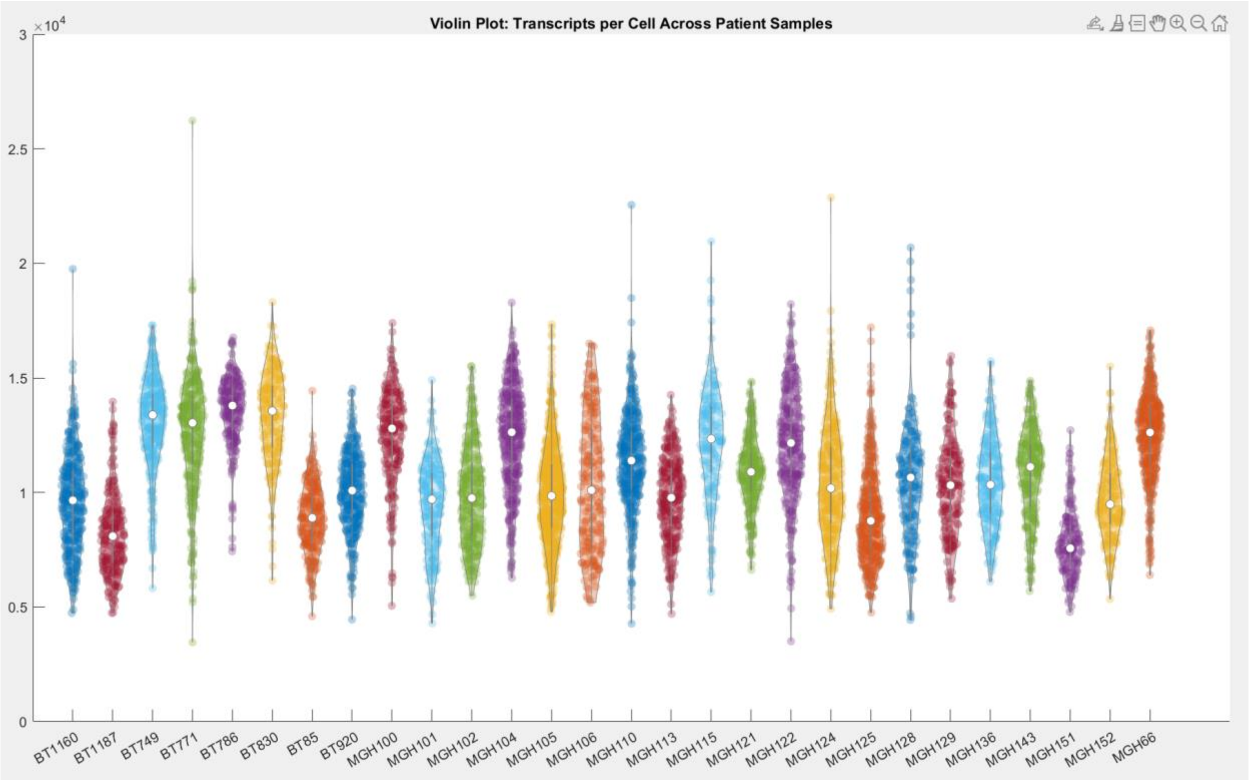
Marcus Ibrahim / Harry Kabodha / Ayman Khaleq

