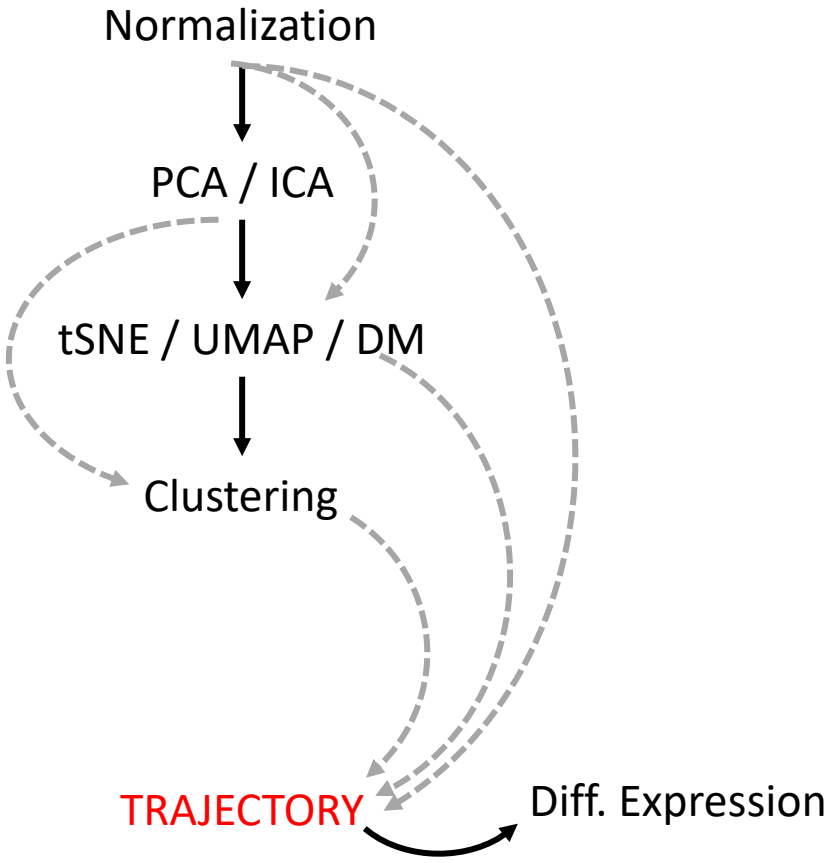
Trajectory inference analysis

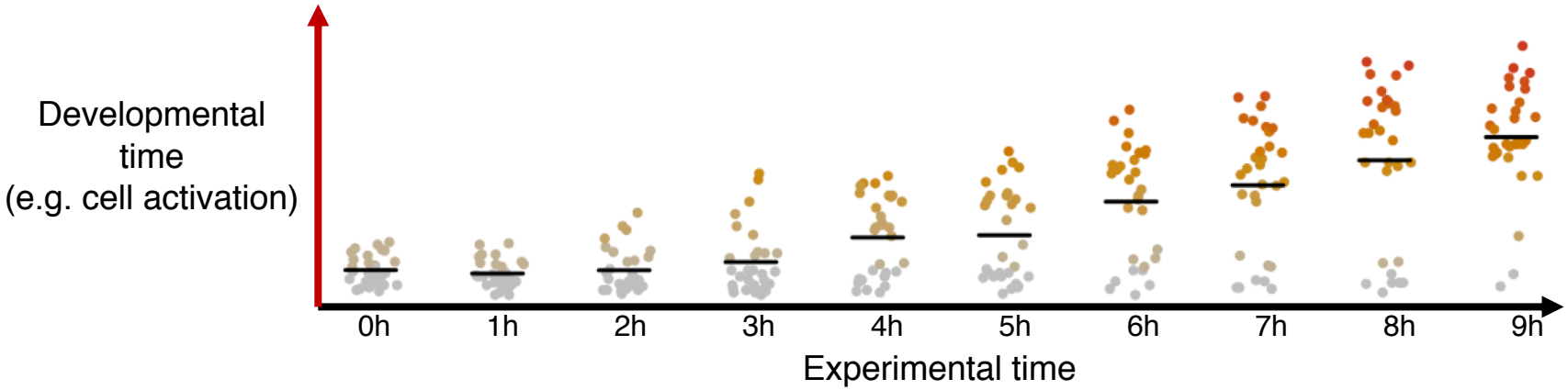
Paulo Czarnewski, **ELIXIR-Sweden (NBIS)**
Åsa Björklund, **ELIXIR-Sweden (NBIS)**
Gilet Jules, **ELIXIR-France**

European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org

Why trajectory inference?



What is trajectory inference / pseudotime?



- Cells that differentiate display a continuous spectrum of states

  Transcriptional program for activation and differentiation

- Individual cells will differentiate in an unsynchronized manner

  Each cell is a snapshot of differentiation time

- Pseudotime – abstract unit of progress

  Distance between a cell and the start of the trajectory

Should you run Trajectory Inference?

Are you sure that you have a developmental trajectory?

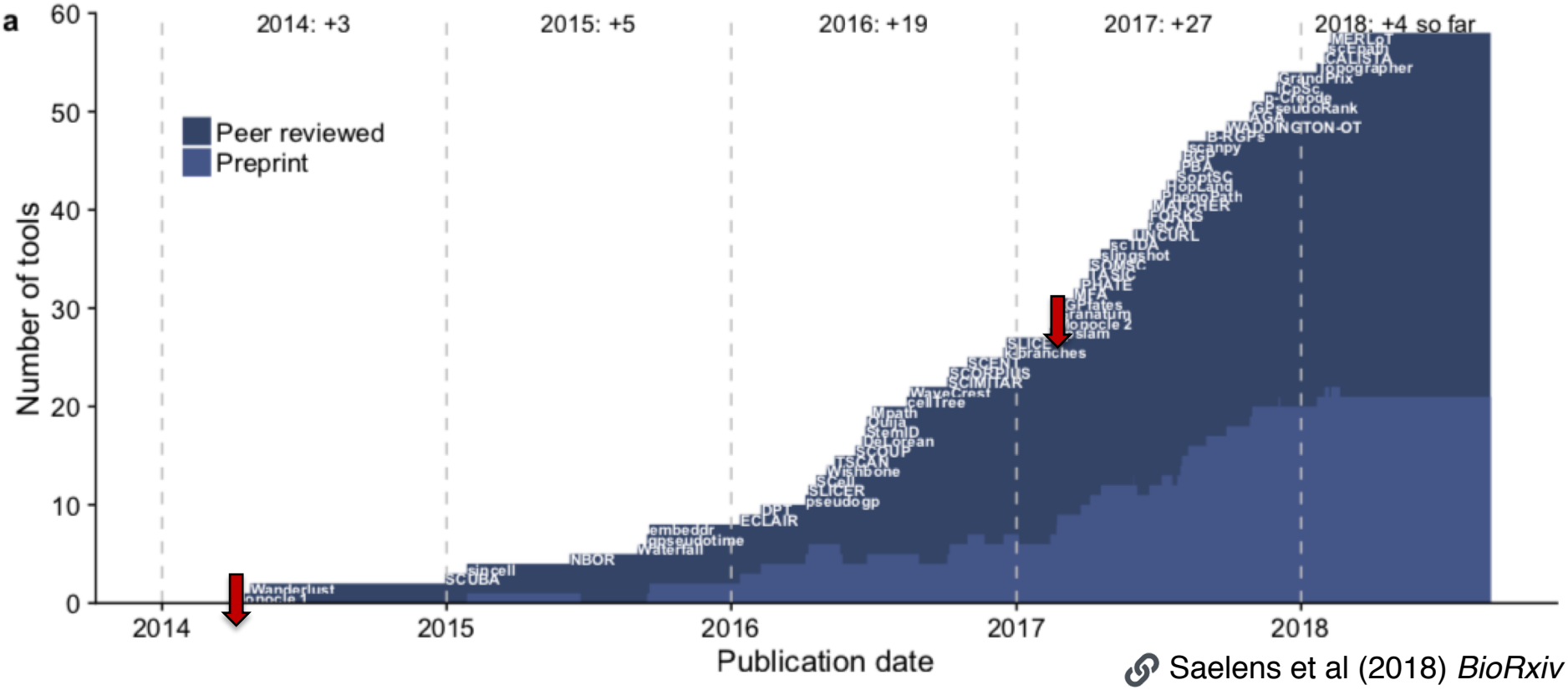
Do you have intermediate states?

Do you believe that you have branching in your trajectory?

! Be aware, any dataset can be forced into a trajectory without any biological meaning!

! First make sure that gene set and dimensionality reduction captures what you expect.

