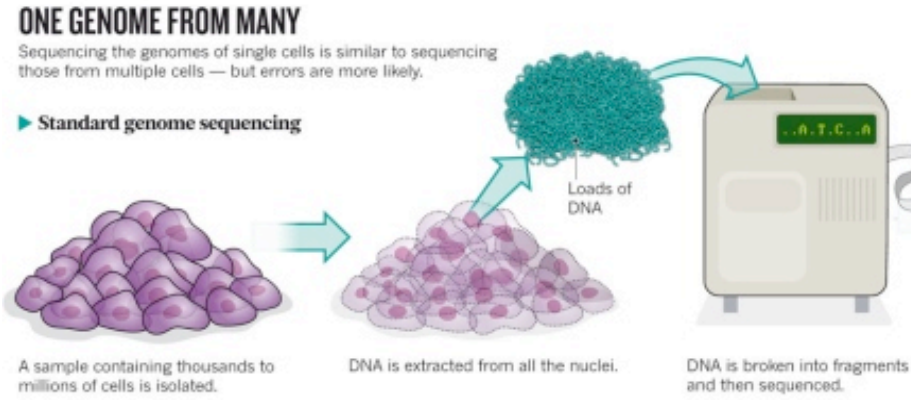
LINEAGE ESTIMATION WITH SINGLE CELL MRNA-SEQ DATA

Elizabeth Purdom
Associate Professor
Department of Statistics, UC Berkeley

Statistical and Computational Challenges in Large Scale Molecular Biology
Banff International Research Station
March 28, 2017

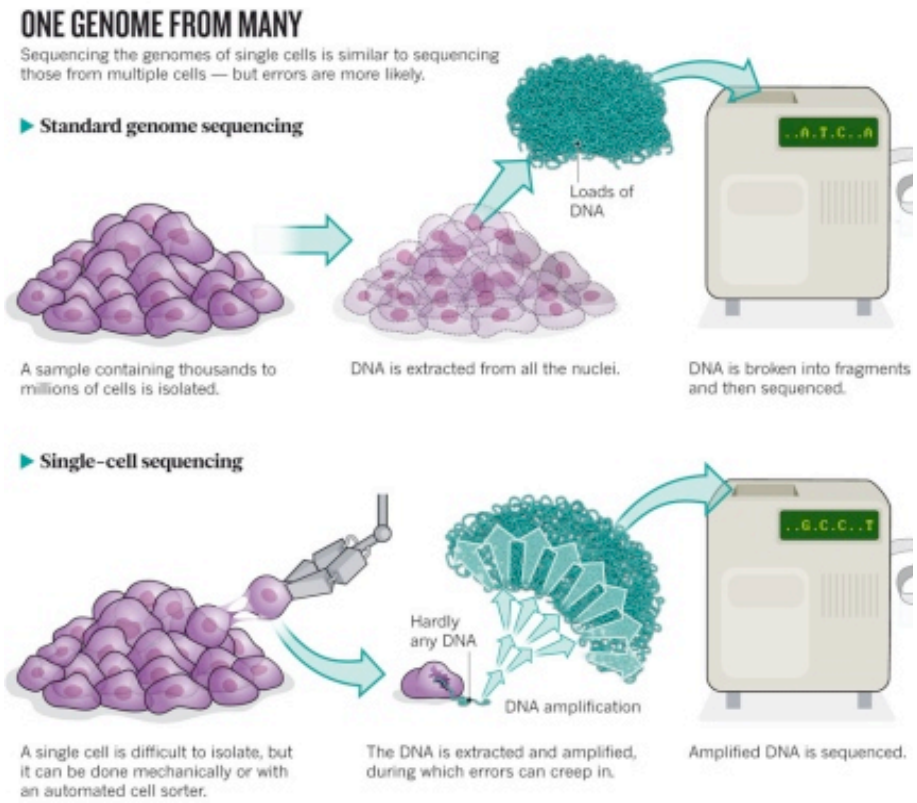
Single Cell sequencing

- Standard mRNA-Seq on bulk populations



Single Cell sequencing

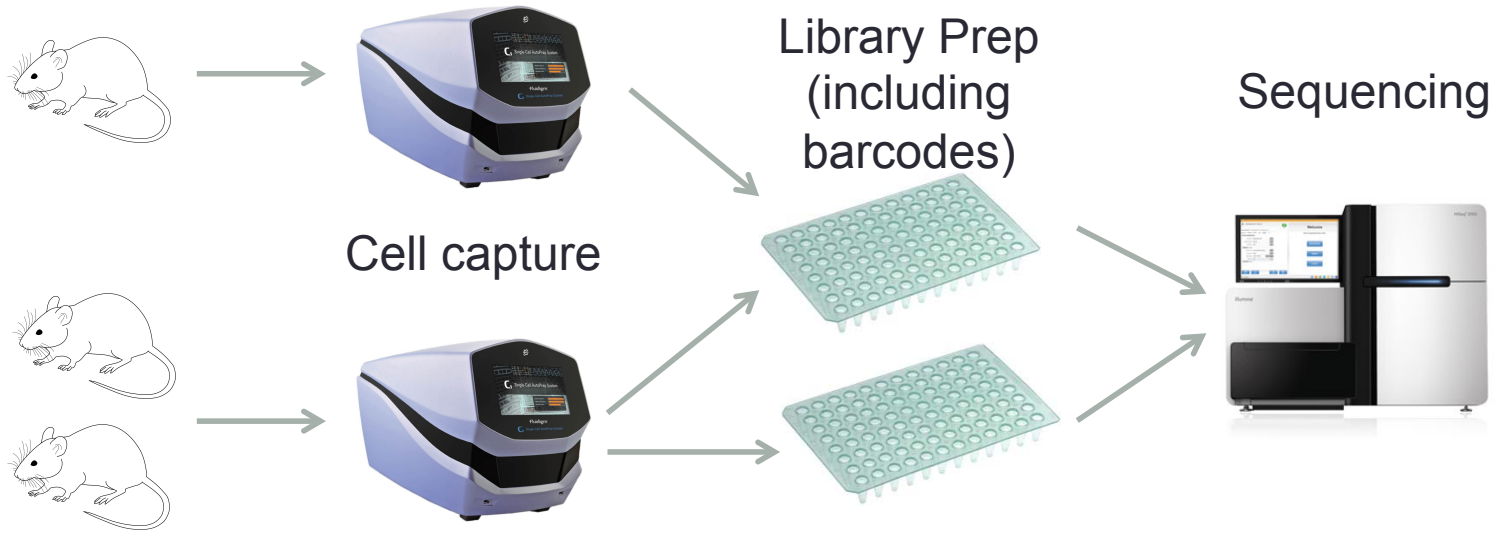
- Standard mRNA-Seq on bulk populations
- Single cell: allows to see diversity of individual cells



Owens (2012) “Genomics: The single life” Nature News

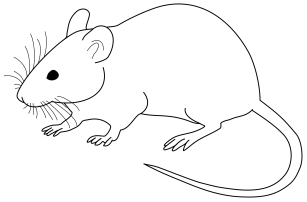
Experimental process

- Isolate cell
  - Micropipette
  - FACS : Fluidigm C_1
    ≤96 cells per run*,
    good: 60-70% capture rate
  - Droplet
- Library Prep
  - Amplification: small input material, high amplification
- Sequencing
  - Low seq. depth: e.g. 96 per lane (1M reads)

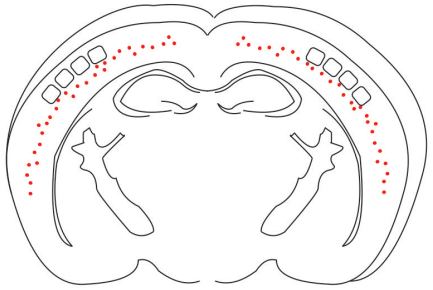


S1 cortex in mice (NIH BRAIN Initiative Cell Census Consortium)

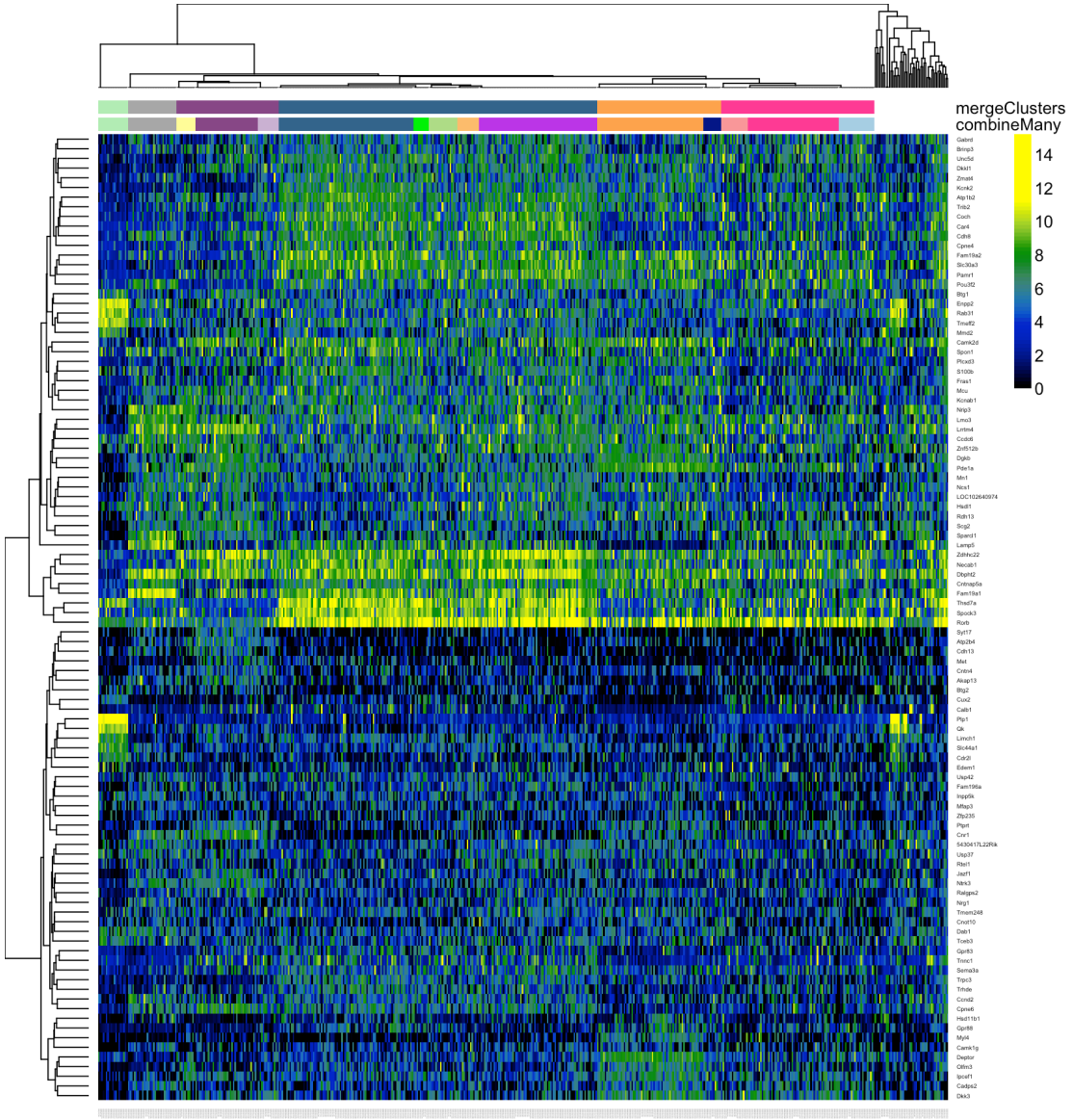
- FACS sorting of the S1 cortex (Layer 4/5/6)



