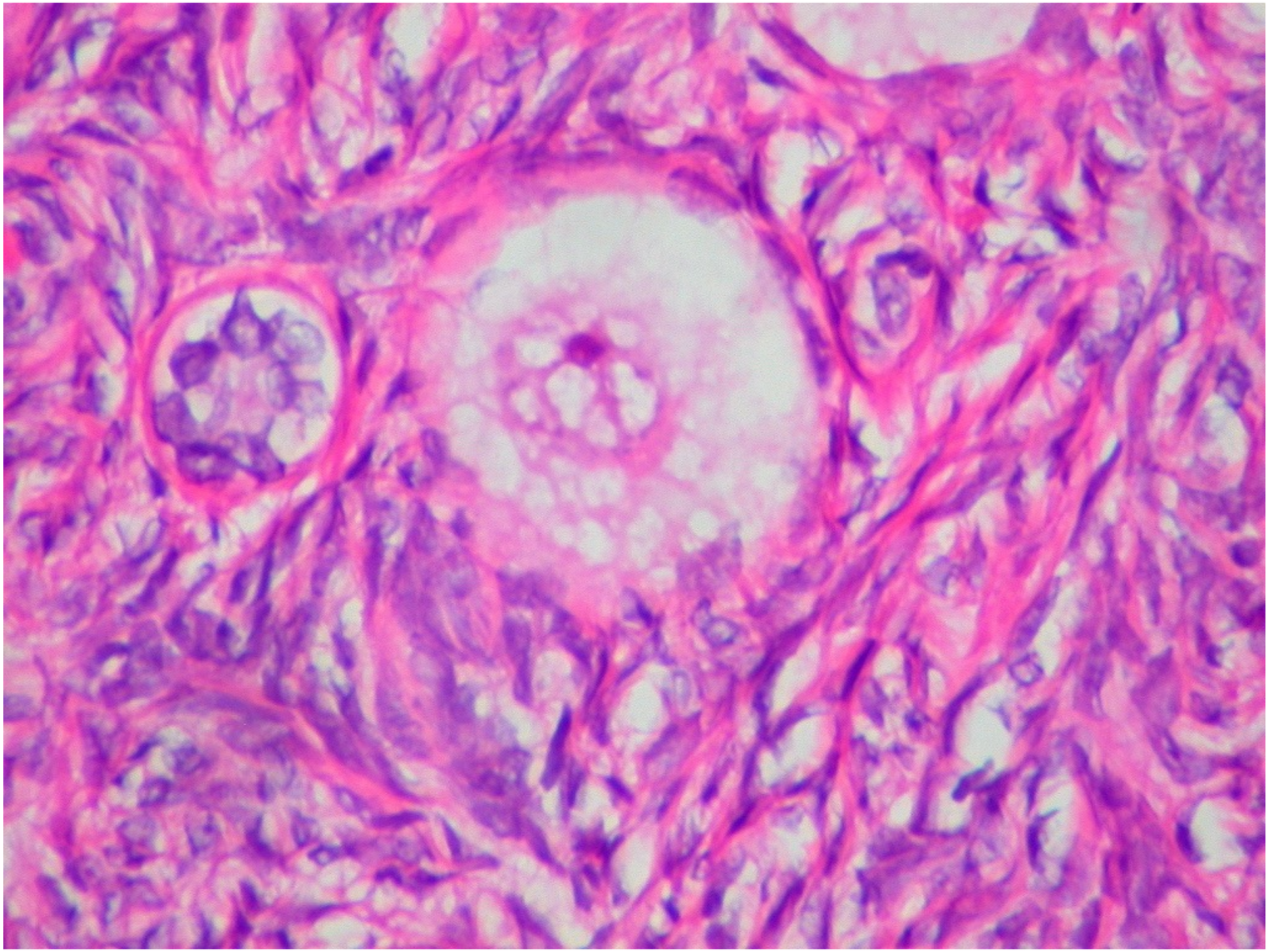
Trajectory or pseudotime analysis of single-cell RNA-seq data

Davis McCarthy
NHMRC Early Career Fellow
Stegle Group

Physalia Course, Berlin
08 February 2018

www.ebi.ac.uk

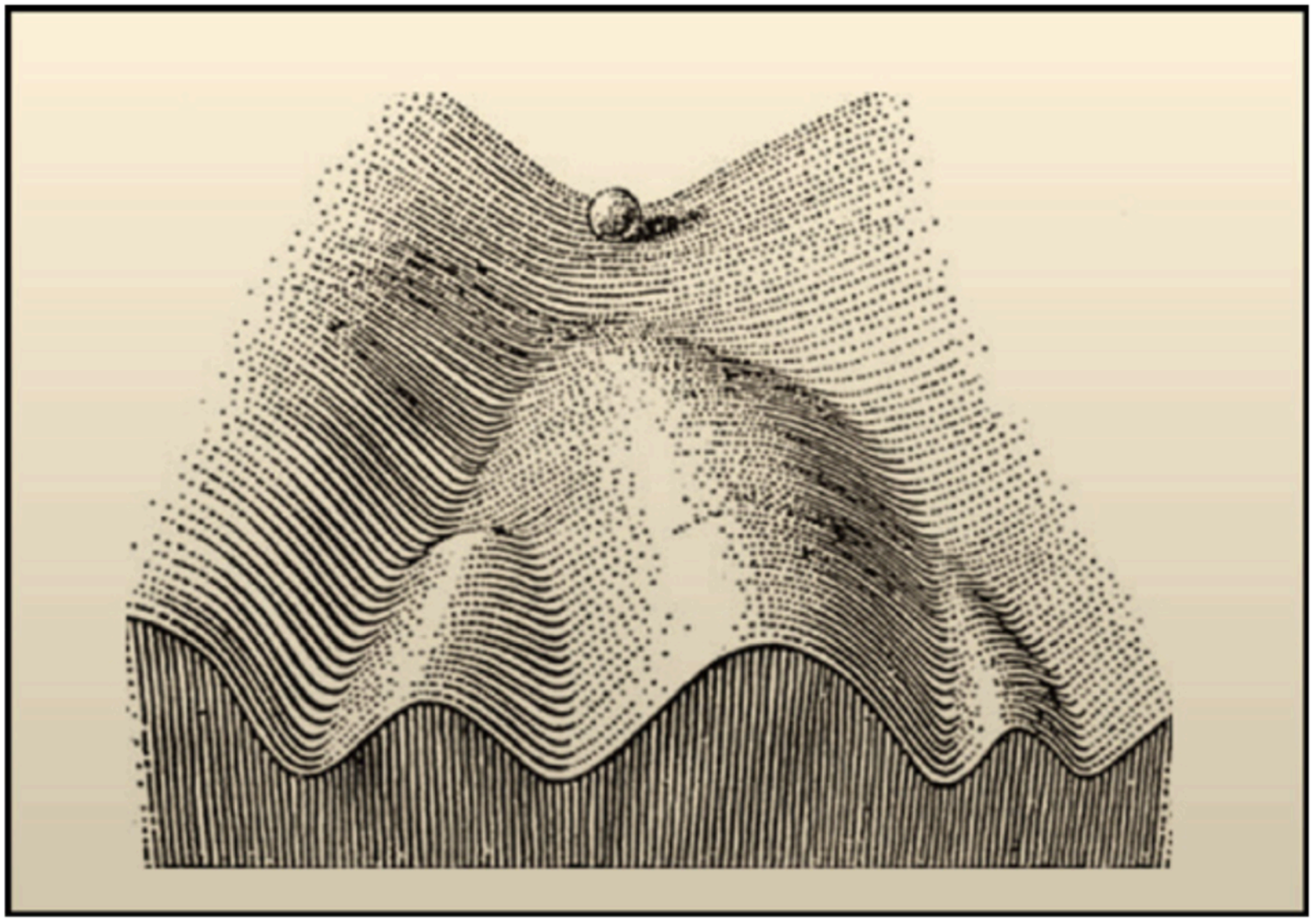
Importance of studying development



Photograph: Ed Uthman, Flickr Creative Commons

How do cells “make decisions” about their development?

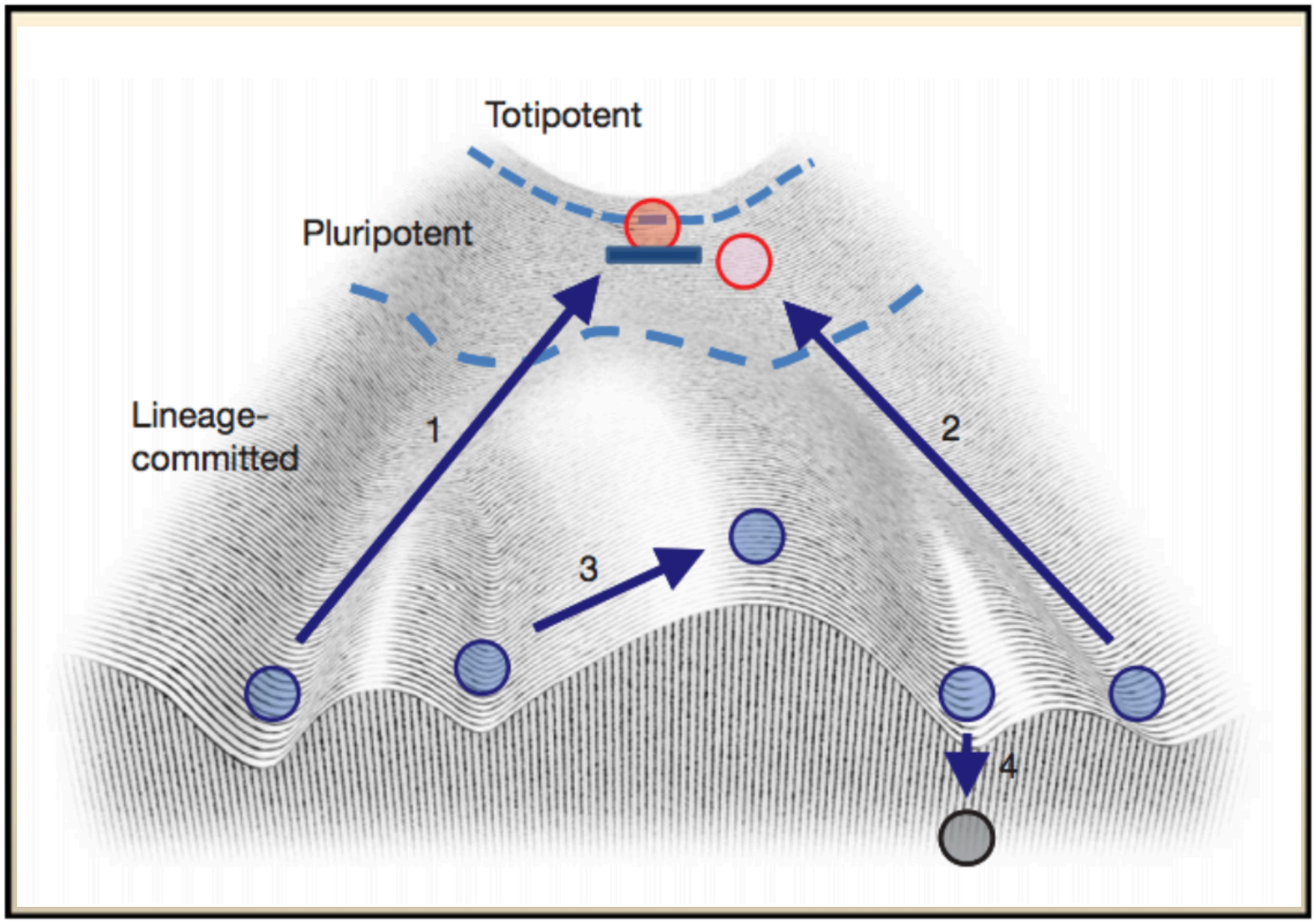
Waddington’s landscape



Waddington's epigenetic landscape is a metaphor for how gene regulation modulates development.

Waddington, 1957 / Yamanaka, *Nature*, 2009.

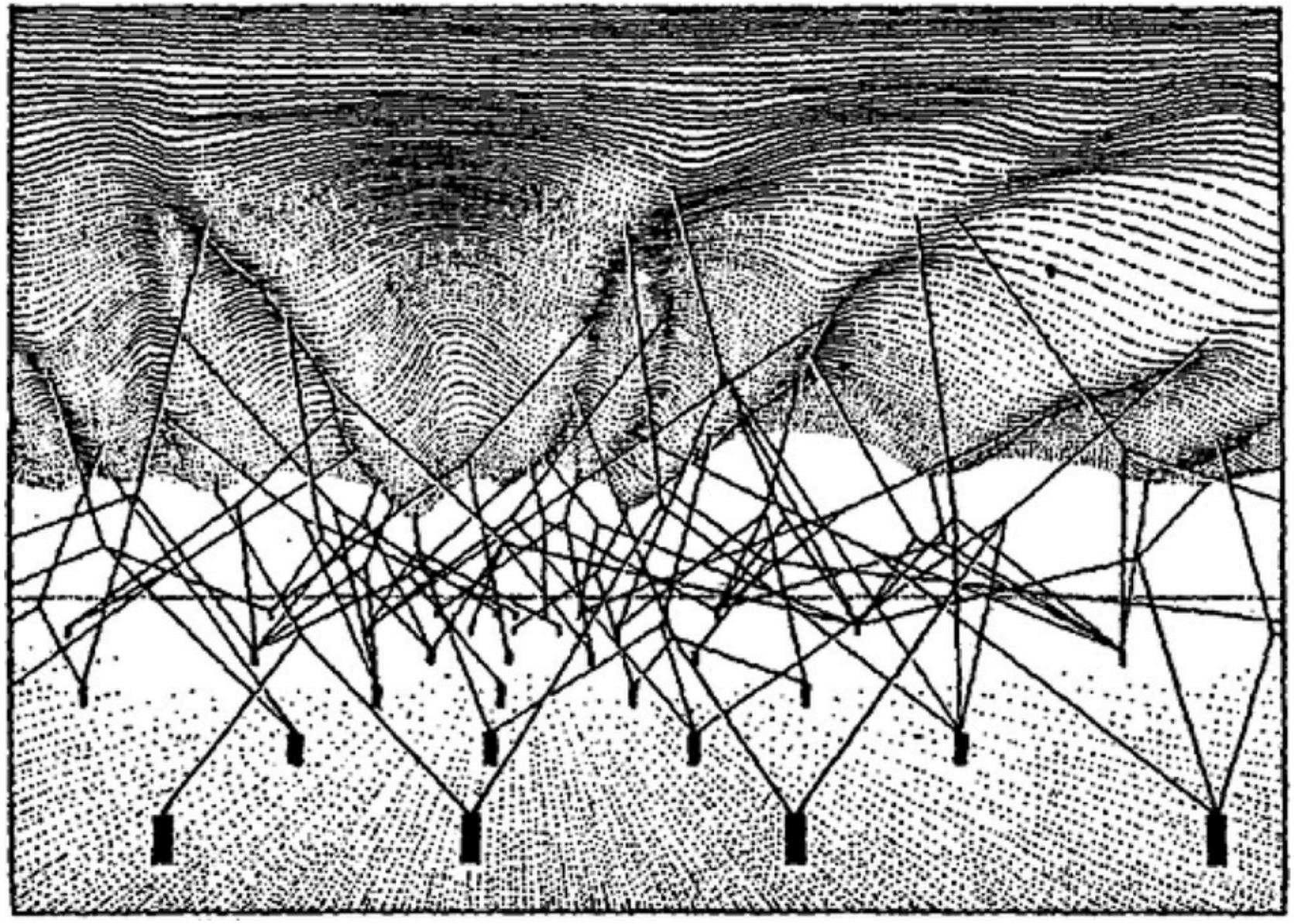
Waddington’s landscape



Waddington's epigenetic landscape is a metaphor for how gene regulation modulates development.

Waddington, 1957 / Yamanaka, *Nature*, 2009.

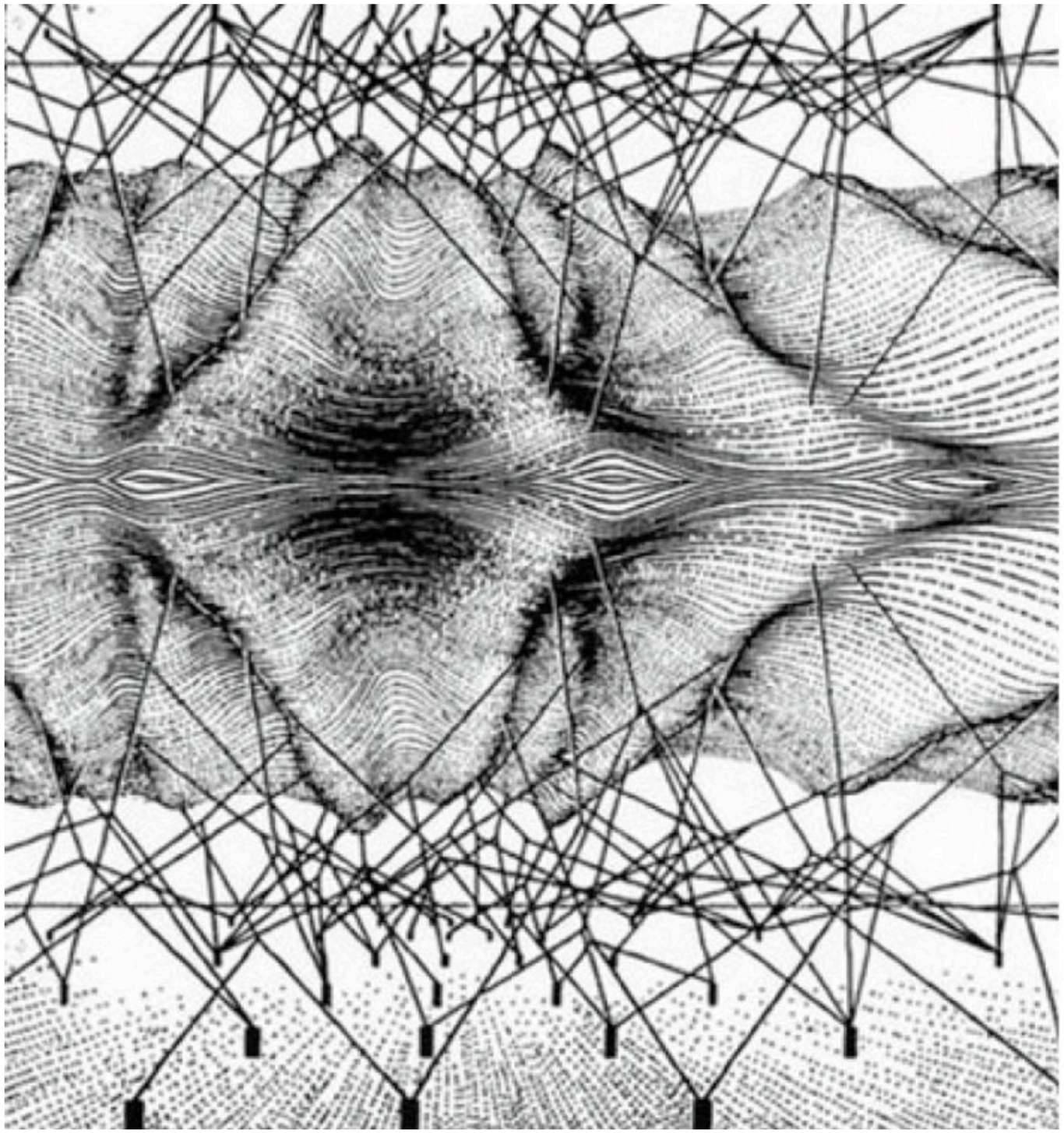
Waddington’s landscape



Waddington's epigenetic landscape is a metaphor for how gene regulation modulates development.