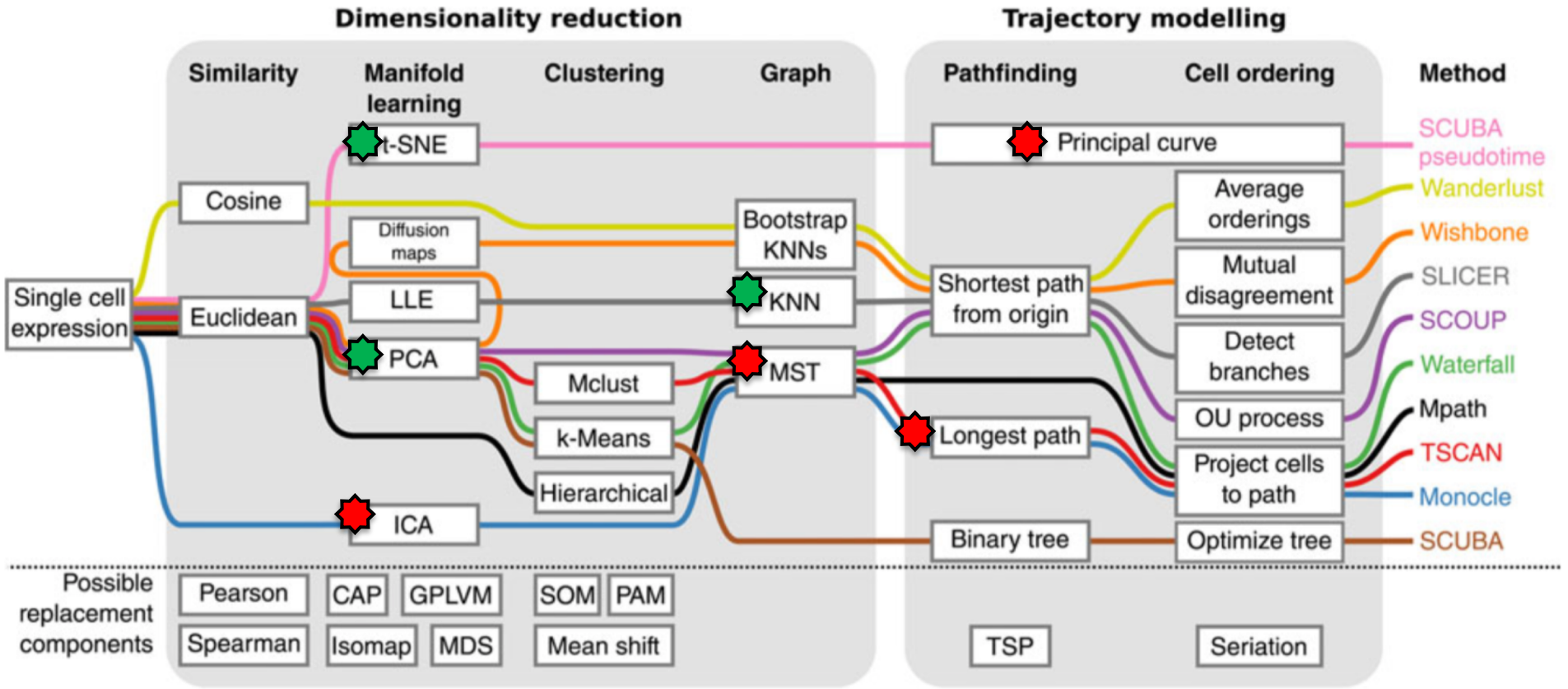
FAST development of Trajectory Inference