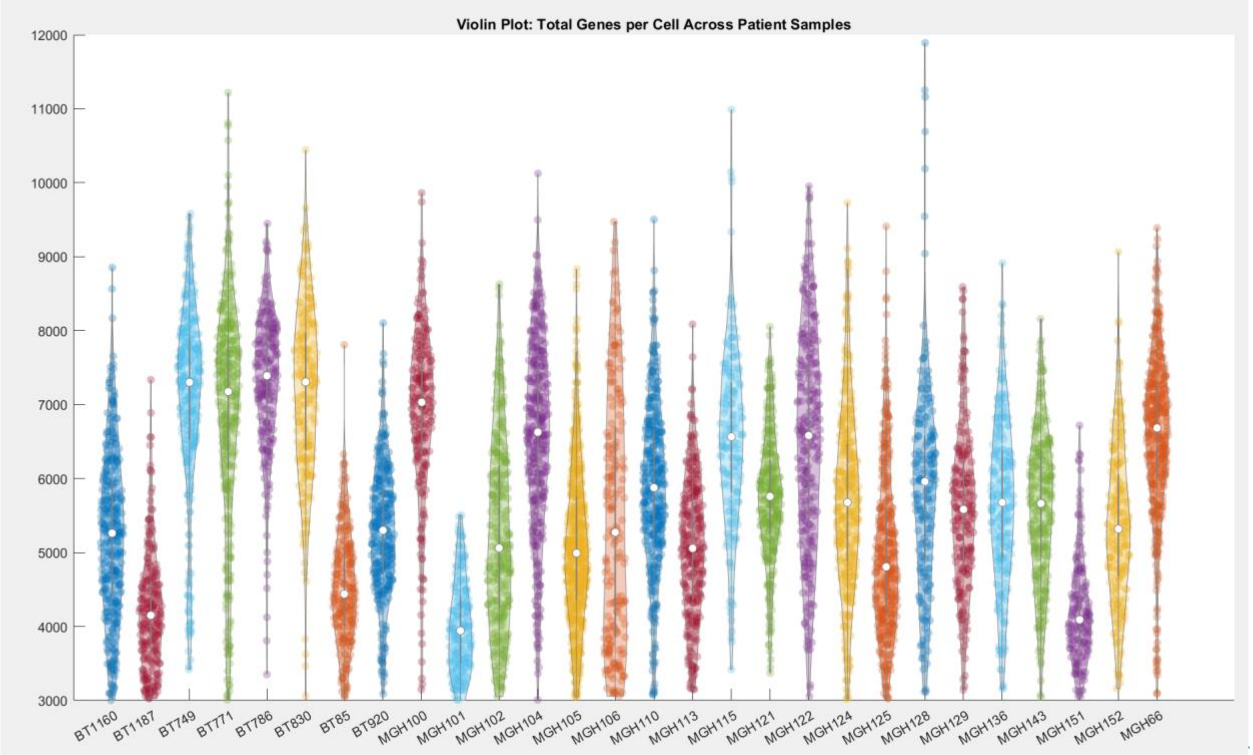
Q1-1. Each row of the expression matrix corresponds to a gene for which transcripts were quantified and each column corresponds to a cell. Examine the dimensions of your expression matrix. How many genes are in the dataset? How many individual cells were captured in the experiment? What does each individual count represent? [5 pts]

There are 23,686 genes in the dataset, as indicated by the length of the genelist and there are 7,930 cells, as indicated by the length of the barcodelist. Each individual count in the expression matrix X represents the inferred mRNA count per cell, as stated in the problem statement.

Q1-2. This dataset contains single-cell transcriptomes obtained from glioblastoma brain tumor samples across multiple patients. Determine the number of patients in your datasets. The "Sample" column, in the metadata table, refers to each patient.

There are 20 MGH variables and 8 unique BT variables giving a total of 28 different variables in the column, meaning that there are 28 unique patients in this study.


Violin Plot: Transcripts per Cell Across Patient Samples
×10^4
3
2.5
2
1.5
1
0.5
0
BT1160
BT1187
BT749
BT771
BT786
BT830
BT85
BT920
MGH100
MGH101
MGH102
MGH104
MGH105
MGH106
MGH110
MGH113
MGH115
MGH121
MGH122
MGH124
MGH125
MGH128
MGH129
MGH136
MGH143
MGH151
MGH152
MGH66


Violin Plot: Total Genes per Cell Across Patient Samples
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
BT1160
BT1187
BT749
BT771
BT786
BT830
BT85
BT920
MGH100
MGH101
MGH102
MGH104
MGH105
MGH106
MGH110
MGH113
MGH115
MGH121
MGH122
MGH124
MGH125
MGH128
MGH129
MGH136
MGH143
MGH151
MGH152
MGH66

Q1-3. As a pre-processing step, normalize your expression matrix using the function sc_norm() with normalization method type ‘deseq’.

Done check code

Q1-4. The graph below depicts the total number of transcripts (n.umi) for each cell barcode in a single-cell experiment (cell_rank refers to cell barcodes ranked from the largest to lowest number of total transcripts per barcode). Which area of this graph is more likely to contain low-quality cells (i.e., cell barcodes corresponding to ruptured cells or cell debris)? Which area of this graph is more likely to contain doublets or multiplets (i.e., multiple cells labeled by one cell barcode)?

The area at the end of the graph where the count is lowest is the area that contains the low quality ruptured cells. There is a steep decline on the right side around the 4-5 x axis value, which represents the cells with very few transcripts. This can be due to ruptured cells, debris, and bad cell quality. The area of the graph on the left where the log transformed transcript counts per cell are the highest. The sharp increase at the beginning of the graph suggests that these barcodes have a much higher than typical number of transcripts. This shows the doublets or multiplets because the transcript counts are aggregated from two or more cells that were not separated properly before sequencing. This would result in a higher transcript count due to one barcode being associated with multiple cells rather than just one.