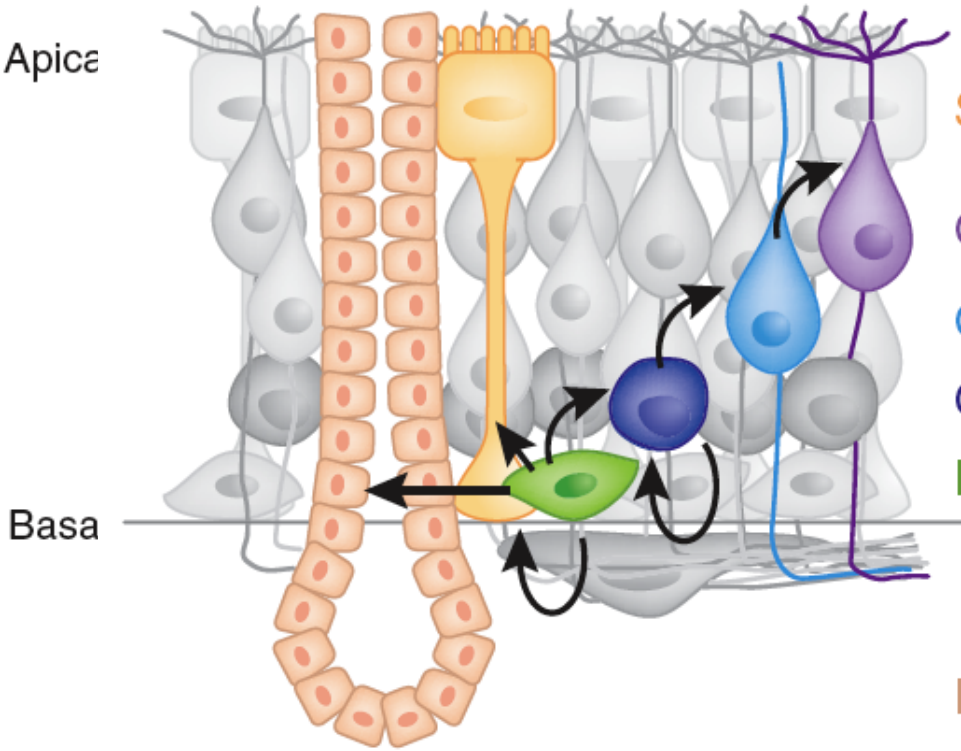
Layer 5 Cre x tdTomato reporter



Layer 5 cells (Glial contaminents removed)



Olfactory Epithelium (OE)



Sus	**Sustentacular cells**
ORN_m	**Mature olfactory neurons**
ORN_i	**Immature olfactory neurons**
GBC	**Globose basal cells**
HBC	**Horizontal basal cells**
BG	**Bowman's gland**

Quick snapshot of the data

Data Set	Olefactory	Brain
# mice	51	41
# C1 Batches	61	40
# Illumina Lanes	19	7
# cells	2,627*	1,249
# cells pass QC	2,190	1,042
# Sequenced Reads	4,001 Million	1,500 Million

* Many conditions: in this talk, only 904 total (687 after sequencing)

Overview

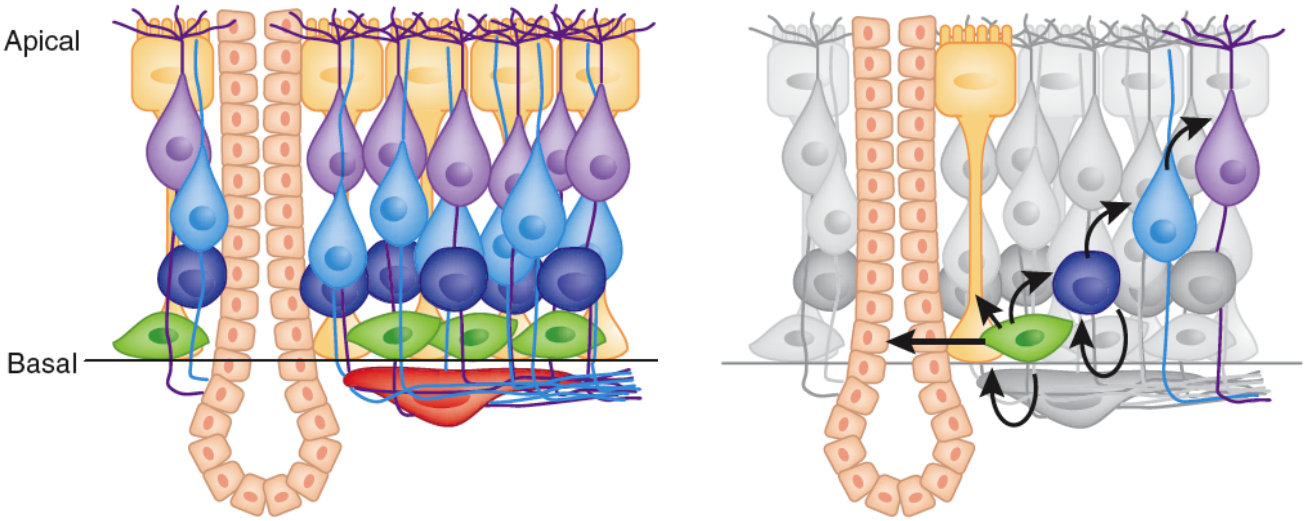
- SCONE
  - Data specific choice of normalization strategy
  - Via comprehensive comparison *in every dataset*
  - Metrics to rank normalized data
- RSEC
  - Robust clustering strategy to find heterogeneity in scSeq data
  - Subsampling and sequential clustering, merging of clusters, …
  - Part of `clusterExperiment` package for common clustering tasks (e.g. pairwise DE, plotting with clustering information)
- Slingshot
  - Estimation of developmental lineages

Overview

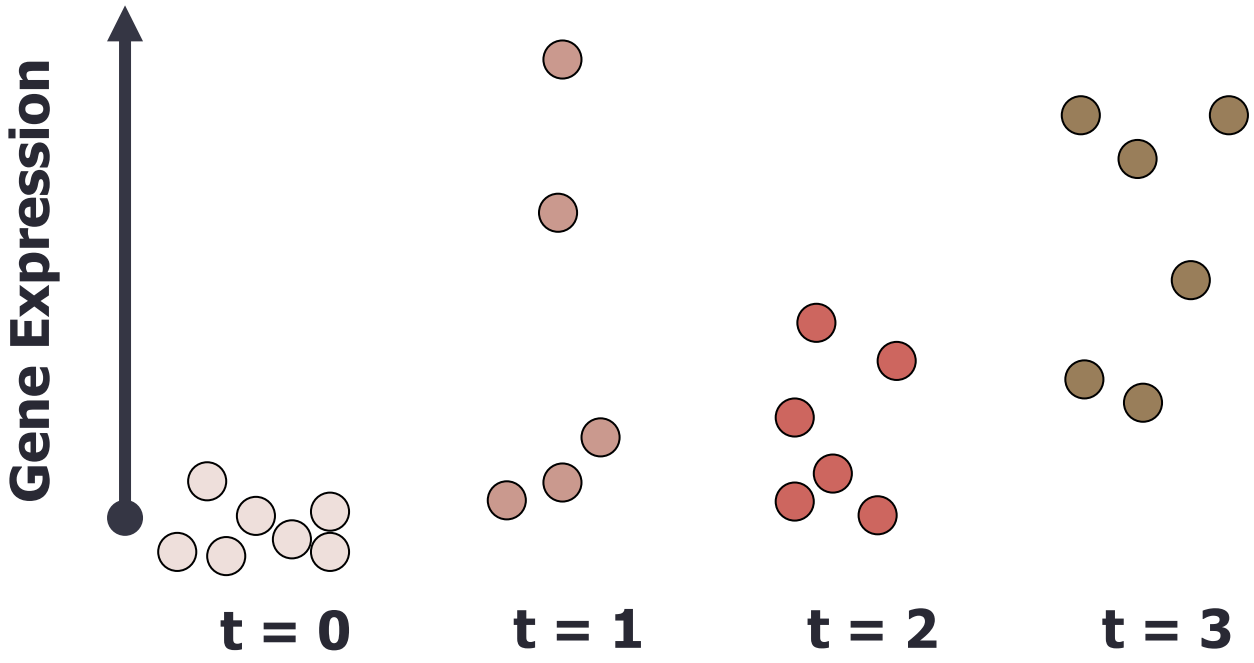
- SCONE
  - Data specific choice of normalization strategy
  - Via comprehensive comparison *in every dataset*
  - Metrics to rank normalized data
- RSEC
  - Robust clustering strategy to find heterogeneity in scSeq data
  - Subsampling and sequential clustering, merging of clusters, …
  - Part of `clusterExperiment` package for common clustering tasks (e.g. pairwise DE, plotting with clustering information)
- Slingshot
  - Estimation of developmental lineages

Experiments (two):

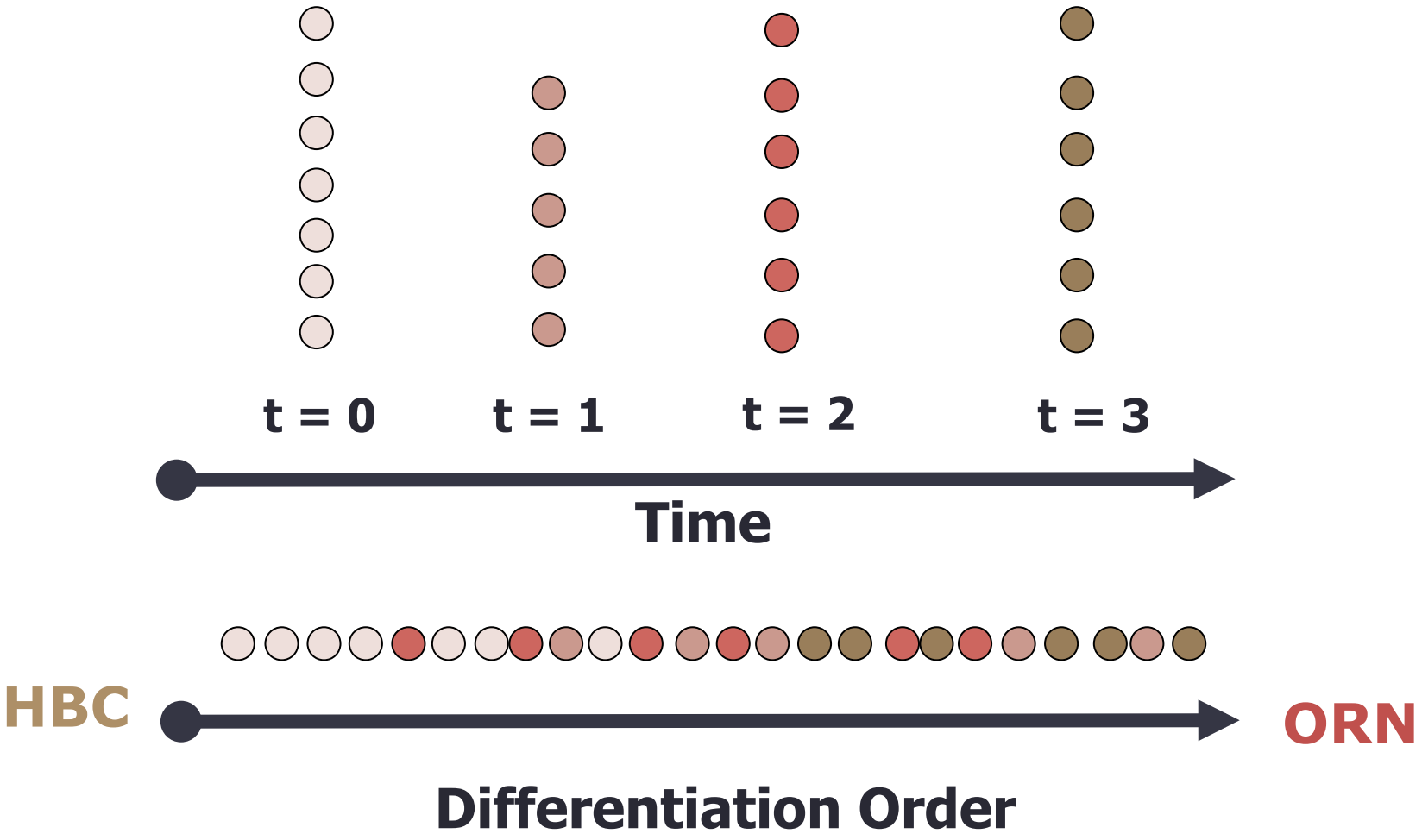
- Capture descendant cells at several time points after regeneration
  - Destroy all but HBC and watch them regenerate: 145 cells (of 175)
  - Lineage tracing after inducing HBC: 542 cells (of 729)
- Sequence the individual cells to determine what is changing
- Goal: characterize the differentiation process and at what point cell fate is chosen



