Ph. D. Qualifying Examination

Committee:
Elizabeth Purdom (chair)
Sandrine Dudoit
John Ngai
Maya Petersen

Hector Roux de Bézieux
Group in Biostatistics
Sandrine Dudoit's lab
GitHub: HectorRDB
Website: http: //hectorrdb. github. io

Overview

I/ scRNA–Seq: an essential tool for biology

II/ Differential Expression with `tradeSeq`

III/ Improving cluster replicability with **Dune**

I/ Single cell RNA–Sequencing: an essential tool for biology

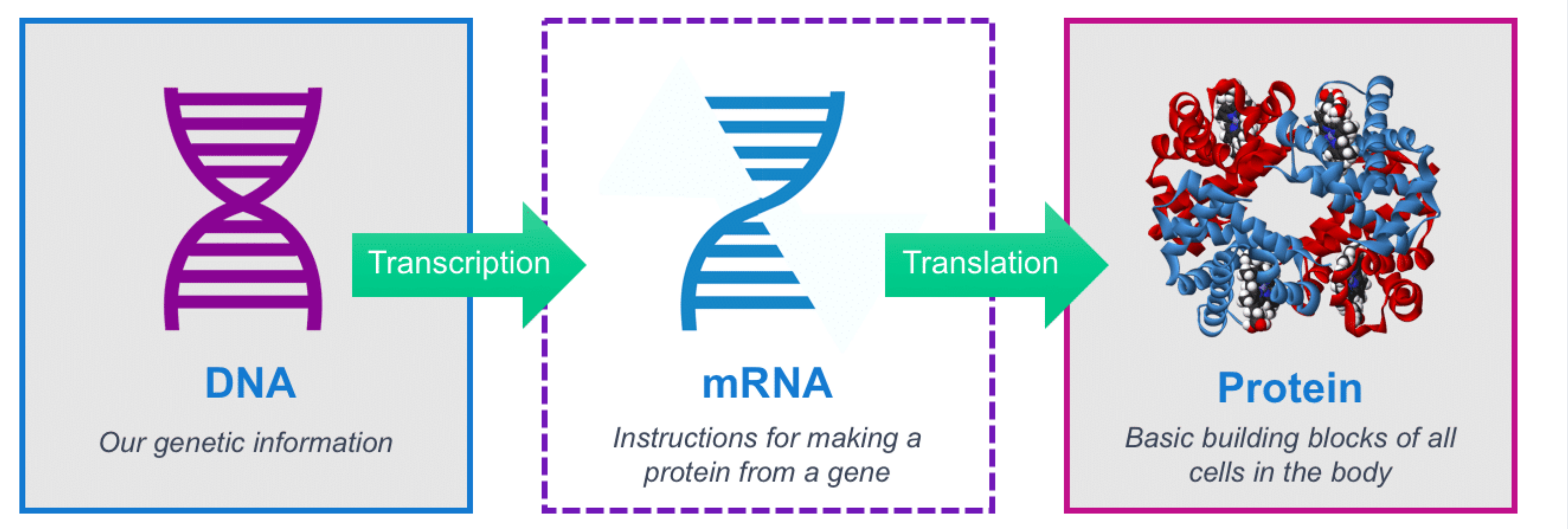
i/ A quick intro to scRNA–Seq technology

ii/ Datasets used for this presentation

iii/ Trajectory inference and **Slingshot**

I. i/ Single cell RNA-Sequencing: an essential tool for biology

Central Dogma of biology



DNA

Our genetic information

mRNA

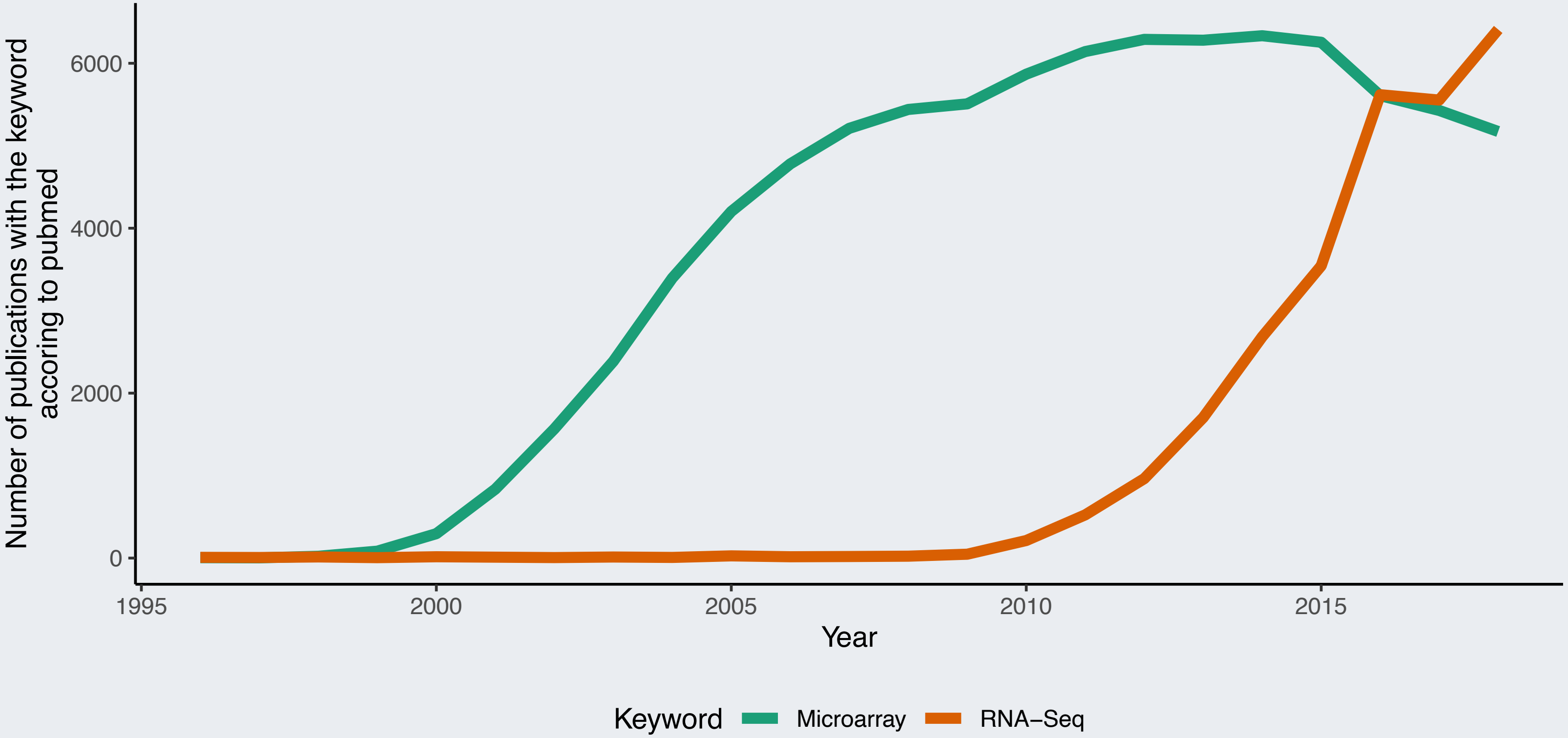
Instructions for making a protein from a gene

Protein

Basic building blocks of all cells in the body

https://translate.bio/rna-therapeutics/central-dogma-for-web-4-3/

From micro-array to bulk RNA-Seq



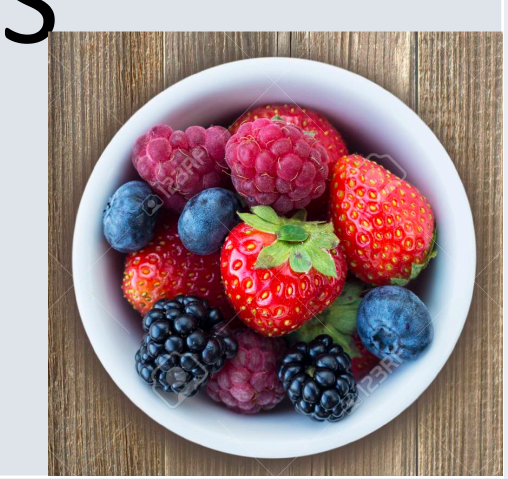
- Sequencing the mRNAs aims to capture gene expression level, mostly as a proxy for protein levels

- RNA-Seq enables whole transcriptomic sequencing without *a-priori* need for a reference genome

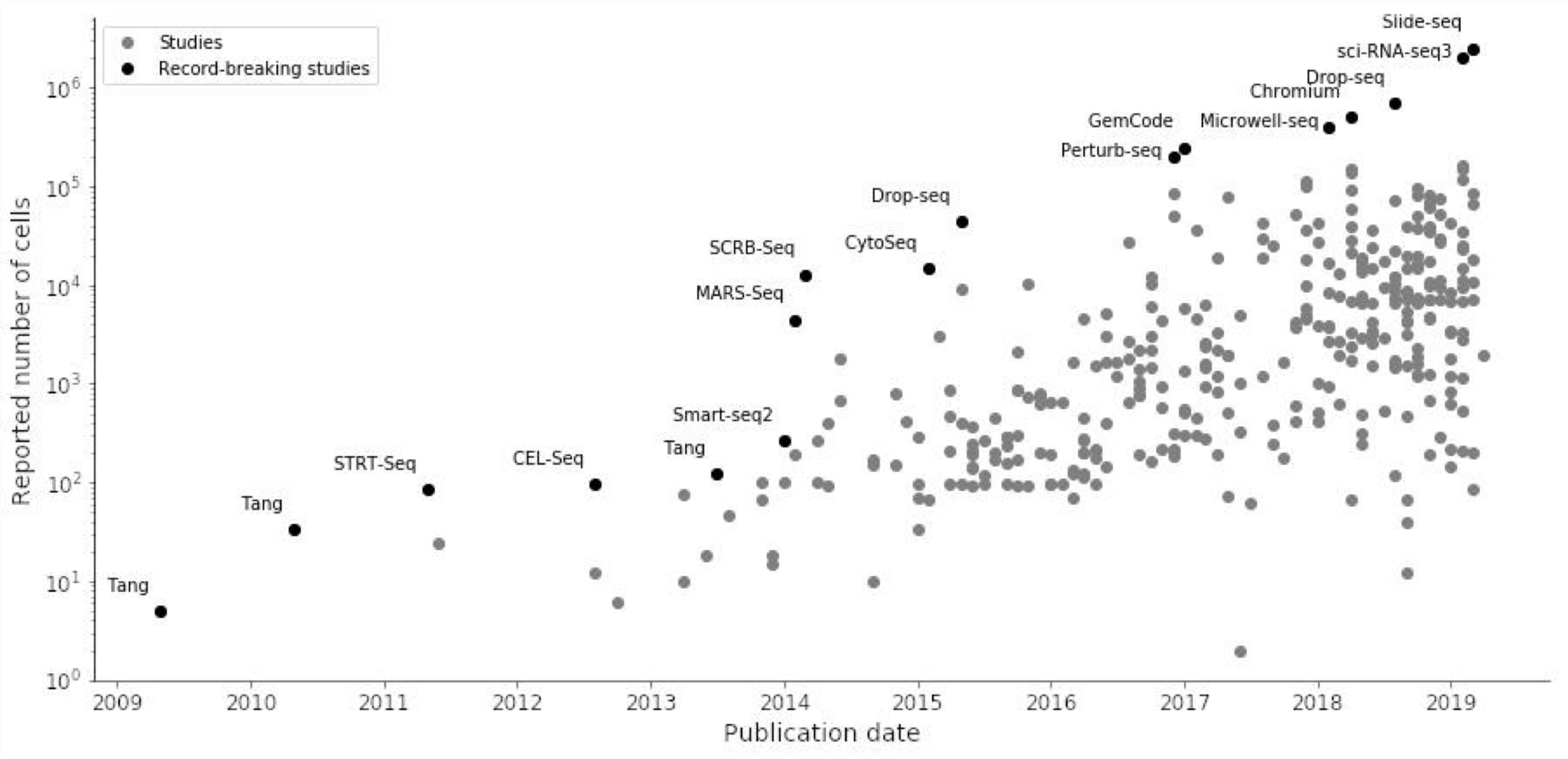
Data from PubMed taken on 11/10/2019

Single–cell RNA–Seq

Unmixing the smoothie

Bulk RNA - Seq	Single–cell RNA - Seq
 VS 	 VS 

Recent explosion in scRNA–Seq

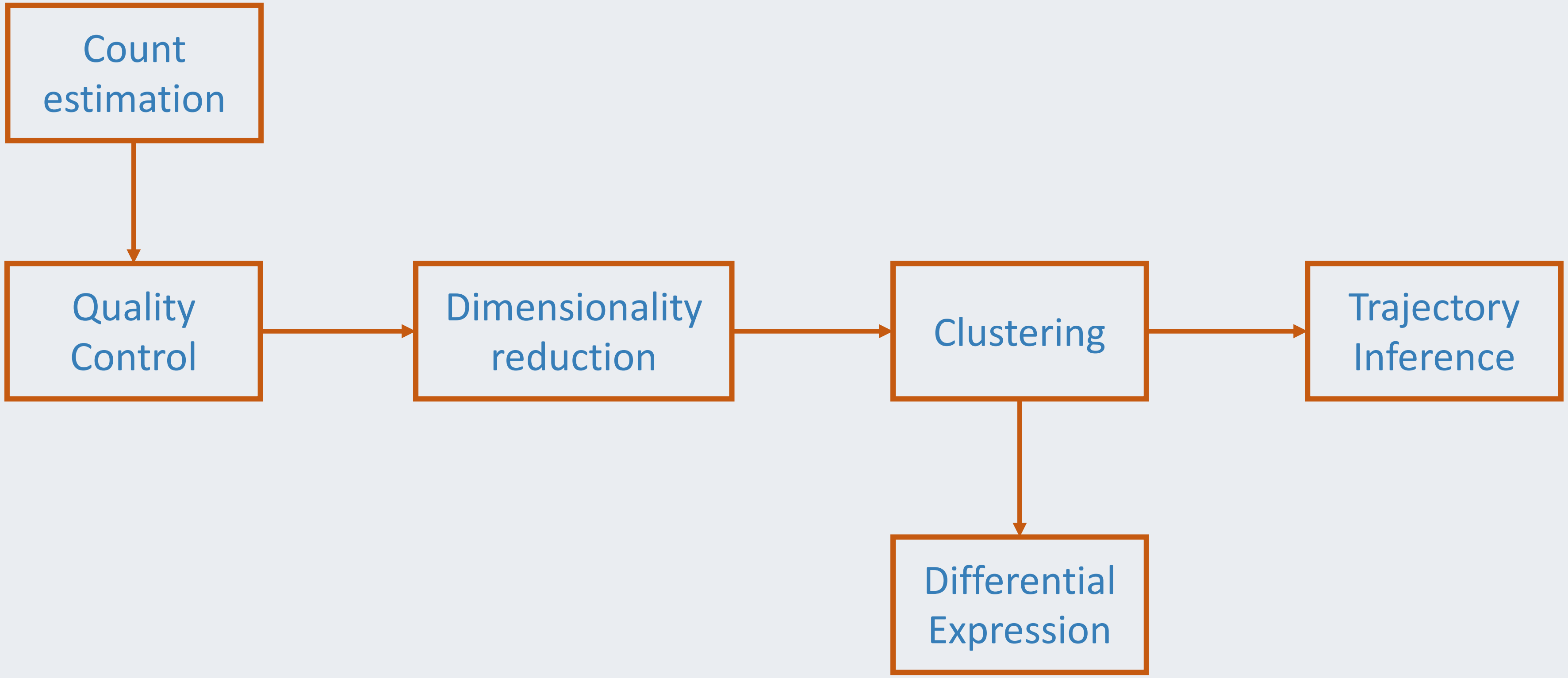


https://twitter.com/vallens/status/1113982015517282304

Data structure

	Cell 1	***Cell 2***	***Cell 3***	**…**	***Cell n***
Gene 1	0	28	25	…	2
Gene 2	0	3	8	…	36
Gene 3	5	0	0	…	0
…	…	…	…	…	…
Gene G	12	8	0	…	11

Common workflow



Count estimation

Quality Control

Dimensionality reduction

Clustering

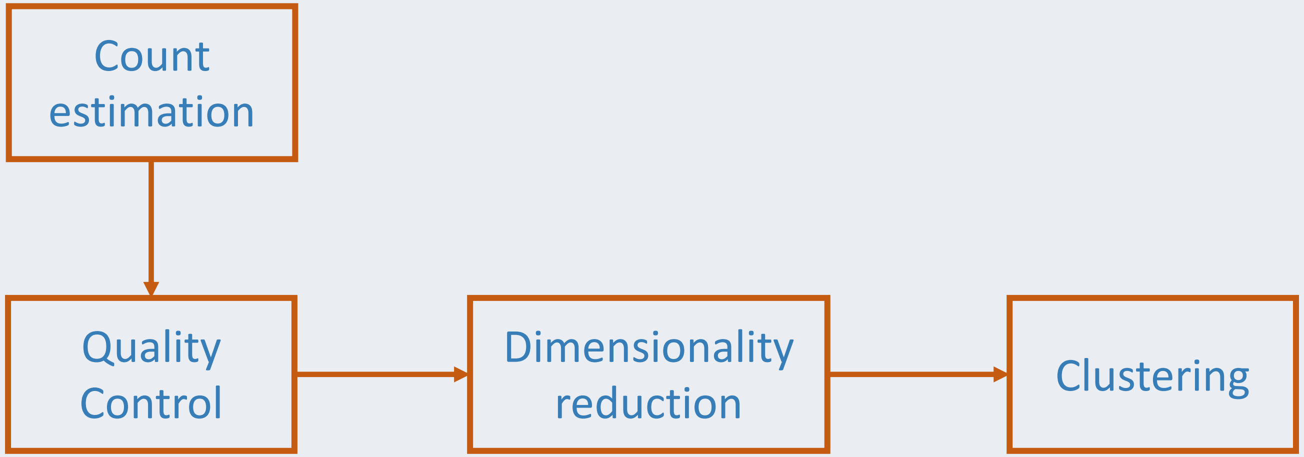
Trajectory Inference

Differential Expression

I. ii/ Case studies and example datasets

Dataset characteristics

- Number of cells
- Number of genes
- Low-dimensionality representation colored by clusters



Count estimation

Quality Control

Dimensionality reduction

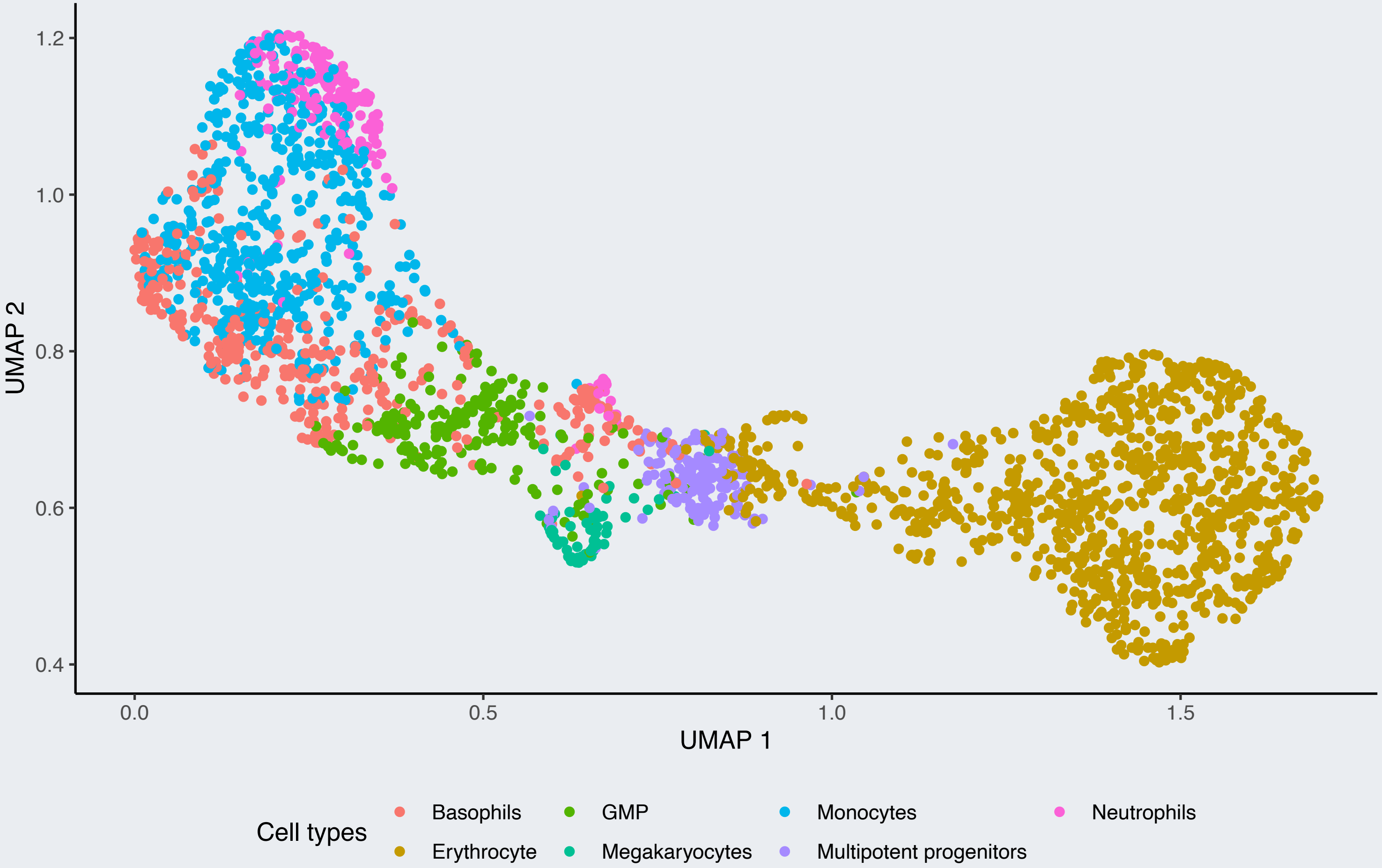
Clustering

Embryogenesis

- Embryogenesis dataset from Deng et al
- 258 cells and 13179 genes
- Dimensionality reduction with PCA
- Clusters from the original publication

Deng, Q., Ramsköld, D., Reinius, B., & Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*. https://doi.org/10.1126/science.1245316

Bone marrow stem cells



Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv
, 2 2018. URL
http: //arxiv. org/abs/1802. 03426

- Bone-marrow stem cells from the [monocle 3 vignette]
- 2660 cells and 3004 genes
- Dimensionality reduction with UMAP
- Clusters from the original publication

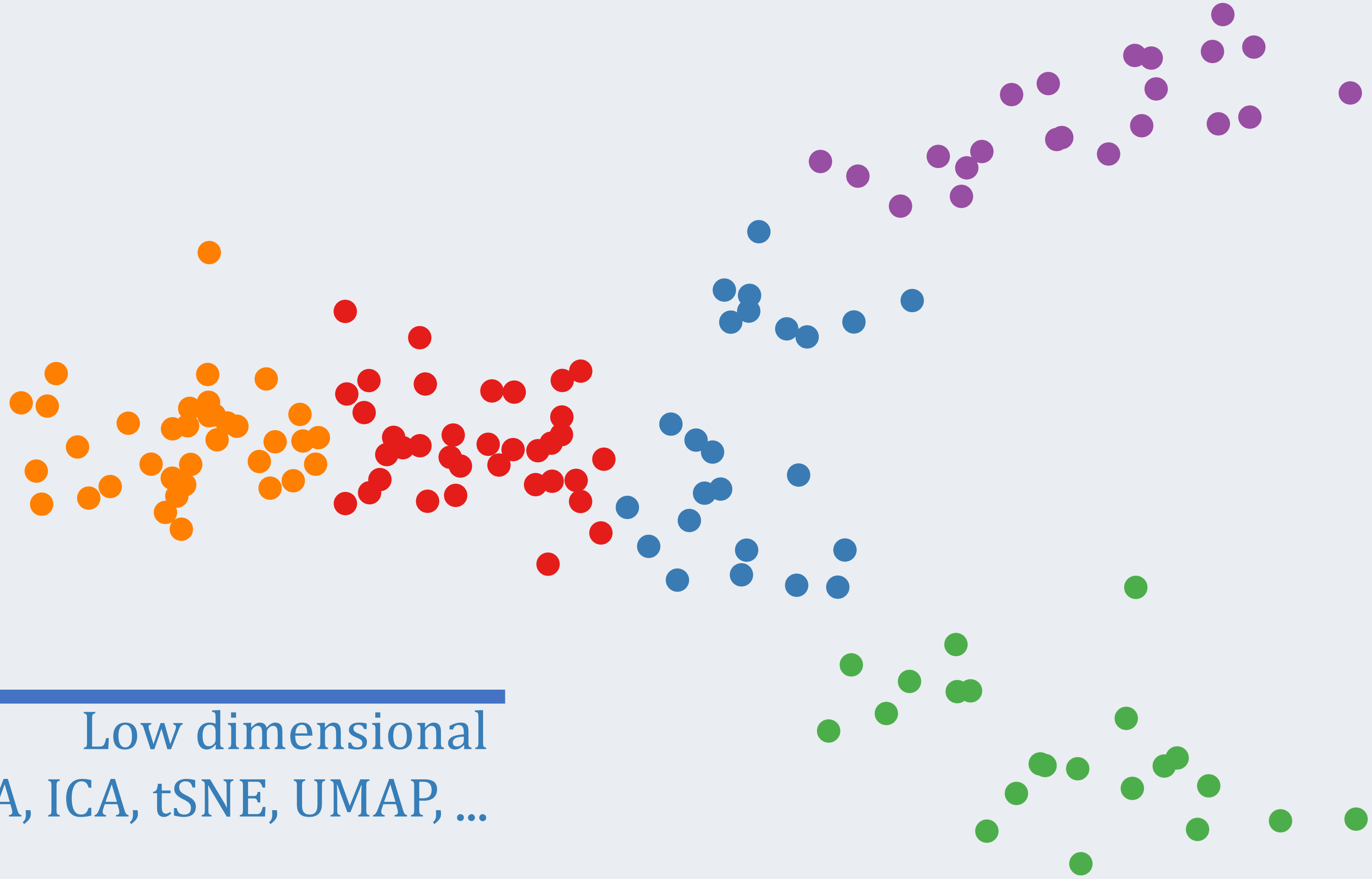
I. iii/ Trajectory inference with **Slingshot**

