Tissue cells clustering by stereo-seq data

Anton Eliseev

Aim and issues

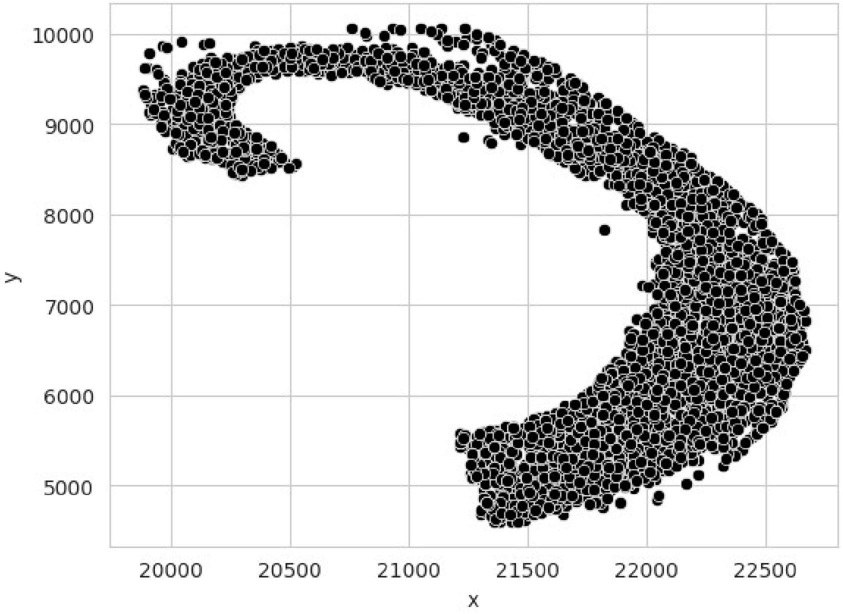
The **aim** of this project to provide clusterization method for cells of midbrain tissue by its stero-seq data

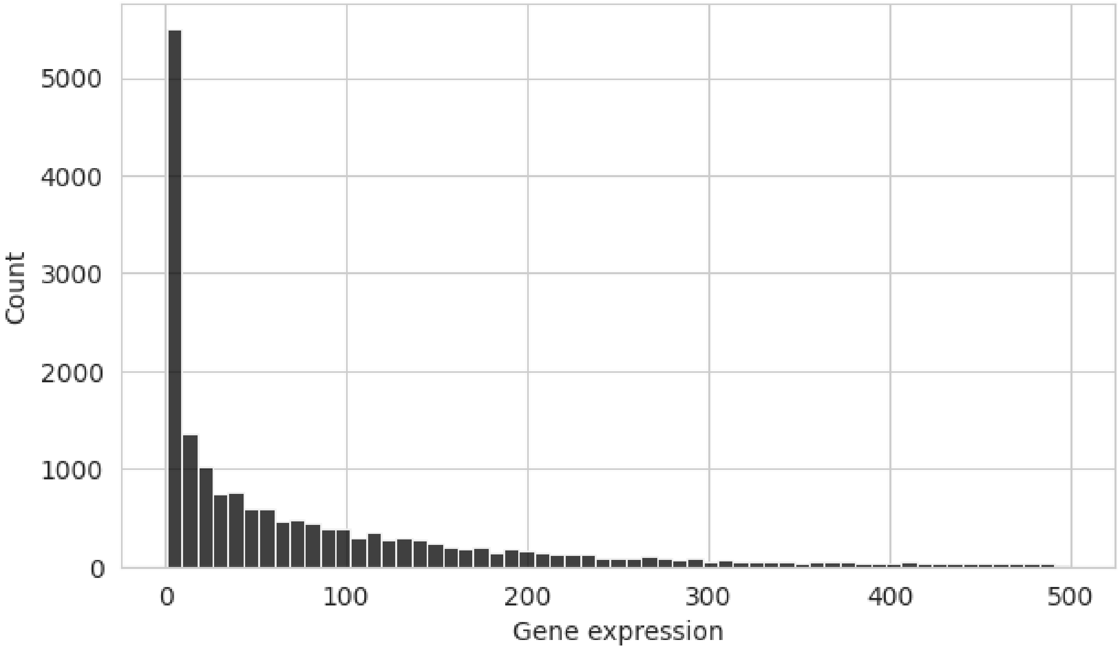
Issues:

- Review and preprocess stereo-seq data
- Use classical unsupervised approaches on preprocessed data with and without predefined cluster amount
- Develop a new algorithm of stereo-seq data clusterization based on clusterization of weighted graph
- Apply the new algorithm to midbrain tissue stereo-seq data
- Propose other possible algorithms and ways to improve presented approaches