Saelens et al (2018) *BioRxiv*

Saelens et al (2019) *Nat Biotechnology*

Trajectory Inference Overview



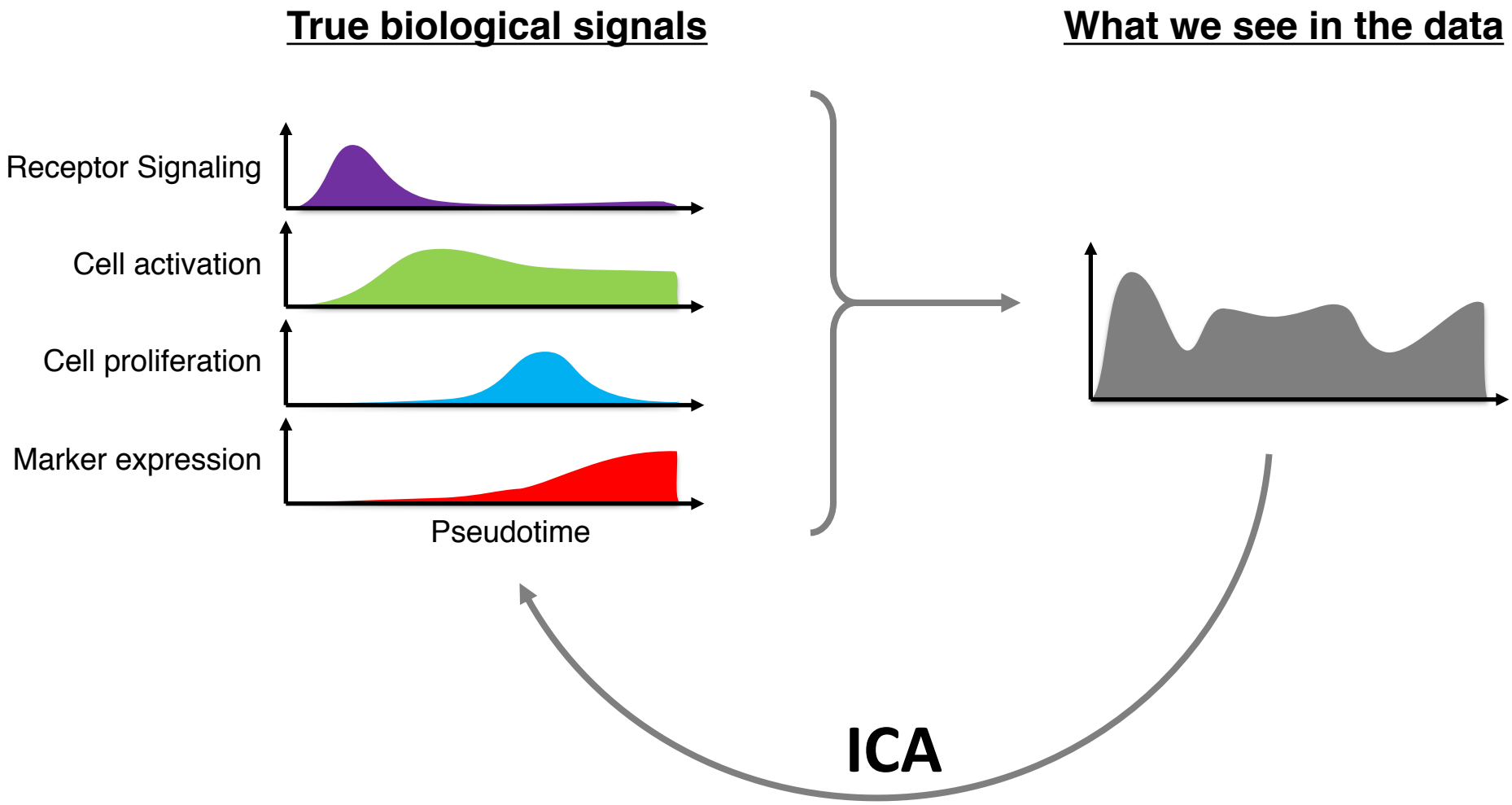
Cannoodt et al (2016) *Eur J Immunol*

ICA

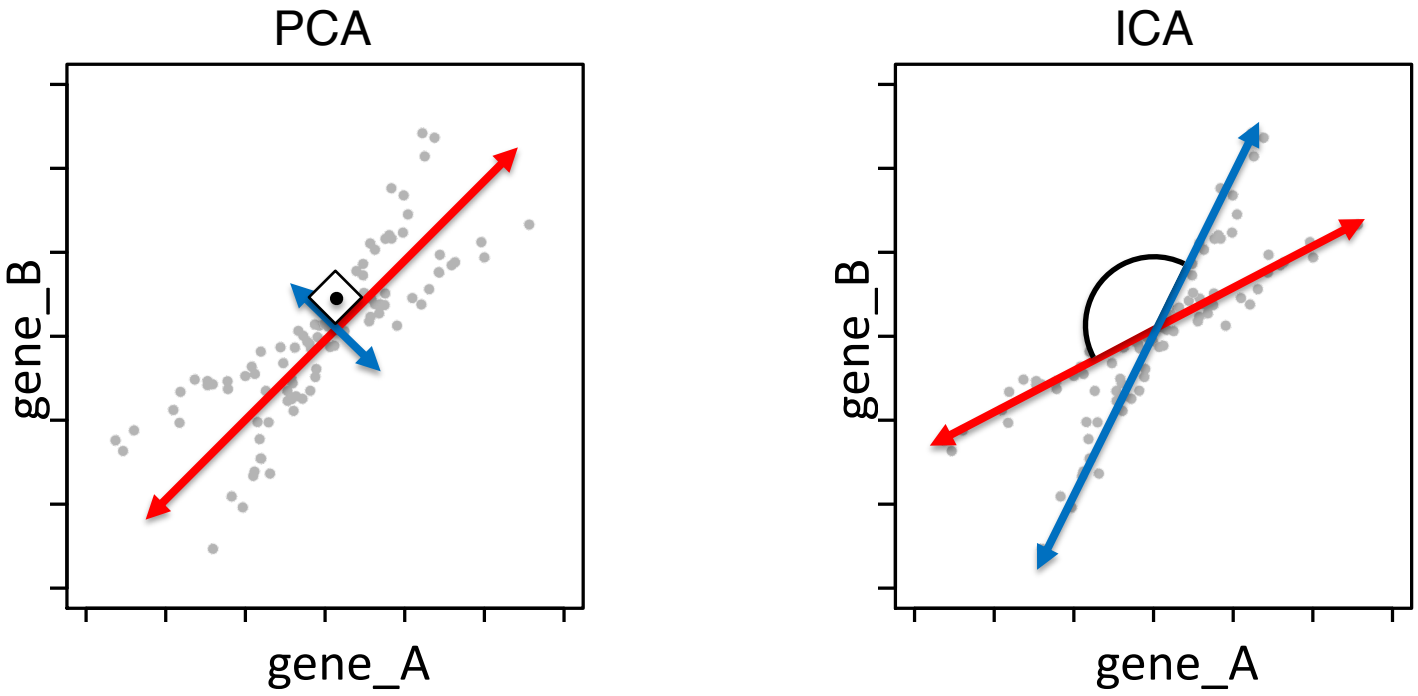
Independent Component Analysis

A method for decomposing the data

Why ICA?


True biological signals
What we see in the data
Receptor Signaling
Cell activation
Cell proliferation
Marker expression
Pseudotime
ICA

How does ICA work?

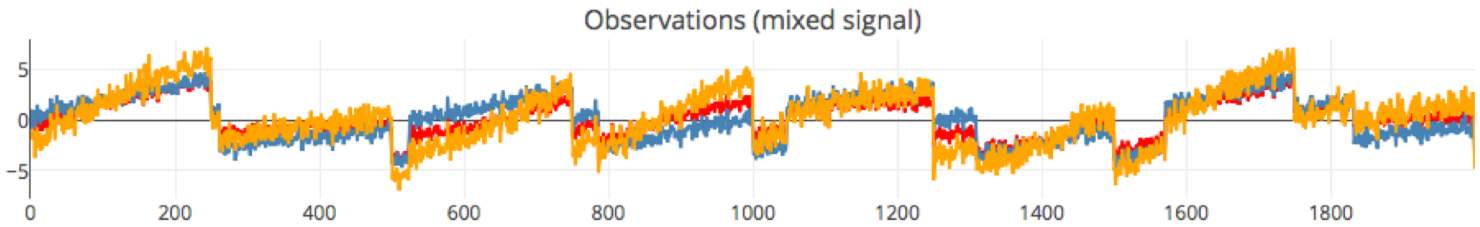


ICA assumes that:

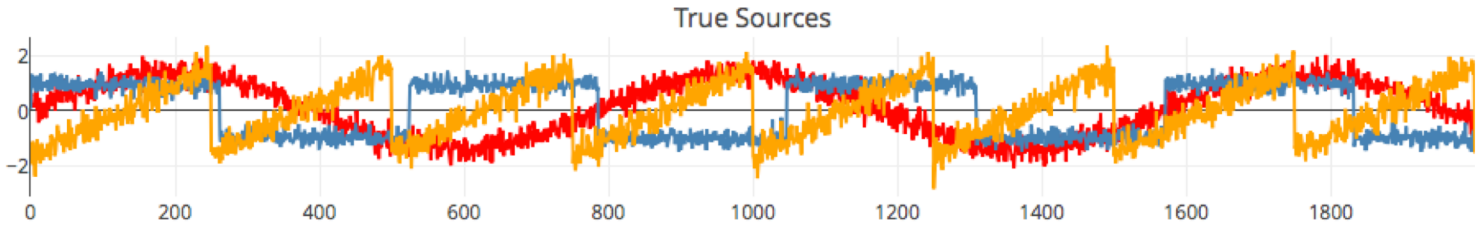
1. The source signals are independent of each other.
2. The values in each source signal have non-Gaussian distributions.

A visual intuition for ICA

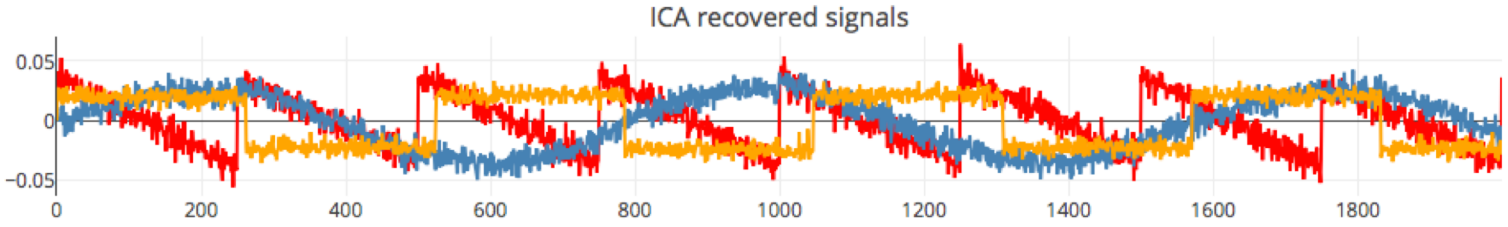
Analogy to single cell



Raw Gene Expression



Biological Processes (source)



ICA

https://plot.ly/scikit-learn/plot-ica-blind-source-separation/

ICA: summary

It is a LINEAR method of dimensionality reduction.

ICA is used to estimate the sources that compose the data.

The sources are assumed to be independent of each other
This might not be true for single cell

Problems with ICA for single cell data:

Assumes that the data distribution is non-Gaussian
This might not be true for single cell

Each component has equal importance
Unlike PCA where they are sorted by variance
ICA cannot identify the actual number of source signals

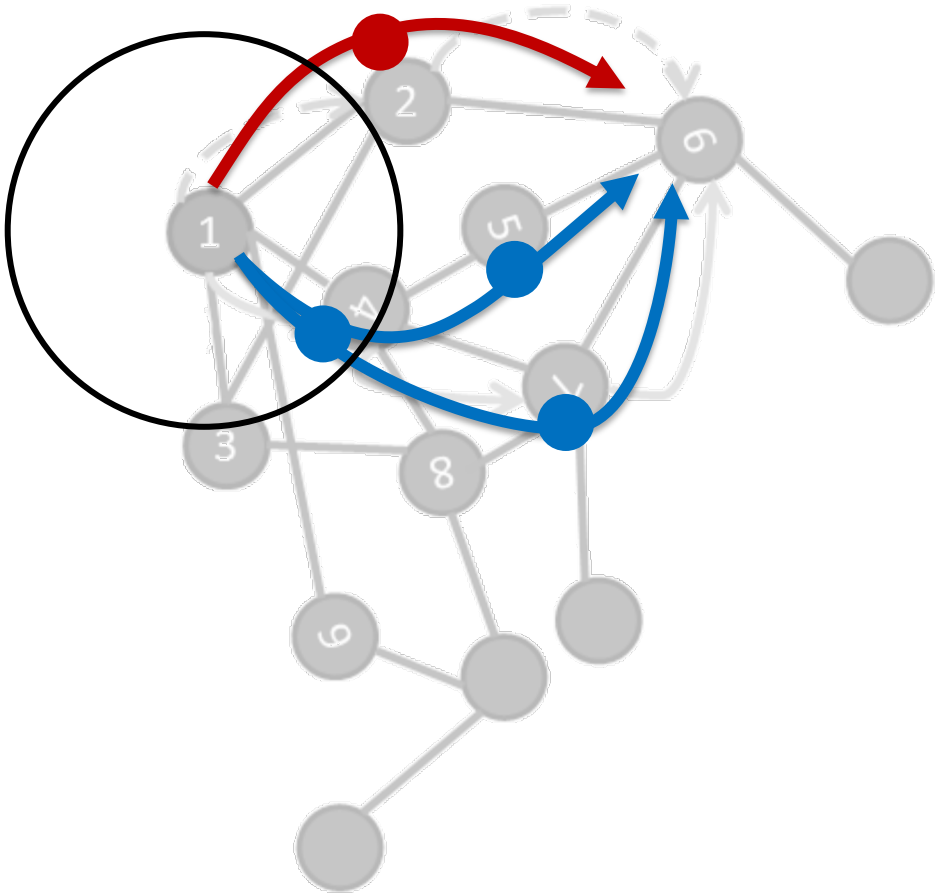
Diffusion Maps

in brief

How Diffusion Maps work?

Diffusion maps is a non-linear dimensionality reduction algorithm

The distance between points A and B is defined as the probability of going through the nodes using K steps.