Waddington, 1957.

Waddington's landscape



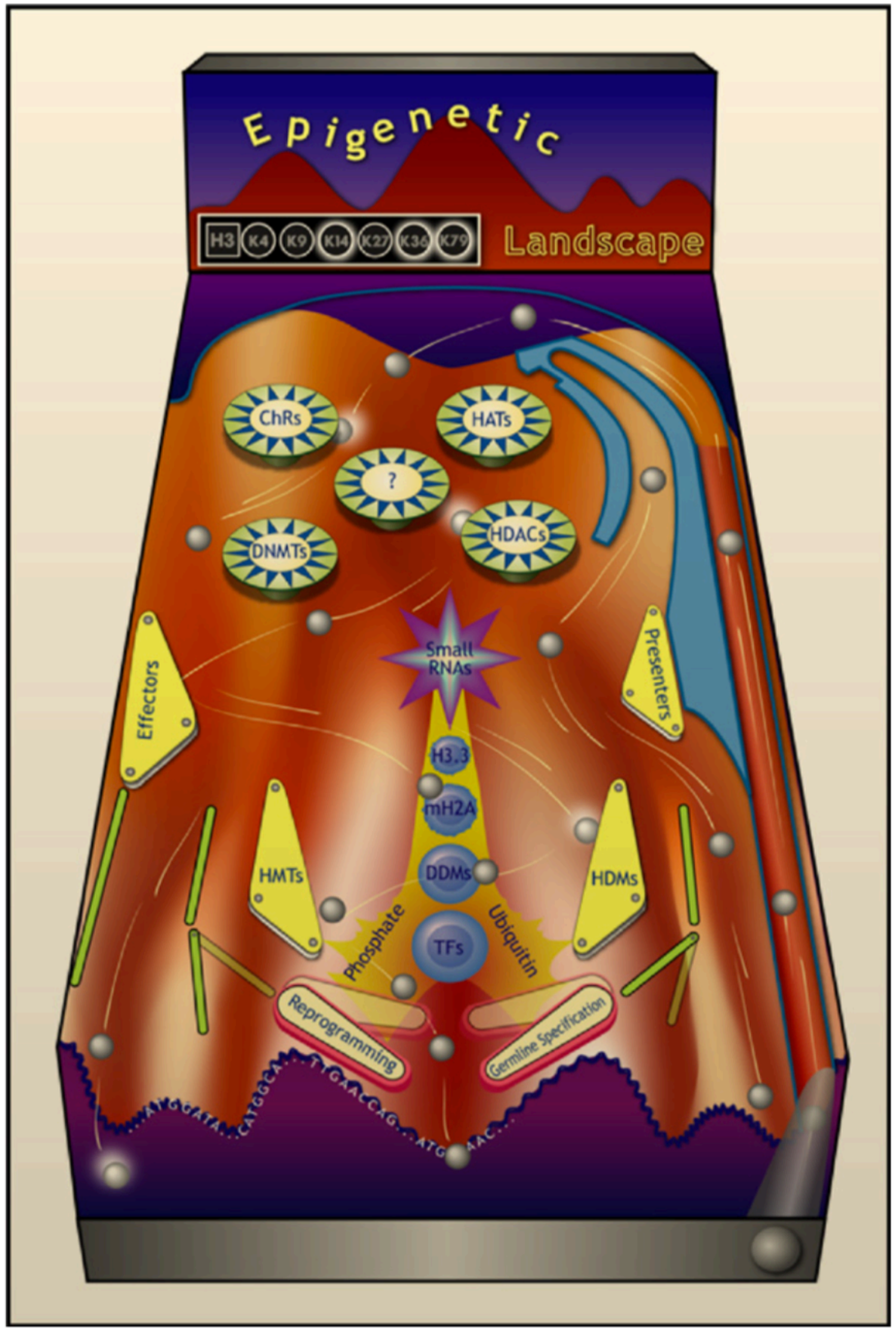
Influence of environment

Developmental landscape

Functional networks

Genes

Waddington's pinball landscape



Goldberg et al, *Cell*, 2007.

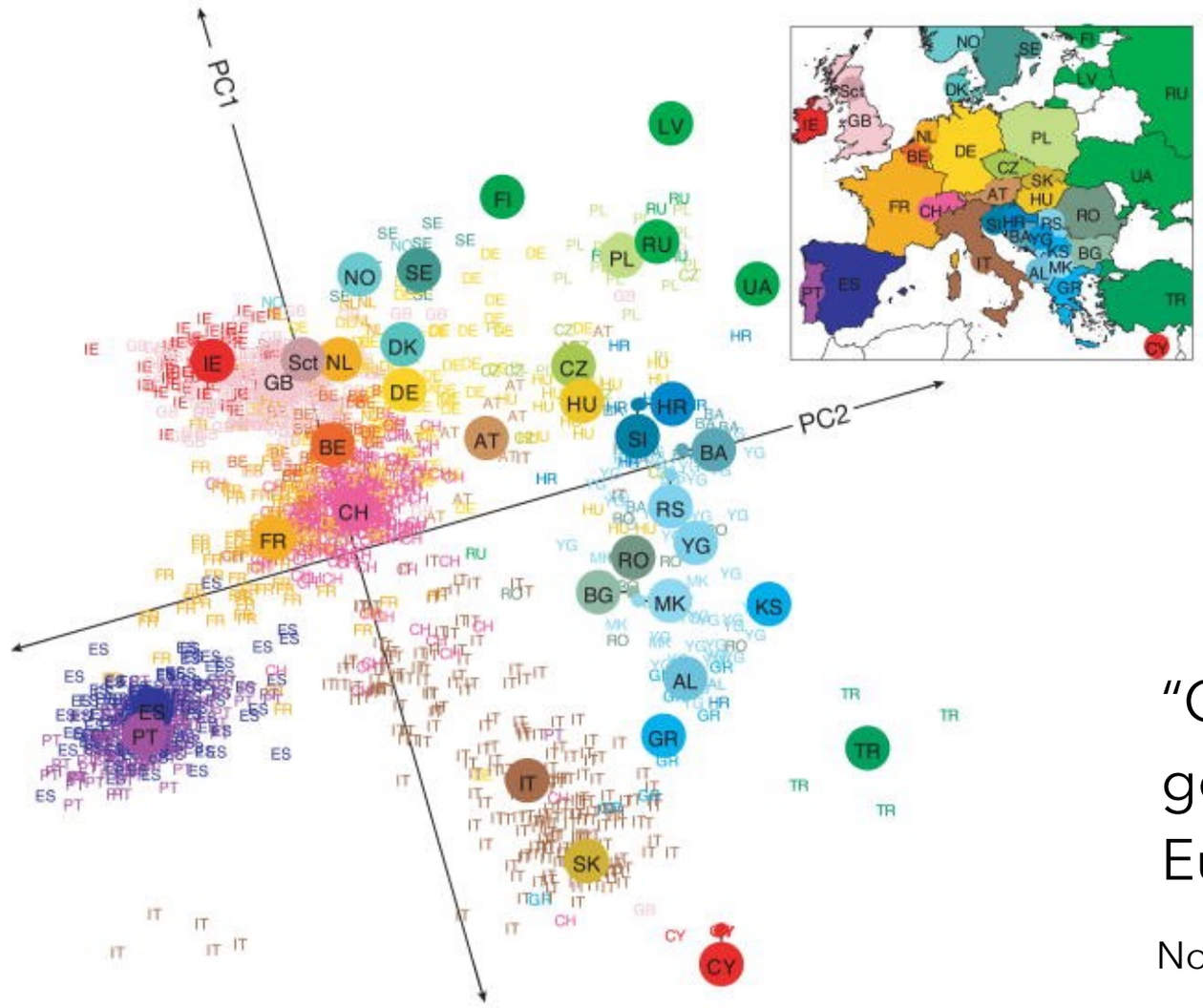
van Leeuwenhoek's microscope (http://www.history-of-the-microscope.org/)

How do we represent extremely high dimensional data in a way that we can interpret?

The "crowding problem". What do we lose? What can go wrong?

PCA: linear combinations of gene expression values to maximise variance between cells

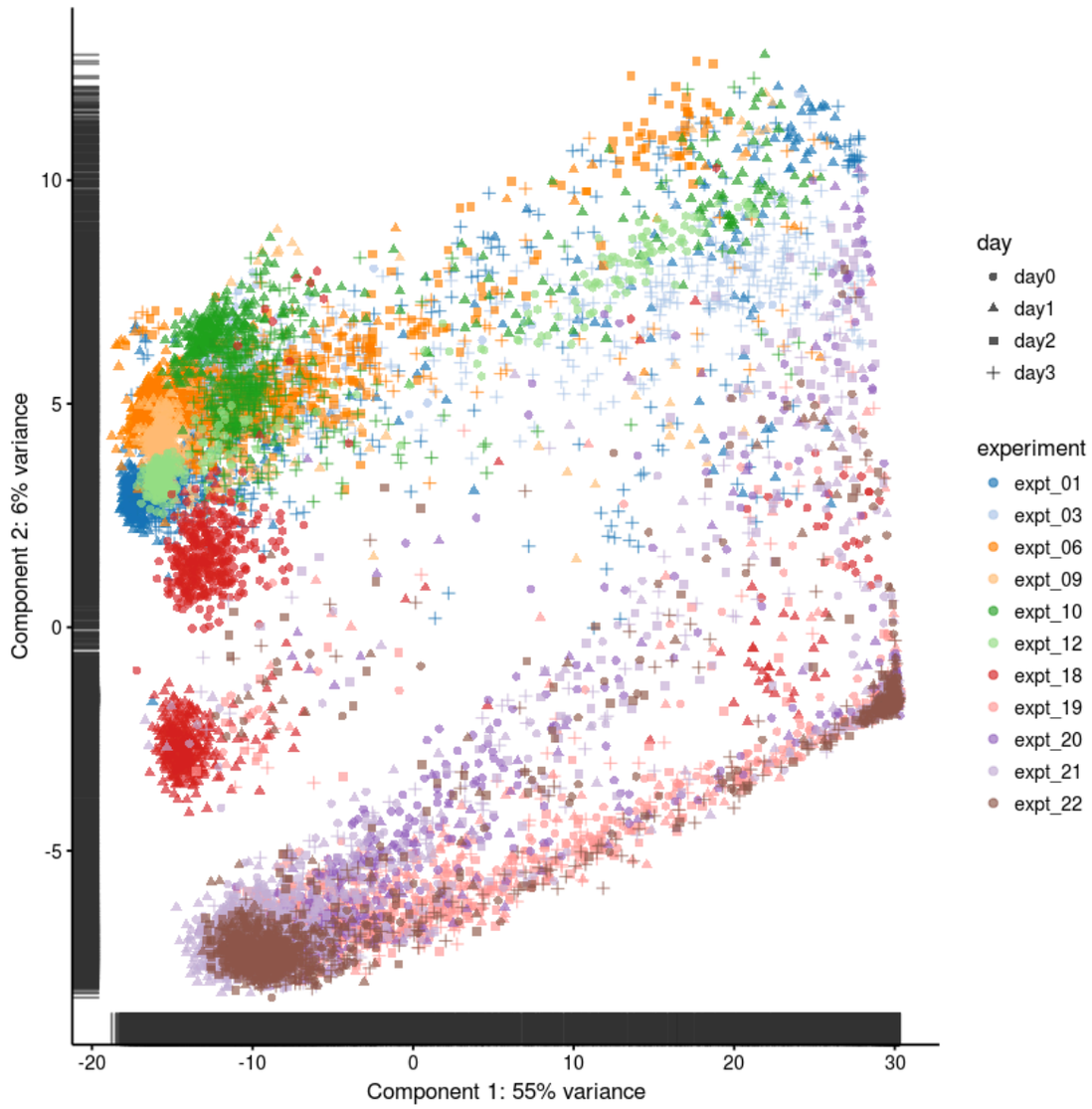
PCA: linear combinations of gene expression values to maximise variance between cells



"Genes mirror geography within Europe"

Novembre et al, *Nature,* 2008

PCA on single-cell data typically captures technical effects, most often number of genes detected



First principal component explains 55% of variance, very strongly correlated with number of genes detected per cell.

PCA extremely useful for QC.

Produced with the R/Bioconductor package *scater* (McCarthy et al, *Bioinformatics*, 2017)

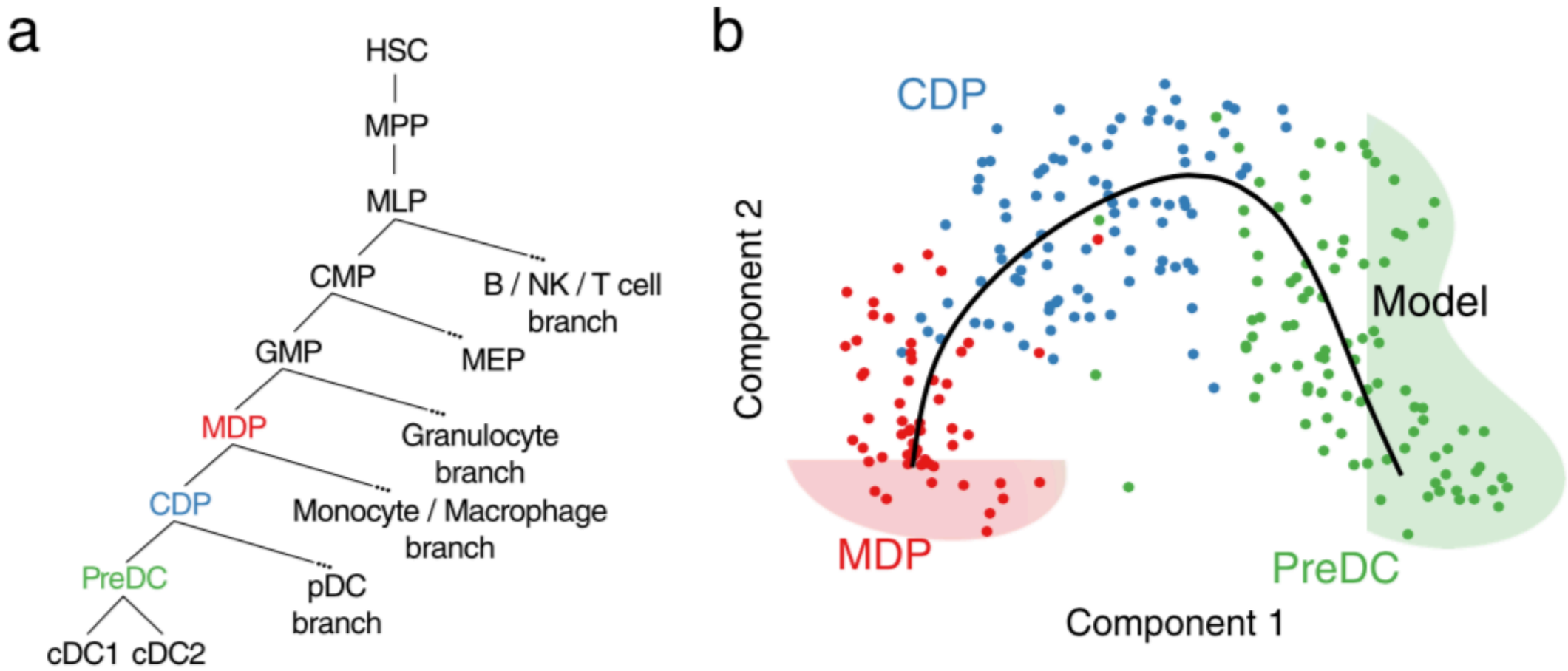
PCA on single-cell data typically captures technical effects, most often number of genes detected

First principal component explains 55% of variance, very strongly correlated with number of genes detected per cell.

PCA extremely useful for QC.

Produced with the R/Bioconductor package *scater* (McCarthy et al, *Bioinformatics*, 2017)

Trajectory analysis



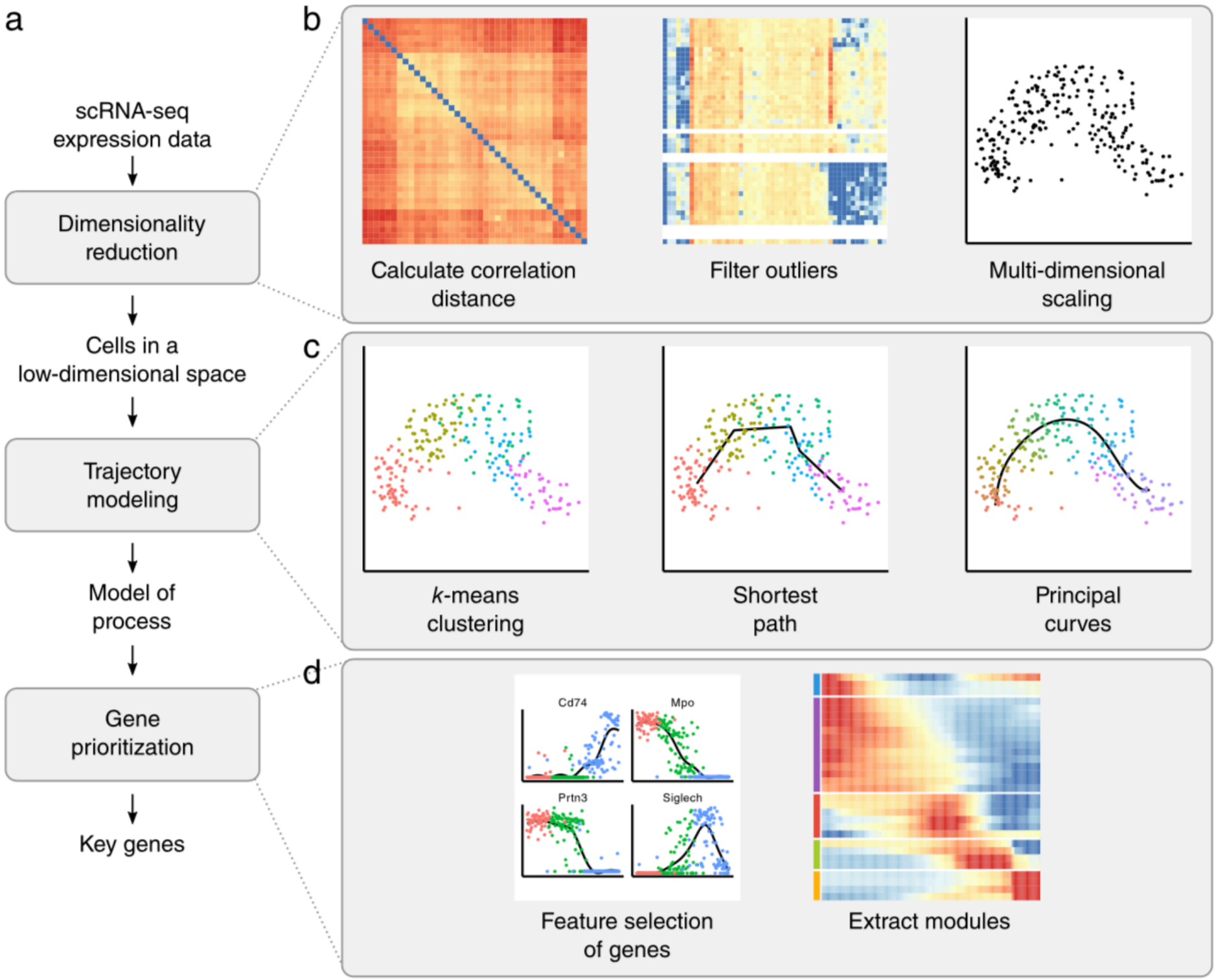
Tools:

monocle, SCORPIUS, TSCAN, SLICER, ouija, PhenoPath, GrandPrix (GPLVM)
+ many more