Q2-1. Using the sc_plotcells() function, generate a UMAP embedding of the single-cell transcriptomes in your dataset using the top 2000 over-dispersed genes as features. A list of over- dispersed genes can be retrieved using the sc_hvg() function. Note: PCA dimensionality reduction is performed prior to UMAP dimensionality reduction within the sc_plotcells() function. Refer to the ‘Appendix: Glossary of Useful Functions’ at the end of this document for guidance on how to apply the sc_plotcells() function.

BMEN6003_HW3

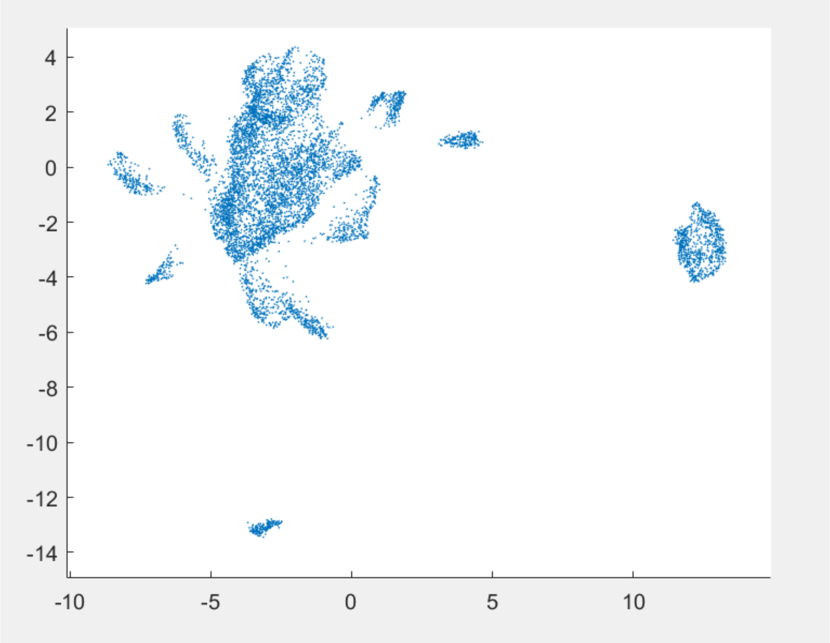
Marcus Ibrahim / Harry Kabodha / Ayman Khaleq

Why are highly variable genes frequently chosen as features when modeling single-cell transcriptomics data?

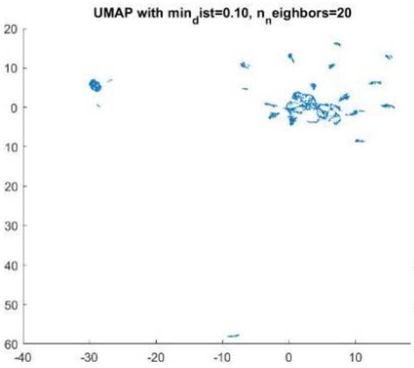
In single-cell transcriptomics, we choose HVGs as features because they tend to capture the most significant biological variations among individual cells. Indeed, these genes are more likely to be differentially expressed across various cell types. Therefore, by focusing on them, we can enhance the ability of our model to detect and highlight the underlying structure of the cell populations by distinguishing between subgroups. Finally, this also reduces the dimensionality of the dataset, which minimizes computational complexity.

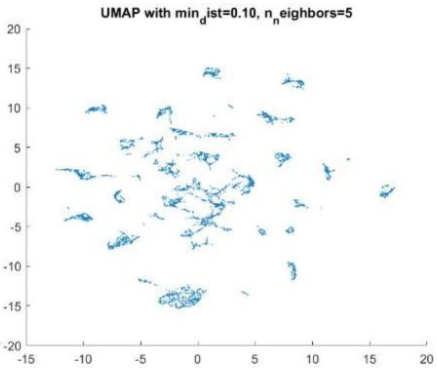
What is UMAP attempting to summarize? Categorize your answer by referring to the local and global differences in the distribution of individual samples in UMAP space

