Derek Host

Merscope Tutorial

[dth65@case.edu](mailto:dth65@case.edu)

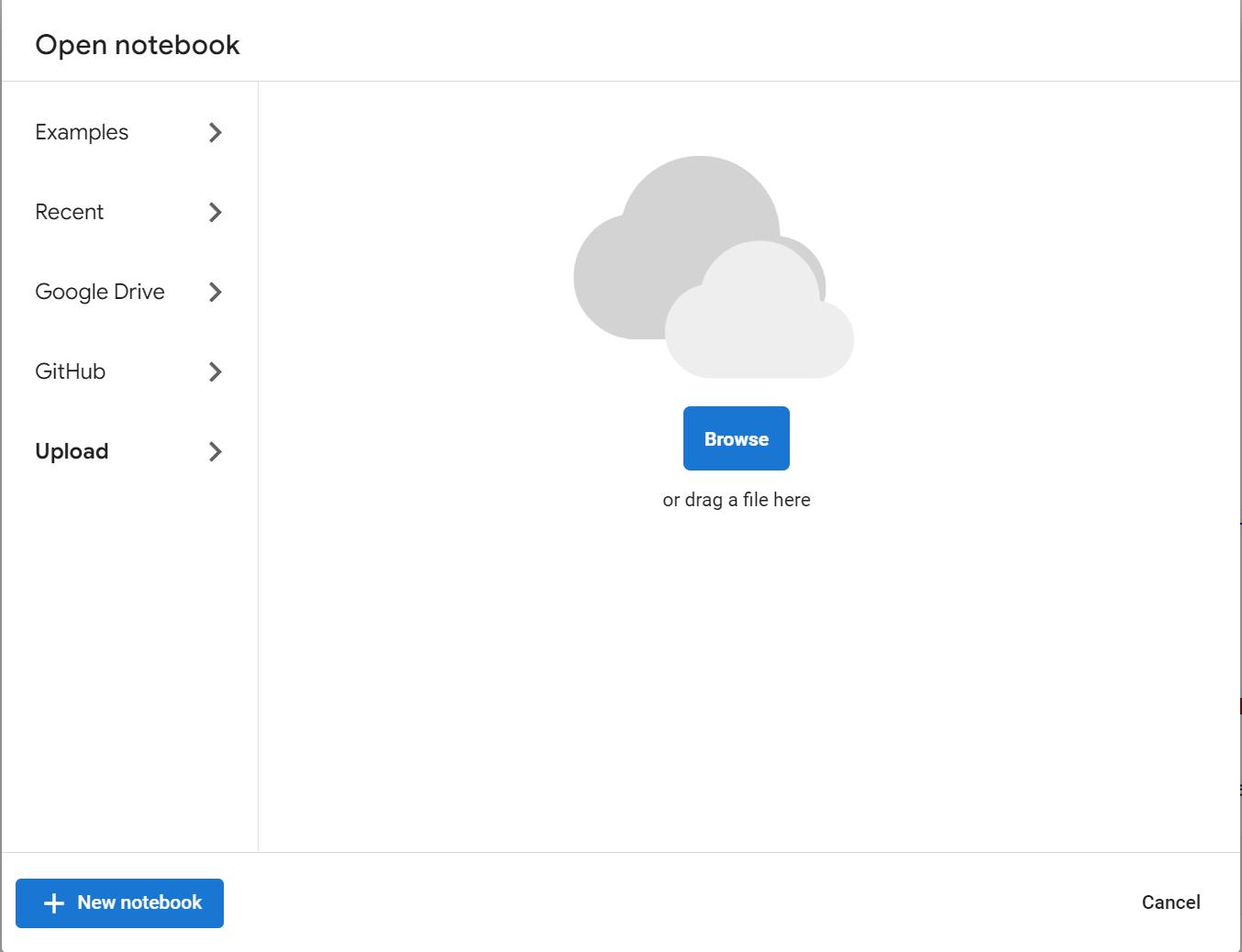
**Spatial**

**Transcriptomics**

**Tutorial**

**Quick Start**

1. Go to Google Colab (<https://colab.google/>) and select ‘Open Collab‘, this will be near the center of the page.
2. Select the upload button and drag the file named Spatial Transcriptomics Tutorial.ipynb to the middle section.



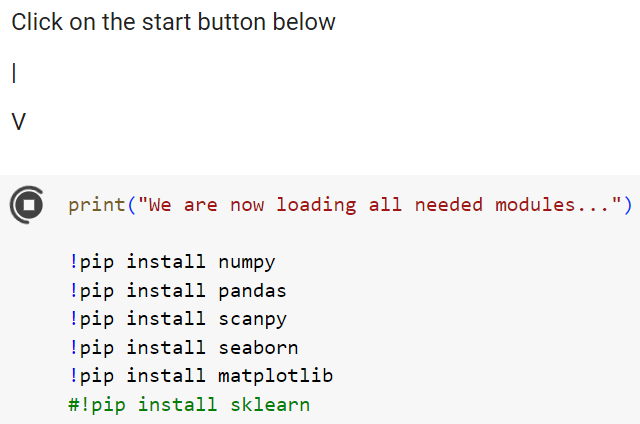


1. On the left side of the page, drag both .csv files

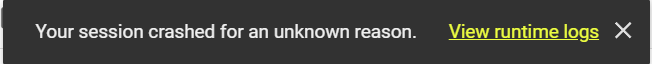
onto the sidebar (shown on the right), these

files may take a minute or two to load depending on the size. These are the output files within each ROI folder in the merscope output.

1. Make sure the files are named cell\_by\_gene.csv and cell\_metadata.csv, they should already be that name or something similar but just verify.



