BGI Test presentation

Yunona Pospelova
17.11.22

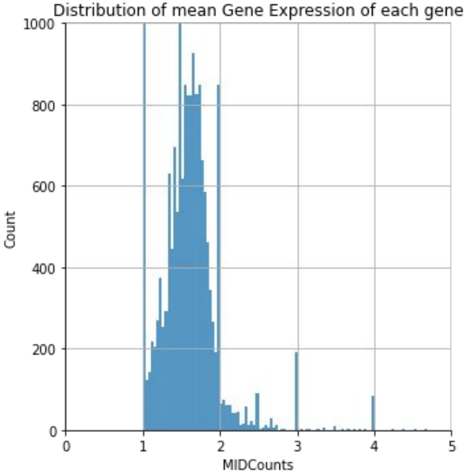
GitHub link

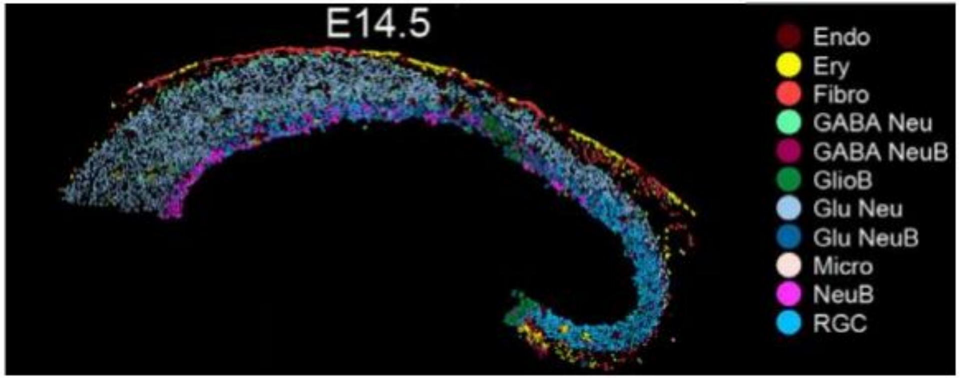
Literature Review and Data exploration

Stereo-seq enables large field-of-view spatial transcriptomics at cellular resolution

Exploration:

- **Amount of cells**: 4872
- **Unique genes:** 18698





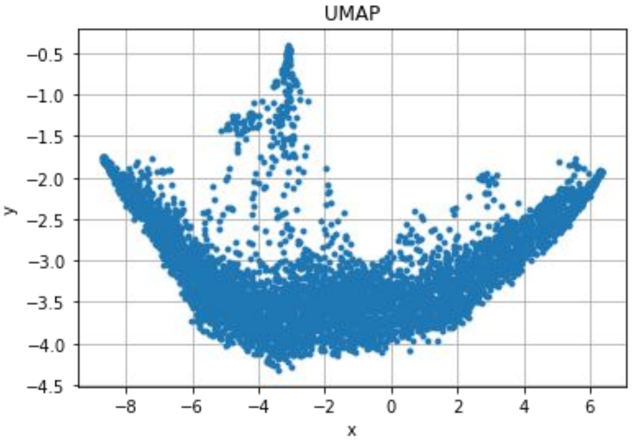
Ground truth image from the paper (Ao Chen et al., 2022)