#2 Steps (1|6):
P(1|2) * P(2|6) = 0.2

#3 Steps (1|6):
P(1|4) * P(4|5) * P(5|6) +
P(1|4) * P(4|7) * P(7|6)

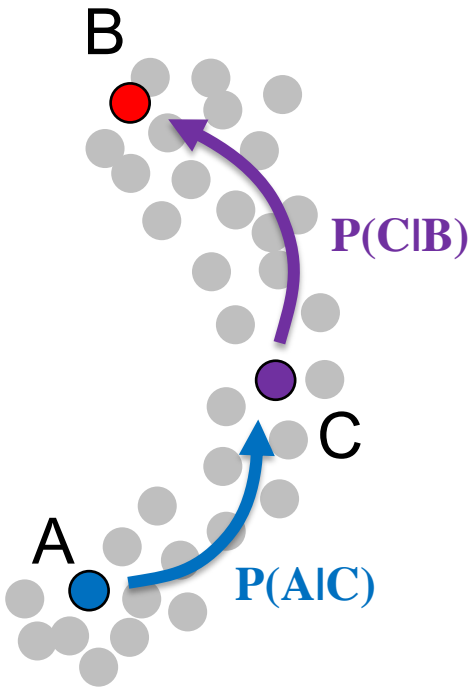
de la Porte et al (2008)
Coifman et al (2005) PNAS

How Diffusion Maps work?

To transform probabilities to distance, diffusion maps calculates the difference in probabilities to an intermediate point:

$$\text{diff\_dist}(A|B) = P(A|C) - P(C|B)$$

If the $P(A|C) \approx P(C|B)$, $\text{dist}(A|B)$ approaches 0, indicating that **A** and **B** are <u>well</u> connected via the intermediate point **C**.



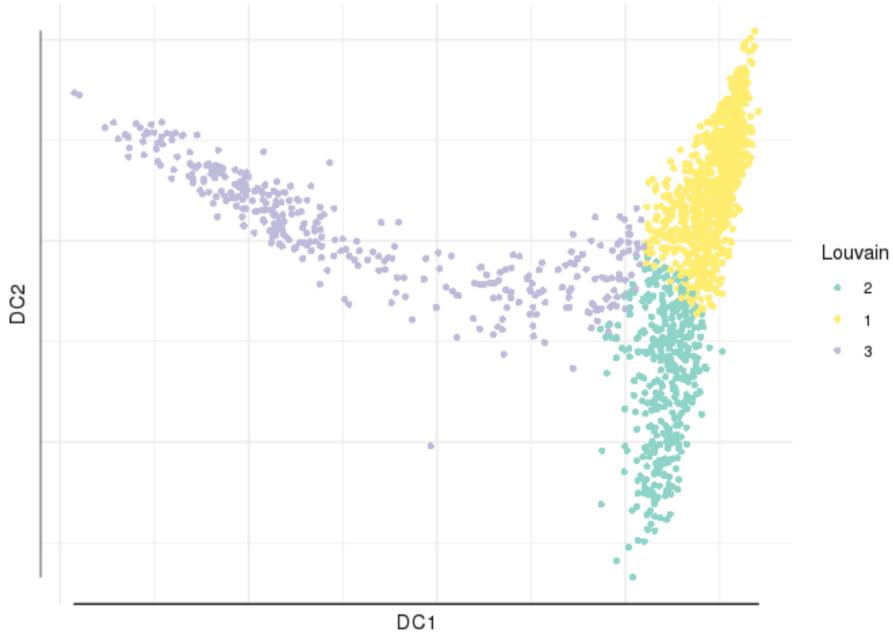
Dimensionality reduction is done by eigenvalue decomposition (like PCA does). The dimensions should be selected by the contribution to each dimension (like PCA).

Diffusion Maps: summary

It is a NON-LINEAR method of dimensionality reduction.

The distances between points are measured as probability from going from one to another.

The data must present connectivity (transitional cells).



MST

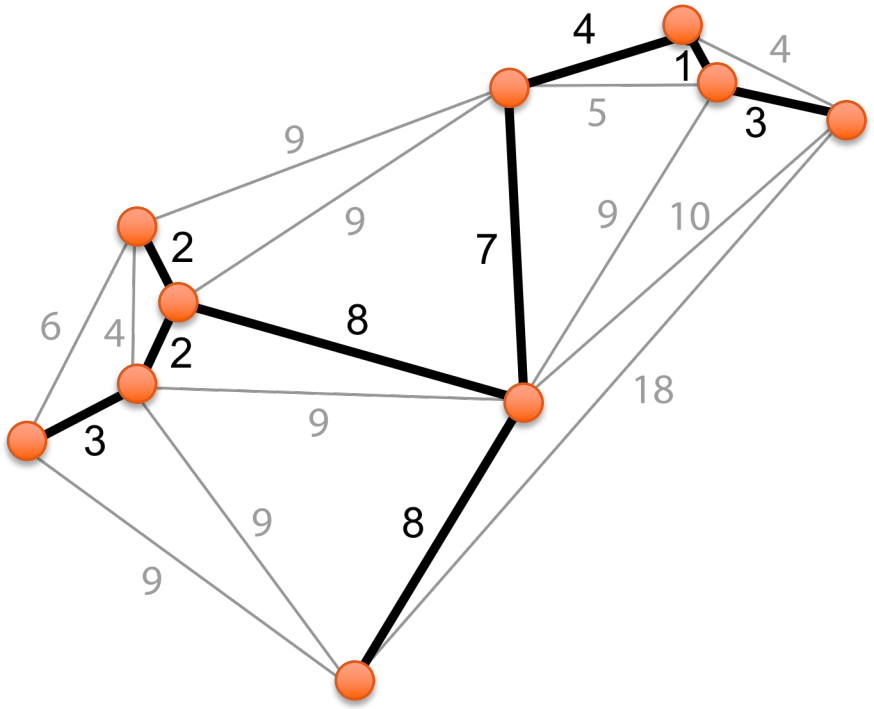
Minimum spamming tree

What is a minimum spanning tree (MST)?

Given a set of points,
how do we connect them so that the total sum of all distances is minimized?

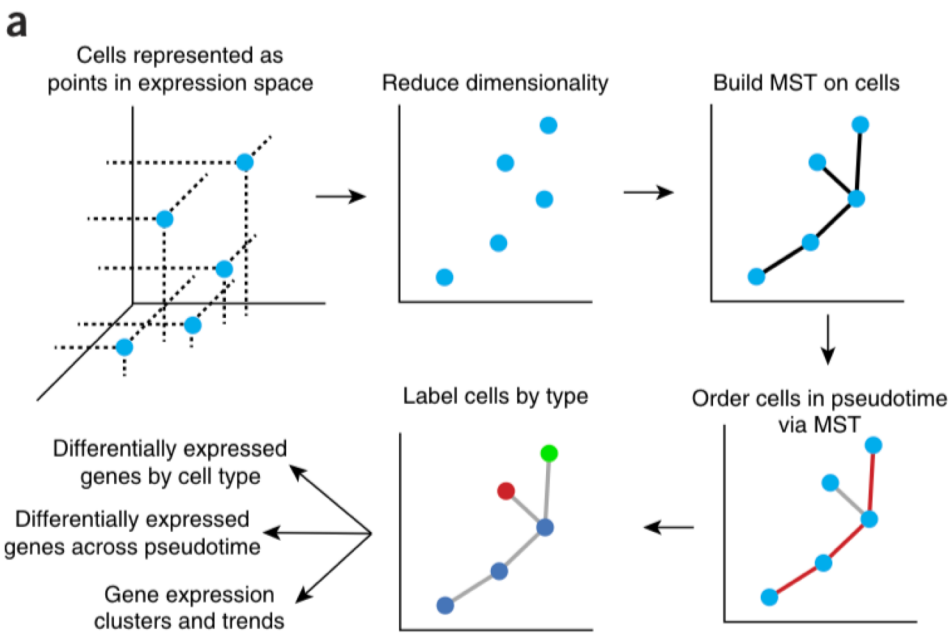
Having more transitional cells
improves the definition of the tree

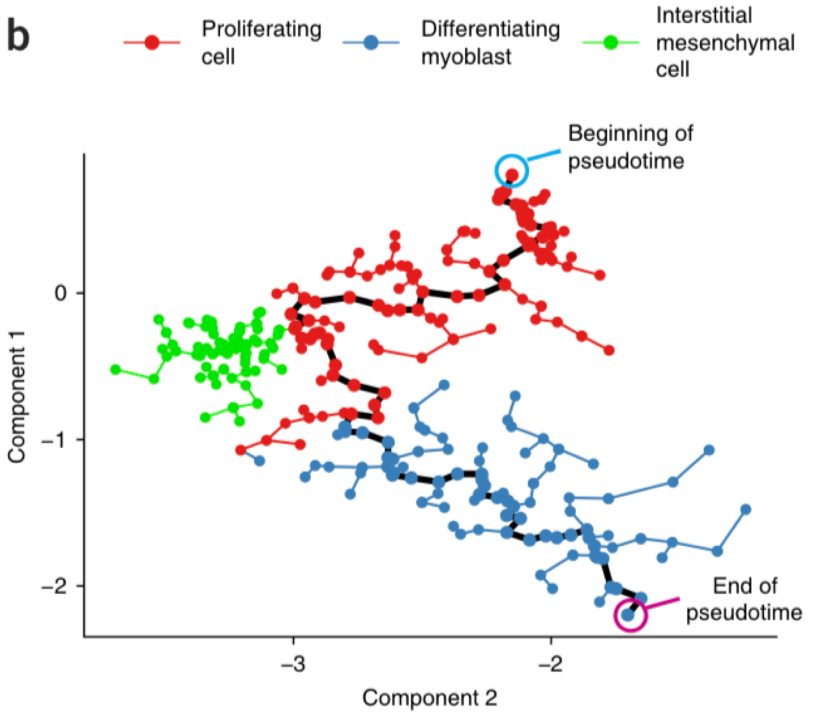
The weights can be the
distance is the ICA space
or a correlation between
cells, etc.