Motivation

- It is possible to distinguish a trajectory in the reduced space that tracks biological development
- Trapnell et al. introduces the concept of pseudotime in 2014

Trapnell, C. , Cacchiarelli, D. , Grimsby, J. , Pokharel, P. , Li, S. , Morse, M. , ... Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. https: //doi. org/10. 1038/nbt. 2859

Input Data



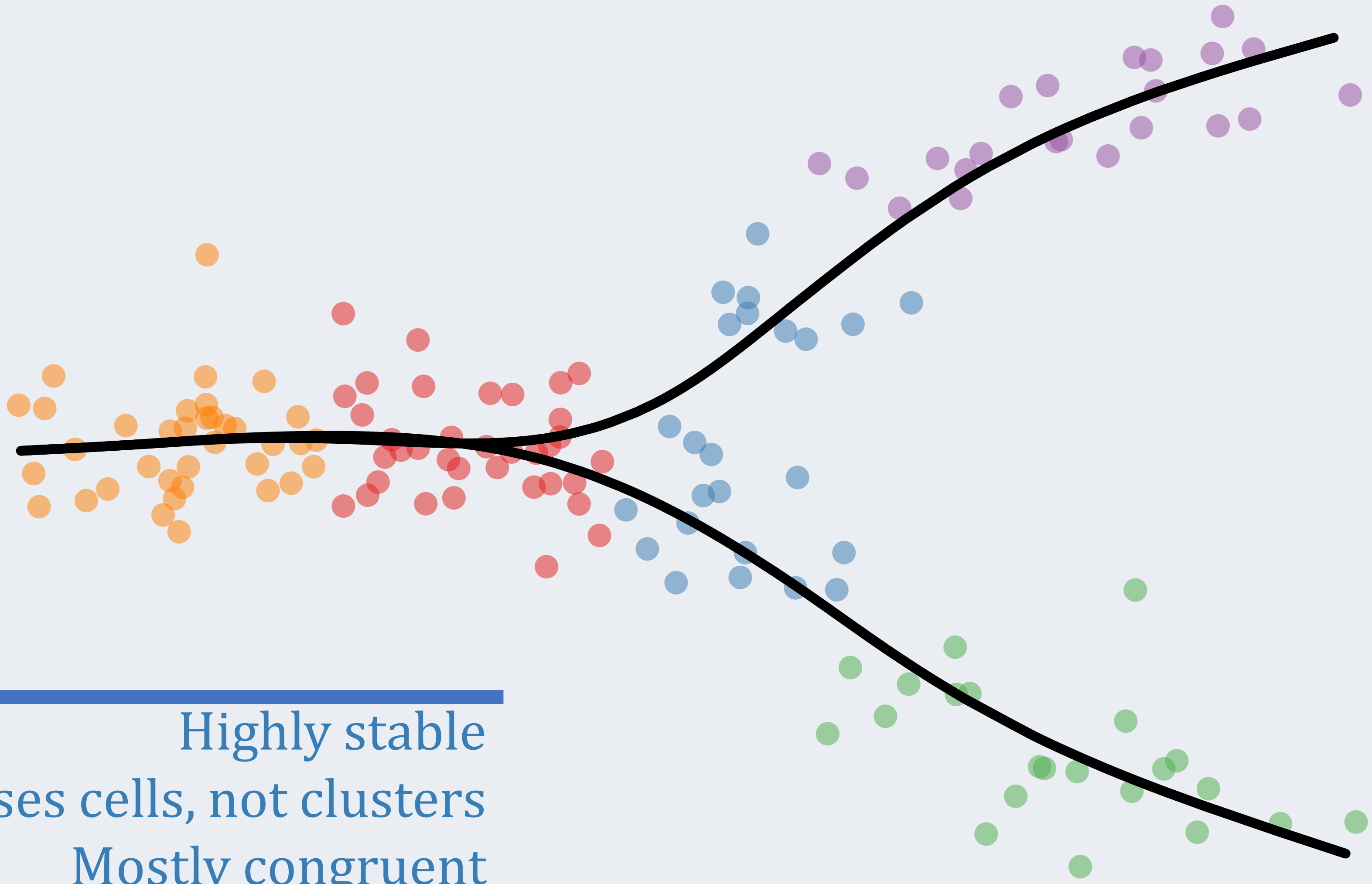
Low dimensional

PCA, ICA, tSNE, UMAP, ...

Clustered

SC3, Seurat, RSEC, ...

Simultaneous Principal Curves

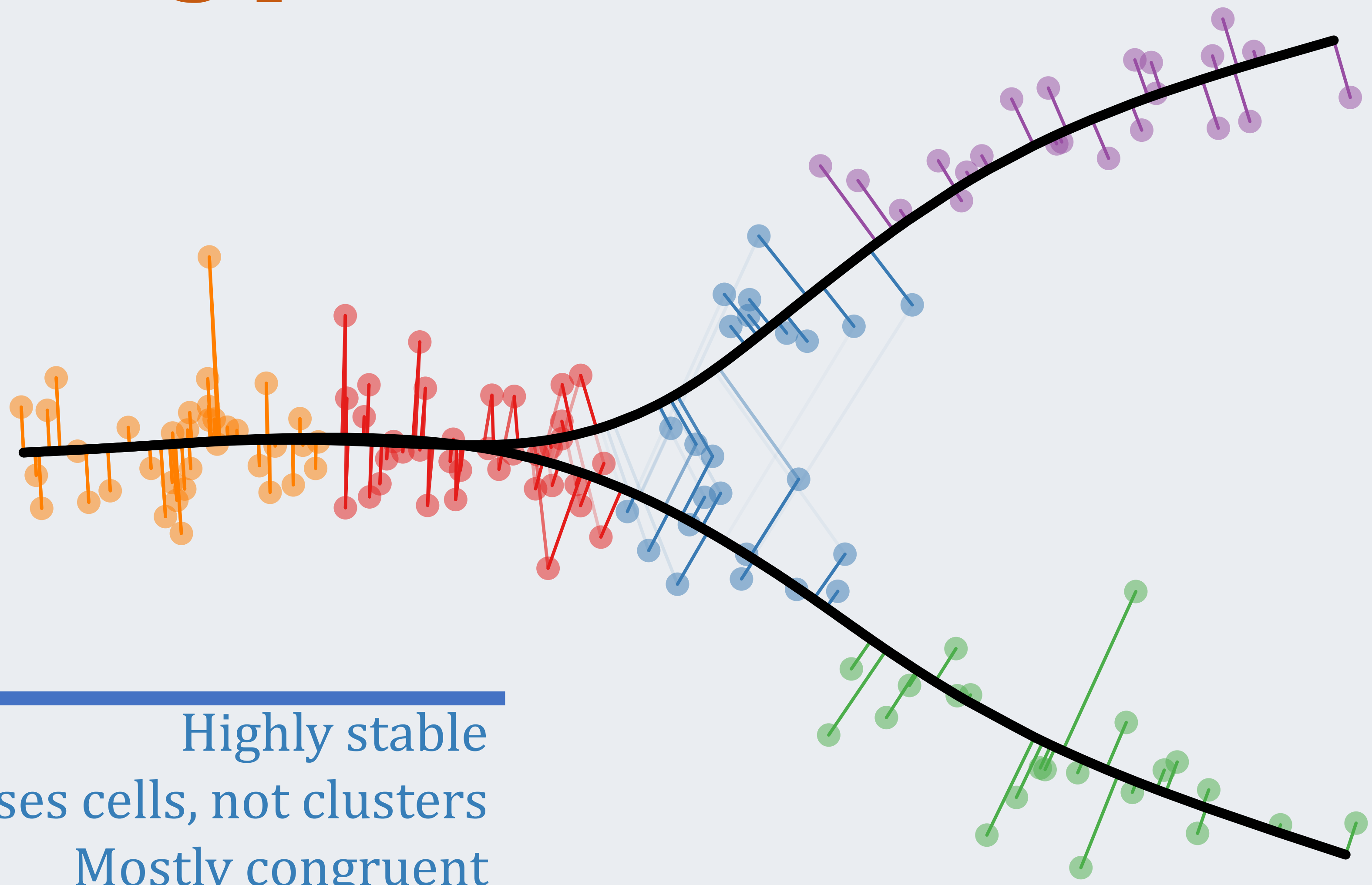


Highly stable

Uses cells, not clusters

Mostly congruent across branches

Computing pseudotimes



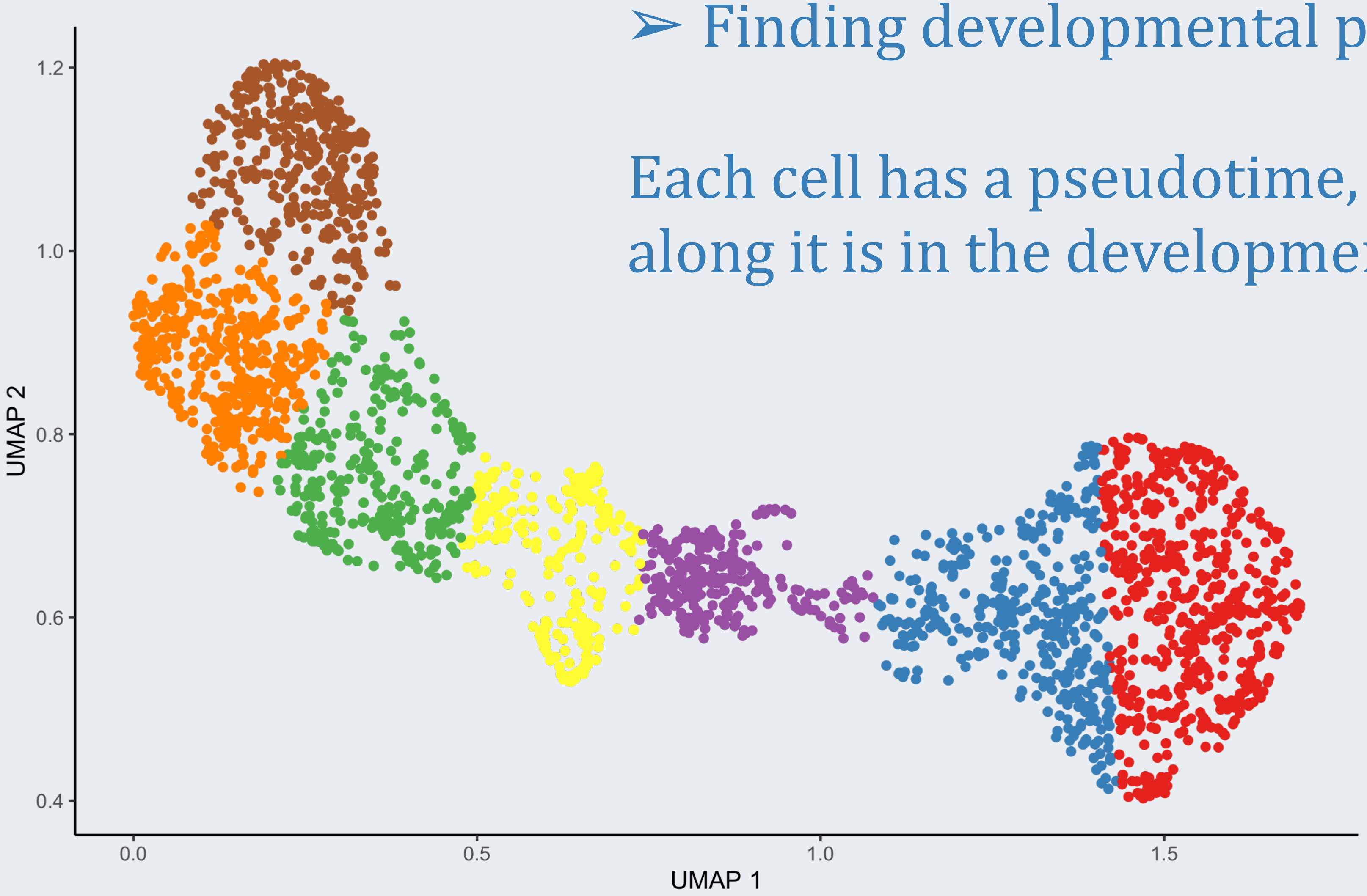
Highly stable

Uses cells, not clusters

Mostly congruent across branches

Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef
,Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime
inference for single-cell transcriptomics.
BMC Genomics
, 19(1): 477, 12 2018. ISSN 1471-2164. doi: 10. 1186/S12864-018-4772-0. URL
https: //bmcgenomics. biomedcentral. com/articles/10. 1186/S12864-018-4772-0

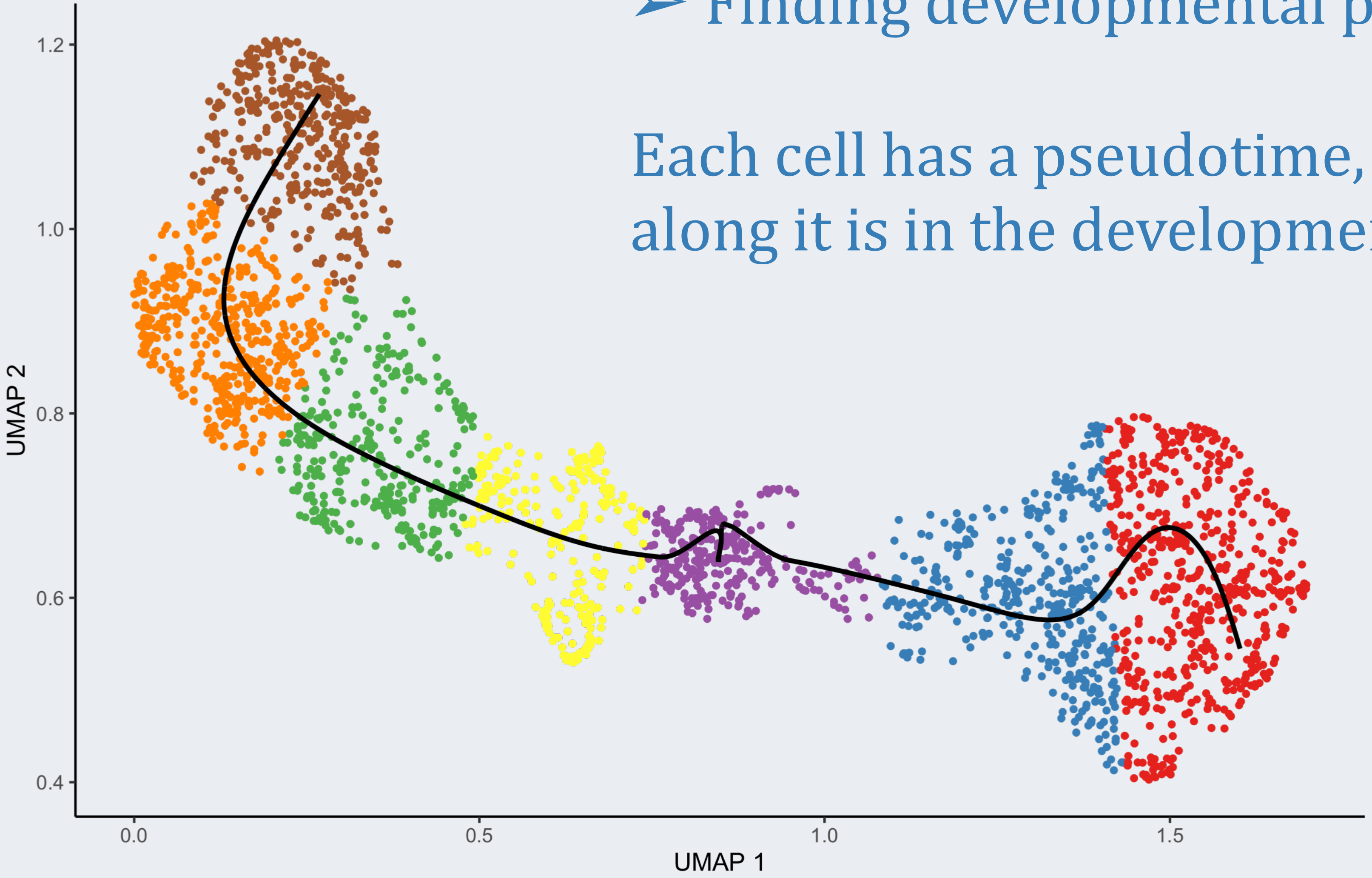
Application to bone–marrow



➢ Finding developmental paths

Each cell has a pseudotime, which measure how far along it is in the developmental process

Application to bone–marrow



- ➢ Finding developmental paths

Each cell has a pseudotime, which measure how far along it is in the developmental process

II/ Differential Expression with `tradeSeq`

i/ Motivation

ii/ Statistical framework

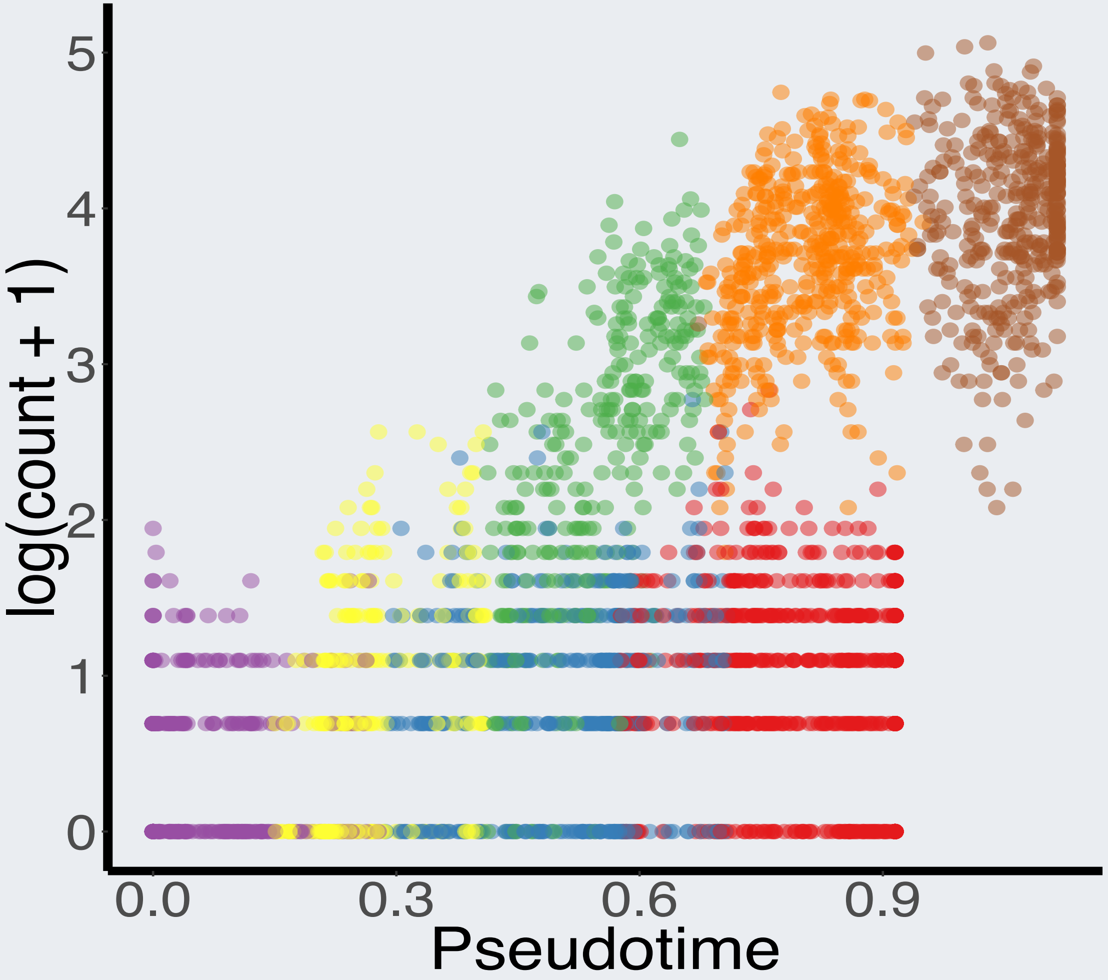
iii/ Results

II. i/ Motivation

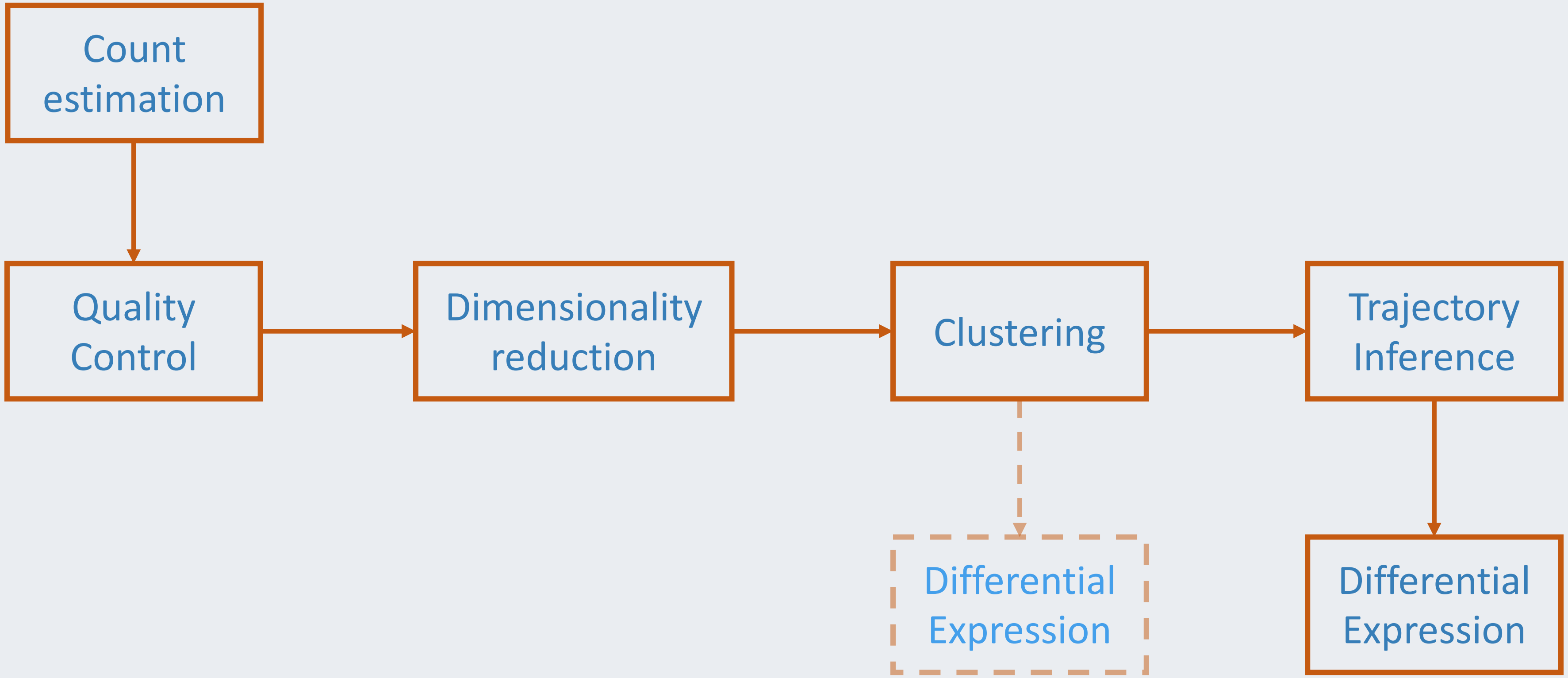
cluster–based DE is artificial

Genes are now expressed in a continuous manner (since 2014)

Differential Expression is still cluster–based, i. e. discrete.