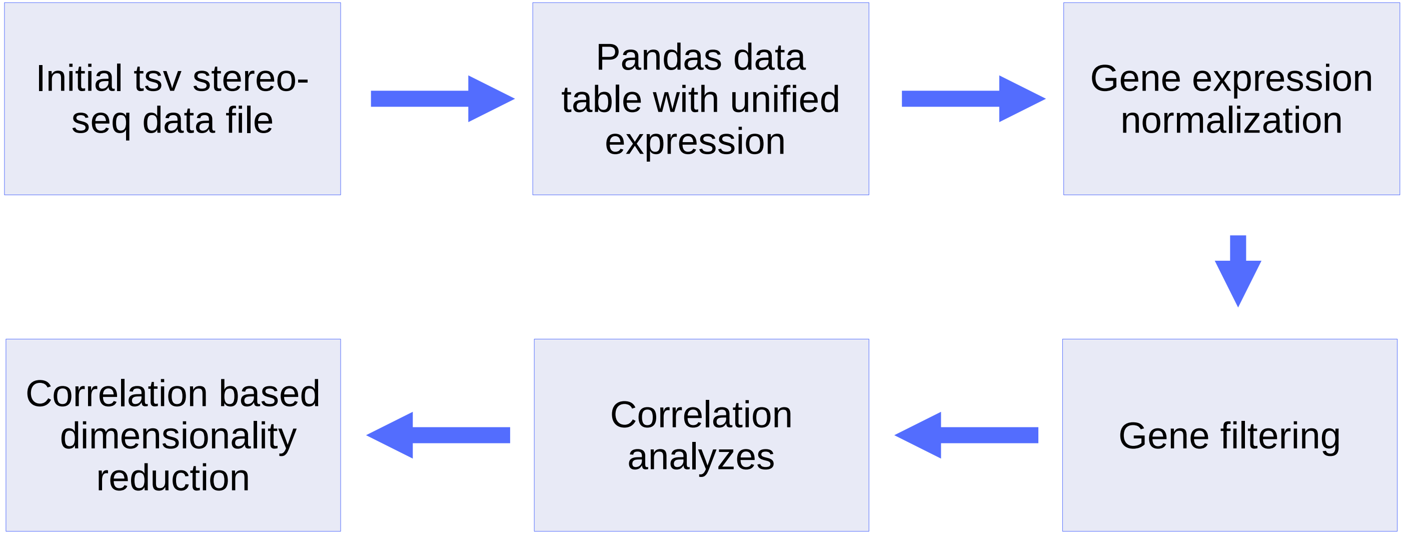
Data review

The data includes cell ids, its coordinates in tissue and gene expressions. The tissue sample and gene expression looks like it’s presenter of following scatterplot and histogram. As it shown expression distributed exponentially and it has to be log normalized.