1. Once those have been successfully uploaded you can start the analysis by clicking on the start button on the top left of a ‘cell’, this is shown on the right side.
2. There is a chance that the first run will result in an error, if you don't receive an error skip to step 8. If you encounter an error, continue to the next step. The error would like like down below

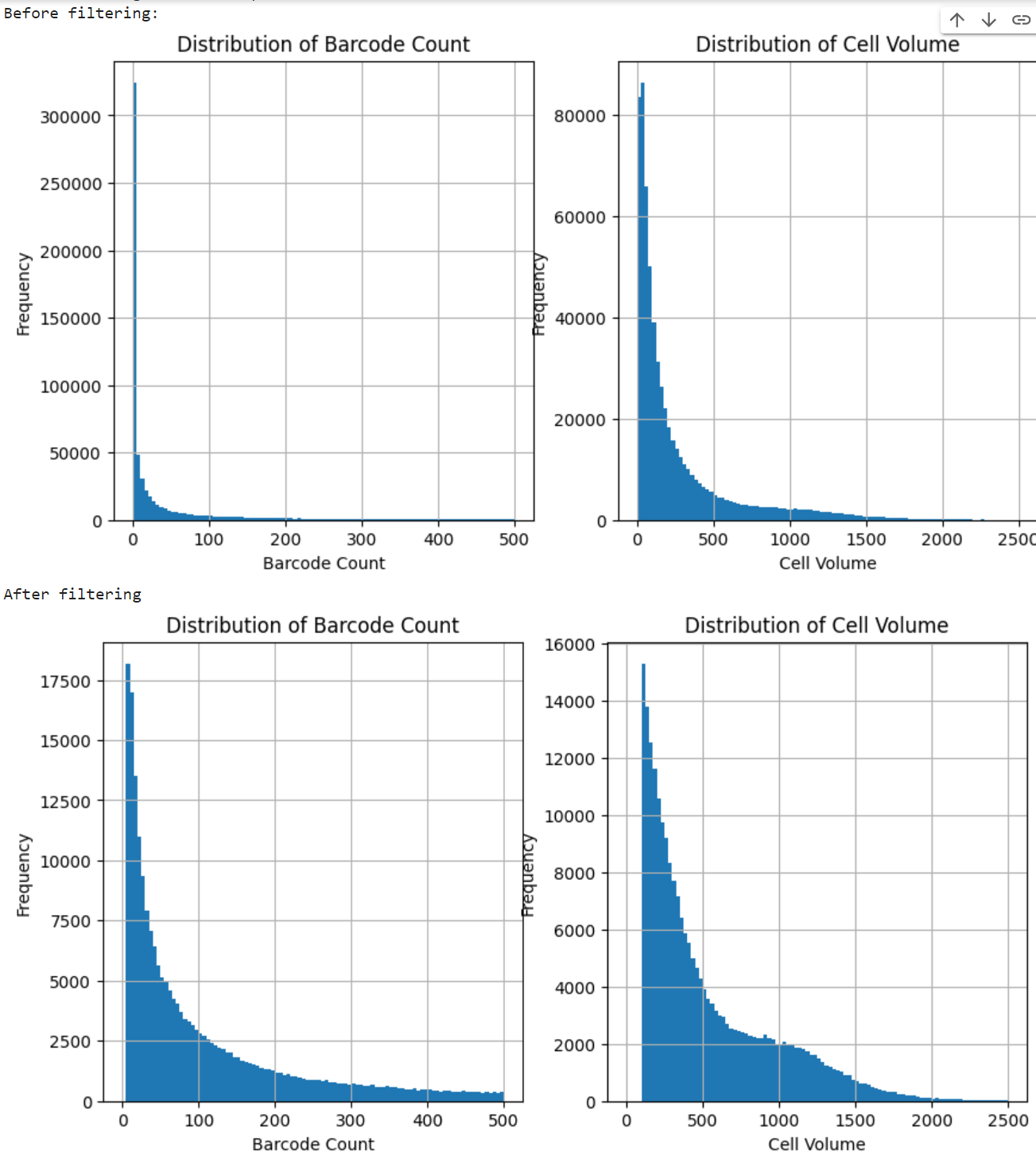


This will appear on the bottom left of the webpage

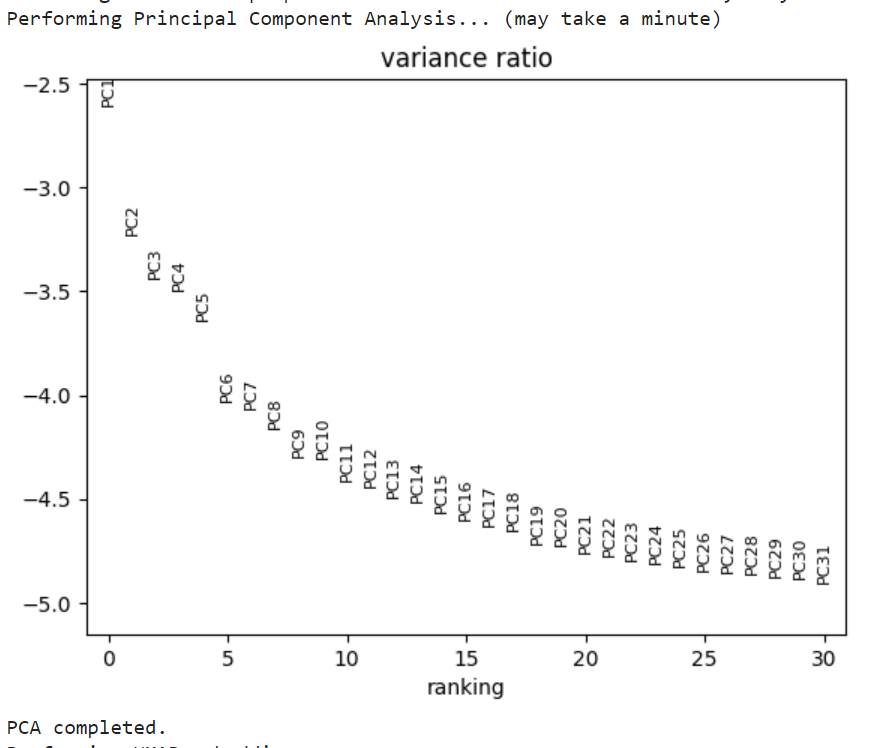


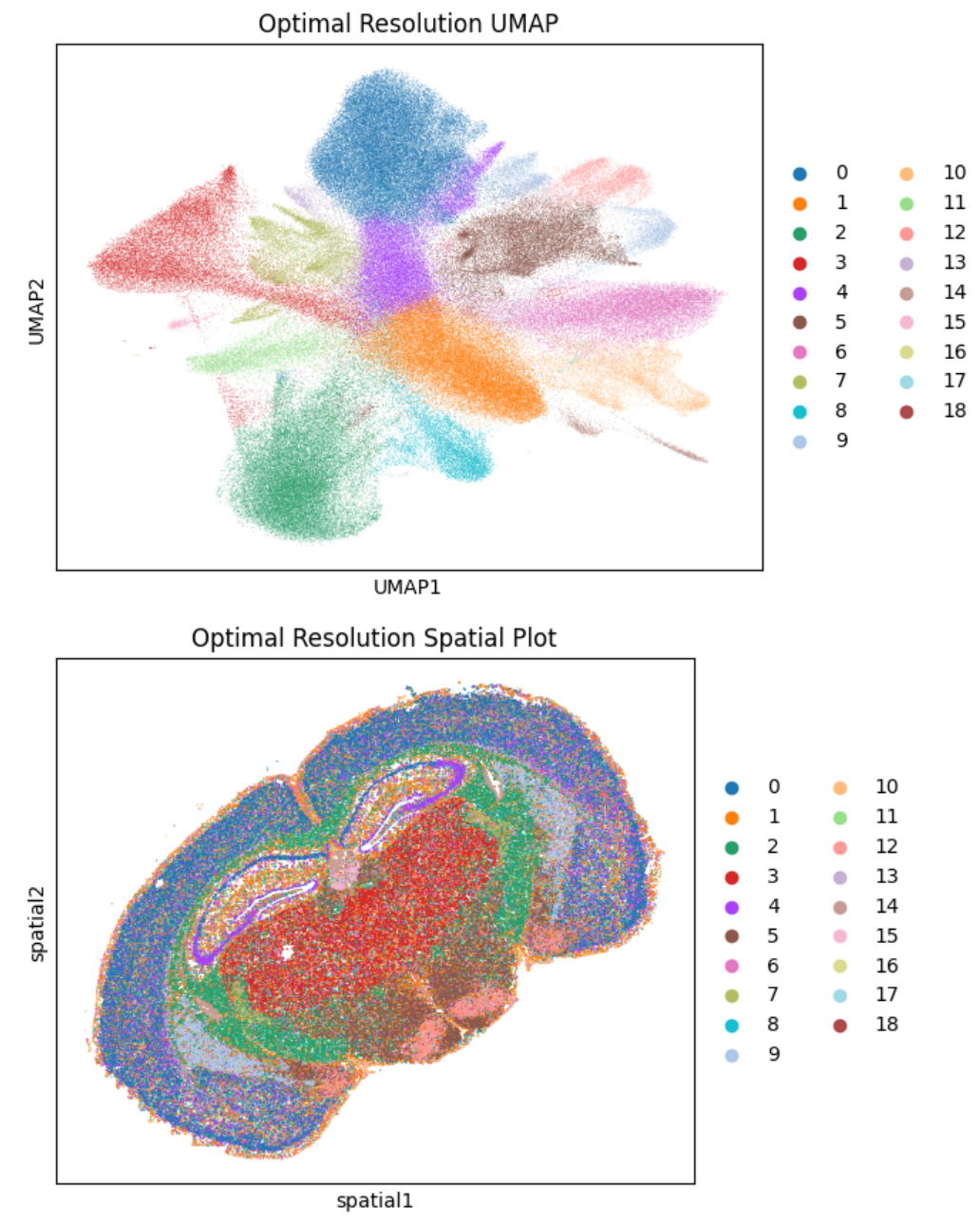
The play button will also turn red, indicating that an error occurred and the code block has stopped running.

1. If this has happened, simply click on the play button in step 5 and continue with the rest of the tutorial.
2. As the data is analyzed, its progress will be printed out on the screen. The printout will be below the code that is running, you can scroll down to find it.
3. The first thing that is done is having the data loaded in, a brief summary of the cells and genes that are included such as the total number of cells (unfiltered) and genes (in the panel) that were measured.
4. Graphs displaying basic distributions of the data are now shown before and after the filtering process. The filtering process is just removing all the cells that are below a certain predetermined size (much smaller than ‘real’ cells) and cells that have less than a certain number of transcripts present in the cell (cells with less than ~5 or so transcripts are likely not real cells) This is consistent with many other procedures and suitable for a quick analysis. This is shown below.



