Single-cell RNA Seq Trajectory Inference

DISC `omics study group workshop April 2023

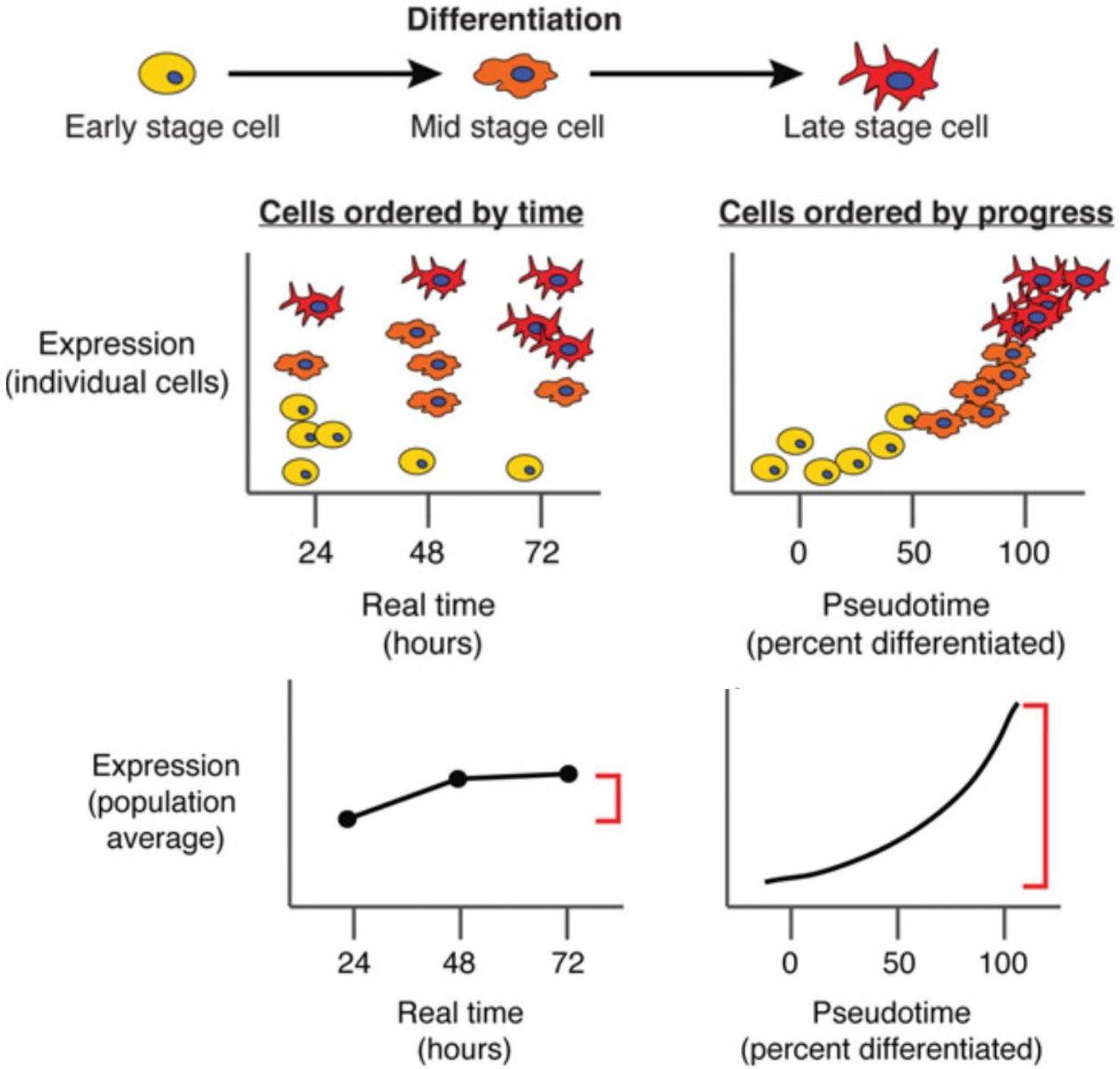
Rebecca Batorsky, Senior Data Scientist, Data Intensive Studies Center

Rebecca.Batorsky@tufts.edu

Jason Laird, Bioinformatics Scientist, TTS Research Technology

Albert Tai, Research Assistant Professor of Immunology, TUSM

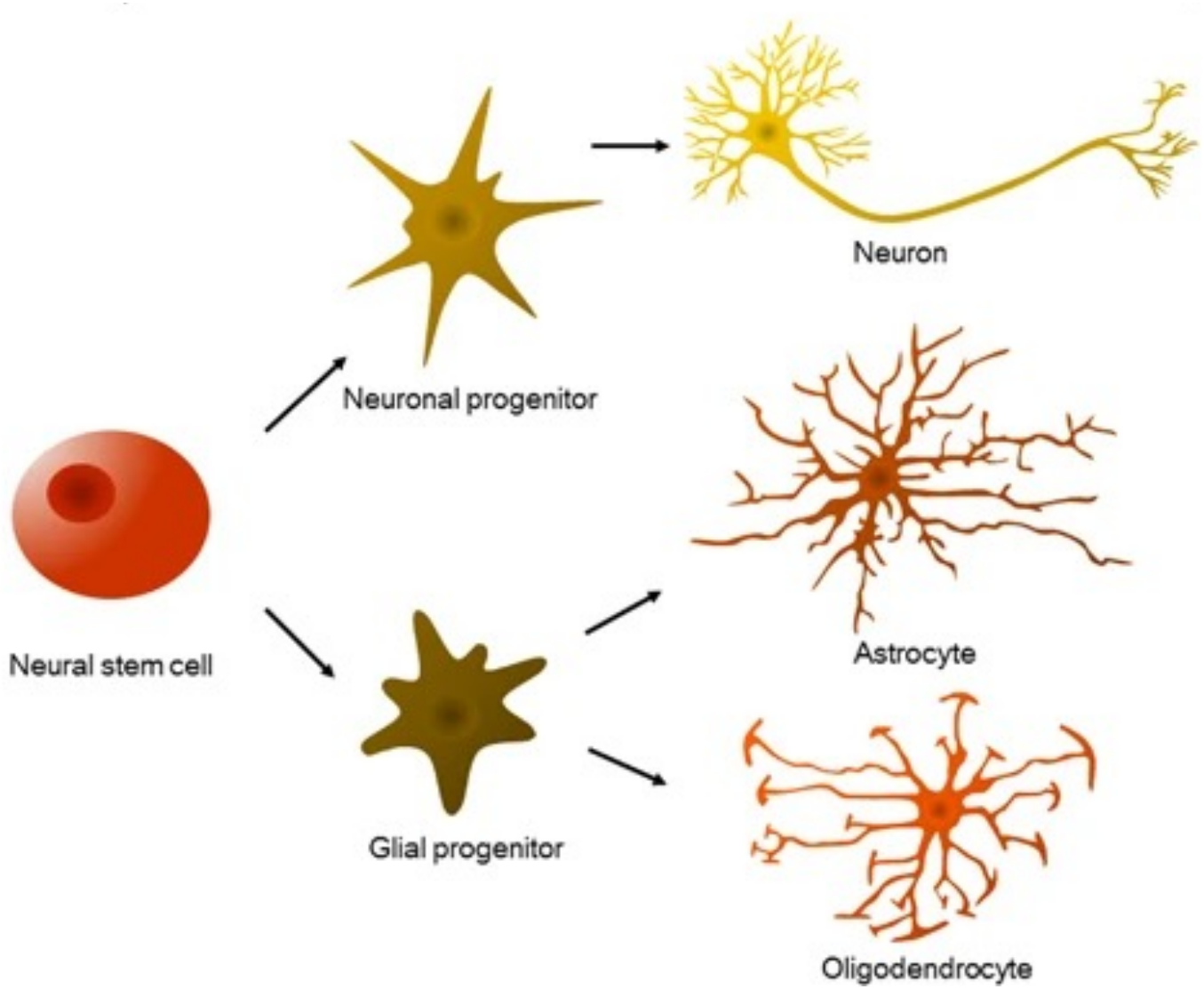
How to study gene expression dynamics during development?



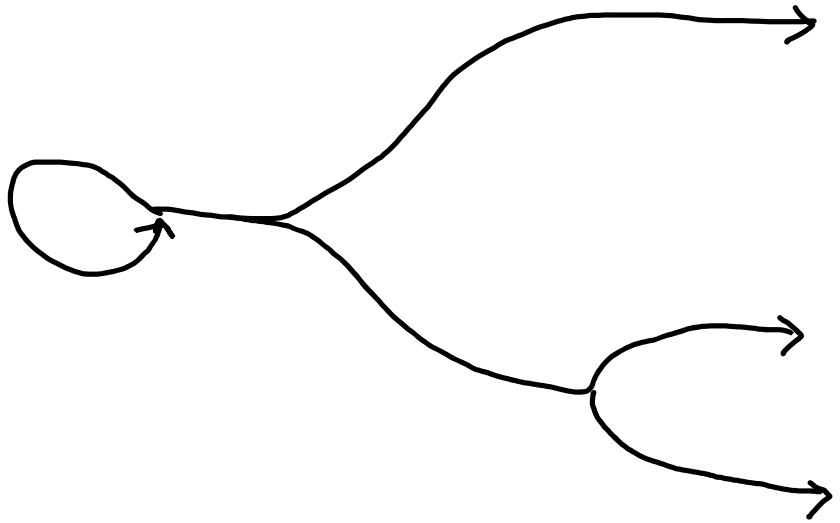
Trapnell **Defining cell types and states with single-cell genomics** 2015

Multiple cell fates in Neuron Differentiation

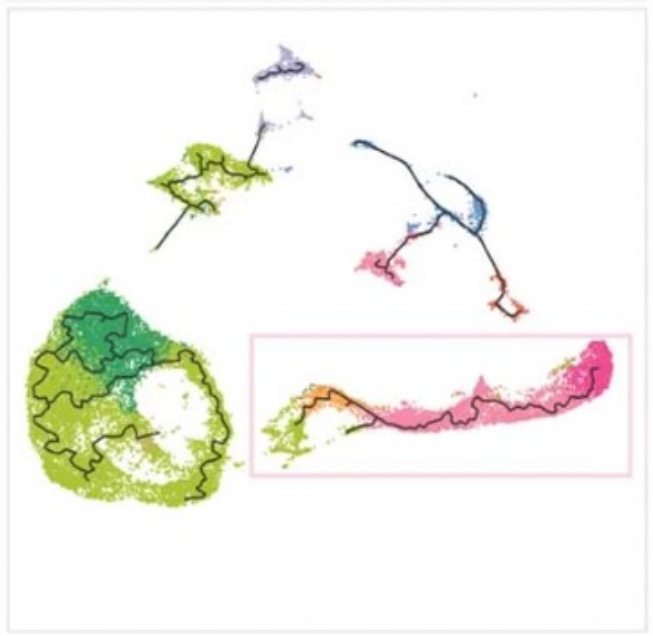
Early stage cell **Mid stage cell** **Late stage cell**