Trajectory based Differential Expression



Count estimation

Quality Control

Dimensionality reduction

Clustering

Trajectory Inference

Differential Expression

Differential Expression

Trajectory-based DE

We developed `tradeSeq`, an algorithm that leverages the continuous nature of scRNA-Seq.

- Available as an **R** package on Bioconductor.
- Modular tool that work with any dimensionality reduction and trajectory inference method.

II. ii/ Statistical framework

Data structure

$Y =$

	Cell 1	**Cell 2**	**Cell 3**	**…**	**Cell n**
Gene 1	0	28	25	…	2
Gene 2	0	3	8	…	36
Gene 3	5	0	0	…	0
…	…	…	…	…	…
Gene G	12	8	0	…	11

n cells per G genes

$T = [T_0, .., T_n] \subset (\mathbb{R}^L)^n$

Pseudotimes for each cell

$Z = [Z_0, .., Z_n] \subset [0{:}\,1]^{L \times n}$

Lineage assignment weights

Statistical model

$$Y_{gi} = NB(\mu_{gi}, \phi_g)$$

Negative–binomial model common for RNASeq count data

Sample and gene–specific mean

Gene–specific dispersion parameter

Statistical model

$$Y_{gi} = NB(\mu_{gi}, \phi_g)$$

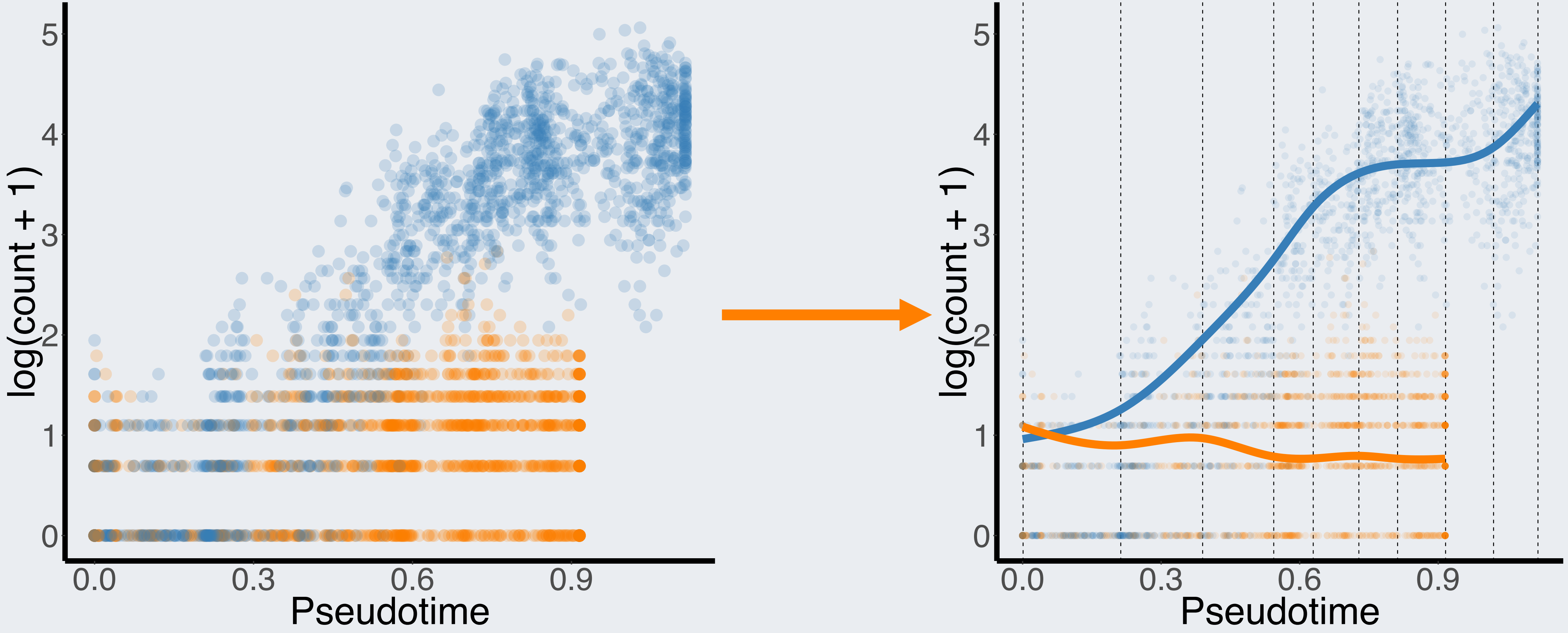
$$\mu_{gi} = \boldsymbol{U}_i \alpha_g + \log(N_i) + \sum_{l=1}^{L} s_{gl}(T_i)\, Z_{li}$$

Can accommodate:

Design matrix

Different sequencing depths

Statistical model



Fitting the smoothers

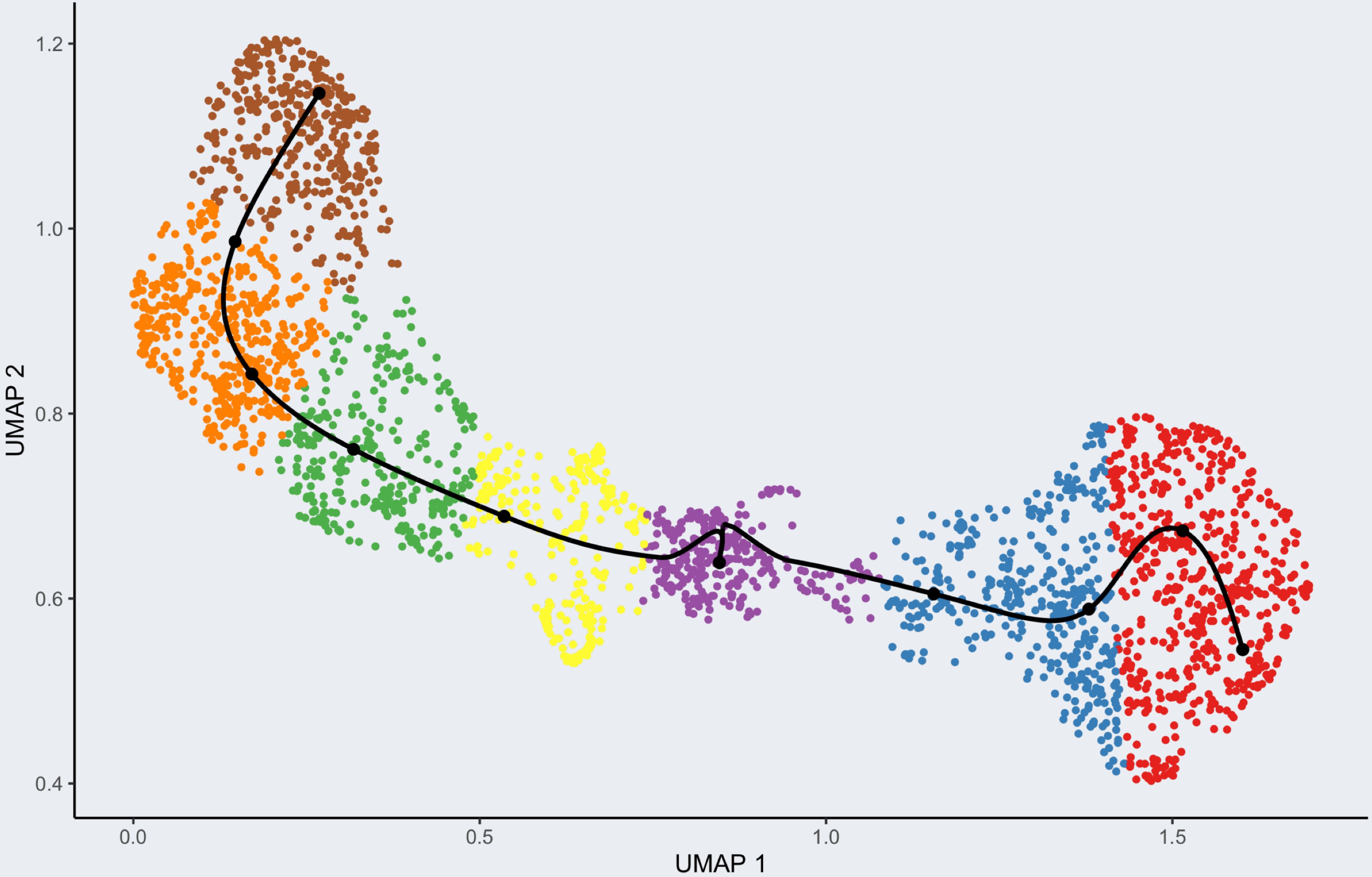
$$\mu_{gi} = \boldsymbol{U}_i \alpha_g + \log(N_i) + \sum_{l=1}^{L} s_{gl}(T_i)\, Z_{li}$$

$$s_{gl}(T_i) = \sum_{k=1}^{K} b_k(t) \beta_{glk}$$

We rely on recent implementations for fitting smoothers in the **mgcv** package

Wood S.N., N. Pya and B. Saefken (2016)
Smoothing parameter and model selection
for general
smooth models (with discussion). Journal
of the American Statistical Association
111:1548-1575.

Knots location



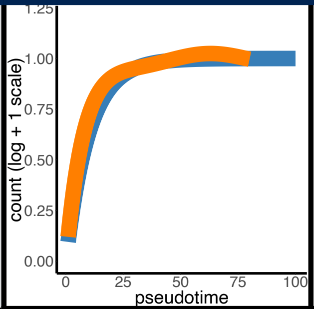
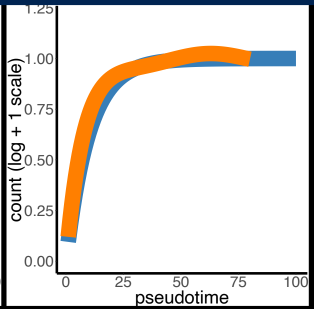
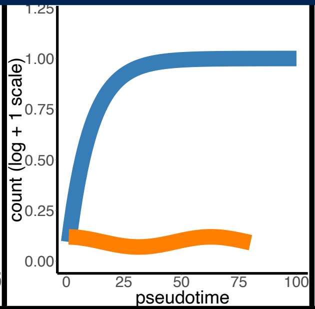
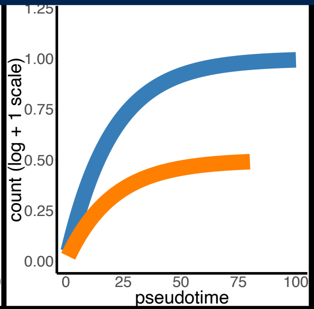
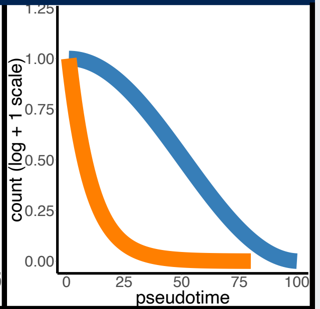
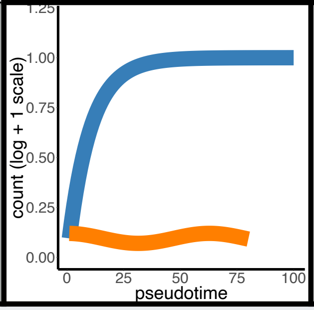
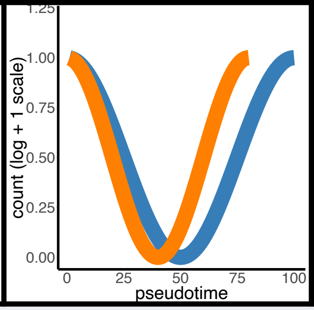
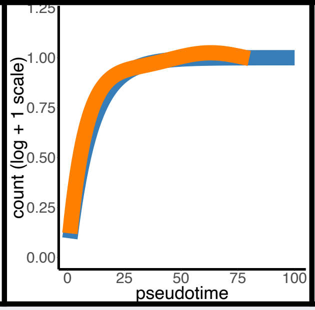
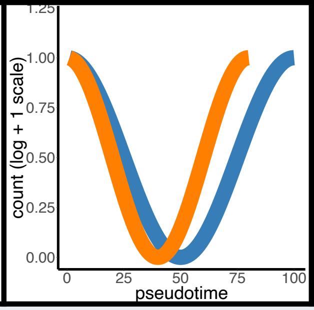
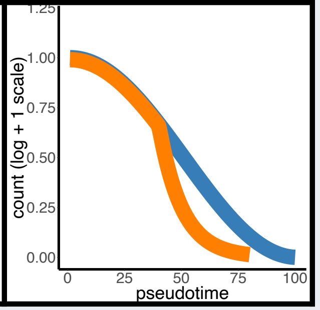
Testing framework

$$s_{gl}(T_i) = \sum_{k=1}^{K} b_k(t)\beta_{glk}$$

Testing null hypotheses of the form: $H_0\colon \boldsymbol{C}^T\beta_g = 0$

Using the Wald Statistics $W_g = \boldsymbol{C}^T\hat{\beta}_g\left(\boldsymbol{C}^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{g}}\boldsymbol{C}\right)^{-1}\hat{\beta}_g^T\boldsymbol{C}^{-T}$

An investigation tool

	Within the orange lineage		Between the orange and blue lineages		
Results	association Test	startVsEnd Test	diffEnd Test	pattern Test	earlyDE Test
DE					
Not DE					

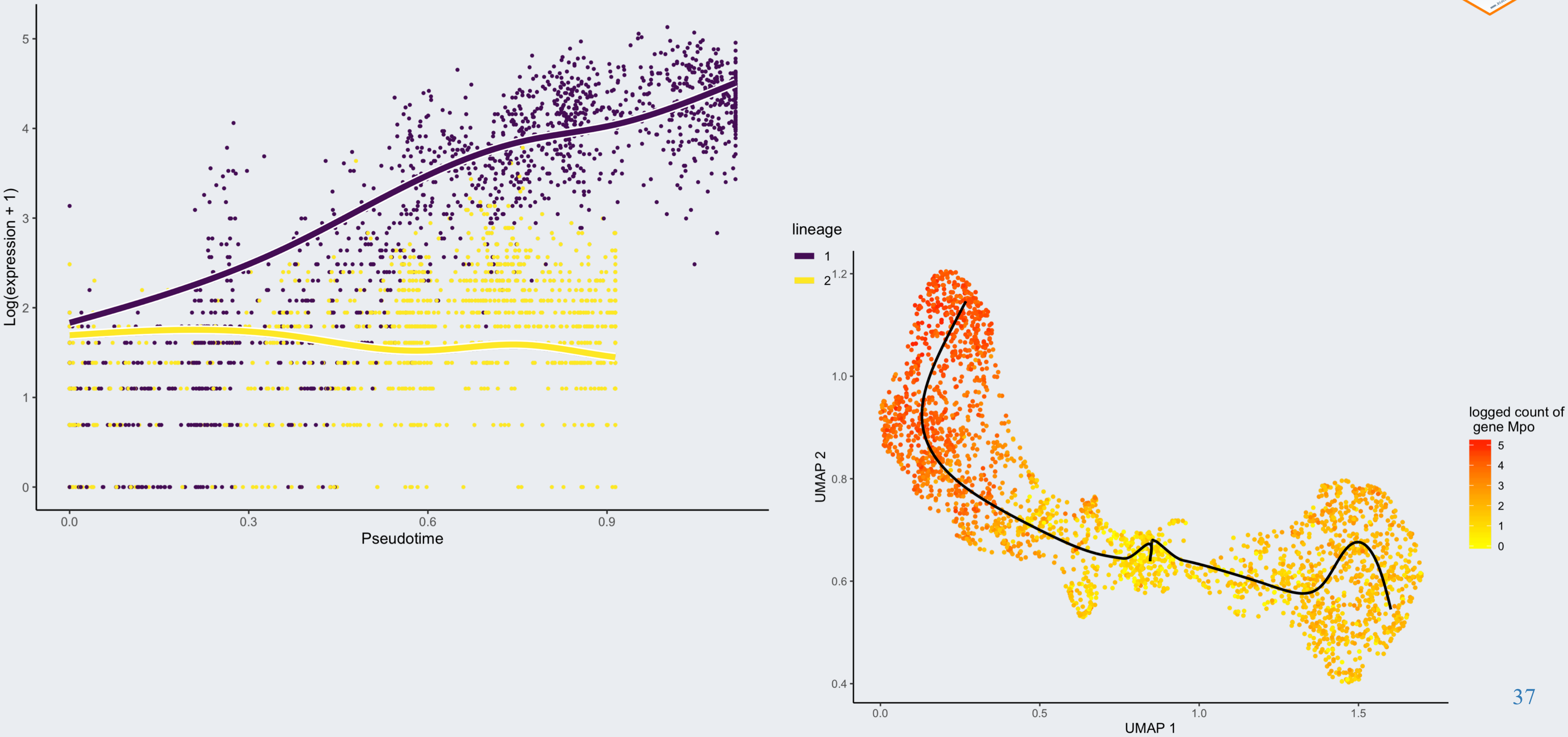
Association test

$$H_0: \beta_{lkg} = \beta_{lk'g} \; for \; all \; k \neq k'$$

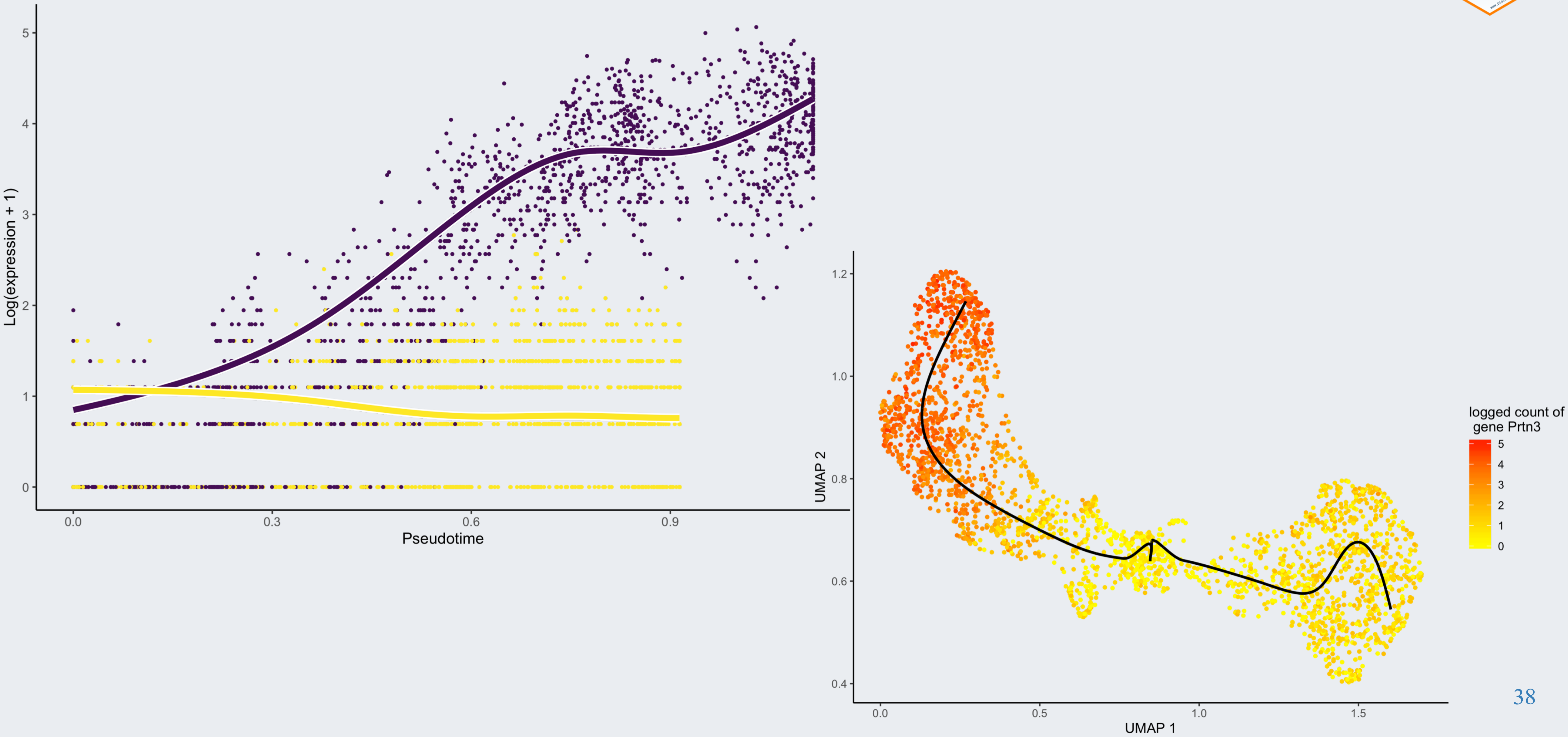
Contrast matrix

β_{l1g}	β_{l2g}	β_{l3g}	…	β_{lKg}
1	-1	0	…	0
0	1	-1	…	0
0	0	1	…	0
…	…	…	…	…
0	0	0	…	-1

StartVsEndTest

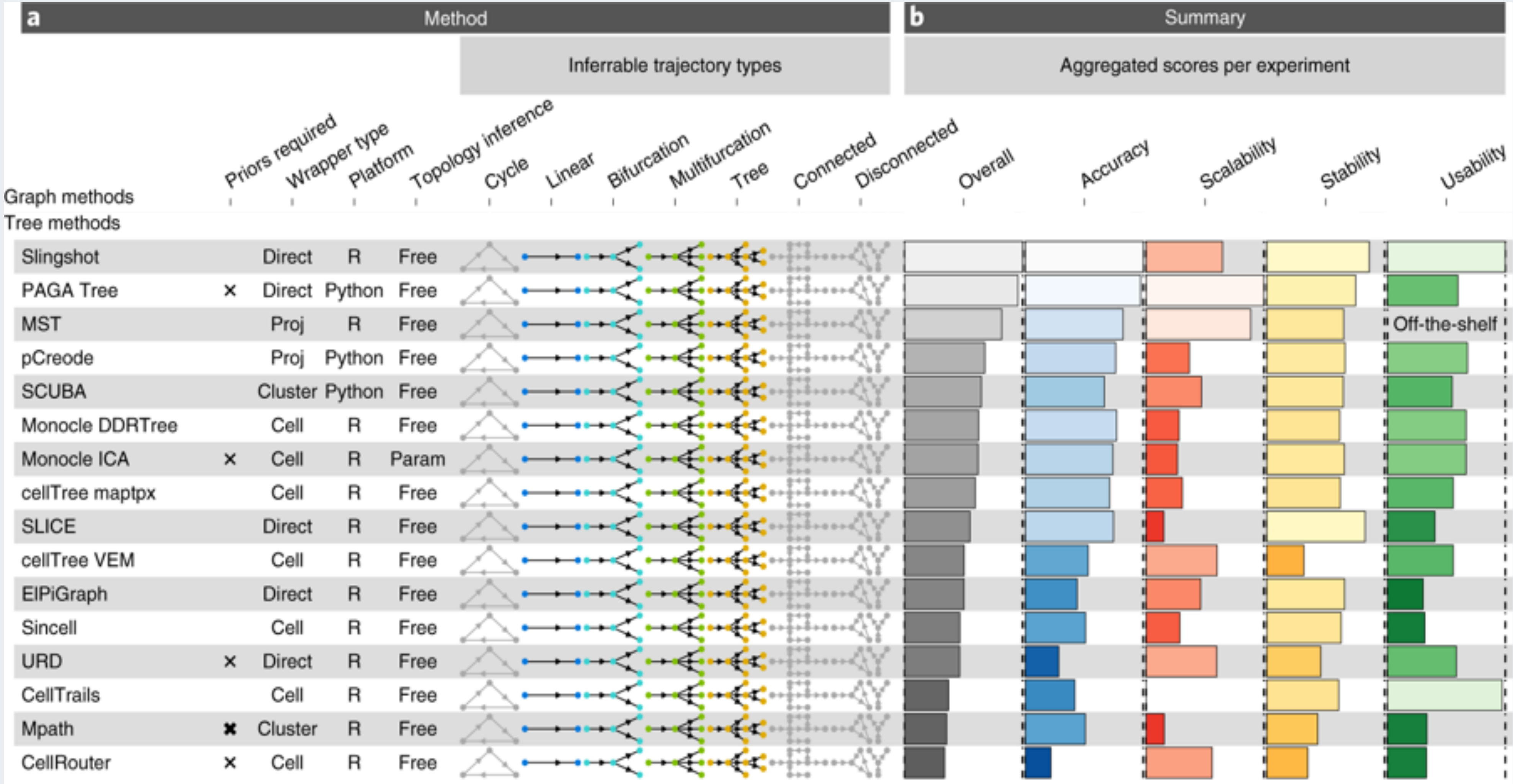


diffTest


Log(expression + 1)
5
4
3
2
1
0
0.0
0.3
0.6
0.9
Pseudotime
UMAP 2
1.2
1.0
0.8
0.6
0.4
0.0
0.5
1.0
1.5
UMAP 1
logged count of
gene Prtn3
5
4
3
2
1
0