1. Once this is completed it will begin working on Principal Component analysis. “Principal component analysis (PCA) is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data.” It is simply a way to reduce how many dimensions are present in our dataset, currently, there are a lot of dimensions (genes) and this technique reduces this into principal components that are easier to work with computationally while retaining the majority of the data.



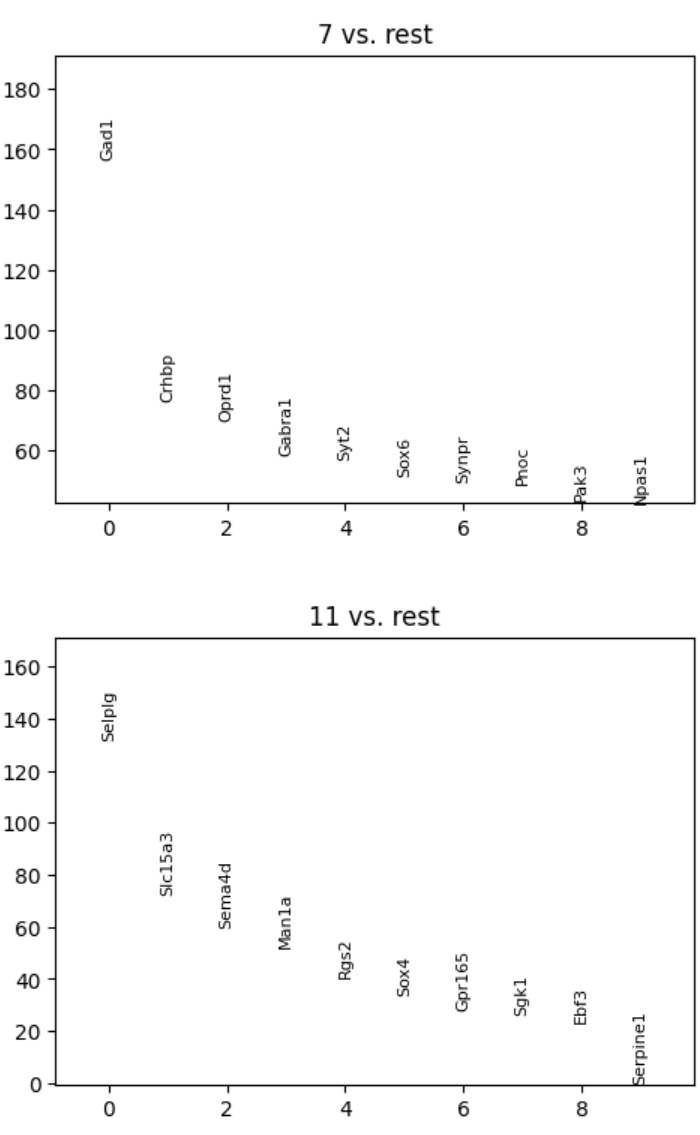
1. After this, a Uniform Manifold Approximation and Projection (UMAP) embedding is computed. This is where the computer projects the multidimensional data into a 2D plane. This allows for visualization of data and allows for the interpretability of it as well. This has a few parameters that can be changed if you continue learning more about the analysis that allows for capturing more local relationships or more global relationships between the cells.
2. Those were the calculations that allowed us to cluster the cells and visualize the result. The UMAP projection and Spatial projection are now displayed to you. The UMAP projection shows you the different clusters where every dot on the plot is a cell and every color is a different cell cluster. These groupings are based on similar gene expression between the cells within the same cluster. On the spatial plot, each dot is now overlaid in its respective spatial location. This allows us to see the distribution and localization of different cell types across the sample. To begin with, the clusters will be unnamed but the next steps will begin to involve the cell type annotation. An example image is shown to the right of a sample output for this step.