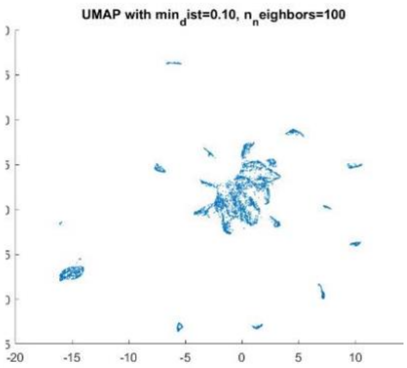
UMAP is trying to summarize the high-dimensional structure of the single-cell data in a way that is faithful to the original dataset. In other terms, it aims to preserve the local structure, meaning that cells that are similar in high-dimensional space remain close in the low-dimensional projection. This local fidelity allows us to identify and interpret cell clusters that share similar gene expression profiles. Not to mention UMAP also reflects the global structure, but with a “less rigid” approach than local relationships; it helps convey the overall data layout, revealing broader relationships between more distinct cell populations.

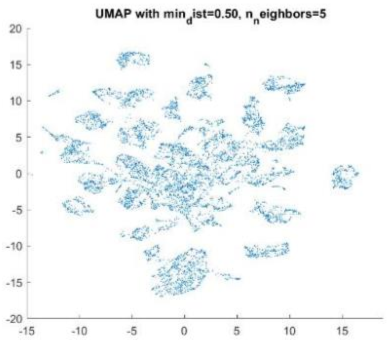


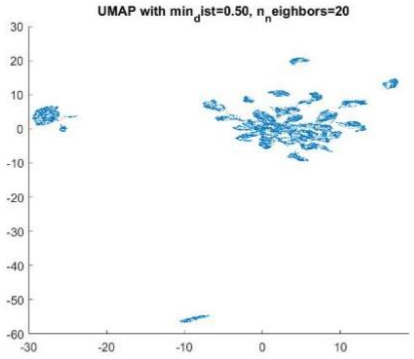
Q2-2. Regenerate the UMAP embedding by varying the min_dist (at 0.1, 0.5 and 0.79) and n_neighbors (5, 20 and 100 neighbors) parameters in sc_plotcells(). Let ndim=2, donorm=false and dolog1p=true for each variation of min_dist and n_neighbors in this question. [5 pts]
How do these parameters alter the resulting embedding? Do smaller values for min_dist optimize for summarizing local or large-scale relationships? [5 pts]