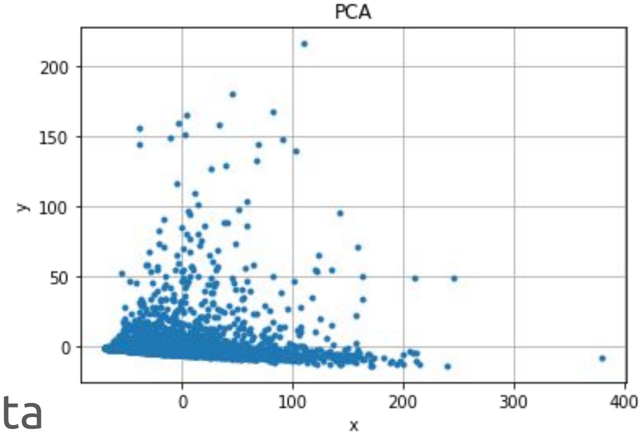
Dimension reduction

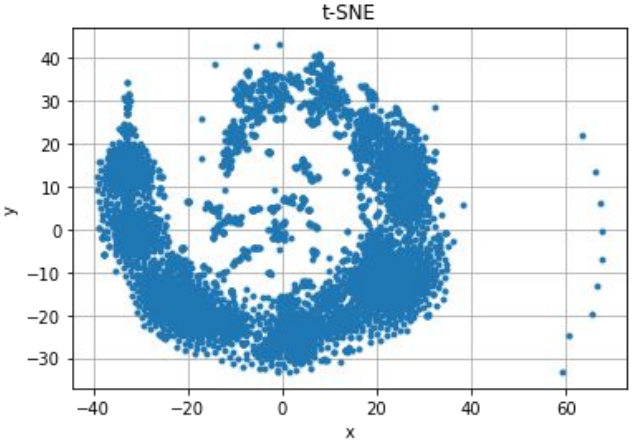
3 techniques were used:

1) UMAP
2) t-SNE
3) PCA

These dimension reduction methods showed that visually data doesn’t split to clusters







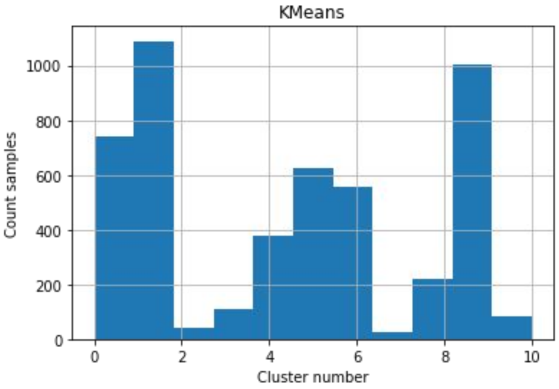
KMeans clusterization

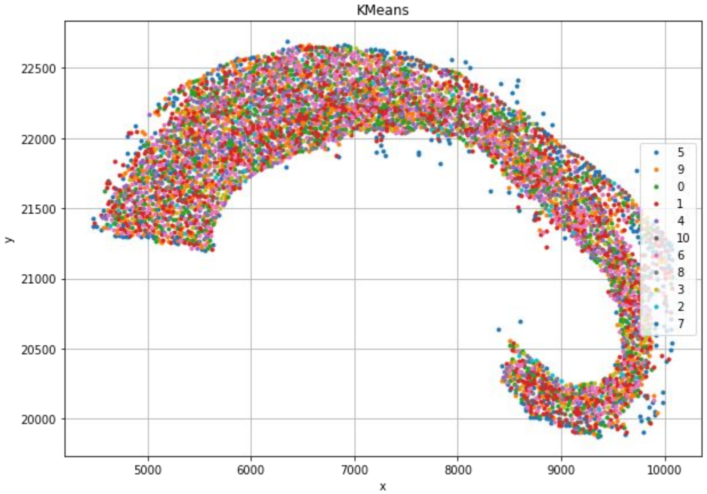
Partition-based clustering

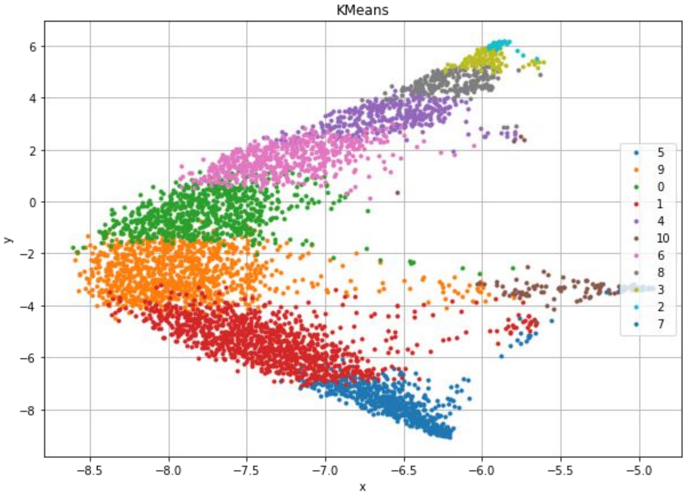
KMeans algorithm pipeline:

- specify number of clusters
- initialize centroids
- assign each point to the centroid via square distance

It is noticeable that ground truth image significantly differs from the image obtained via KMeans clusterization.