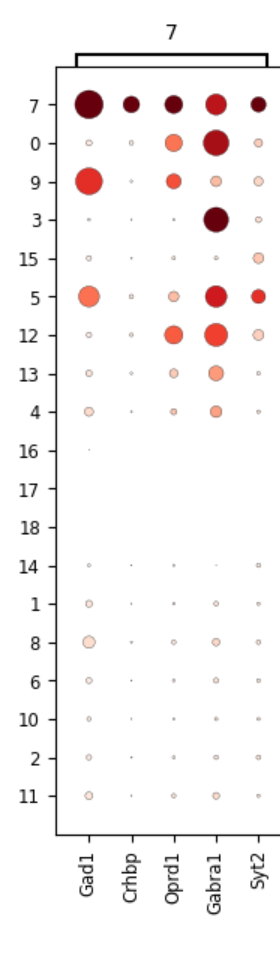
Here is a link to an interactive website that models and explains UMAPs in more detail:

<https://pair-code.github.io/understanding-umap/>

Here is a YouTube video on the topic of UMAPs for those that prefer videos

<https://www.youtube.com/watch?v=6BPl81wGGP8&ab_channel=AICoffeeBreakwithLetitia>

1. We can now begin to look at differentially expressed genes within the clusters, highly differentially expressed genes are genes that are predominantly expressed only in that cluster relative to the rest of the cells in the sample. These graphs can be read by looking for marker genes more differentially expressed ( higher score, further up) and then labeling the cell clusters accordingly (instructions given on the manual annotation page). In this example on the right we can see that for cluster 7 the gene Gad1 is differentially expressed and for cluster 11, Selplg is highly differentially expressed. This could allow us to tentatively label cluster 7 as GABAergic neurons and cluster 11 as microglial cells based on techniques described further below, it is important to always double-check these annotations to ensure accuracy. 



1. A dot plot is now displayed showing differential gene expression in a different way. Now the top five genes are displayed for each cluster, the genes being labeled on the bottom of the graph. Each row represents the gene expression of each gene shown on the graph. This allows users to quickly find genes that are only expressed in a few cell types. Each dot in the column represents the levels of expression of a specific gene for a single cluster. The size of the dot relates to the percentage of cells that express that gene, the bigger the dot the larger the proportion of cells in the cluster that express the gene. The darker, more red the color is, the higher the mean expression is. This is normalized to values 0-1. An example of this is shown on the right. Here we can confirm that Gad1 is differently expressed in cluster 7, this graph also shows us that there is also a small amount of expression in clusters 9 and 5. This could be because they are similar cell types or for various other reasons. These previous two graphs, used in unison greatly help you manually annotate cell types.
2. The next steps are to now review the following pages to understand how to produce basic graphs and charts with the data. Follow the examples the best you can to avoid any errors, if they do occur read the bottom of the page for tips on how to solve those problems, if you ever encounter any errors feel free to reach me at [dth65@case.edu](mailto:dth65@case.edu) with any questions.

**Variables to Change in Gene Section**

**Genes**

* Definition/Meaning - List of genes to plot information on (violin plot, matrix plot, etc...)
* Possible values - Any gene in the gene panel (MUST BE IN PANEL)

Sample structure

* genes = ['Gene1']
* genes = ['Gene1','Gene2']
* genes = ['Gene1','Gene2',....,'Gene N']

**Name**

* This is the name of the experiment, it can be any string of characters as long as it is

Examples:

* name = 'BrainSlide1'
* name = 'CF\_Replicate1'
* name = 'Test'

**Group**

* Definition/Meaning - This is the grouping that is used for the naming of the cell clusters, this can either be the default clusters the computer found earlier manually annotated cell types.
* Possible values - 'leiden' for original groups, 'manual' for manually annotated clusters.
* **Must use 'leiden' if you have not finished manually annotating the cell types**

Examples:

* group = ‘leiden’
* group = ‘manual’

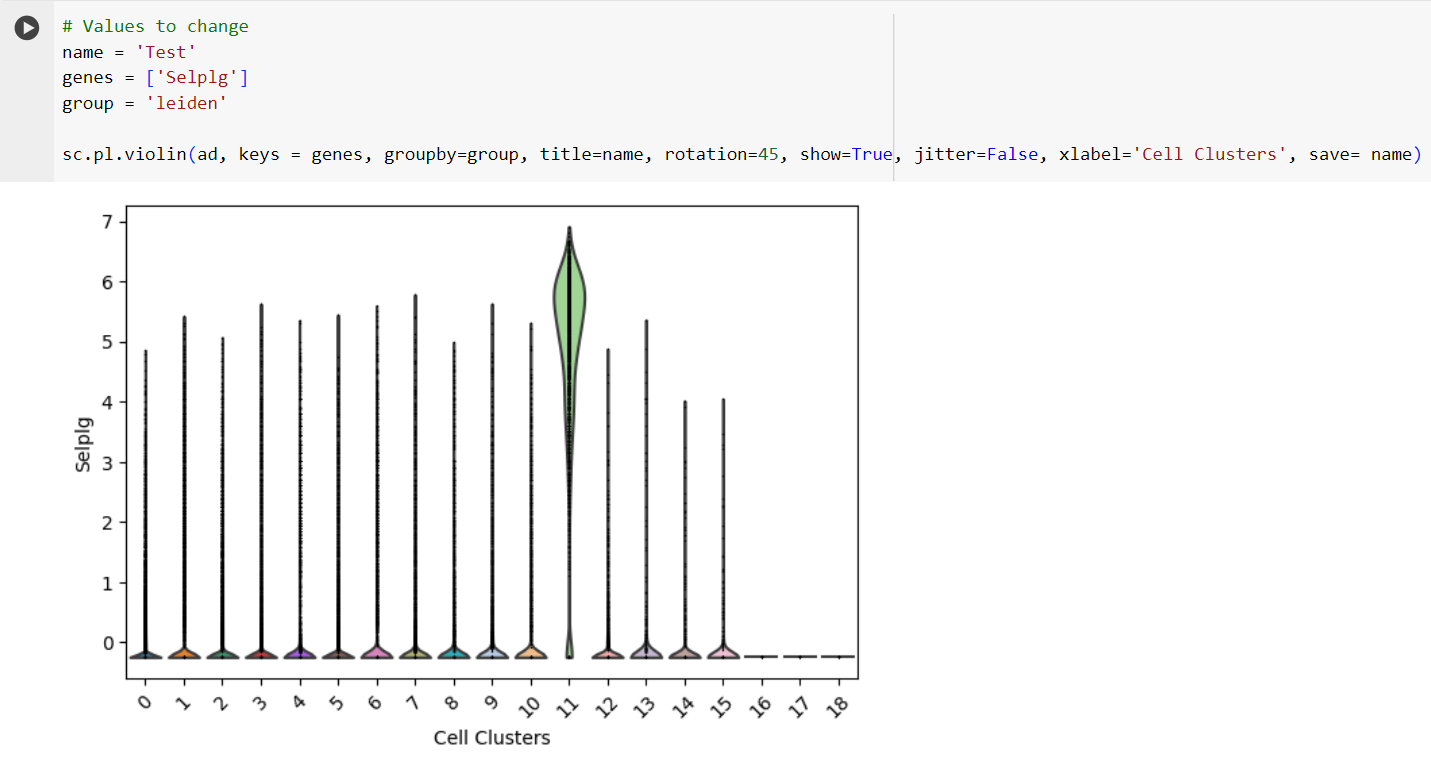
Here is a quick video on basic Python syntax if you are struggling (first 3 minutes is enough)

