Topological analysis of the Human Prefrontal cortex

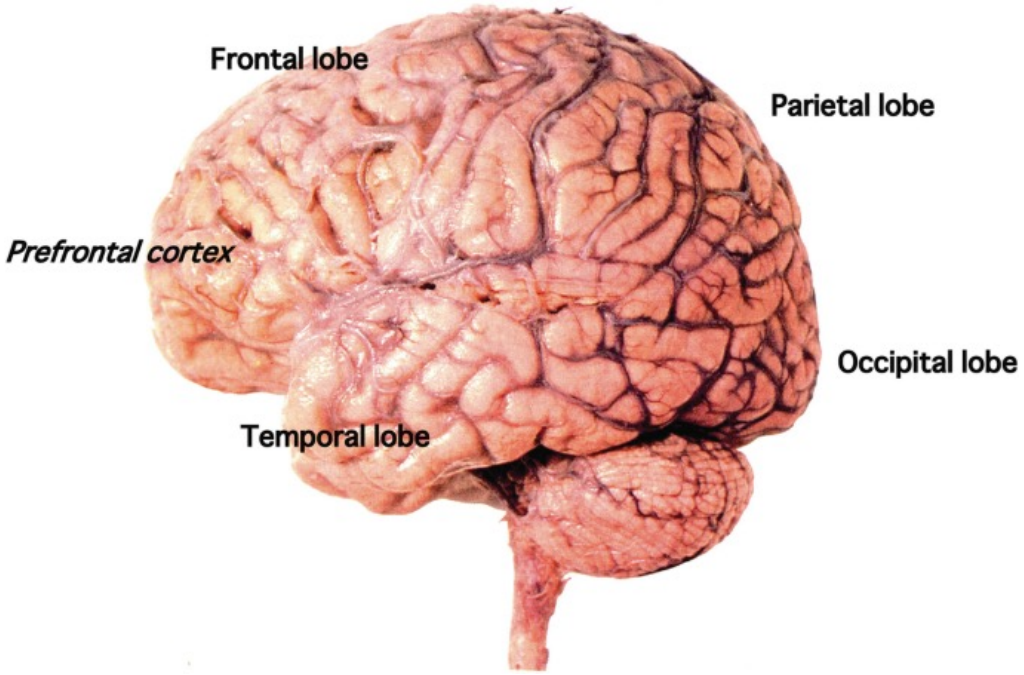
CSIC5011 Final project

LEE Jooran, WANG Minghao

Mammalian prefrontal cortex

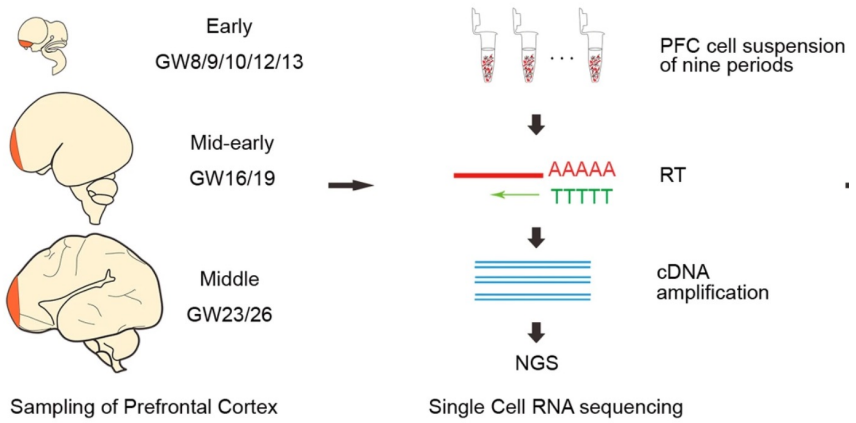
- A highly specialized brain areas containing billions of cells and serves the centre of the highest-order cognitive functions.

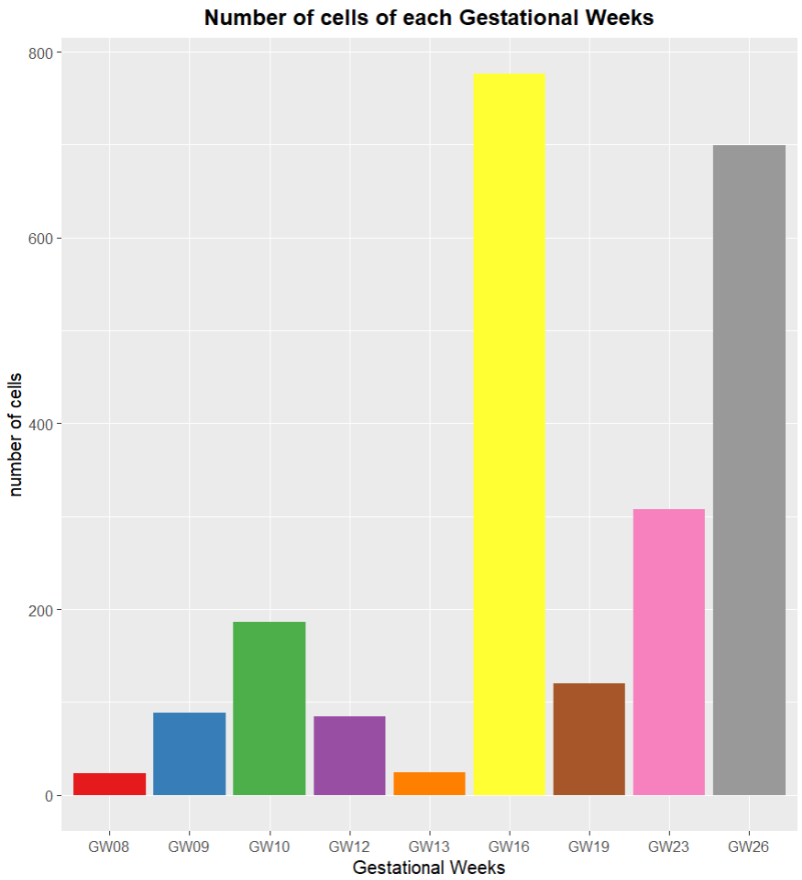
- Dysfunction of the prefrontal cortex contributes to cognitive deficits, however, it is still difficult to identify cell types in the developing human prefrontal cortext and to distinguish their features.



Dataset and software

- 2,309 single-cells in the developing human prefrontal cortex from gestational weeks (GW) 8 to 26 using RNA sequencing.
- TPM (Transcripts per million) count matrix of the 24,153 matrix was used as a gene expression dataset.
- Software: R, Python





	GW8	GW9	GW10_01	GW10_02	GW10_03	GW12	GW13	GW16	GW19	GW23_01	GW23_02	GW26	Sum
Gender	Female	Female	Male	Female	Female	Male	Female	Female	Female	Male	Female	Female	
Sequenced cells	23	88	48	95	48	88	24	789	120	143	181	747	2,394
Filtered cells	23	88	47	92	47	85	24	776	120	132	176	699	2,309

Research questions

- Which type of plot is the optimal for the data projection of the human PFCs?
- What is the major cell types in the developing PFCs?
- Differences in trajectory analysis results from several different approaches?
- What are the key genes in regulating the prefrontal cortex in the developmental trajectories?