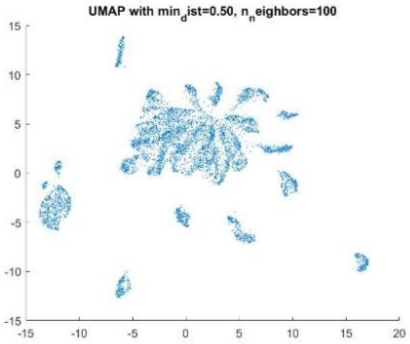


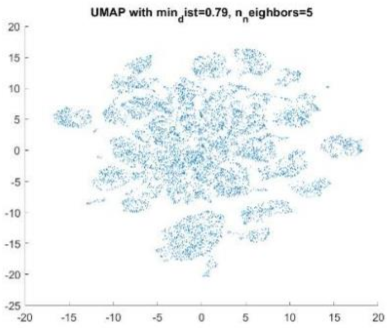


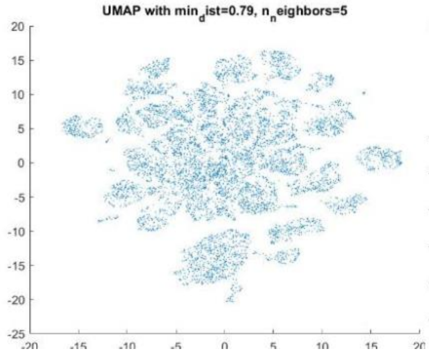


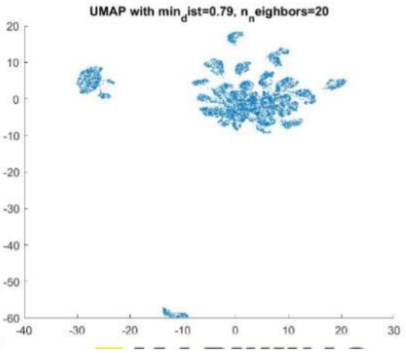








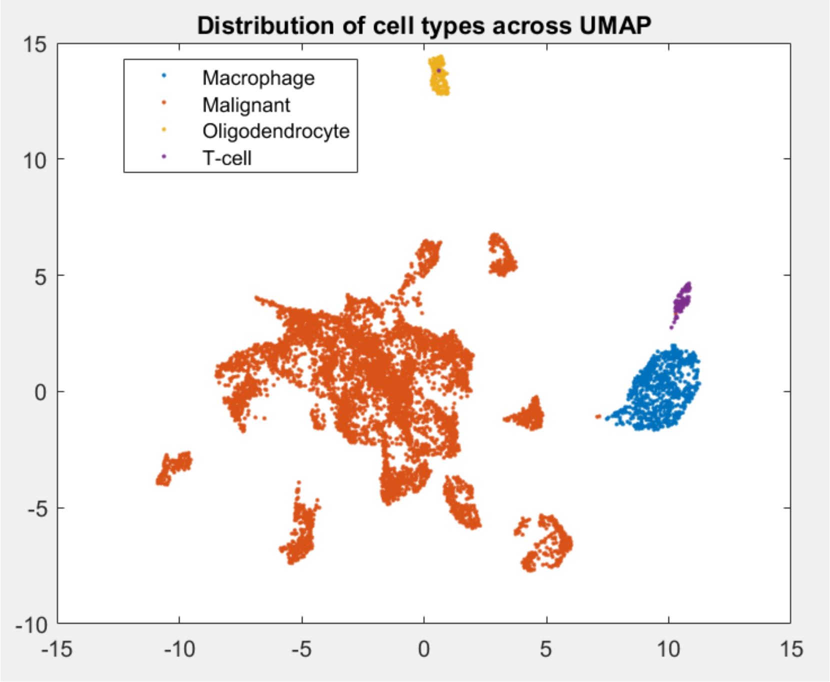


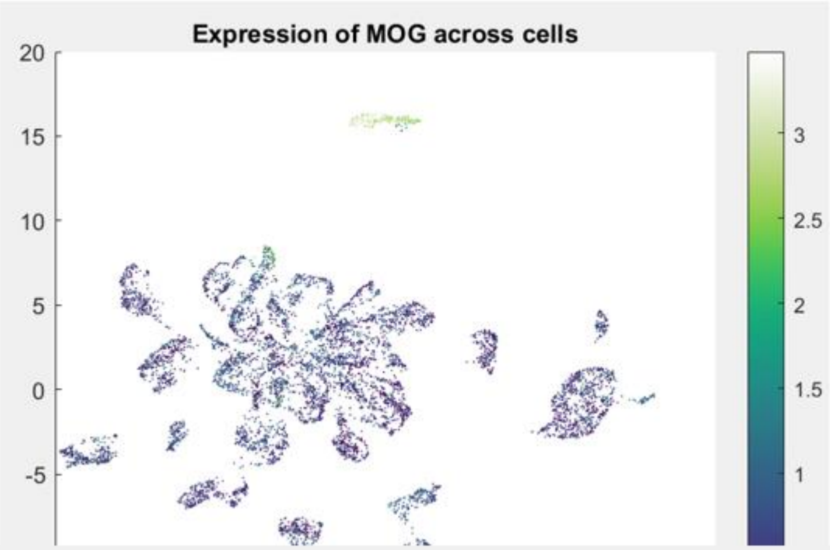


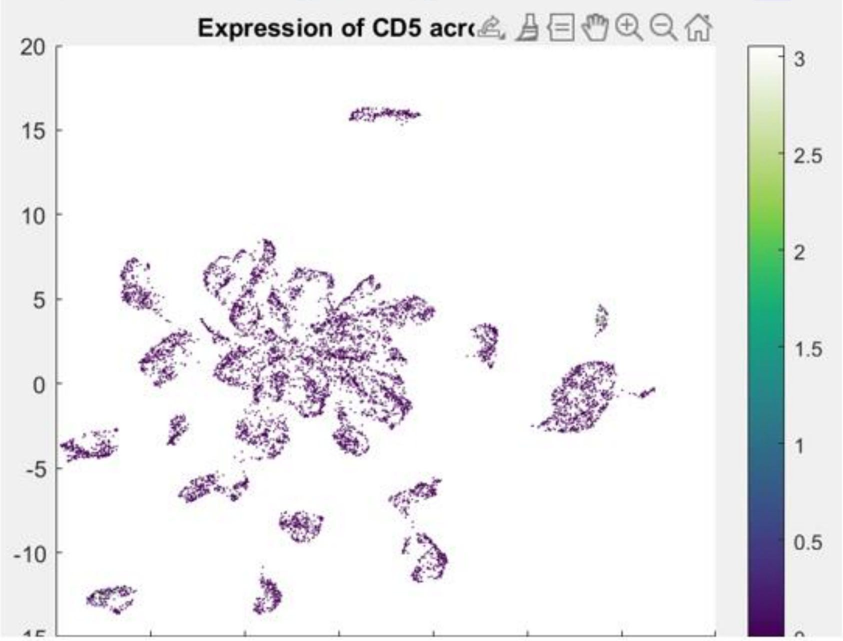
When we set a small min_dist (like 0.1), the clusters are very compact, which shows the finer details of the data's structure. It's great for when we need to focus on the subtleties between closely related cells. But, as we increase the min_dist value, the clusters start to spread out. By the time we get to 0.79, there's a noticeable blending of the clusters. The separation isn't as distinct, implying that higher min_dist values are better for exploring broader patterns rather than the fine-grained details.