Cannoodt R, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, et al. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development [Internet]. bioRxiv. 2016. p. 079509. doi:10.1101/079509

EMBL-EBI

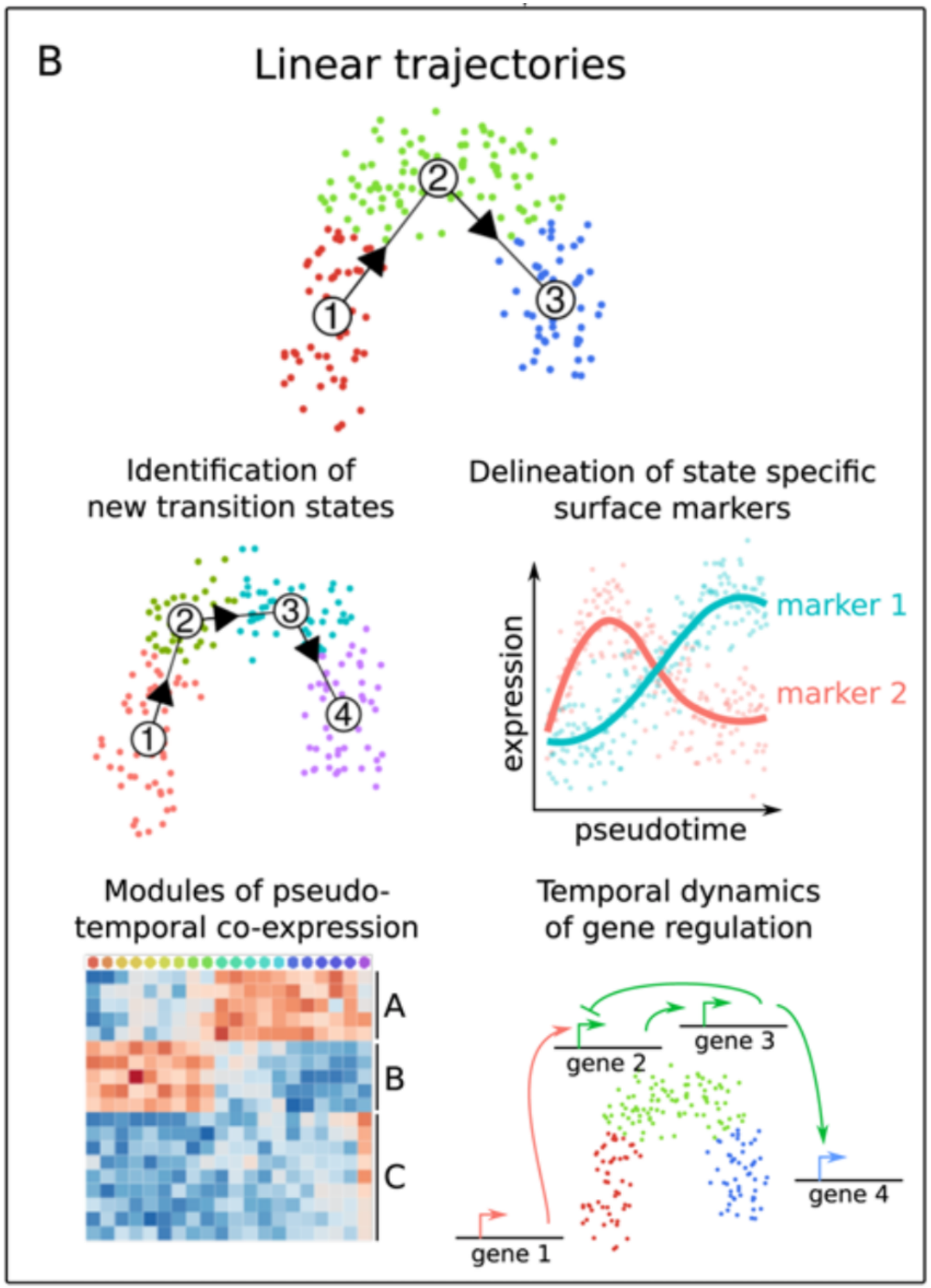
Trajectory analysis



Cannoodt R, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, et al. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development [Internet]. bioRxiv. 2016. p. 079509. doi:10.1101/079509

EMBL-EBI

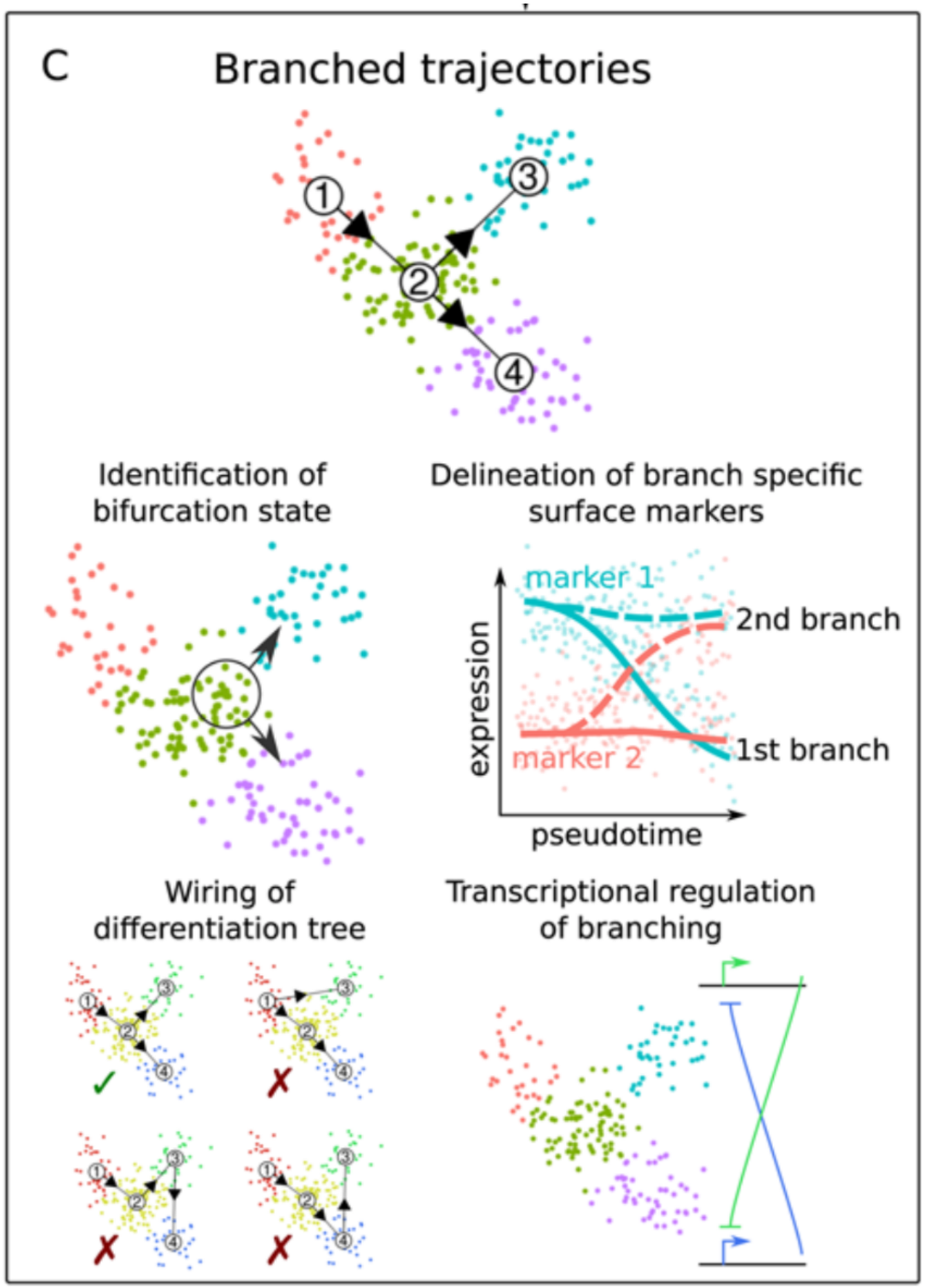
Linear trajectories



Cannoodt et al, *Eur. J Immunology,* 2016

EMBL-EBI

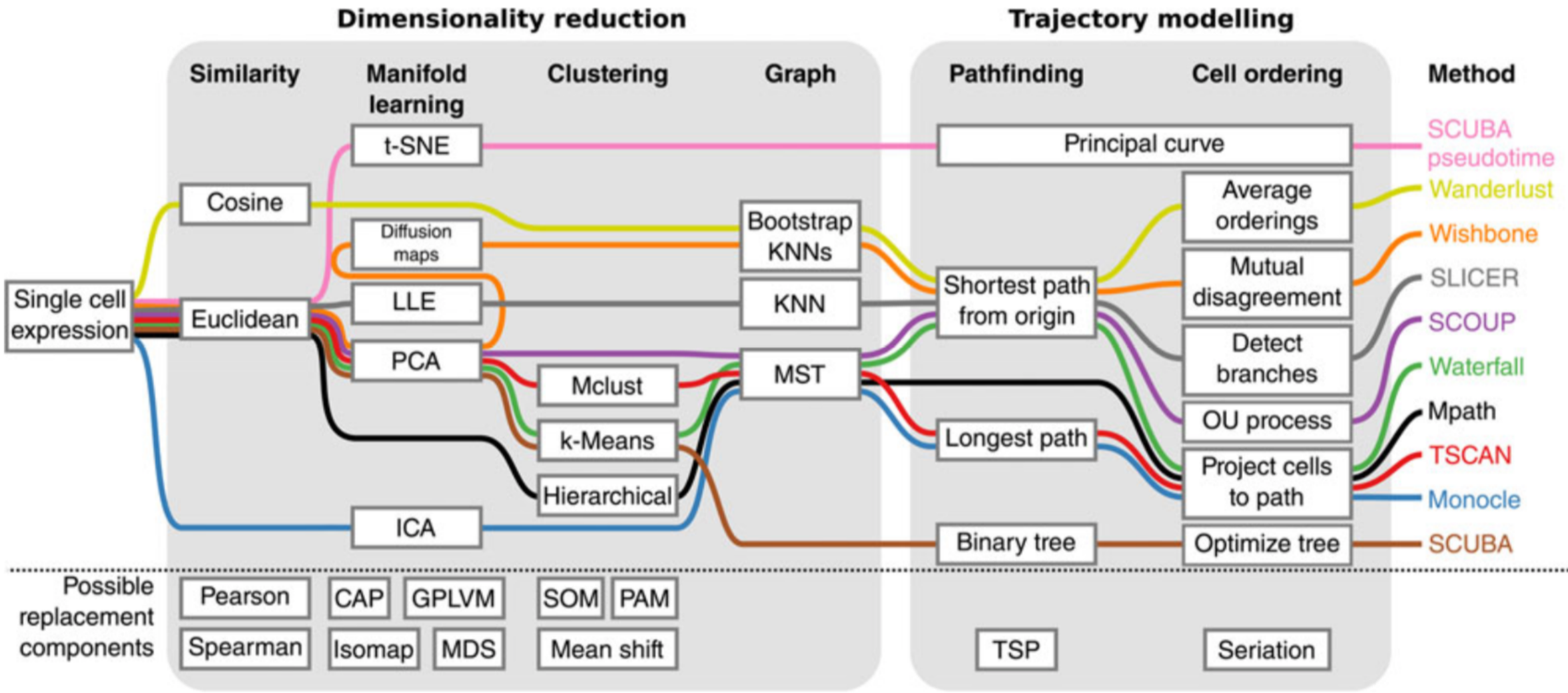
Branched trajectories



Cannoodt et al, *Eur. J Immunology,* 2016

EMBL-EBI

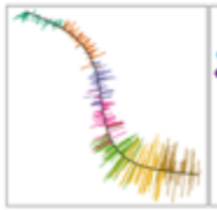
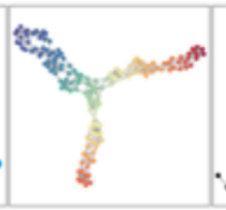
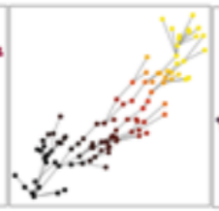
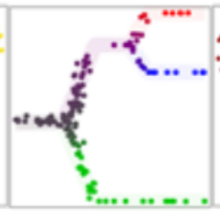
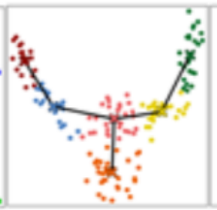
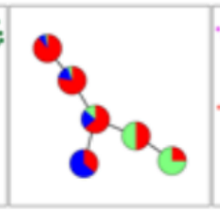
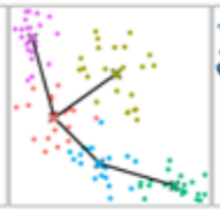
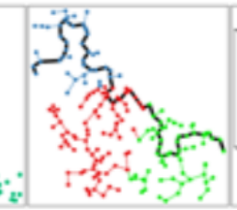
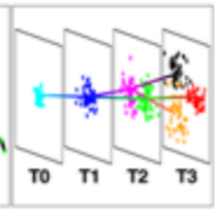
Generalised trajectory inference



Cannoodt et al, *Eur. J Immunology,* 2016

EMBL-EBI

Some trajectory inference methods

Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	−	+	−	−
Scalability w.r.t. cells	−	−	±	±	−	±	+	+	−	±
Scalability w.r.t. genes	+	+	+	+	−	+	±	±	±	+
Code and documentation	−	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	−	±	−	+	+	+

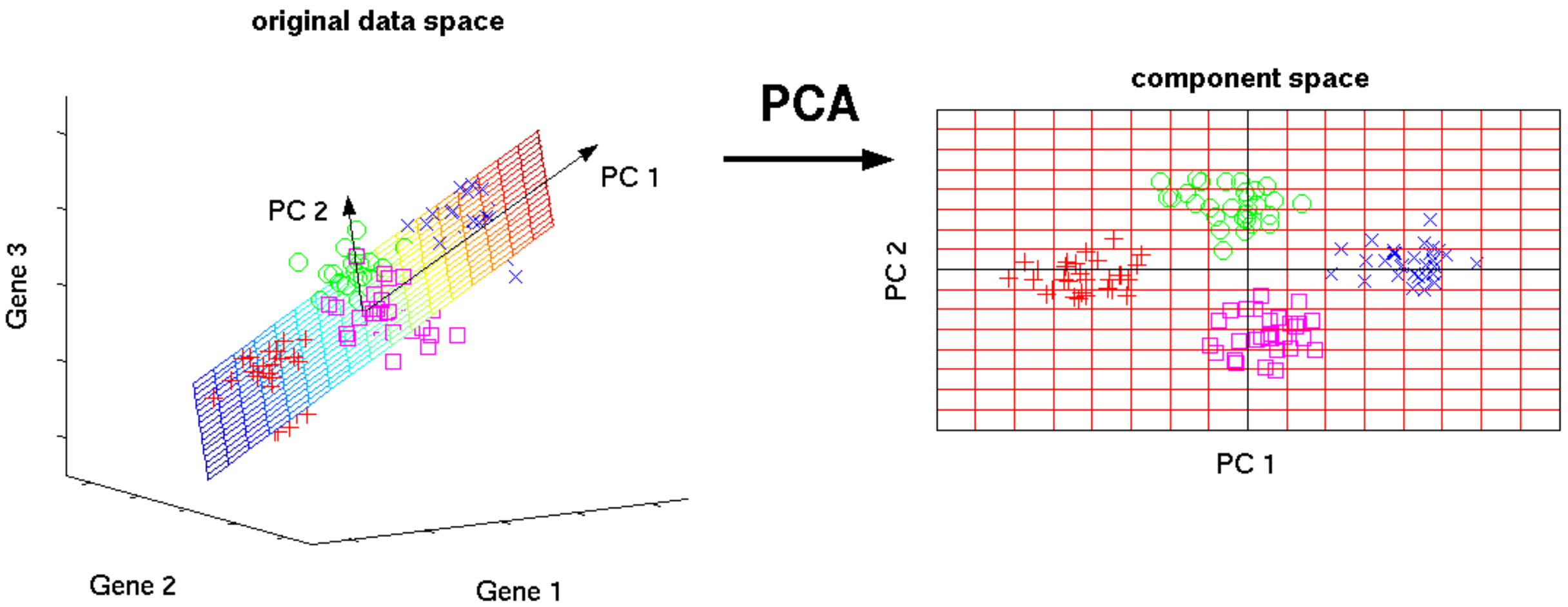
Cannoodt et al, *Eur. J Immunology*, 2016

EMBL-EBI

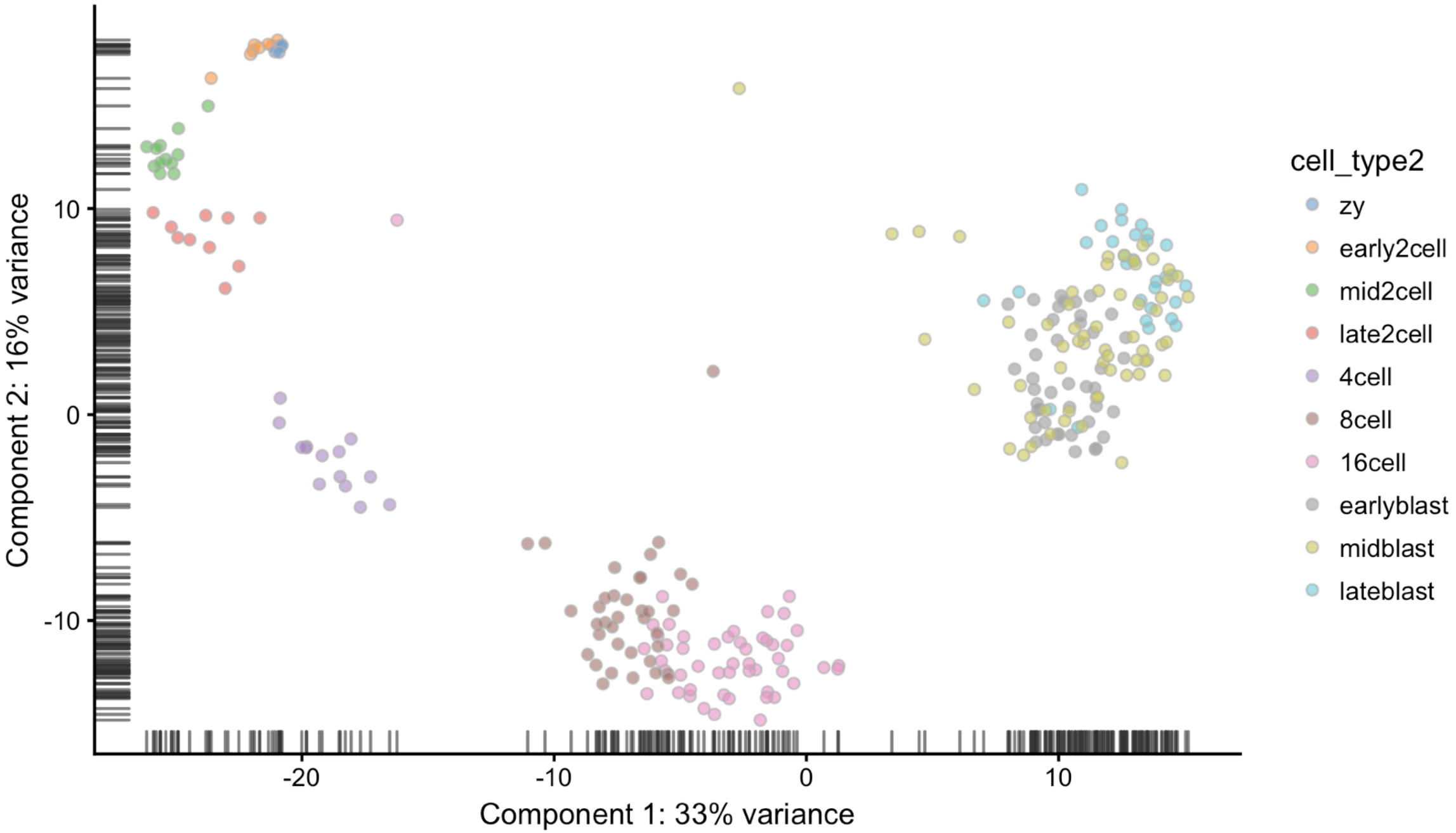
Trajectory inference methods we will use:

- Naive PCA
- Diffusion maps
- Monocle
- TSCAN
- SLICER
- Ouija

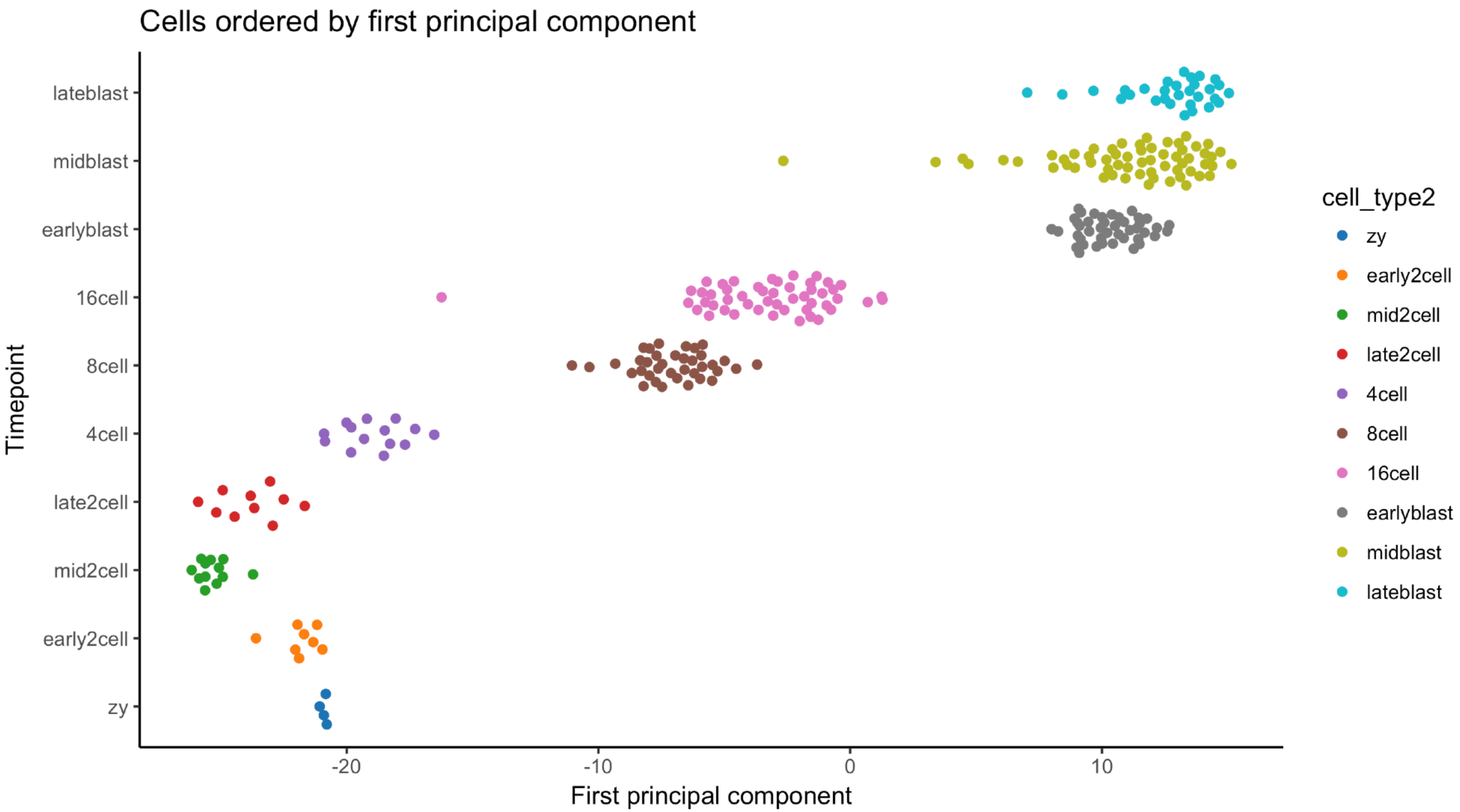
PCA for naive trajectory inference



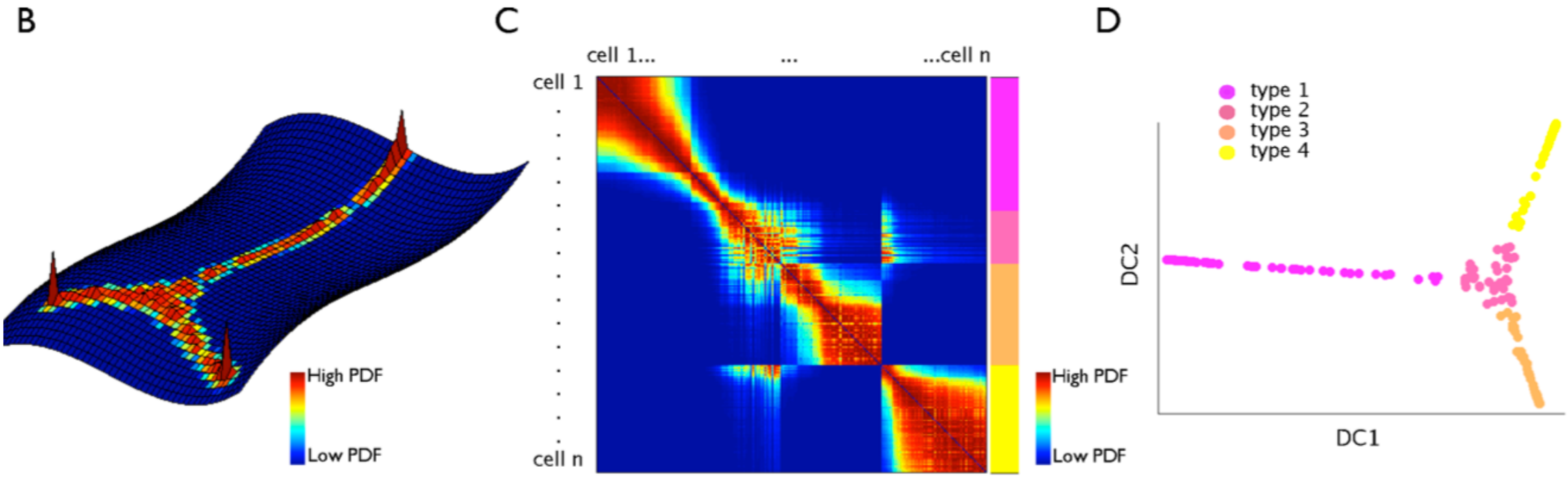
PCA for naive trajectory inference



PCA for naive trajectory inference

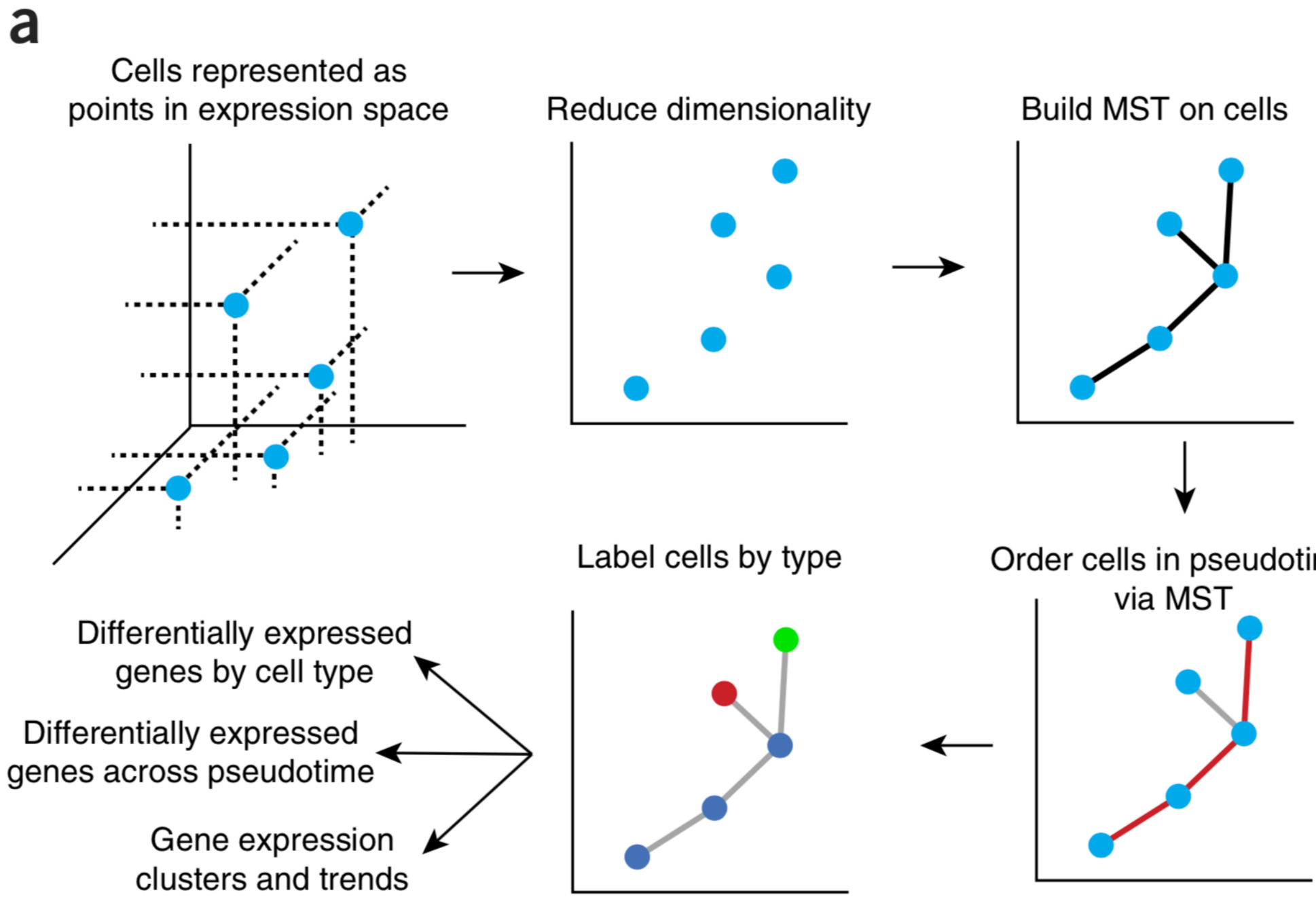


Diffusion Maps

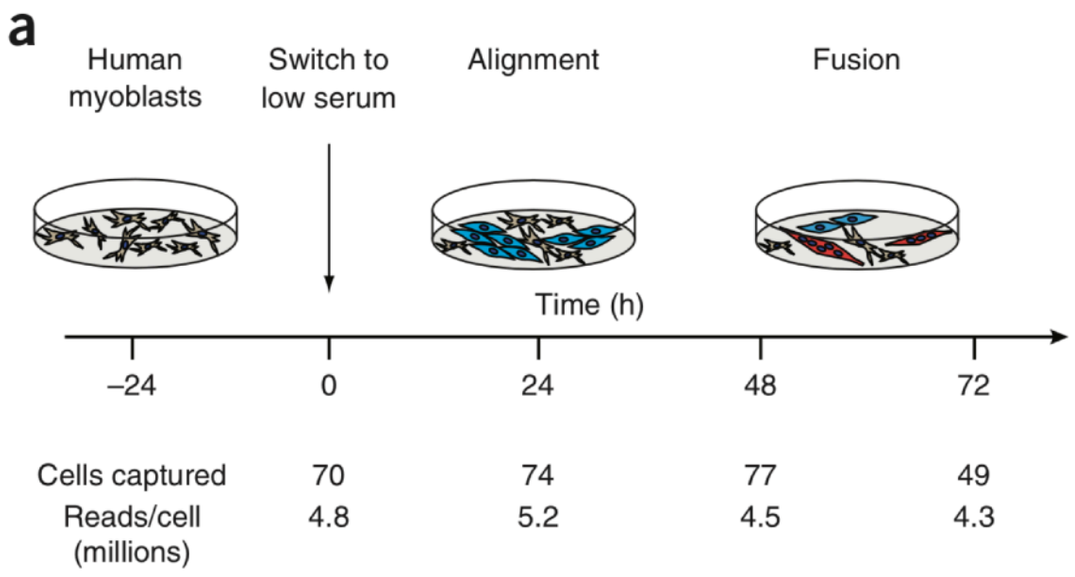


see also “dpt”: diffusion pseudotime

Monocle


a
Cells represented as
points in expression space
Reduce dimensionality
Build MST on cells
Order cells in pseudotime
via MST
Label cells by type
Differentially expressed
genes by cell type
Differentially expressed
genes across pseudotime
Gene expression
clusters and trends