By definition, a MST has no cycles
So you cannot use MST to define cyclic trajectories (i.e. cell cycle)

Monocle ICA (v1)





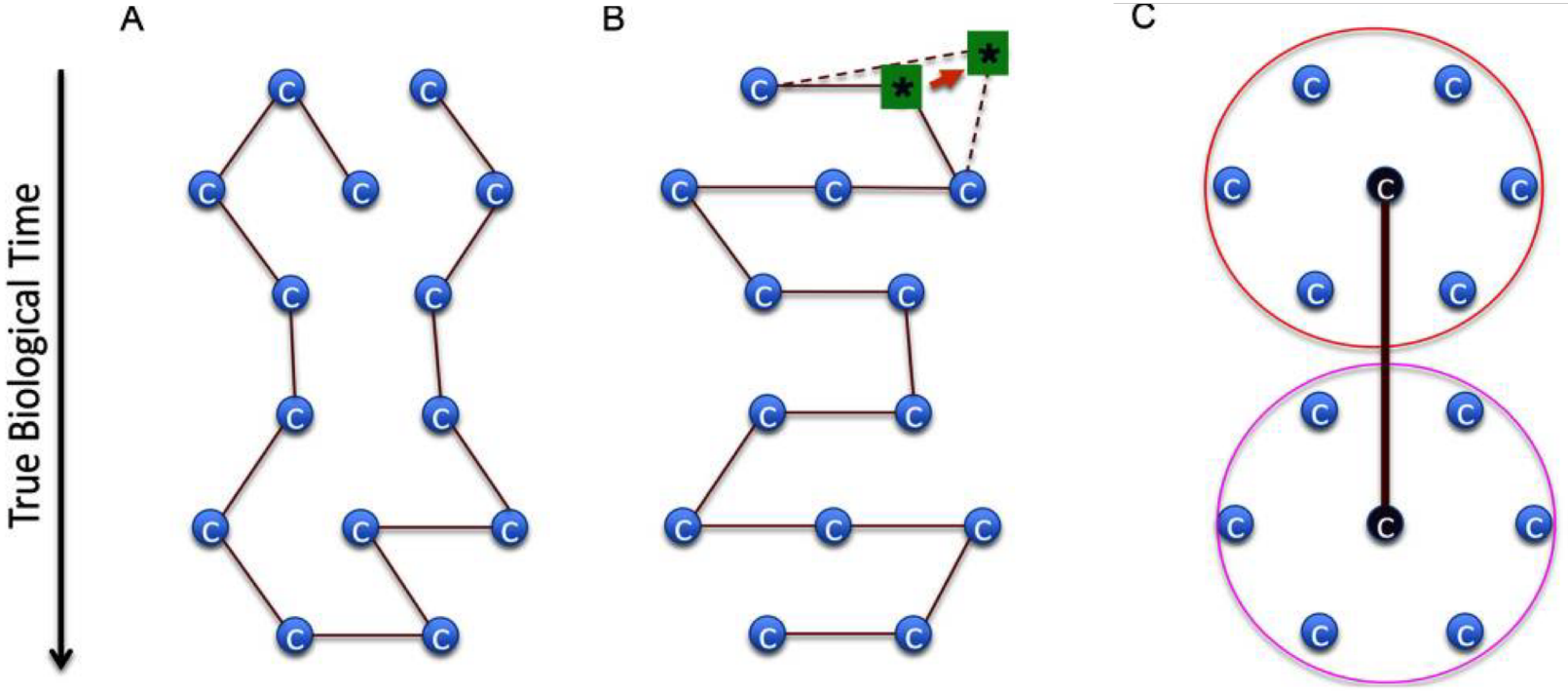
Trapnell et al (2014) *Nat Biot*

Reverse graph embedding

(RGE)

i.e. DDRTree and others

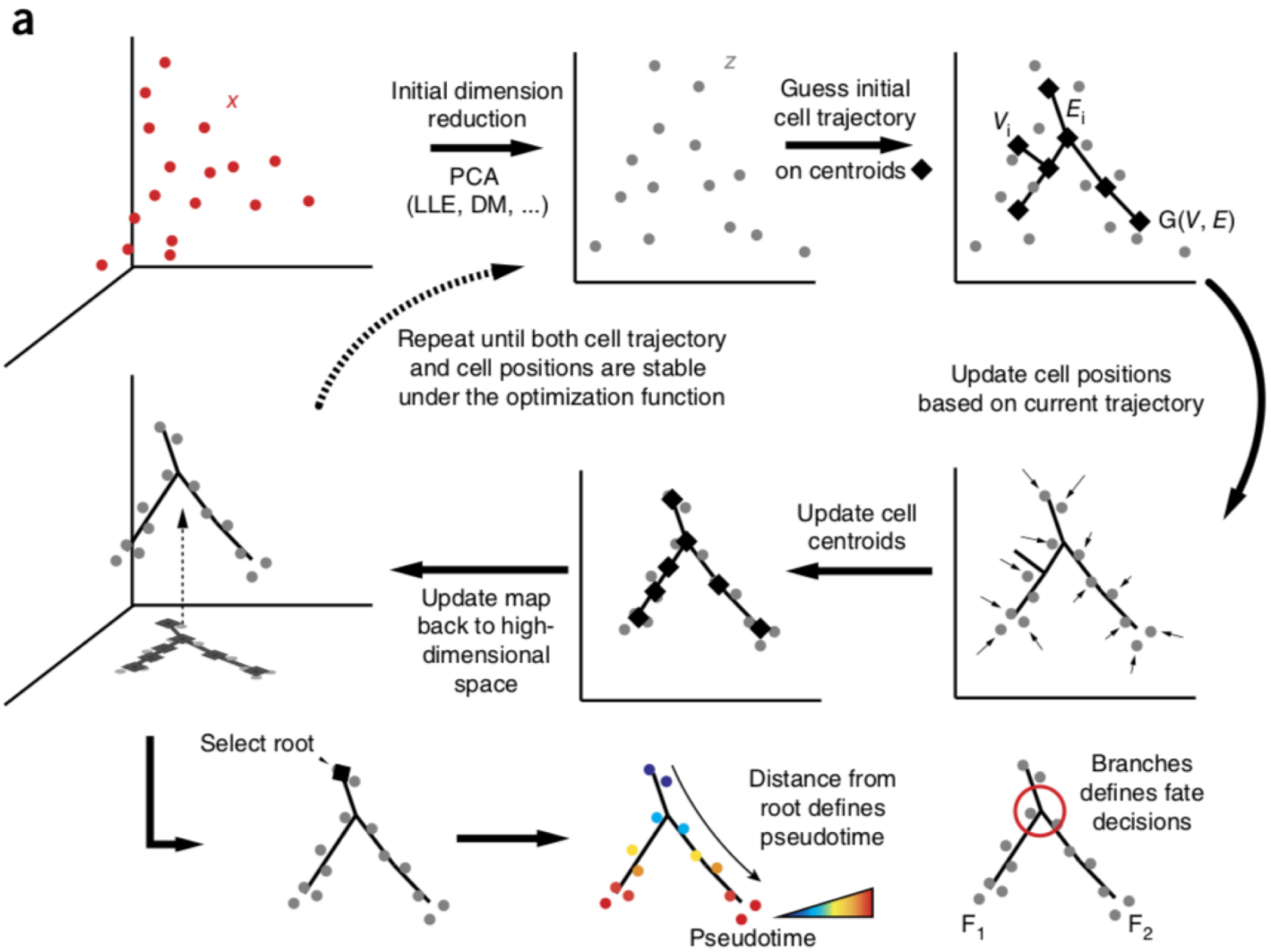
The limitation of MST



Zhicheng et al (2016) *Nuc Acid Res*

Trajectory construction using MST is highly dependent on single data points

Monocle DDRTree (v2)