Topological data analysis: Original data and Foundation Model embeddings

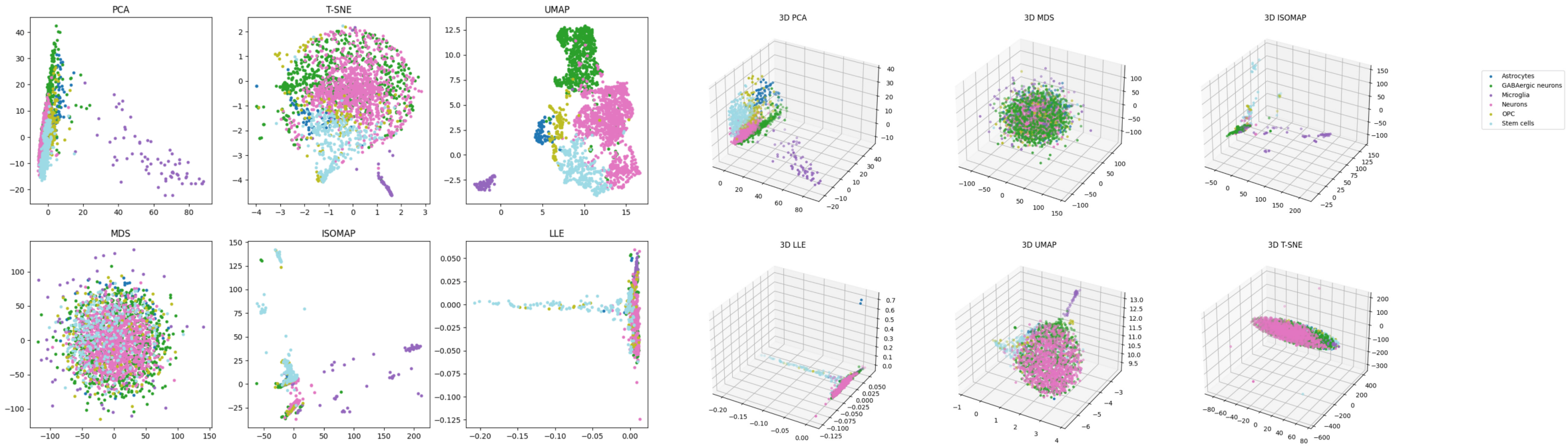
- What is foundation model?

A foundation model is a type of **large-scale machine learning model** that is pretrained on **extensive amounts of data**. These models have the ability to **adapt to a wide variety of tasks** without needing to be retrained from scratch.

- What is foundation model embedding?

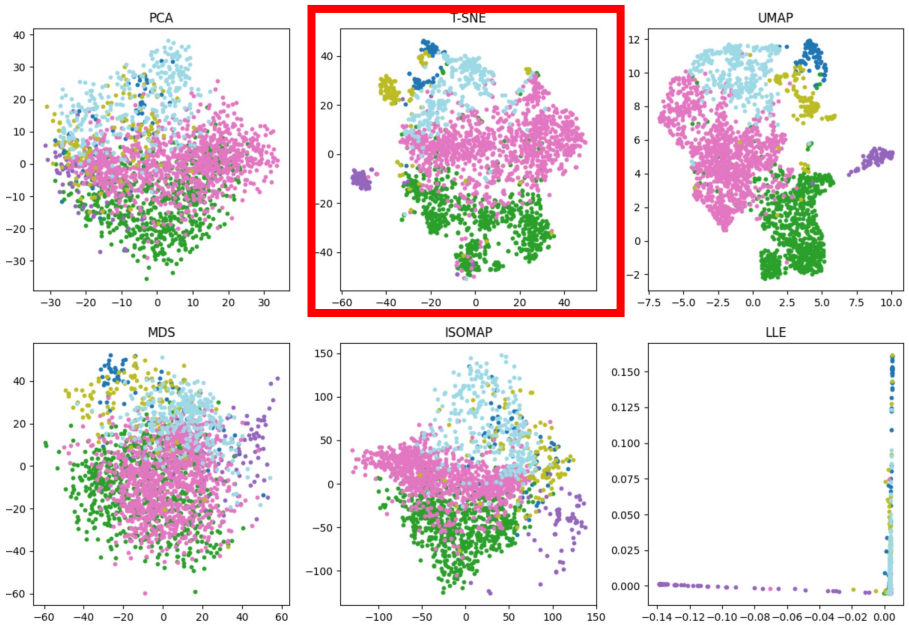
Foundation model embeddings are the **dense numerical representations** that these models generate from input data. An embedding is a vector that represents a word, sentence, or document in a way that captures its semantic properties. These embeddings allow the model to understand and process the input data effectively.

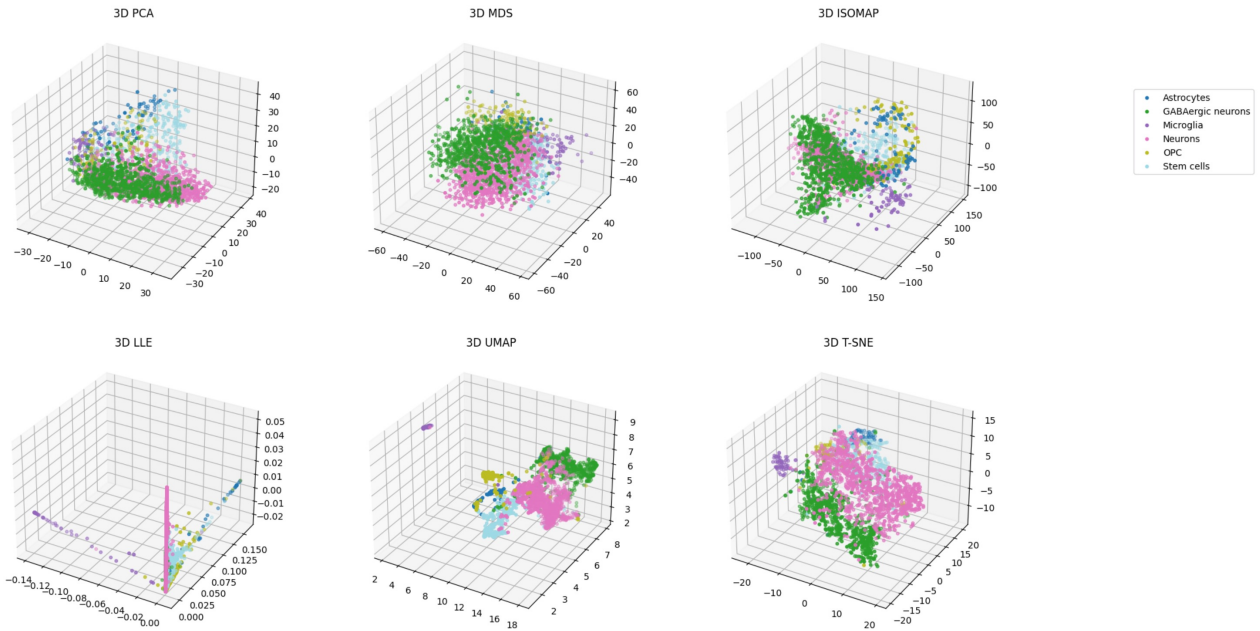
Different topology methods on original data



Only UMAP is the best without any parameter tuning.