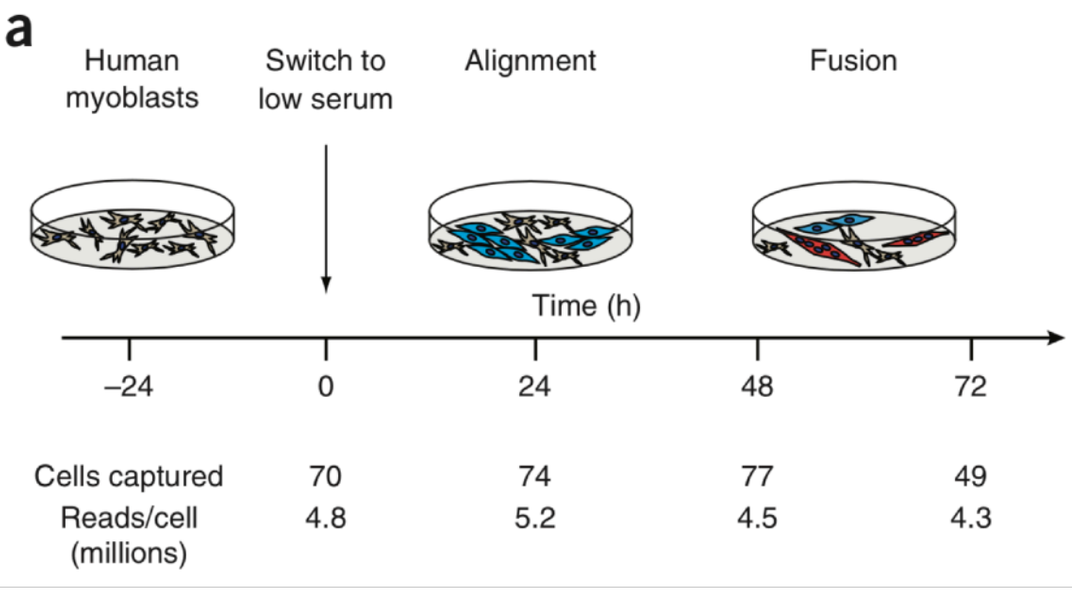
Monocle

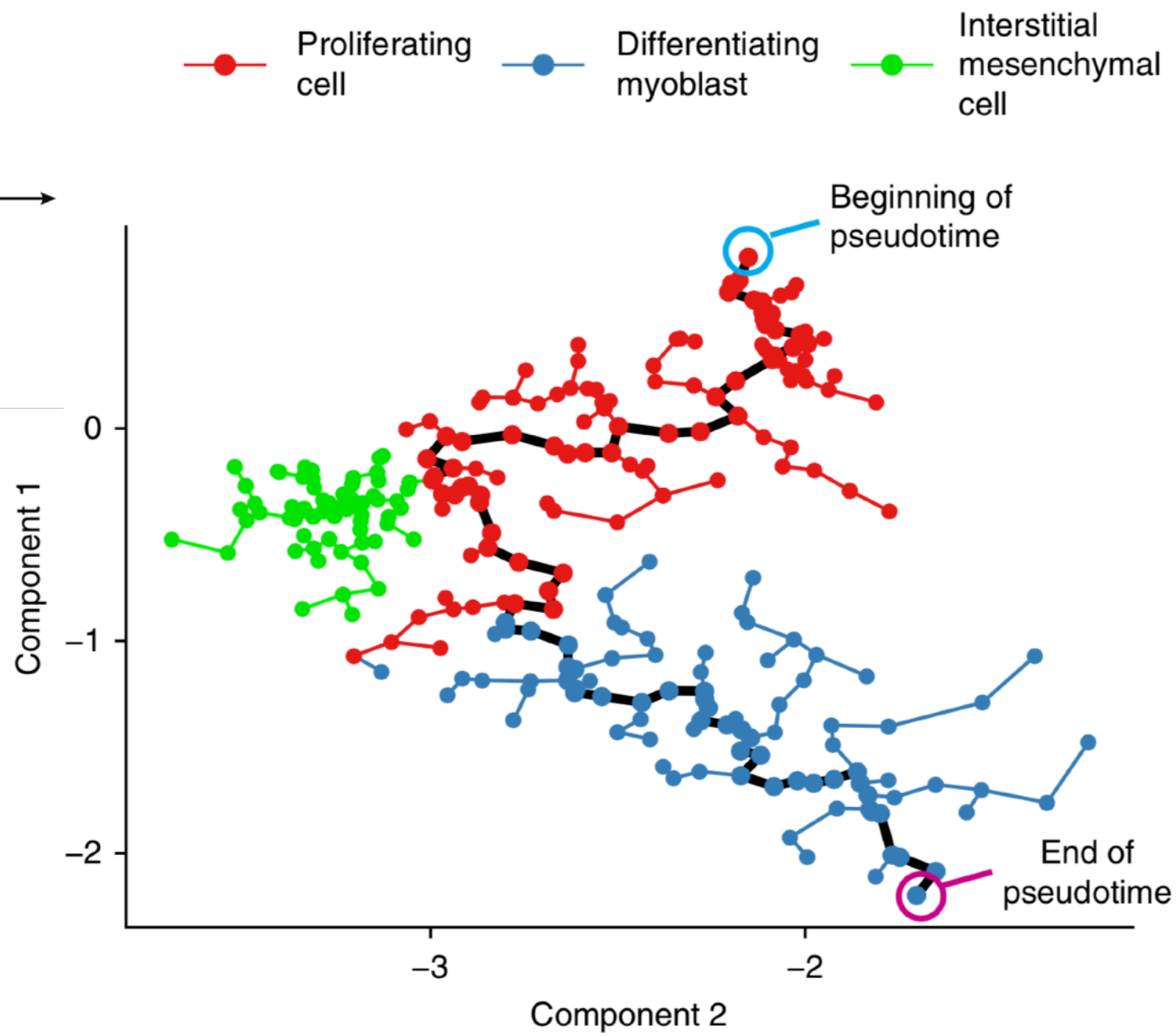


Trapnell et al, *Nature Biotech.*, 2014

EMBL-EBI

Monocle

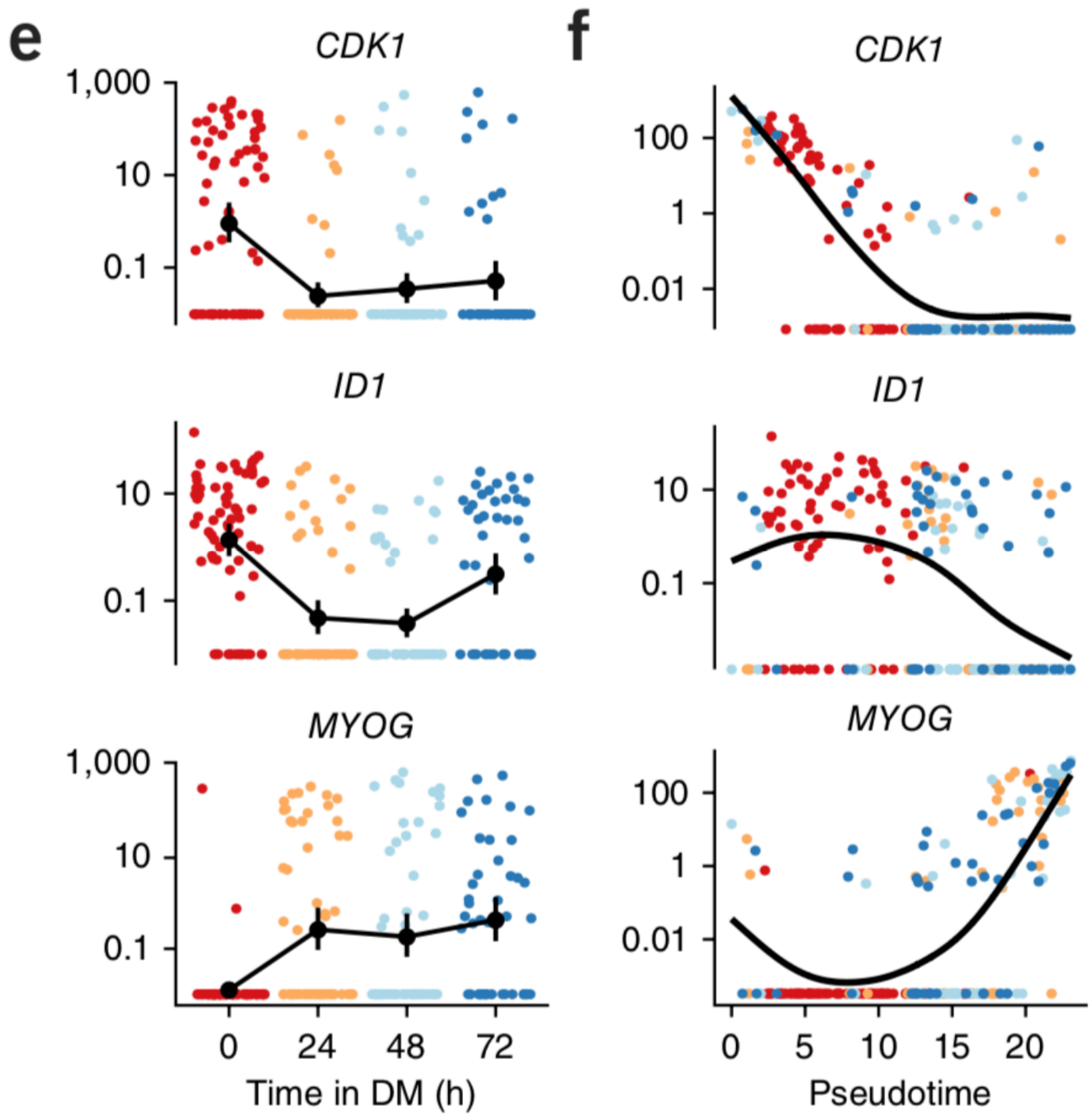




Trapnell et al, *Nature Biotech.*, 2014

EMBL-EBI

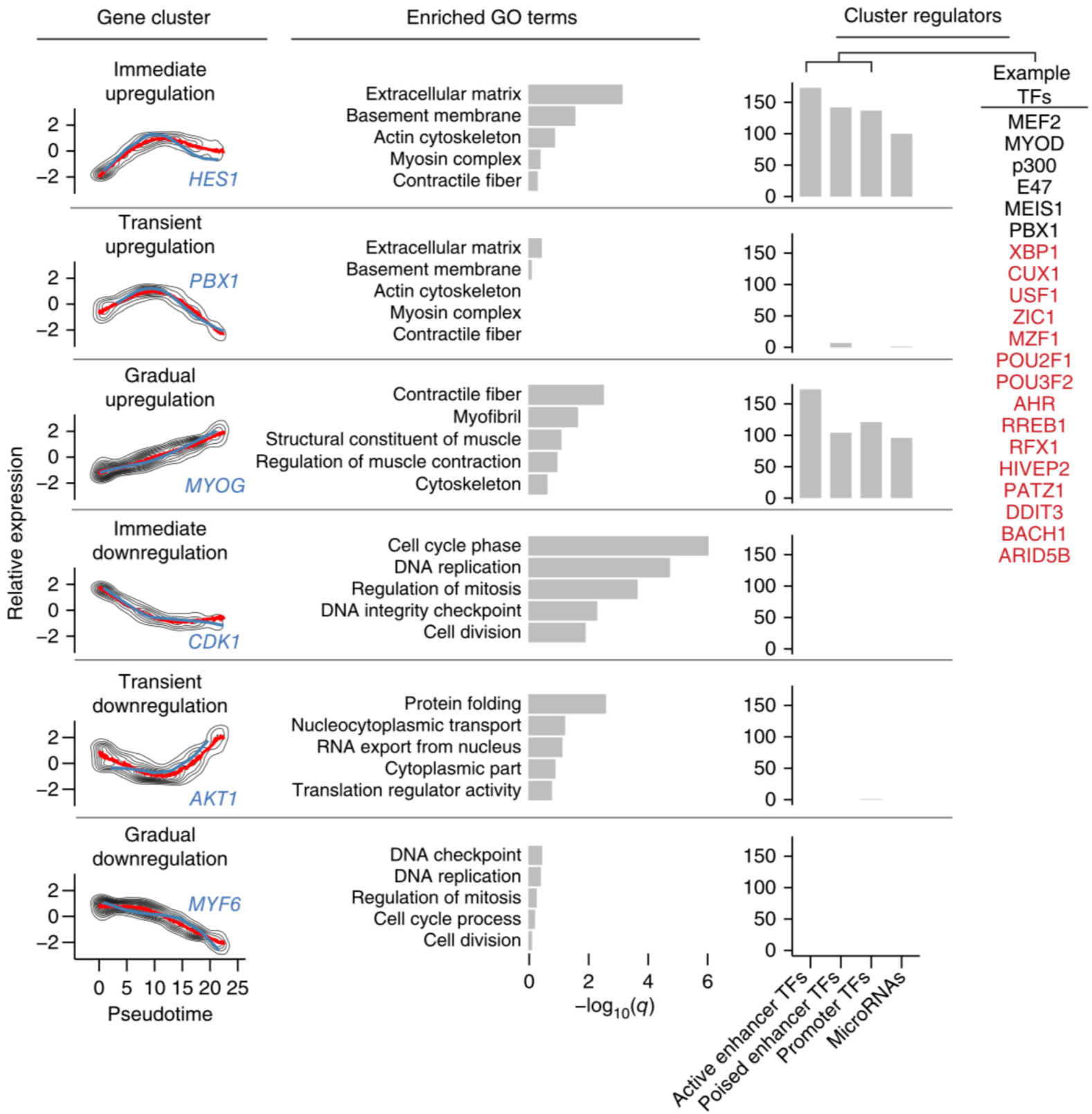
Monocle



Trapnell et al, *Nature Biotech.*, 2014

EMBL-EBI

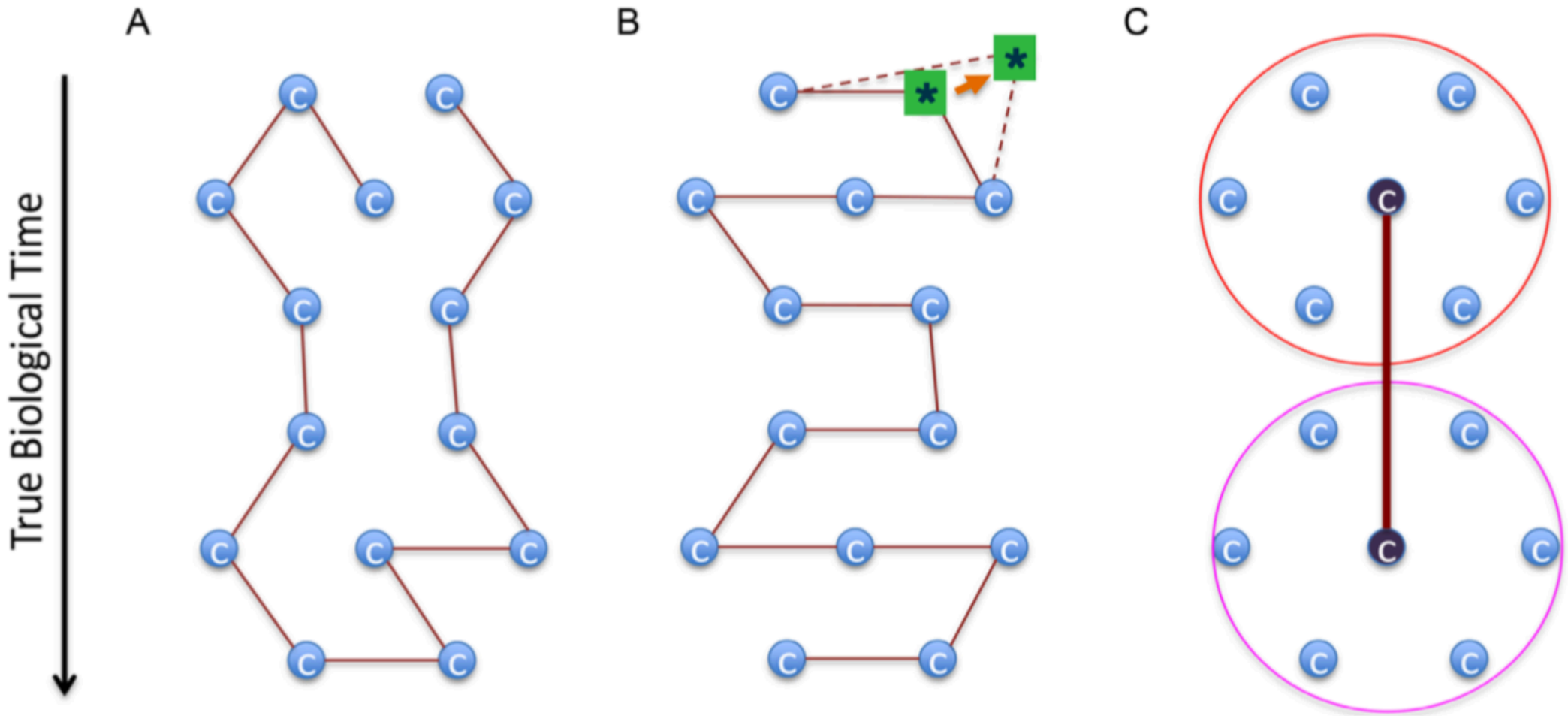
Monocle



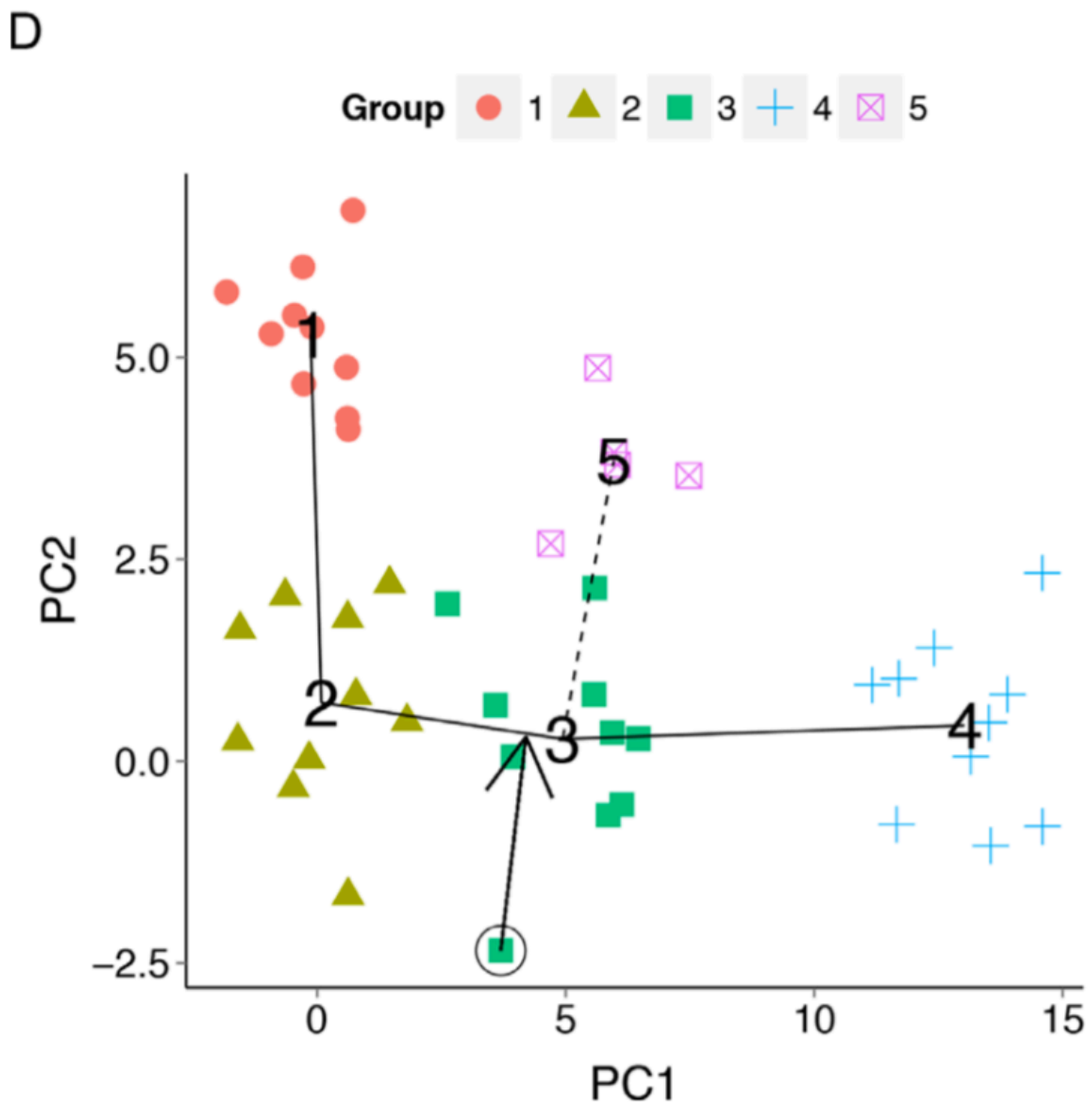
Trapnell et al, *Nature Biotech.*, 2014

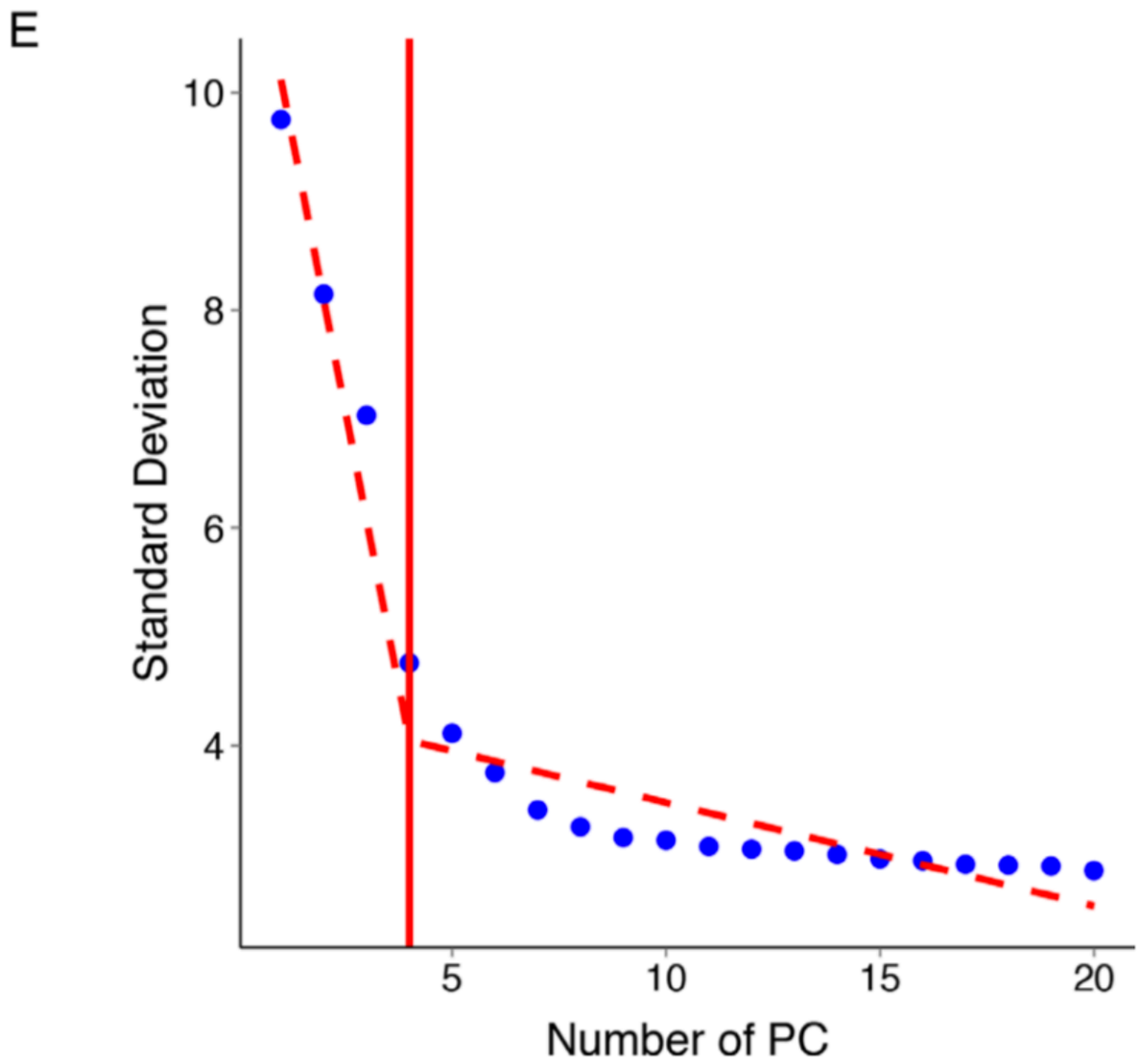
EMBL-EBI

TSCAN

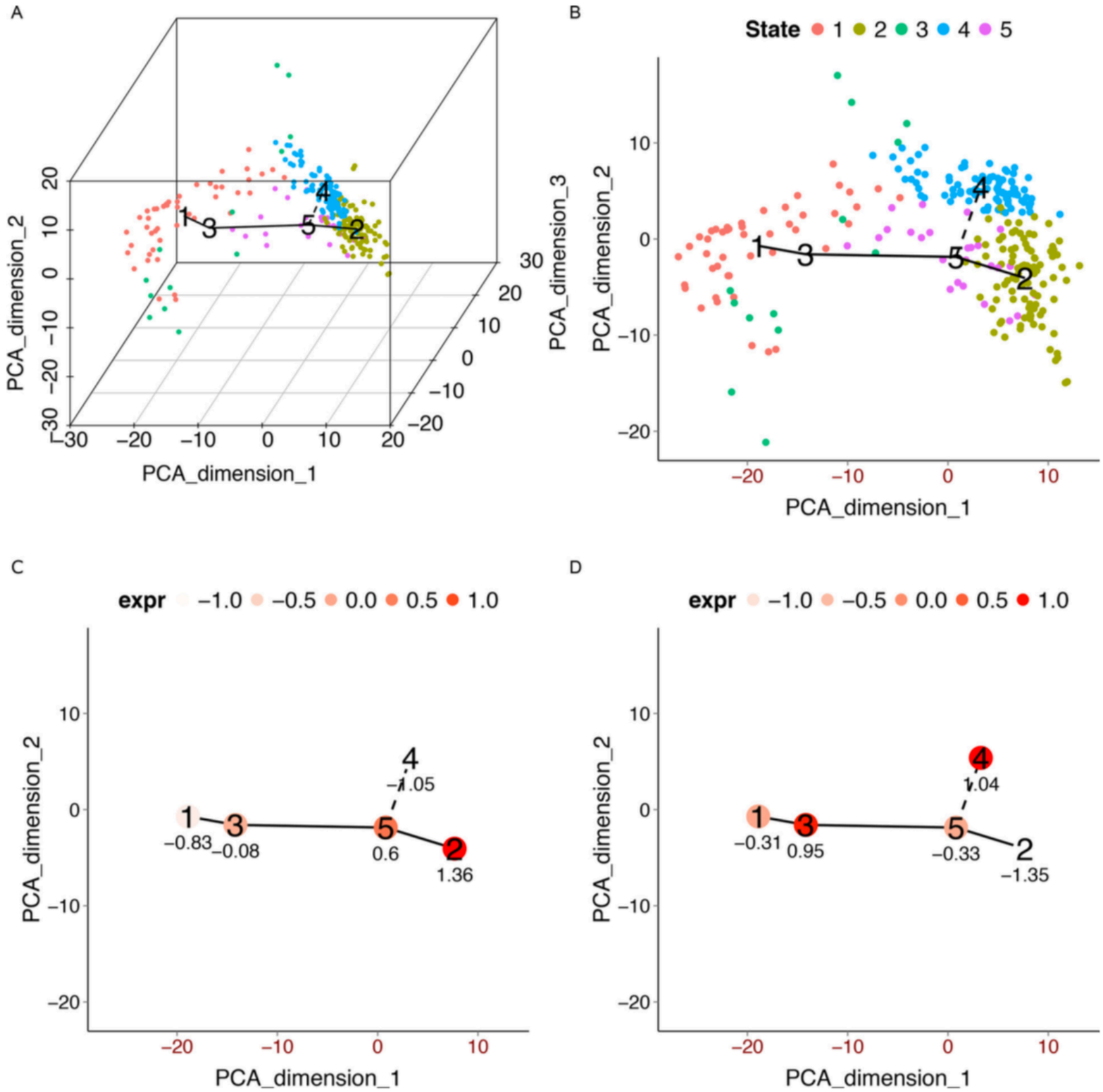


TSCAN

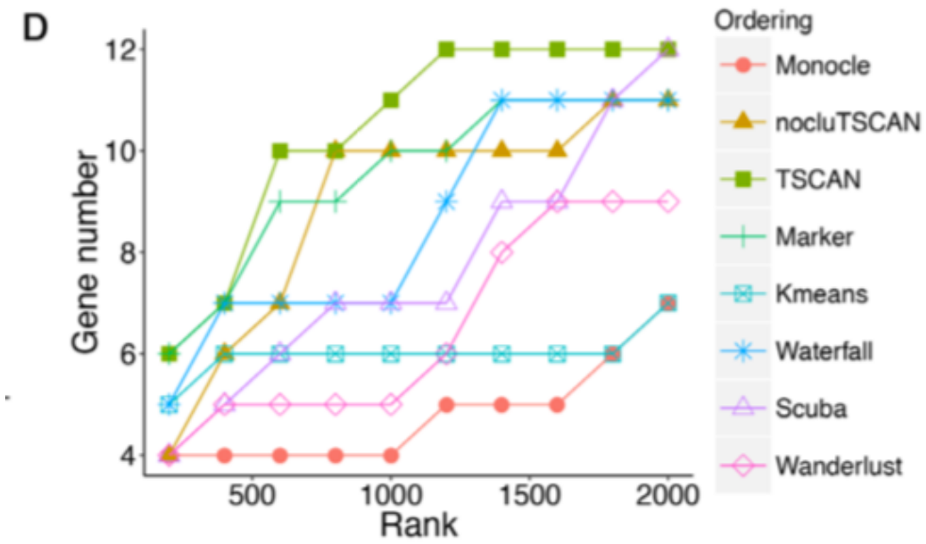

D
Group 1 2 3 4 5
PC2
PC1
5.0
2.5
0.0
−2.5
0
5
10
15
1
2
3
4
5


E
Standard Deviation
Number of PC
10
8
6
4
5
10
15
20