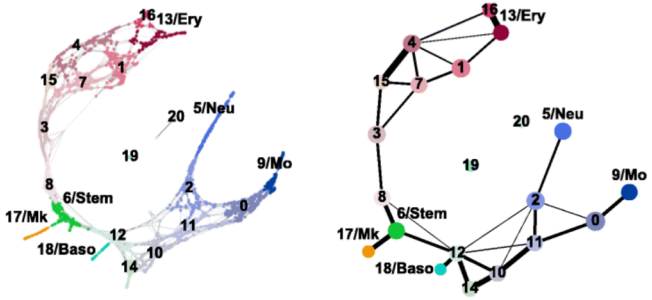


Qiu et al (2017) *Nat Methods*

Many methods derived from RGE idea

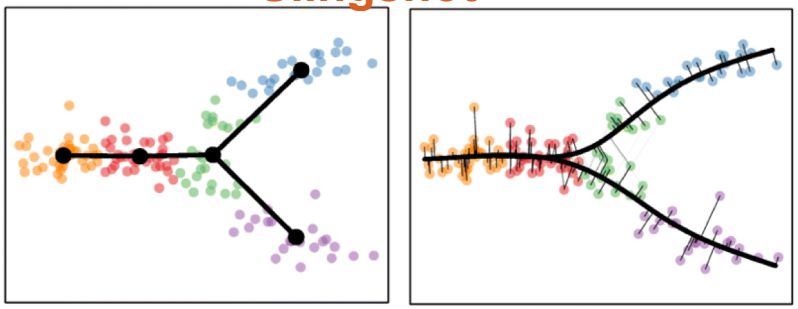
The methods differ on the dimensionality reduction used, the clustering method or the way the tree is constructed

PAGA



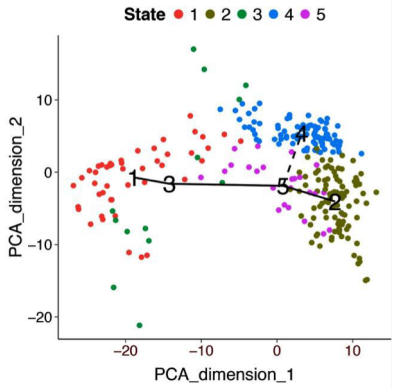
Wolf et al (2019) *Genome Biology*

Slingshot



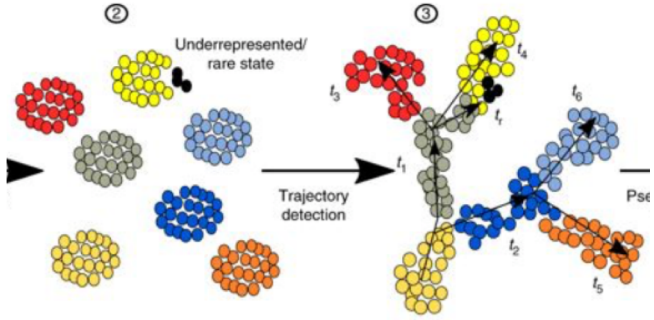
Street et al (2018) *BMC Genomics*

TSCAN



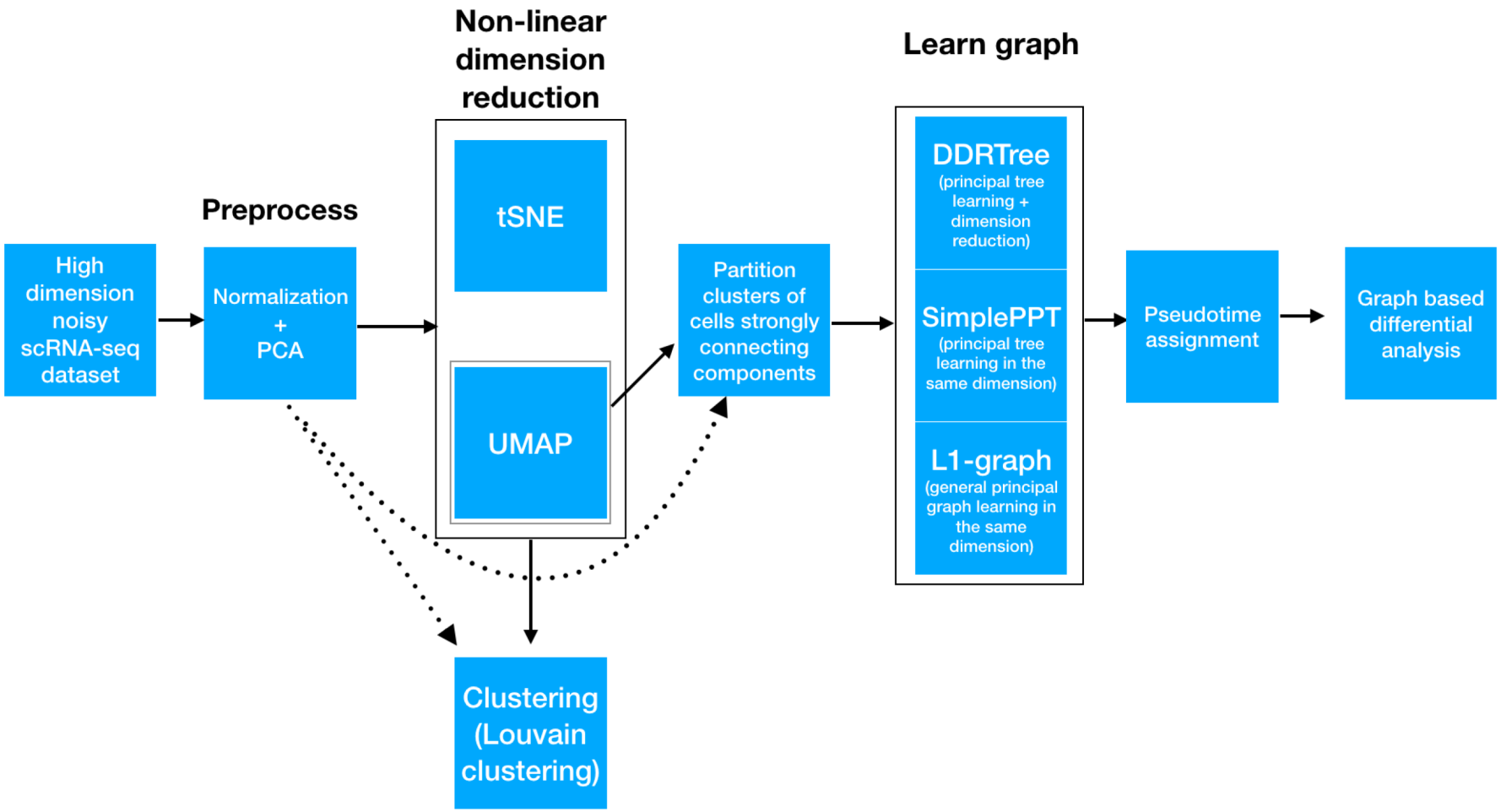
Zhicheng et al (2016) *Nuc Acid Res*

Cell Router



Da Rocha et al (2018) *Nat Commun*

Monocle UMAP (v3)