Different topology methods on Foundation model embedding





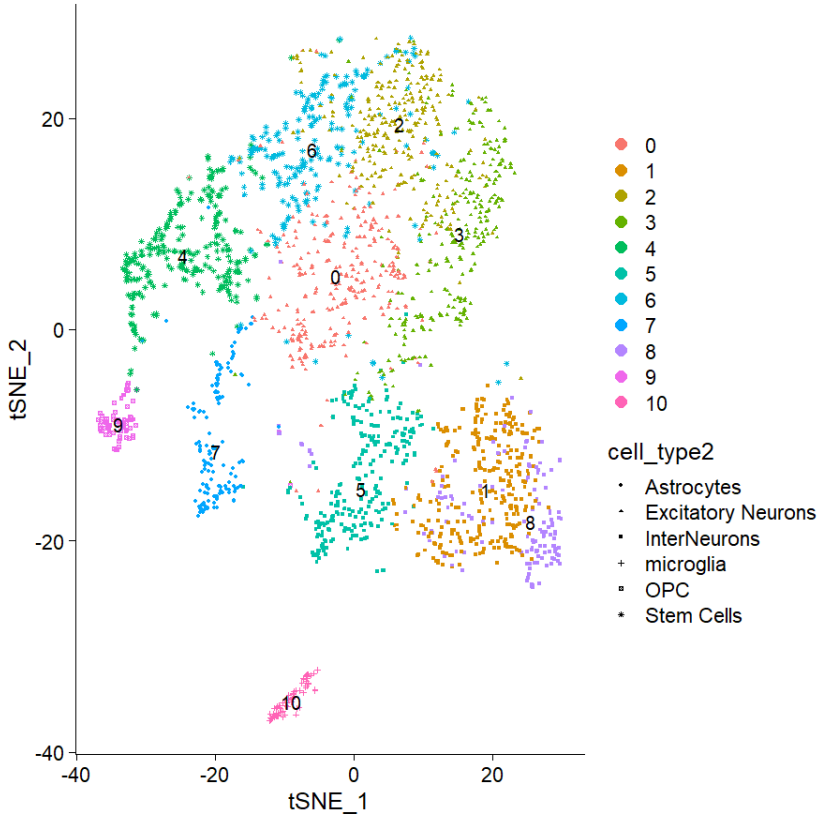
1. Foundation model embeddings can automatically project the data into a unified space such that no permutation are required on topological analysis.

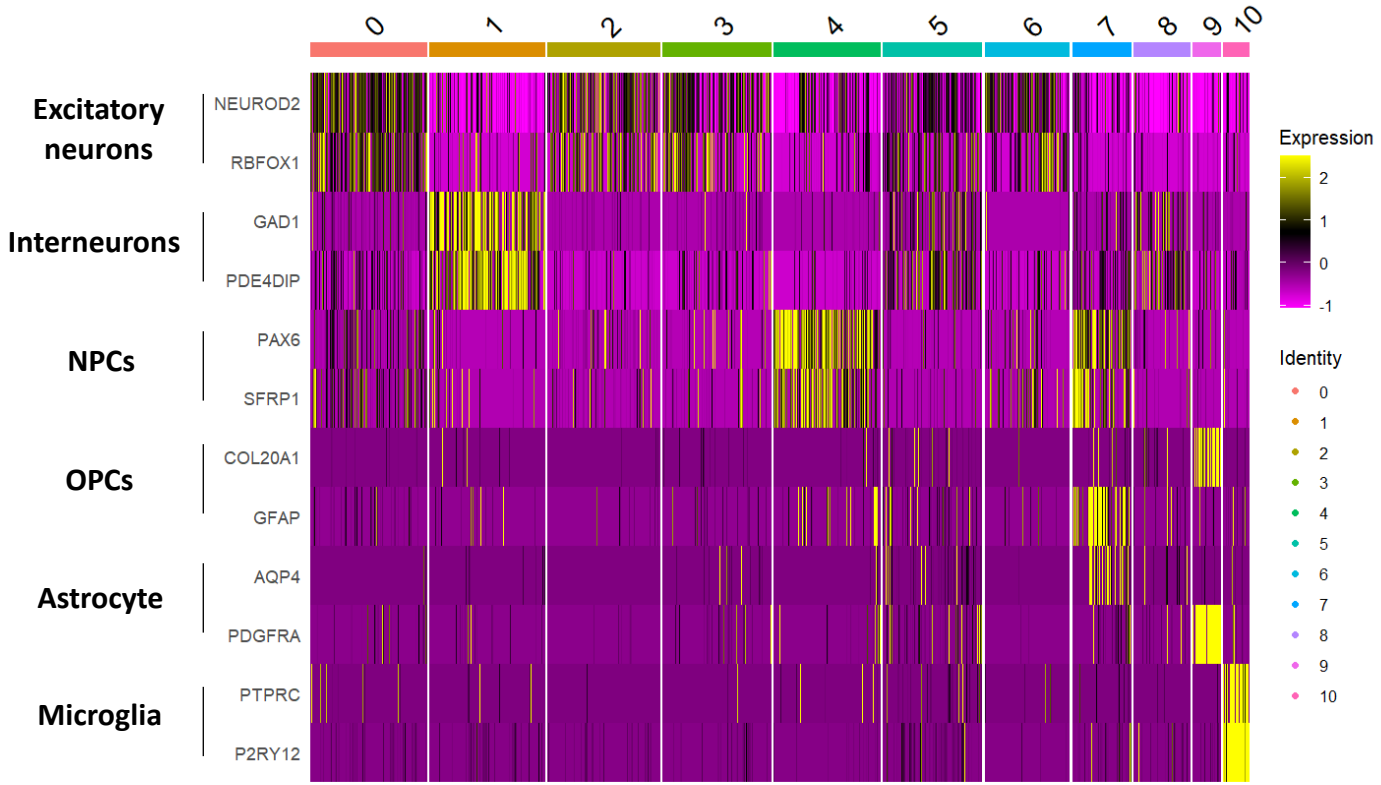
2. T-SNE and UMAP are the most suitable methods for biological data analysis.

Identification of the cellular subgroups

- To classify the major cell types in the developing PFC, the t-distributed stochastic neighbour embedding (T-SNE) analysis using Seurat was done and identified six major clusters.

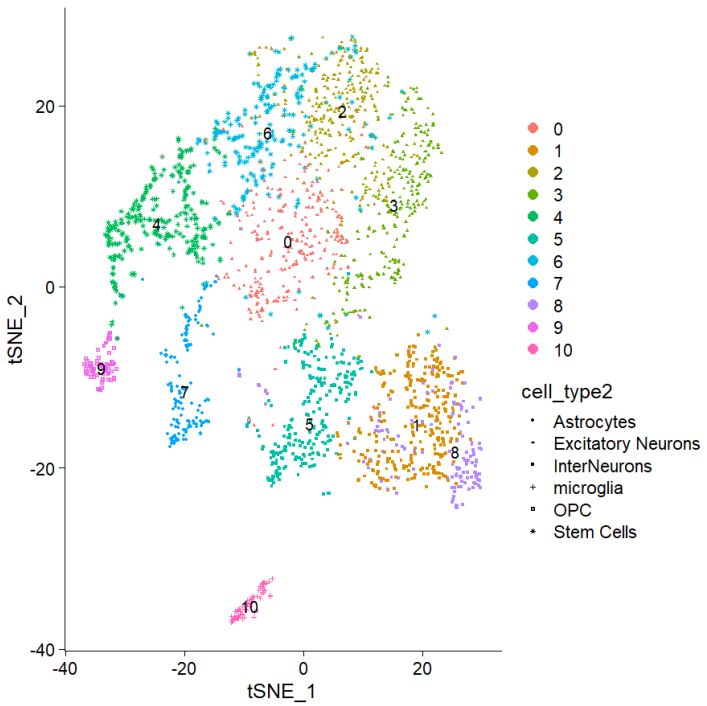
- The T-SNE plot was plotted after dimensional reduction, where the cell-types were manually labelled based on the marker gene expressions.