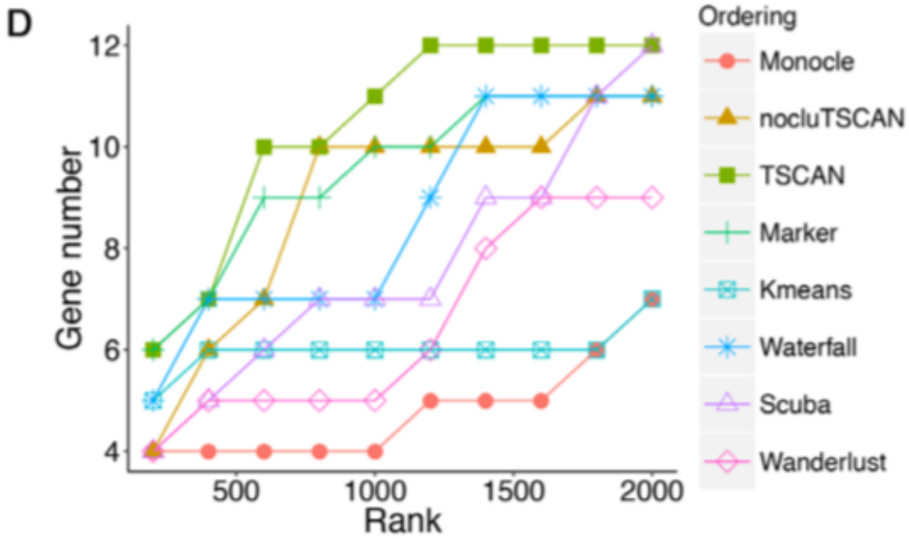
TSCAN

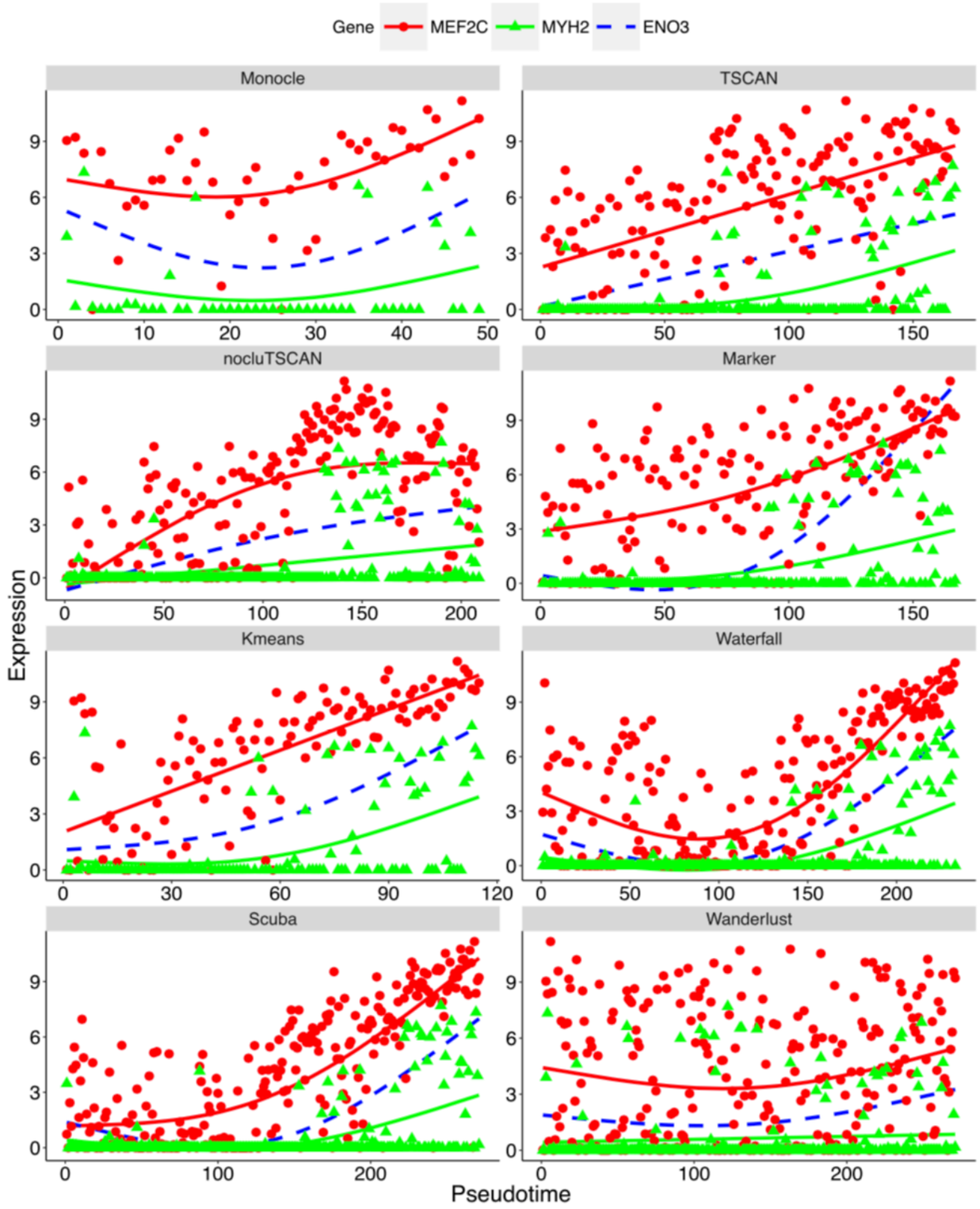

A
PCA_dimension_2
PCA_dimension_1
B
State 1 2 3 4 5
PCA_dimension_3
PCA_dimension_2
PCA_dimension_1
C
expr −1.0 −0.5 0.0 0.5 1.0
PCA_dimension_2
PCA_dimension_1
−0.83
−0.08
0.6
1.36
−1.05
D
expr −1.0 −0.5 0.0 0.5 1.0
PCA_dimension_2
PCA_dimension_1
−0.31
0.95
−0.33
−1.35
1.04

TSCAN

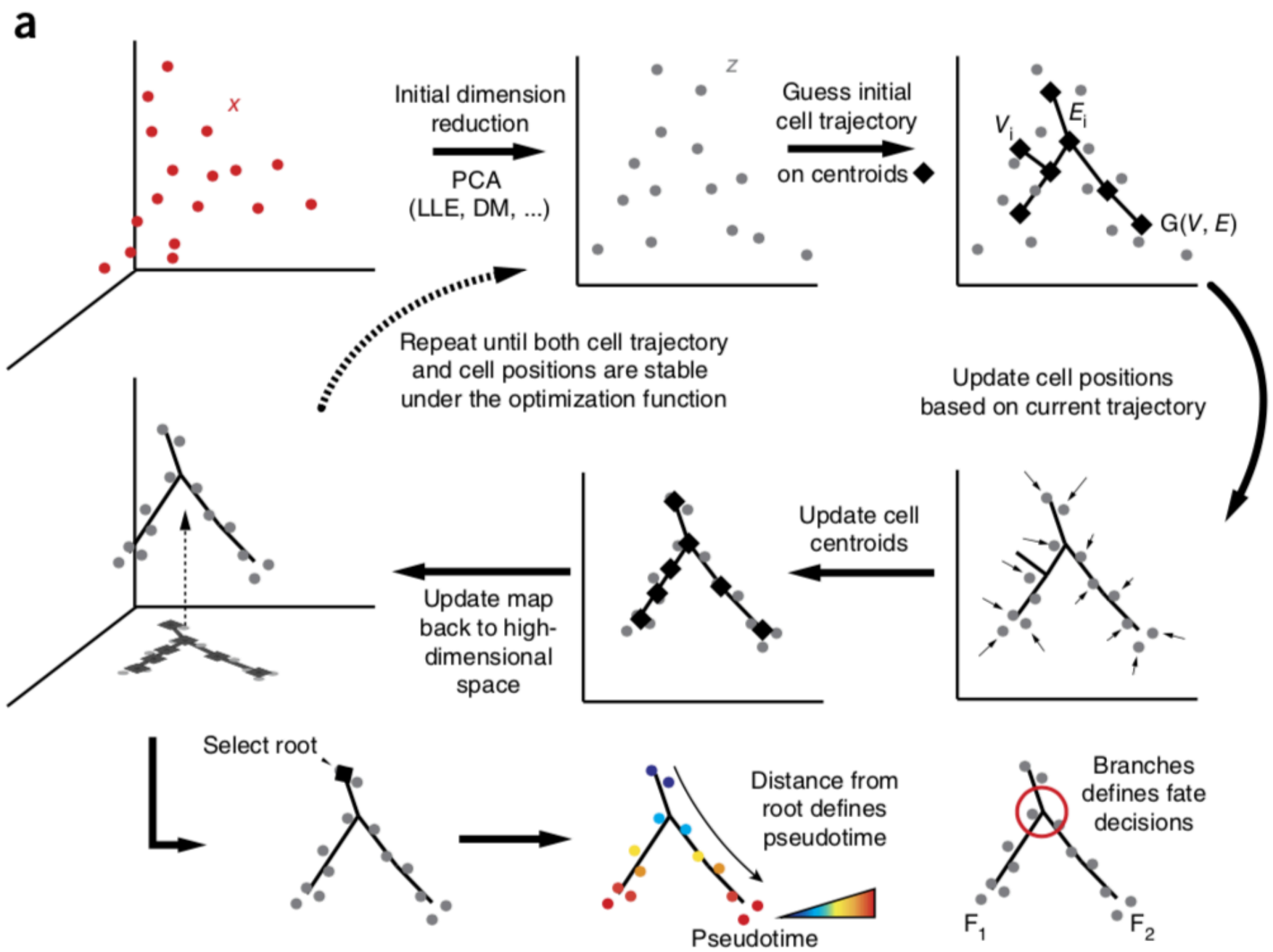

D
Gene number
12
10
8
6
4
500
1000
1500
2000
Rank
Ordering
Monocle
nocluTSCAN
TSCAN
Marker
Kmeans
Waterfall
Scuba
Wanderlust

TSCAN


D
Gene number
Rank
Ordering
Monocle
nocluTSCAN
TSCAN
Marker
Kmeans
Waterfall
Scuba
Wanderlust


Gene
MEF2C
MYH2
ENO3
Monocle
TSCAN
nocluTSCAN
Marker
Kmeans
Waterfall
Scuba
Wanderlust
Expression
Pseudotime

