Single-cell transcriptomics in plants

Standard and advanced scRNA-seq data analysis workflows

Michael Schon

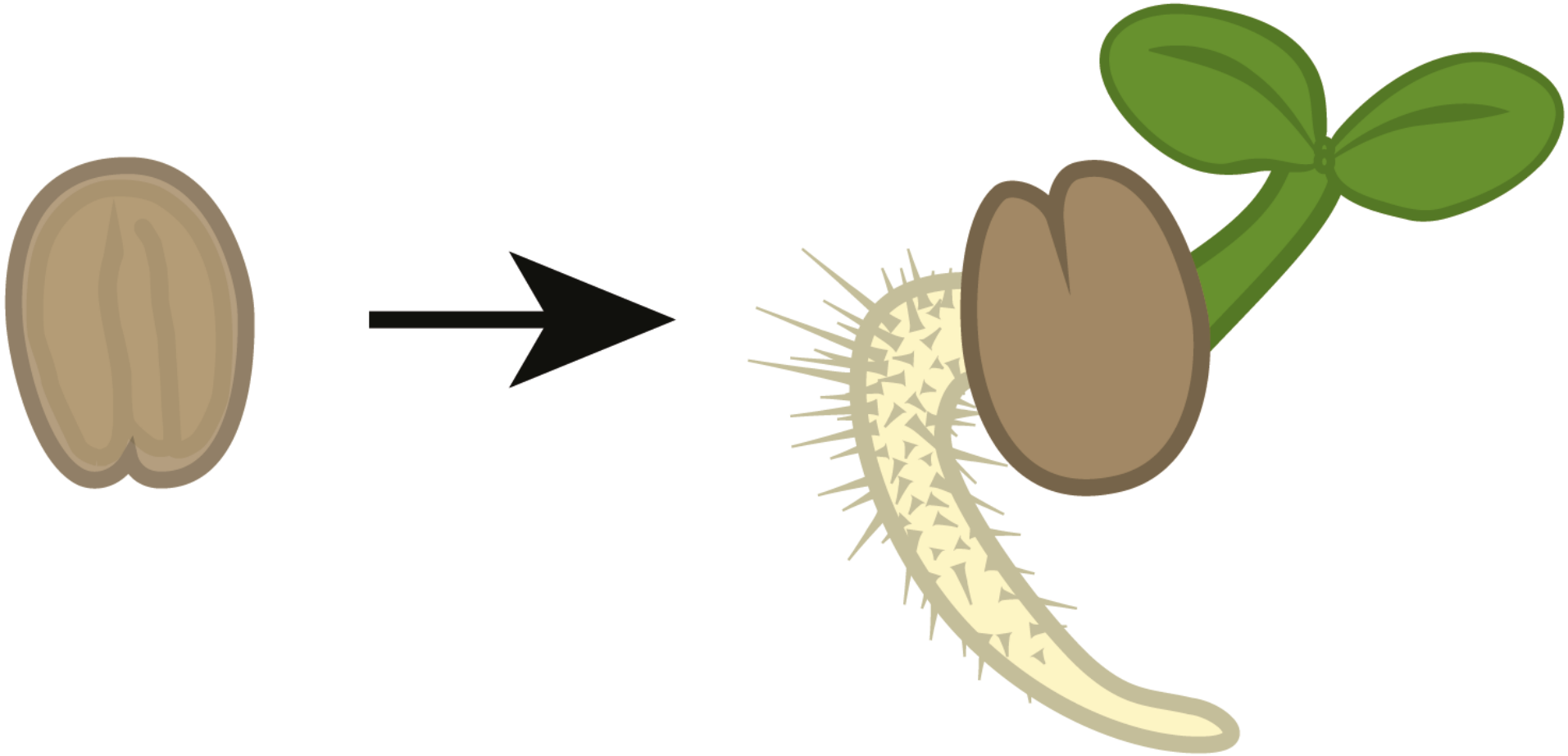
Postdoc, Nodine Group, Laboratory of Molecular Biology
Wageningen University & Research
23 March 2023

WAGENINGEN
UNIVERSITY & RESEARCH

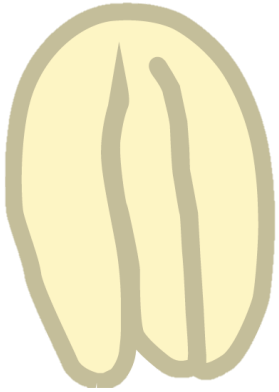
Overview

- Why single cells?
- How-to:
  - Biology → Hardware → Chemistry → Software → *Knowledge?*
- The standard workflow
  - Reads → Counts → Clusters → Markers
  - Dimensionality reduction
- Advanced techniques

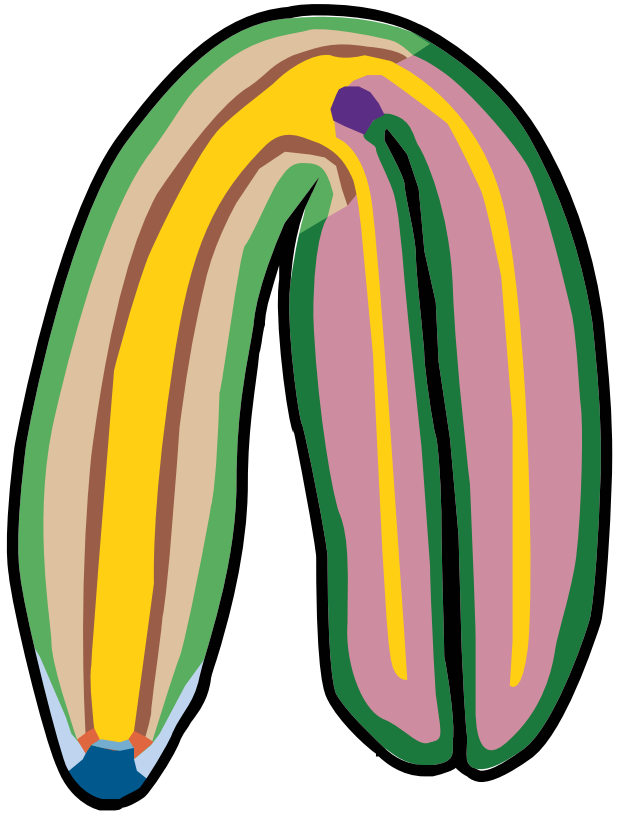
Single cell RNA-seq: Why single cells?



Single cell RNA-seq: Why single cells?

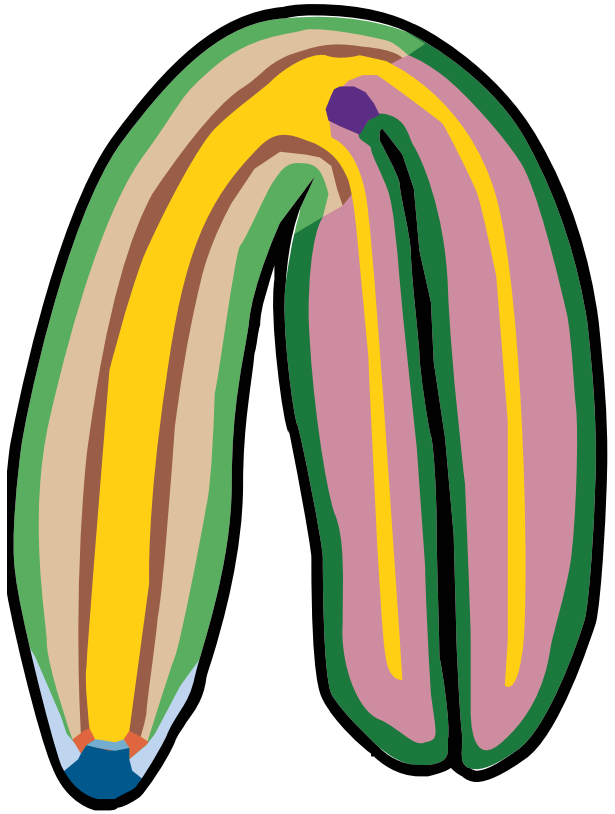


- Vascular
- Ground tissue
- Hypophysis
- Cotyledon
- Shoot meristem
- Cortex
- Endodermis
- Quiescent center
- Columella
- Lateral root cap