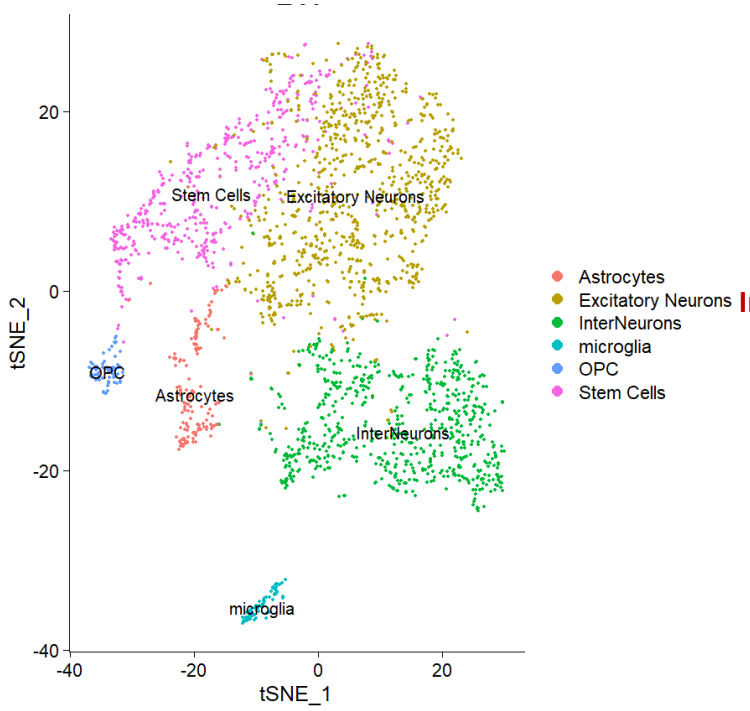


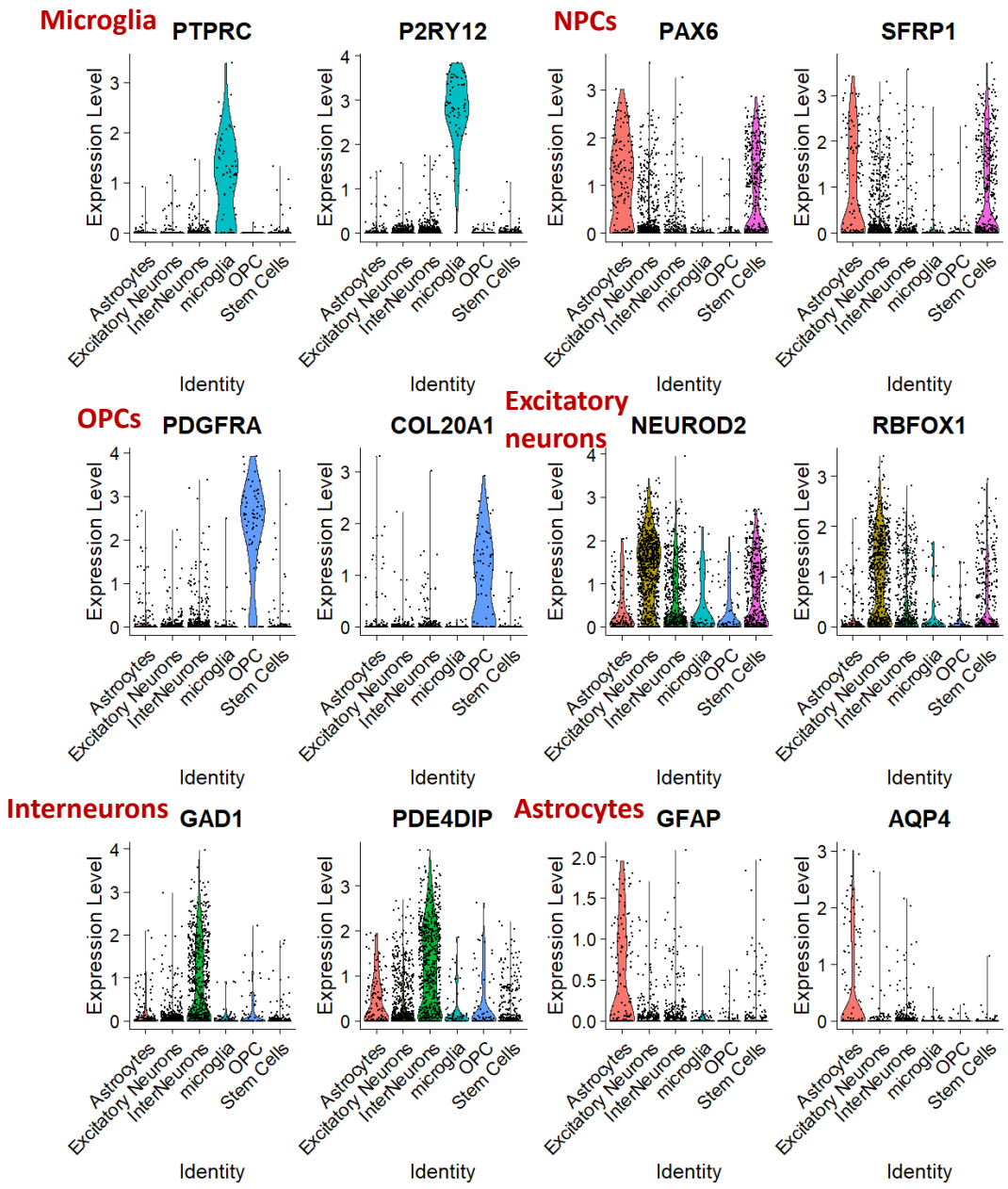


Cell-type labelling of the subgroups

- The T-SNE plot was plotted after dimensional reduction, where the cell-types were manually labelled based on the marker gene expressions.



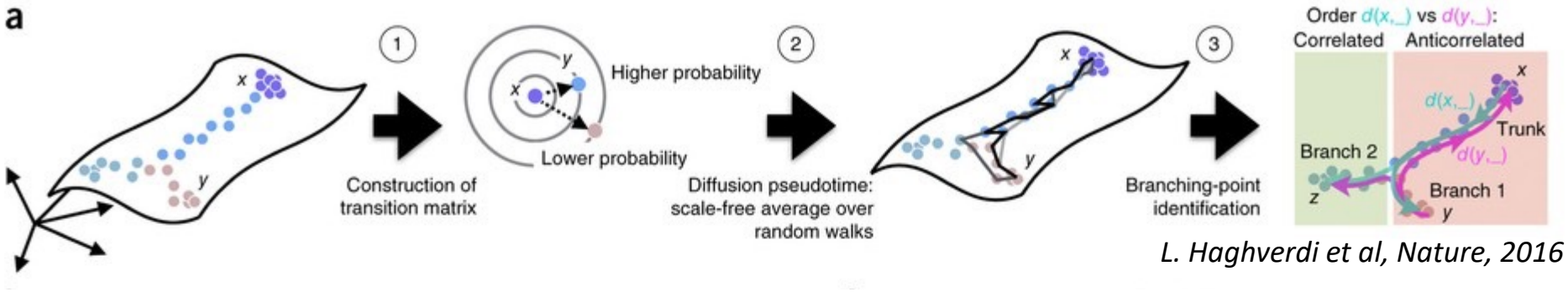




Trajectories inference methods

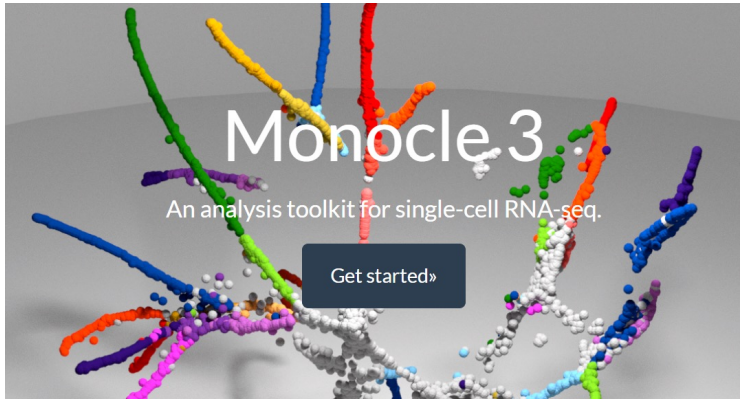
the measure of how much progress an individual cells has made through a process such as cell differentiation.