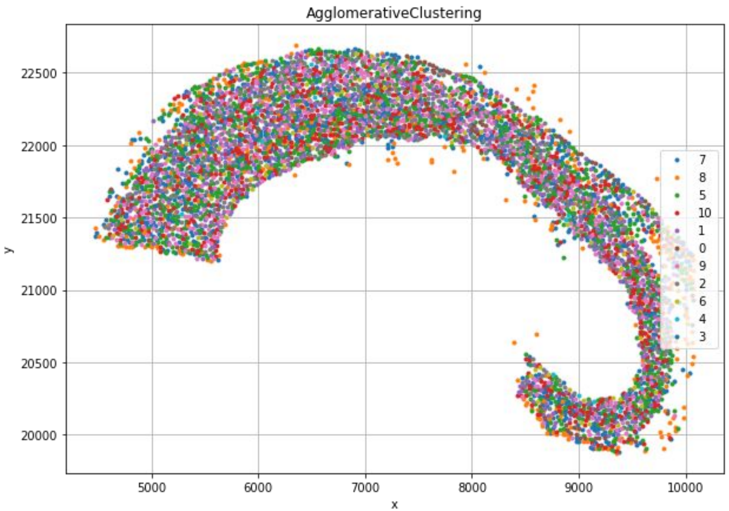


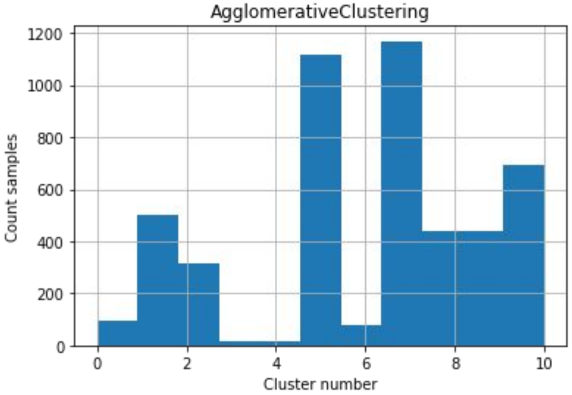
Agglomerative clustering

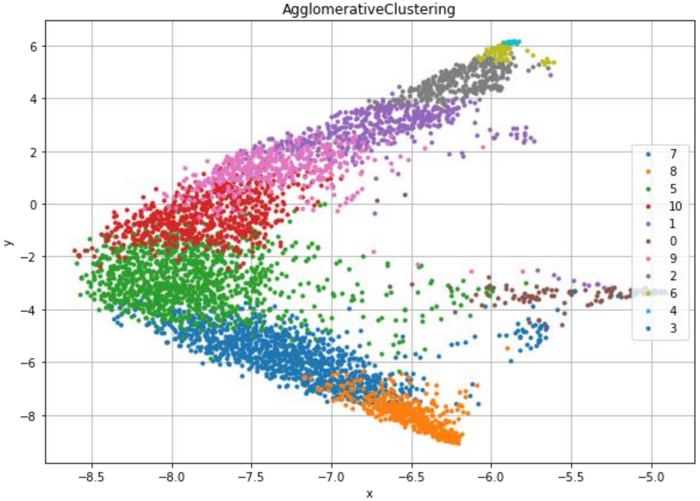
Hierarchical clustering

In this algorithm data is clustered in the bottom-up way.

This approach is also not successful.