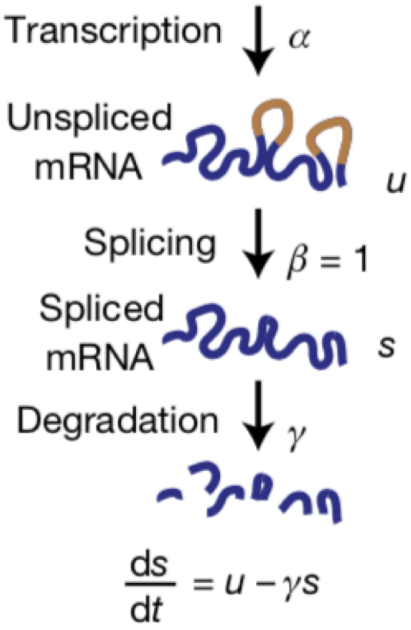


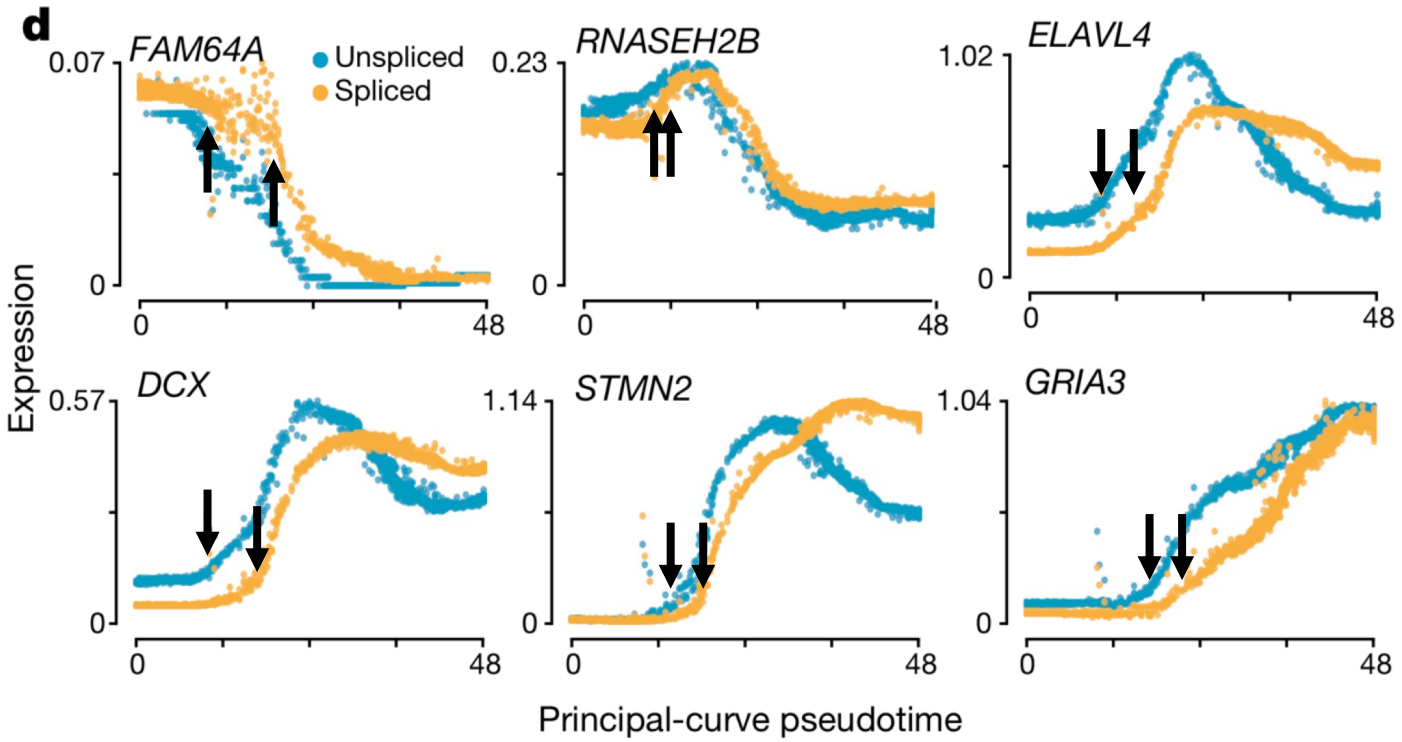
RNA velocity

gene expression trajectory

How does RNA velocity work?

It uses the proportion spliced/unspliced reads to predict the future state of a cell





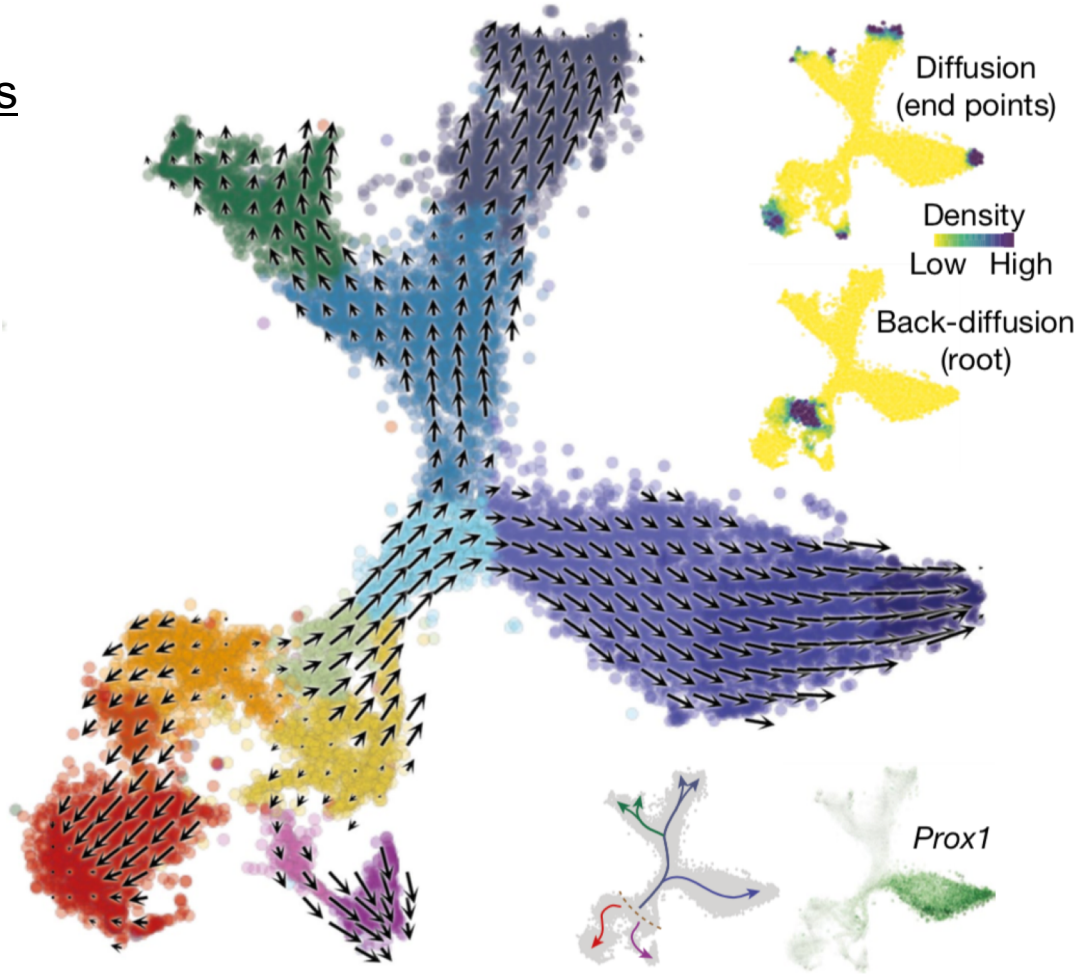
La Manno et al (2018) *Nature*

How does RNA velocity work?

RNA velocity allows a biologically-driven identification of cell transcriptional trajectories:

Defines start, ends and bifurcations

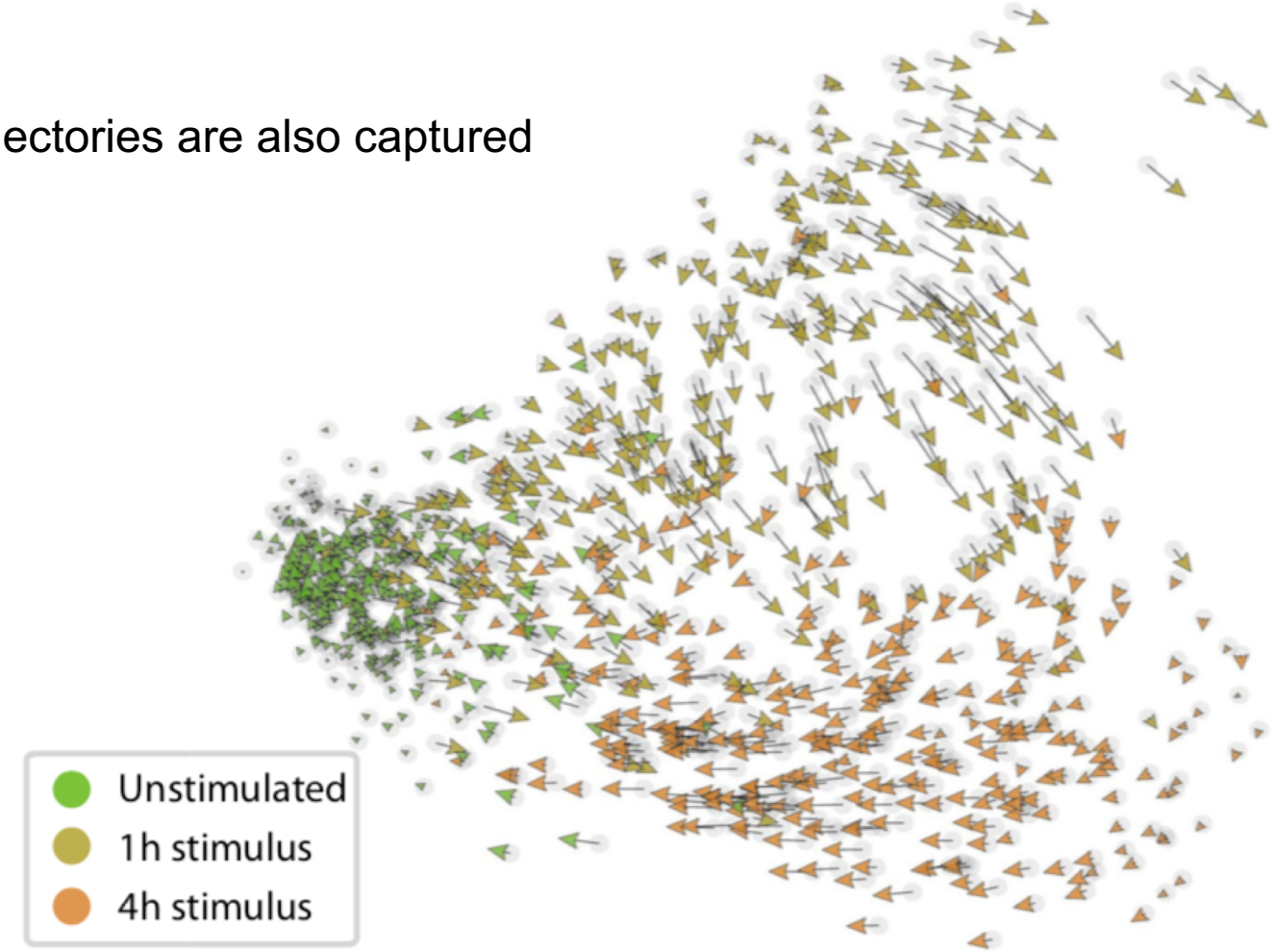
The position of the spliced is represented by the arrow-head



La Manno et al (2018) *Nature*

How does RNA velocity work?