1. Diffusion pseudo-time



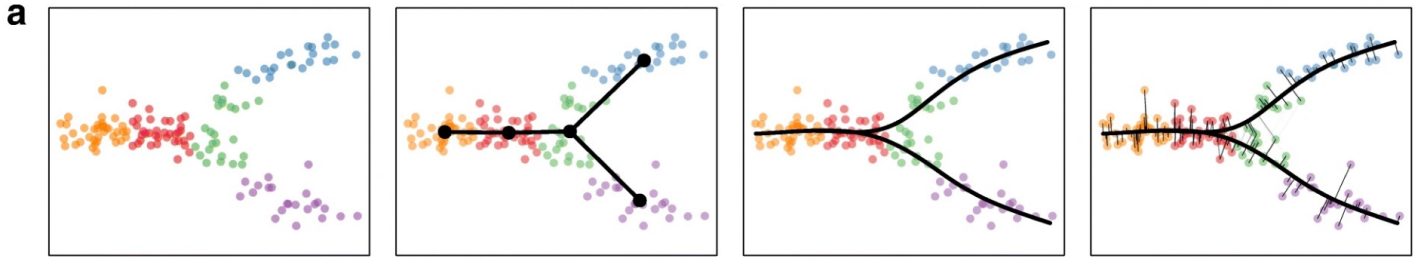
L. Haghverdi et al, Nature, 2016

2. Monocle3



J. Cao et al, Nature, 2019

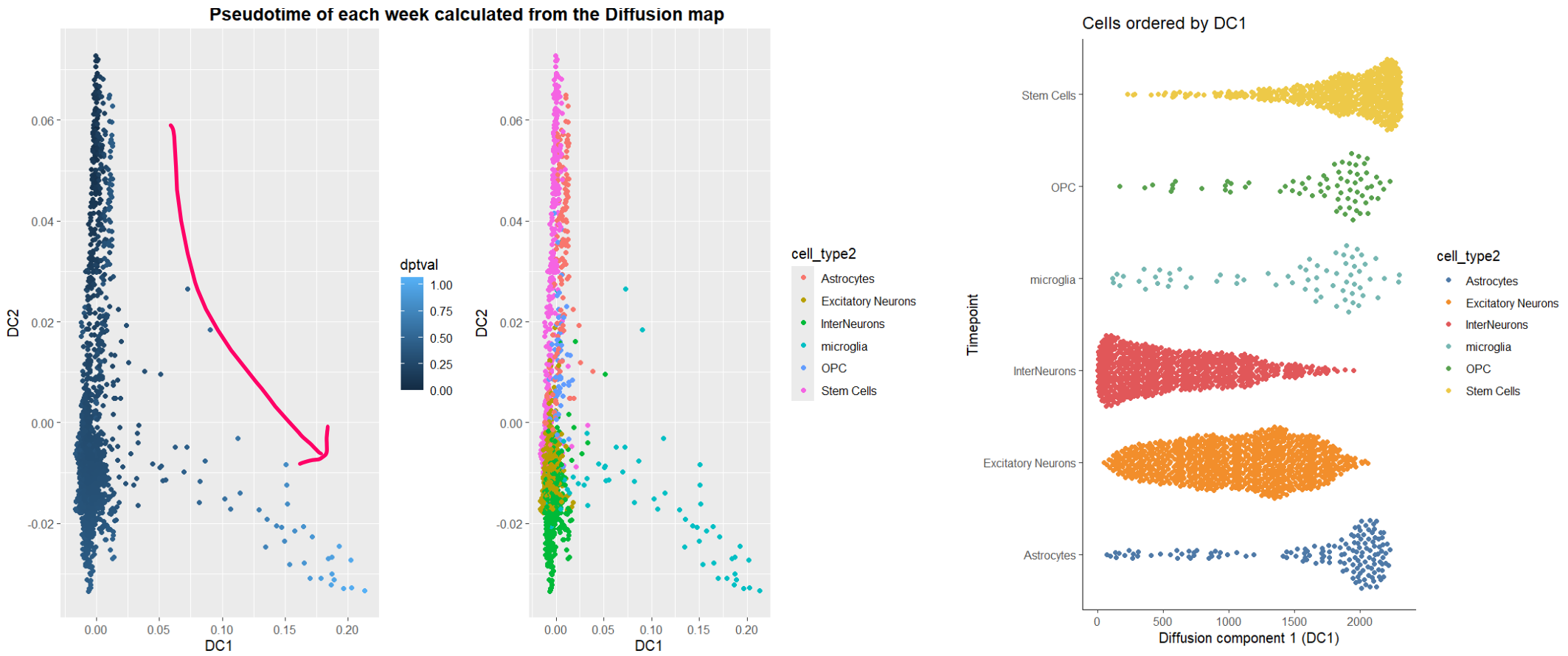
3. Slingshot



K. Street et al, BMC Genomics, 2018

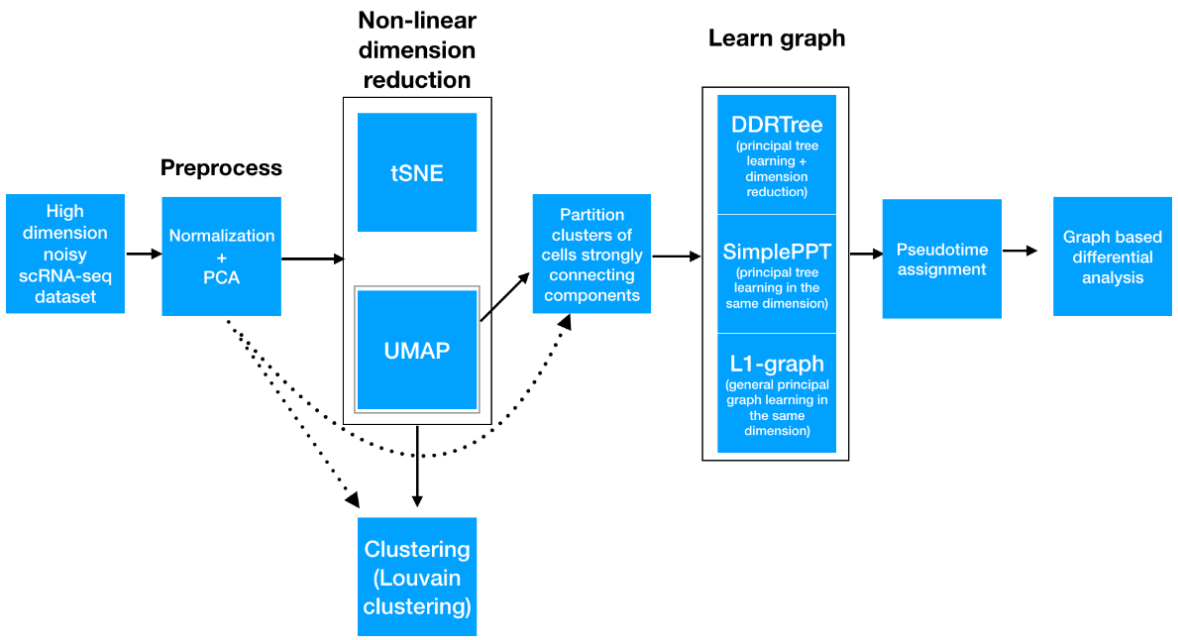
Trajectories inference using Diffusion map