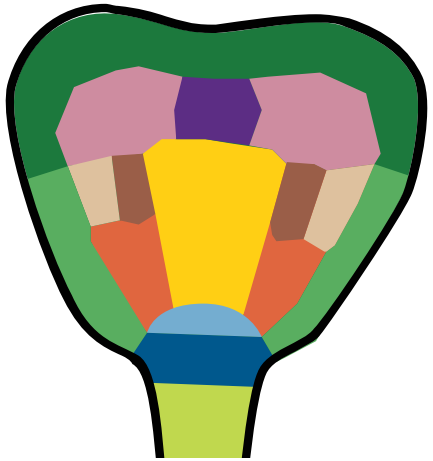


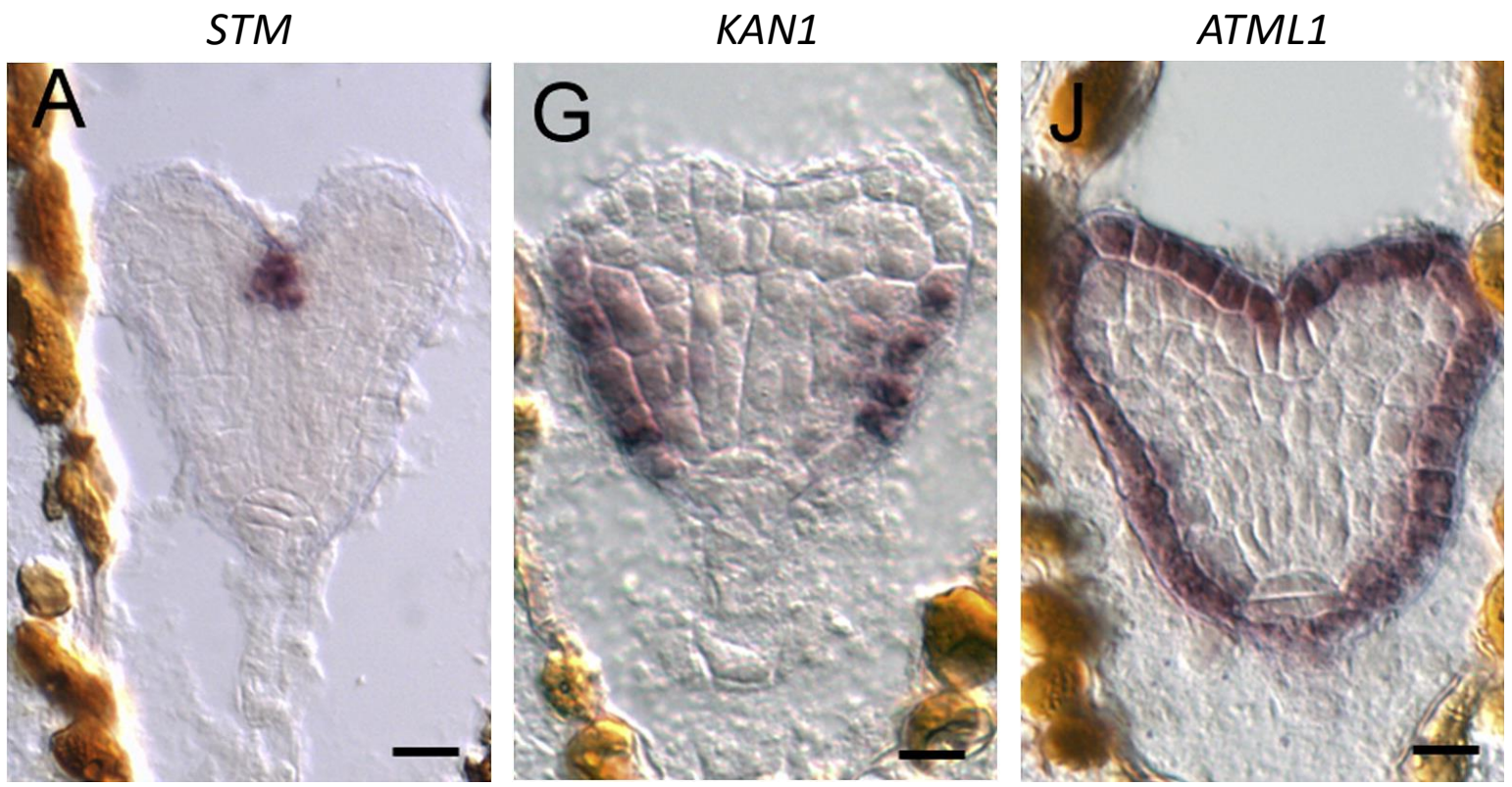
Single cell RNA-seq: Why single cells?

- Vascular
- Ground tissue
- Hypophysis
- Cotyledon
- Shoot meristem
- Cortex
- Endodermis
- Quiescent center
- Columella
- Lateral root cap



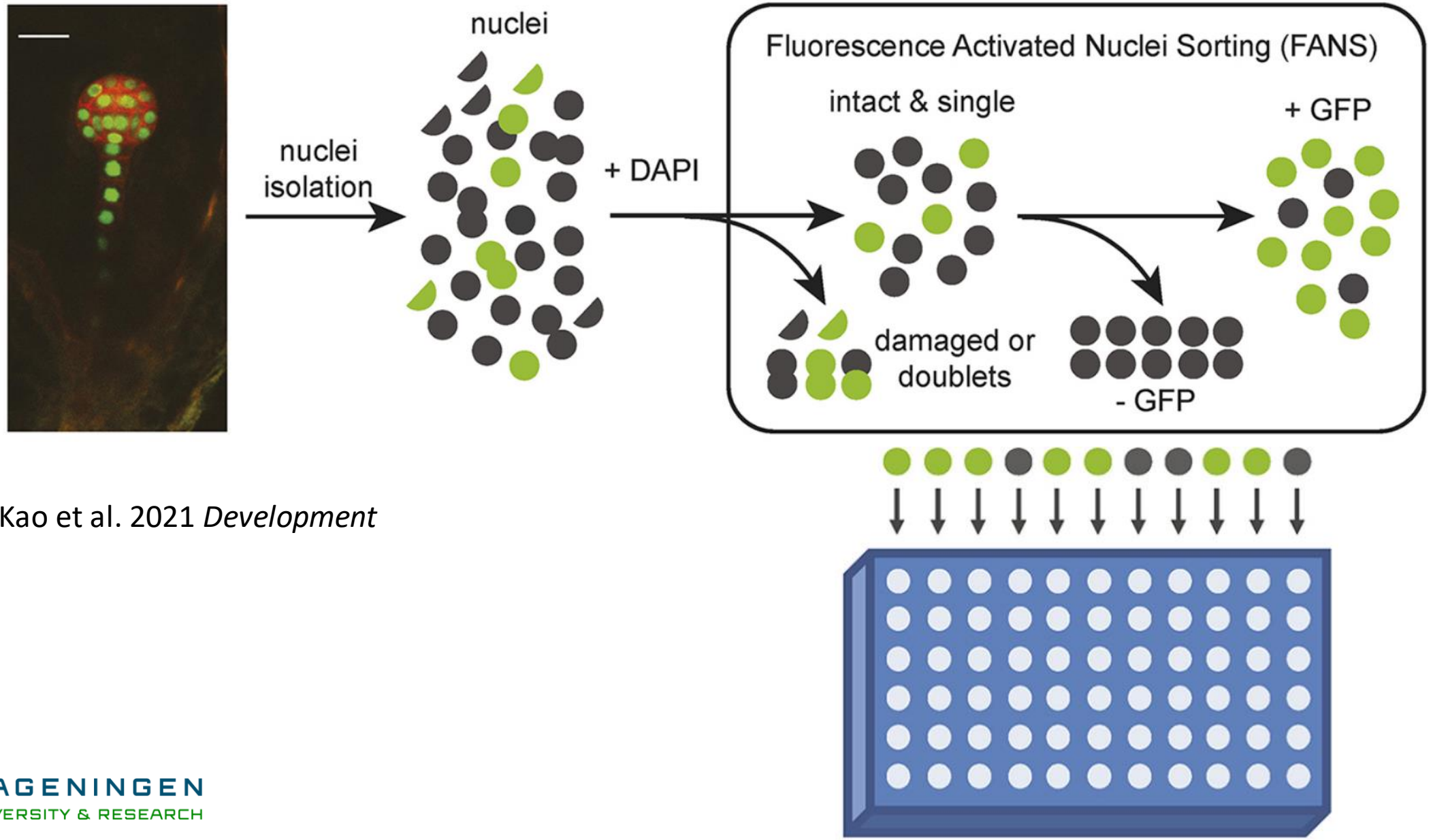
Single cell RNA seq: Why RNA?