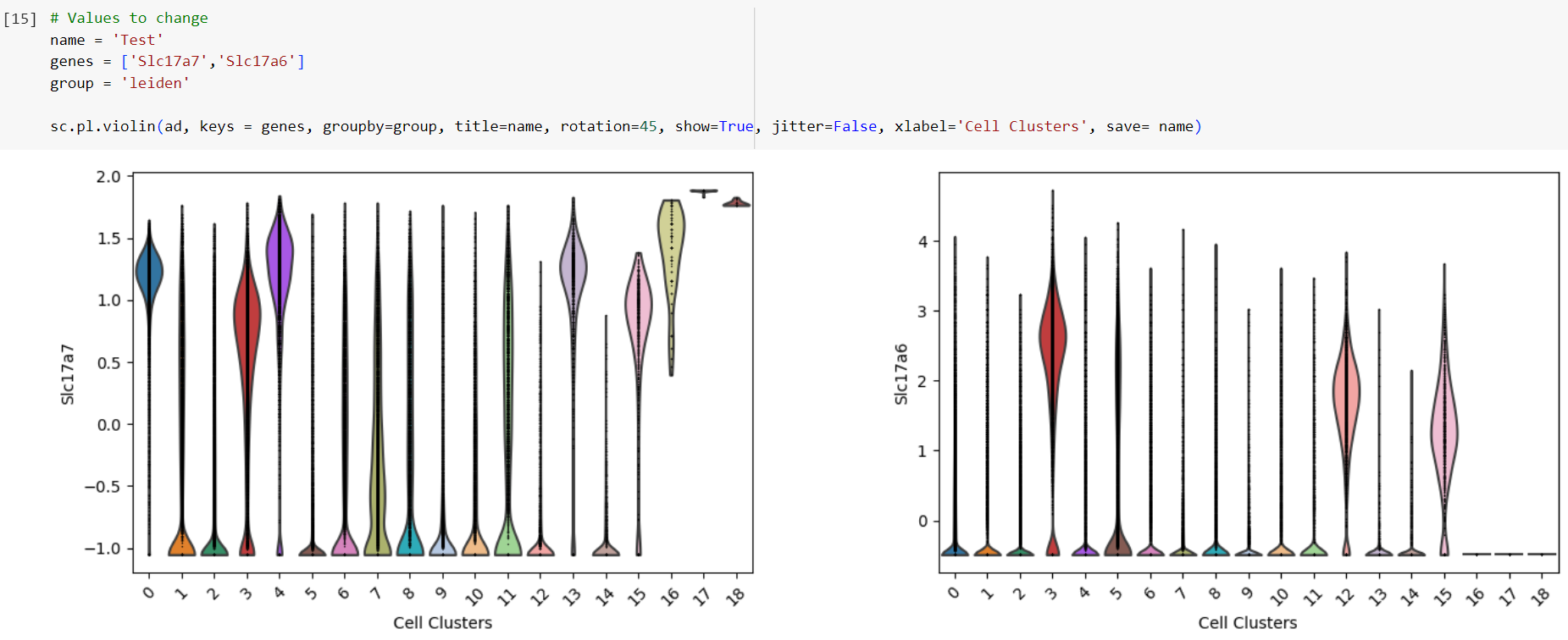
<https://www.youtube.com/watch?v=cQT33yu9pY8&ab_channel=ProgrammingwithMosh>

**Violin Plots**

This code block is used to generate violin plots of gene expression within each cell cluster all on one graph side by side. Can be useful to see gene expression of key marker genes quickly. Can plot multiple of these plots in succession if there are multiple genes of interest at the same time. There are multiple values that can be changed in this cell, name, genes, and group, be sure to follow the directions above for proper syntax regarding the variables. Once you have the parameters what you desire, just click the top left play button when ready like before.