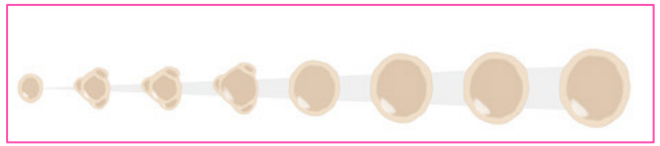
Trajectory



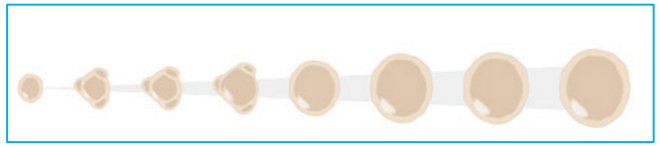
Brain organoid



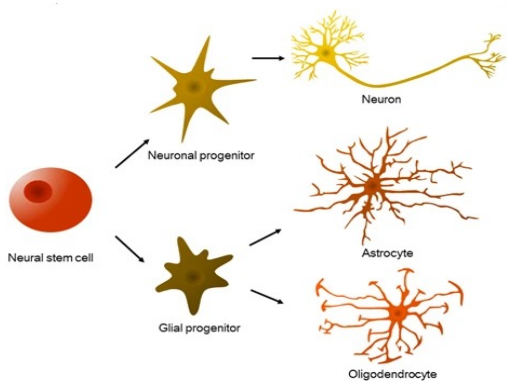
SUV420H1+/- ASD risk gene



Control

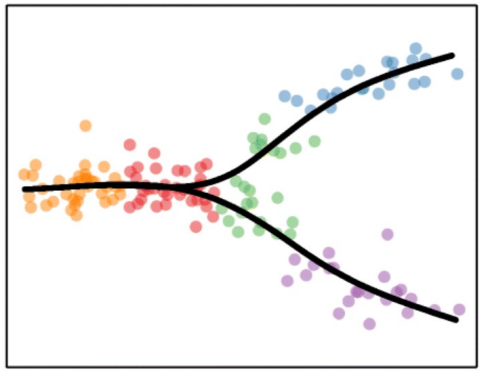


Workshop Goal



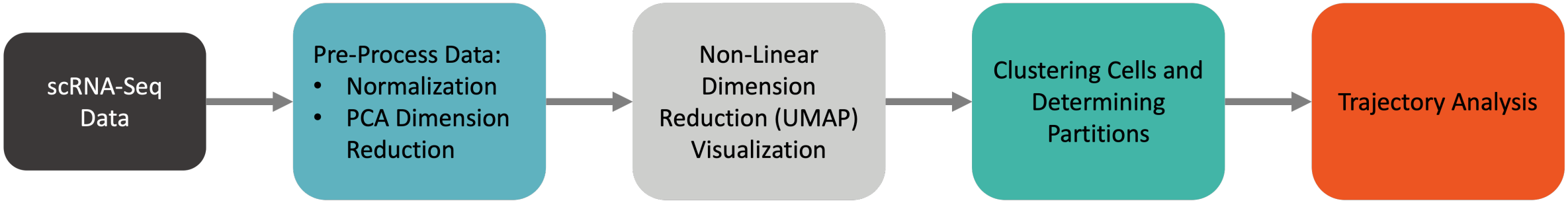
- Characterize progression of unsynchronized cells going through differentiation to potentially multiple cell fates
- Also applicable to other dynamic processes such as response to treatment, disease progression

Method

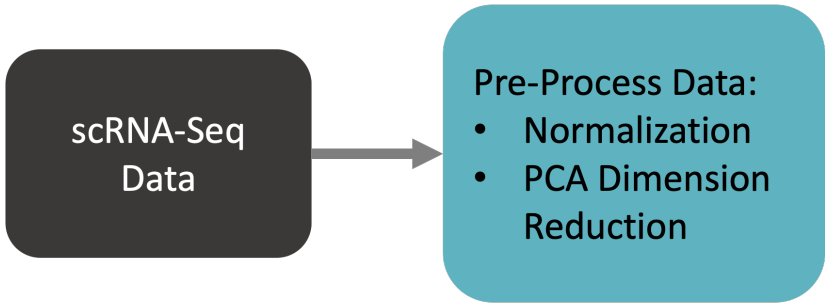


- Unsupervised ordering of cells using scRNAseq transcriptional profiles
- Open Source tools (Monocle 3 and Tidyverse R libraries)
- Develop intuition to understand algorithms

Workflow Outline



Data Pre-Processing



Normalization

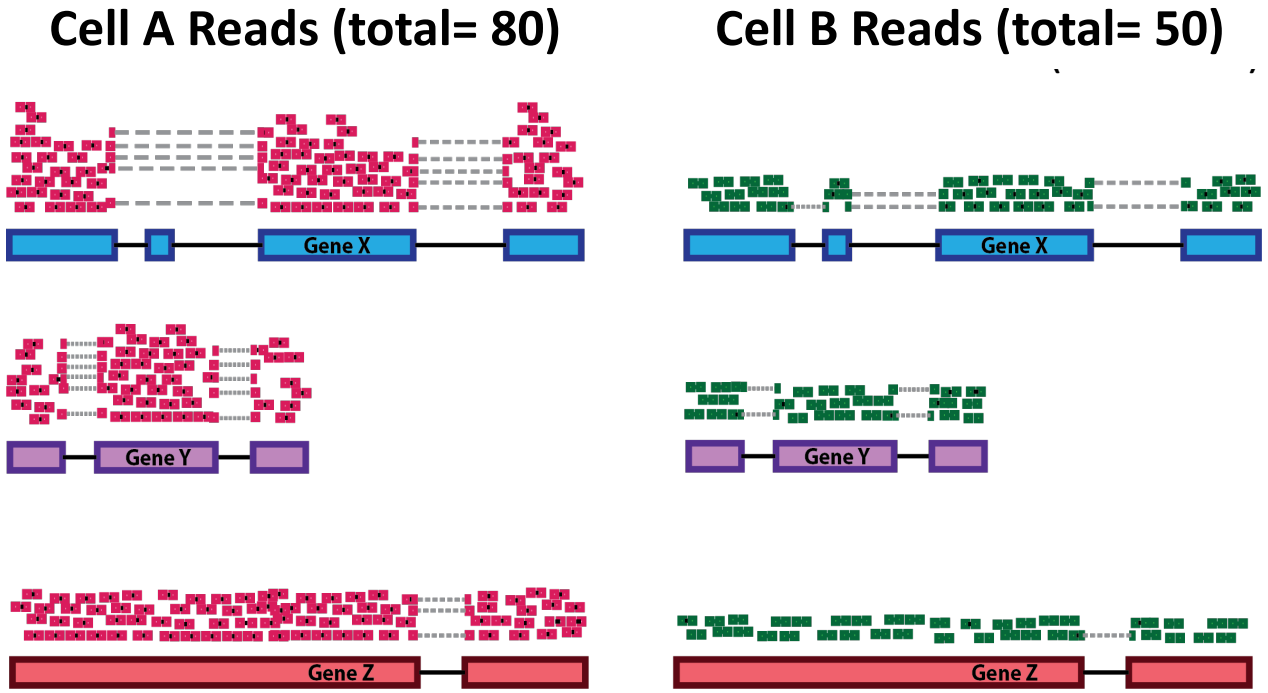
- Raw Count != Expression strength

- Simple Norm

$$\frac{K_i}{Total\ Counts\ Per\ Sample} * Scale\ Factor$$

If $Scale\ Factor$ is 10^6 then
Simple Norm is CPM

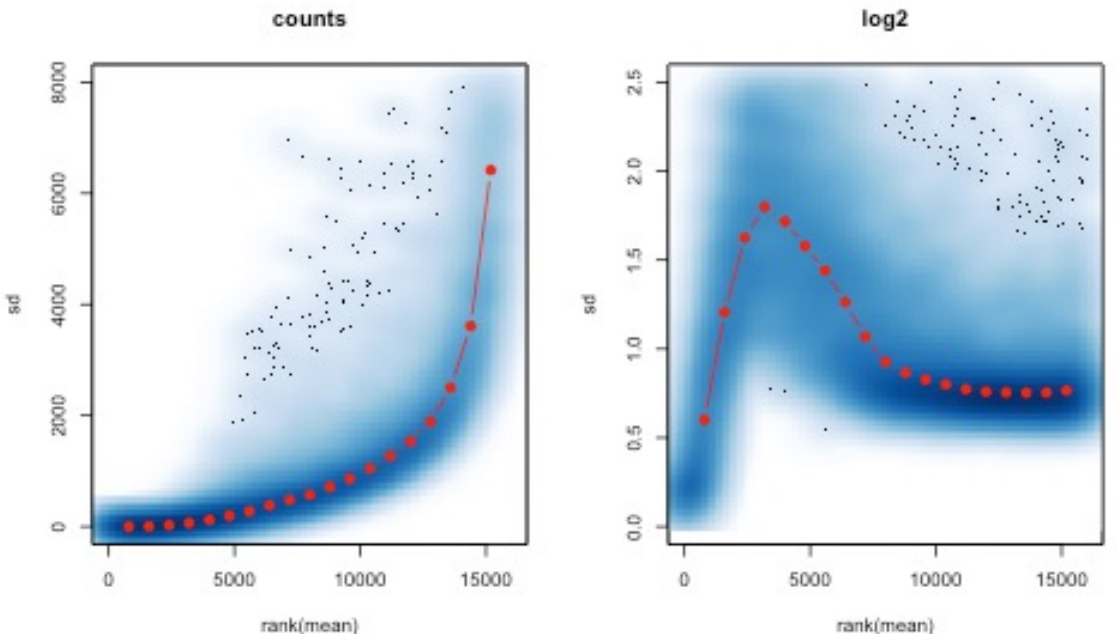
- Monocle 3 uses a cell-specific size factor