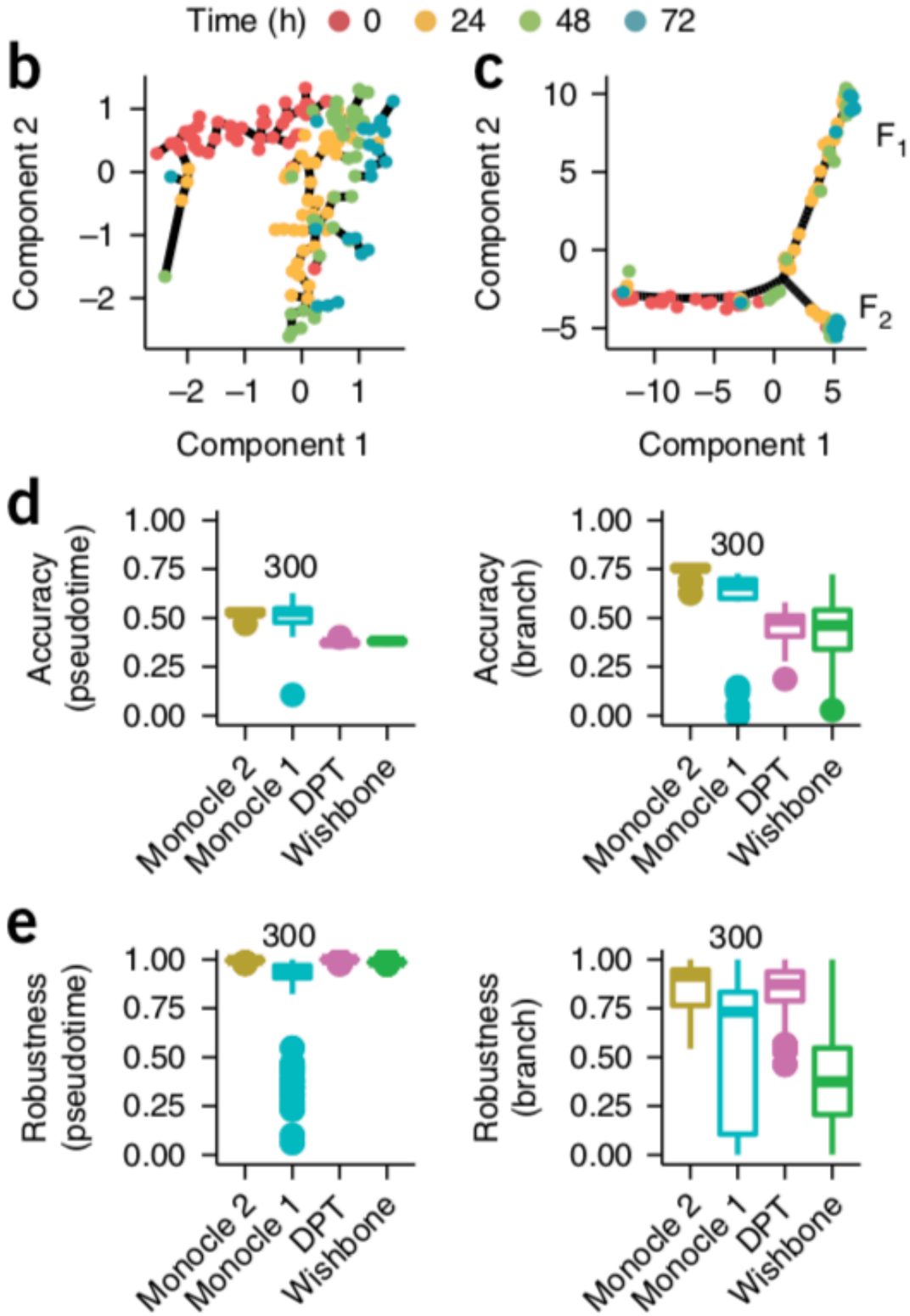
Monocle2



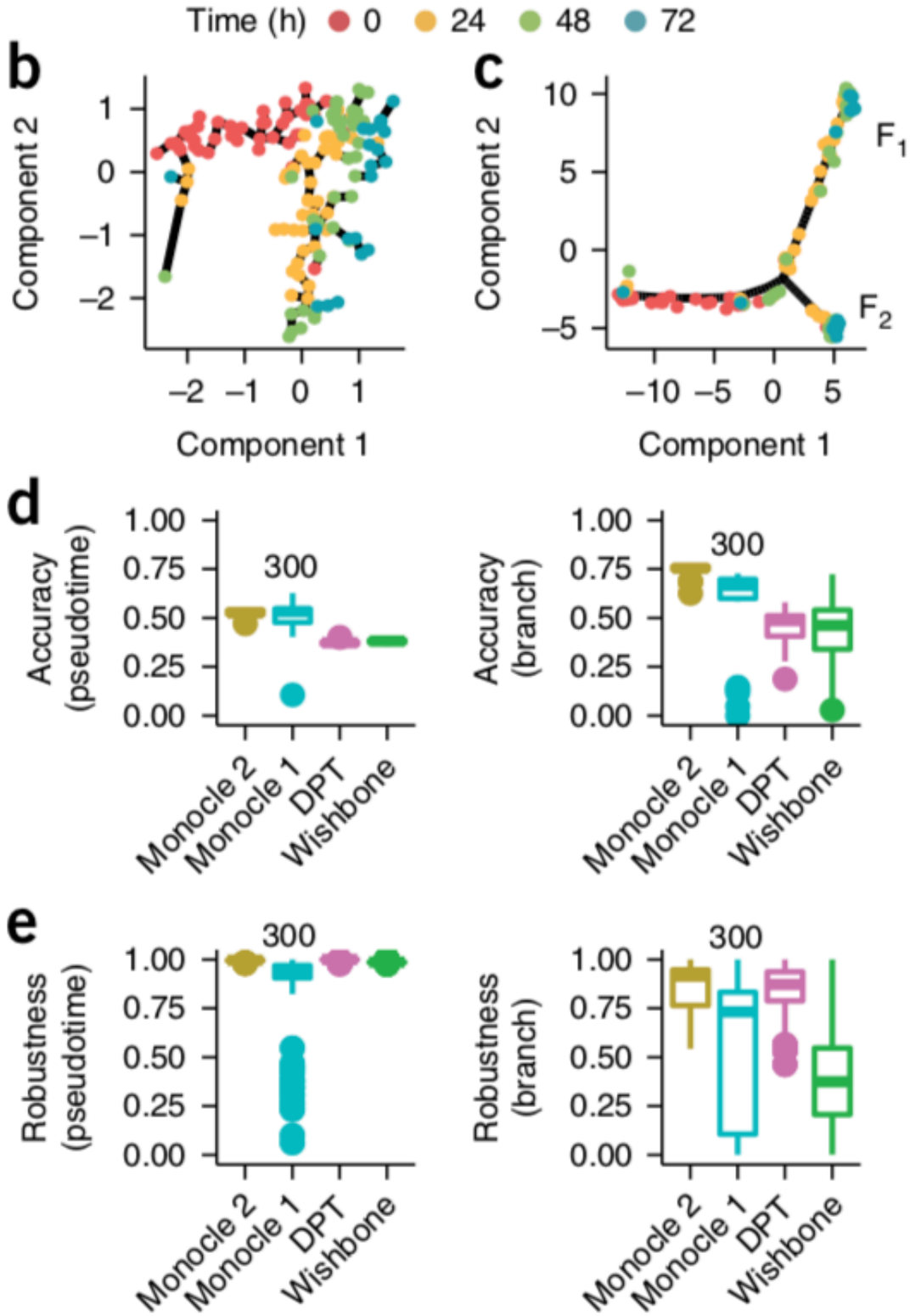
Qiu et al, *Nature Methods*, 2017

EMBL-EBI

Monocle2


Time (h) 0 24 48 72
b
Component 2
Component 1
c
Component 2
Component 1
F1
F2
d
Accuracy (pseudotime)
Accuracy (branch)
300
Monocle 2
Monocle 1
DPT
Wishbone
e
Robustness (pseudotime)
Robustness (branch)

Monocle2


Time (h) 0 24 48 72
b
Component 2
Component 1
c
Component 2
Component 1
F1
F2
d
Accuracy (pseudotime)
Accuracy (branch)
300
Monocle 2
Monocle 1
DPT
Wishbone
e
Robustness (pseudotime)
Robustness (branch)

SLICER

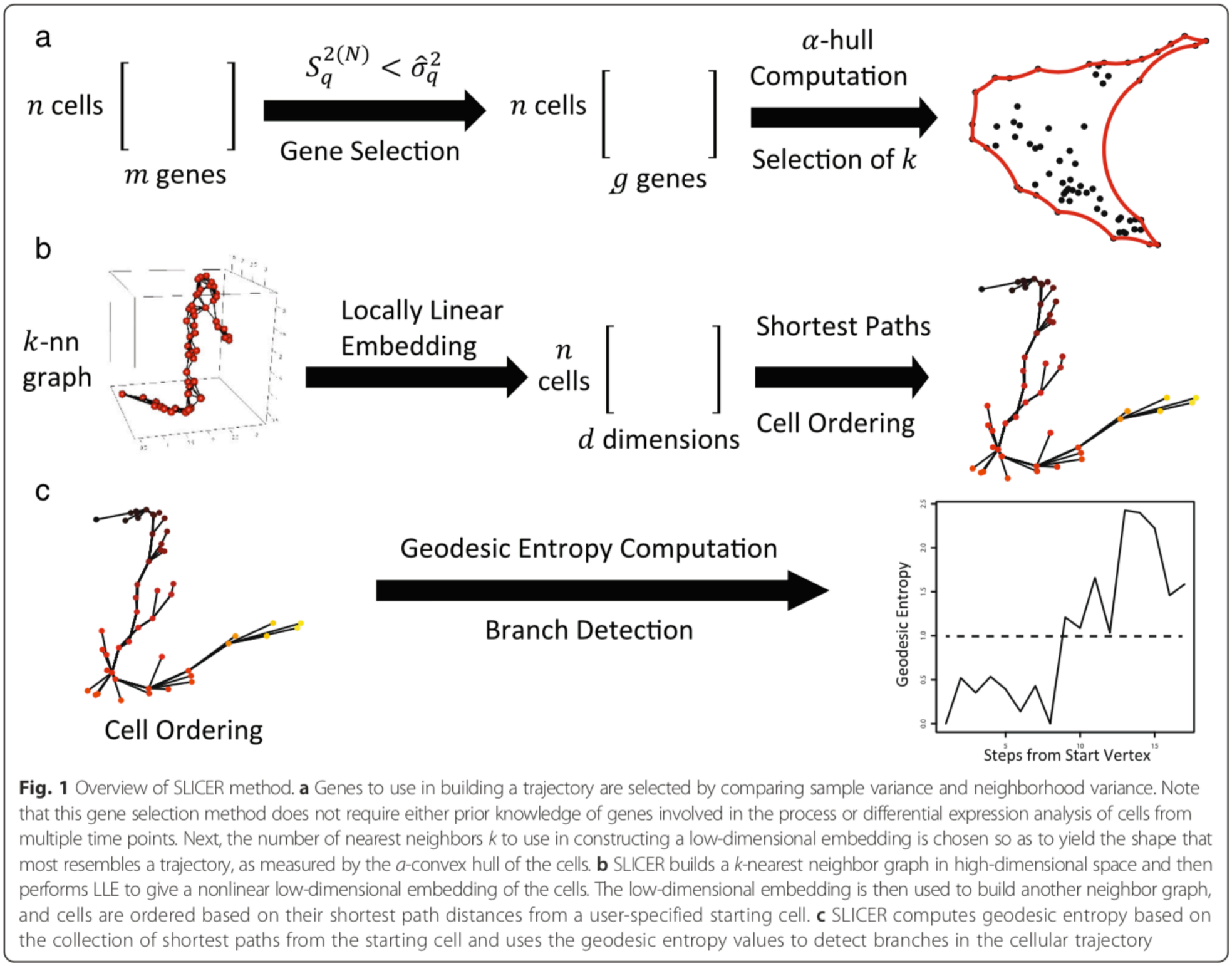
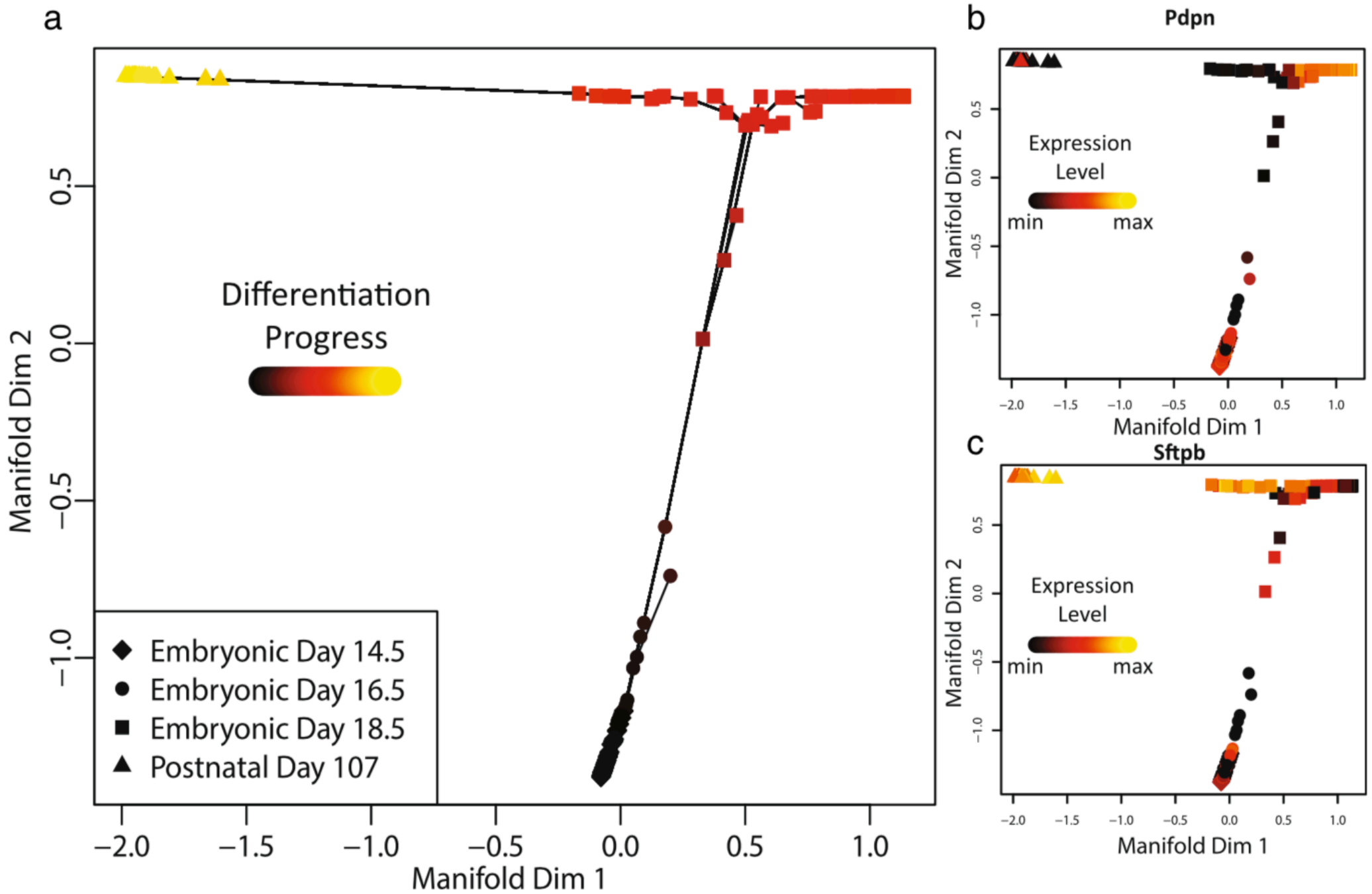


Fig. 1 Overview of SLICER method. **a** Genes to use in building a trajectory are selected by comparing sample variance and neighborhood variance. Note that this gene selection method does not require either prior knowledge of genes involved in the process or differential expression analysis of cells from multiple time points. Next, the number of nearest neighbors *k* to use in constructing a low-dimensional embedding is chosen so as to yield the shape that most resembles a trajectory, as measured by the *a*-convex hull of the cells. **b** SLICER builds a *k*-nearest neighbor graph in high-dimensional space and then performs LLE to give a nonlinear low-dimensional embedding of the cells. The low-dimensional embedding is then used to build another neighbor graph, and cells are ordered based on their shortest path distances from a user-specified starting cell. **c** SLICER computes geodesic entropy based on the collection of shortest paths from the starting cell and uses the geodesic entropy values to detect branches in the cellular trajectory