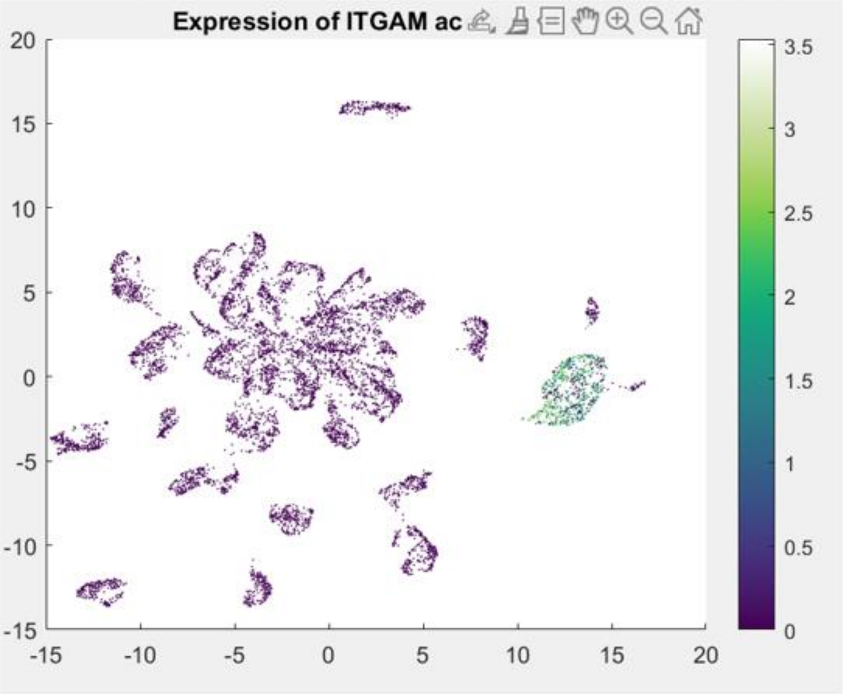
Adjusting the n_neighbors shows a similar trend. A low n_neighbors count like 5 makes the UMAP sensitive to the local neighborhood, which is why the plots with this setting have a fragmented look; they're picking up on very specific local differences. However, increasing n_neighbors smooths out these differences. With 100 neighbors, the UMAP seems to prioritize the bigger picture, bringing out more global structures and relationships between cells.

Q2-3. i. Use sc_plotcells() to plot the expression of each of these marker genes (ITGAM, CD5, MOG) across cells in your UMAP embedding from 2-1. Re-plot the embedding using the 'cell_assignment' label in the metadata




Expression of MOG across cells
20
15
10
5
0
-5
3
2.5
2
1.5
1


Expression of CD5 acr
20
15
10
5
0
-5
-10
3
2.5
2
1.5
1
0.5


Expression of ITGAM ac
20
15
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
15
20
3.5
3
2.5
2
1.5
1
0.5
0

Does it appear that one marker is sufficient to identify all cell types in this dataset? [5 pts]

Looking at the graphs above, we can tell that no single marker is sufficient to identify all cell types in the dataset. Indeed, even though ITGAM seems to match the macrophage cluster to some extent, there is no apparent distinction between CD5 and MOG for the T-cells and oligodendrocytes.