Data preprocessing pipeline



Initial tsv stereo-seq data file

Pandas data table with unified expression

Gene expression normalization

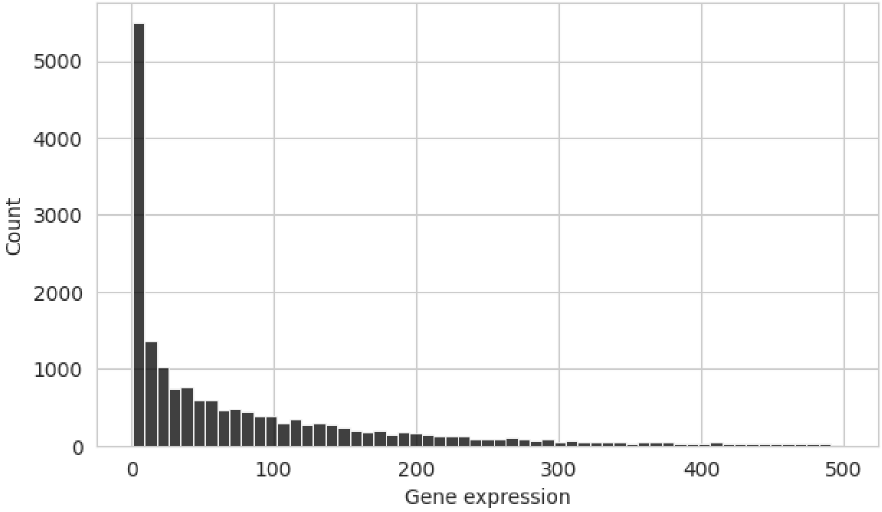
Gene filtering

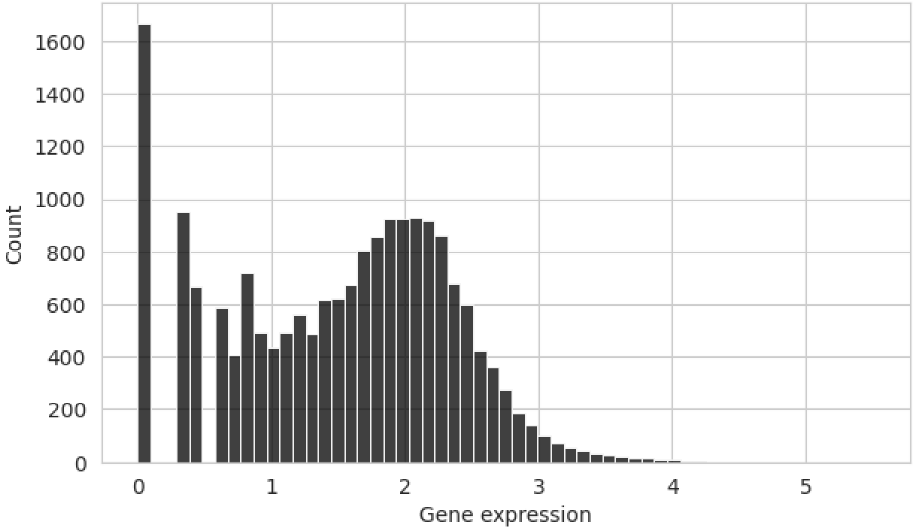
Correlation analyzes

Correlation based dimensionality reduction

Normalization

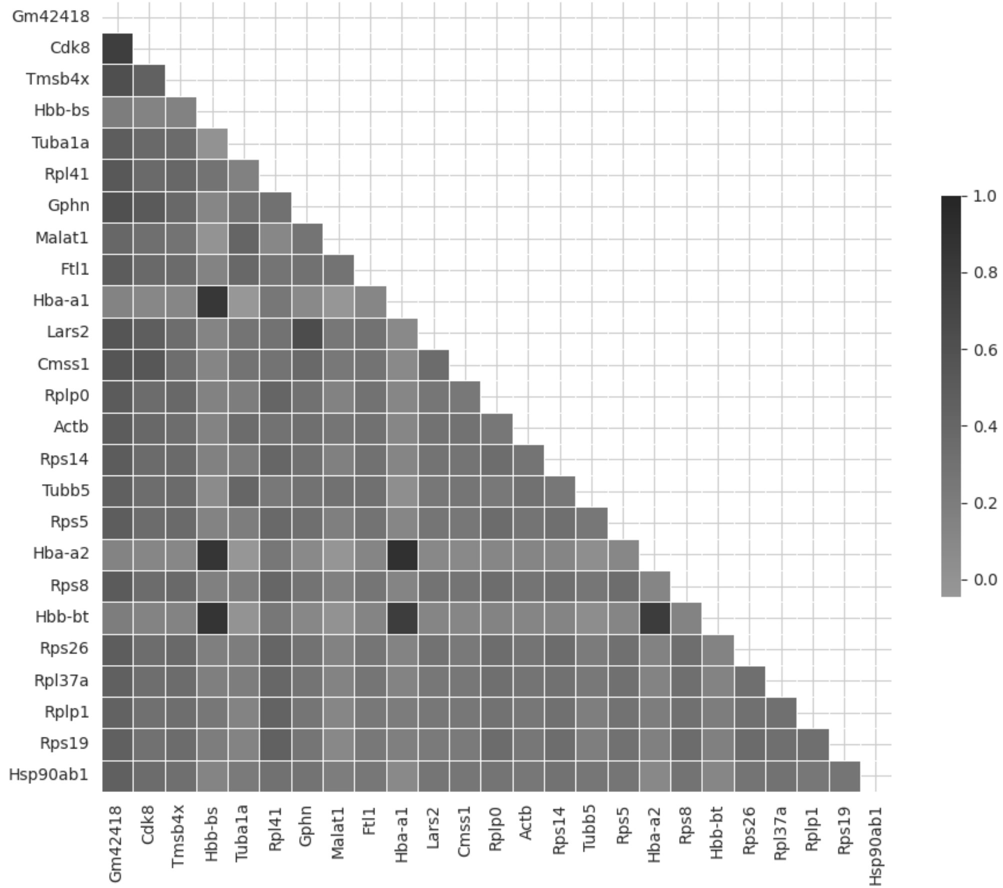
After log normalization (details in notebook data_preprocessing) gene expression looks like mix of two distributions, exponential and Poisson, let’s suppose than exponential data provided by errors (very rare genes), so the first step is removing rear genes with expression coefficient lower then 1. Such filtering removes approximately 30% of genes with less then 1% of common expression.





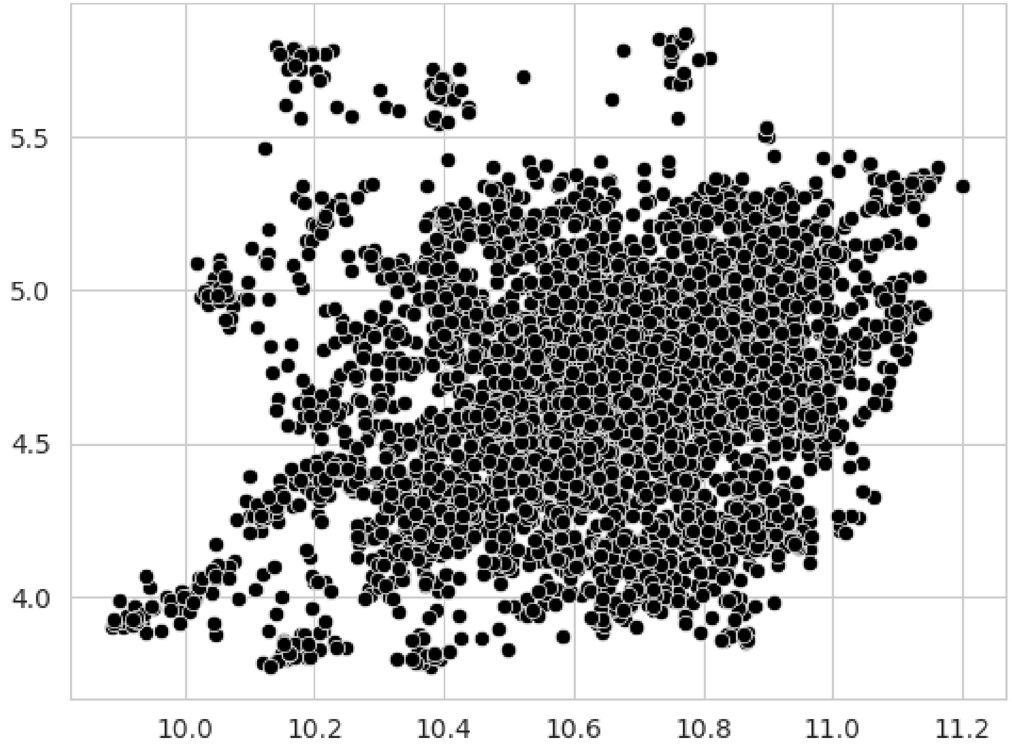
Correlation analyzes

As it shown on the correlation map even 25 the most expressed genomes are correlated. It means that their expression in different cells remains the same and thus the expression of such genes not provides information to differ cells. Moreover, there is 14e3 genes, which means hight dimension of features. So the next step is dimensionality reduction.



Dimensionality reduction

By applying UMAP algorithm to reduce dimensionality to 20 features by using correlation as a metric we have reduced dataset with uncorrelated features. On the scatterplot two first components visualized.