Grigg et al. 2009 *Current Biology*

Capturing cells



Kao et al. 2021 *Development*

Capturing the transcriptome

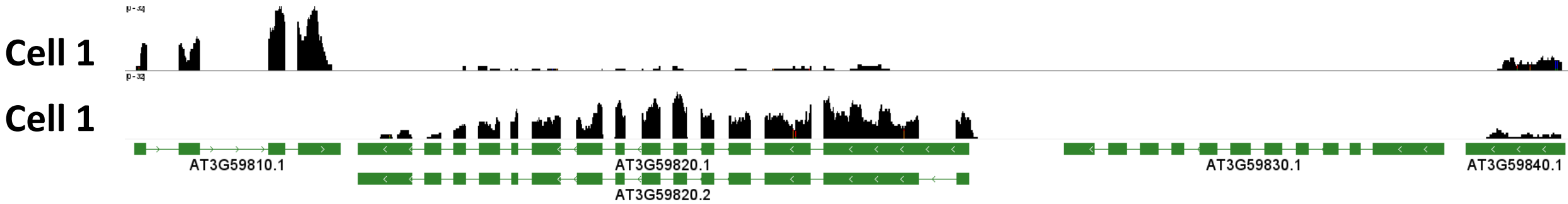
Capturing the transcriptome

	Protocol	Total RNA Input (ng)
Unstranded RNA-seq	TruSeq	≥100
	Ovation	≥0.5
	SMARTer	≥0.01 (single cell)
	Smart-seq2	≥0.01 (single cell)
Strand-specific RNA-seq	Not-so-random (NSR)	≥1,000
	TruSeq Stranded (dUTP)	≥100
	SMART-Seq Stranded	≥10
Barcoded 3′ end sequencing	QuantSeq	≥1
	10X Chromium	≥0.01 (single cell)
	MARS-seq	≥0.01 (single cell)
Poly(A) site sequencing	3P-seq	≥20,000
	TAIL-seq	≥100,000
	mTAIL-seq	≥50
5′ cap sequencing	CAGE	≥50,000
	nanoCAGE	≥50
	STRT-seq	≥0.01 (single cell)
Degradome (5′P) sequencing	PARE	≥50,000
	GMUCT2	≥5,000
	5P-seq	≥5,000
Long-read sequencing	Pacbio Iso-Seq	≥300
	ONT Direct RNA	~25,000 (≥500ng polyA+)

WAGENINGEN
UNIVERSITY & RESEARCH

RNA x Cells... Now what?

- Align sequenced reads with transcripts
  - Genome alignments (STAR, HiSat)
  - Pseudoalignment (Kallisto, Salmon)



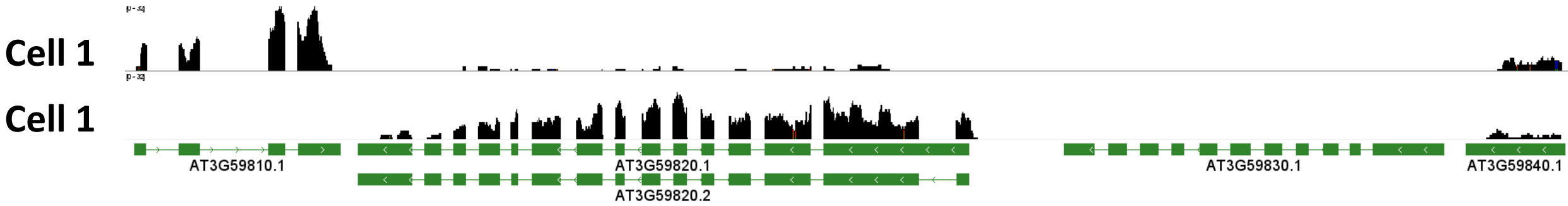
	Cell 1	Cell 2
AT3G59810	110	0
AT3G59820	20	304
AT3G59830	0	0
AT3G59840	22	13

More information in Practical 1!

WAGENINGEN
UNIVERSITY & RESEARCH

RNA x Cells... Now what?

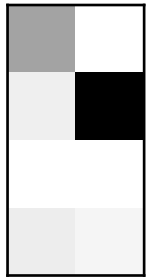
- Align sequenced reads with transcripts
  - Genome alignments (STAR, HiSat)
  - Pseudoalignment (Kallisto, Salmon)



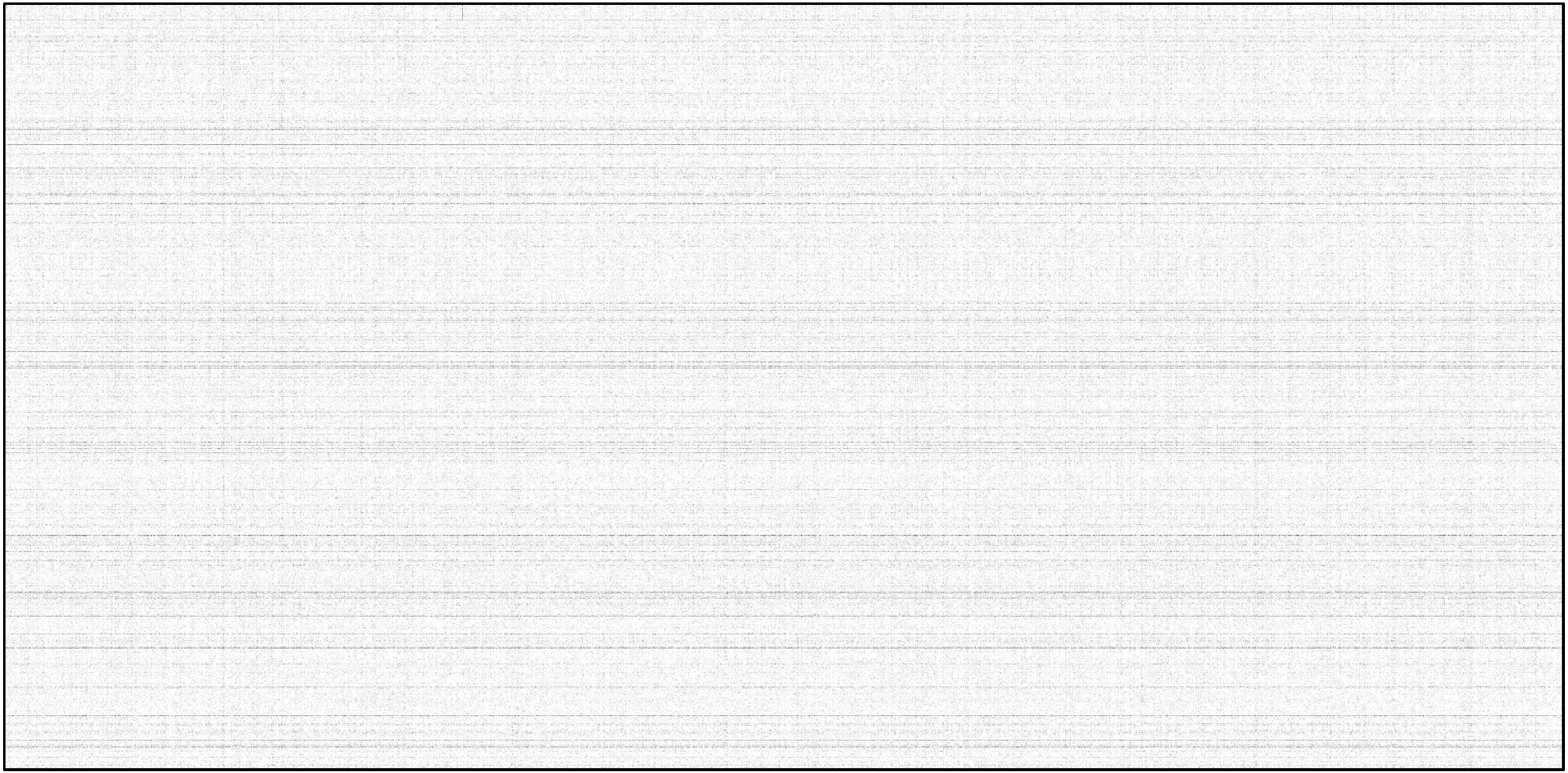
	Cell 1	Cell 2
AT3G59810	110	0
AT3G59820	20	
AT3G59830	0	0
AT3G59840	22	13