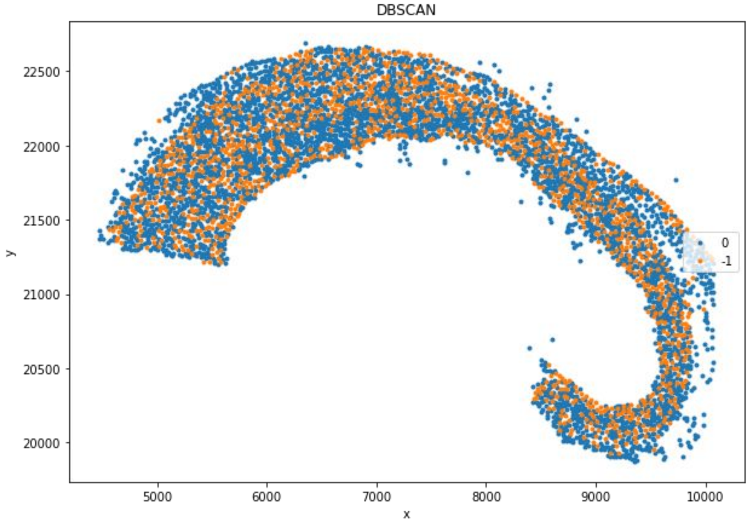


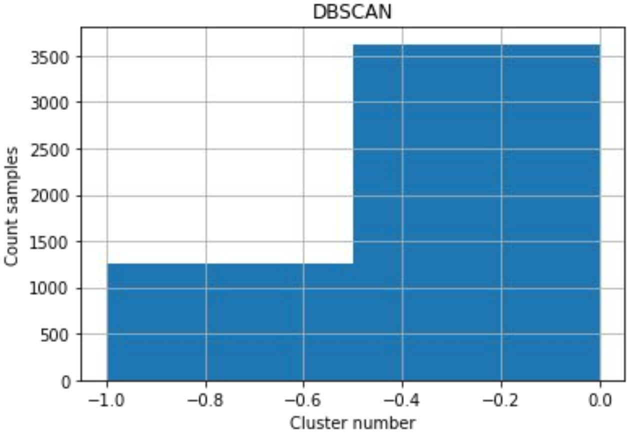
DBSCAN clusterization

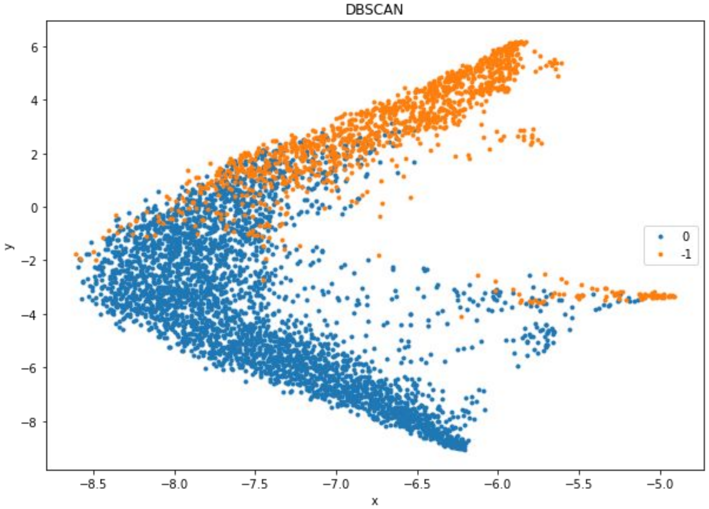
Density-based clustering

This algorithm tries to search the amount of clusters by itself, but it still needs distance between neighbours and minimum number of points to define cluster

DBSCAN algorithm has defined only 2 clusters.