Adapted from https://hbctraining.github.io/DGE_workshop

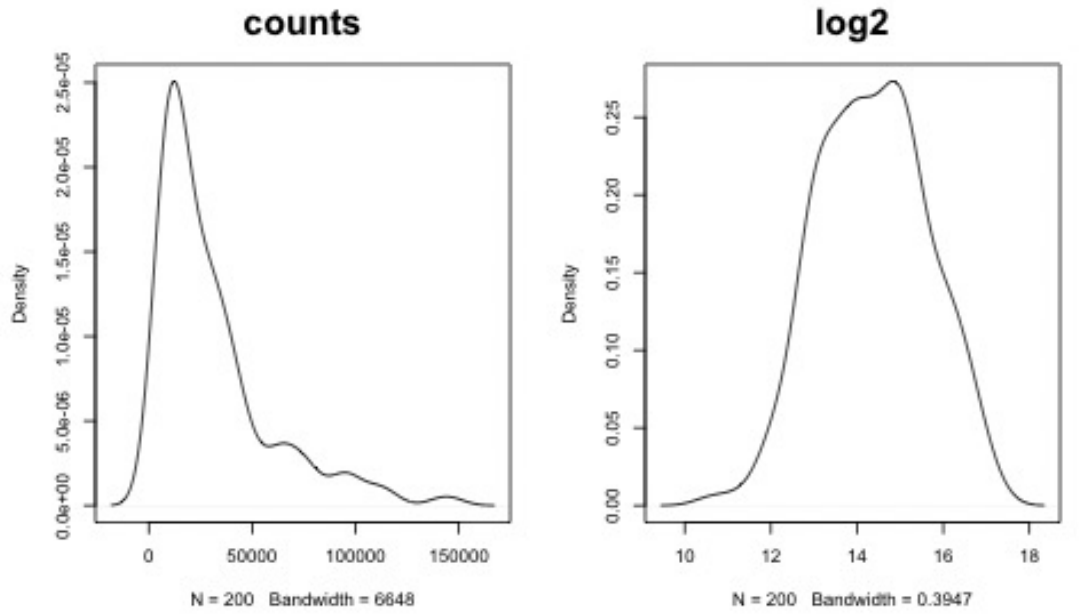
Transformation and Scaling

- Log2-transform (norm_counts + 1)
- Scale counts to unit variance and zero mean



Rational:

- **Fold-Changes** rather than additive differences
- Reduce the mean-variance dependency
- Better approximation of Normal Distribution



Principal Component Analysis (PCA)

Gene Expression Matrix

G x N

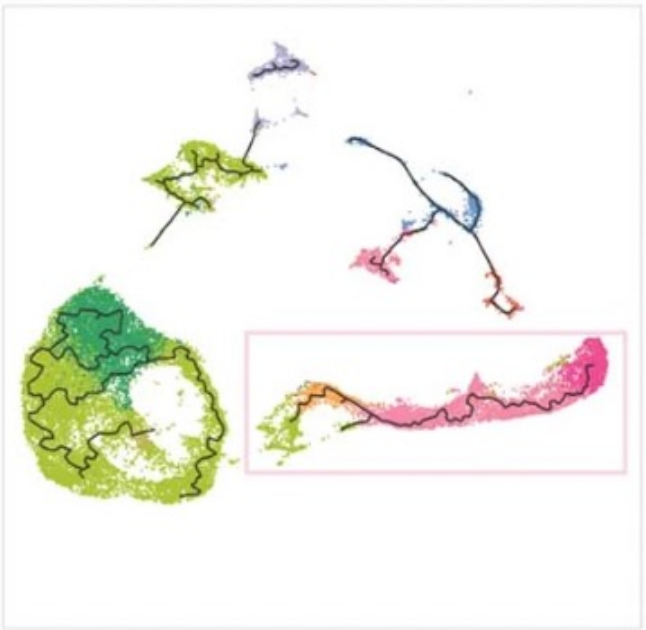
Feature selection

Reduce number of dimensions while preserving important characteristics of the data

PCA Matrix

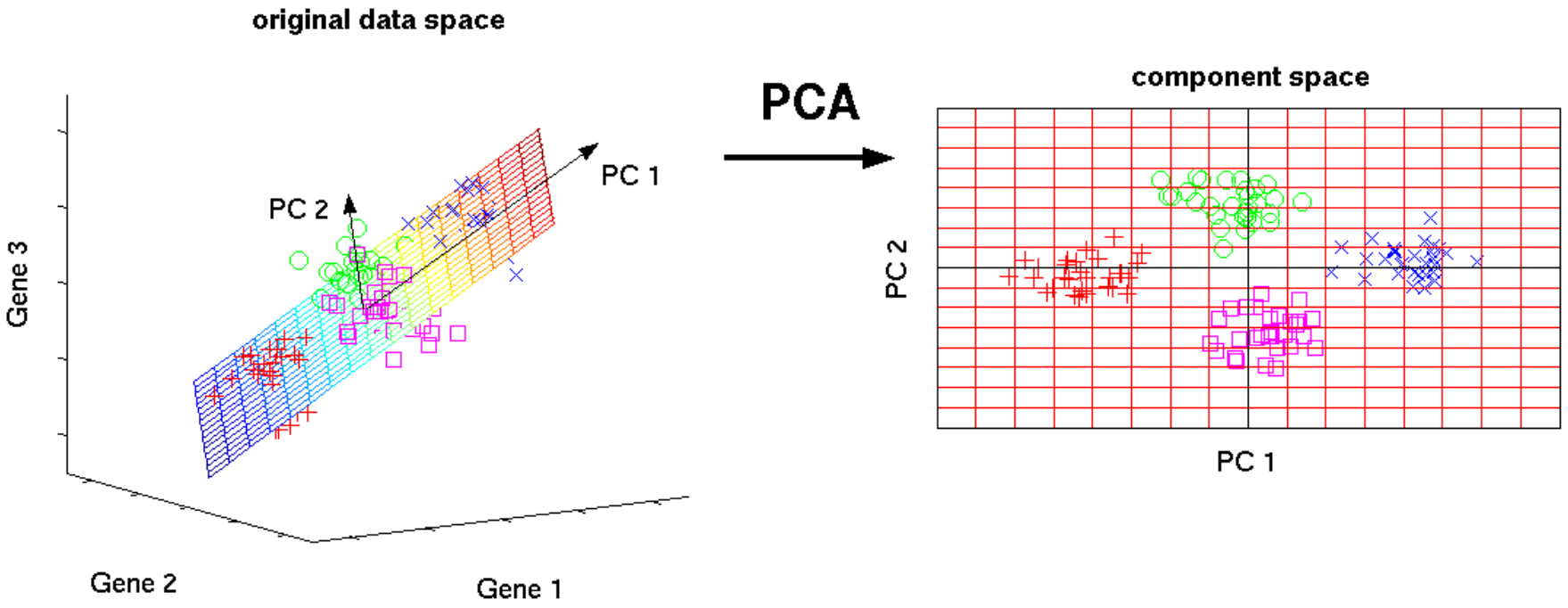
P x N

Gene expression between cell types is highly correlated, so we don't need all ~30,000 genes

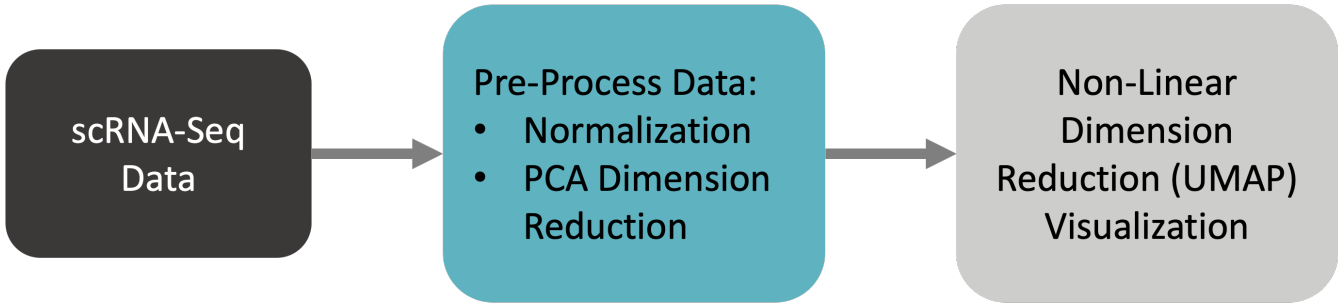


Principal Component Analysis (PCA)

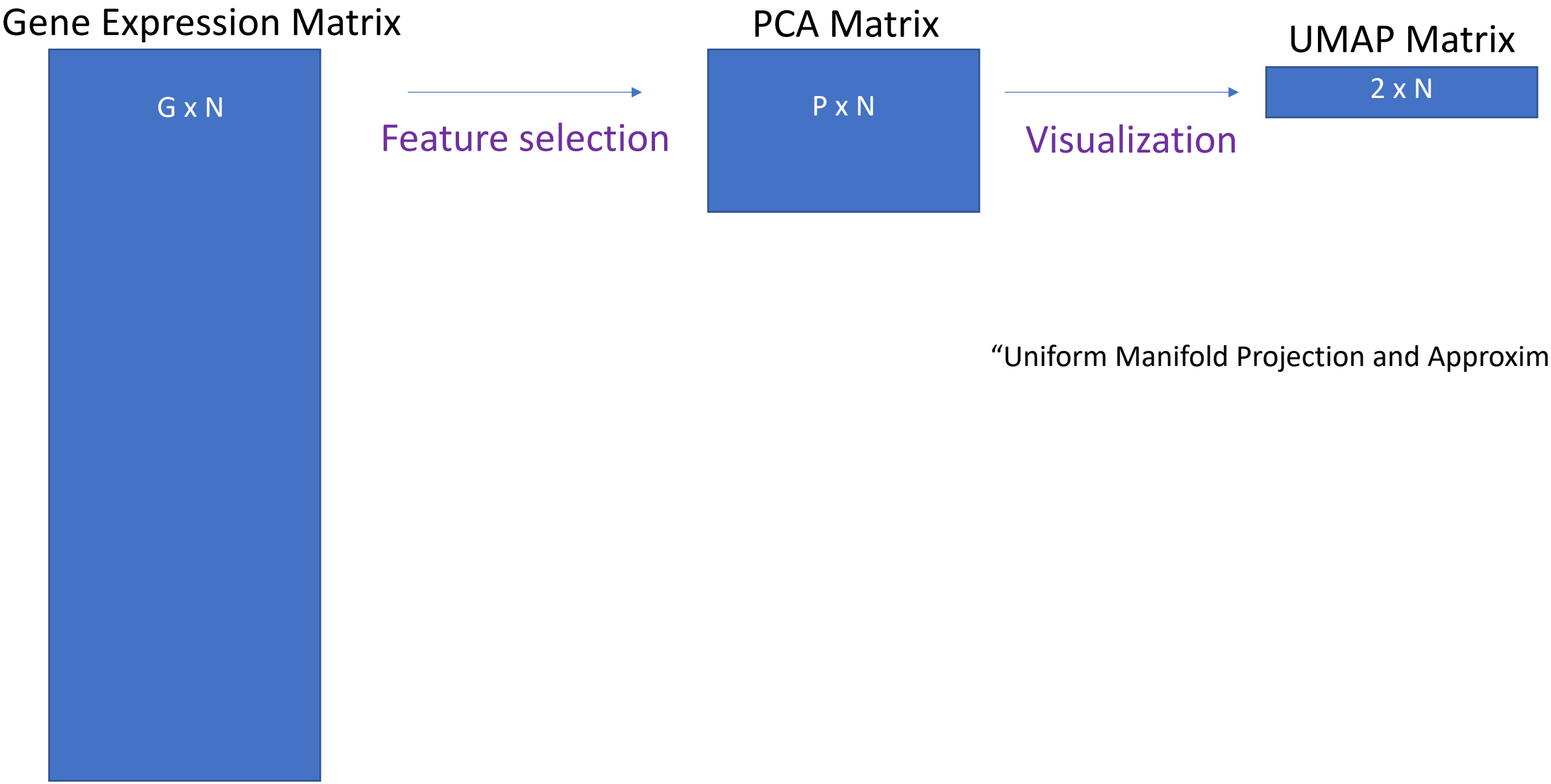
Gene	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8
Gene 1	8.9	8.9	8.9	9.0	8.9	8.9	9.0	6.8
Gene 2	0.6	-1.0	0.6	-1.0	0.6	-1.0	0.6	3.8
Gene 3	4.1	11.9	4.1	-0.5	4.1	8.7	4.0	4.4



Nonlinear Dimension Reduction



Why do we need 2-step Dimension Reduction?