II. iii/ Results

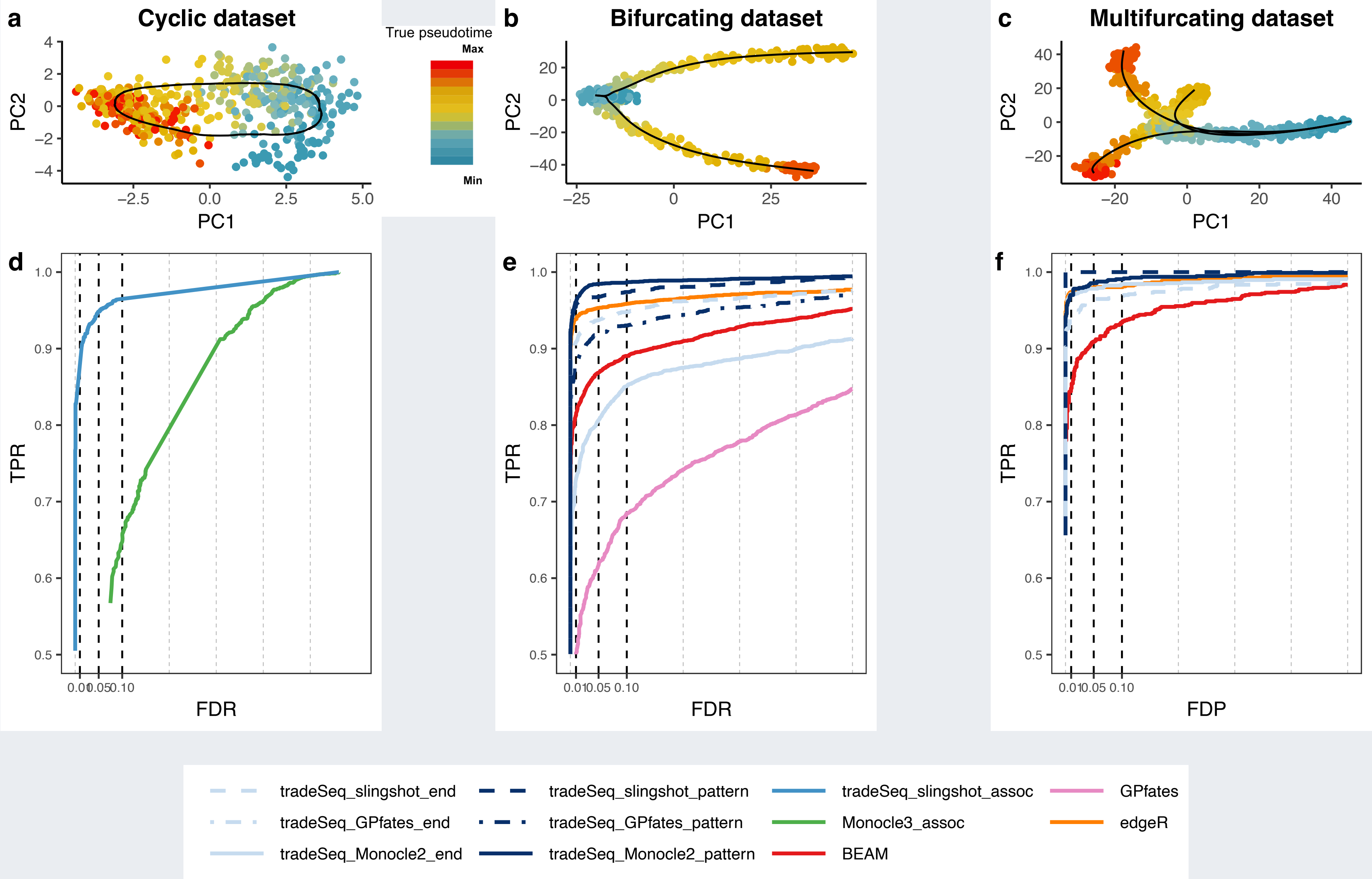
Simulation framework: dynverse



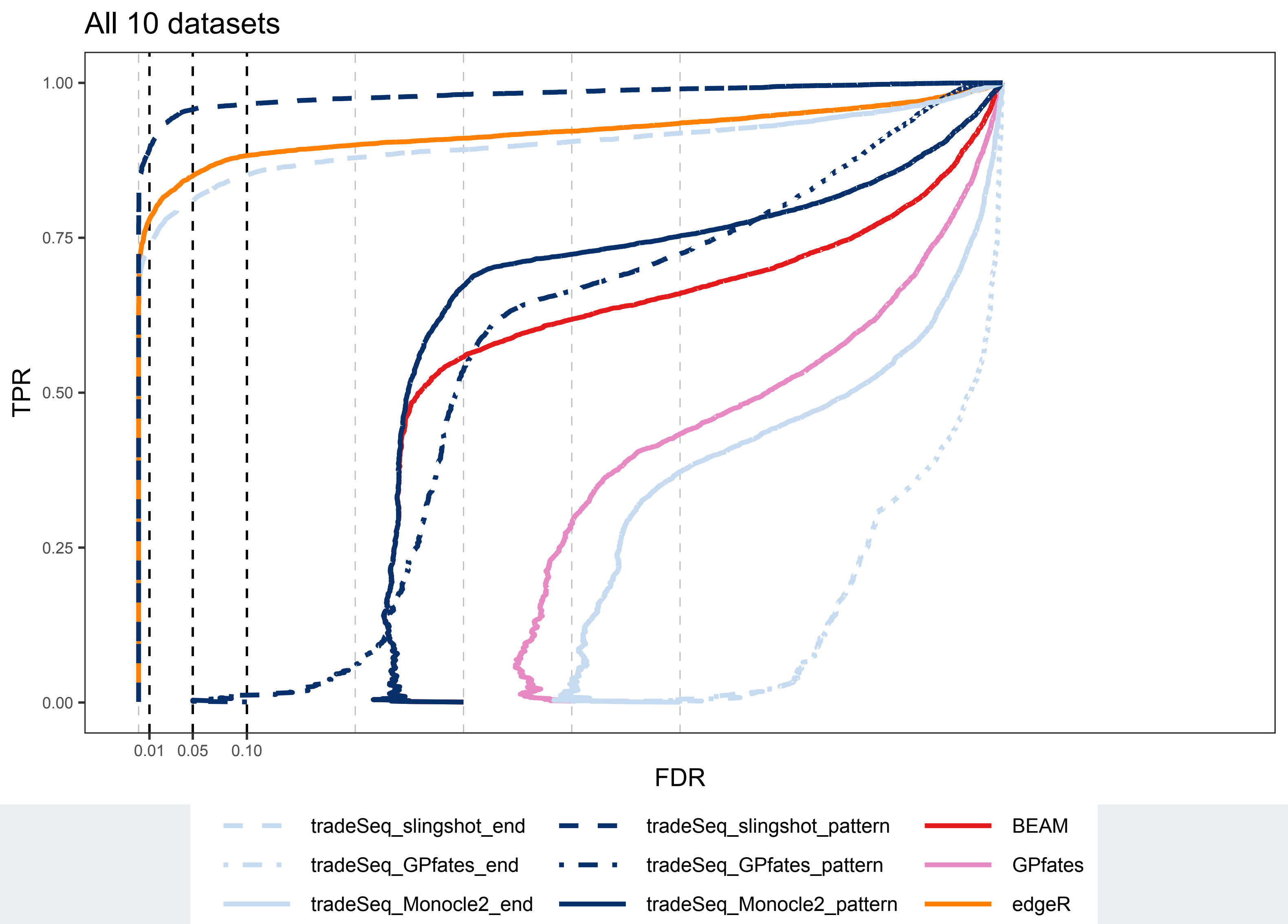
	Priors required	Wrapper type	Platform	Topology inference
Slingshot		Direct	R	Free
PAGA Tree	×	Direct	Python	Free
MST		Proj	R	Free
pCreode		Proj	Python	Free
SCUBA		Cluster	Python	Free
Monocle DDRTree		Cell	R	Free
Monocle ICA	×	Cell	R	Param
cellTree maptpx		Cell	R	Free
SLICE		Direct	R	Free
cellTree VEM		Cell	R	Free
ElPiGraph		Direct	R	Free
Sincell		Cell	R	Free
URD	×	Direct	R	Free
CellTrails		Cell	R	Free
Mpath	✖	Cluster	R	Free
CellRouter	×	Cell	R	Free

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. Acomparisonofsingle–cell trajectory inference methods. NatureBiotechnology , page 1, 4 2019. ISSN 1087–0156. doi: 10. 1038/ s41587–019–0071–9. URL http: //www. nature. com/articles/s41587 –019–0071–9

Outperforms existing methods



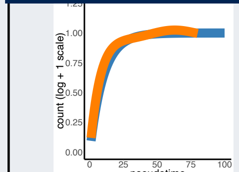
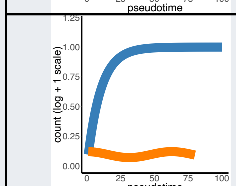
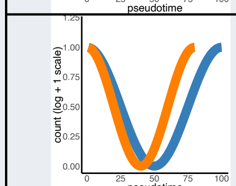
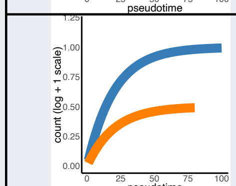
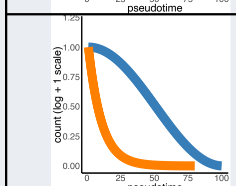
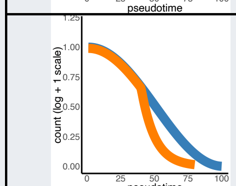
Outperforms existing methods



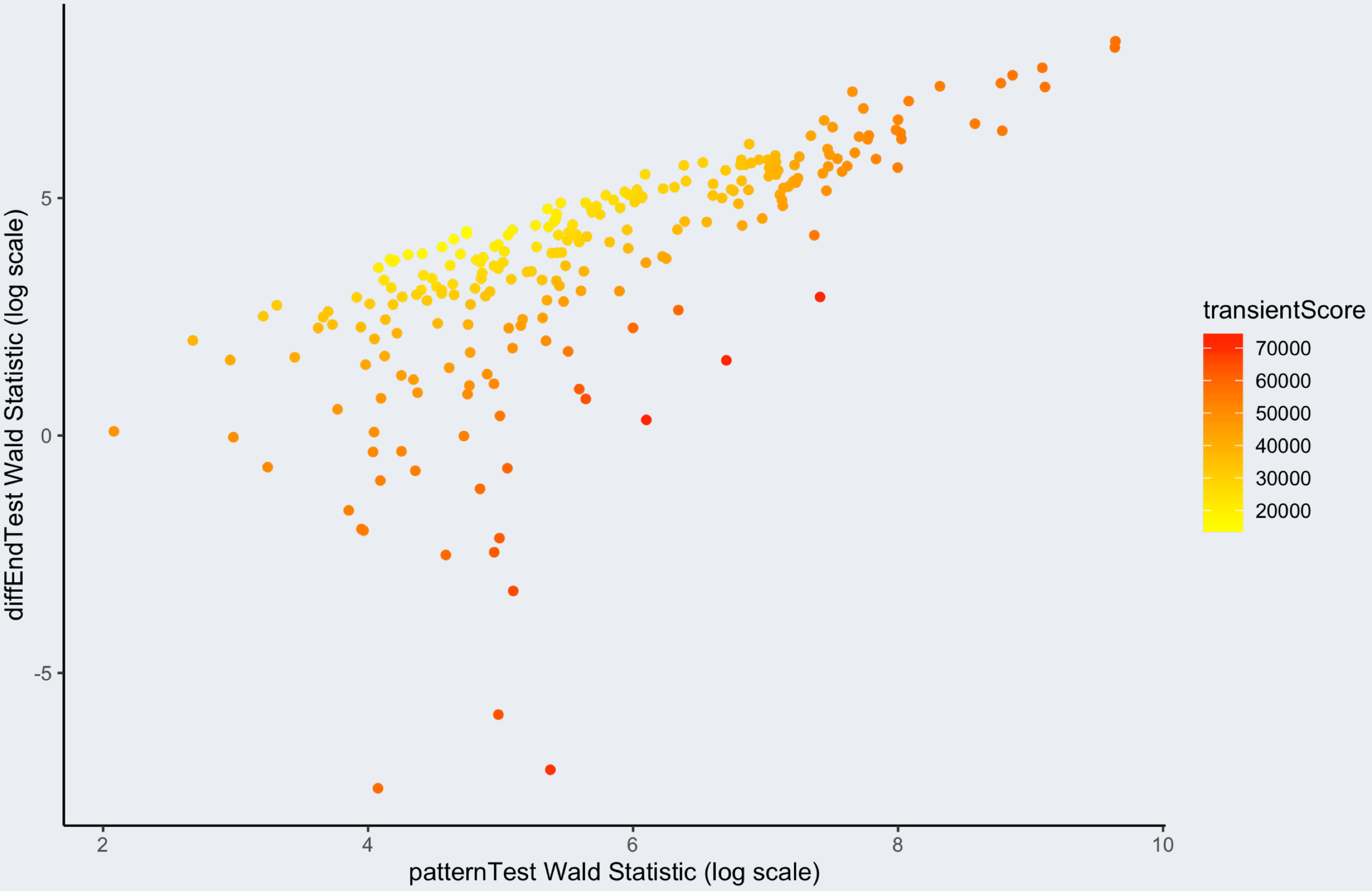
Provides unique insights

Differential Expression Tests

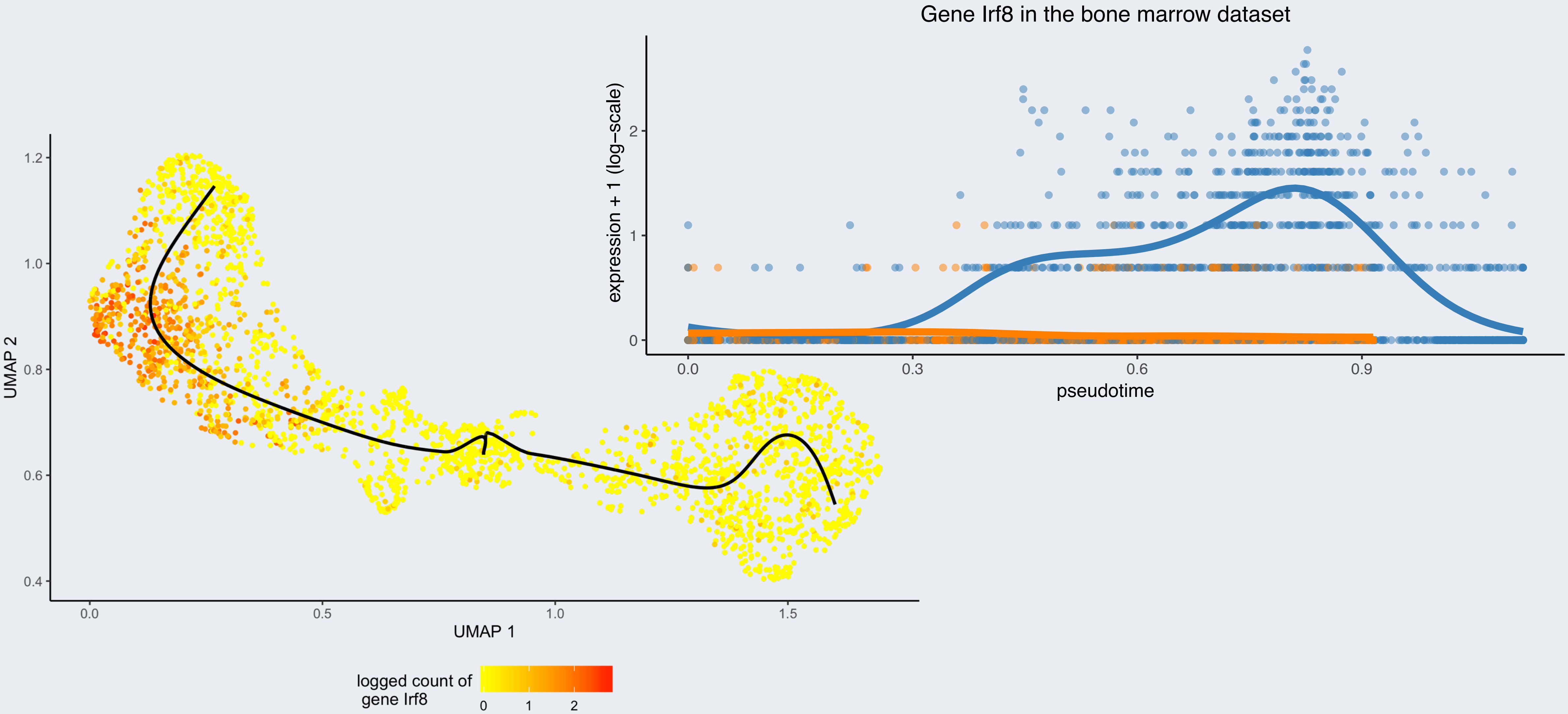
c)

Lineages	Within the orange lineage		Between the orange and blue lineages		
	association Test	**startVsEnd Test**	**diffEnd Test**	**pattern Test**	**earlyDE Test**
	DE	DE	Not DE	Not DE	Not DE
	Not DE	Not DE	DE	DE	DE
	DE	Not DE	Not DE	Not DE	Not DE
	DE	DE	DE	DE	Not DE
	DE	DE	Not DE	DE	DE
	DE	DE	Not DE	DE	Not DE

Provides unique insights



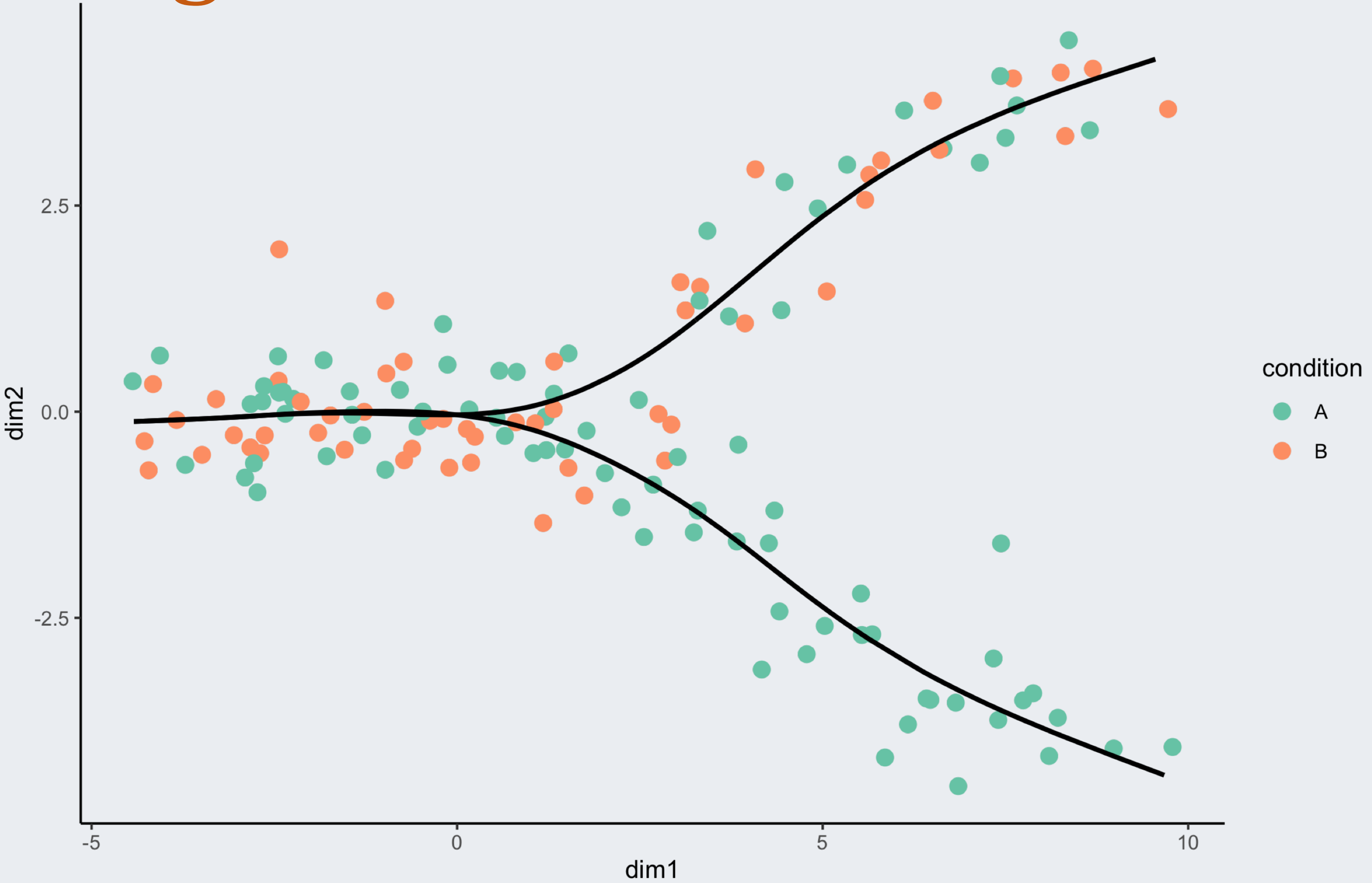
Provides unique insights


Gene Irf8 in the bone marrow dataset

Perspectives for tradeSeq

- Possible to develop new tests, especially to look at speed or acceleration of gene changes.
- Expand the framework to test lineage × condition interaction
- Publish the paper

Lineage × condition



III/ Improving cluster replicability with **Dune**

i/ Motivation

ii/ Datasets

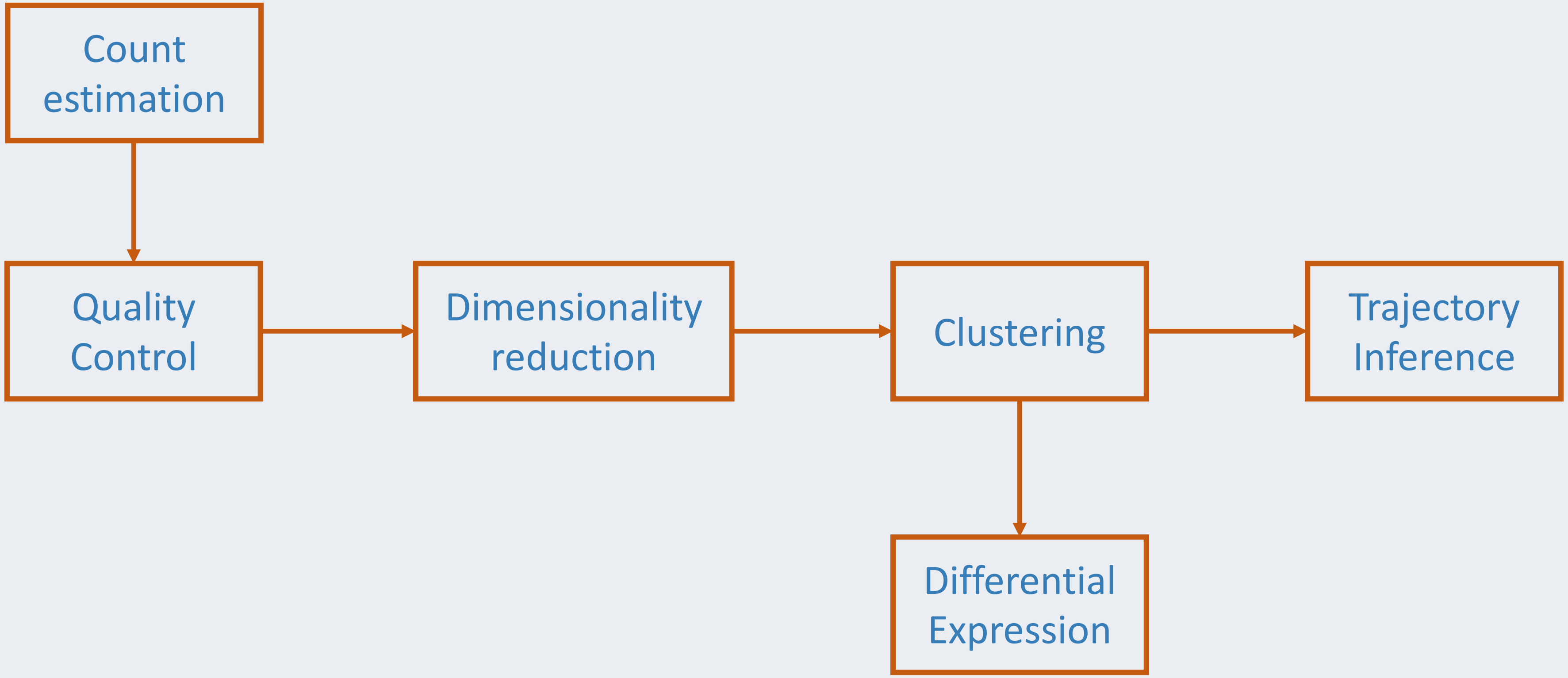
iii/ Measuring replicability

iv/ Method

v/ Results

III. i/ Motivations

Motivation



Count estimation

Quality Control

Dimensionality reduction

Clustering

Trajectory Inference

Differential Expression

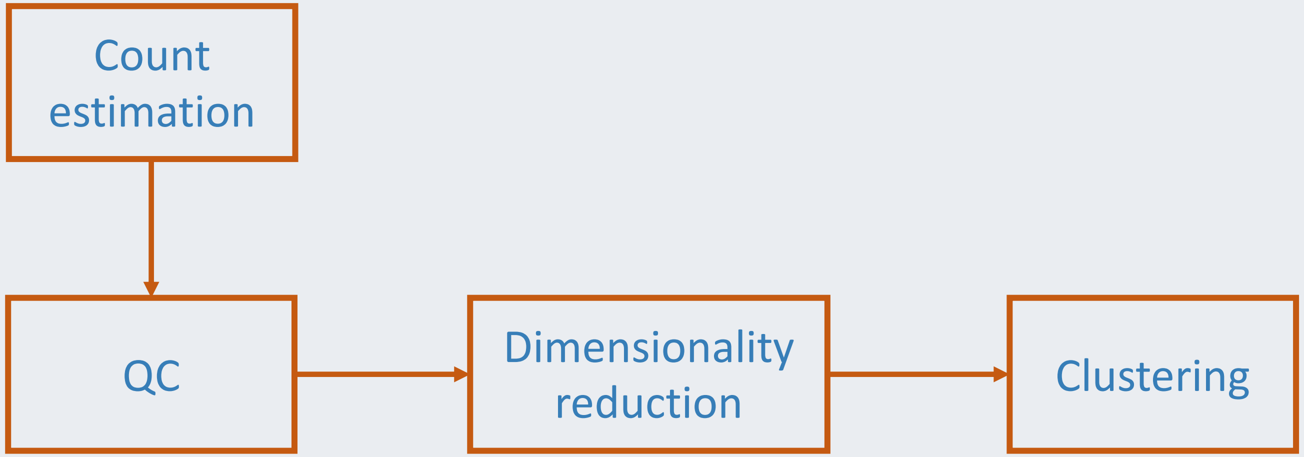
Clustering in scRNA–Seq

Here, we talk about clustering of cells, not genes

Clustering is used to detect cell–types, i. e. cells with a distinct common transcriptomic signature.