Cyclic trajectories are also captured



La Manno et al (2018) *Nature*

Wrap-up

Final Considerations

- In reality, distance in multidimensional space reflects difference in transcriptional landscape, not actual time.

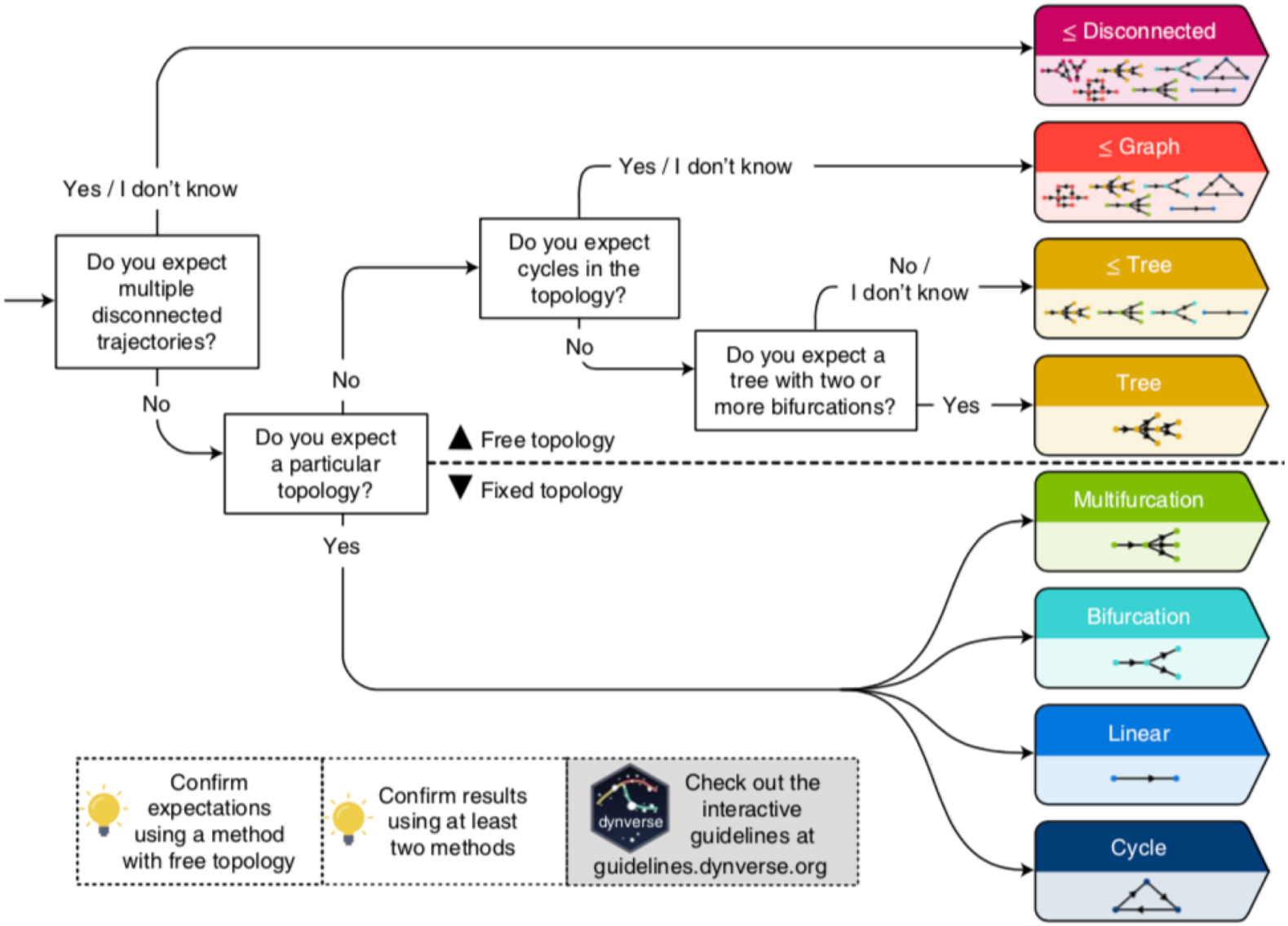
- Necessary to have a continuum of states among your cells

  Will not work well with 2 distinct clusters.

- May work with single time-point if ongoing differentiation process

  It is better to have multiple experimental time points.

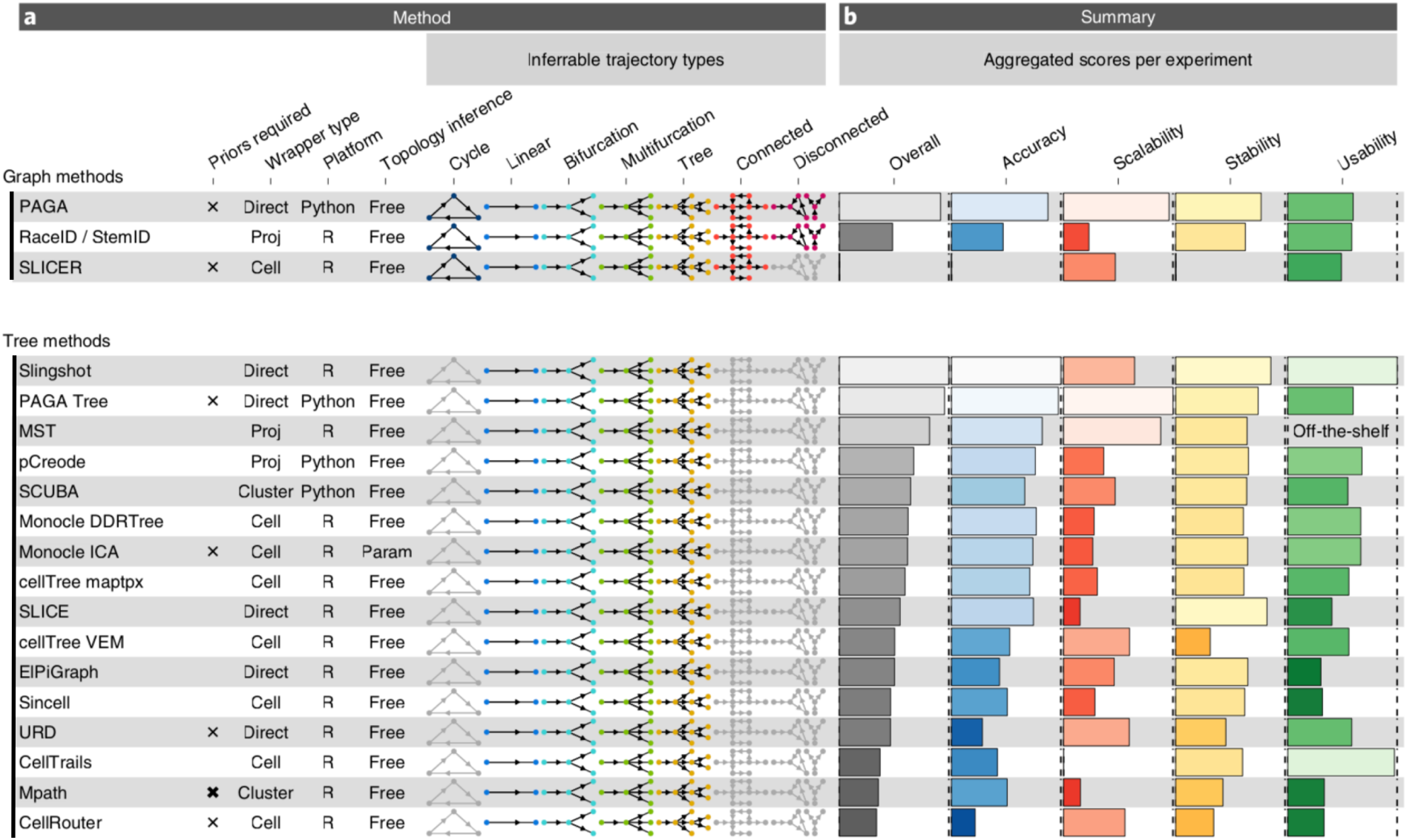
Which method should I use?



http://guidelines.dynverse.org

Saelens et al (2019) *Nat Biotechnology*

Which method should I use?



http://guidelines.dynverse.org

Saelens et al (2019) *Nat Biotechnology*