**Example Input and Output**

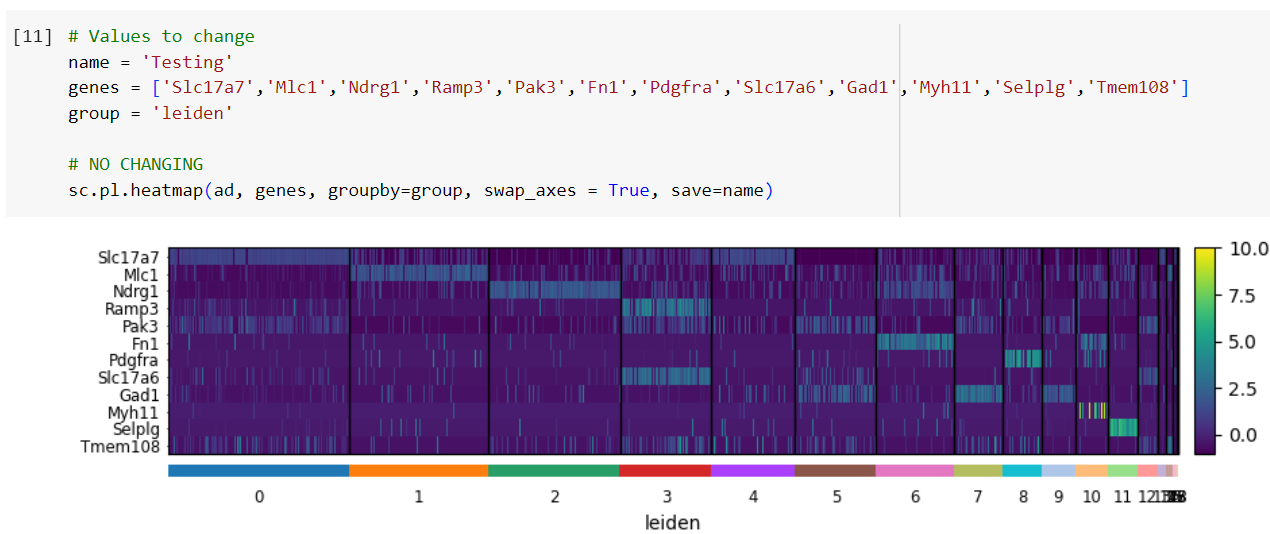


**Common issues:**

* Gene not in the panel used - change the gene name
* group = ‘manual’ when manual annotation not performed - change to ‘leiden’
* Invalid syntax - look for missing parenthesis, quotes etc.. (follow guide above)

**Heat Map**

Heat maps are a graphical way to visualize gene expression in each cluster where the colors represent the levels of expression in every cell. This is similar to a matrix plot with a few differences, first, The size of the clusters is preserved so the larger the cluster (most cells) will appear the largest because each cell is represented in a heat map but not a matrix plot. There are multiple values that can be changed in this cell, name, genes, and group, be sure to follow the directions above for proper syntax regarding the variables. Once you have the parameters what you desire, just click the top left play button when ready like before.

**Example input and output**

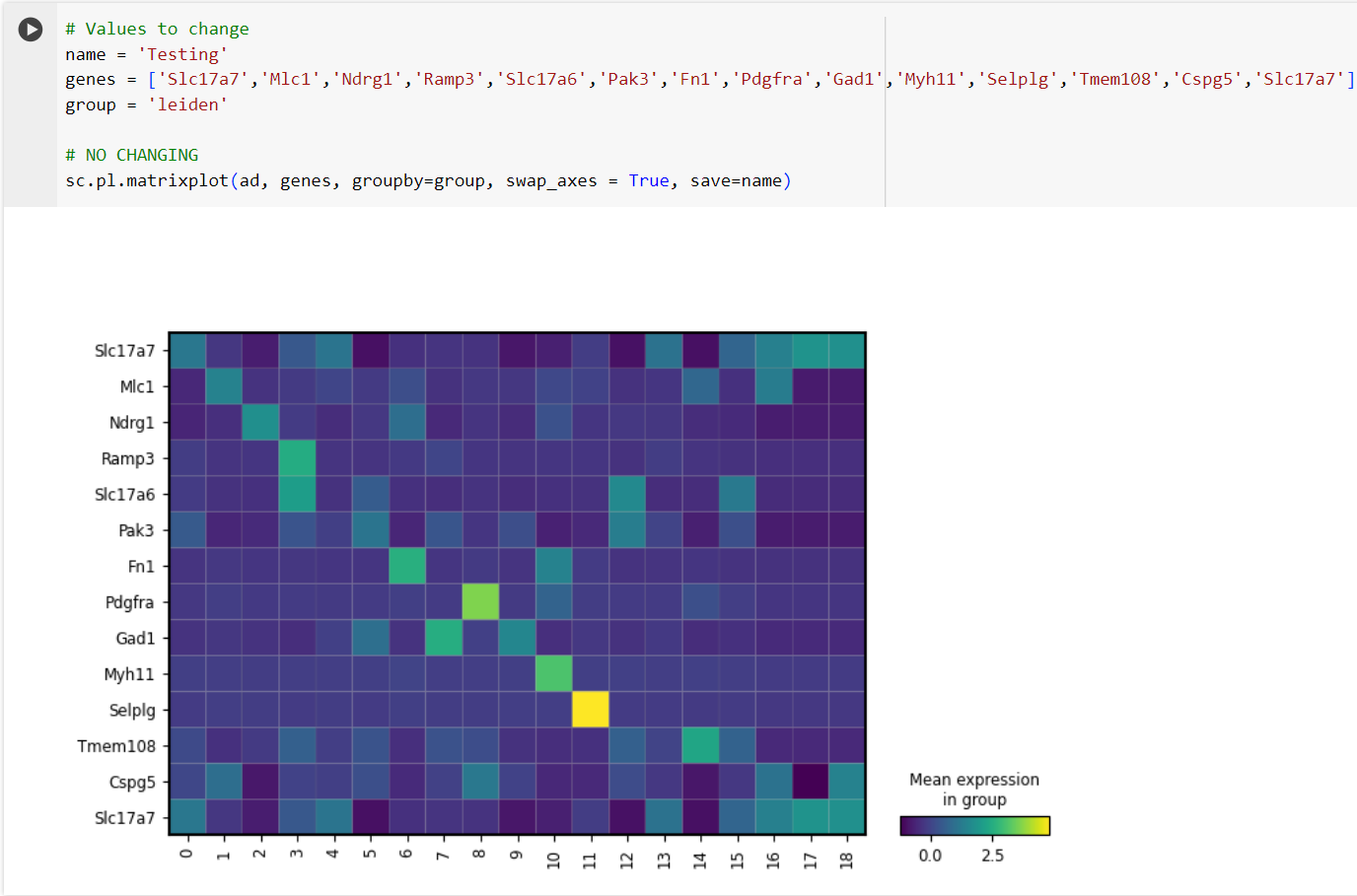
**Common issues:**

* Gene not in the panel used - change the gene name
* group = ‘manual’ when manual annotation not performed - change to ‘leiden’
* Invalid syntax - look for missing parenthesis, quotes, etc.. (follow the guide above) can be especially confusing with the genes variable

**Matrix Plot**

Very similar to the heat map above except the matrix plot has the same size area for each cluster regardless of size, this could be used to better represent clusters with very low cell counts that could represent rare cell types. There are multiple values that can be changed in this cell, name, genes, and group, be sure to follow the directions above for proper syntax regarding the variables. Once you have the parameters what you desire, just click the top left play button when ready like before.