WAGENINGEN
UNIVERSITY & RESEARCH

RNA x Cells... Now what?

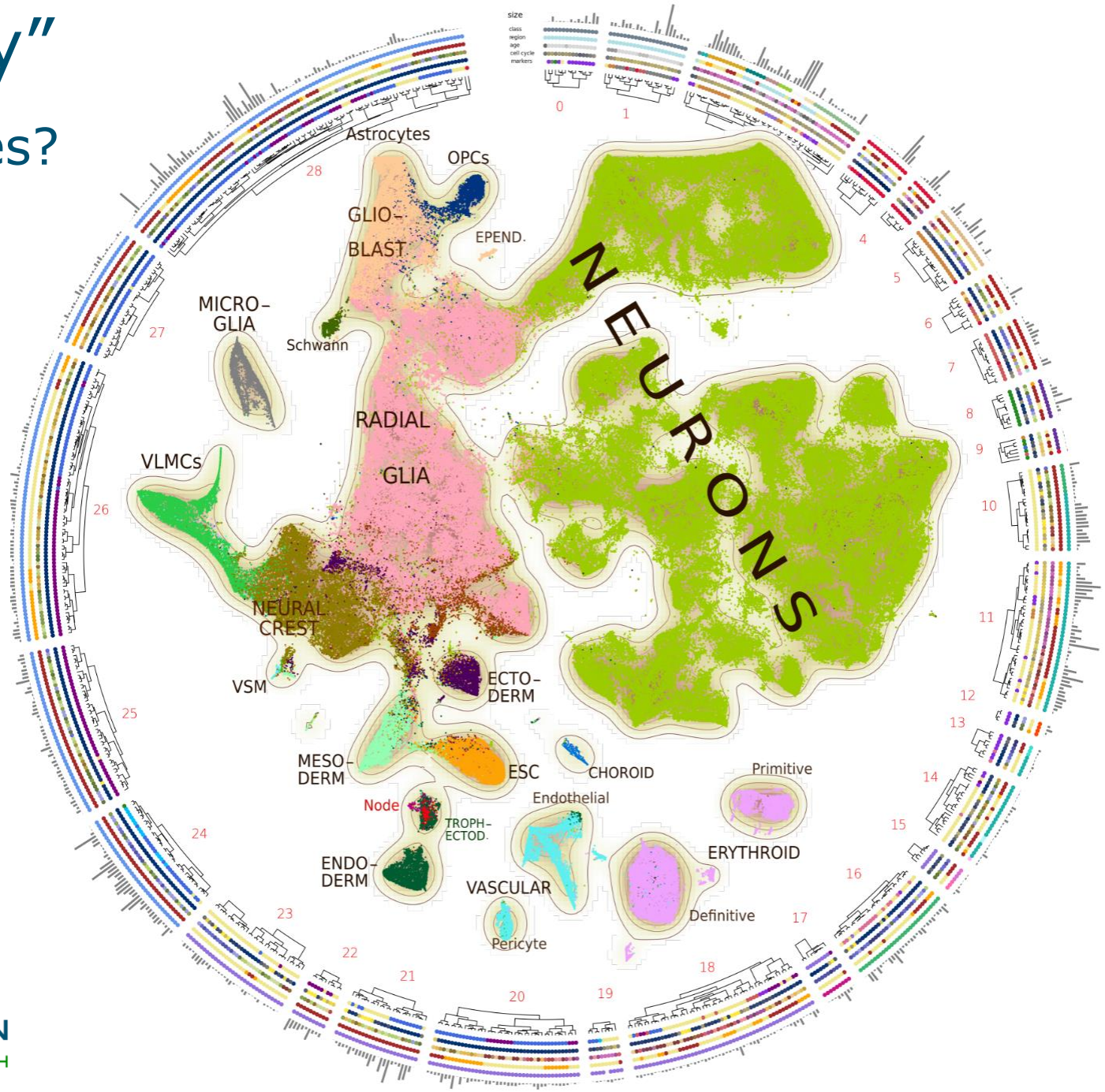


RNA x Cells... Now what?



“Similarity”