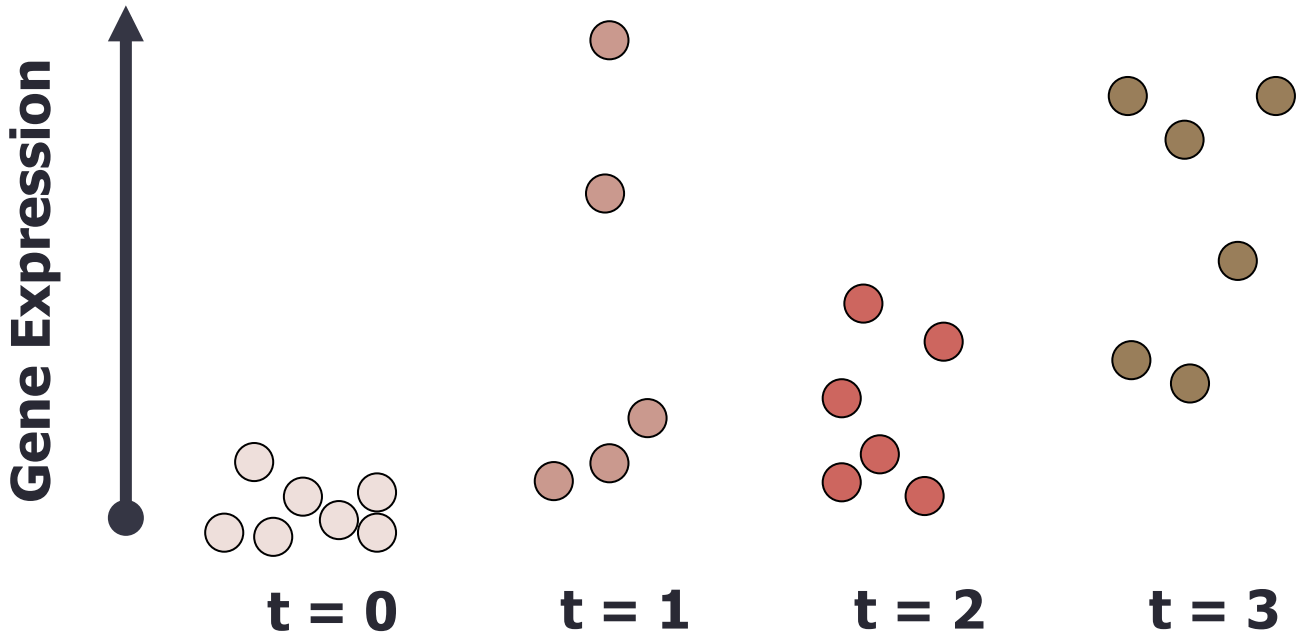
Find genes related to differentiation



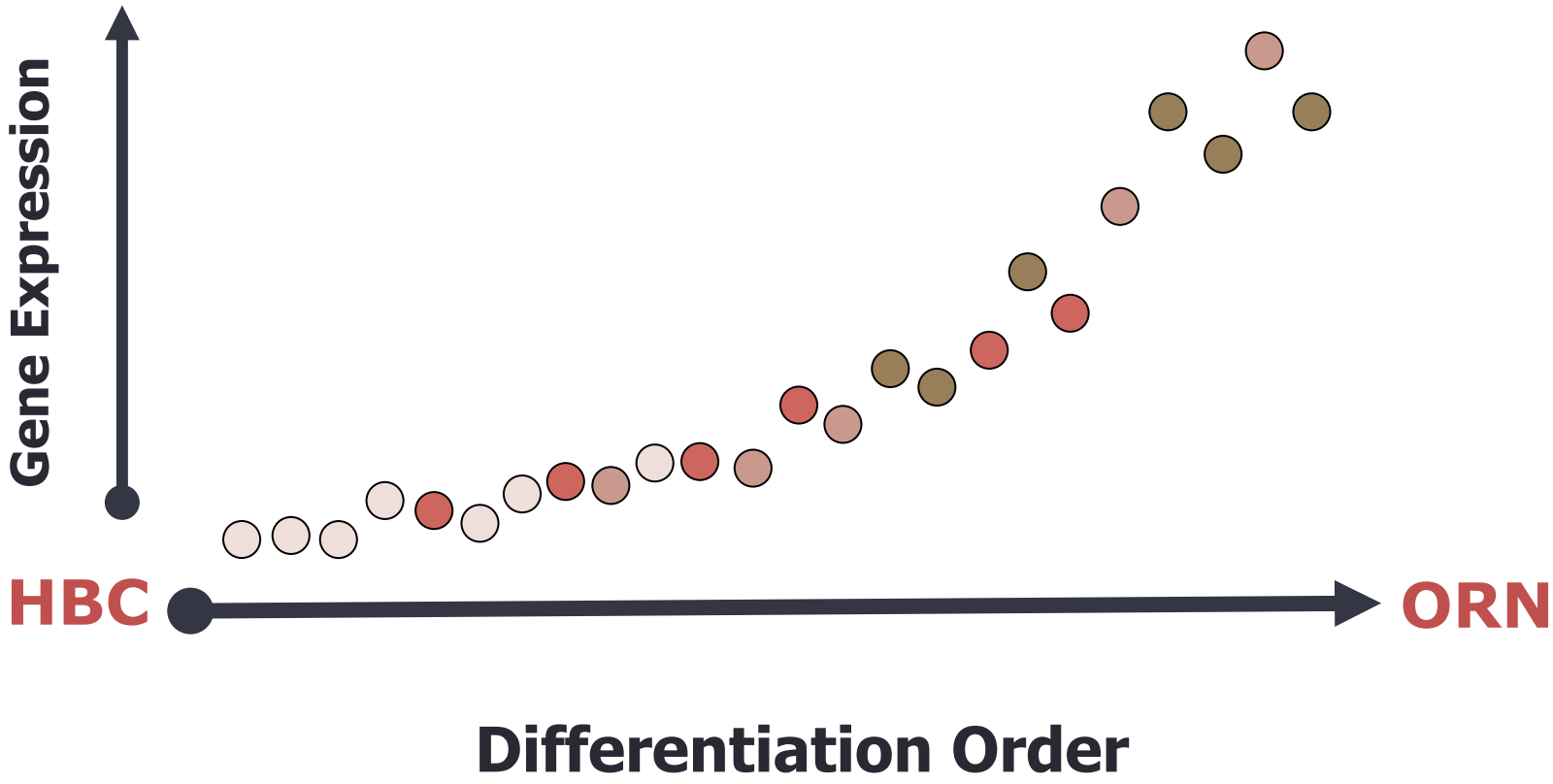
But observed time is not differential state



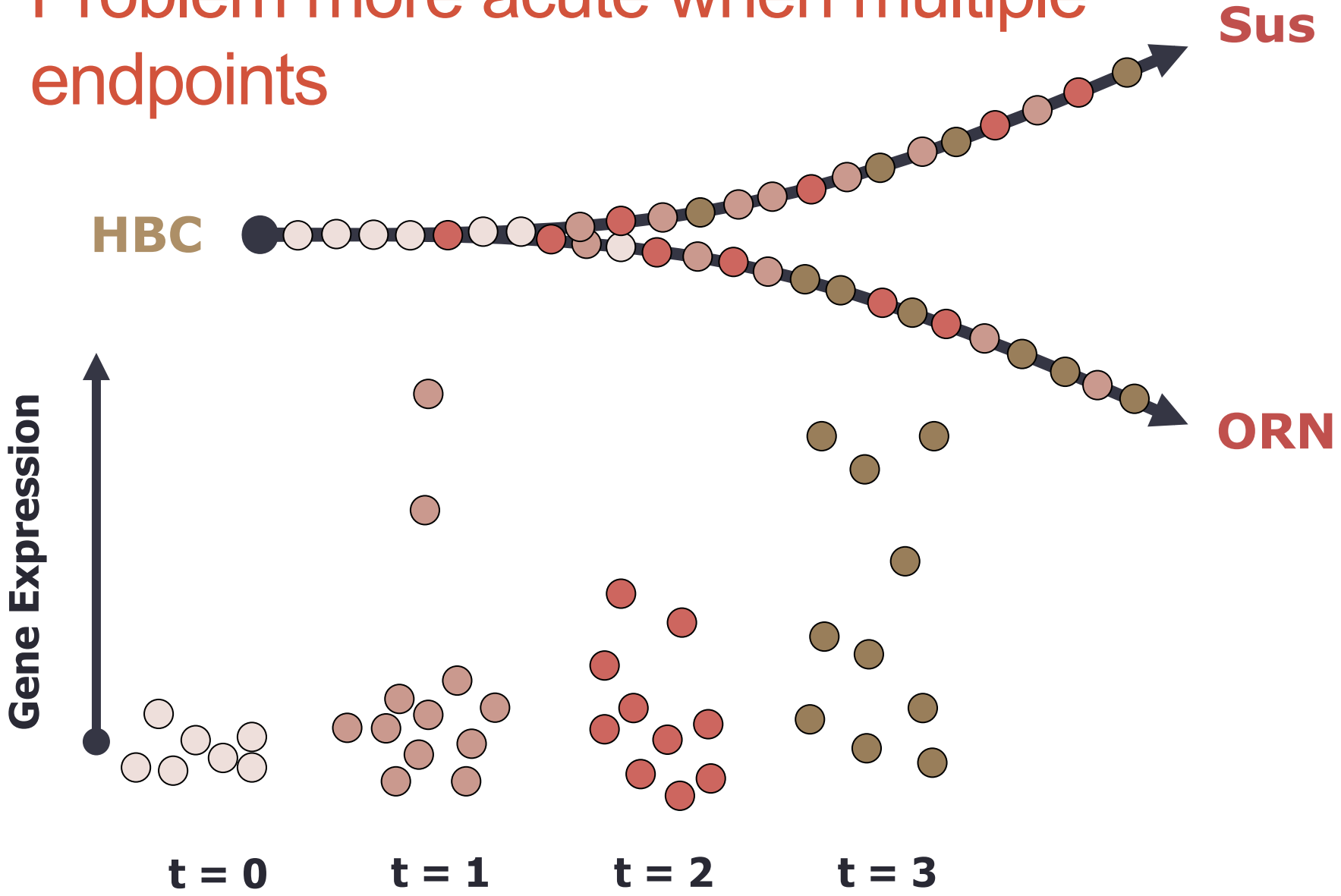
Better representation if order cells by differentiation state