"Uniform Manifold Projection and Approximation"

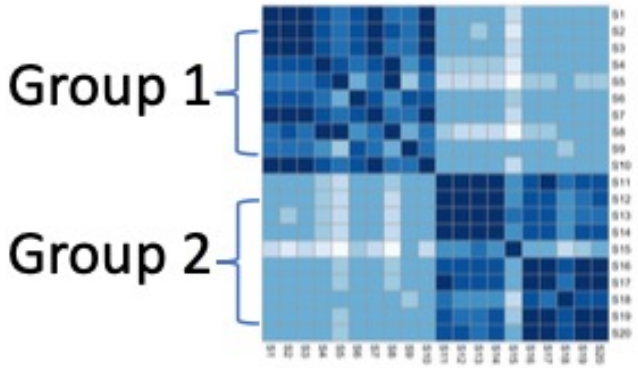
UMAP

PCA Matrix

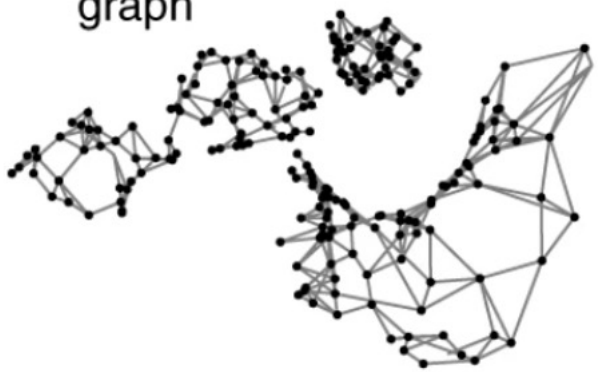
P x N

Calculate the cell-cell pairwise distance matrix

N x N



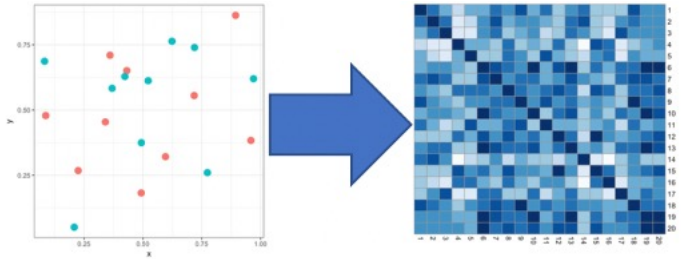
k-nearest-neighbors graph



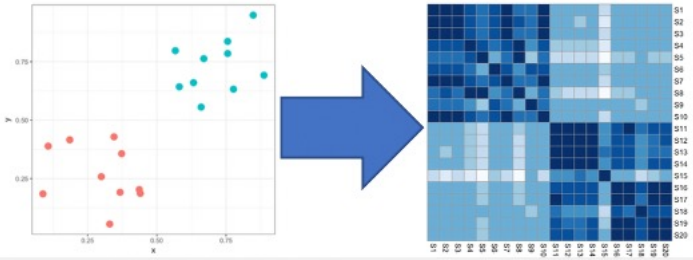
$$\mathcal{G} = (V, \varepsilon)$$

Use optimization to find a 2D graph with similar cell-cell pairwise distance matrix

Initial



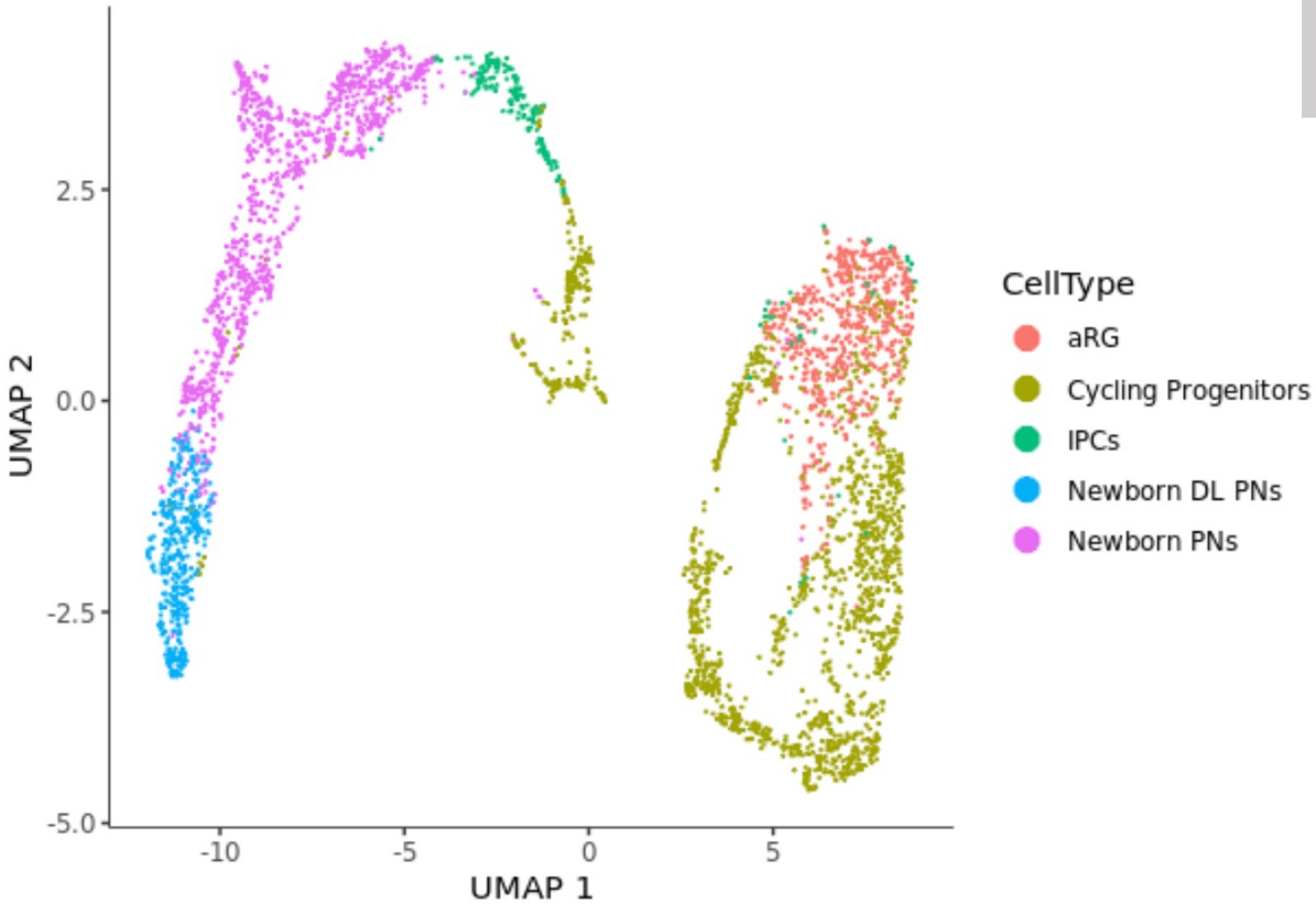
Final



UMAP Matrix

2 x N

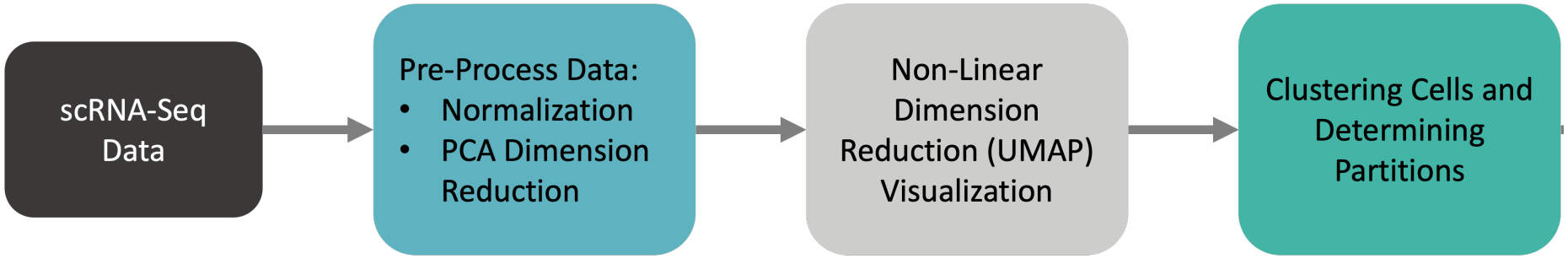
UMAP of brain organoids




```
plot_cells(cds,
           color_cells_by = "CellType")
```

Today we are not covering cell-type identification ☺

Clustering Cells and Determining Partitions



Clustering

- **Goal**: Find Highly related groups of cells (communities)
- **Input**: k-nearest neighbor graph

- **Method:** Optimize **Modularity** by merging cells into communities iteratively to maximize the within cluster connections and minimize the between cluster connections