Correlation analyzes

Reduced features could be used to find correlation with initial features to find the most significant initial features for UMAP algorithm. Here is the information about the mosto correlated

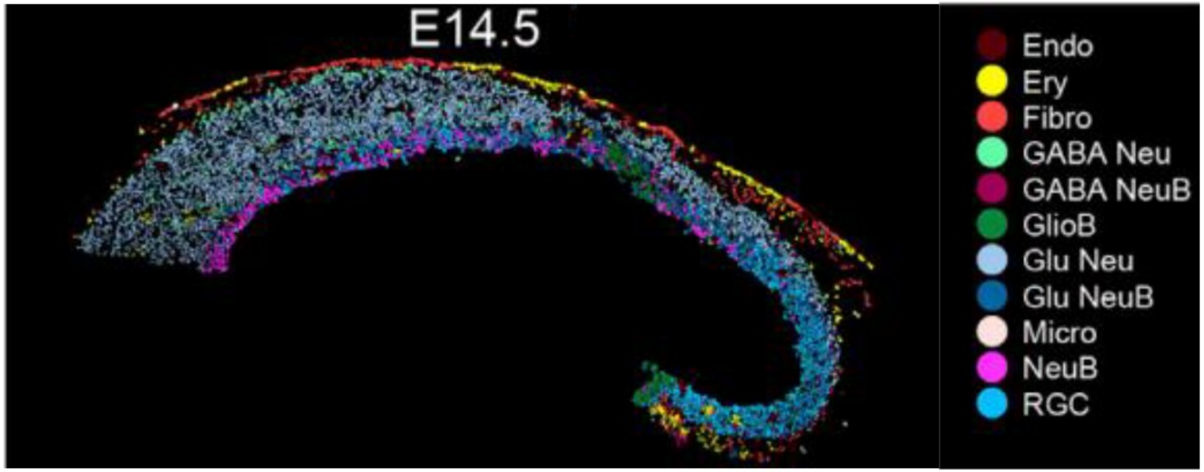
- Rpl6: This gene encodes a protein component of the 60S ribosomal subunit. This protein can bind specifically to domain C of the tax-responsive enhancer element of human T-cell leukemia virus type 1, and may participate in tax-mediated transactivation of transcription. .
- Lamtor5: As part of the Ragulator complex it is involved in amino acid sensing and activation of mTORC1, a signaling complex promoting cell growth in response to growth factors, energy levels, and amino acids.
- Polr2e: This gene encodes the fifth largest subunit of RNA polymerase II, the polymerase responsible for synthesizing messenger RNA in eukaryotes. This subunit is shared by the other two DNA-directed RNA polymerases and is present in two-fold molar excess over the other polymerase subunits.
- Ipo5: Nucleocytoplasmic transport, a signal- and energy-dependent process, takes place through nuclear pore complexes embedded in the nuclear envelope.

All of this genes are pretty different by their functionality, probably it's a reason why they are more informative for UMAP algorithm.

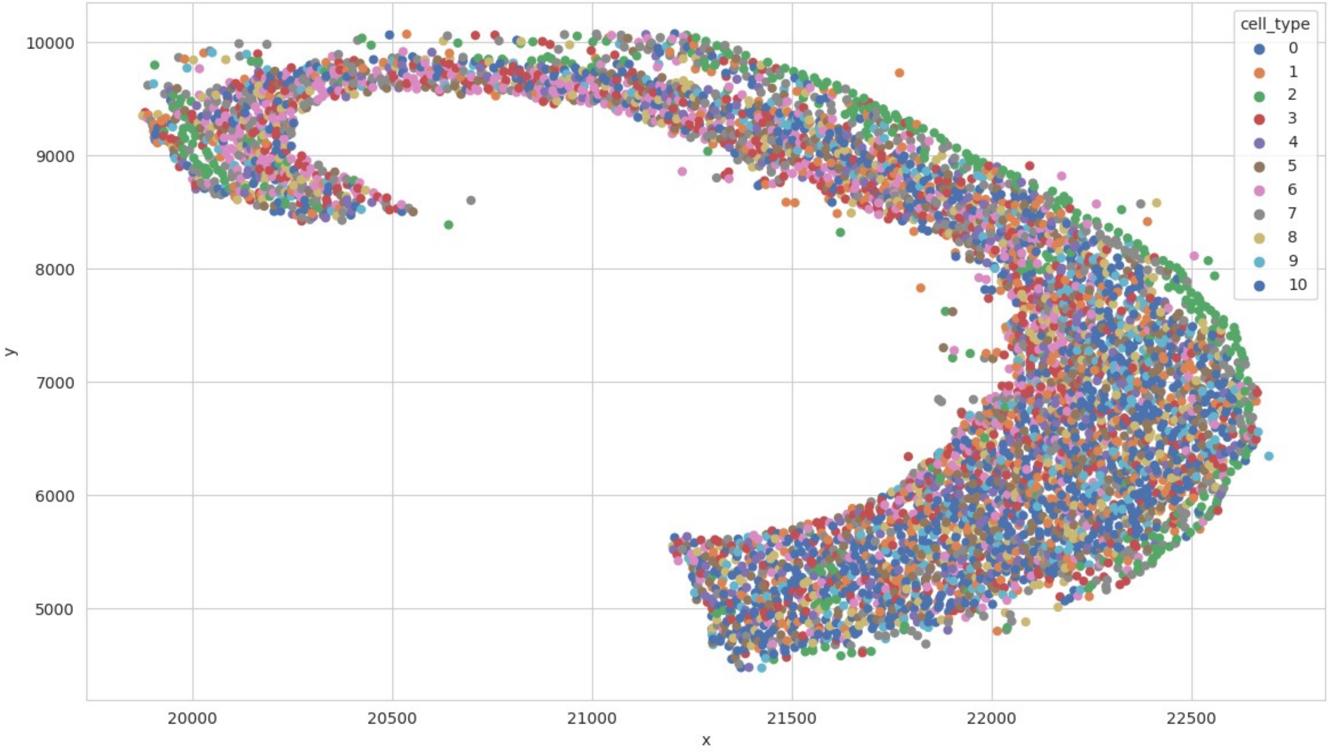
Unsupervised clusterization

Reduced features could be used to clusterization by such methods as k-means and DBSCAN with and without predefined cluster amount correspondingly.