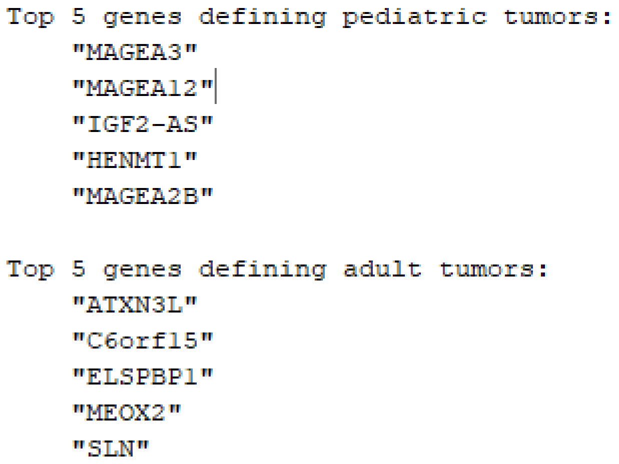
ii) Compare the distribution of cells expressing these markers with the distribution of tumor cells. How do tumor and normal cells partition across the UMAP? How do you think that distribution relates to the transcriptional similarity of normal and tumor cells across patients? [10 pts]

We see that cells expressing CD5, ITGAM, and MOG are found in distinct regions separate from the majority of malignant cells. This distribution reflects a significant transcriptional dissimilarity between the normal immune cells and the malignant cells. Notably, the clusters of normal cells marked by these immune markers are discrete from the dense malignant cluster, so we can effectively say that there is a clear transcriptional distinction from the tumor cells.

Moreover, the UMAP color intensities represent the expression levels of the markers, this could mean that We see clusters with varying shades, indicating different levels of expression within the cell populations, which can also indicate clustering of different types of cells. This transcriptional diversity across normal and malignant cells thus suggests that despite the mixed cellular landscape of a tumor, specific cell types maintain their unique gene expression profiles. This would mean that tumor cells have different gene expression than normal other cell populations, which can explain why different physical and physiological effects take place.

Q3-1. In this question, use the normalized expression matrix and its corresponding gene list as opposed to the expression matrix with just the 2000 highly variable genes. Create a subset of the dataset containing only malignant cells (i.e., remove macrophage, T-cells and oligodendrocytes). Perform a differential expression test using the

default Mann-Whitney U test in sc_deg() to identify genes that are differentially expressed as a function of cells being from pediatric or adult tumors. Inspect the output table. What are the top 5 genes that define pediatric tumors according to the average log2-fold change? Apply a p_val_adj cutoff at 0.001 and ignore Inf and NaN values. What are the top 5 genes that define adult tumors? [5 pts]



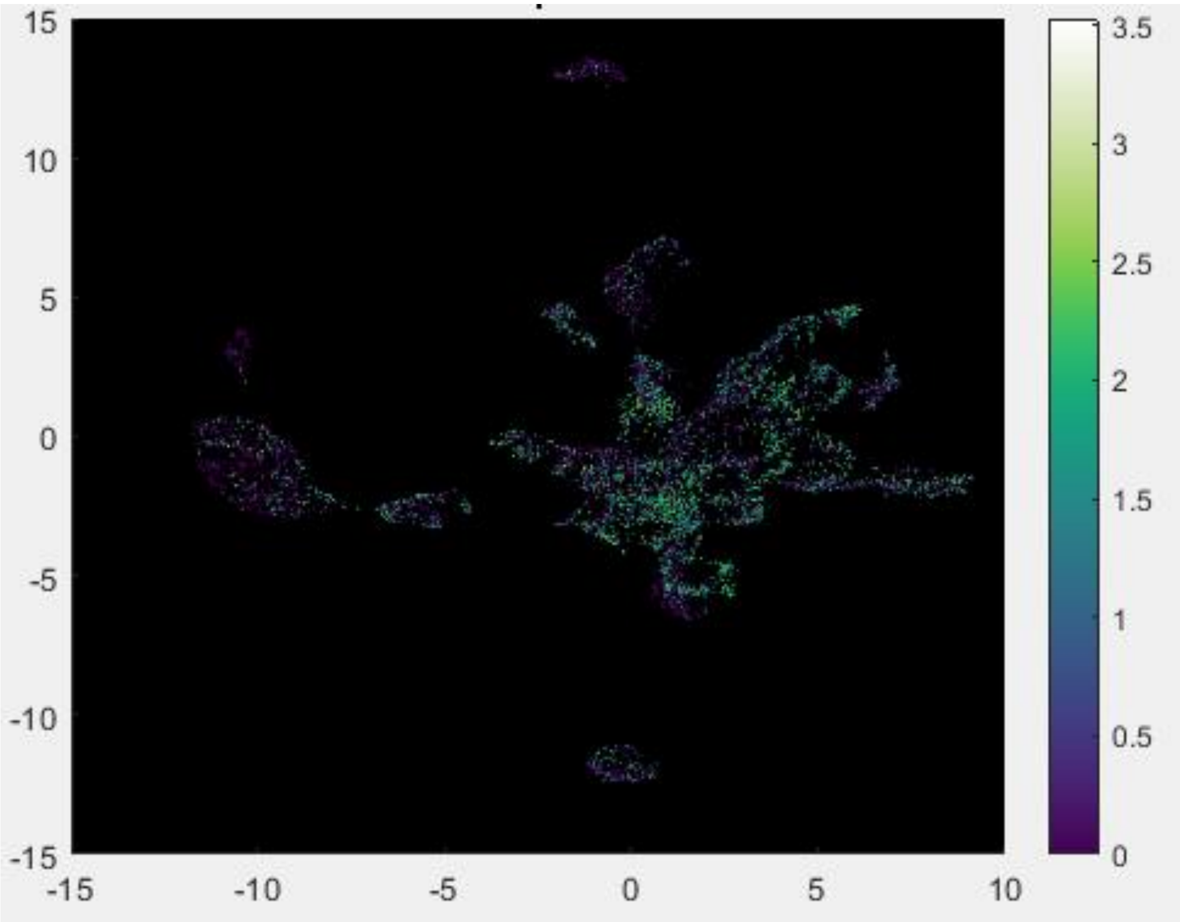
```
Top 5 genes defining pediatric tumors:
    "MAGEA3"
    "MAGEA12"
    "IGF2-AS"
    "HENMT1"
    "MAGEA2B"

Top 5 genes defining adult tumors:
    "ATXN3L"
    "C6orf15"
    "ELSPBP1"
    "MEOX2"
    "SLN"
```