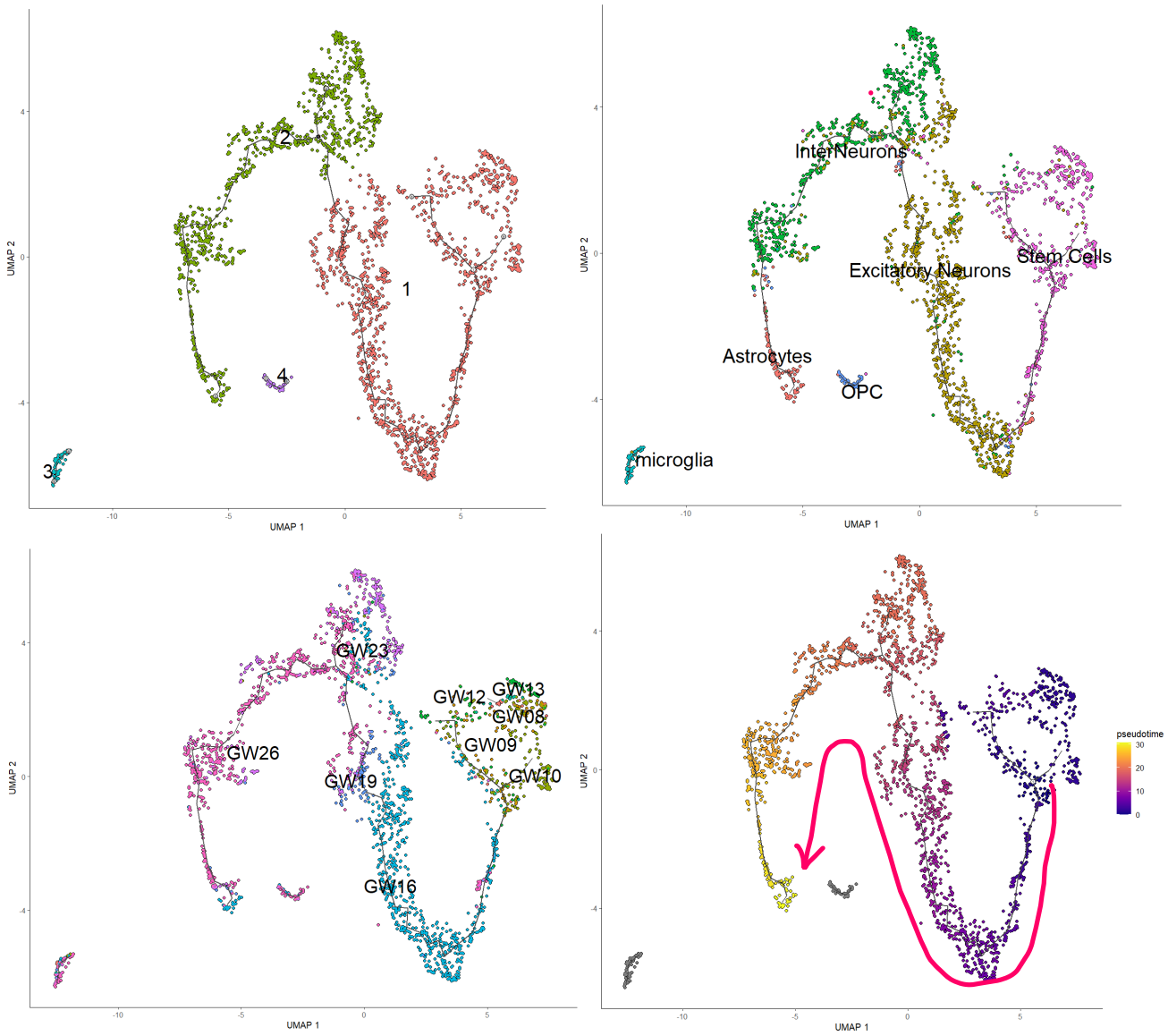
- **Diffusion map pseudo-time** performs at placing cells along the expected differentiated trajectory.
  - This method assumes that the data are sampled from a diffusion process, and it infers the low-dimensional manifold by estimating the eigenvalues and eigenvectors for the diffusion operator related to the data.
  - However, none of the PCA/UMAP/T-SNE effectively segregated clusters with different cell-types.



Trajectories inference using Monocle3

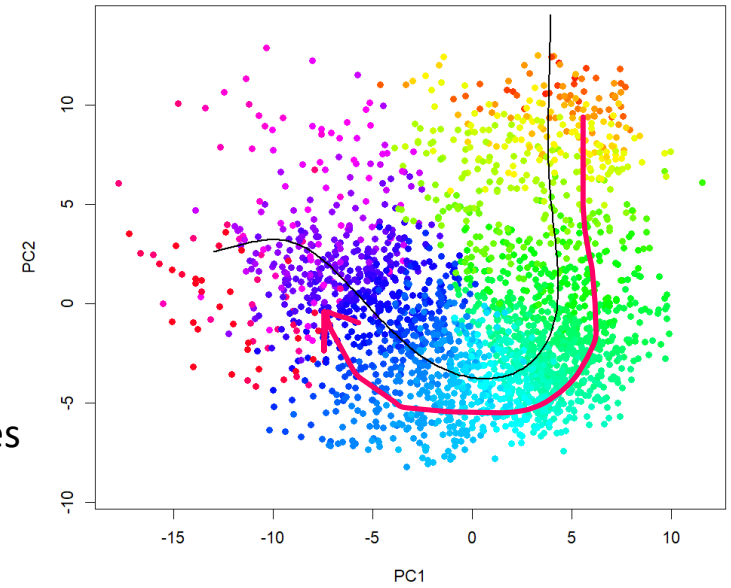
- After the non-linear dimensional reduction step, the Monocle3 provides three different ways to organize cells into trajectories, all of which are based on the concept of “principal graph embedding”.