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

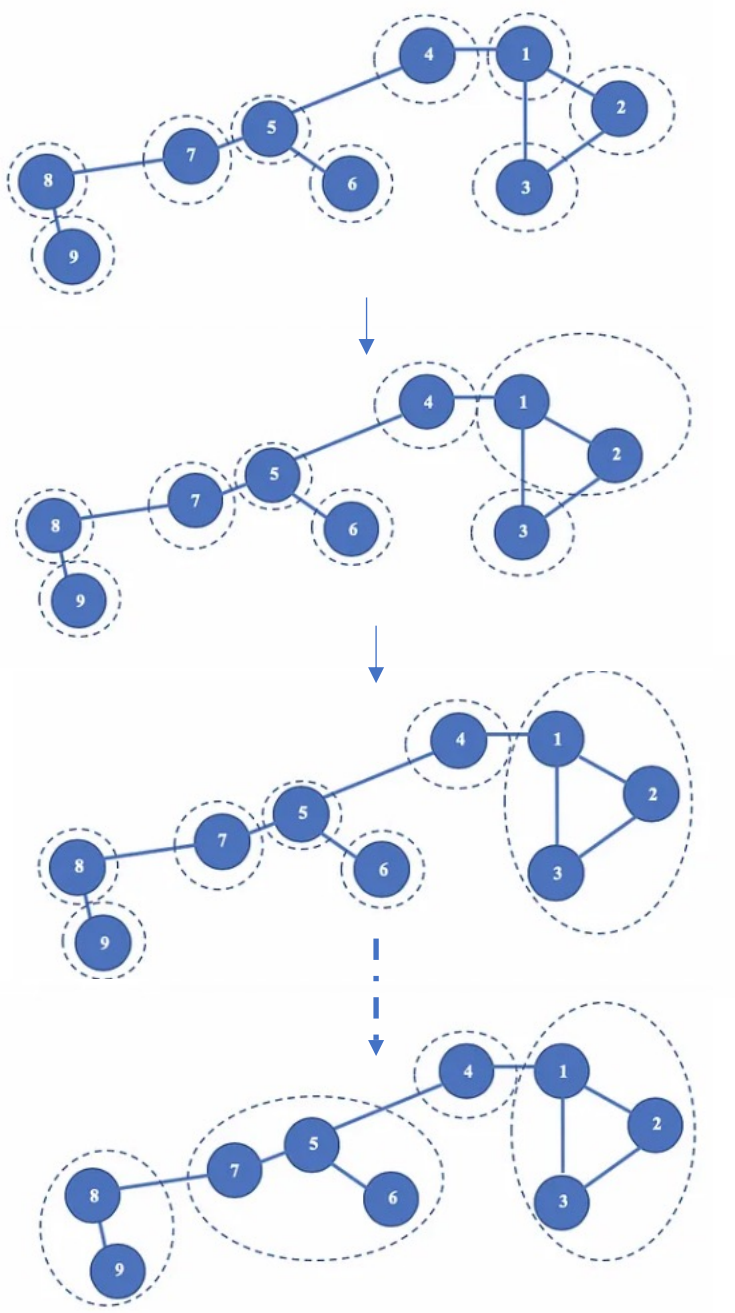
A_{ij} = Weight of edge between cells *i* and *j*

k_i, k_j = *Degree of cells i and j*

$\delta(c_i, c_j)$ =

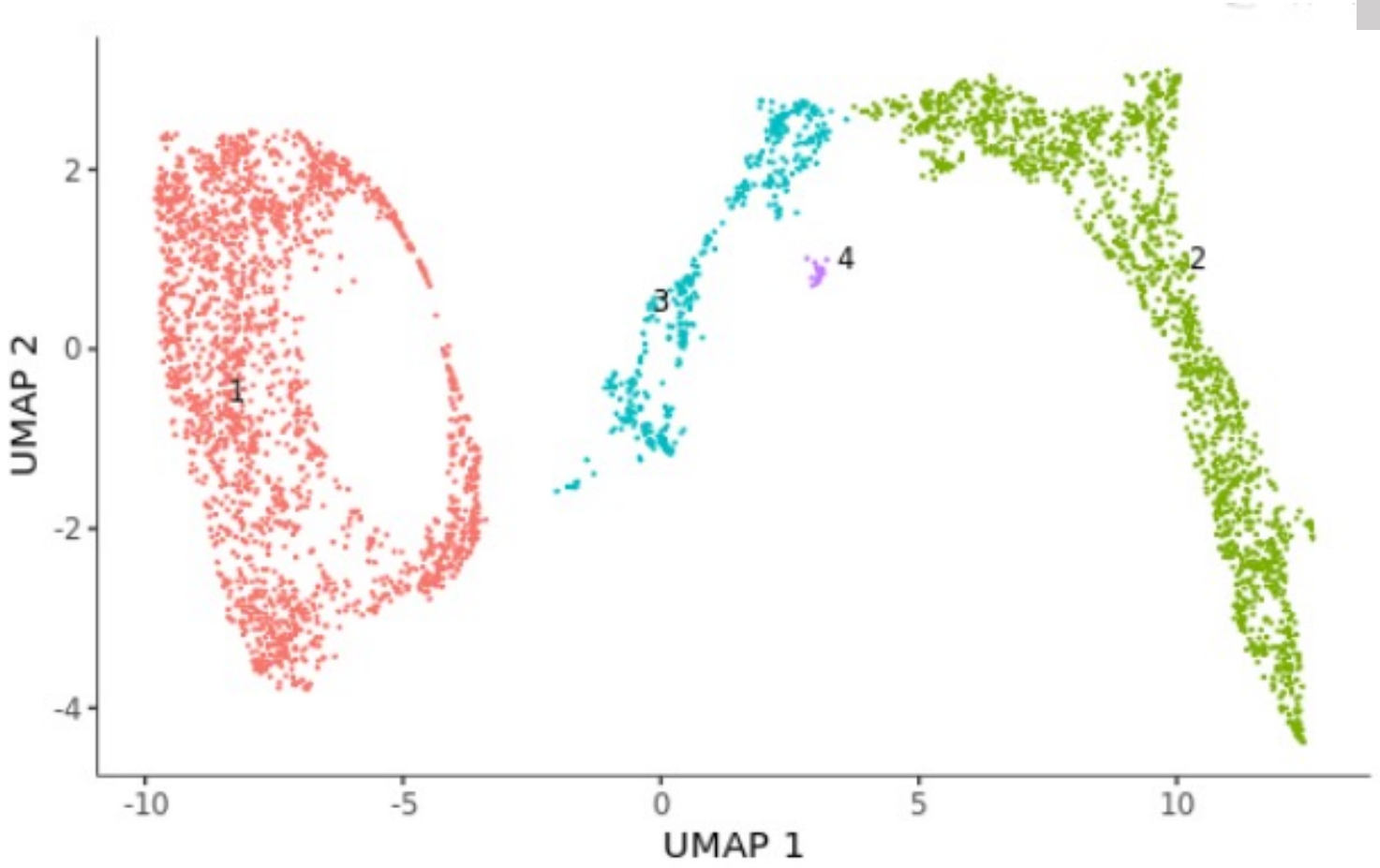
- 1 if *cells i and j are in the same community*
- *0 otherwise*

- m total number of links



Cell Clusters

```
plot_cells(cds,
           color_cells_by = "cluster")
```

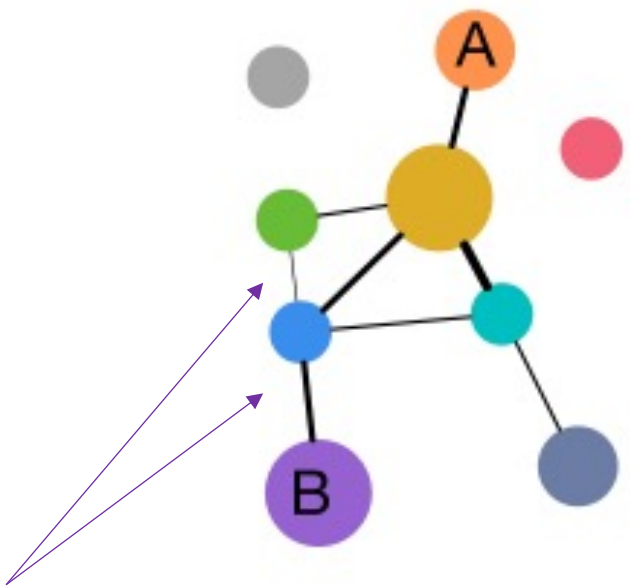


Partition

- **Goal**: Find highly connected clusters, likely to be developmentally related
- **Input**: Clustered, k-nearest neighbor graph
- **Method:** Compare the number of connections **between two cluster**s to the number of connections expected by chance.

single-cell graph

Abstracted graph with partitions



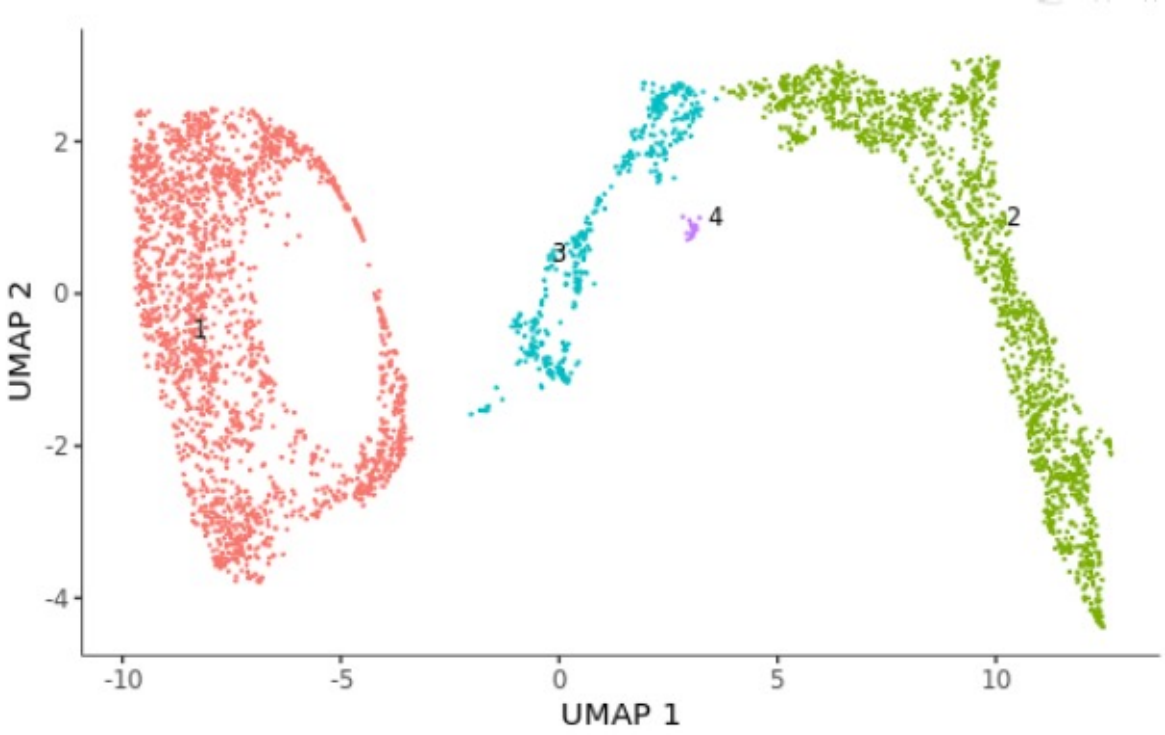
Partitions are not connected in the abstracted graph