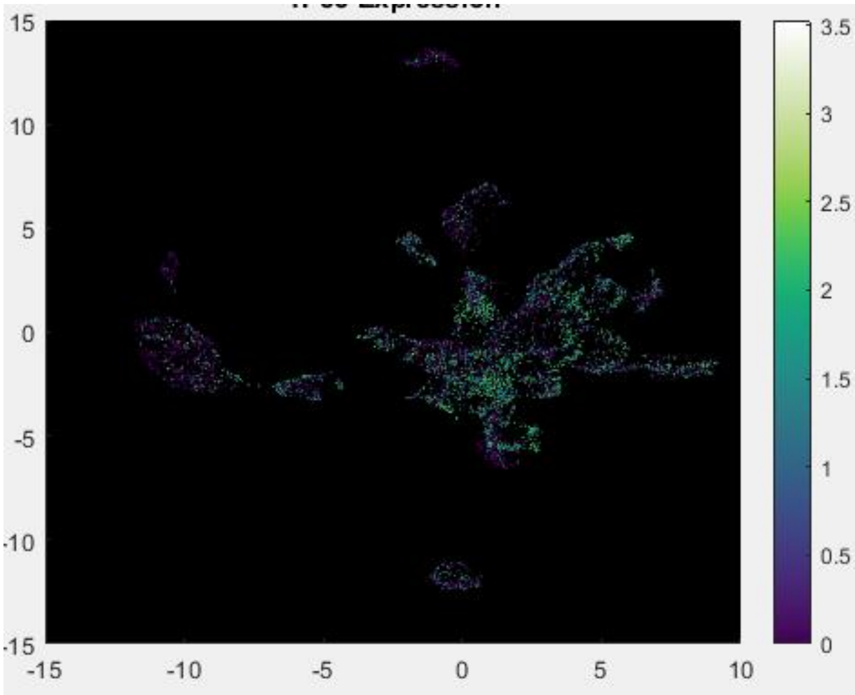
Bonus Q3-2. Examine the expression of the top pediatric and the top adult defining gene across UMAPs containing only malignant cells. Use the ‘viridis_white’ colormap for marker gene expression. Is expression confined only to pediatric and adult tumors? Is there evidence for heterogeneity in expression across cells of a single patient? [5 pts (only applied to the total of this assignment)]. Note: Keep in mind there may be technical factors that may underlie observed Heterogeneity.

There seems to be evidence that heterogeneity in expression crosses cells of a single patient by showing a heightened level of expression in the graph. From the graph shown below, expression ranges from 0-3. This would also mean that expression is greater in both pediatric and adult tumors. Both of the graphs are extremely similar, showing it is not confined only to adults or pediatrics.

Adult



Pediatric


15
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
3.5
3
2.5
2
1.5
1
0.5
0