Many clustering methods are used in scRNA–Seq, ranging from direct application of existing clustering methods to adaptation of those methods for scRNA–Seq specific purposes.

Motivation



PCA
UMAP
t–SNE
ZINBWave
…

k–means
hierarchical
RSEC
SC3
…

And all those methods have parameters of their owns

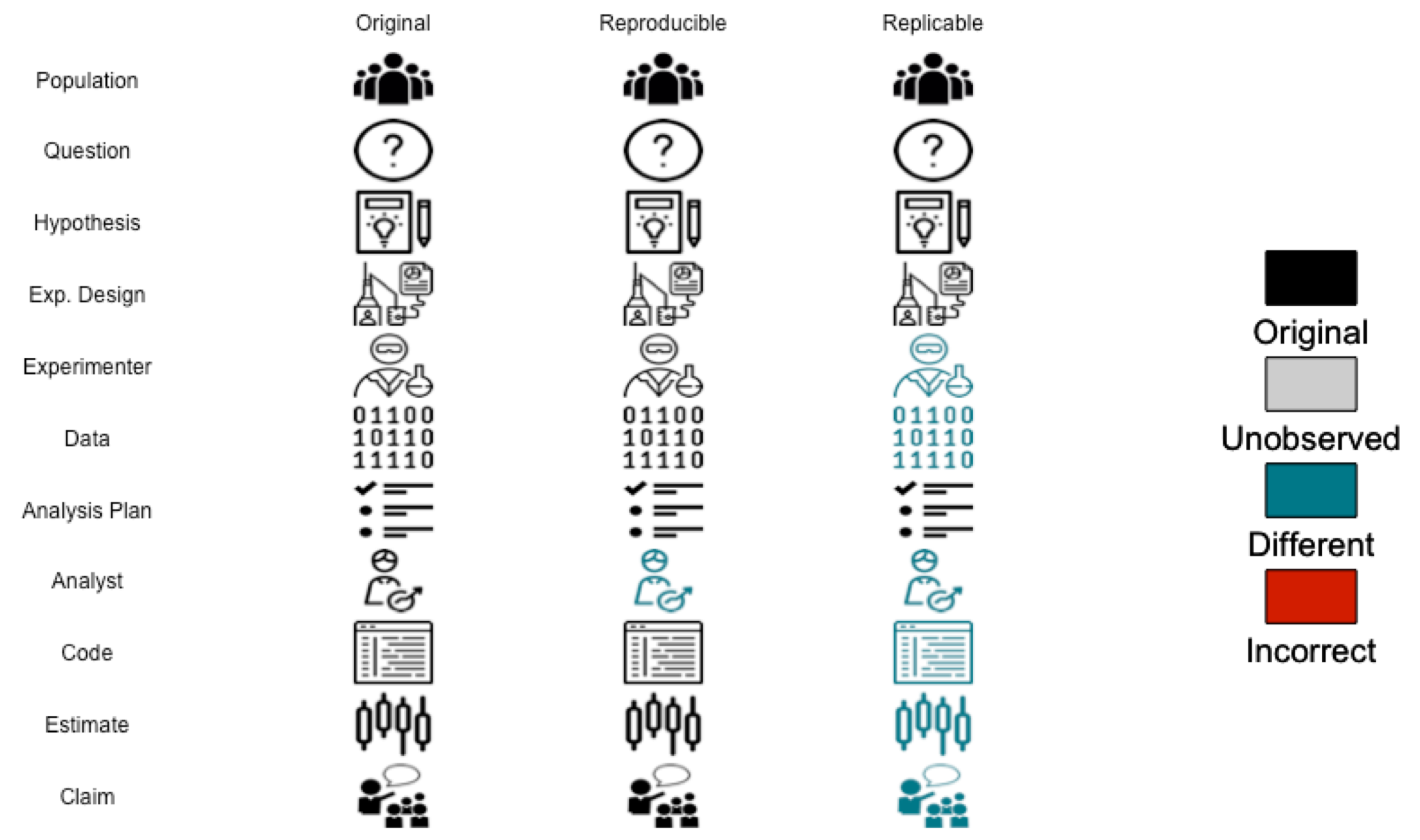
How to benchmark clustering?

The first and most common way is to compute agreement with a gold standard, usually using the adjusted Rand Index (ARI)

	X_1	X_2	…	X_r	*Sums*
Y_1	n_{11}	n_{12}	…	n_{1r}	a_1
Y_2	n_{21}	n_{22}	…	n_{2r}	a_2
…	…	…	…	…	…
Y_s	n_{s1}	n_{s2}	…	n_{sr}	a_s
Sums	b_1	b_2	…	b_r	

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\left(\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right)}{\binom{n}{2}}}{\frac{1}{2}\times\left(\sum_i\binom{a_i}{2} + \sum_j\binom{b_j}{2}\right) - \frac{\left(\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right)}{\binom{n}{2}}}$$

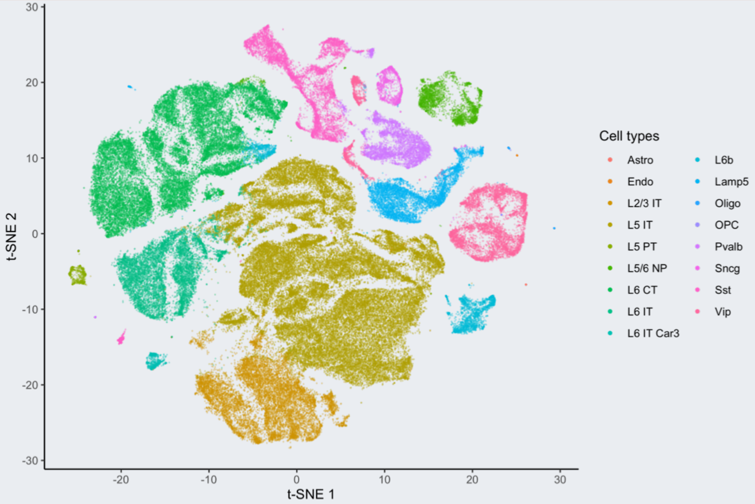
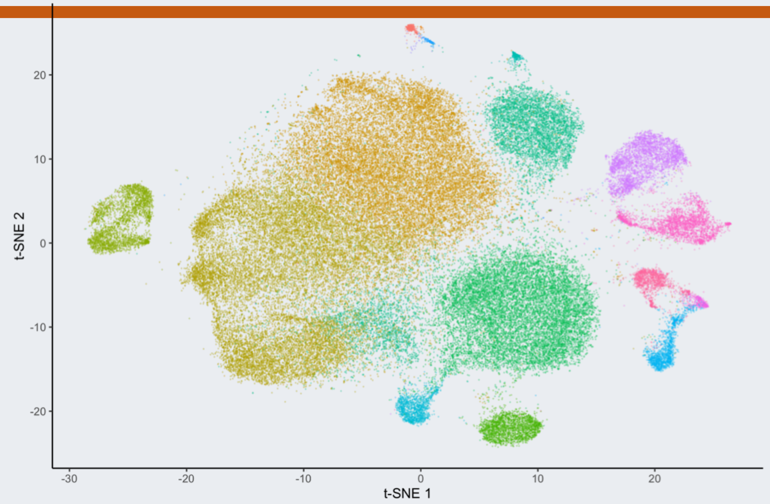
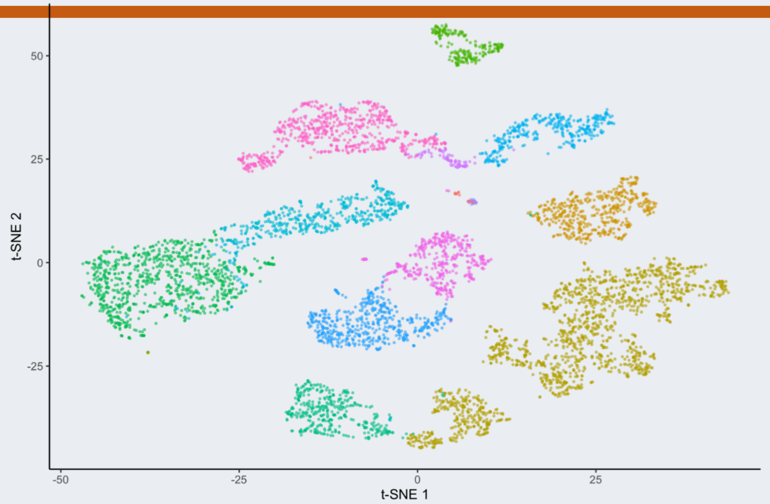
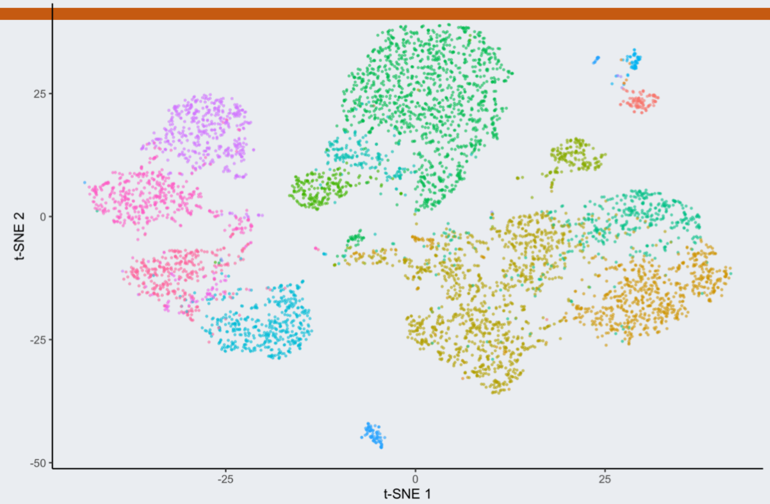
Replicability and reproducibility



Patil, P. , Peng, R. D. & Leek, J. T. A visual tool for defining reproducibility and replicability. *Nat Hum Behav* 3, 650–652 (2019) doi: 10. 1038/s41562-019-0629-z

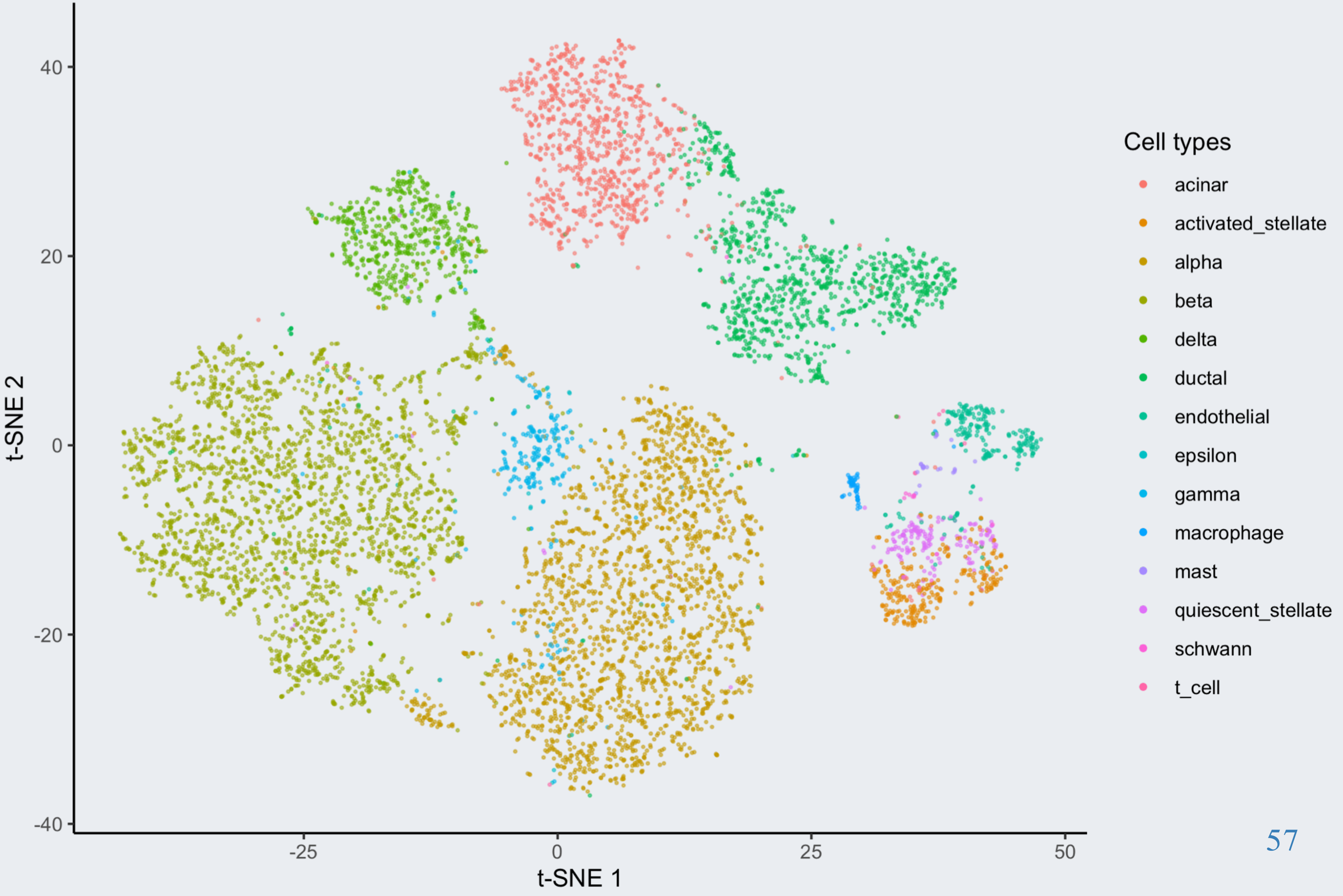
III. ii/ Datasets

Mouse Brain: 4 platforms

	Cells	Nuclei
10x		
Smart-Seq		

Pancreas data from Baron et al.

- 8569 cells and 5124 genes
- Dimensionality reduction with zinbWave + t-SNE
- Clusters from the original publication



Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., … Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, *3*(4), 346–360. e4. https://doi.org/10.1016/j.cels.2016.08.011

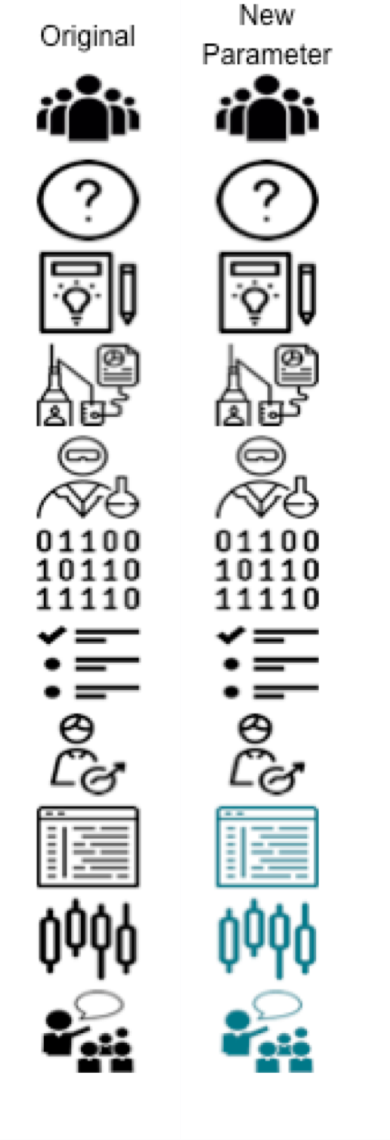
Pancreas data from Segerstople et al.

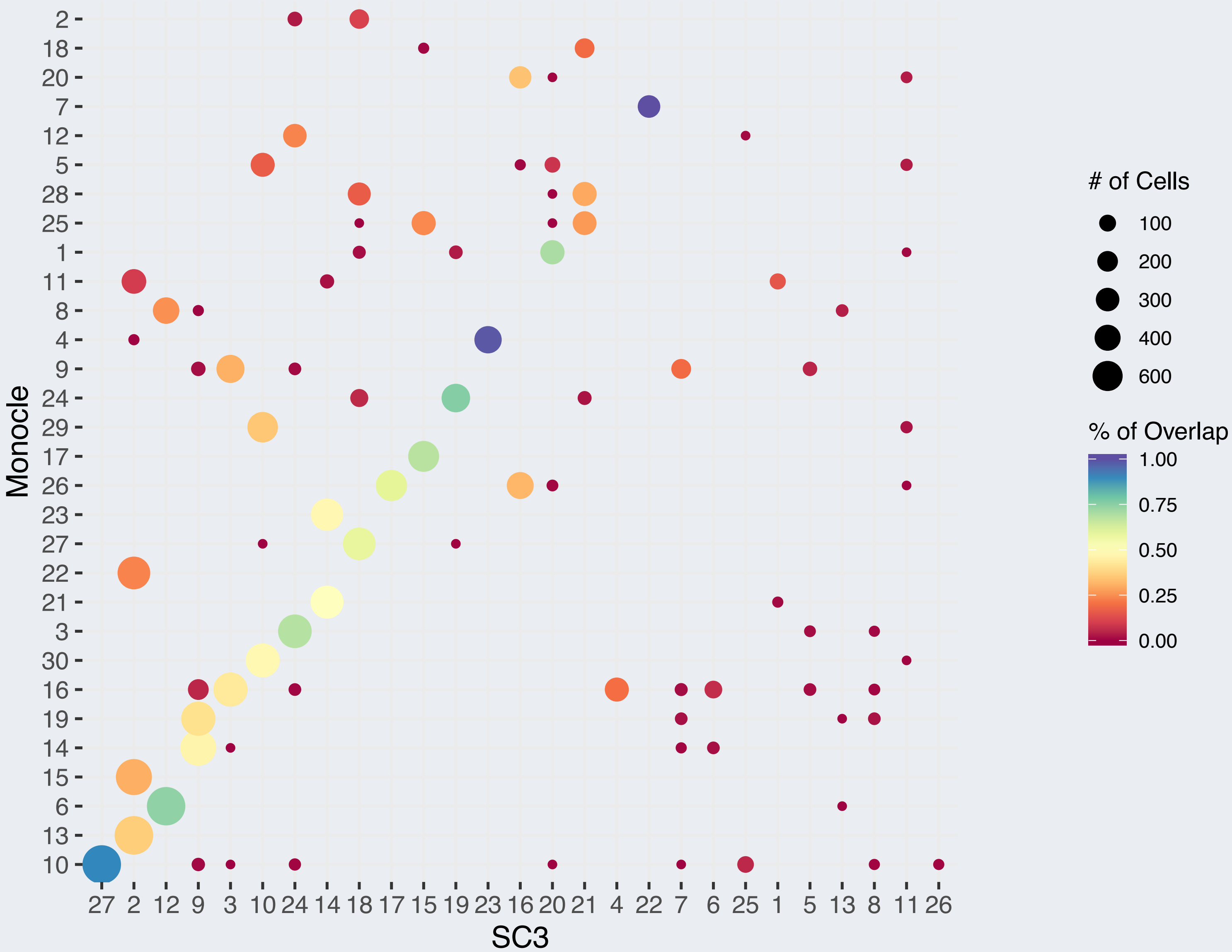
- 2136 cells and 7764 genes
- Dimensionality reduction with zinbWave + t-SNE
- Clusters from the original publication

Segerstolpe, Å. , Palasantza, A. , Eliasson, P. , Andersson, E. M. , Andréasson, A. C. , Sun, X. , ... Sandberg, R. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, *24*(4), 593–607. https: //doi. org/10. 1016/j. cmet. 2016. 08. 020

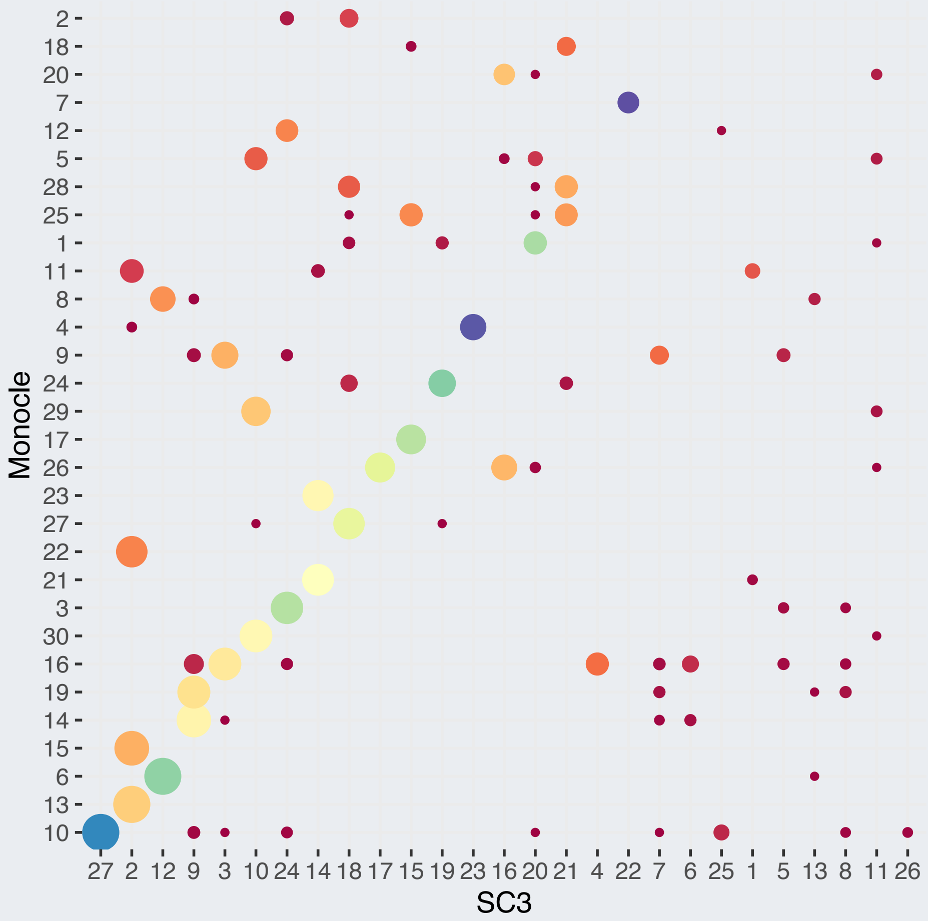
III. iii/ Measuring replicability

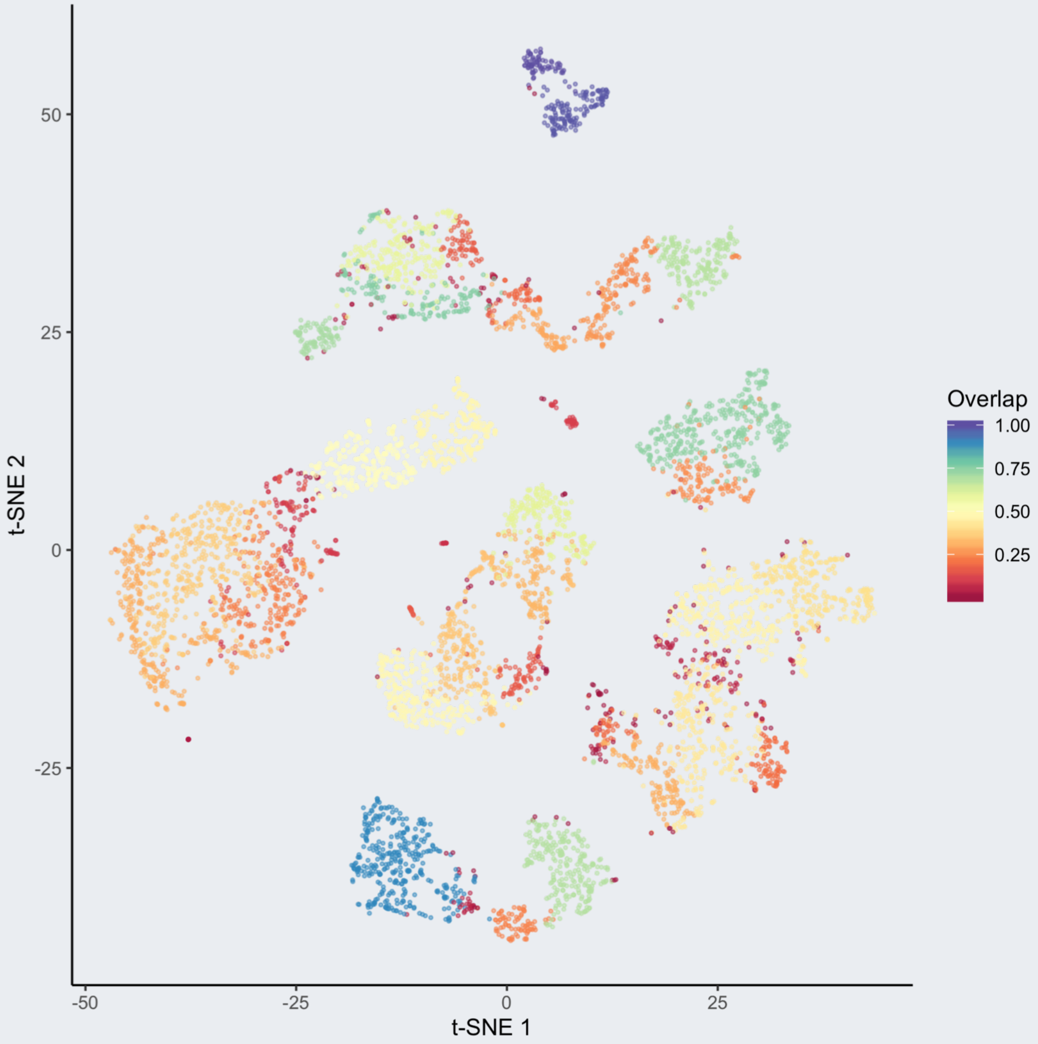
Clustering is highly sensitive to analysis choices