Quantify the strength of connections between clusters

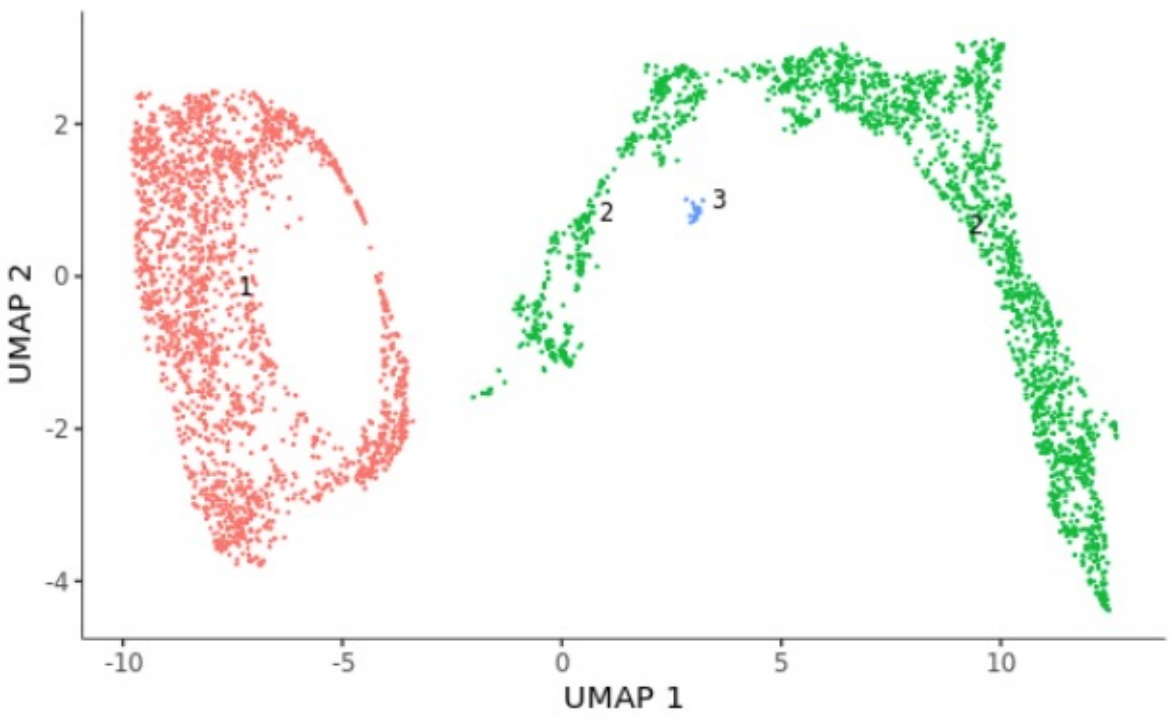
Cluster vs. Partition

Clusters are often used to identify cell types. Partitions are larger, more well separated groups of cells than clusters and will be used to create trajectories

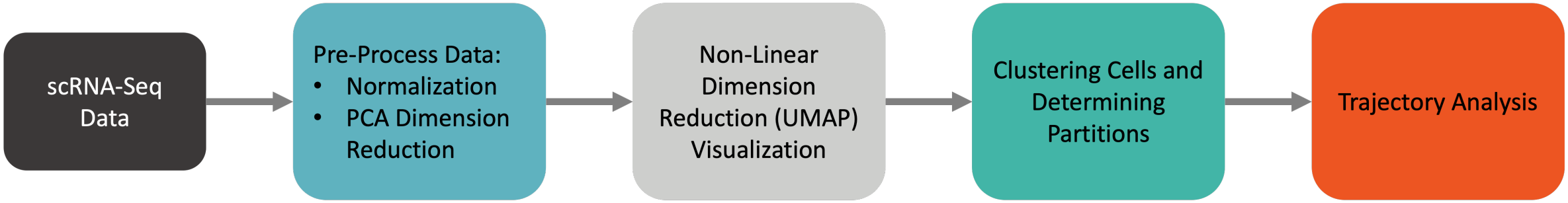
```
plot_cells(cds,
           color_cells_by = "cluster")
```




```
plot_cells(cds,
           color_cells_by = ‘partition’)
```



Trajectory Analysis

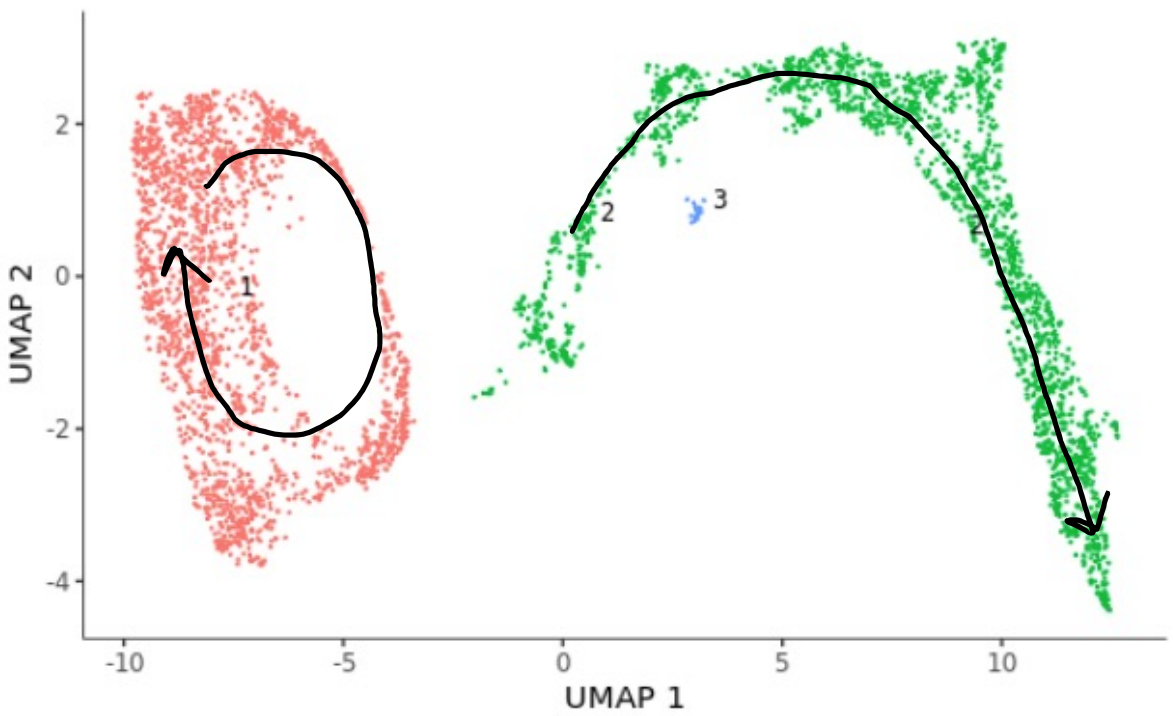


Trajectory Analysis

Goal: Uncover underlying structure in our low-dimensional representation of the data. It may be curved and have branches and/or loops.

Input: PCA Matrix and UMAP Matrix

Method: Reversed graph embedding (SimplePPT)

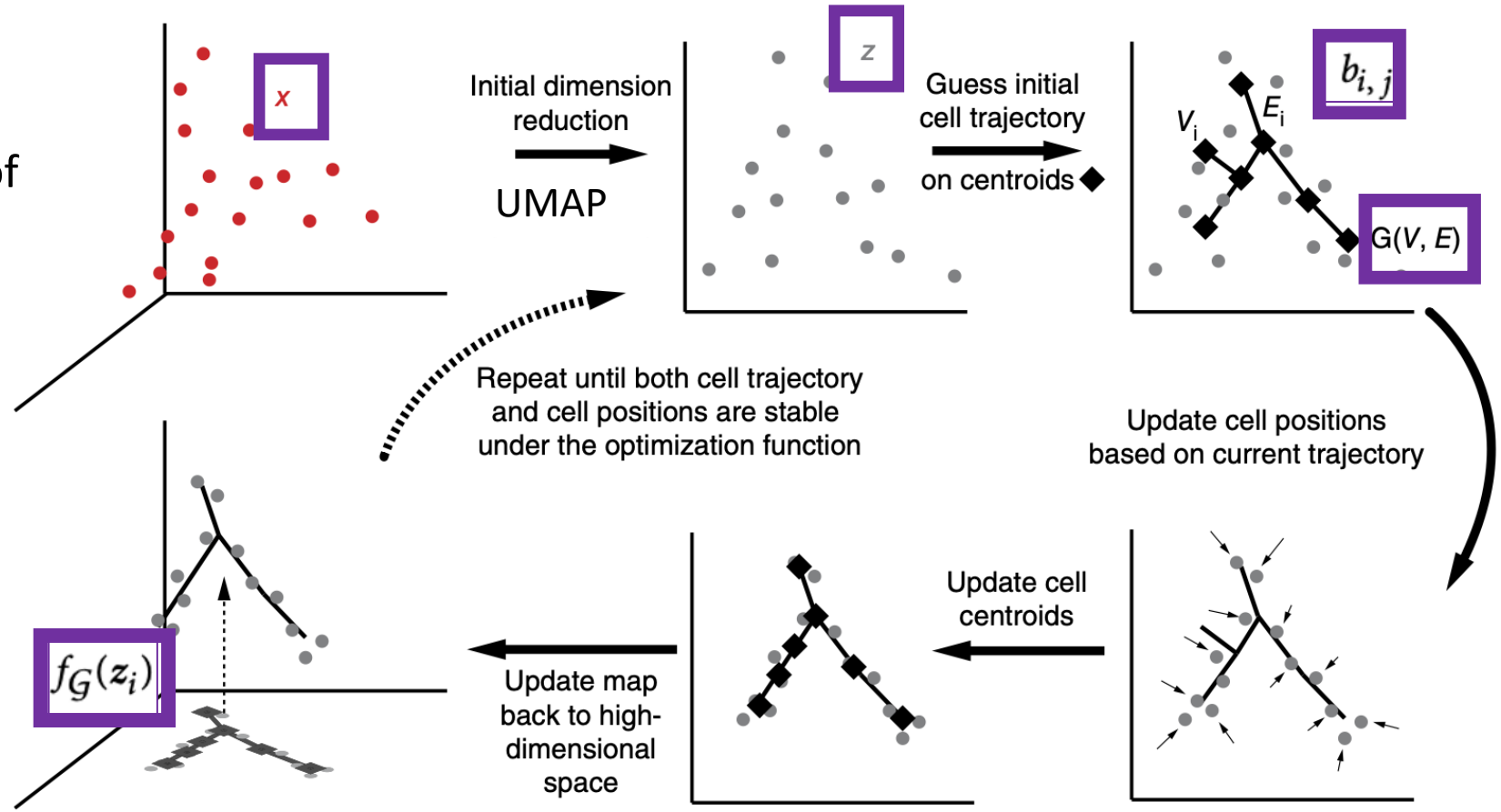


Trajectory Analysis

Goal: Uncover underlying structure in our low-dimensional representation of the data. It may be curved and have branches and/or loops.