- Marker genes?
  - *NANOG*

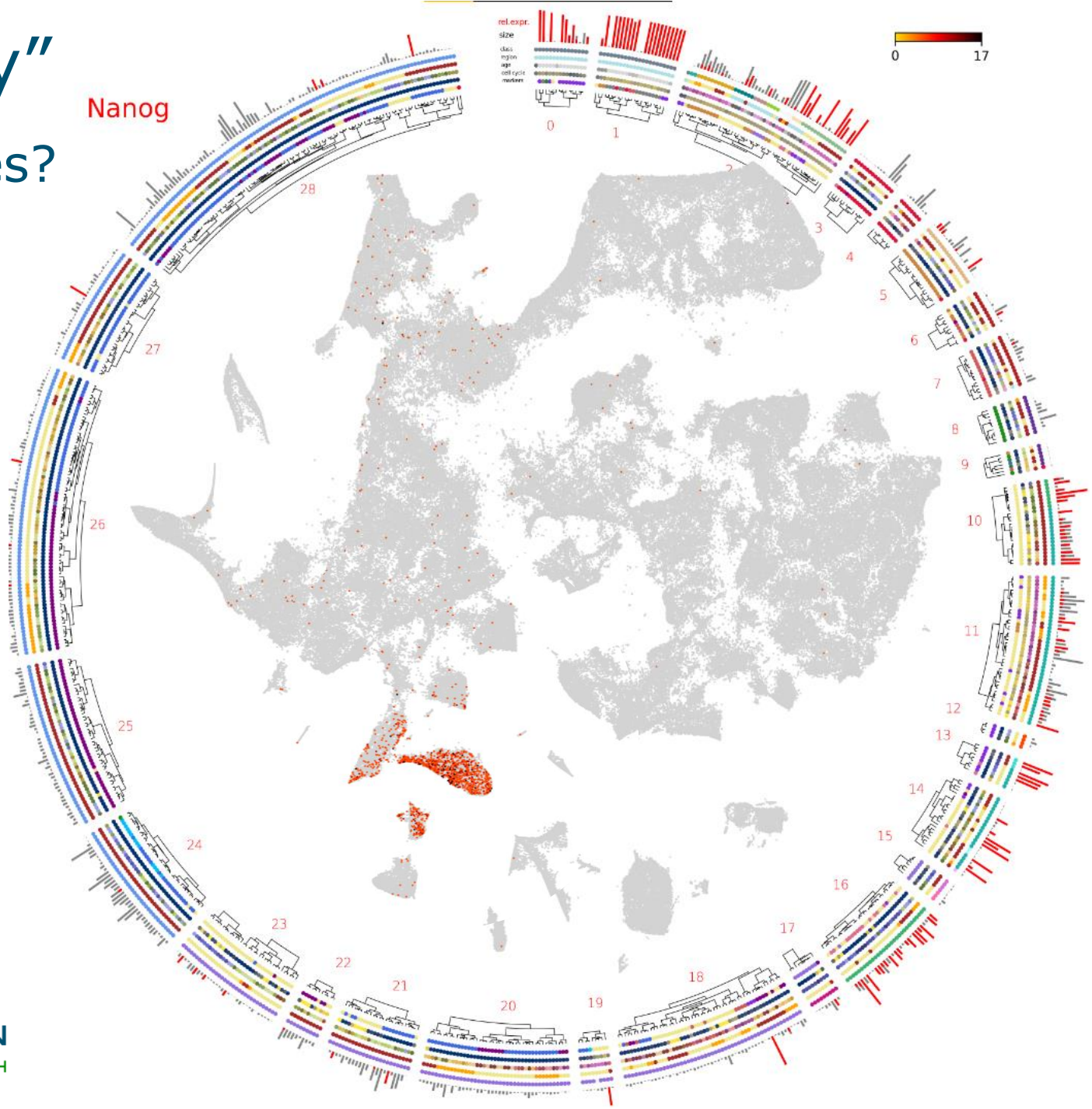


WAGENINGEN
UNIVERSITY & RESEARCH

La Manno et al. 2021 *Nature*

“Similarity”

- Marker genes?
  - *NANOG*



WAGENINGEN
UNIVERSITY & RESEARCH

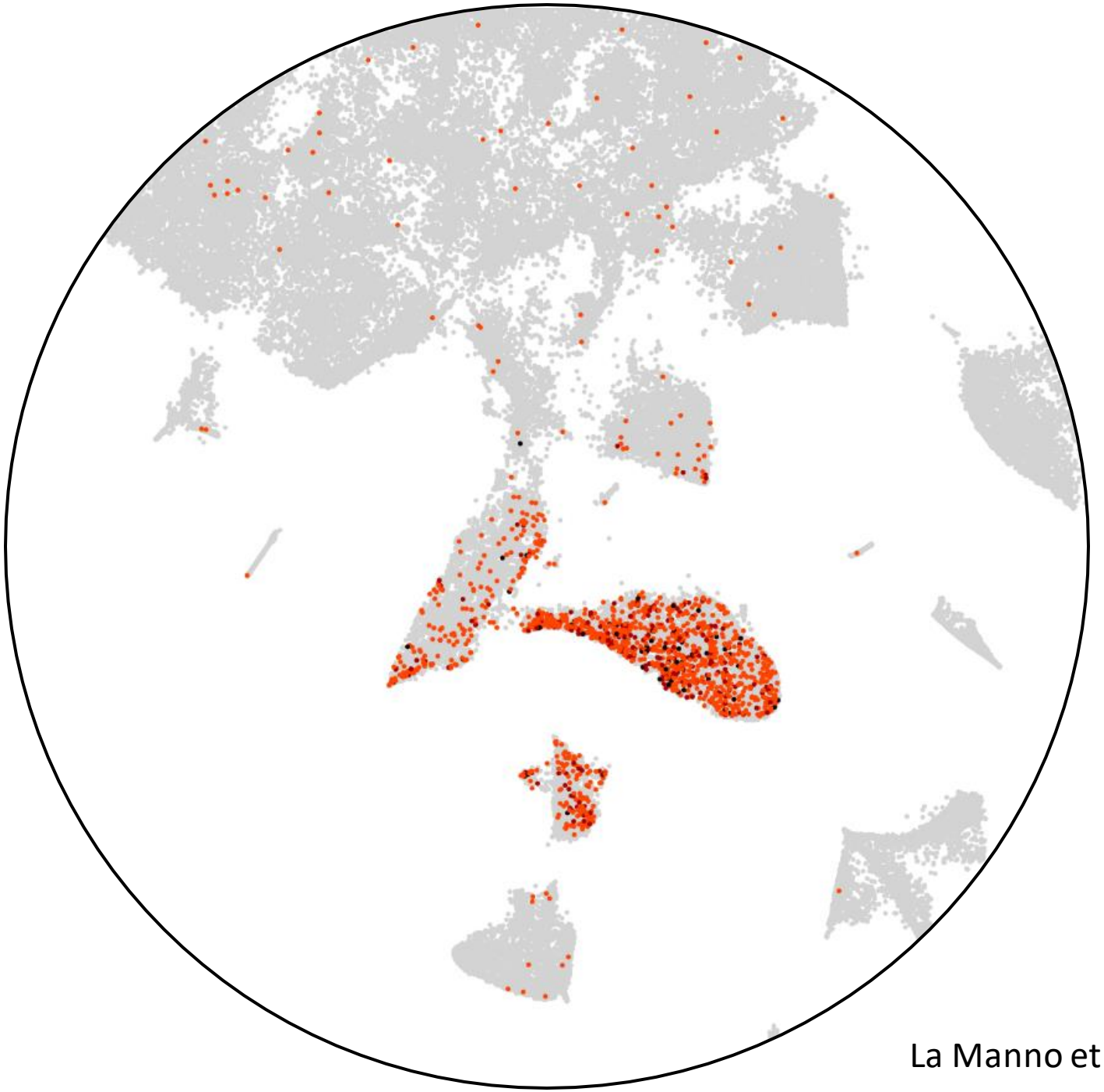
La Manno et al. 2021 *Nature*

“Similarity”