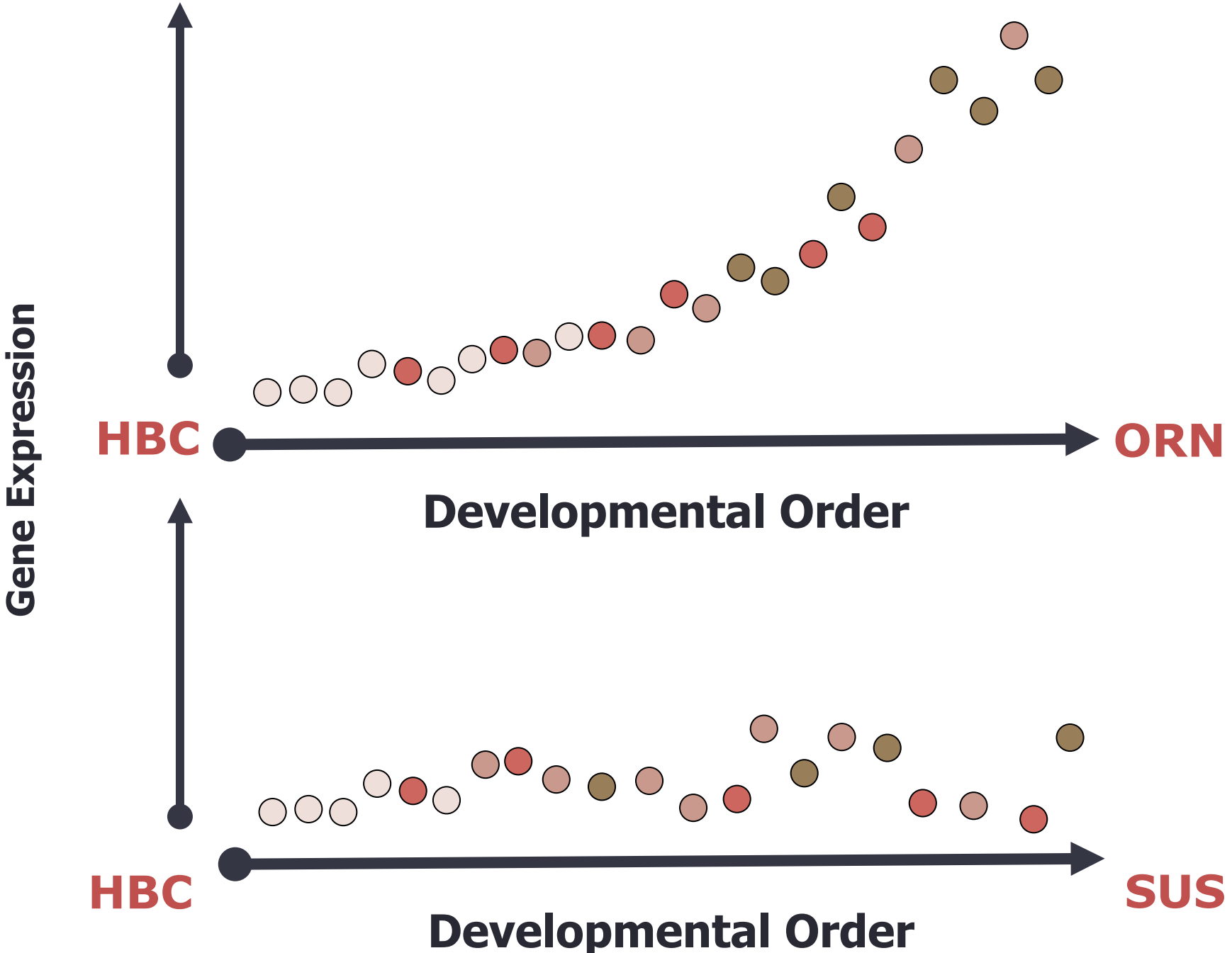


Better representation if order cells by differentiation state



Problem more acute when multiple endpoints



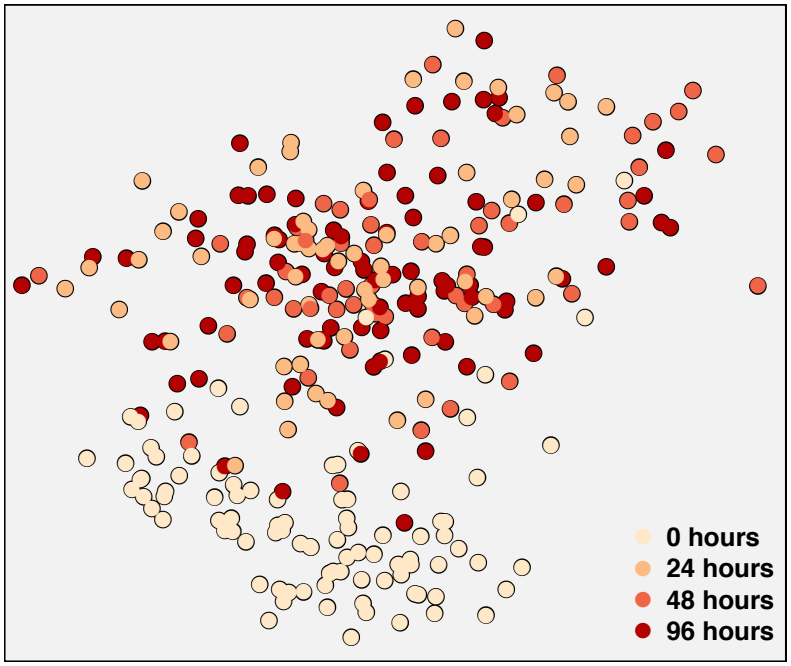

Gene Expression
HBC
ORN
Developmental Order
HBC
SUS
Developmental Order

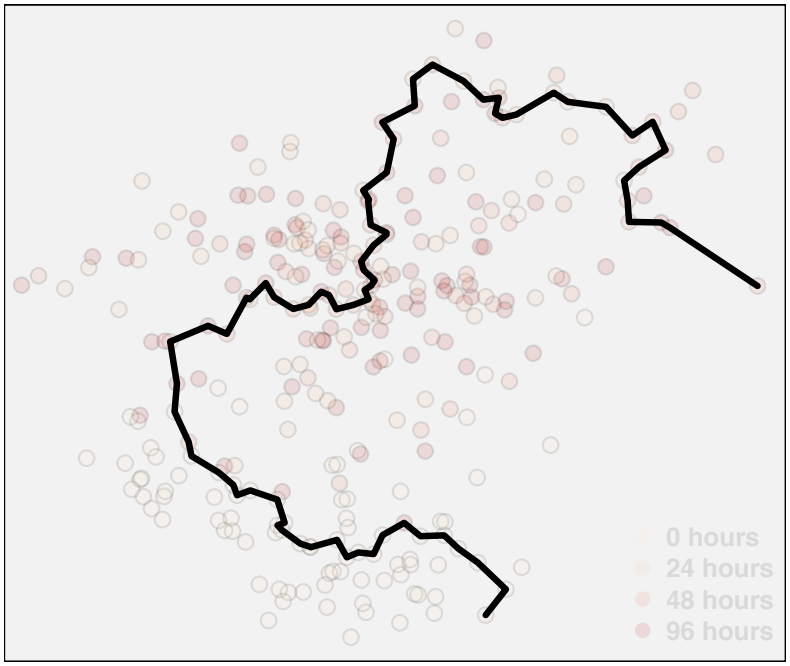
Many Strategies for One lineage

- Assume distance gives differentiation order, at some level
- Find a ‘path’ (lineage) through space of gene expression data
- Order individual cells on the path
  - E.g. orthogonal projection
- Many “details” hard-coded in, make comparisons difficult
  - Dimensionality of space (e.g. 2 dimensions)
  - How find low dimensions (ICA / PCA / Laplacian Embedding)

Path Choices

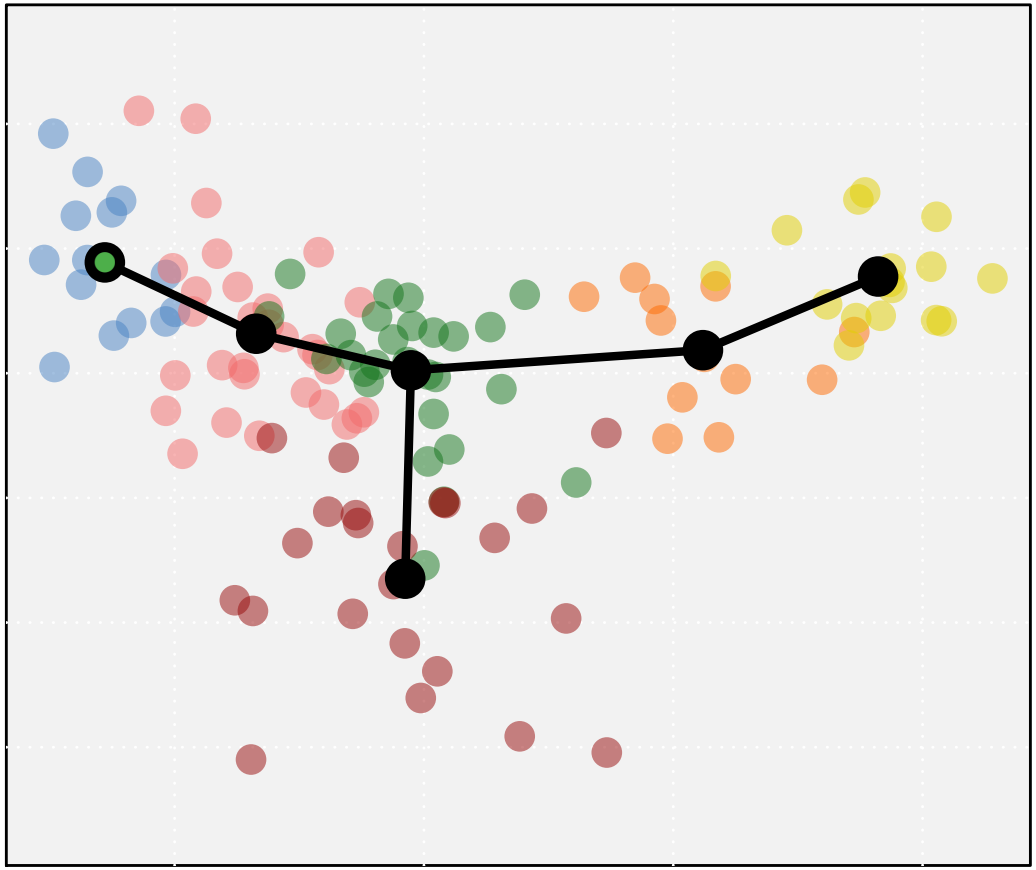
- MST through individual cells, take longest path (Monocle) ‘Project’ onto path via where branch off path





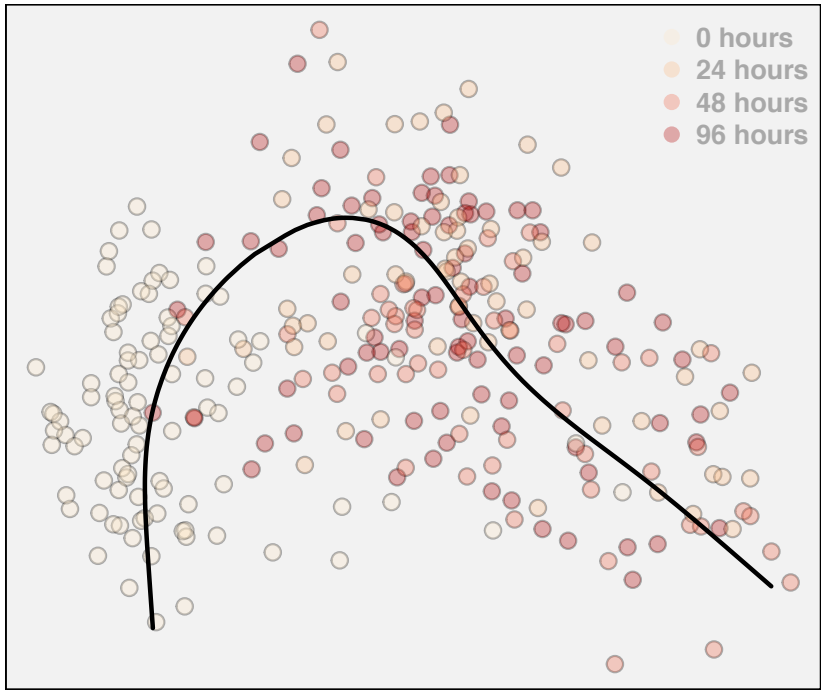
Path Choices

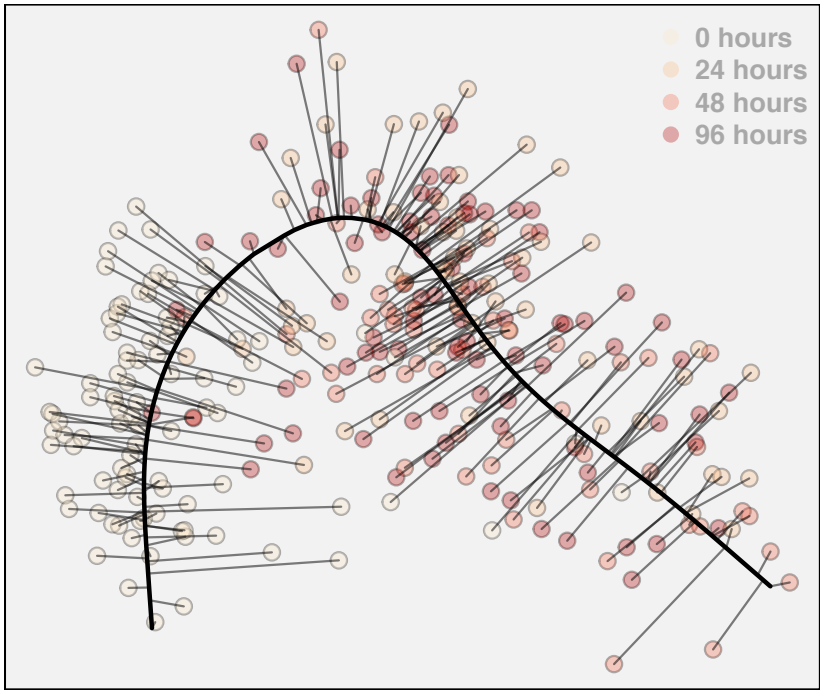
- MST through individual cells, take longest path (Monocle)
- MST on Clusters, orthogonal projection (Waterfall / TSCAN)



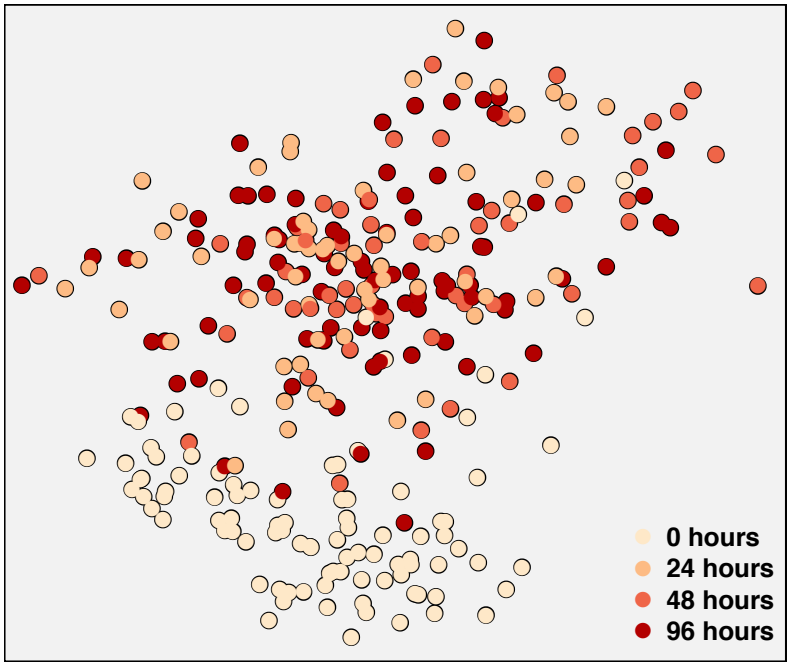
Path Choices

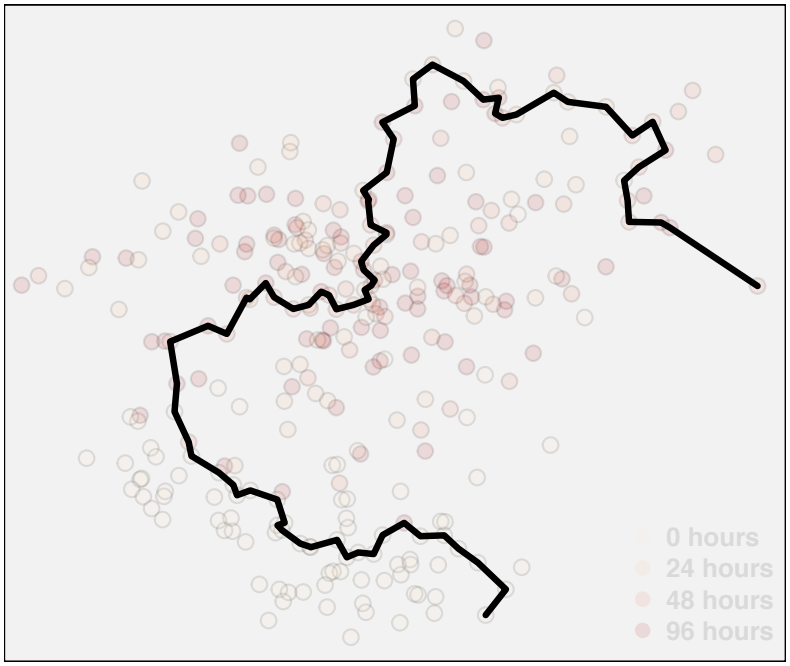
- MST through individual cells, take longest path (Monocle)
- MST on Clusters, orthogonal projection (Waterfall / TSCAN)
- Principal Curves, orthogonal projection (Embedder)