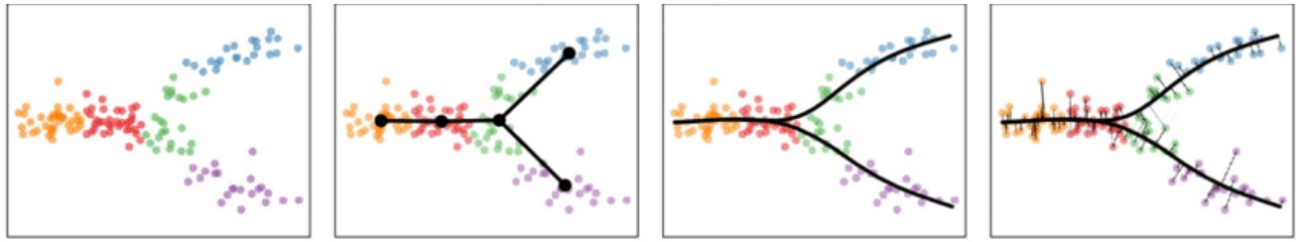




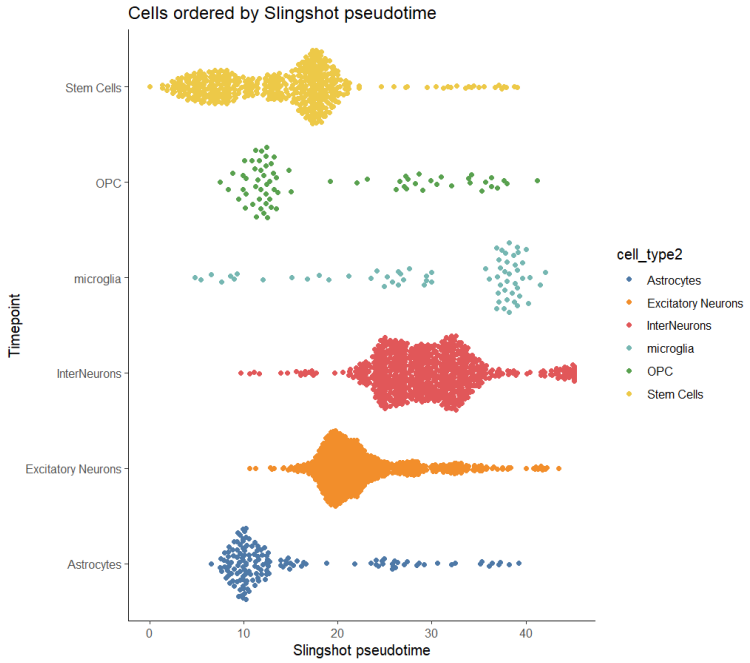
Trajectories inference using Slingshot

- The goal of slingshot is to use clusters of cells to uncover global structure and convert it into smooth lineages represented by one-dimensional variables, called “pseudo-time”.
    - Step1) minimum spanning tree is constructed on the clusters
    - Step2) obtain the smooth representations of each lineage
    - Step3) pseudo-time values are calculated by orthogonal projection onto the curves



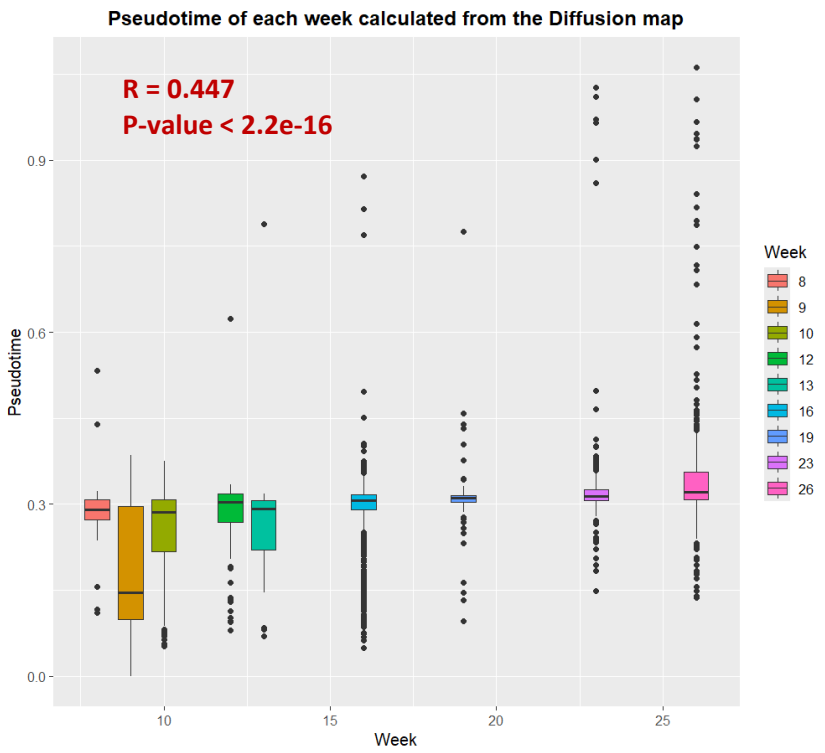


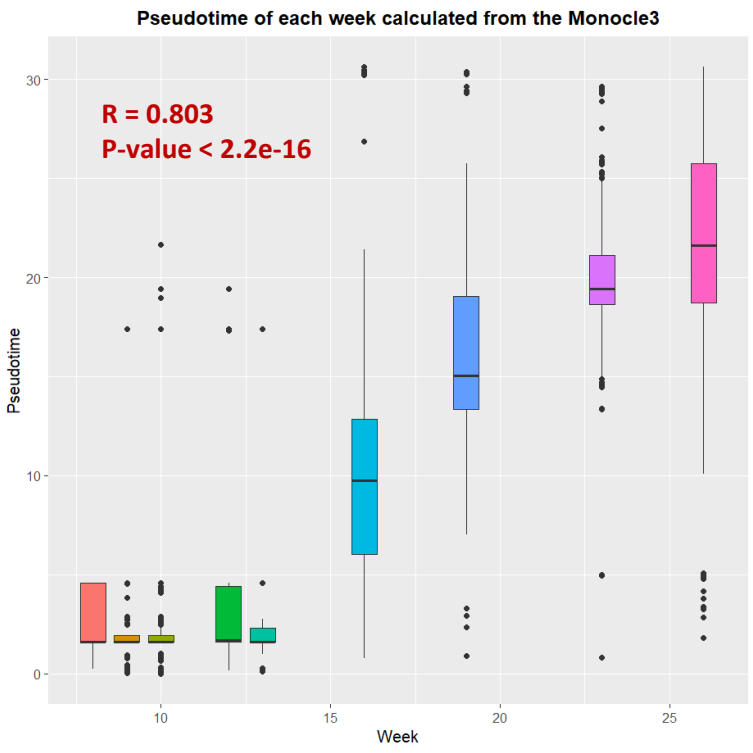
K. Street et al, BMC Genomics, 2018

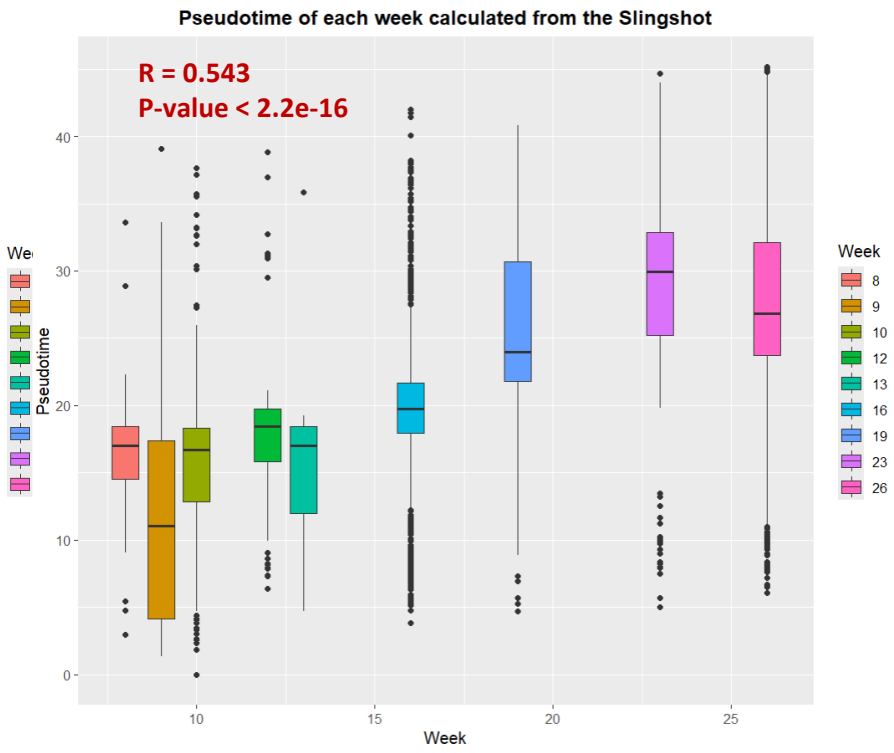


- Some of the cell-types, such as “Stem cells”, “OPC”, “microglia”, and “Astrocytes” were not effectively clustered, resulting in sparse clustering results.