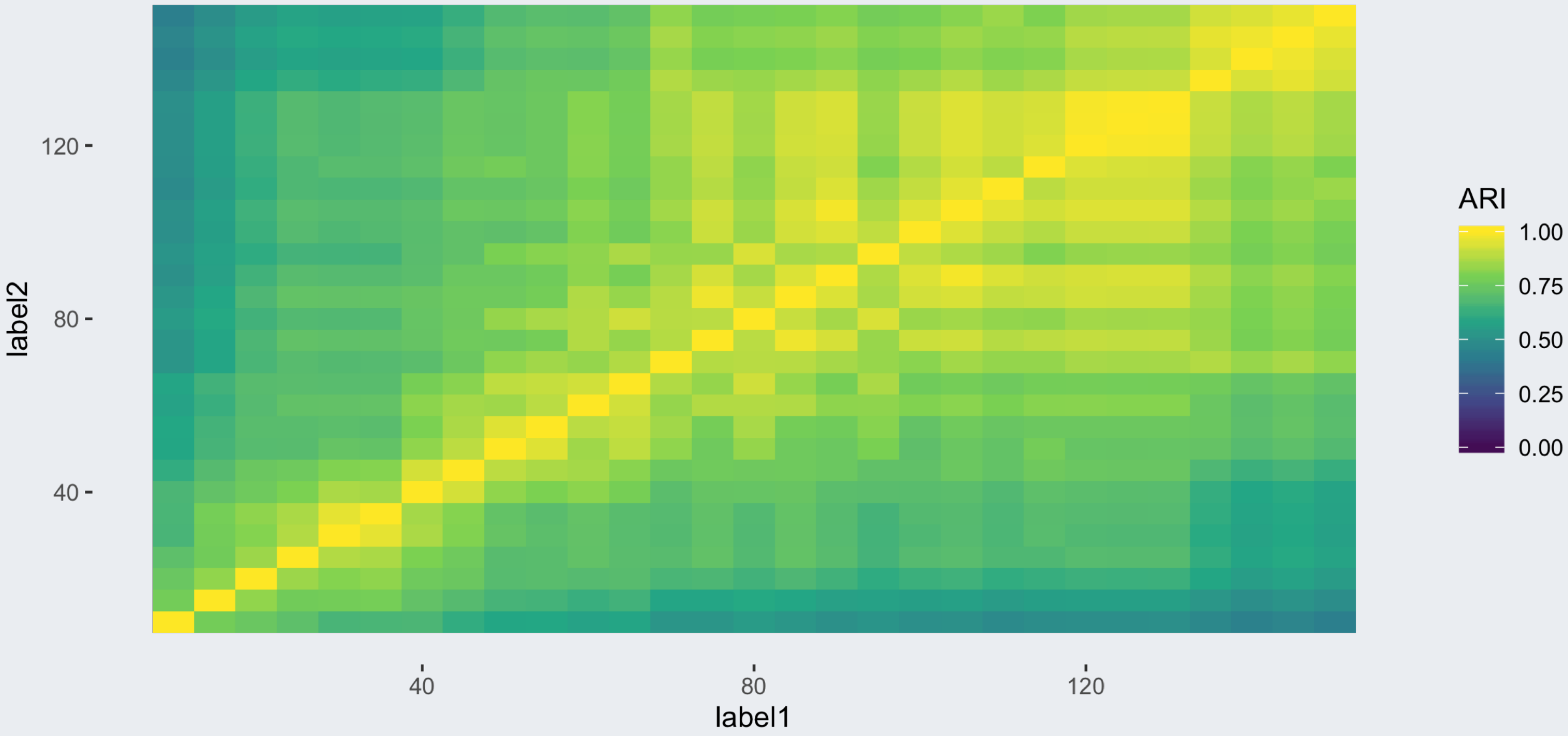


Clustering is highly sensitive to analysis choices



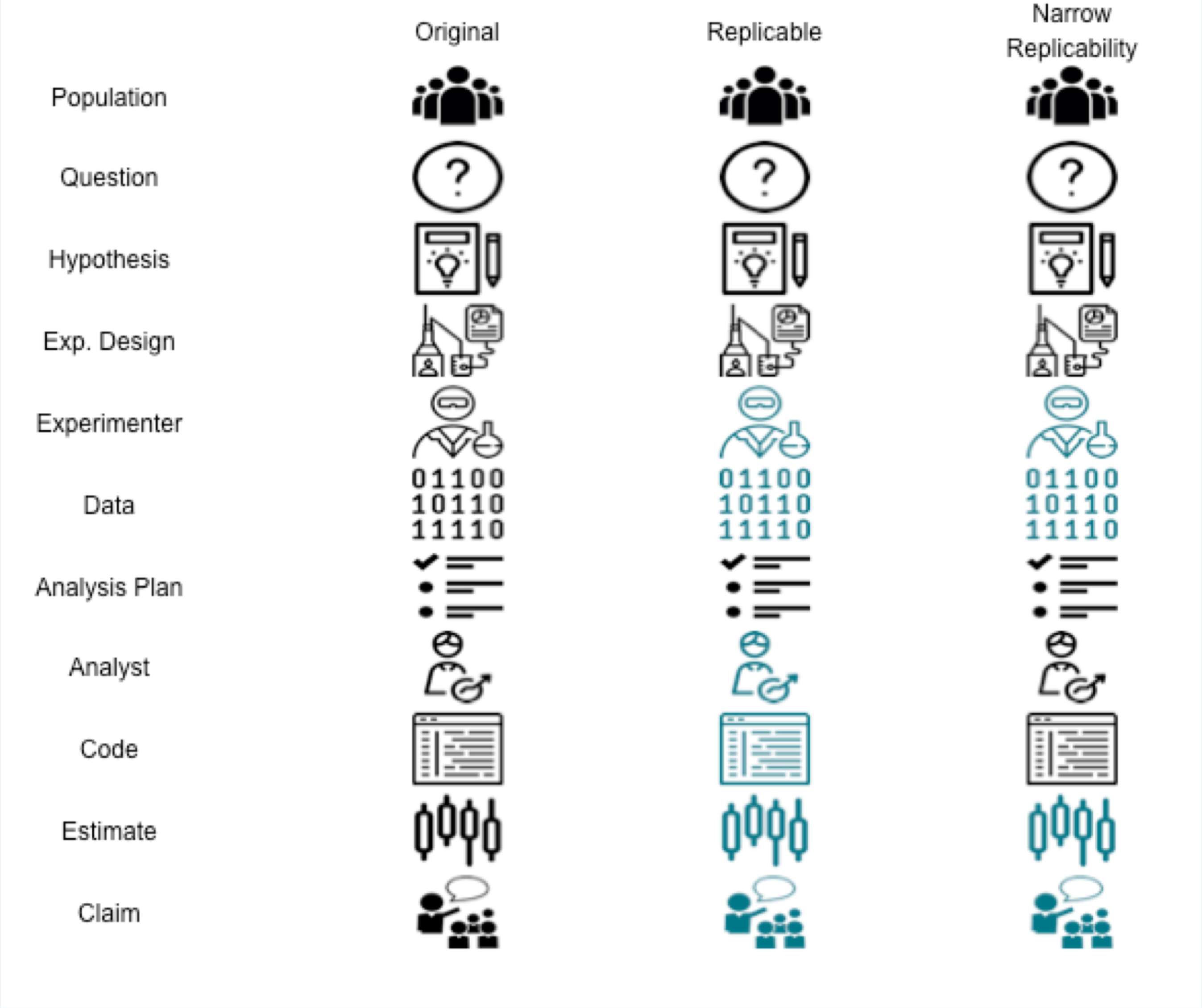


Clustering is highly sensitive to analysis choices



Monocle ran on the same dataset while changing one parameter

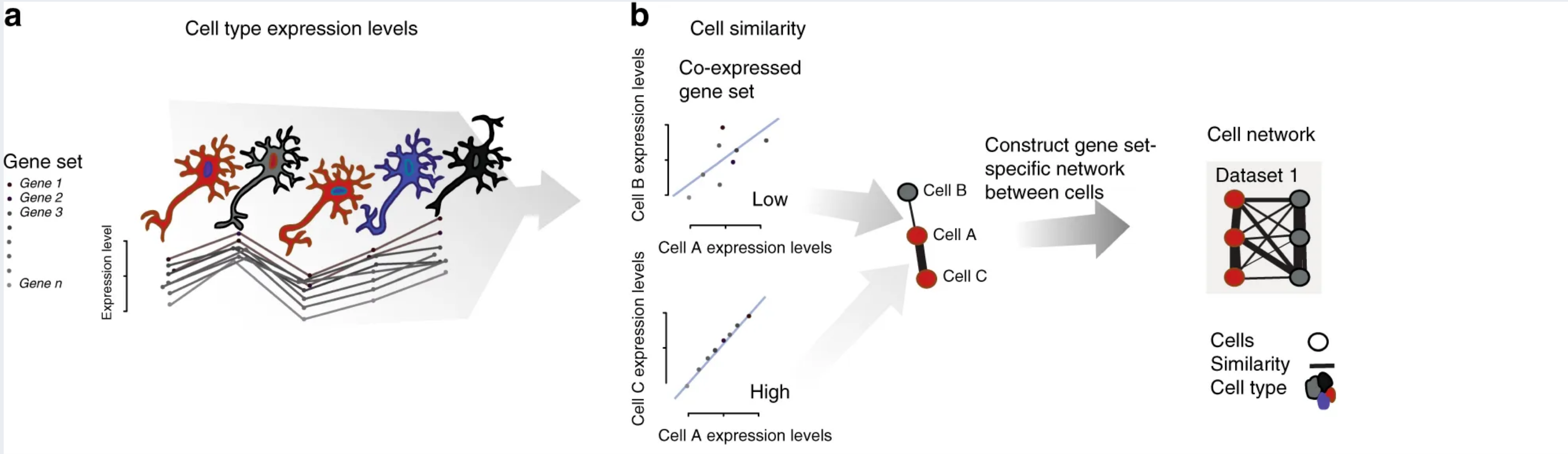
Measuring replicability with MetaNeighbour



For example, we compare how the output from *"running ZINB-Wave + Monocle with* $k = 45$*"* replicates over two datasets

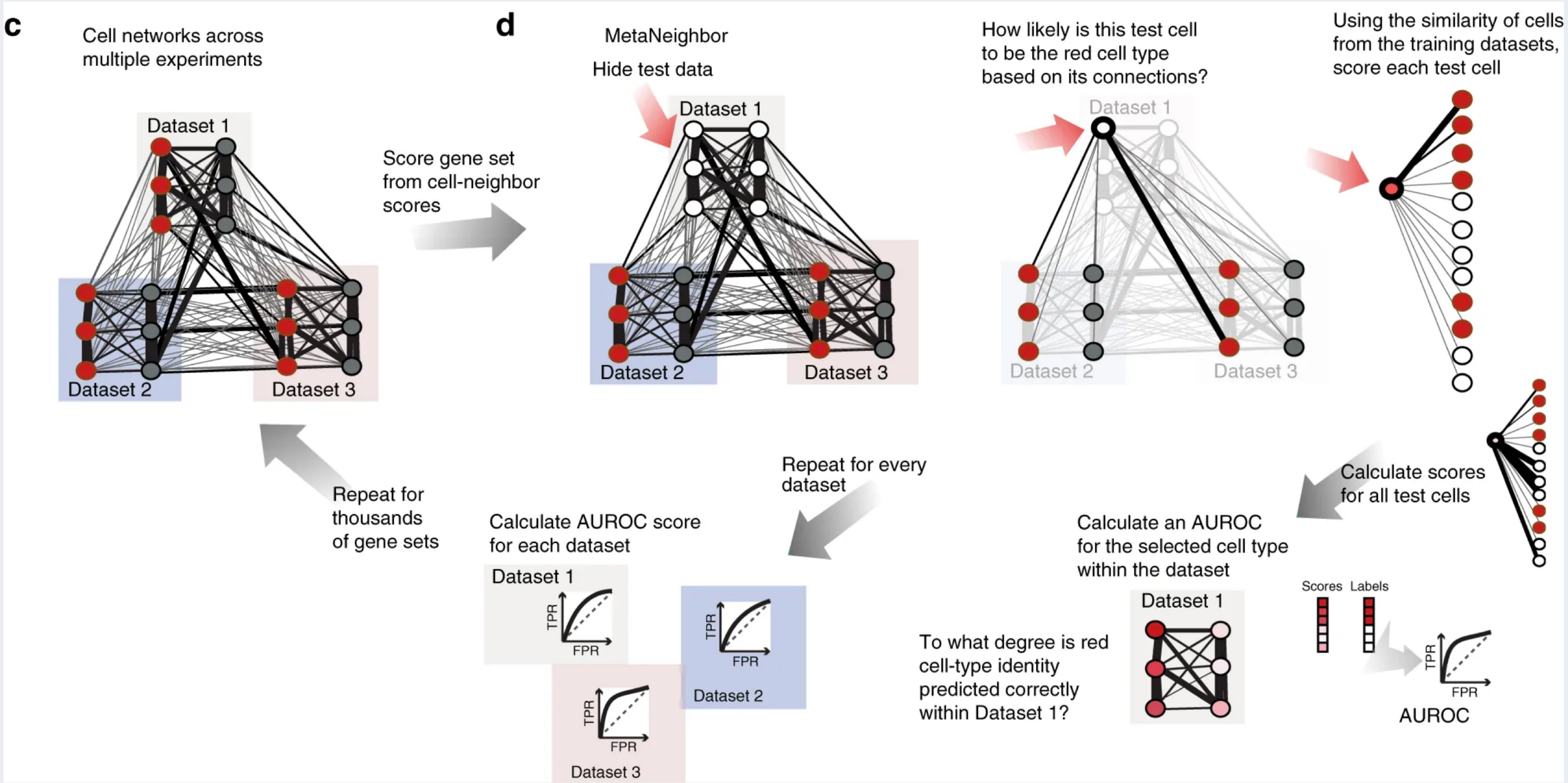
Crow, M. , Paul, A. , Ballouz, S. , Huang, Z. J. , & Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nature Communications*, *9*(1), 884. https: //doi. org/10. 1038/s41467-018-03282-0

Measuring replicability with MetaNeighbour



Crow, M. , Paul, A. , Ballouz, S. , Huang, Z. J. , & Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nature Communications*, *9*(1), 884. https://doi.org/10.1038/s41467-018-03282-0

Supervised Metaneighbour



Supervised Metaneighbour: cluster labels are shared among datasets

Crow, M. , Paul, A. , Ballouz, S. , Huang, Z. J. , & Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nature Communications*, *9*(1), 884. https: //doi. org/10. 1038/s41467–018–03282–0

Unsupervised Metaneighbour

- Greedy approach: look at every pair of clusters, score each cluster by how well it predict the other

- Define the AUROC for a pair of clusters by the minimum of the two AUROC

- A cluster is replicable if AUROC > cutoff

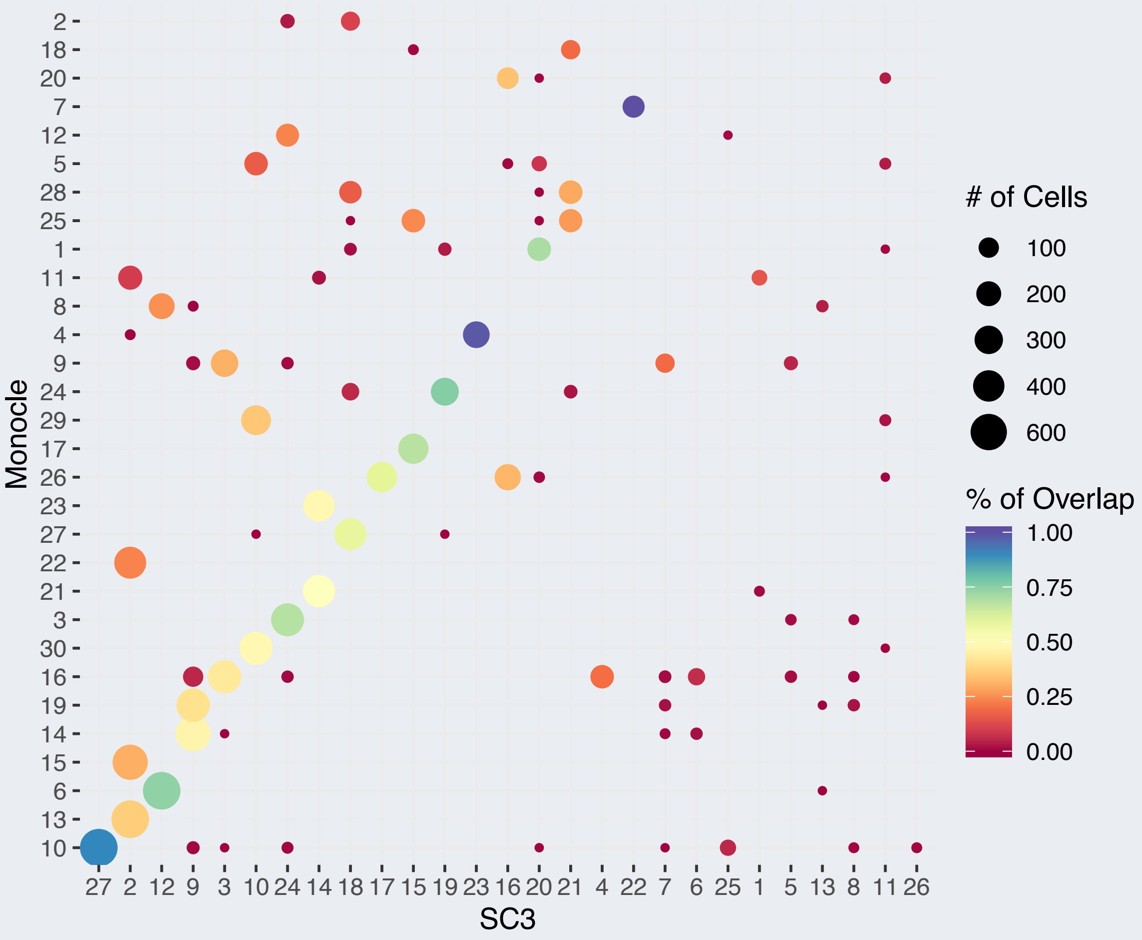
Metaneighbour classify a pair of clusters are replicable if each cluster of the pair is well predictive of the other, and more predictive than any other cluster.

