This folder contains a pipeline for labeling cell types in an automated fashion.

First, user inputs the genes that are known to associate with individual cell types into the document cell\_type\_labels.tsv.

Second, user copies all files from this google Drive folder into a new folder on their hard drive.

Third (optional) user builds docker image and runs it to install seurat, e.g. with something like:

docker build . -t image\_name

to build seurat docker image, and run seurat from the image using

docker run image\_name

Fourth, user runs

python3 automated\_Merfish\_Seurat\_driver.py

The script will ask for user to input desired 'resolution' of clusters and desired number of principal components to analyze (displaying a helpful table of significant principal components).

The script generates a list of genes that are positively associated with each cluster, similar to the included table called 'markers.tsv'). The script then uses this list in conjunction with cell\_type\_labels.tsv to match genes with cell types, and automatically relabel clusters (so that a cluster known to be positively associated with genes known to associate with a cell type becomes labeled as that corresponding cell type). The script will generate an output file (e.g. actual\_cluster\_labels.txt) and prompt the user to edit this file before continuing. clusters that associate with multiple cell types are given a slash to represent this. The script also generates dotplots that the user can view before continuing.

Not implemented yet: theoretically, the data used to generate the dotplots could be exported to an output text file. A user could then write a script that would set thresholds for expression level of a gene of interest and fraction of cells expressing the gene of interest and lack of other competing identifications that the cluster might associate with. Ideally, these thresholds would be edited until they agreed closely with a human's subjective judgment.

In my case (with merfish datasets), 2 dimensional images are then generated along with umaps for the individual cell types.