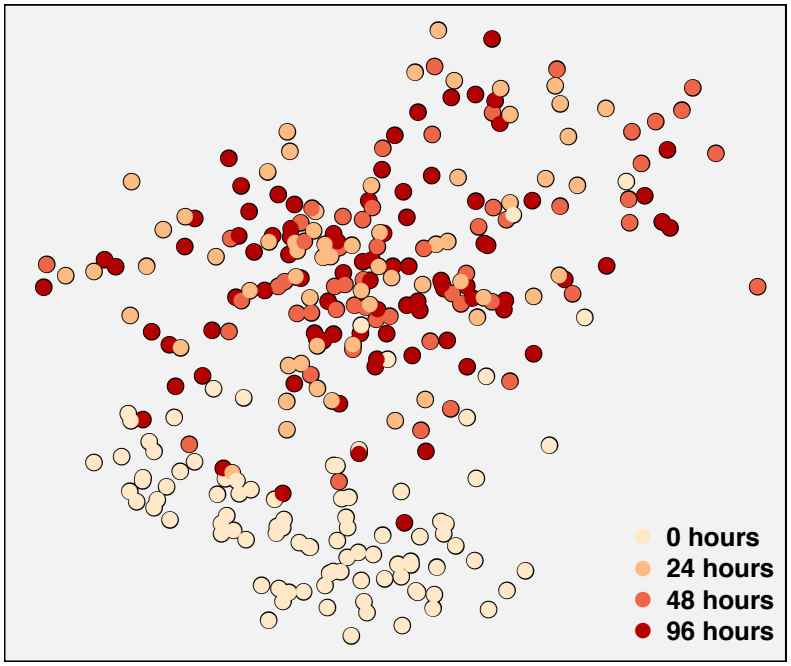


Monocle not robust

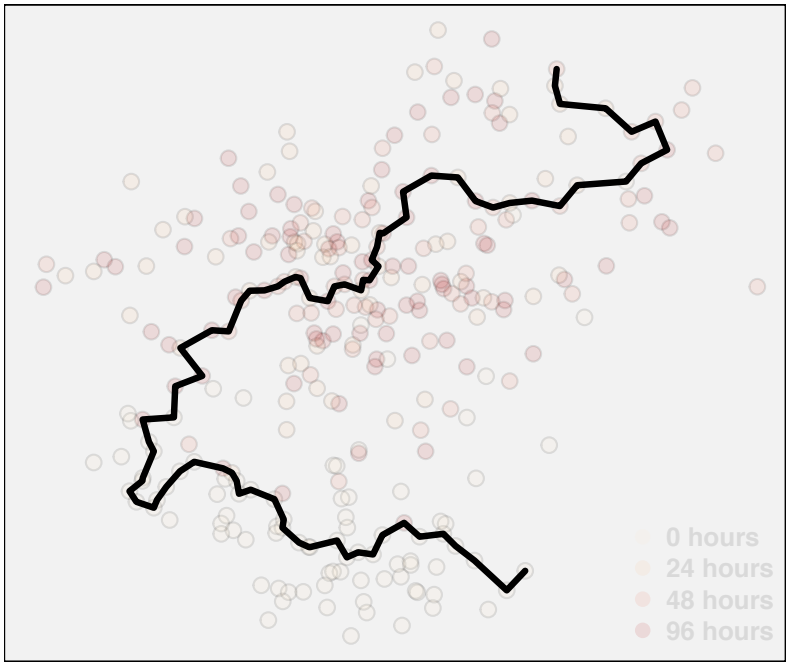




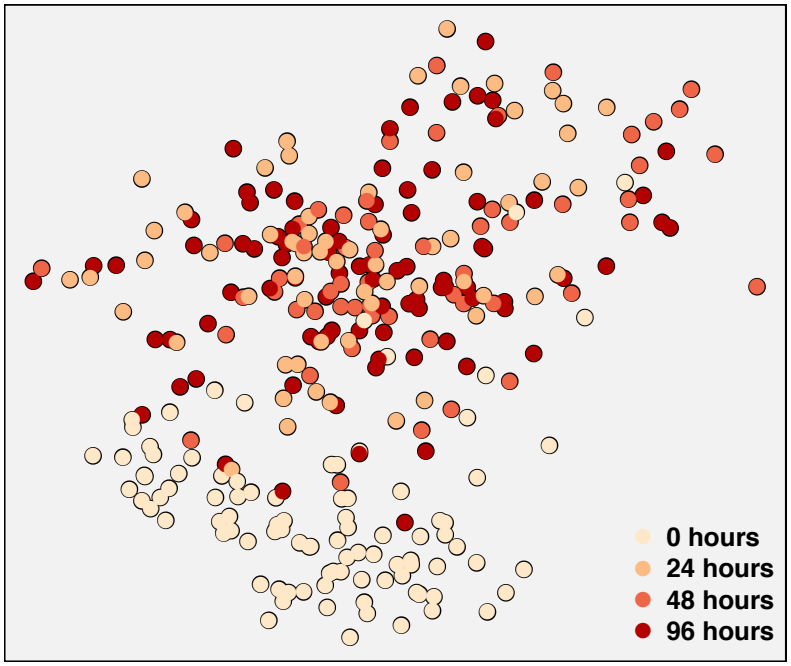
Monocle not robust



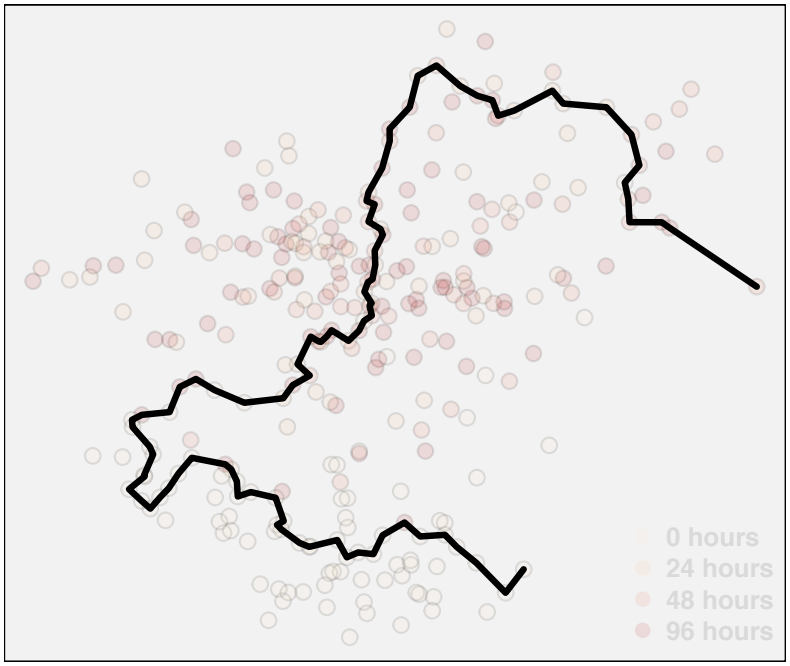
(Jittered)



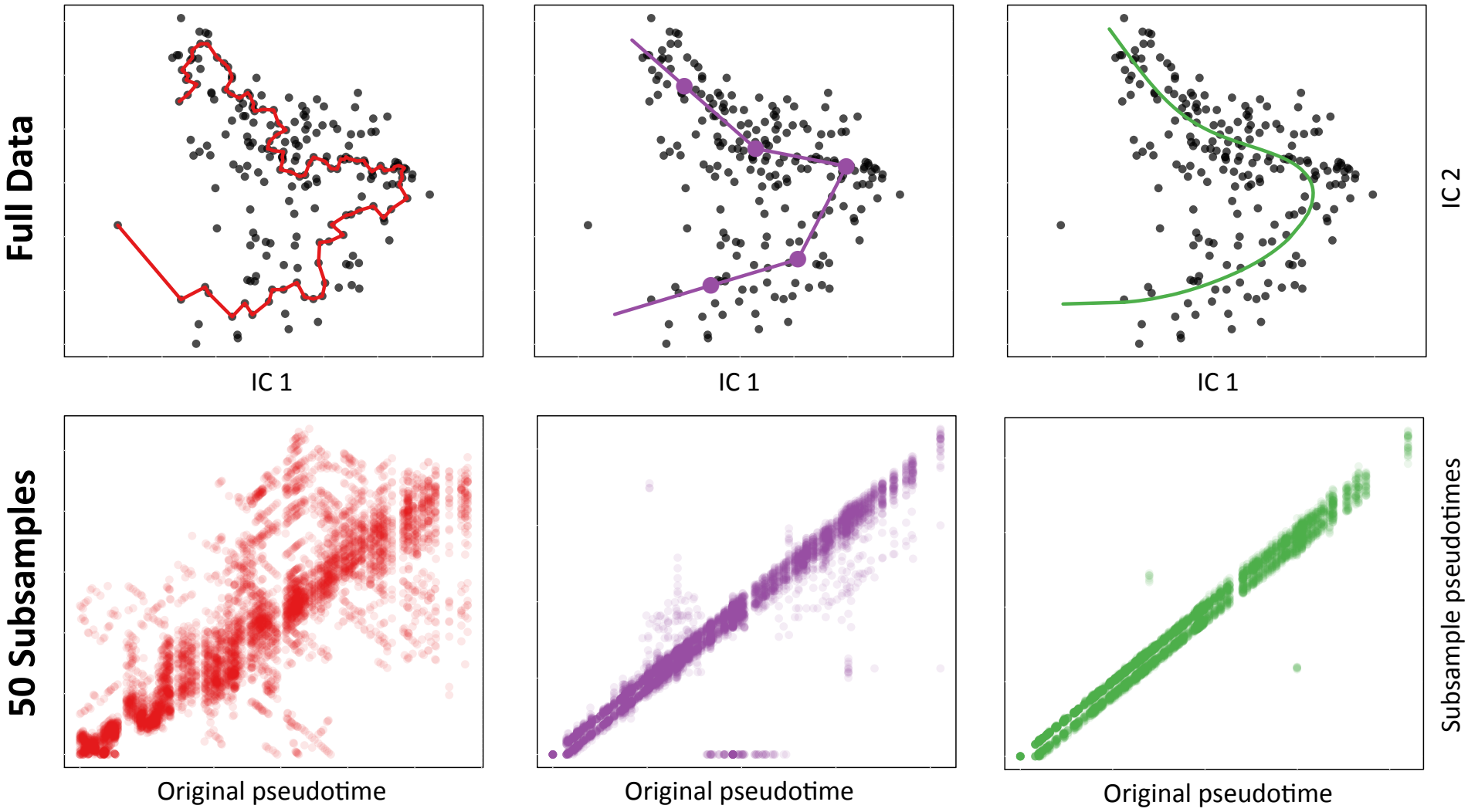
Monocle not robust



(Jittered)