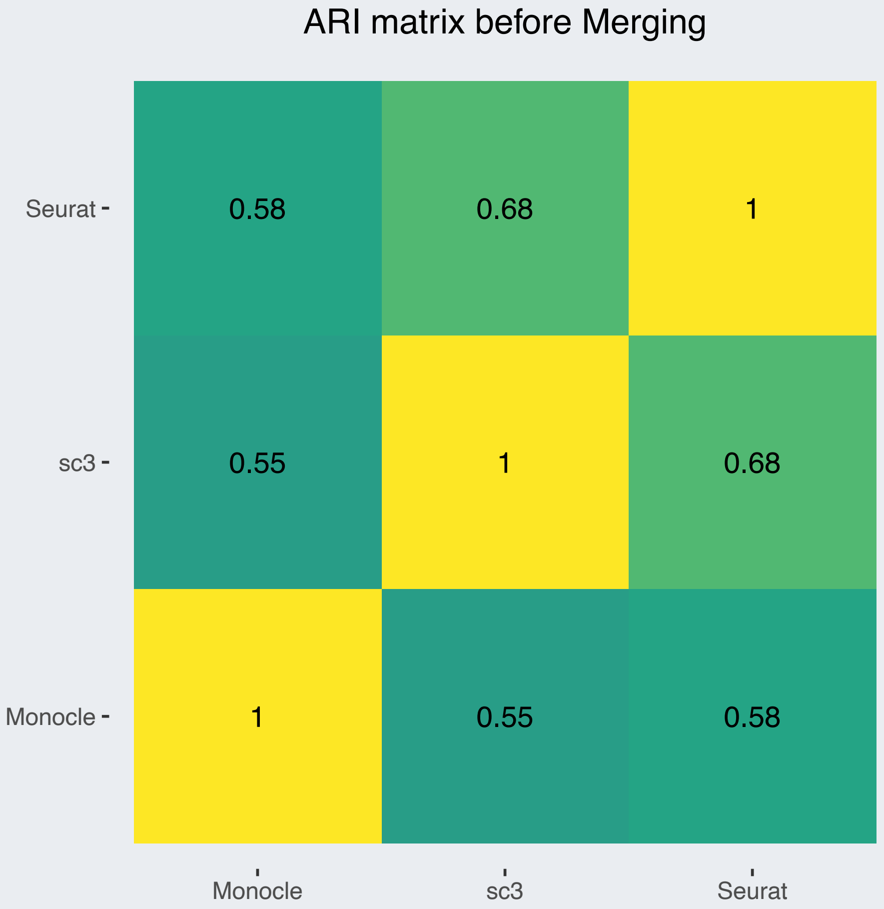
Crow, M., Paul, A., Ballouz, S., Huang, Z. J., & Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nature Communications*, *9*(1), 884. https://doi.org/10.1038/s41467-018-03282-0

Resolution– Replicability trade–off

III. iv/ Methods

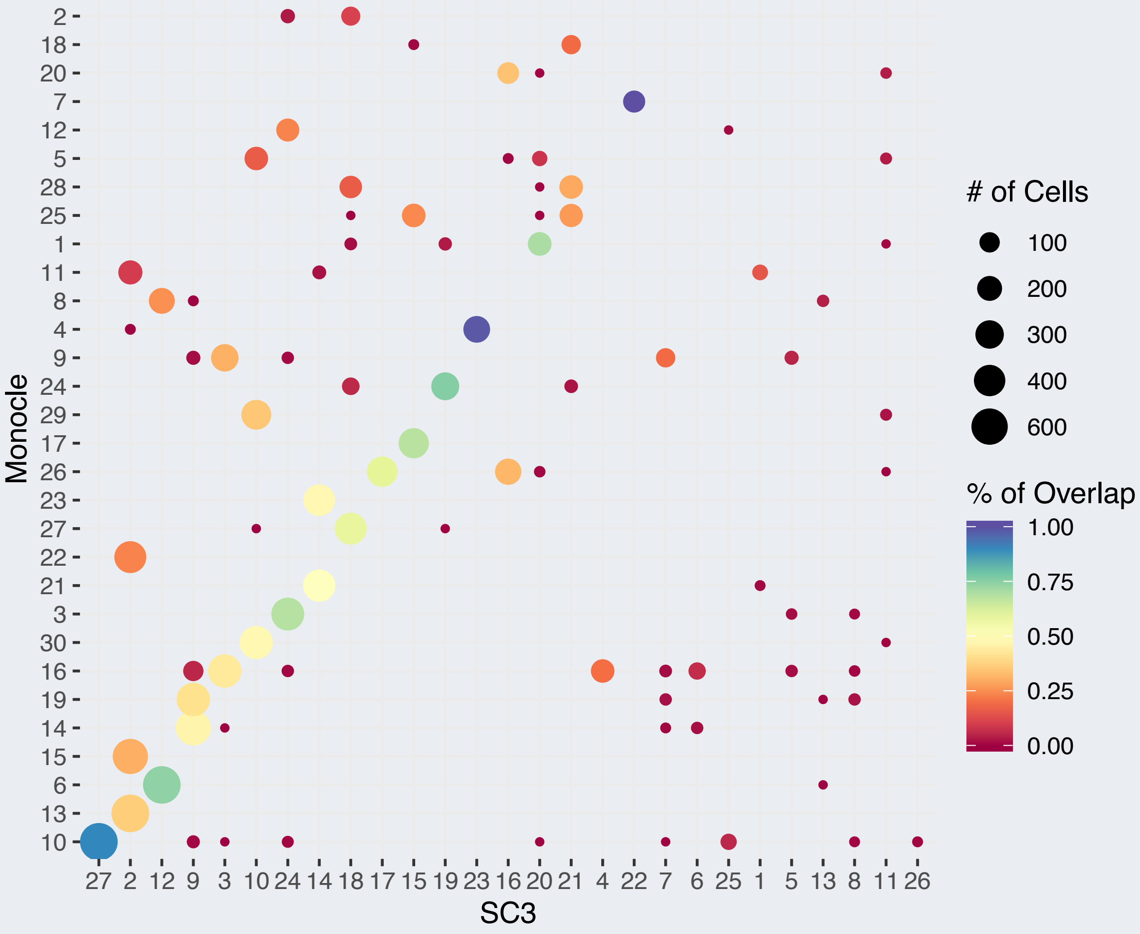
Pairwise merging to improve the ARI




ARI matrix before Merging

	Monocle	sc3	Seurat
Seurat	0.58	0.68	1
sc3	0.55	1	0.68
Monocle	1	0.55	0.58

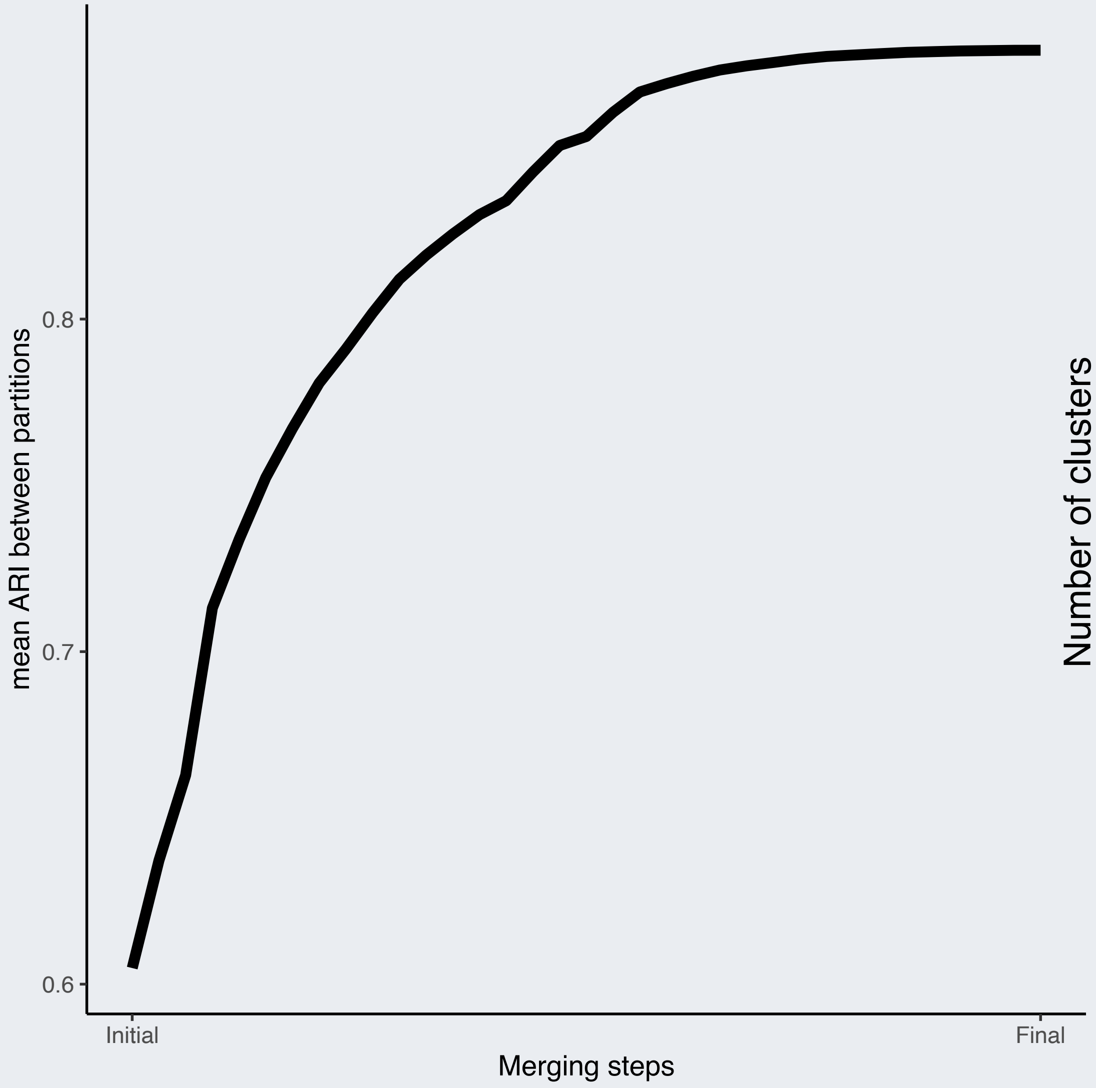
Pairwise merging to improve the ARI

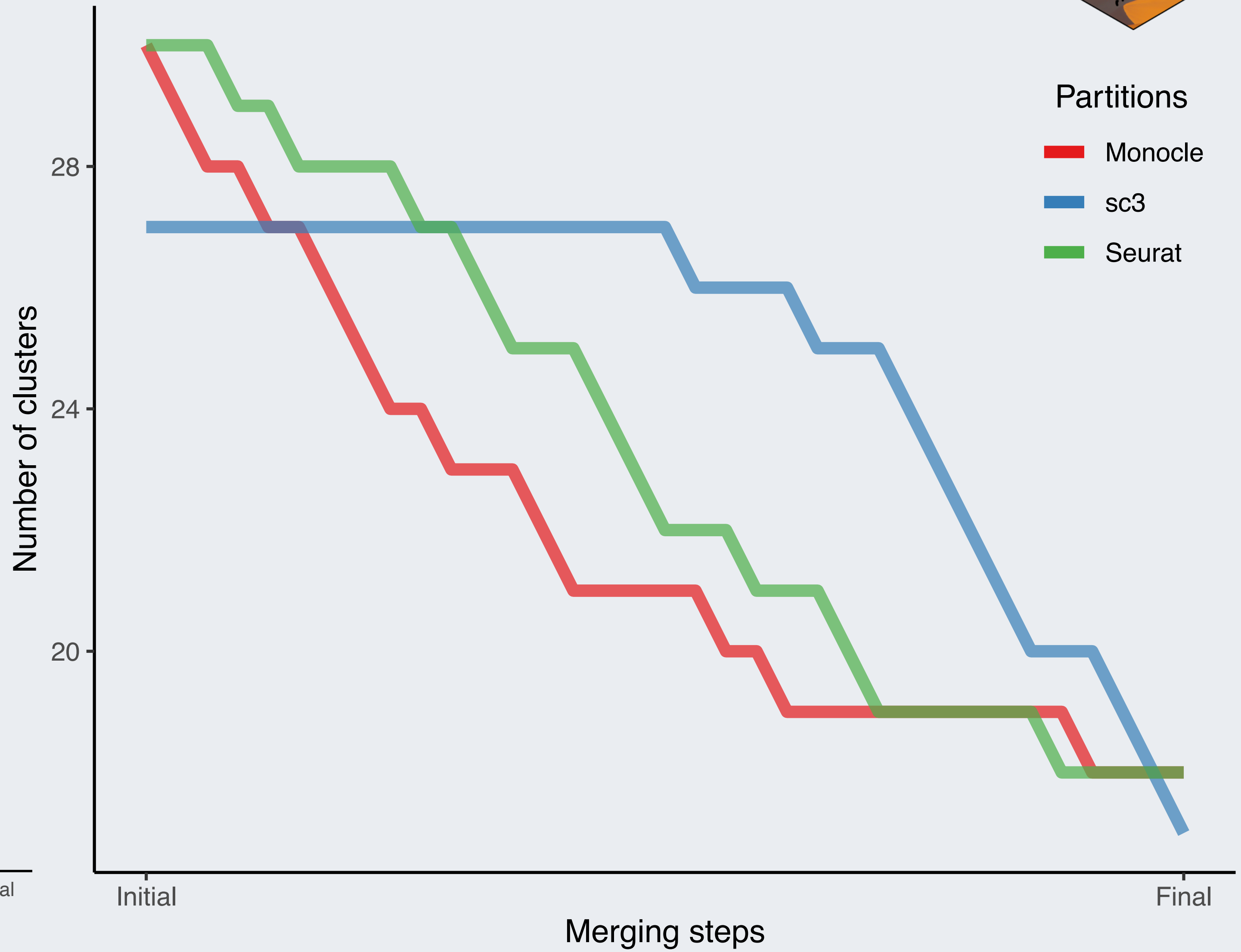


- Look at every pair of clusters over all cluster labels (or partitions)
- Merge the pair and recompute the mean ARI
- Find the pair where this improves the mean ARI the most
- Actually merge that pair
- Iterate

- Stop when you cannot improve the ARI anymore

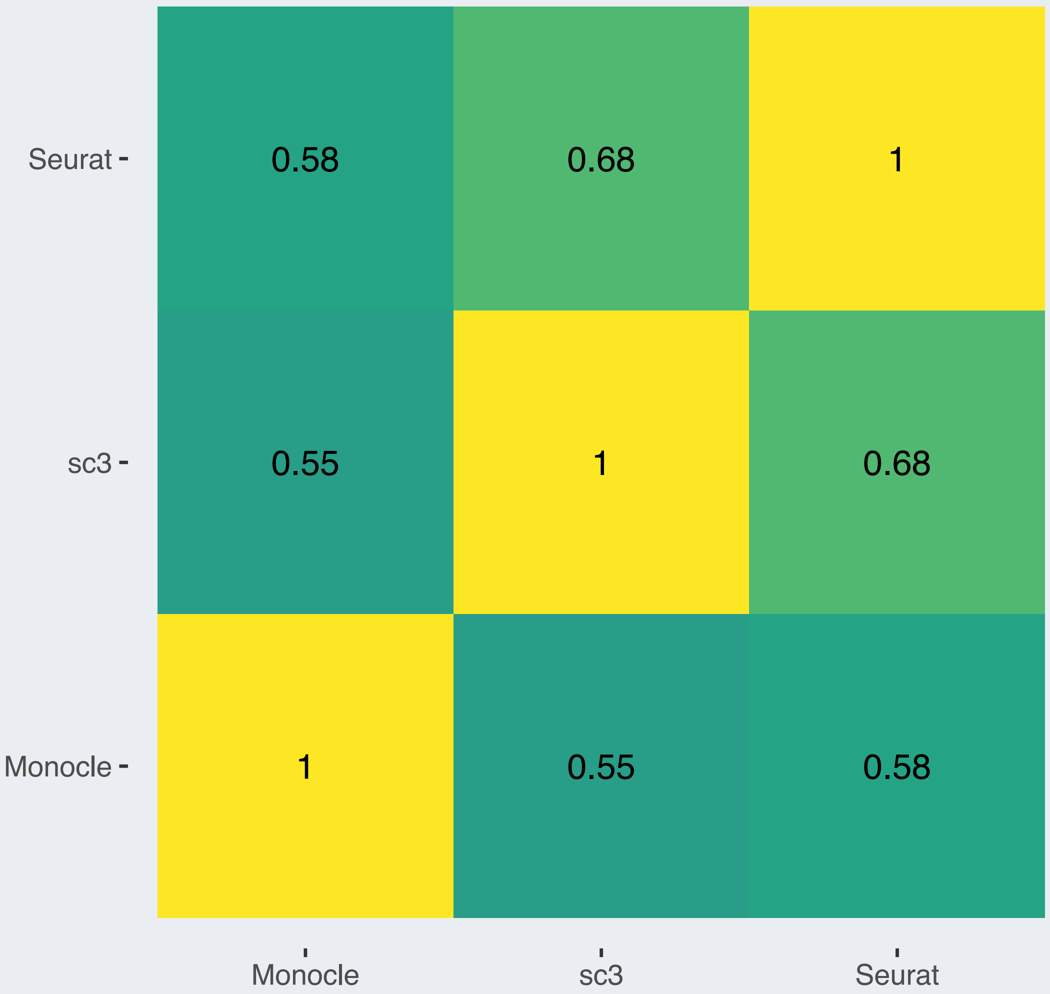
Dune improves mean ARI





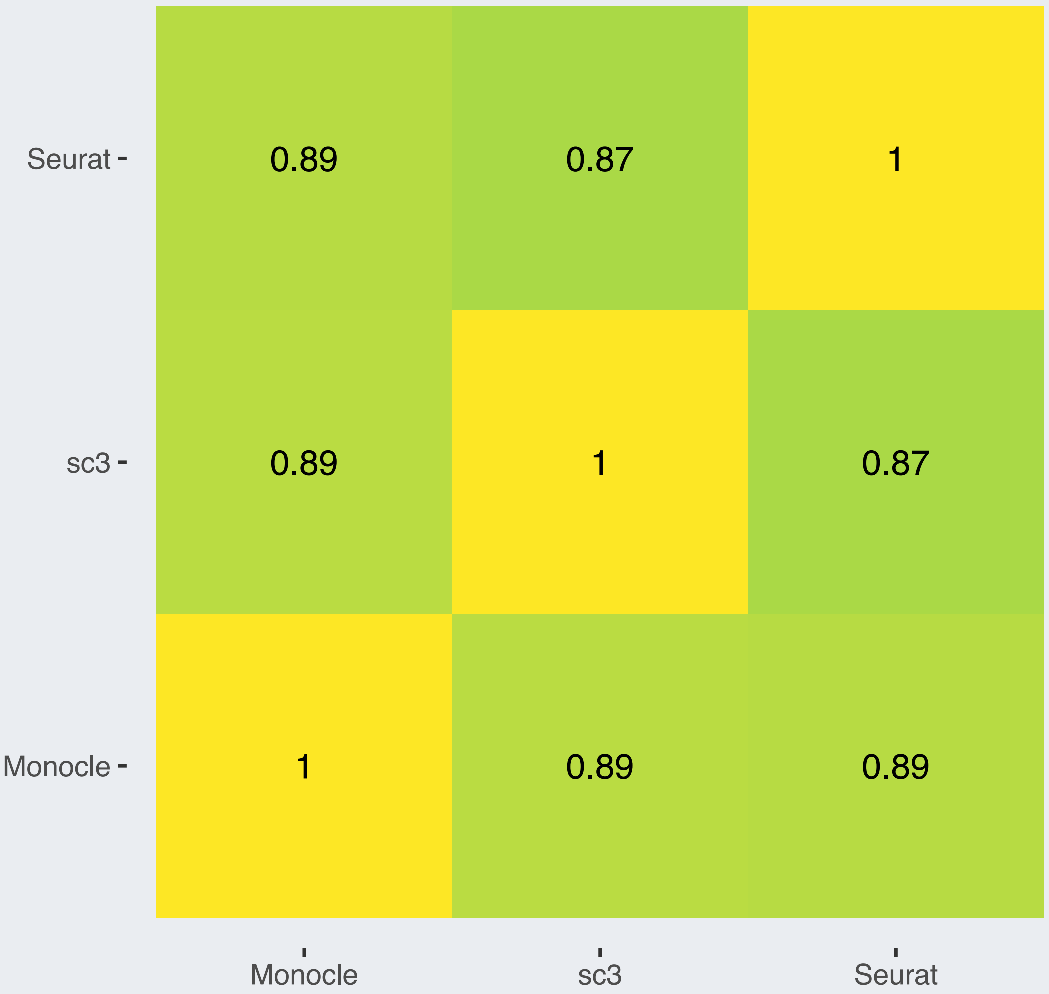
Dune improves mean ARI

ARI matrix before Merging



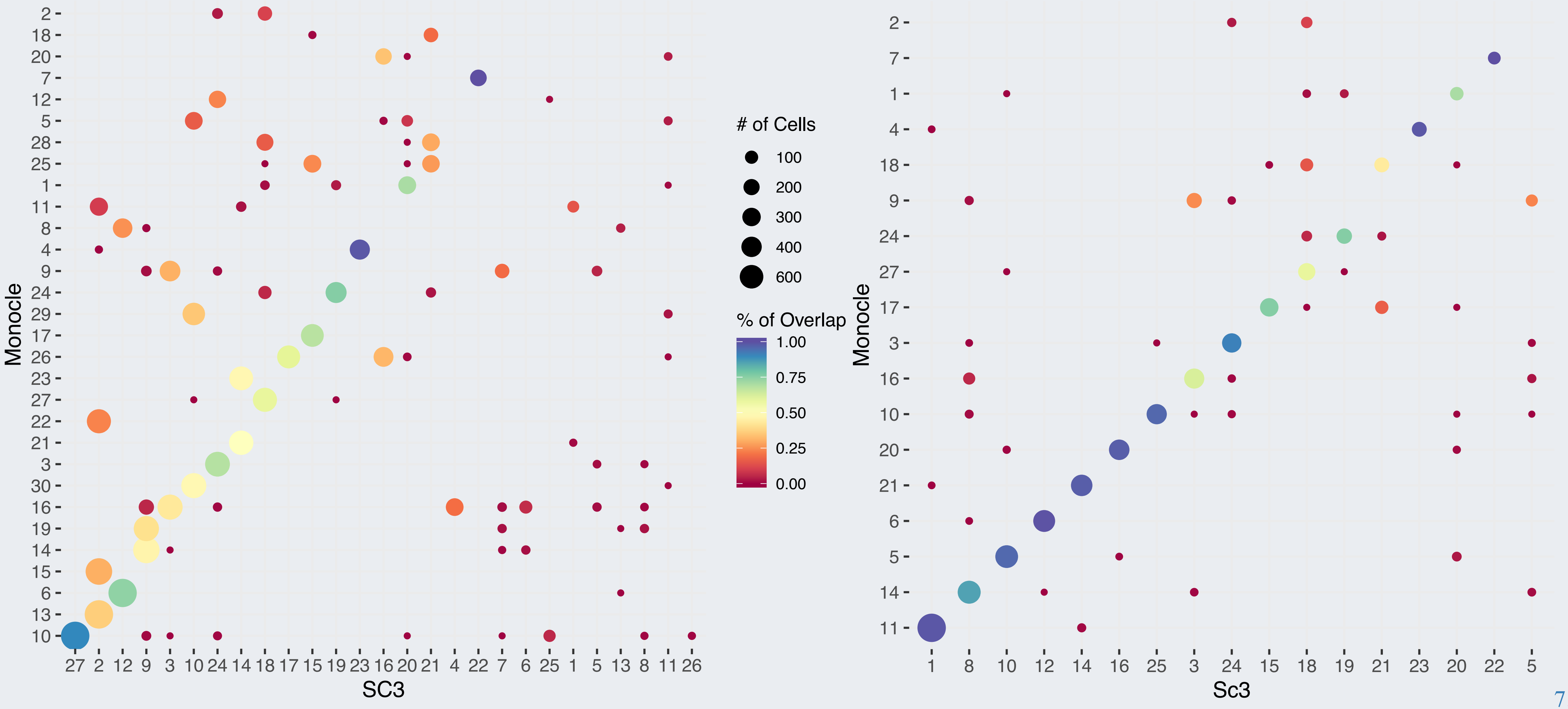
	Monocle	sc3	Seurat
Seurat	0.58	0.68	1
sc3	0.55	1	0.68
Monocle	1	0.55	0.58

ARI matrix after Merging



	Monocle	sc3	Seurat
Seurat	0.89	0.87	1
sc3	0.89	1	0.87
Monocle	1	0.89	0.89

Dune improves mean ARI



Other methods of merging

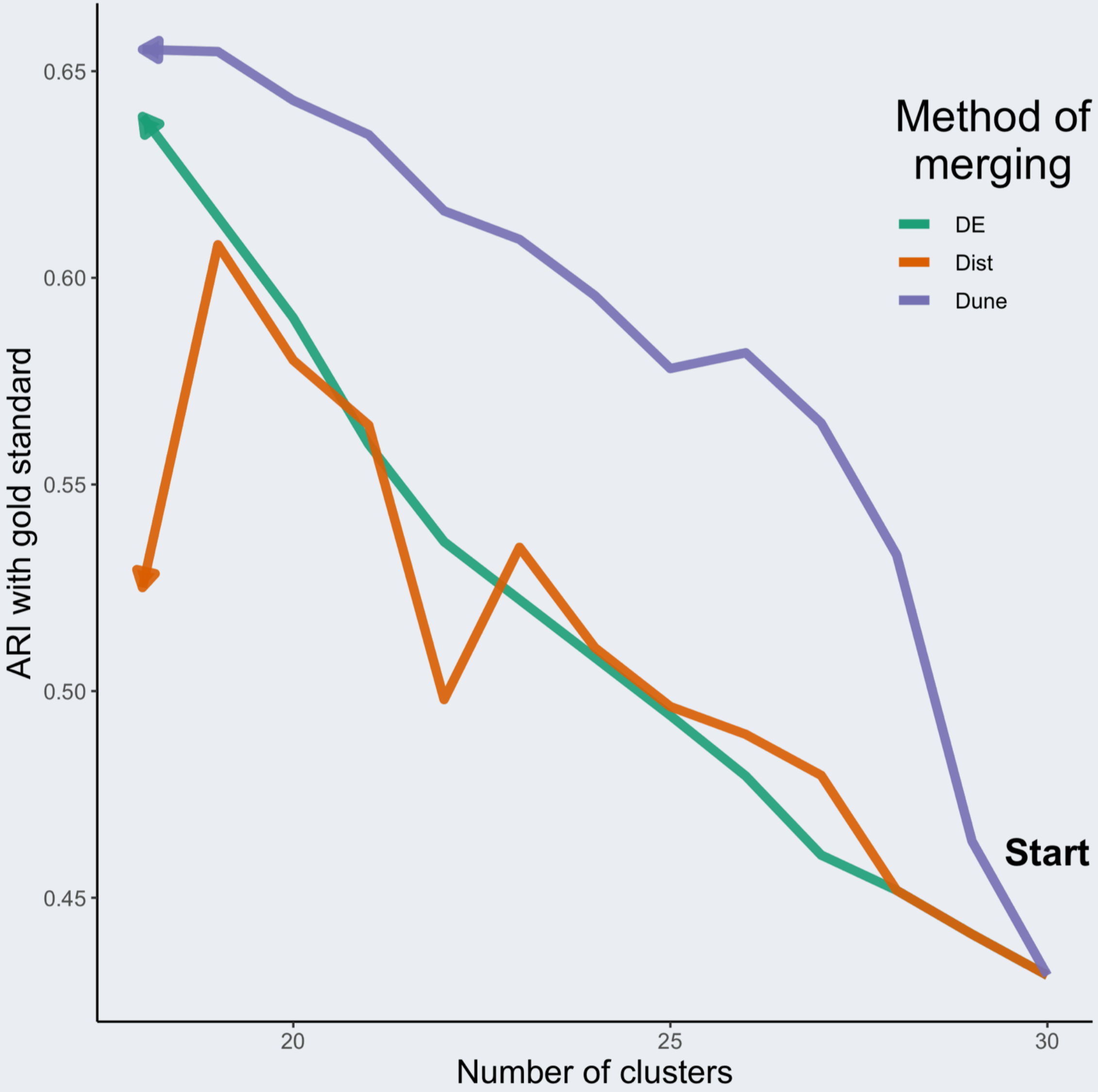
1) Build a hierarchy on the clusters based on a distance metric and linkage.