**Example input and output**

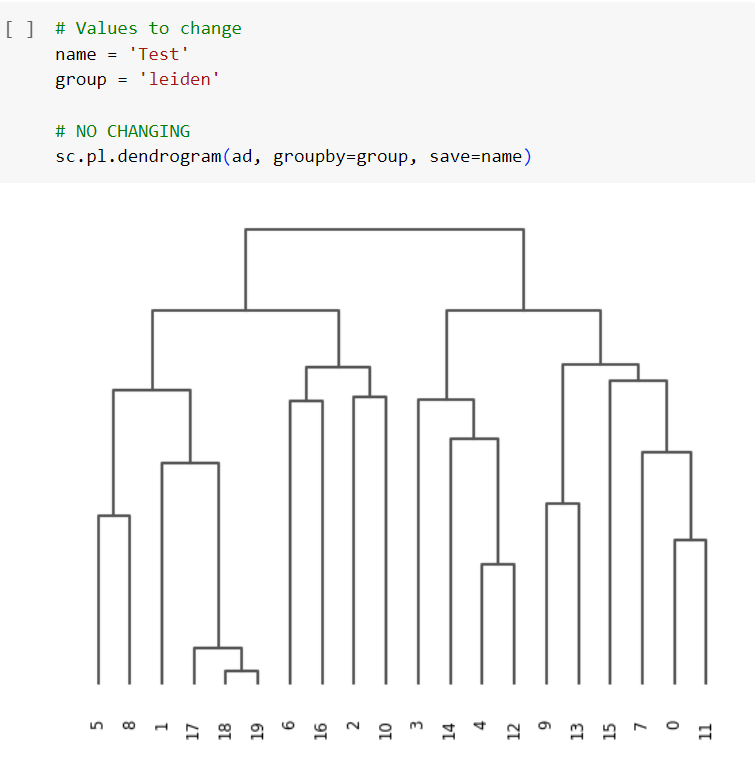


**Common issues:**

* Gene not in the panel used - change the gene name
* group = ‘manual’ when manual annotation not performed - change to ‘leiden’
* Invalid syntax - look for missing parenthesis, quotes, etc.. (follow the guide above) can be especially confusing with the genes variable

**Dendrogram**

Draws a dendrogram to represent the hierarchical clustering of the cell clusters. This could display what cells share a large amount of similarity to each other. There are multiple values that can be changed in this cell, name, and group, be sure to follow the directions above for proper syntax regarding the variables. Once you have the parameters what you desire, just click the top left play button when ready like before.

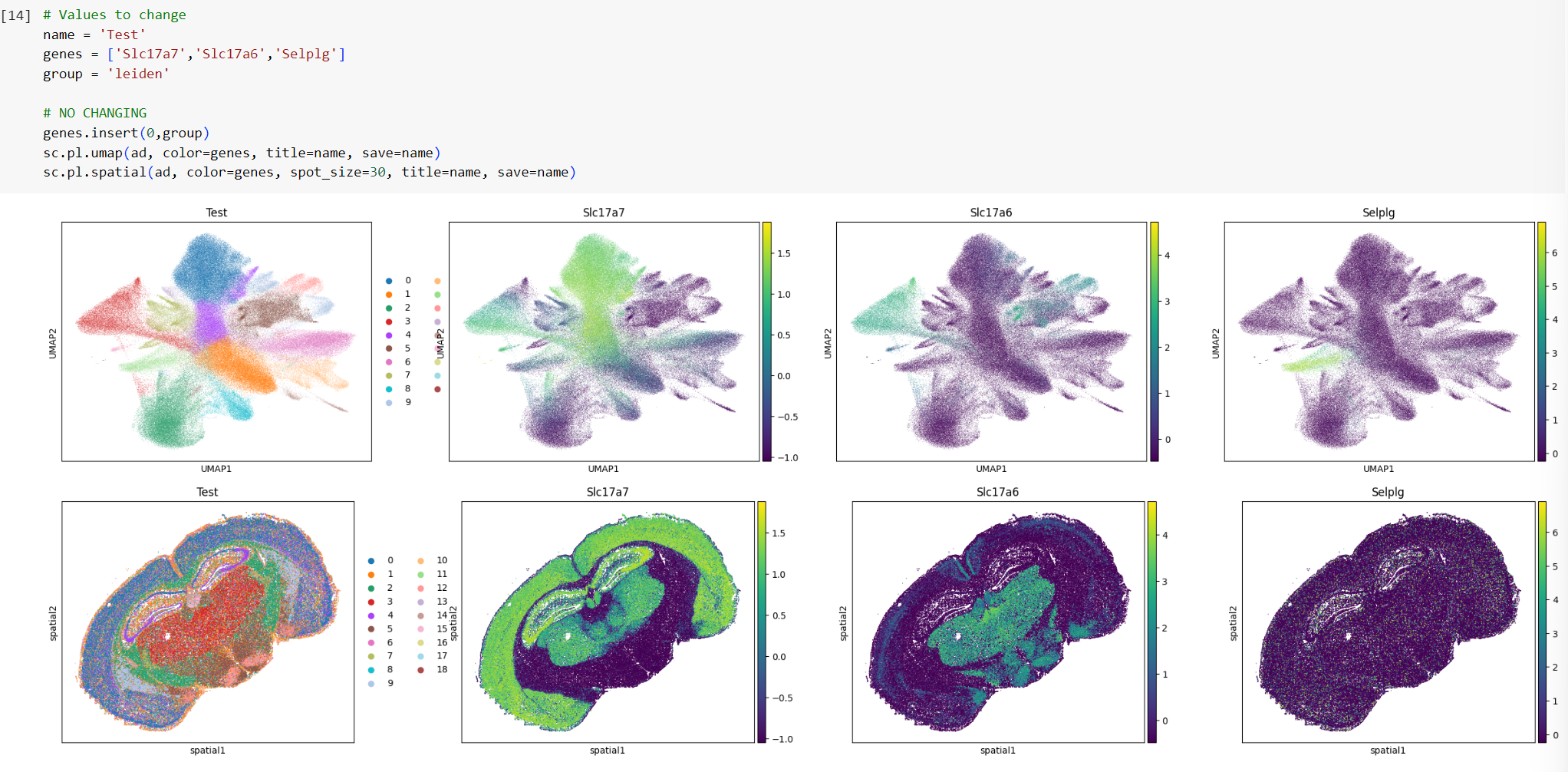
**Example input and output**

**Common issues:**

* group = ‘manual’ when manual annotation not performed - change to ‘leiden’
* Invalid syntax - look for missing parenthesis, quotes etc.. (follow the guide above) can be especially confusing with the genes variable

**Spatial gene Expression**

Plots both the UMAP and Spatial projection of whatever grouping you want (Leiden by default, can do manual annotations after completed). This allows you to visualize what cell clusters (now manually annotated with their cell types) are expressing specific genes with the UMAP projection and allows us to see where the gene is being expressed spatially with the spatial projection.