- Marker genes?
  - *NANOG*

⚠ Not all ESCs have *NANOG*

⚠ Not all *NANOG*+ cells are ESCs



WAGENINGEN
UNIVERSITY & RESEARCH

La Manno et al. 2021 *Nature*

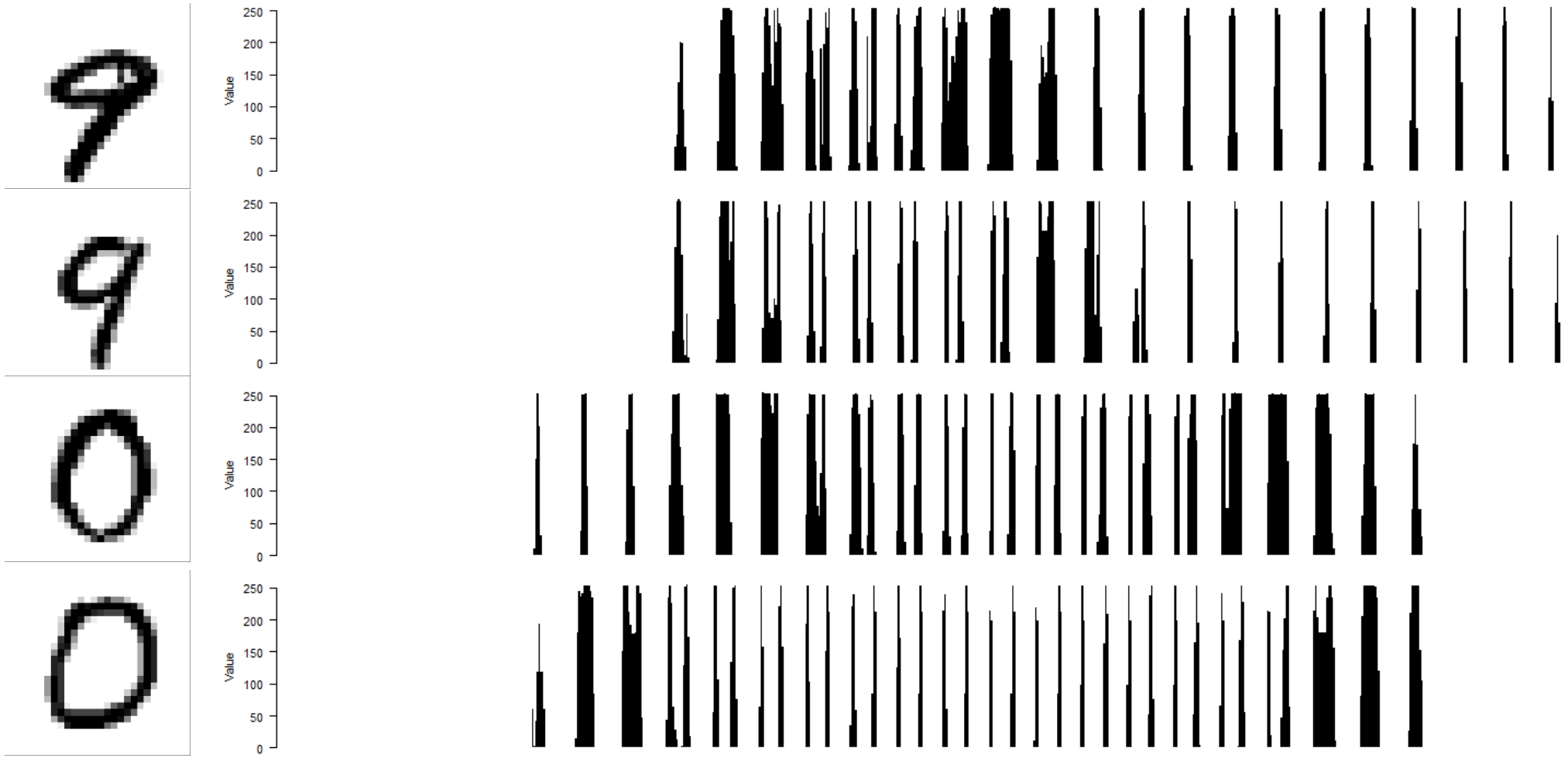
Neighbors

- t-SNE: stochastic neighbor embedding
  - Define neighbors in high-D space
  - Minimize a loss function
  - Many-body simulation
  - High initial learning rate (should scale with input)
- UMAP: uniform manifold approximation
  - Also minimizes neighbor distance, but by “folding space”

Neighbors in 784 dimensions

MNIST (70,000 hand-drawn digits, 28x28 pixels)

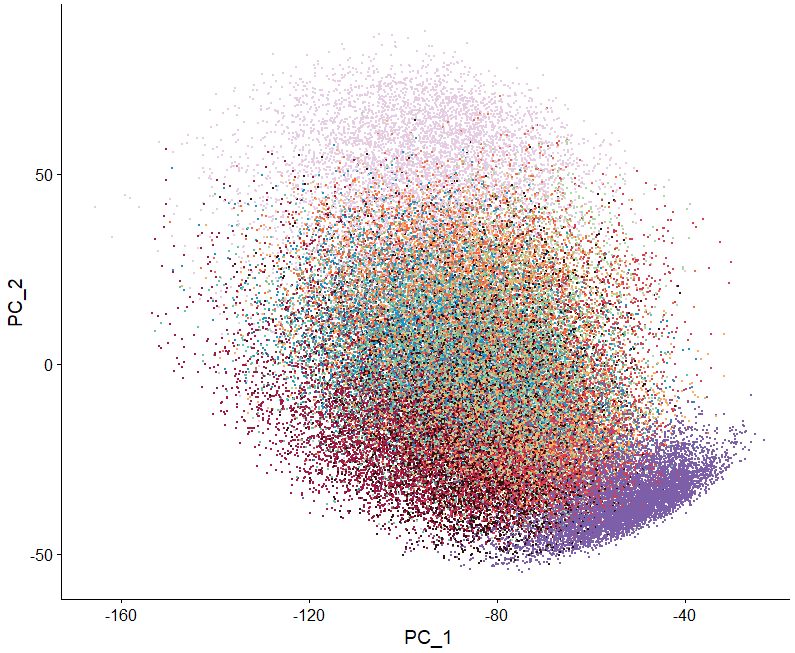
What dimensions are we reducing?

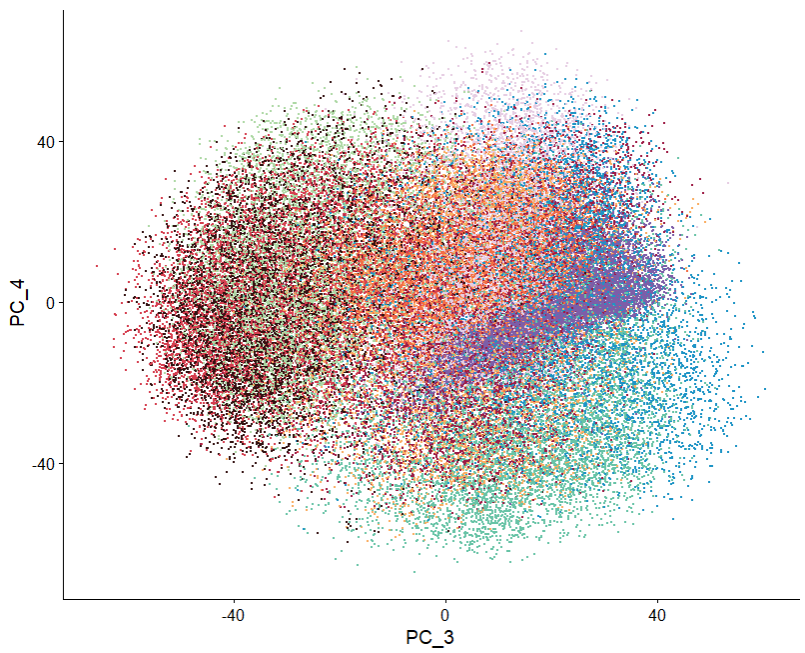


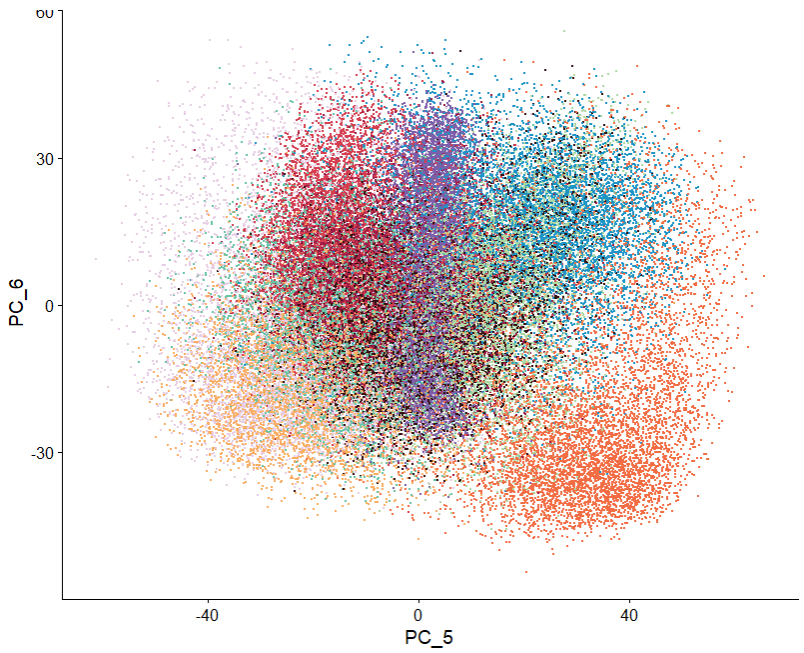
Pixel values

What dimensions are we reducing?