Input: PCA Matrix and UMAP Matrix

Method: Reversed graph embedding (SimplePPT)



Optimization function:

$$\min_{G \in G_b} \min_{f_G \in \mathcal{F}} \min_{Z} \sum_{i=1}^{N} \| x_i - f_G(z_i) \|^2$$

$$+ \frac{\lambda}{2} \sum_{(V_i, V_j) \in \varepsilon} b_{i,j} \| f_G(z_i) - f_G(z_j) \|^2$$

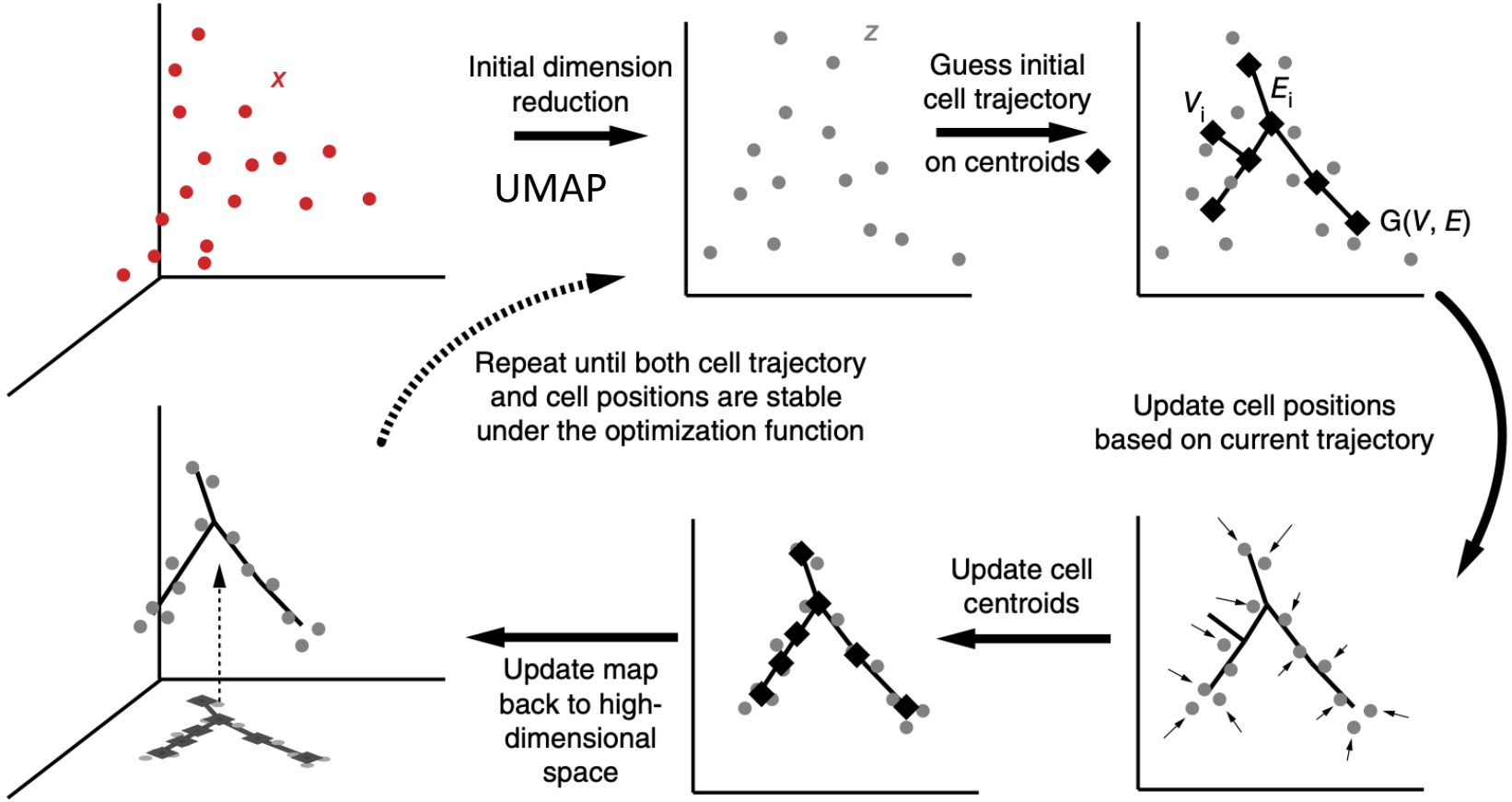
"reverse embedding"

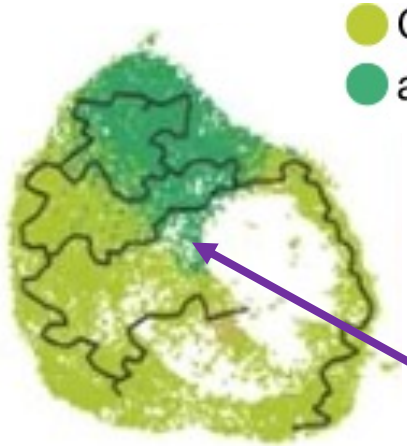
Trajectory Analysis

Goal: Find the **tree structure** that describes the data, allowing for loops

Input: Gene Expression Matrix + UMAP Matrix

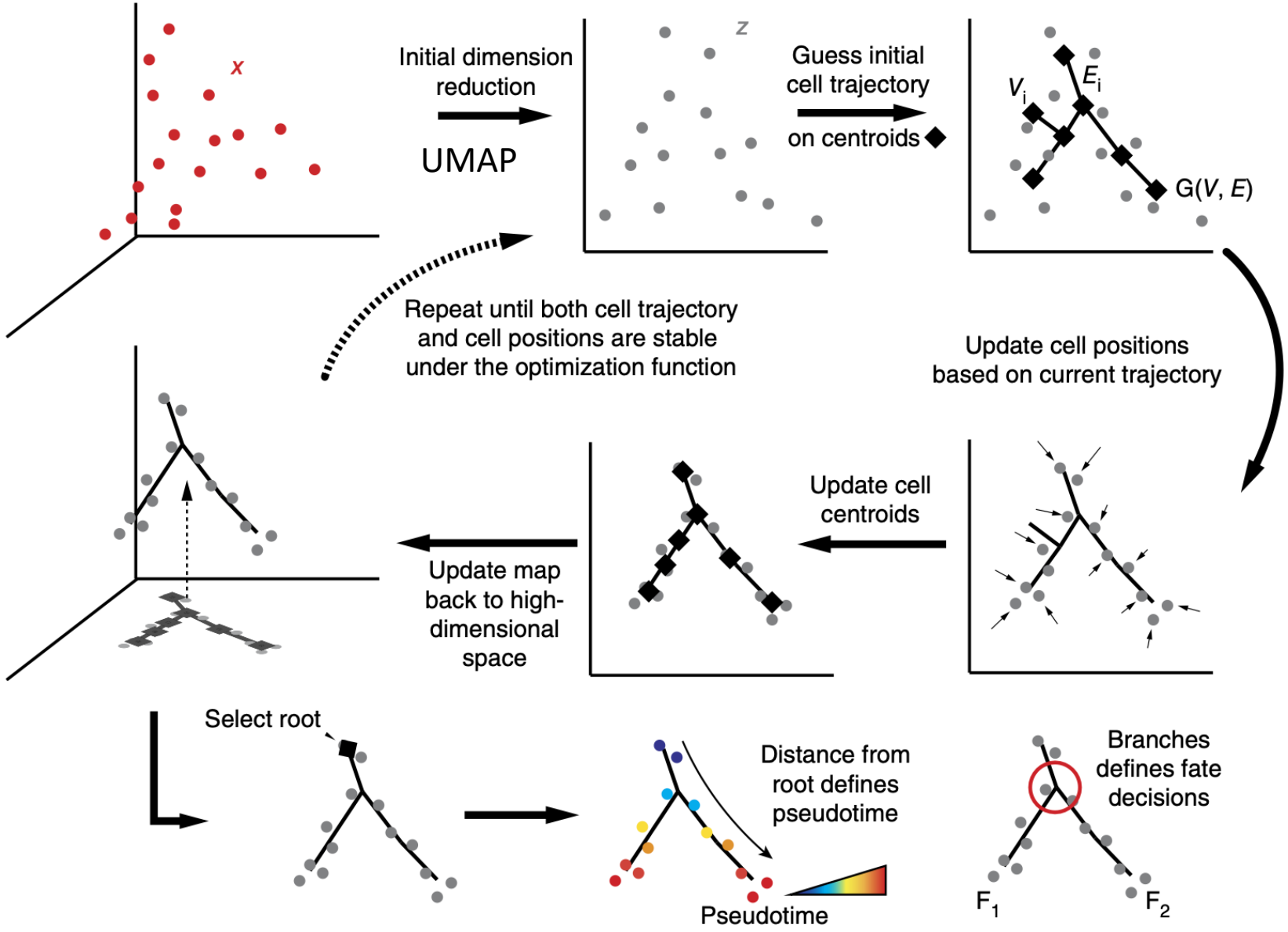
Method: Reversed graph embedding (SimplePPT)





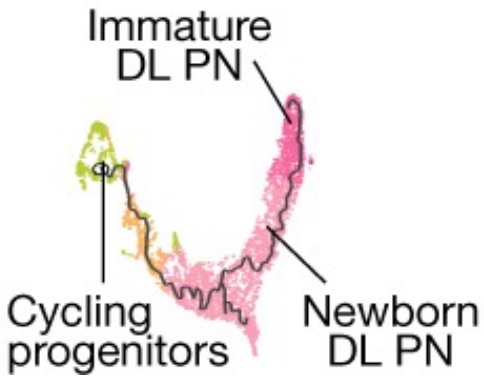
Note: loops are allowed!