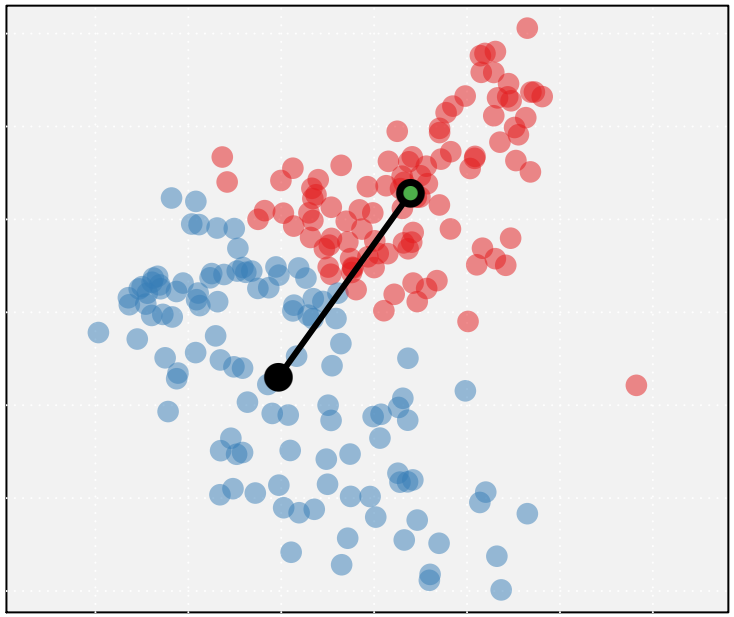
Principal Curves More Stable



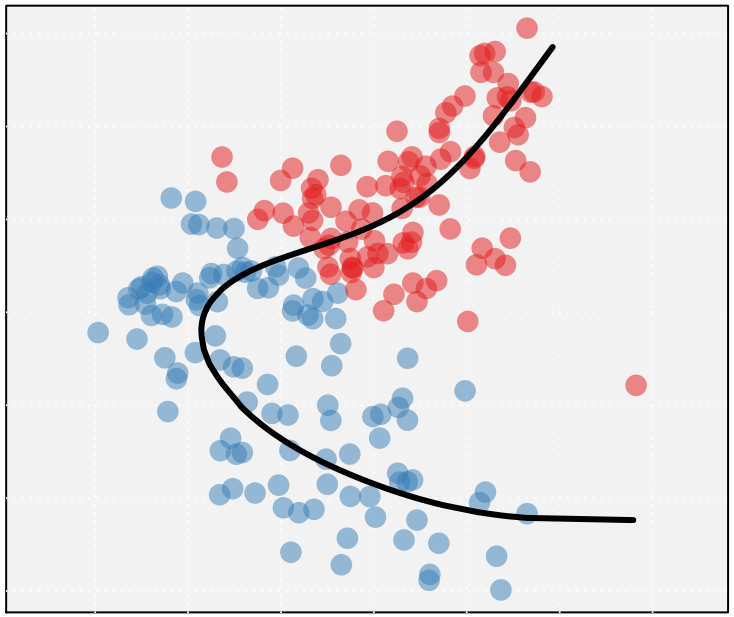
Principal Curves Not Reliant on Clustering

- MST on clusters can be sensitive to choice of clusters

MST on Clusters



Principal curves



Slingshot: Multiple Lineages

- MST useful for broad shapes, finding branching
  Clustering often uses more dimensions – more information
- Principal curves more robust estimates of ordering

- Slingshot

  → Use MST for assigning clusters of cells to lineages
  → Principal curves within lineages to give ordering

Slingshot: Multiple Lineages

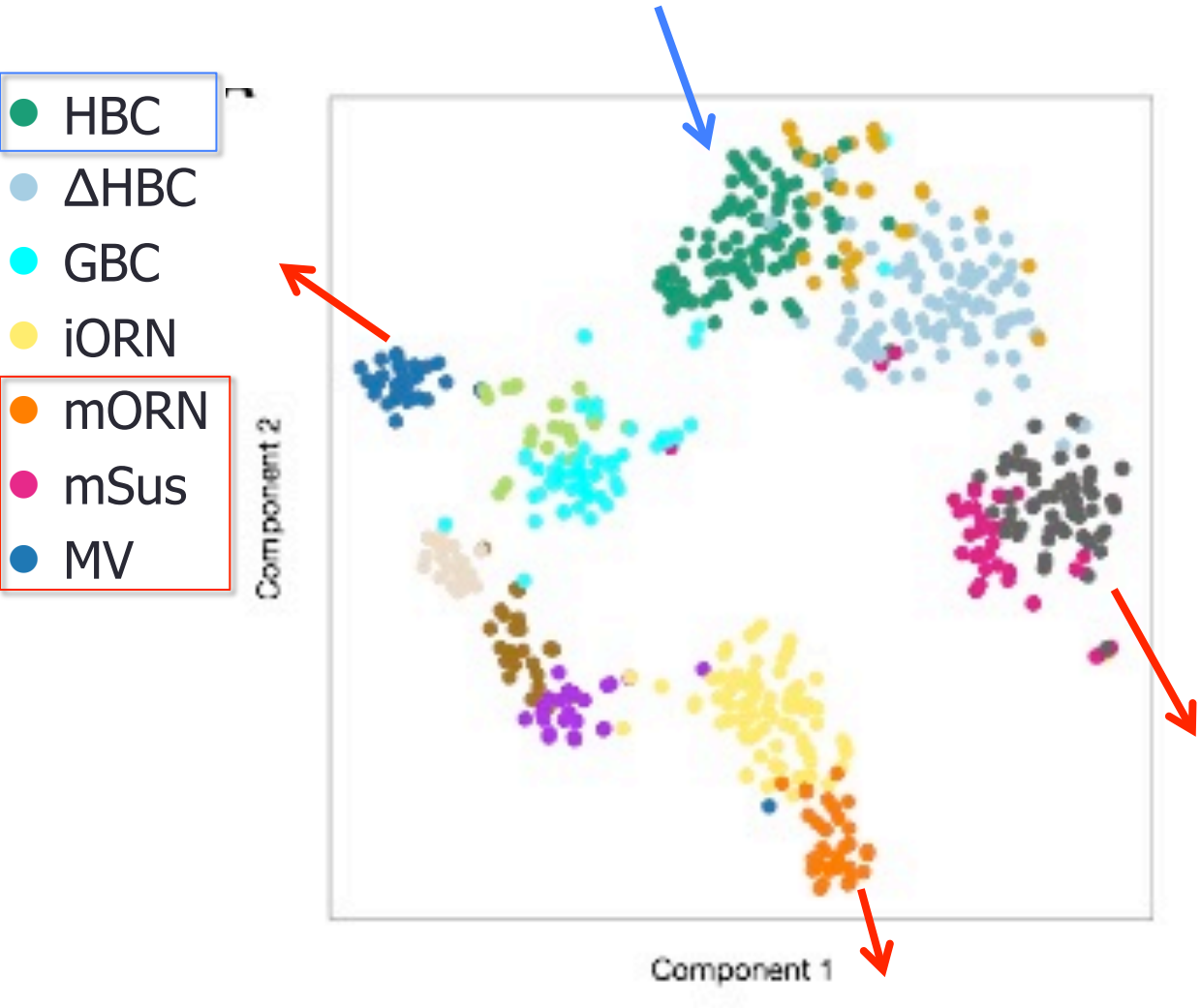
- MST useful for broad shapes, finding branching
  Clustering often uses more dimensions – more information
- Principal curves more robust estimates of ordering

- Slingshot

  → Use MST for assigning clusters of cells to lineages
  → Principal curves within lineages to give ordering

- Additionally
  → **allow for supervision (constrained MST)**

Importance of Constrained MST

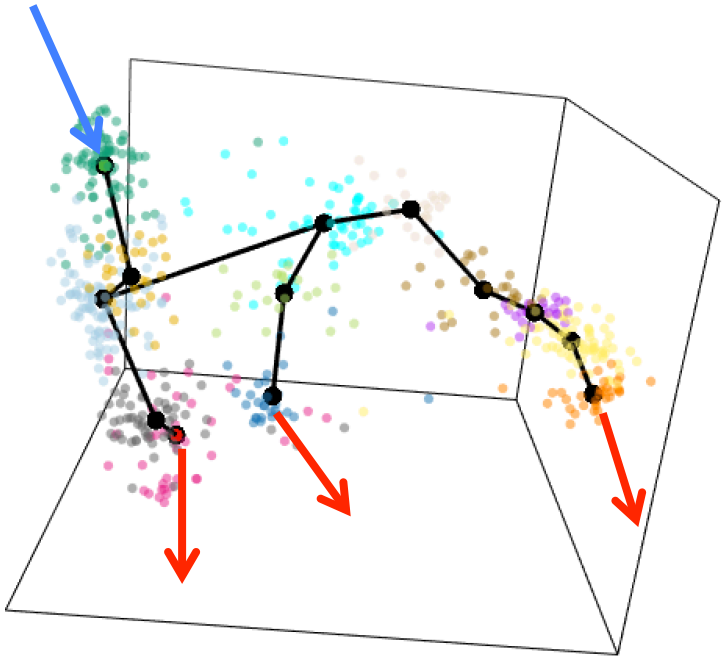


PCA of Gene Expression, with clusters

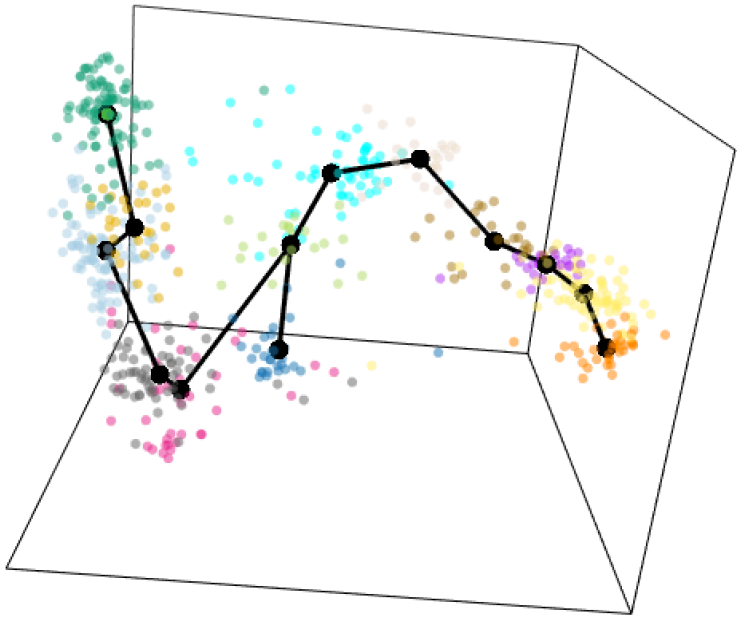
- Huge assumption distance in gene expression = order
- Clustering gives important information
- If know the end points of process, should guide estimation

Constraint keeps these lineages separate

With Constraints



Without Constraints



- HBC
- ΔHBC
- GBC
- iORN
- mORN
- mSus
- MV

Slingshot: Multiple Lineages

- MST useful for broad shapes, finding branching
  Clustering often uses more dimensions – more information
- Principal curves more robust estimates of ordering

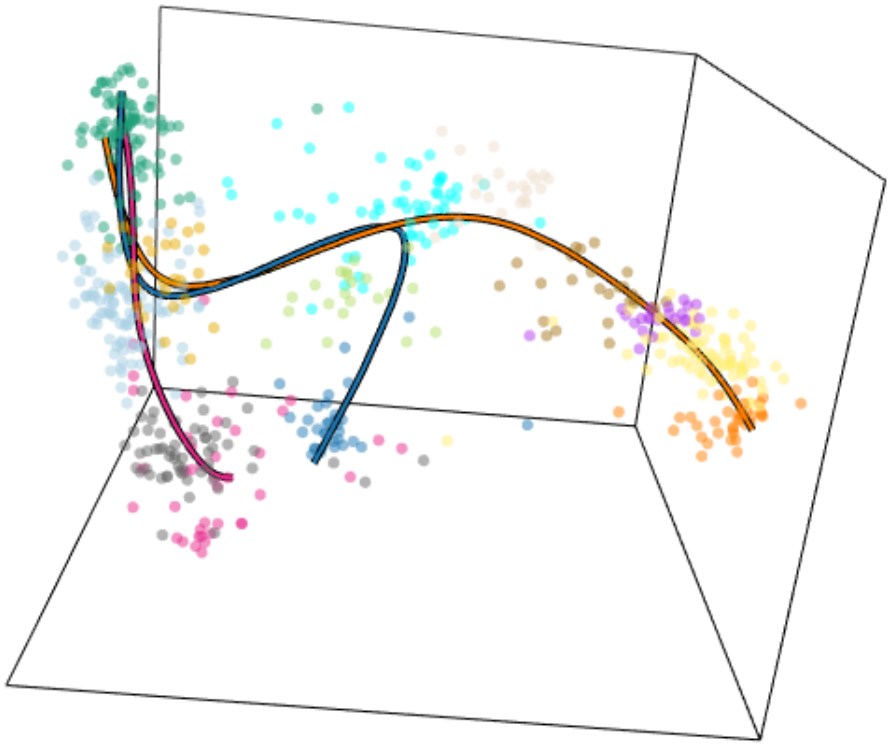
- Slingshot

  → Use MST for assigning clusters of cells to lineages
  → Principal curves within lineages to give ordering