- "Traditional" DR technique: Principal component analysis (PCA)



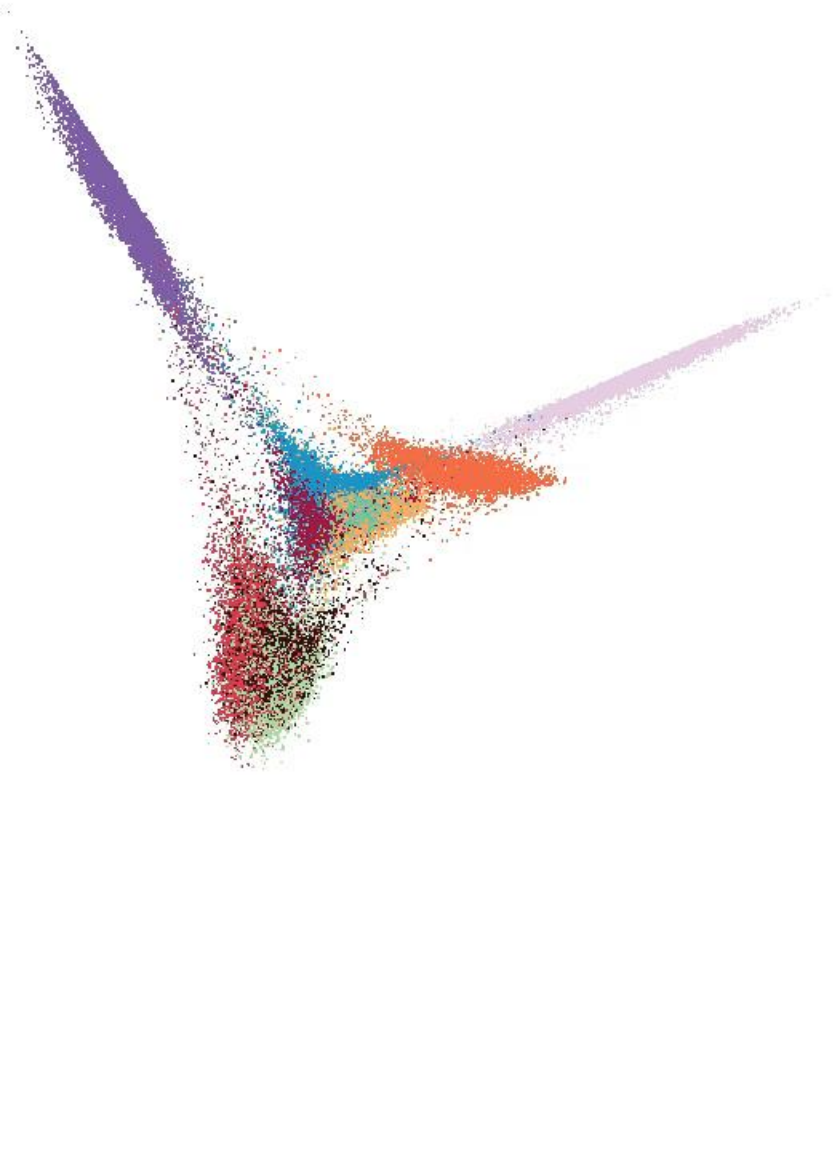




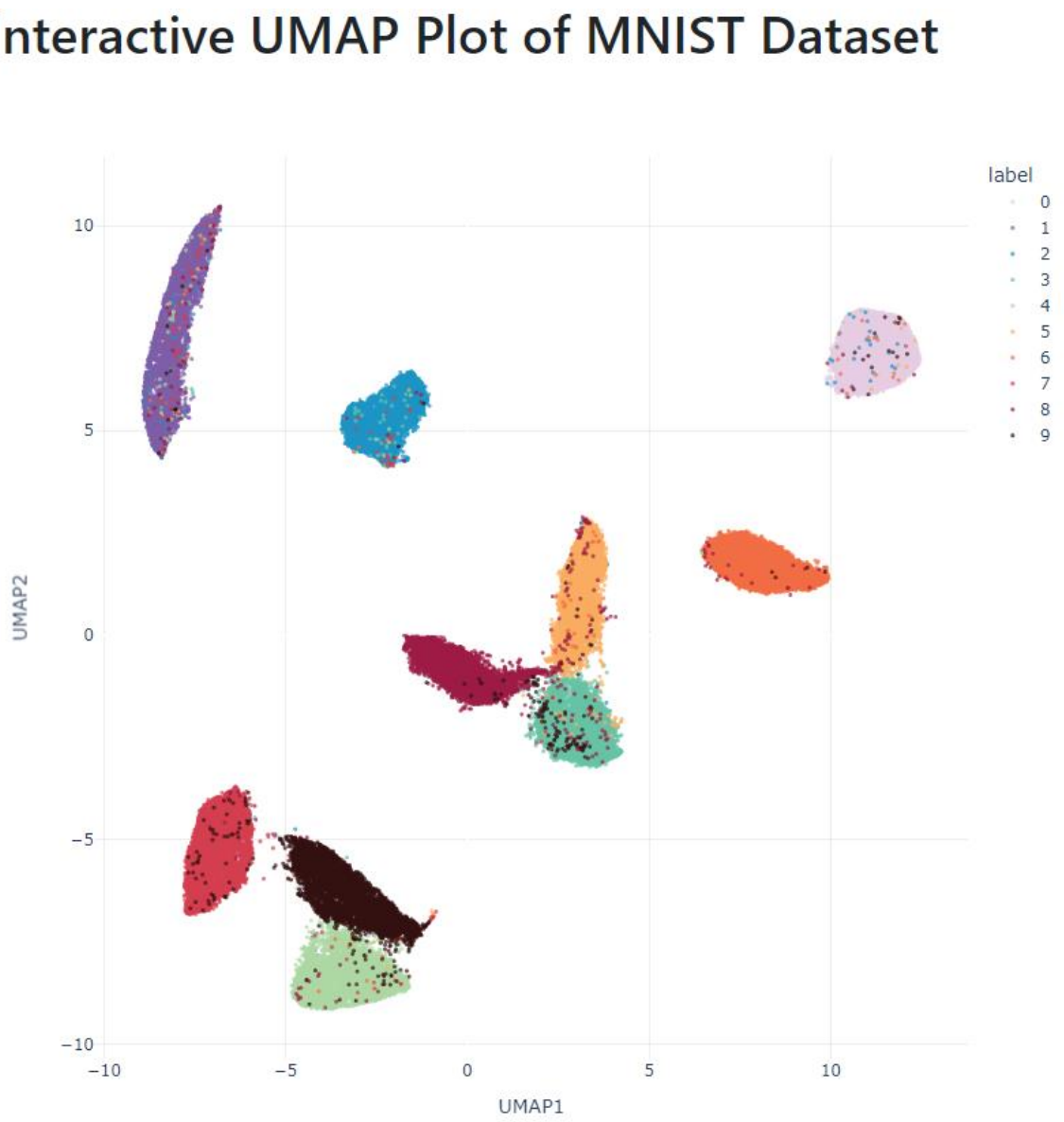
WAGENINGEN
UNIVERSITY & RESEARCH

Dimensional reduction (UMAP)

- Begin with an embedding
  - PCA
  - "Spectral" embedding (Laplacian eigenvectors)
- Minimize a loss function that penalizes high distances between neighbors
- "High-energy" early, followed by annealing