**Example input and output**

**Common issues:**

* Gene/s not in the panel used - change the gene/s name
* group = ‘manual’ when manual annotation not performed - change to ‘leiden’
* Invalid syntax - look for missing parenthesis, quotes, etc.. (follow the guide above) can be especially confusing with the genes variable

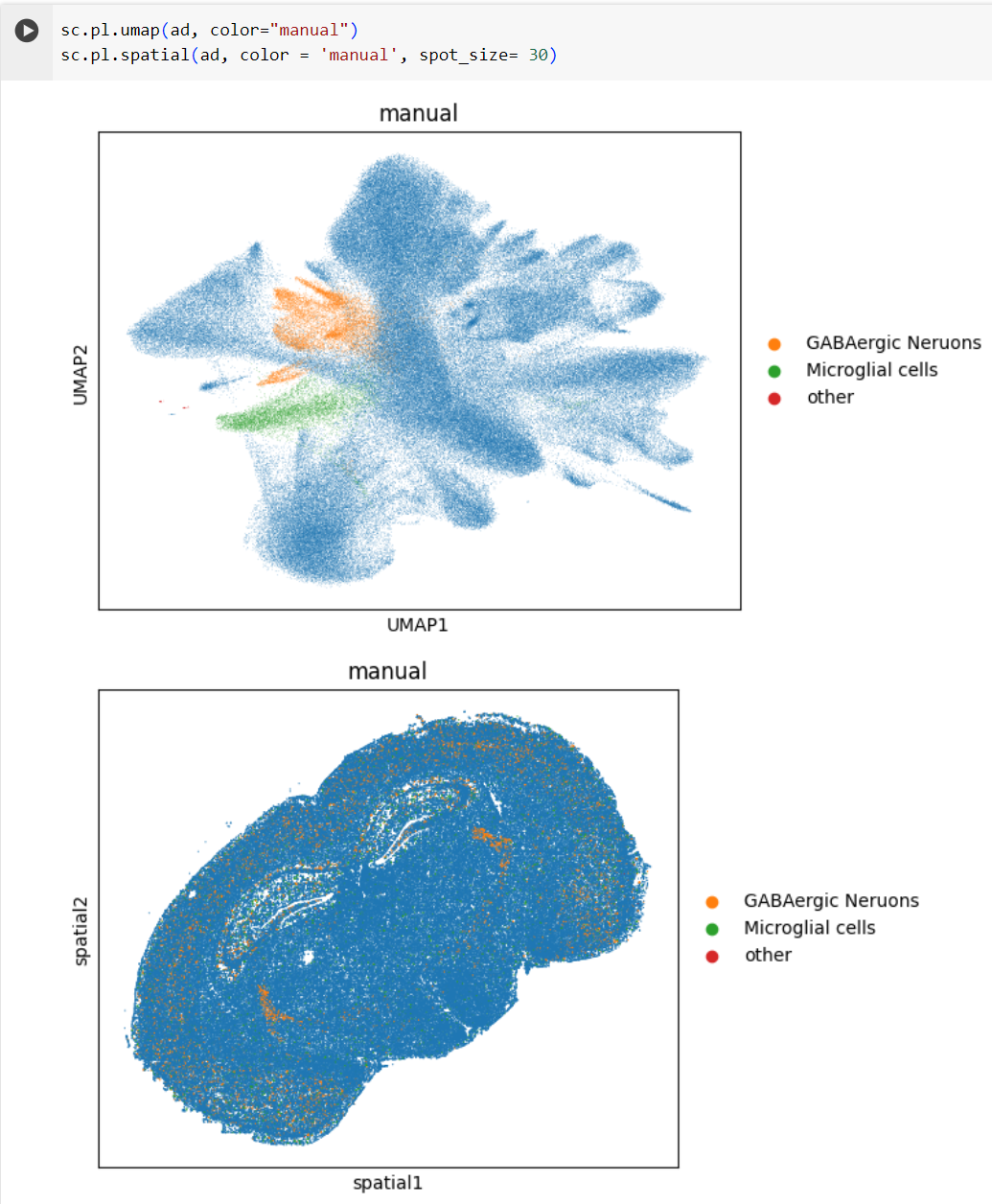
**Manual annotation**

**Useful Websites:**

* <https://www.proteinatlas.org/> (search gene and select single-cell in the top bar and then brain below the graph to view more info)
* <https://www.panglaodb.se/search.html> (search full list of marker genes or individually)
* <http://xteam.xbio.top/CellMarker/index.jsp>

**Quick Instructions:** Once you have found genes that are differentially expressed in specific cell clusters (highly expressed genes/s in only that cluster) you can begin to look to see if it is a marker gene for a particular cell type. For example, if Sst was only expressed in one cluster then you can deduce that cluster is a cluster of SST neurons. Ideally, you would like to confirm the presence of more than one marker gene but with a limited gene panel that can be difficult all the time. As you continue making annotations, the spatial projection becomes more and more complete. Once completed you can go back to the previous cells to make figures with manual annotations instead of the Leiden clustering names

**Formatting:** every new cluster that you are annotating must be in the below format. There must be a comma after every cluster. The order of the clusters doesn't matter but it can be helpful for your own visual sake. For example, the first entry could be cluster 18 or cluster 0 but going in order is always easier to visualize.



**Example syntax:**

**add/remove as needed**

**Further Possibilities**

This notebook was just a small intro into the cell clustering and manual annotation of cell clusters with their cell types, as well as a basic intro into the figures that can be produced as well. Many more possibilities involve just a little more tinkering and time.

You are able to delve into a specific cluster/s of interest and further cluster them out. This would allow you to break apart cell types of interest without worrying about raising the resolution of the overall clustering and affecting them. An example of this would be if you were interested in looking at splitting apart a population of clusters that you know are neurons. You can select only those cells and cluster them out into their subpopulations, this could produce a subset of cells that are GABAergic and another subset of cells that are Glutaminergic neurons. If needed you could then subcluster the GABAergic cells further to break them down into subcategories such as SST neurons (if you need to divide like this twice you might need to change the original resolution though). This isn't the only possibility, you could also attempt to find rare cell types. If there is a rare subtype of smooth muscle cells that you are interested in, you could break down the population of smooth muscle cells into subclasses until you discover what cluster is your cell type.

Automatic annotation is possible with more familiarity with this process, as long as there is a good reference for your tissue that you are working on automatic annotation should be possible. While manual review is still STRONGLY recommended, you could save a considerable amount of time with this method.

While this basic introduction gives you a brief overview of the possibilities for charts and graphs, these are just the most basic possibilities. If you would like to delve into more details yourself it is fairly easy to get familiar with the process and produce the charts and figures that you desire but regardless this notebook should give a basic introduction to the generation of figures for ST analysis.