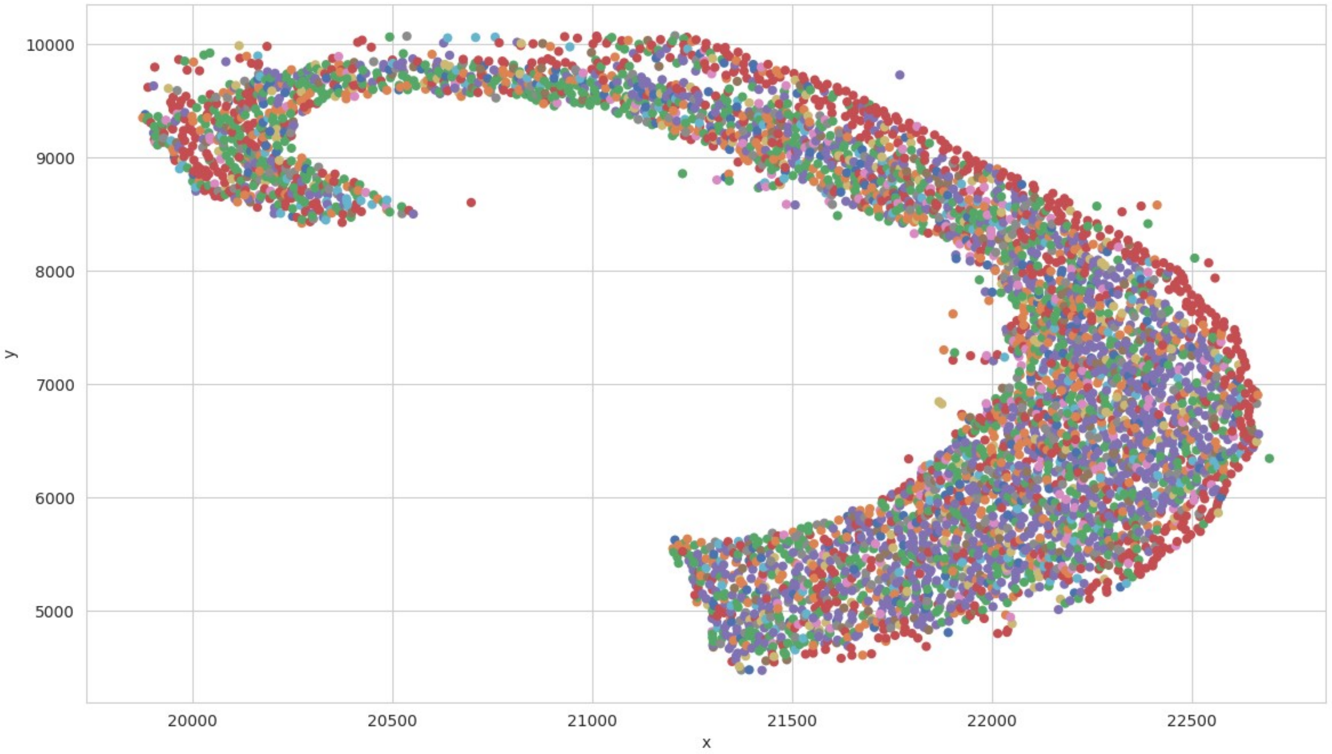
- For k-means algorithm been used k = 11 as it shown on image from the test issue
- DBSCAN clusterization applied to data with information about coordinates and without