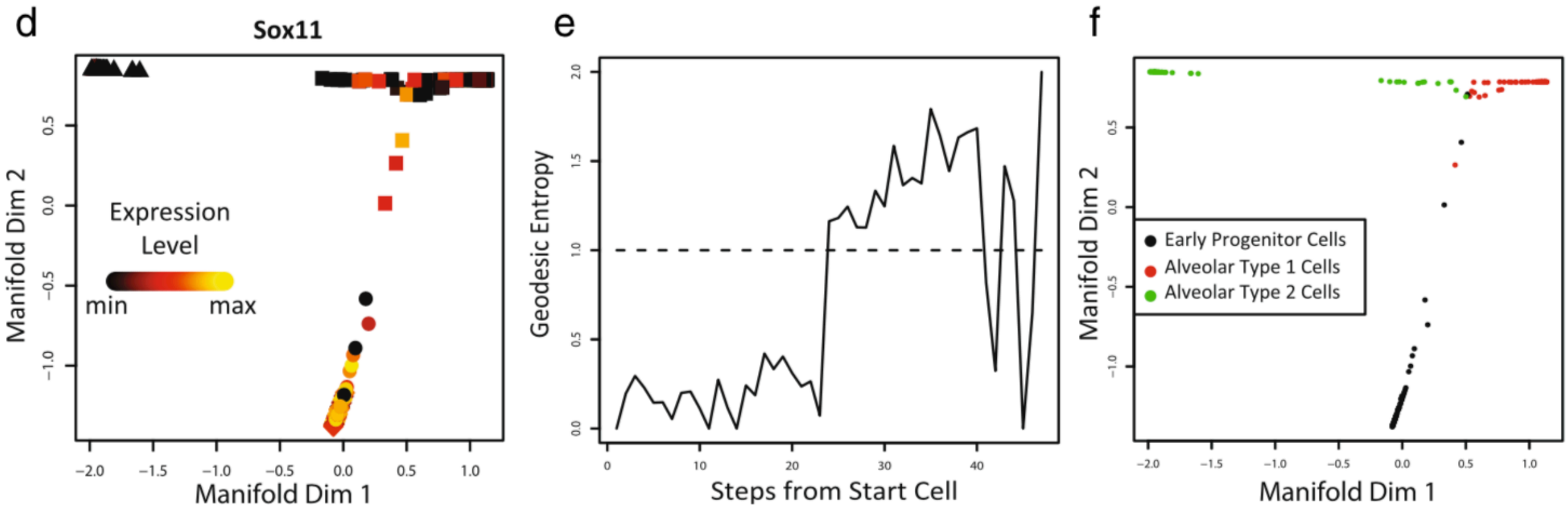
SLICER



Welch et al, *Genome Biology*, 2017

EMBL-EBI

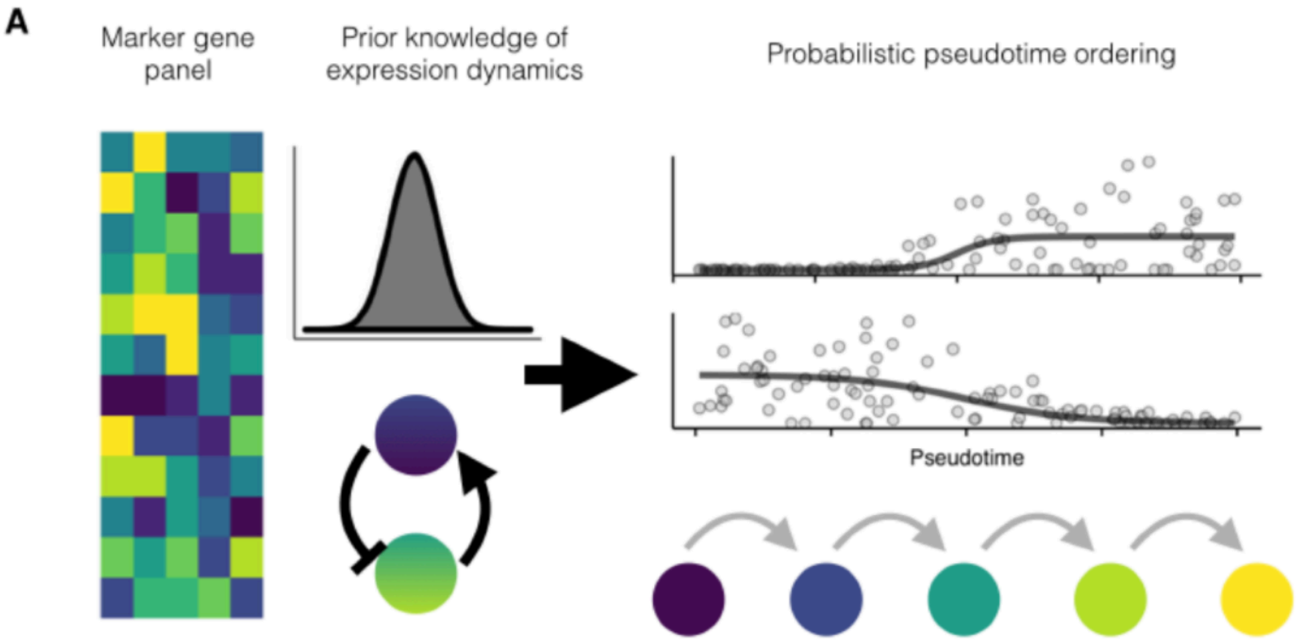
SLICER

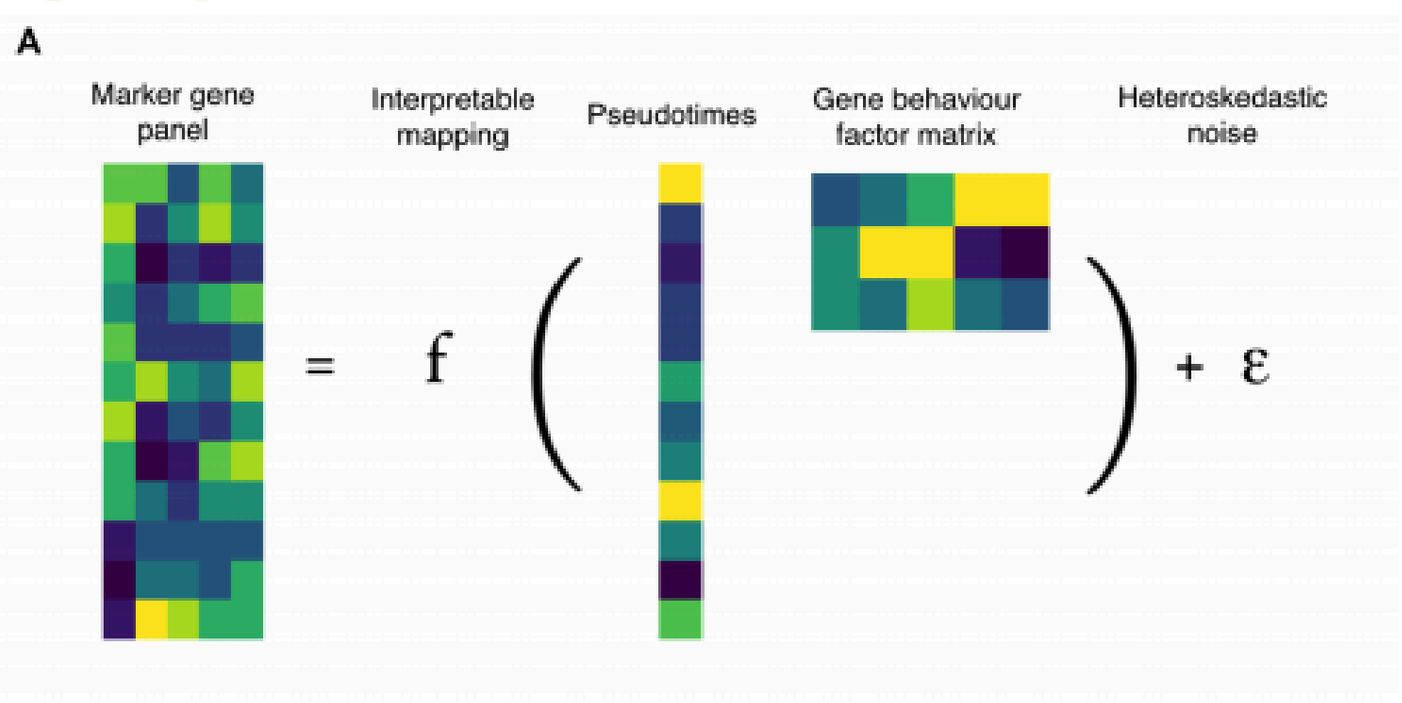


Welch et al, *Genome Biology*, 2017

EMBL-EBI

Ouija: Using knowledge of marker genes to guide pseudo time inference

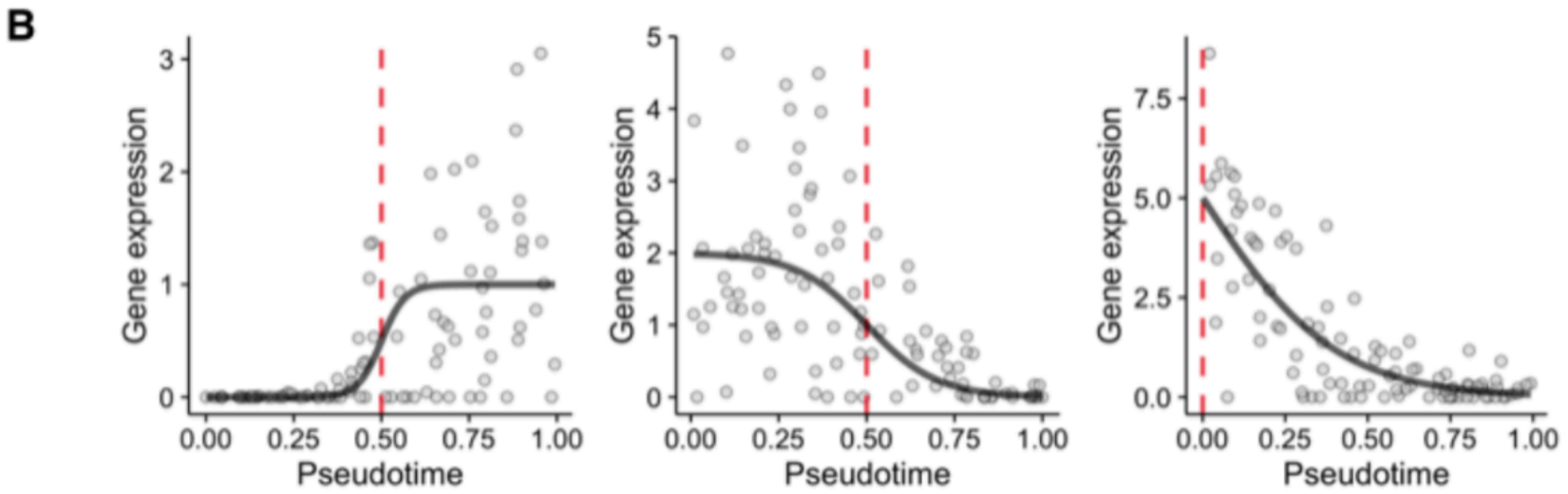




Campbell K, Yau C. Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis. bioRxiv. 2016; 060442.

EMBL-EBI

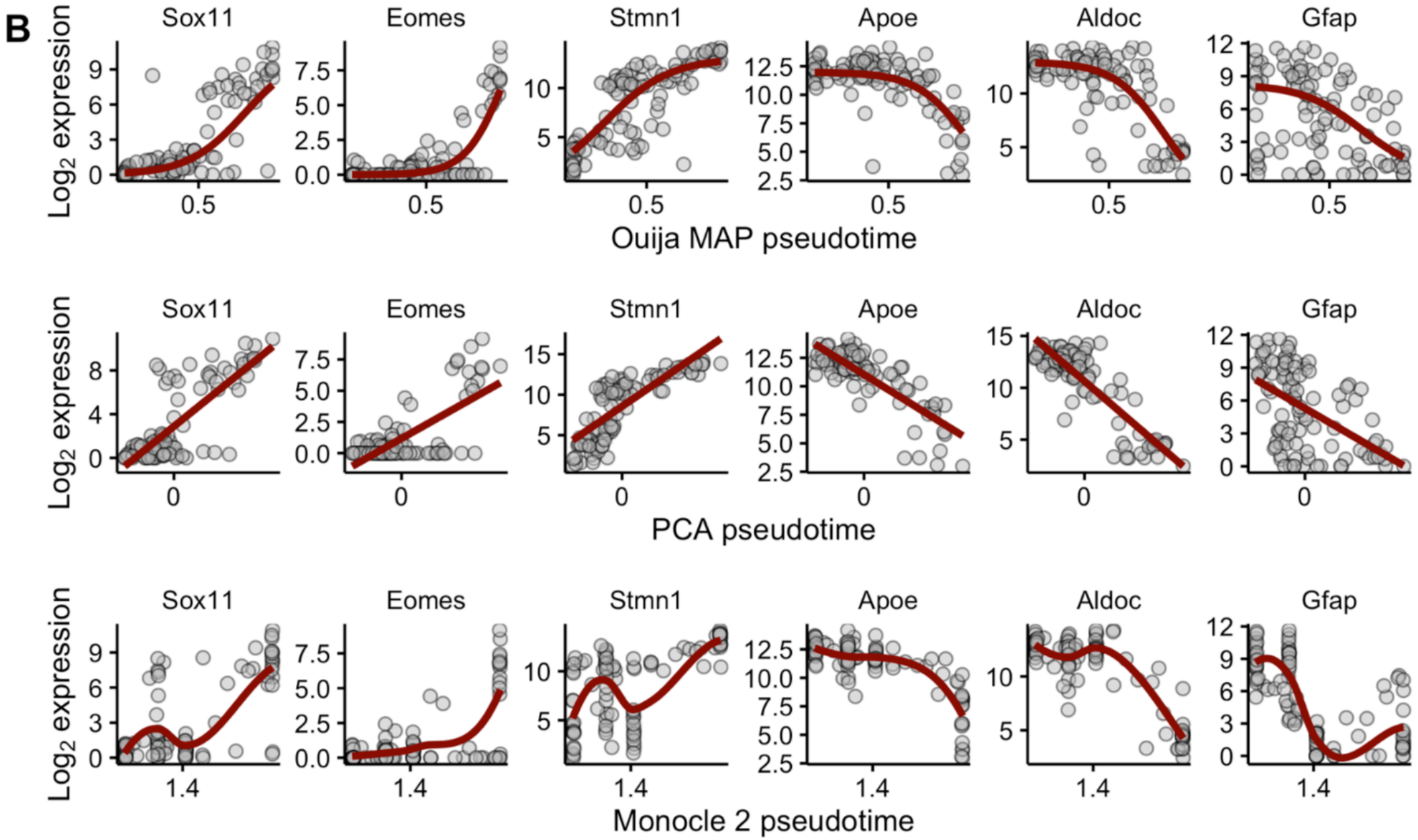
Ouija: Using knowledge of marker genes to guide pseudo time inference



Campbell K, Yau C. Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis. bioRxiv. 2016; 060442.

EMBL-EBI

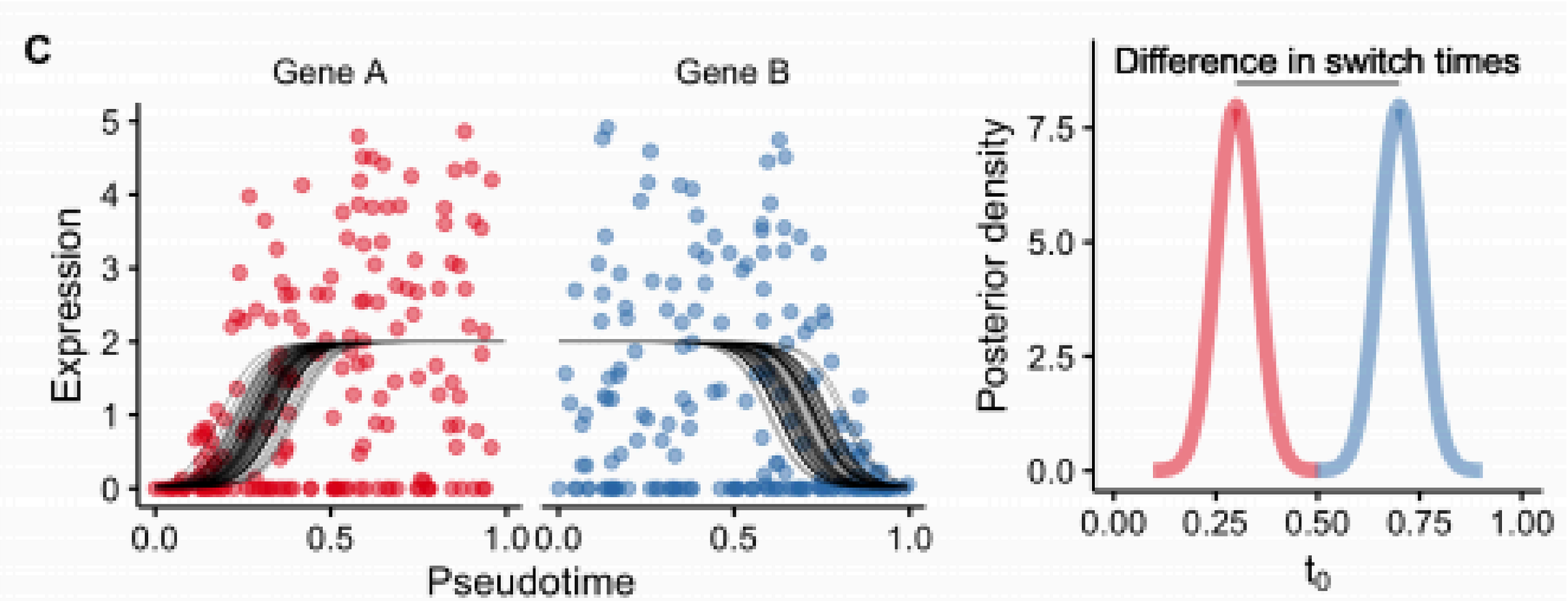
Ouija vs PCA vs Monocle on data from neurogenesis in mice



Campbell K, Yau C. Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis. bioRxiv. 2016; 060442.

EMBL-EBI

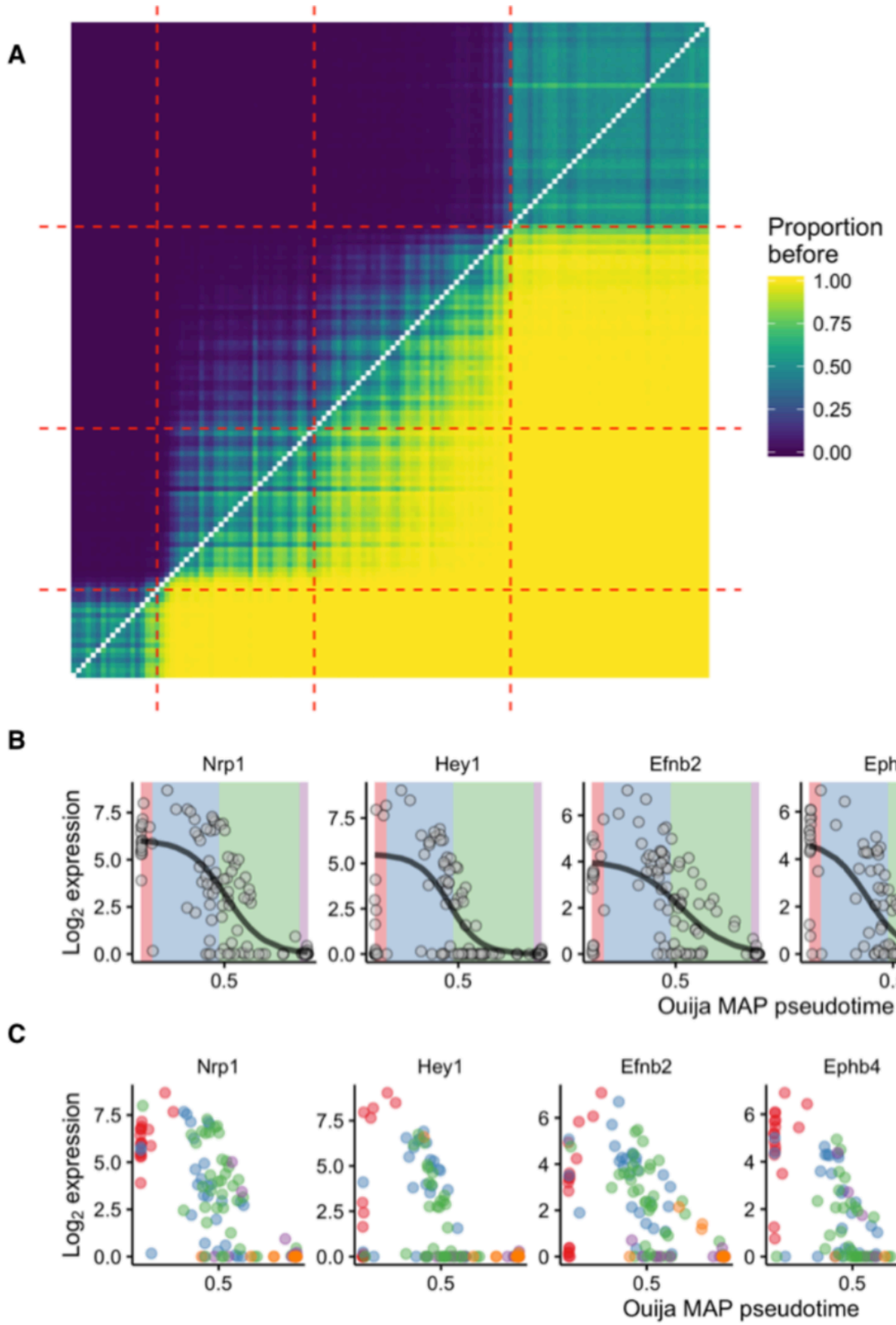
Ouija: difference in switch times (gene regulation)



Campbell K, Yau C. Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis. bioRxiv. 2016; 060442.

EMBL-EBI

Ouija: metastable states



D

Ouija cell type

Cell type	1	2	3	4
EC	15	2	0	0
T1-preHSC	3	15	10	0
T2-preHSC	1	13	23	11
E12-HSC	0	1	6	14
E14-HSC	0	1	3	12

Campbell K, Yau C. Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis. bioRxiv. 2016; 060442.

EMBL-EBI

Trajectory inference methods we have looked at:

- Naive PCA
- Diffusion maps
- Monocle
- TSCAN
- SLICER
- Ouija

Robust pseudotime analysis

- Use multiple approaches
- Consider bootstrapping
- Use orthogonal data (cell markers, etc)
- Use prior biological knowledge
- Be sceptical

Gotchas

- Trajectory methods will return a trajectory, even where none is present
  - Compare with clustering methods (e.g. flowSOM)
- Particular assumptions for a method may introduce bias that prevent the method from finding the correct but unexpected trajectory in the data

Visualisation

Acknowledgements

Stegle Group @ EMBL-EBI:
Oliver Stegle, Florian Buettner, Anna Cuomo, Raghd Rostom, Marc Jan Bonder, Yuanhua Huang, Ricard Argelaguet + many more great colleagues

scater authors:
Aaron Lun, Kieran Campbell, Quin Wills

Hemberg Group @ Sanger Institute:
Vlad Kiselev, Tallulah Andrews, Martin Hemberg

Teichmann Group @ Sanger Institute:
Sarah Teichmann, Valentine Svensson + more

Bioconductor and open-source software developers

Australian Government
National Health and Medical Research Council

EMBL-EBI

Useful resources

- ***scater*:**
  McCarthy et al. *Bioinformatics*. 2017; doi:10.1093/bioinformatics/btw777
  http://bioconductor.org/packages/scater/

- F1000 Research "A **step-by-step workflow** for low-level analysis of single-cell RNA-seq data":
  http://f1000research.com/articles/5-2122/v1

- **Hemberg Lab scRNA-seq course:**
  http://hemberg-lab.github.io/scRNA.seq.course/

- **conquer:** a repository for processed, QC'd single-cell datasets.
  http://imlspenticton.uzh.ch:3838/conquer/

- **scRNA tools database:** https://www.scrna-tools.org/

- **(Long!) list** of single-cell tools and software:
  https://github.com/seandavi/awesome-single-cell