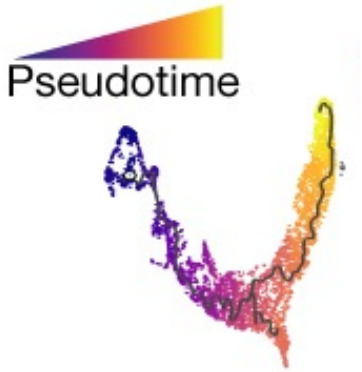
Pseudotime Assignment



Pseudotime is defined for each cell as the distance along the trajectory to user-selected **Root Node.**

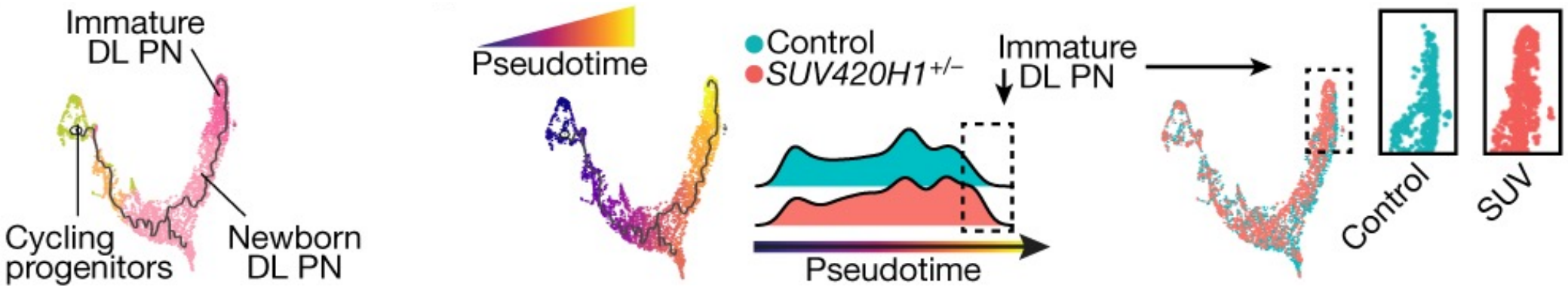
Pseudotime Assignment





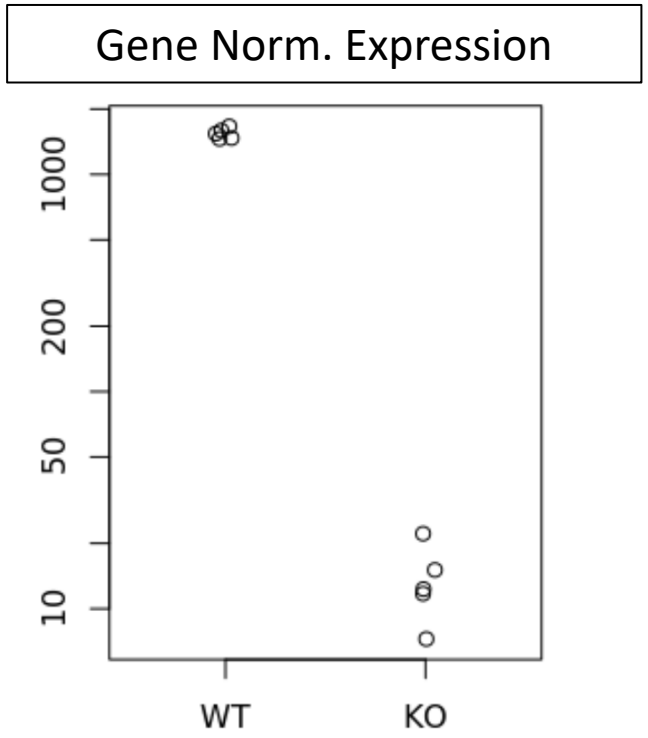
Uzquiano et al, **Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex,** Cell 2022

Identifying Genes that vary over time and with genotype



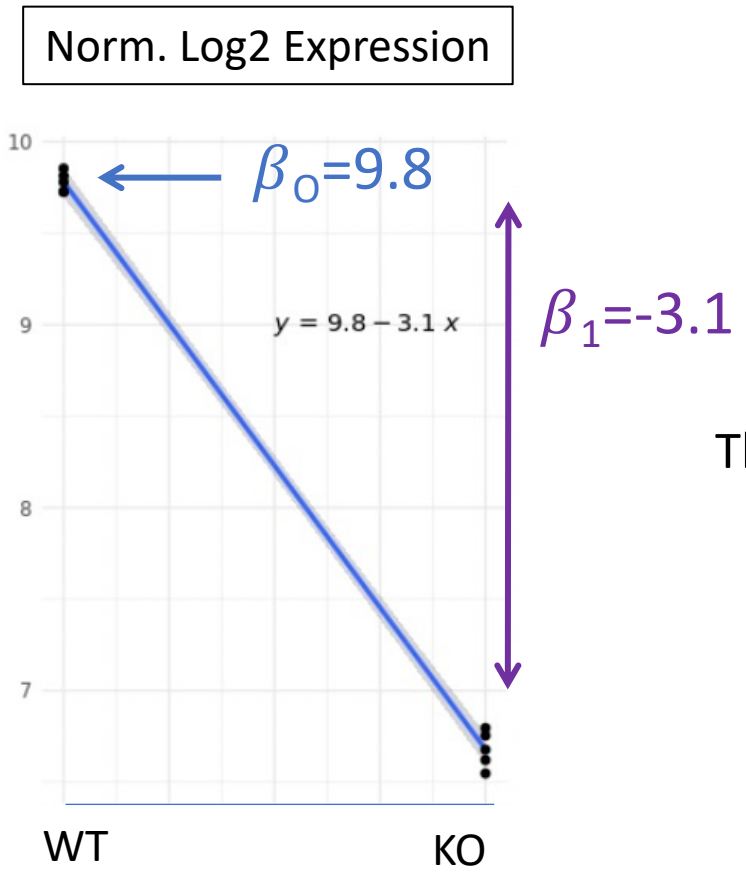
Uzquiano et al, **Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex,** Cell 2022

Modeling Gene Expression Values



Modeling Gene Expression Values

All leading DE tools use **regression models** to estimate the fold change between genotypes for **each gene**
Example, simple linear regression:



Intercept — Condition (0-WT, 1-KO)

$$Y = \beta_o + \beta_1 X + e$$

Log2 Expression Values

Slope: difference between WT/KO

Error

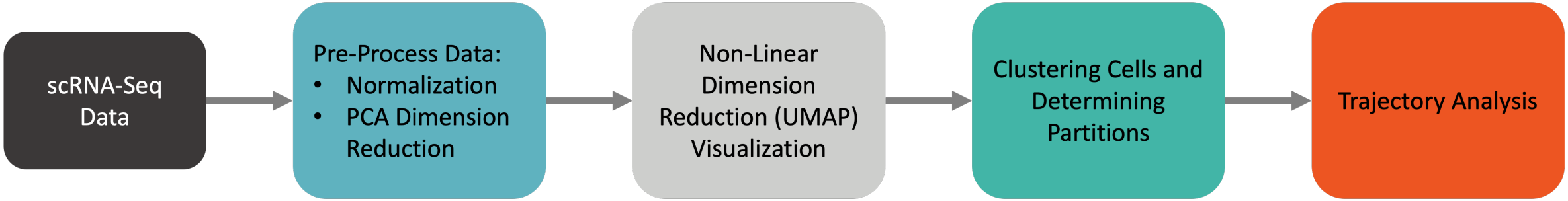
This method can be extended to multiple factors as well as interactions:

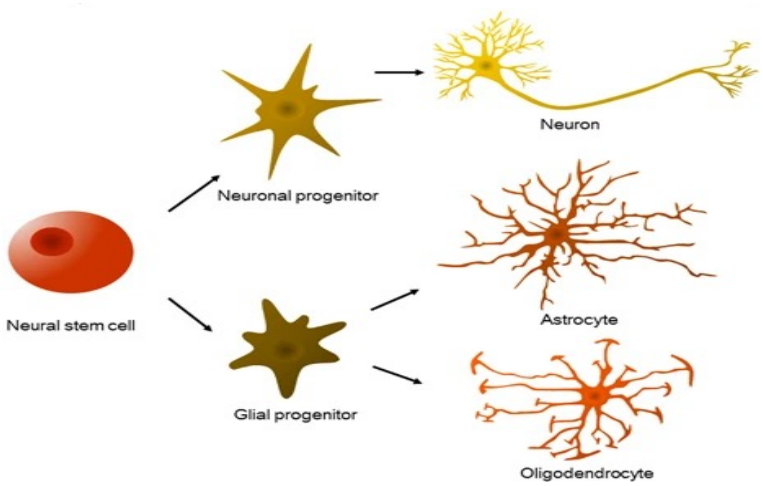
E.g. How does the effect of **Genotype** change over **Pseudotime**

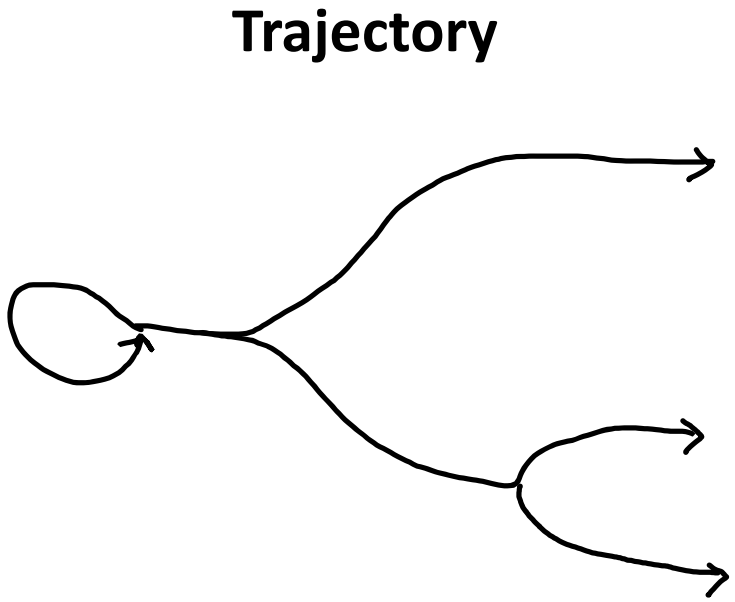
$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$$

Genotype — Time — Interaction Genotype <-> Pseudotime

Conclusion







[Getting Started with Monocle 3]
[COMP 150: Machine Learning for Graph Data Analytics]
https://disc.tufts.edu/
https://it.tufts.edu/bioinformatics

Start Setting up for the Workshop

Workshop webpage: https://go.tufts.edu/trajectory_analysis

Or https://tuftsdatalab.github.io/tuftsWorkshops/ ->2023 Workshops-> Trajectory Analysis

Log on with Tufts Credentials to On Demand on Tufts Cluster https://ondemand.pax.tufts.edu/

Click on `Interactive Apps > RStudio Pax` and you will see a form to fill out to request compute resources to use RStudio on the Tufts HPC cluster. We will fill out the form with the following entries:

- `Number of hours` : `5`
- `Number of cores` : `1`
- `Amount of memory` : `16GB`
- `R version` : `4.0.0`
- `Reservation for class, training, workshop` : `Bioinformatics Workshops`