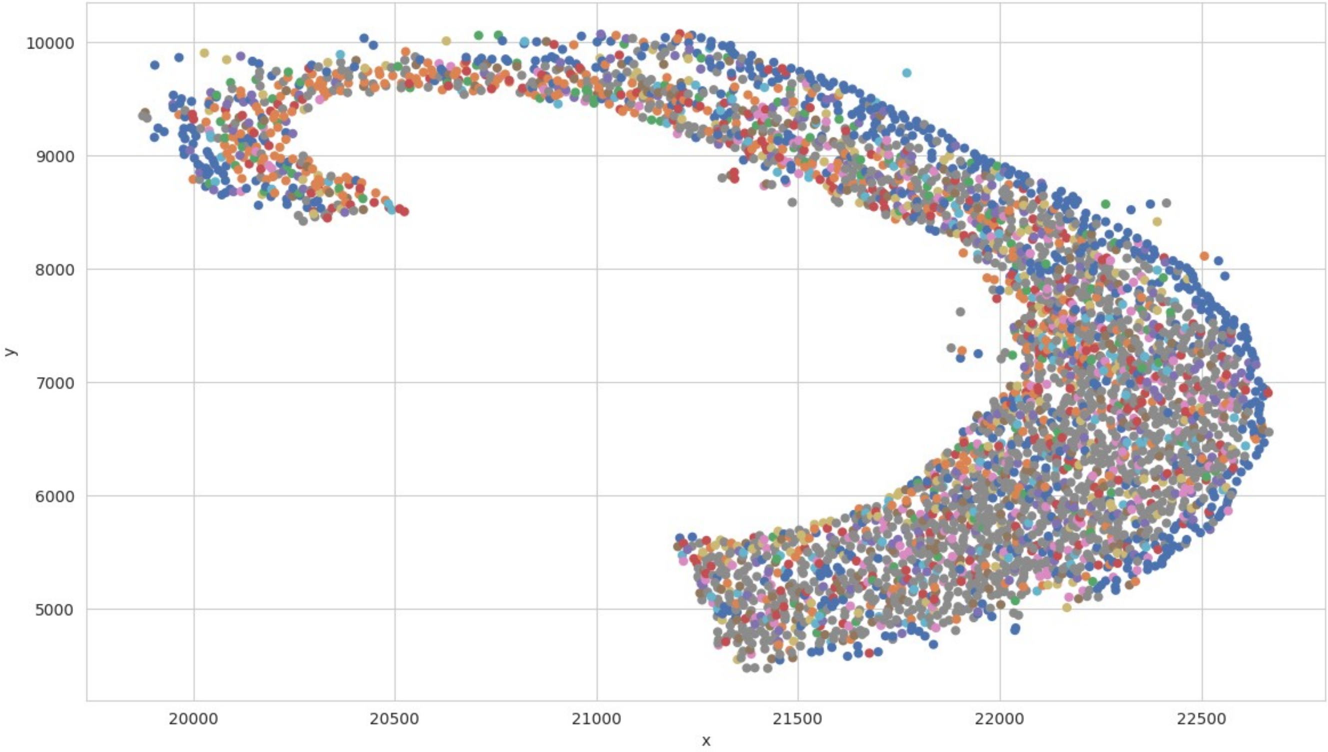
K-means result



DBSCAN without coordinates



DBSCAN with coordinates



Unsupervised clusterization

- All methods allows to detect Endo and Ery types of cell and to differ them
- DBSCAN clusterization provides less then 11 clusters and information about coordinates doesn’t help a lot
- Other cells types except Endo and Ery detects not precise and resulting image is noisy
- Probably UMAP dimensionality reduction is not the best way to do that and DL embeddings trained on labeled data can work better
- Probably it is possible to use T-sne, PCA or any other dimensionality reduction data instead of UMAP
- Information about gene expression and coordinates might be used in a better way based on probabilistic approach as it shown further

Gene expression graph

Let's define expression graph as weighted undirected graph where each node is a cell from tissue and edge u, v exist if u, v cells gene expressions are similar, and besides weight of (u, v) proportional to expression similarity and real(euclidean) distance between cells.

- Probability of edge based on distance:
  Half normal cumulative distribution function with maximal distance to nearest neighbor selected as a variance

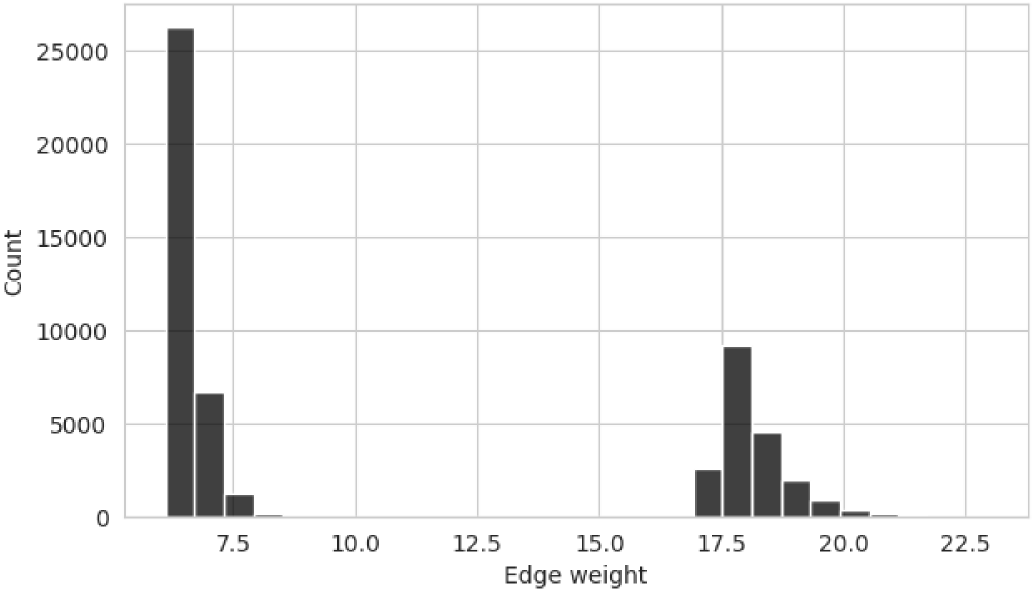
$$P_{dist}(d)=\frac{1}{2}\left[1+erf\left(\frac{d}{\sigma\sqrt{2}}\right)\right]$$

- Probability of edge based on shared gene expression
  For given cells c1, c2 and their expression e1, e2 probability to have an edge defined like 1 minus difference between expression by following equation

$$P_{expression}(e_1,e_2)=\frac{1-\sum_{k=0}^{n}\frac{\left|e_{1,k}-e_{2,k}\right|}{e_{1,k}+e_{2,k}}}{\sum_{k=1}^{n}e_{1,k}\neq 0+\sum_{k=1}^{n}e_{2,k}\neq 0}$$