   Here we picked Euclidian distance in the reduced space for the cluster medoids and complete linkage, as implemented in **RSEC**.

2) Merge along the tree:
   1) Either merge clusters where $\%DE\ genes < cutoff$ **DE**
   2) Or cut tree at various heights $distance_{clusters} < cutoff$ **Dist**

III. v/ Results

ARI with gold standard

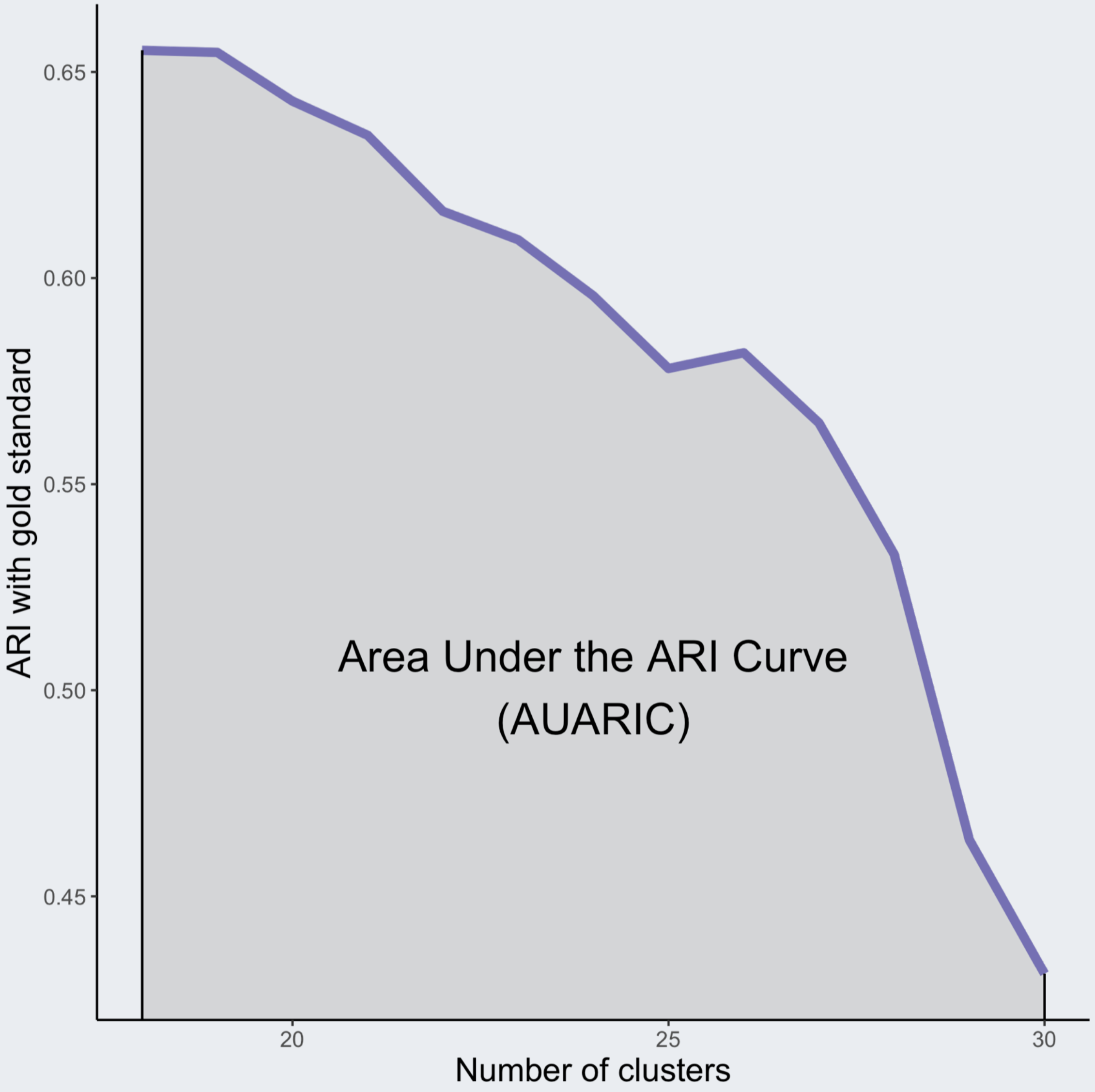


On the Brain Smart–Seq cell dataset, we run **SC**3 with $\theta = 0$ and then merged with the three methods

Note that

- **Dune** is merging using **Monocle** and **Seurat**.
- We use **Dune**'s stopping point to stop the other methods

ARI with gold standard



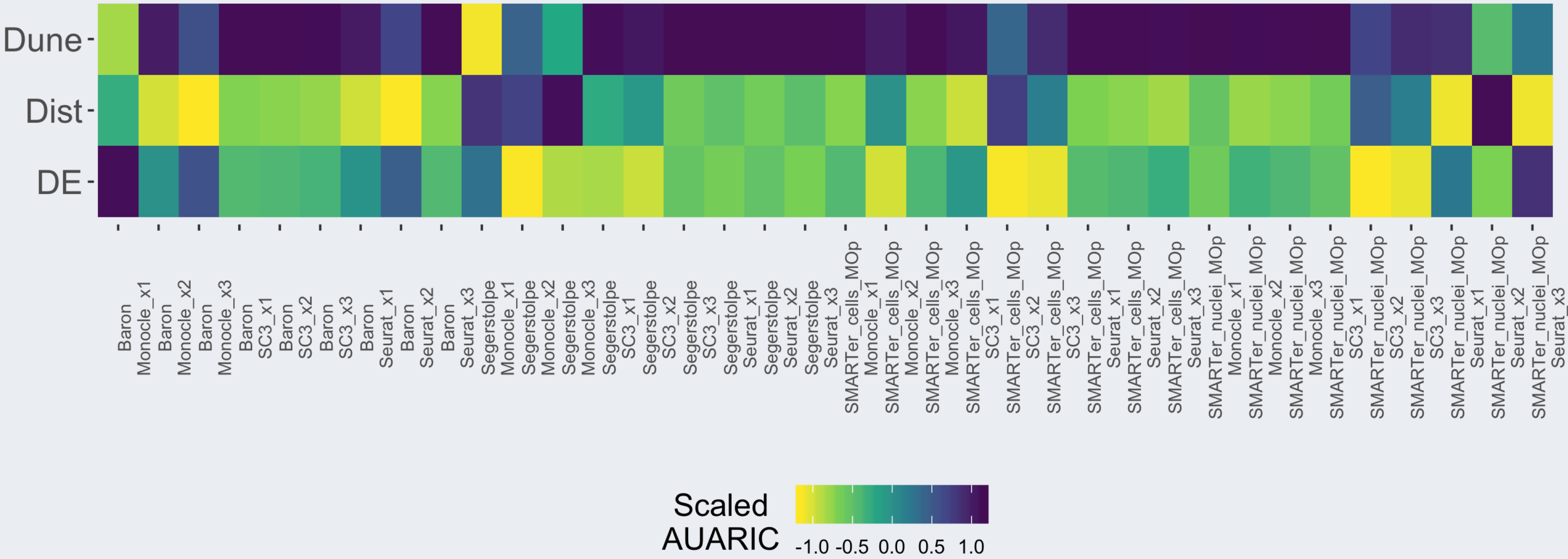
- Compute the AUARIC
- With a different parameter or different clustering method or different dataset, the stopping point will vary

–> We scale the AUARIC values

Method of merging	AUARIC	Scaled AUARIC
DE	6.23	-0.41
Dist	6.28	-0.72
Dune	7.02	1.14

ARI with gold standard

3 clustering methods × 3 θ_{method} × 4 datasets = 36 comparisons

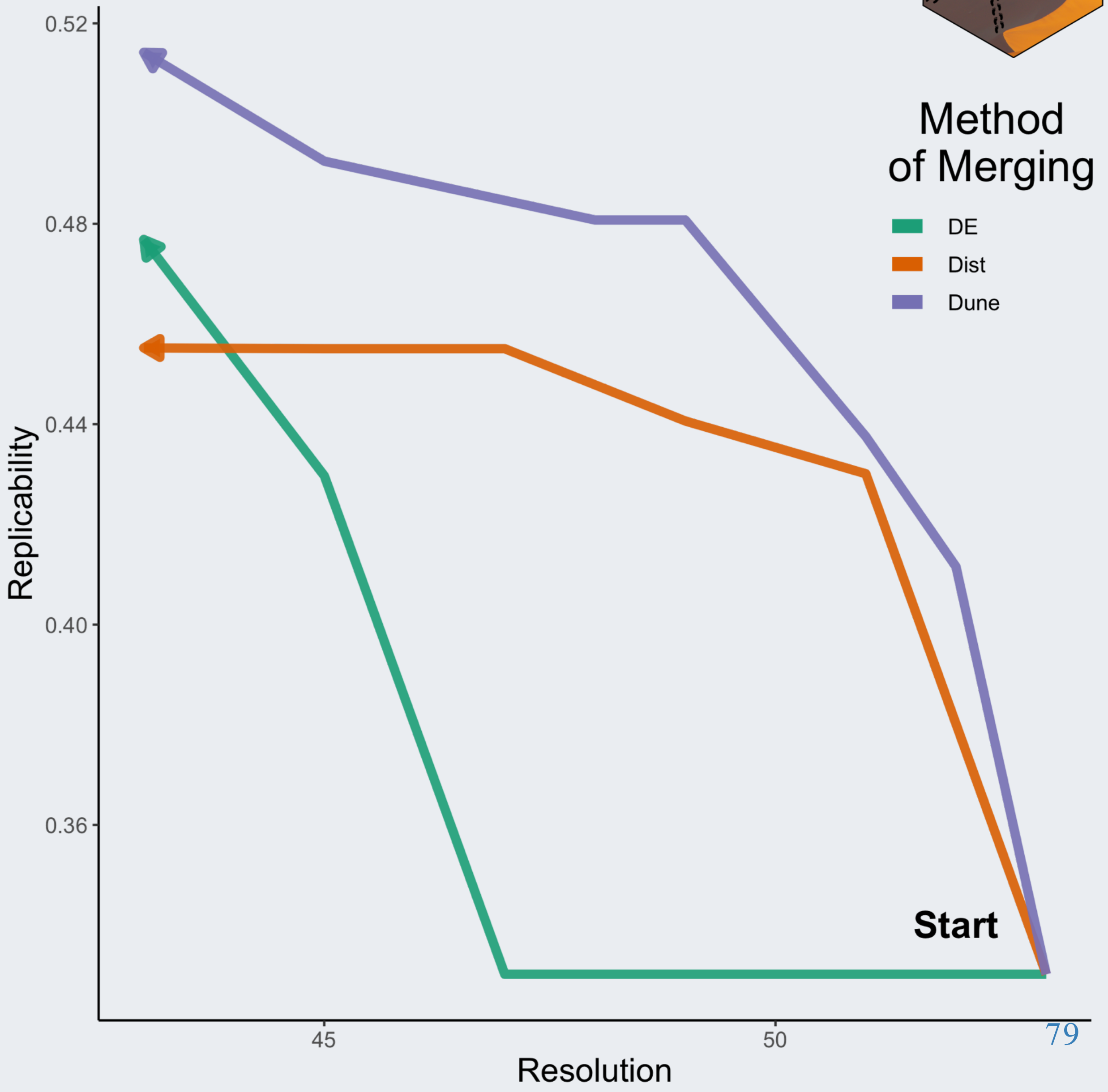


Replicability over pairs of datasets

On the Brain Smart–Seq cell dataset, we run **Seurat** with $\theta = 1.2$ and then merged with the three methods

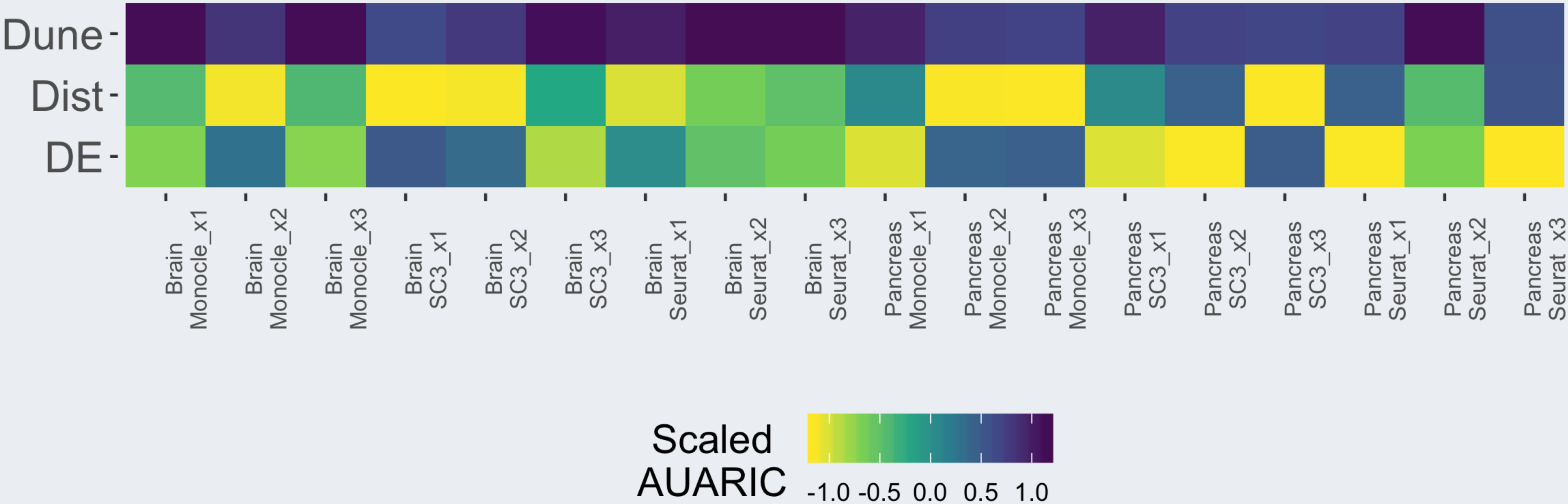
Note that

- **Dune** is merging using **Monocle** and **SC3**.
- We use **Dune**'s stopping point to stop the other methods

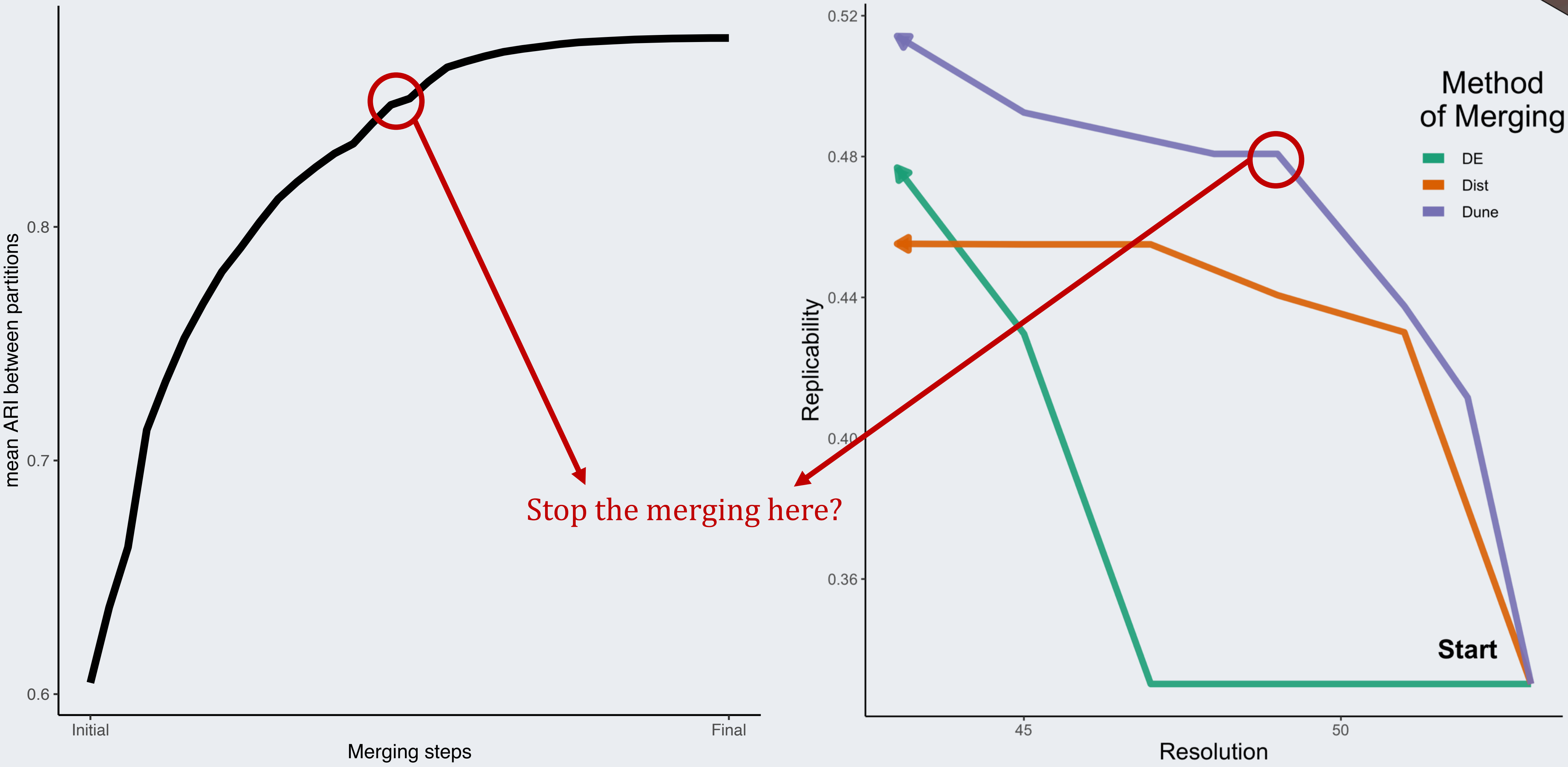


Replicability over pairs of datasets

3 clustering methods × 3 θ_{method} × 2 pairs of datasets = 18 comparisons



Introduce some regularization?



Stop the merging here?

Thanks to the tradeSeq team

Sandrine Dudoit

Lieven Clement

Kelly Street

Koen Van den Berge

Thanks to the Dune team

Sandrine Dudoit

John Ngai

Elizabeth Purdom

Stephan Fischer

Davide Risso

Kelly Street

Koen Van den Berge

Rebecca Chance

Jesse Gillis

Thanks to

- Martin Kinisu and Lin He of the He lab
- Students in the Biostat program for help in preparing this presentation and for many insightful discussions over time.
- Kelly Street for mentoring me at the beginning of my work in the Dudoit Lab
- Sandrine for being my advisor

Thank to all of you
for listening

Questions?

Citations

- Deng, Q. , Ramsköld, D. , Reinius, B. , & Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*. https: //doi. org/10. 1126/science. 1245316
- Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv, 2 2018. URL http: //arxiv. org/abs/1802. 03426
- Franziska Paul, et al . Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell, 163(7): 1663–1677, 12 2015. ISSN 00928674. doi: 10. 1016/J. CELL. 2015. 11. 013. URL https: //www. sciencedirect. com/science/article/pii/S0092867415014932?via%3Dihub#app3
- Fletcher RB, Das D, Gadye L, Street K, Baudhuin A, Risso D, Wagner A, Cole MB, Flores Q, Choi YG, Yosef N, Purdom E, Dudoit S, Ngai J. Deconstructing Olfactory Stem Cell Trajectories at Single-Cell Resolution. *Cell Stem Cell*. 2017; 20(6): 817–30.
- Trapnell, C. , Cacchiarelli, D. , Grimsby, J. , Pokharel, P. , Li, S. , Morse, M. , … Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. https: //doi. org/10. 1038/nbt. 2859
- Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef ,Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics , 19(1): 477, 12 2018. ISSN 1471-2164. doi: 10. 1186/s12864-018-4772-0. URL https: //bmcgenomics. biomedcentral. com/articles/10. 1186/s12864-018-4772-0
- Wood S. N. , N. Pya and B. Saefken (2016) Smoothing parameter and model selection for general smooth models (with discussion). Journal of the American Statistical Association 111: 1548-1575.
- Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. NatureBiotechnology, page 1, 4 2019. ISSN 1087-0156. doi: 10. 1038/s41587-019-0071-9. URL http: //www. nature. com/articles/s41587-019-0071-9
- Tapio Lonnberg et al. Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. Science immunology, 2(9), 3 2017. doi: 10. 1126/sciimmunol. aal2192. URL http: //www. ncbi. nlm. nih. gov/pubmed/2834507 http: //www. pubmedcentral. nih. gov/articlerender. fcgi?artid=PMC5365145.
- Xiaojie Qiu, 832 Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. Nature Methods, 8 2017. doi10. 1038/nmeth. 4402. URL https: //www. nature. com/nmeth/journal/vaop/ncurrent/full/nmeth. 4402. html.
- Lauren E. Byrnes, Daniel M. Wong, Meena Subramaniam, Nathaniel P. Meyer, Caroline L. Gilchrist, Sarah M. Knox, Aaron D. Tward, Chun J. Ye, and Julie B. Sneddon. Lineage dynamics of murine pancreatic development at single-cell resolution. Nature Communications, 9(1): 3922, 122018. ISSN 2041-1723. doi: 10. 1038/s41467-018-06176-3. URL http: //www. nature. com/articles/s41467-018-06176-3.
- Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research, 40(10): 4288{97, 5 2012. ISSN 1362-4962. doi: 10. 1093/nar/gks042. URL http: //www. pubmedcentral. nih. gov/articlerender. fcgi?artid=3378882&tool=pmcentrez&rendertype=abstract

Citations

- David S Fischer, Fabian J Theis, and Nir Yosef. Impulse model-based differential expression analysis of time course sequencing data. Nucleic Acids Research, 46(20): e119{e119, 8 2018. ISSN 0305-1048. doi: 10. 1093/nar/gky675. URL https: //academic. oup. com/nar/advance-article/doi/10. 1093/nar/gky675/5068248.
- Elizabeth Purdom and Davide Risso (2019). clusterExperiment: Compare Clusterings for Single-Cell Sequencing. R package version 2. 6. 1
- Davide Risso, Liam Purvis, Russell B. Fletcher, Diya Das, John Ngai, Sandrine Dudoit, and Elizabeth Purdom. clusterExperiment and RSEC: A Bioconductor package and framework for clustering of single cell and other large gene expression datasets. PLOS Computational Biology, 14(9): e1006378, 9 2018b. ISSN 1553-7358. doi: 10. 1371/journal. pcbi. 1006378. URL http: //dx. plos. org/10. 1371/journal. pcbi. 1006378.
- William M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336): 846—850, 1971. doi: 10. 1080/01621459. 1971. 10482356. URL https: //www. tandfonline. com/doi/abs/10. 1080/01621459. 1971. 10482356
- Lawrence Hubert and Phipps Arabie. Comparing partitions. Journal of Classification, 2(1): 193—218, Dec 1985. ISSN 1432-1343. doi: 10. 1007/BF01908075. URL https: //doi. org/10. 1007/BF01908075
- Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data. Nature Methods, 14(5): 483—486, may 2017. ISSN 1548-7091. doi: 10. 1038/nmeth. 4236. URL http: //www. nature. com/articles/nmeth. 4236
- Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive Integration of Single-Cell Data. Cell, 177(7): 1888—1902. e21, jun 2019. ISSN 10974172. doi: 10. 1016/j. cell. 2019. 05. 031. URL http: //www. ncbi. nlm. nih. gov/pubmed/31178118http: //www. pubmedcentral. nih. gov/ articlerender. fcgi?artid=PMC6687398
- Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology, 37(1): 38—44, jan 2019. ISSN 1087-0156. doi: 10. 1038/nbt. 4314. URL http: //www. nature. com/articles/nbt. 4314
- Patil, P. , Peng, R. D. & Leek, J. T. A visual tool for defining reproducibility and replicability. *Nat Hum Behav* 3**,** 650—652 (2019) doi: 10. 1038/s41562-019-0629-z
- Baron, M. , Veres, A. , Wolock, S. L. , Faust, A. L. , Gaujoux, R. , Vetere, A. , … Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, *3*(4), 346-360. e4. https: //doi. org/10. 1016/j. cels. 2016. 08. 011
- Segerstolpe, Å. , Palasantza, A. , Eliasson, P. , Andersson, E. M. , Andréasson, A. C. , Sun, X. , … Sandberg, R. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, *24*(4), 593—607. https: //doi. org/10. 1016/j. cmet. 2016. 08. 020
- Crow, M. , Paul, A. , Ballouz, S. , Huang, Z. J. , & Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nature Communications*, *9*(1), 884. https: //doi. org/10. 1038/s41467-018-03282-0
- Risso, D. , Perraudeau, F. , Gribkova, S. *et al.* A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* 9**,** 284 (2018) doi: 10. 1038/s41467-017-02554-5