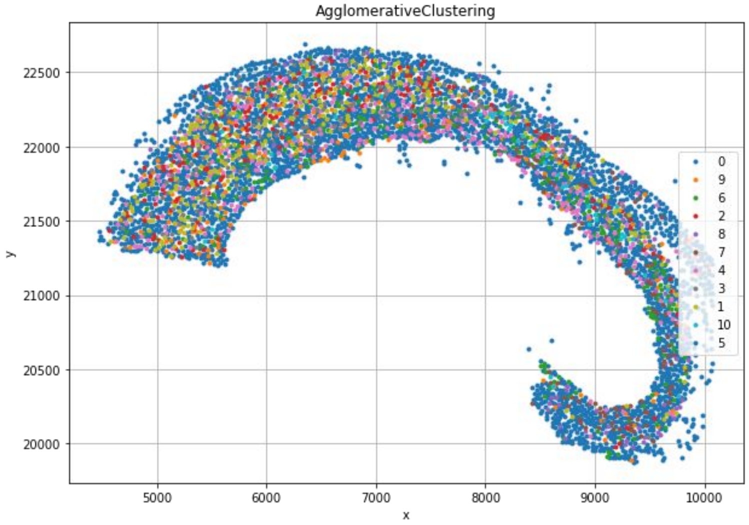


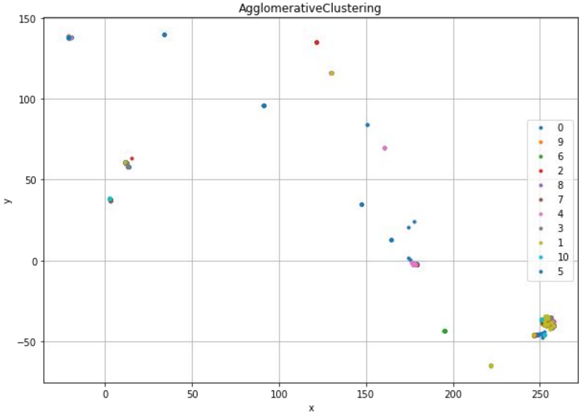




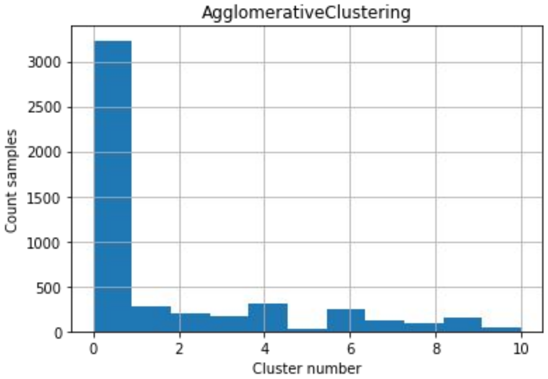
Gene filtering

I have selected several genes, that are the cell-type markers of neural cells. The main idea was to reduce the noise in data. Such filtering did not improve the results of clusterization.



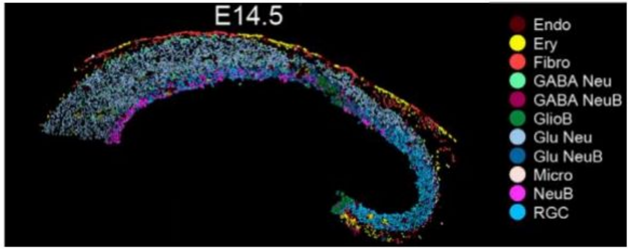


UMAP dimension reduction



List of genes: 'Mki67', 'Dbx1', 'Aldh1l1', 'Efna5', 'Npas3', 'Adgrb3', 'Ccnd2', 'Nes', 'Nrxn3', 'Gad2', 'Tenm3', 'Dcc', 'Nrxn1', 'Nrg3', 'Apoe', 'C1qc'

Explanation



All of the used algorithms has poor performance on such data. Here are some ideas why it happened:

1) Most of the methods **expect clusters to have a spherical shape,** but in case with tissues, cell clusters may have various shapes and sizes (Slides 4-6).
2) The majority of genes transcribed in cells do not define their type. There are only few marker genes that have increased or decreased expression during cell differentiation process. It means that data has **a lot of noise**. However, filtering noise is still not a solution (Slide 7)
3) The other possible difficulty is that the process of **cell differentiation** is observed. For instance, the difference between RGC and NeuB/GlioB in their GE levels is quite narrow at the moment of differentiation. It is better to compare the results obtained in previous time steps. In the above example, **E12.5** RGCs expressed higher levels of genes such as *Lix1, Ccnd1,* and *Gpc1*, and **E14.5** RGCs showed higher expression of *Nfix, Ndrg2, Aldoc*. These could be the markers of RGCs differentiation process. ([1],[2])

Possible solutions

1) Use Supervised methods of Classification
   - Seurat algorithm (two options - supervised and unsupervised)
   - scPred
2) Machine learning algorithms (gene expression data + coordinates)
   - Graph Convolutional Networks
   - Transformer Neural Networks