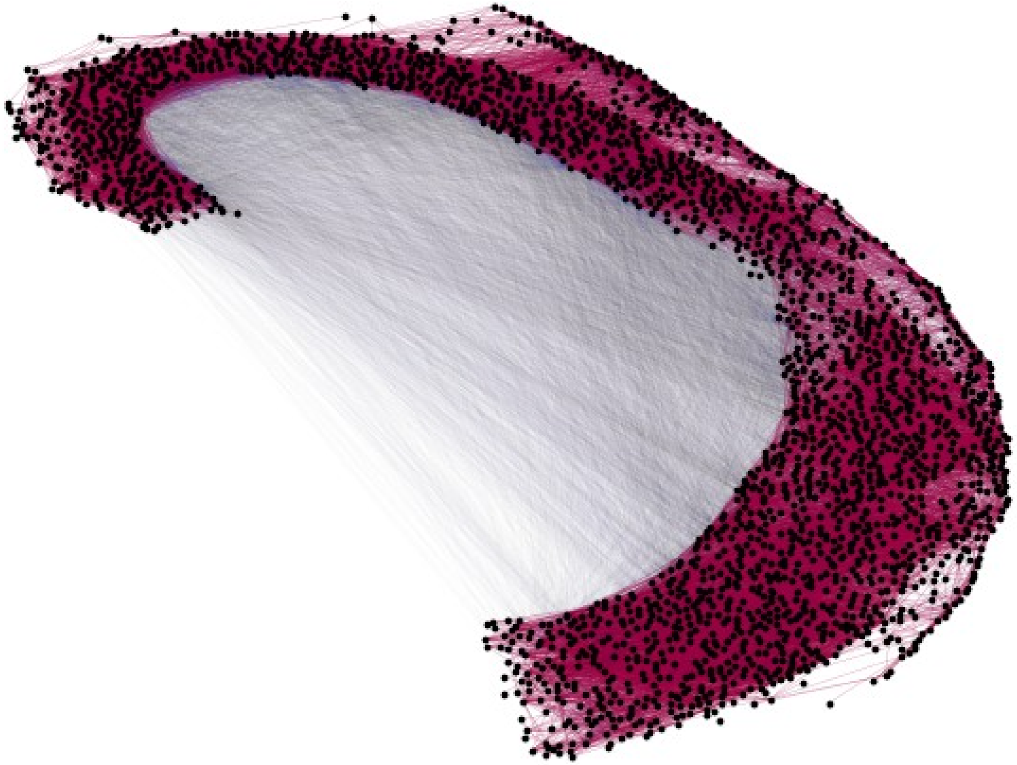
Gene expression graph

Using corresponding probabilities two type of edges could be inserted into the graph. One the one hand there are edges between very close cells with similar expressions, with respect the suggestion that cells with the same type are often close to each other at least because of cell division process, on the other side there are cells which are similar by their expression but separated in space because of fetus development process for example. We can observe two types of edges on weights histogram