Comparison of the Trajectory analysis results

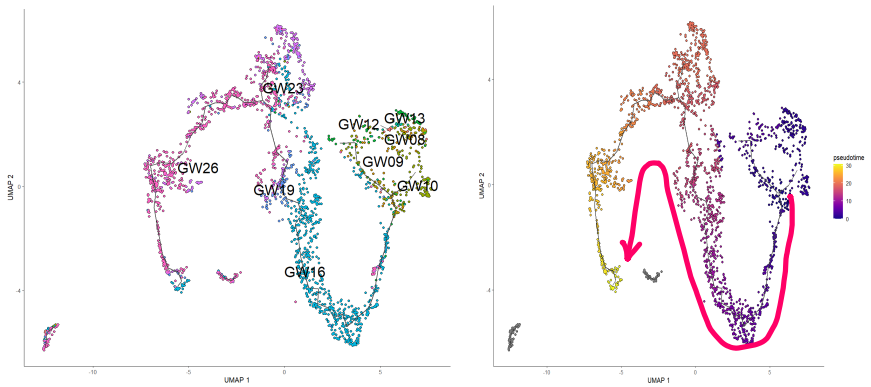


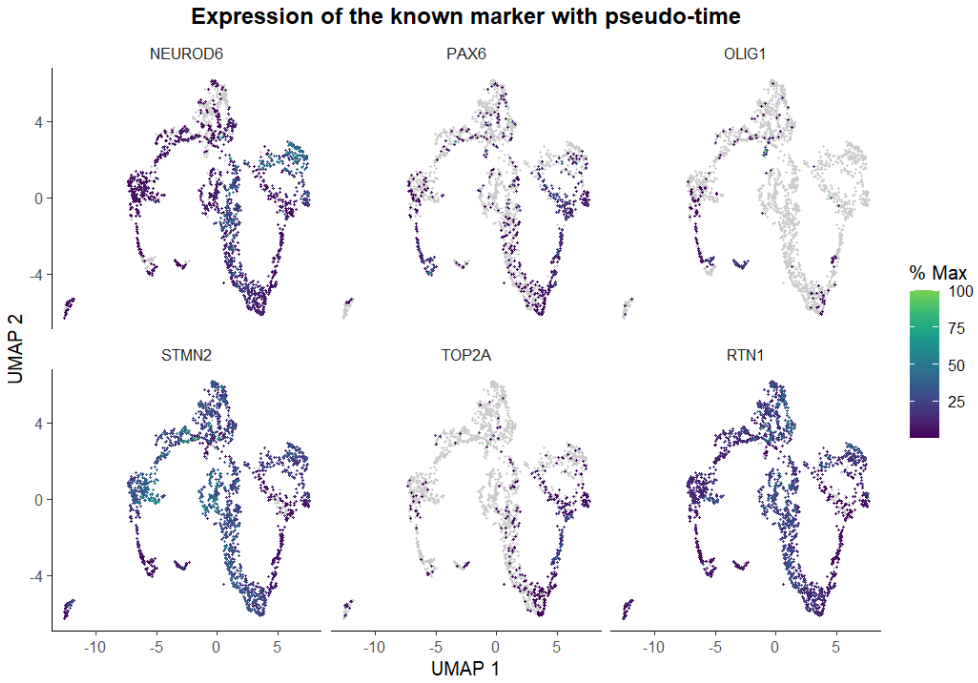


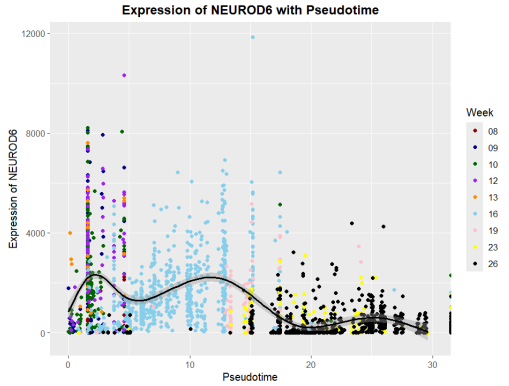


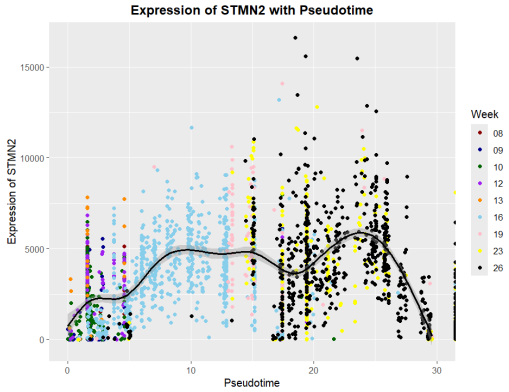
- The spearman rank-order correlation of pseudo-time and week number have been calculated.

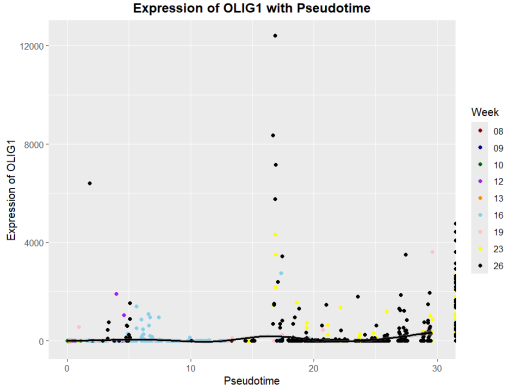
Identifying the differentially expressed genes

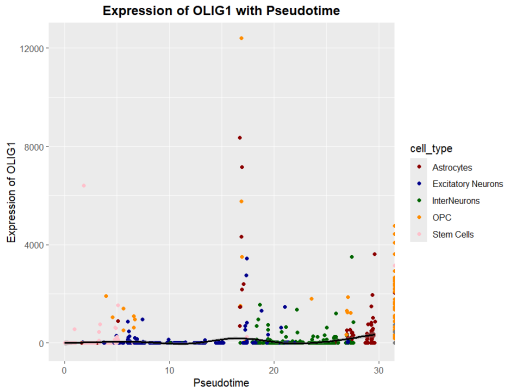












- We have identified the differentially expressed genes with respect to pseudo-time, including the temporally expressed genes whose expression is changing in a continuous manner over pseudo-time.
  - NEUROD6: neurodevelopment and neuronal differentiation.
  - STMN2: microtubule-destabilizing protein that plays a crucial role in regulating microtubules dynamics within neurons.
  - OLIG2:transcription factor that plays a role in the development and specification of oligodendrocytes.