- Additionally
  → allow for supervision (constrained MST)
  → **simulataneous principal curve fitting for overlapping branches**

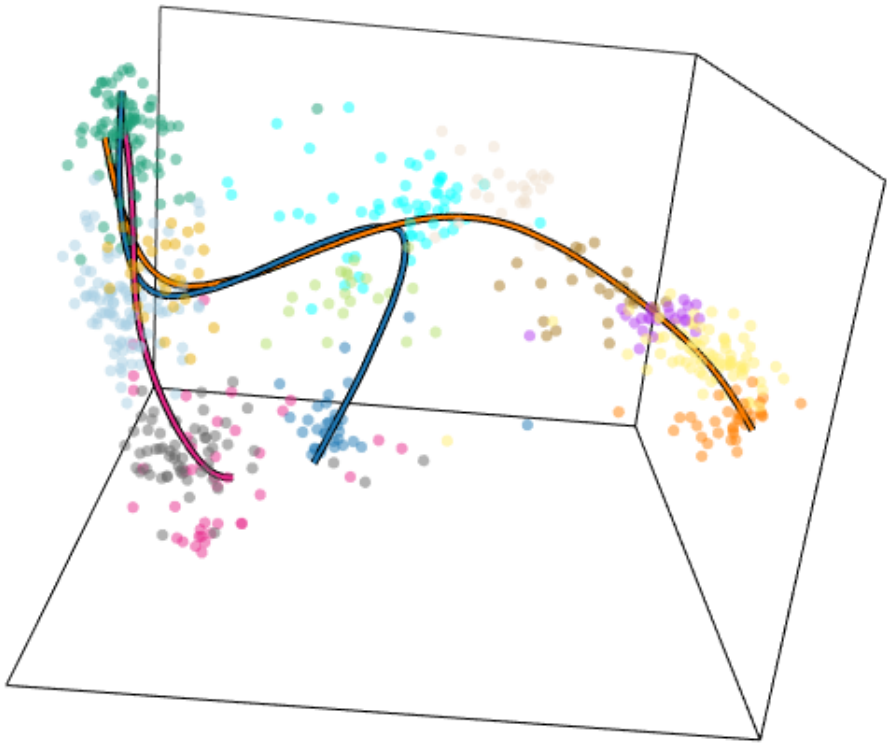
Shrinkage

- Principal curves → multiple pseudotimes for same cells in multiple lineages
- Shrink curves to average based on the density of cells shared across lineages

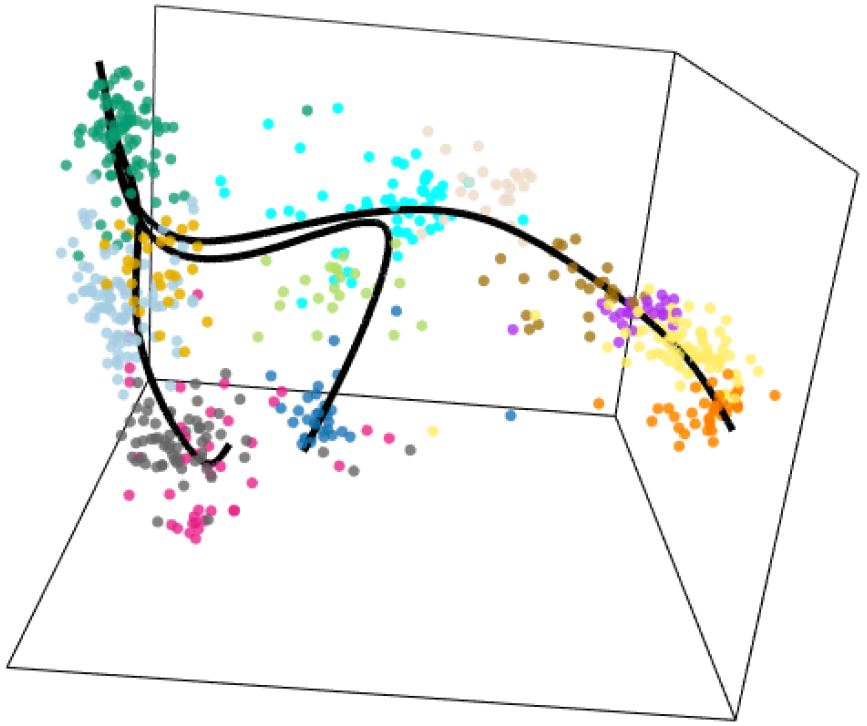


Shrinkage

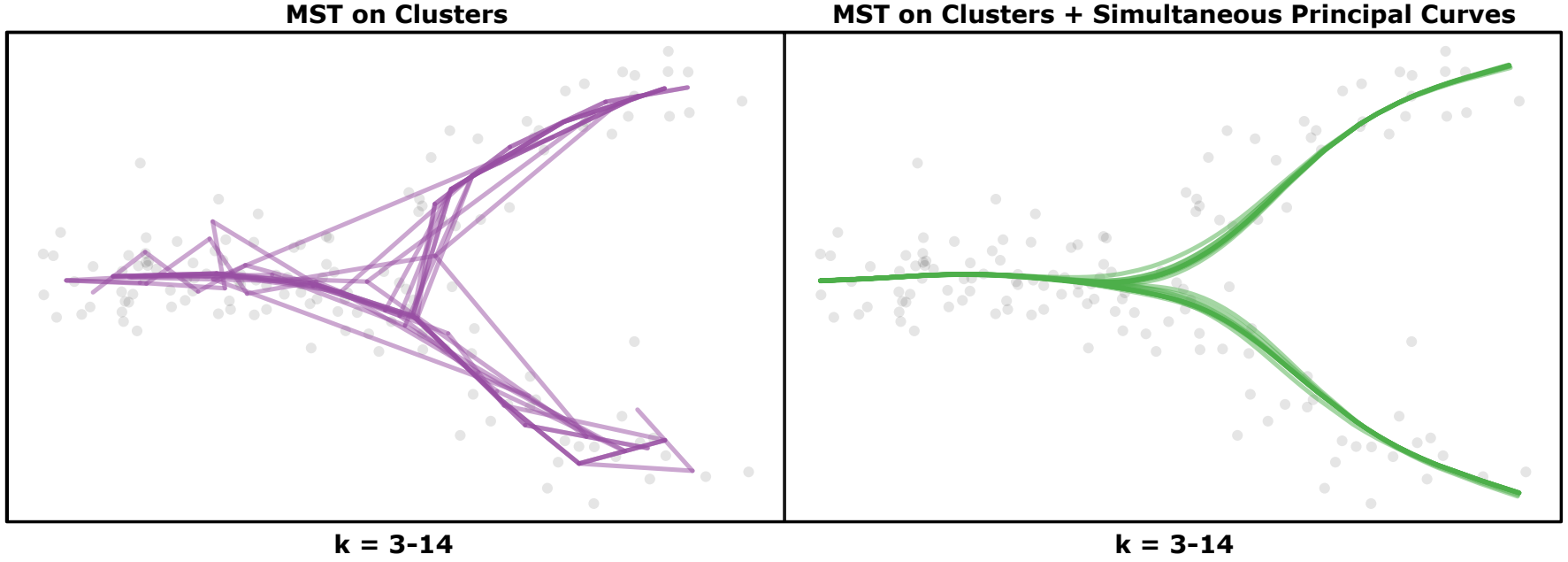
- Principal curves → multiple pseudotimes for same cells in multiple lineages
- Shrink curves to average based on the density of cells shared across lineages



With Shrinkage



Retain robustness of Principal Curves



Slingshot: Multiple Lineages

- MST useful for broad shapes, finding branching
  Clustering often uses more dimensions – more information
- Principal curves more robust estimates of ordering

- Slingshot

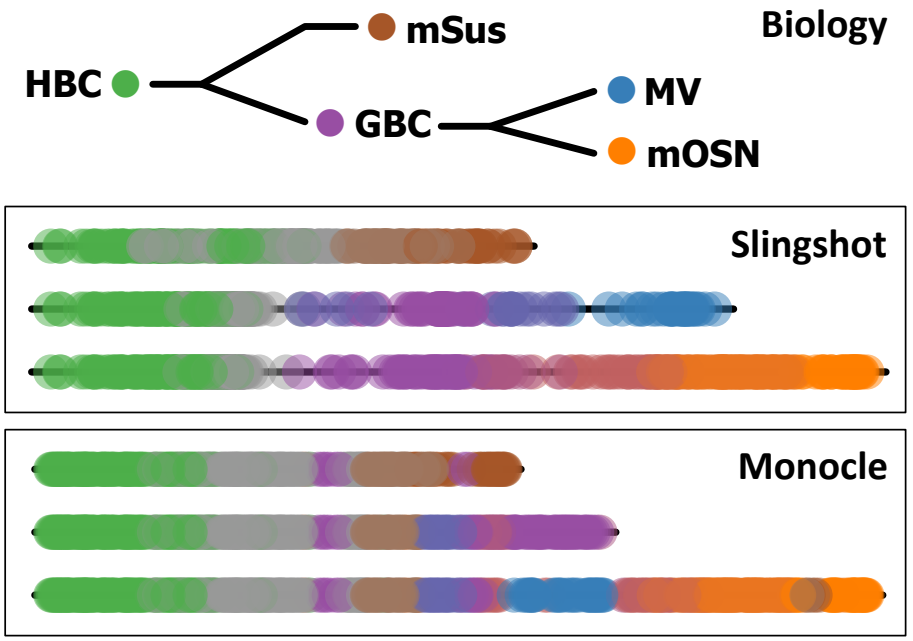
  → Use MST for assigning clusters of cells to lineages
  → Principal curves within lineages to give ordering

- Additionally
  → allow for supervision (constrained MST)
  → simulataneous principal curve fitting for overlapping branches
  → **covariance based distance for MST to capture shape of cluster**