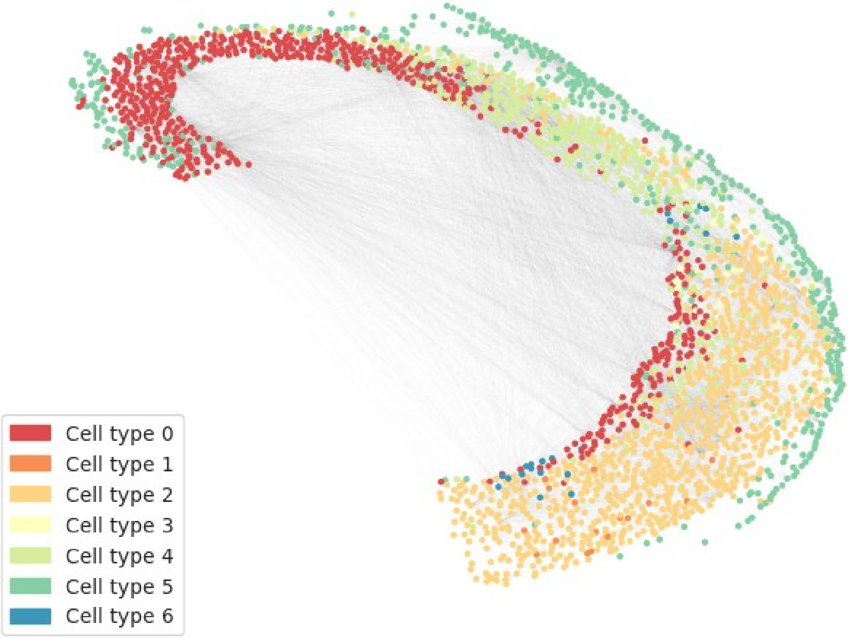
Gene expression graph

Two types of edges can be observed on graph as well



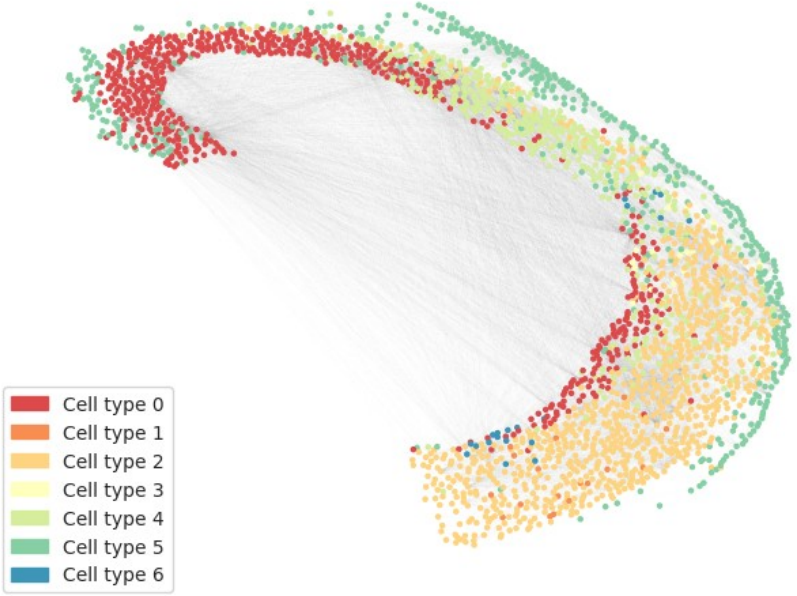
Gene expression graph clusters

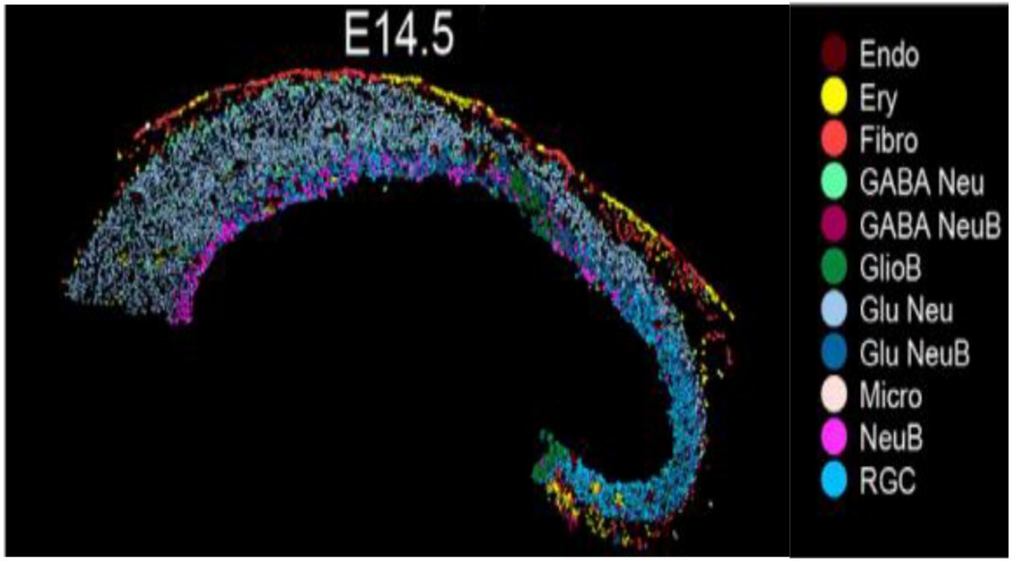
Two find clusters into big graphs with respect to edges weights one of the best algorithm is Louvain community detection algorithm. All details are explained here for example https://en.wikipedia.org/wiki/Louvain_method. Here this algorithm is applied to clusterize midbrain tissue cells into gene expression graph.



Gene expression graph clusters

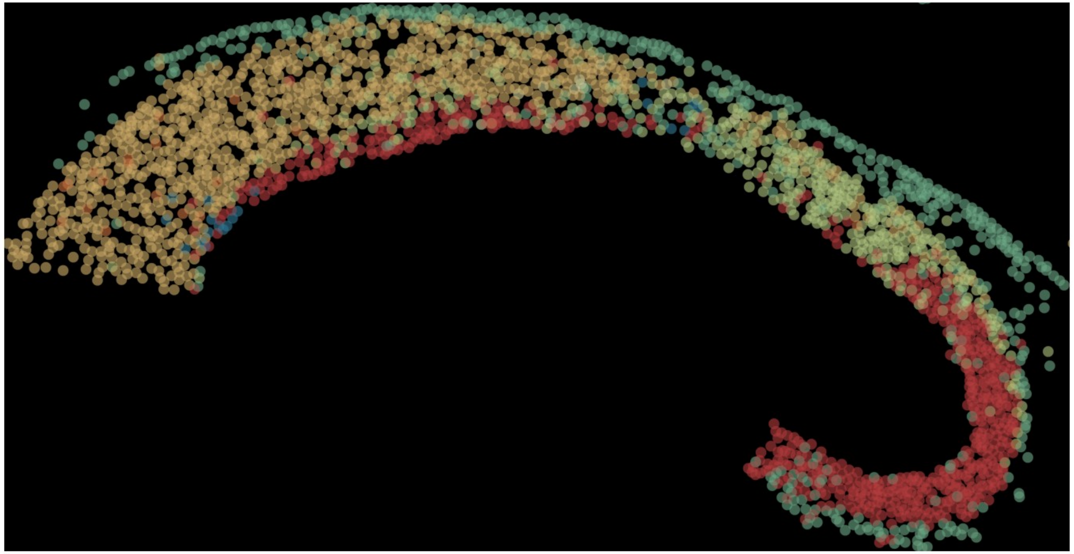
As it shown on following images the developed algorithm performs well to differ such cells types as RGC, Gliu Neu, GlioB, Endo/Ery

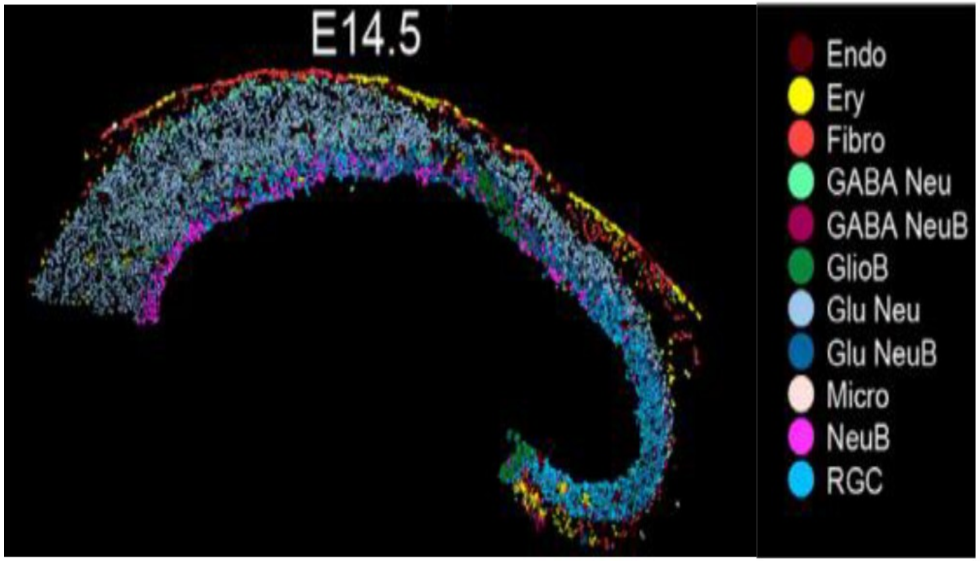




Gene expression graph clusters

As it shown on following images the developed algorithm performs well to differ such cells types as RGC, Gliu Neu, GlioB, Endo/Ery





Conclusion

Results:

- Classical approaches performs quite not well on stereo-seq data to compere with new developed algorithm based on gene expression graph.
- Edges probability based on distance and on gene expression allows to use them flexible and validate different hypothesis about expression and distance importance. More over, weights of this edges might be used for improvement of classical approaches performance.

Space for improvements:

- Improve probabilities, for example use more natural way for gene expression probability
- Use approaches based on MCMC, such as are used for active module detection
- Use labeled data to train embedding model
- Train GNN model on labeled and validated gene expression graph to predict cell types
- Spend more then couple of weekend days on this issue :)