Compare to Other Methods

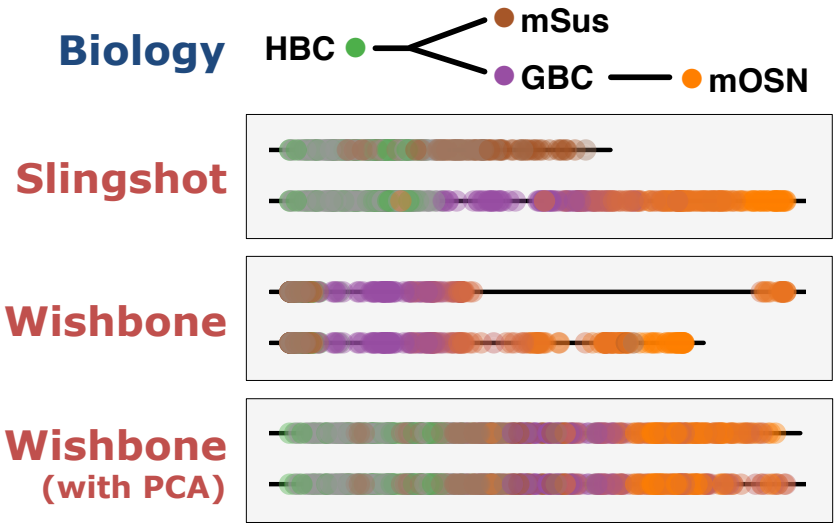
Monocle

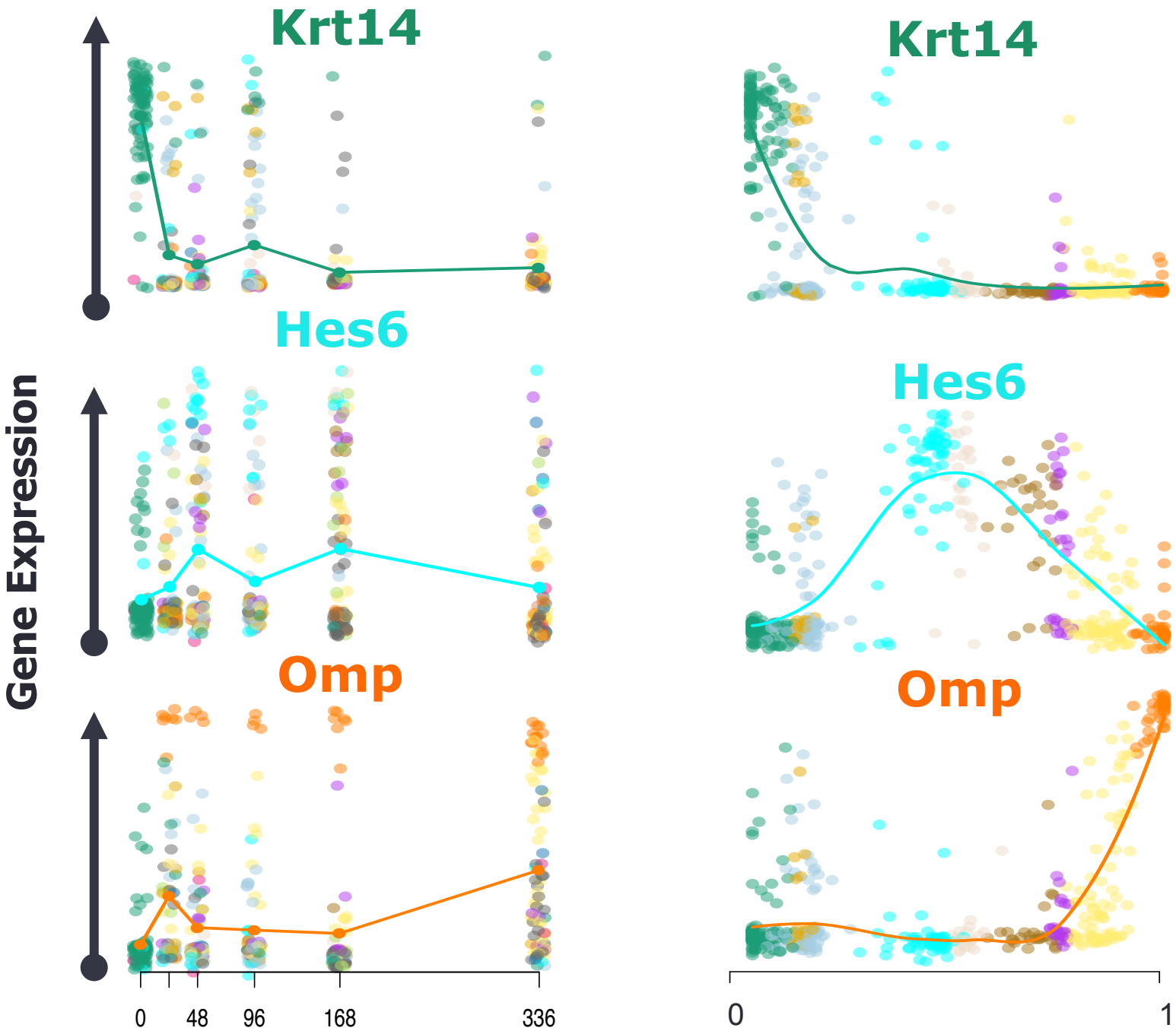
- Must specify # lineages



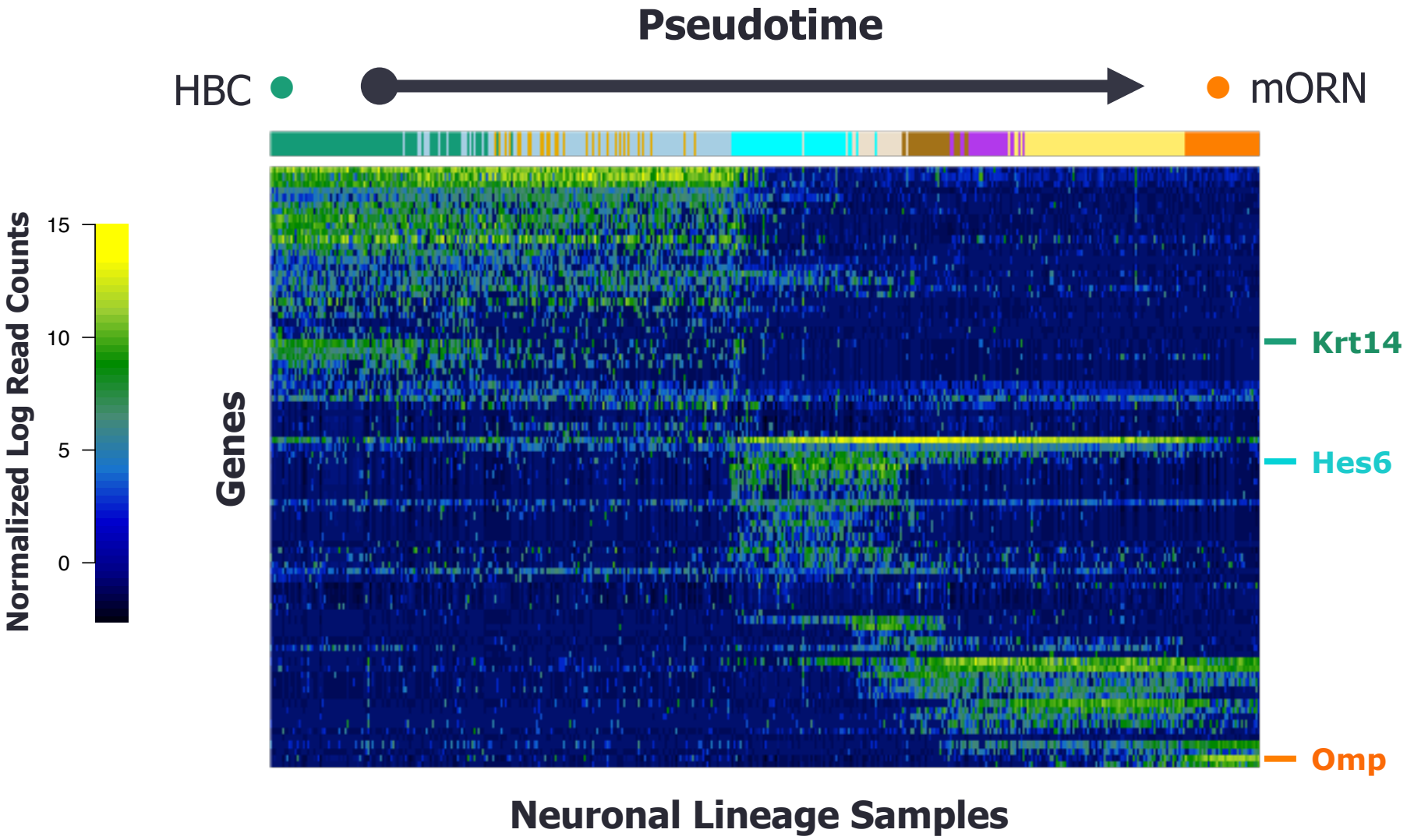
Wishbone

- Only two lineages
- Built-in Dimensionality Reduction




Krt14
Hes6
Omp
Gene Expression
0
48
96
168
336
Krt14
Hes6
Omp
0
1

Olfactory Epithelium



Concluding Remarks

- Robust and flexible method for determining lineage of cells
- However, …
- Very high expectations → Many assumptions
- Processing and dimensionality reduction are also critical components

John Ngai
David Stafford
Jasper Visser
Russell Fletcher
Diya Das
Levi Gadye
Mike Sanchez
Ariane Baudhuin

Hillel Adesnik
David Taylor
Alex Naka

Sandrine Dudoit
Elizabeth Purdom
Davide Risso
Kelly Street

Nir Yosef
Allon Wagner
Michael Cole

Functional Genomics Lab
Justin Choi

CRL Flow Cytometry Core
Hector Nolla

RSEC available as part of clusterExperiment package on bioconductor
SCONE available on bioconductor (dev)
Slingshot available on https://github.com/kstreet13/slingshot

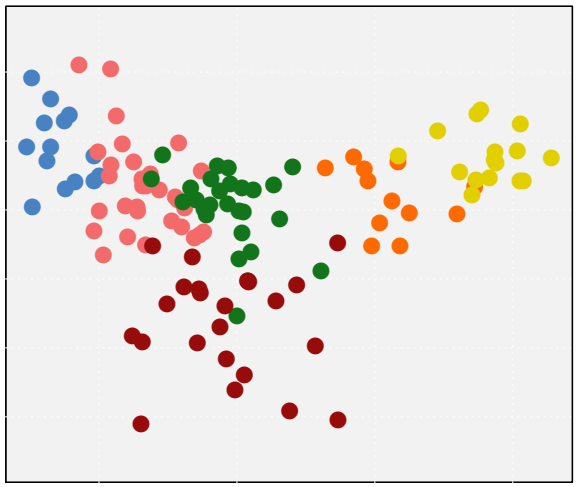
---

NIH BRAIN Initiative Cell Census Consortium
National Institute on Deafness and Other Communication Disorders
National Institute on Aging
National Human Genome Resource Institute
California Institute for Regenerative Medicine

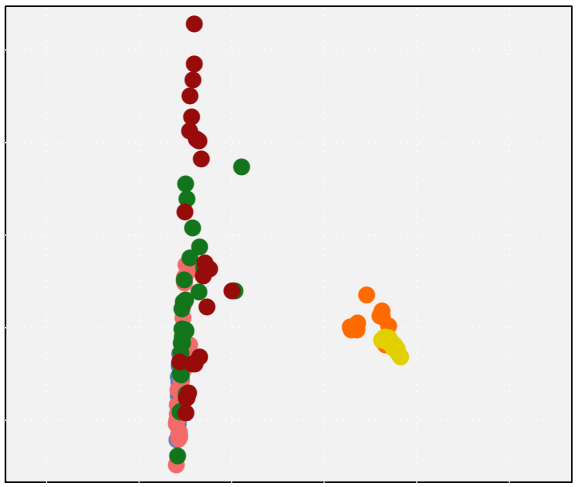
GRAVEYARD

Effect of dimensionality reduction is big

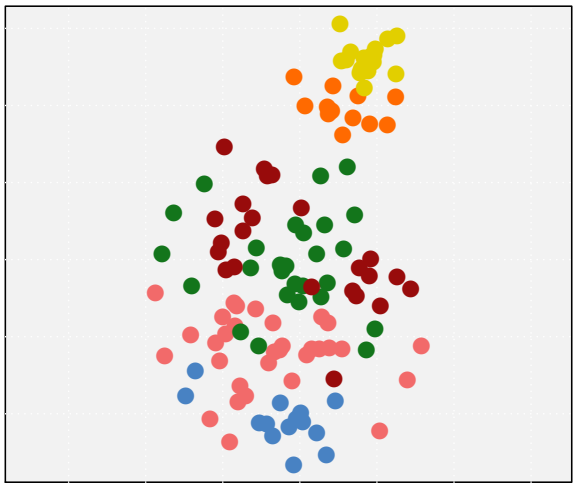
PCA



Laplacian Embedding

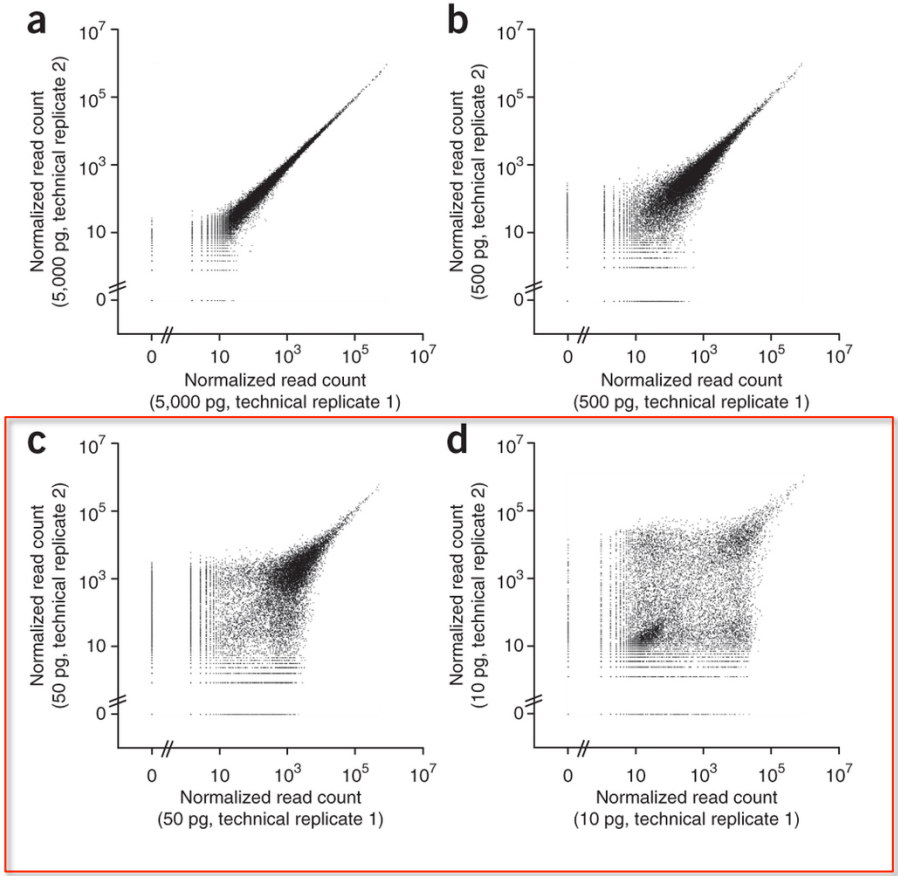


tSNE



Limitations: Noisy data

Dilution of Bulk RNA



Brennecke et al Nature Methods (2013)

Because of low starting input (picograms), large amounts of amplification, other technical problems