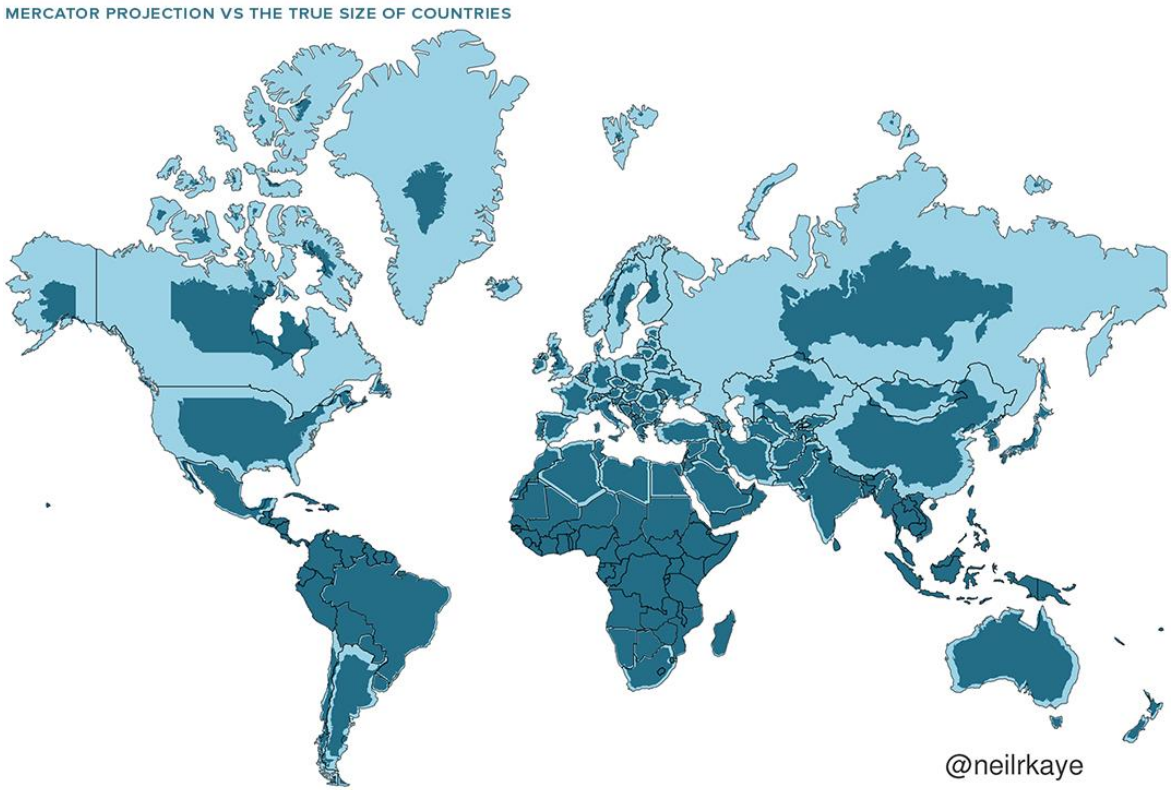


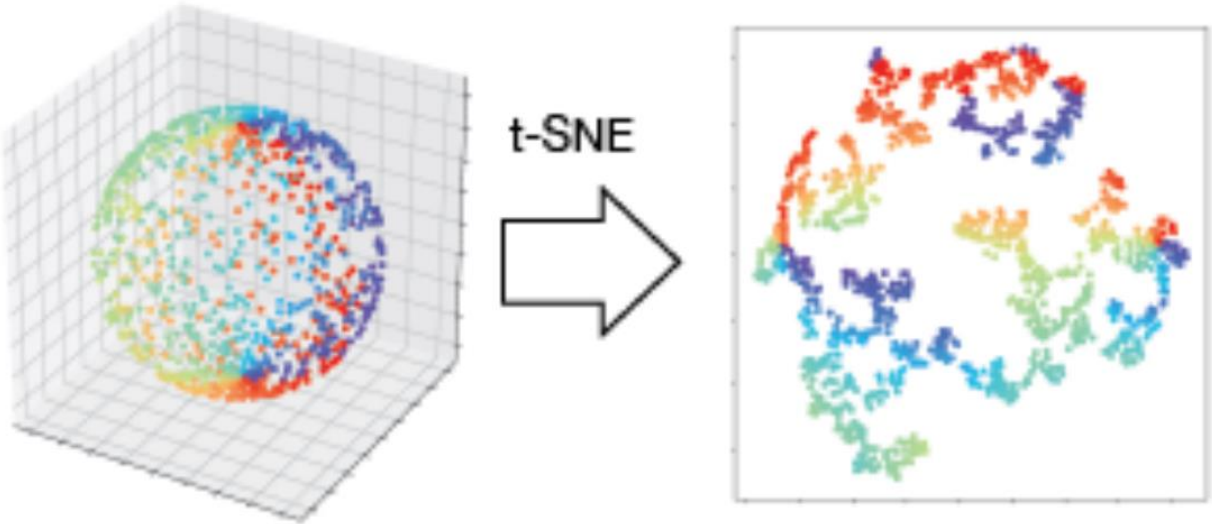
Dimensional reduction (UMAP)



WAGENINGEN
UNIVERSITY & RESEARCH

⚠ Reducing dimensions distorts distance





WAGENINGEN
UNIVERSITY & RESEARCH

Cooley et al. 2021 *bioRxiv*

Initialization matters!

Kobak & Linderman 2021 *Nat Biotech*

Finding causes of variation

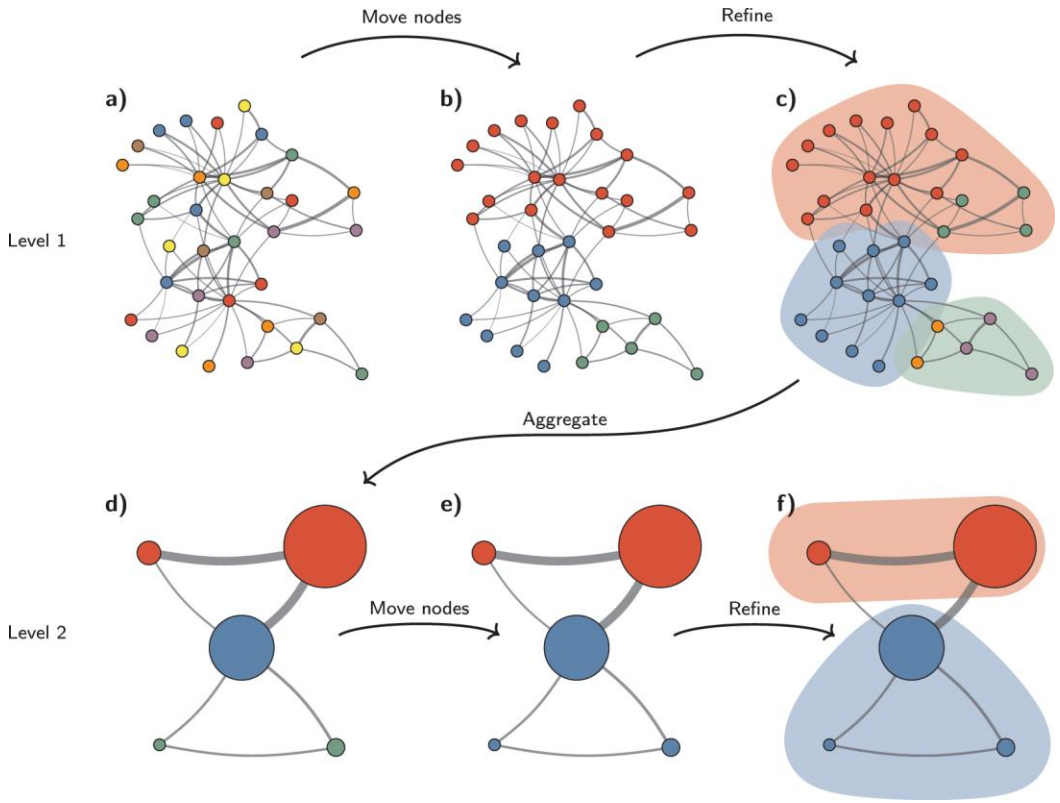
- Differences will either be **discrete** or **continuous**
  - Two unrelated cell types = **discrete**
    - Clustering → marker genes
  - Cellular differentiation = **continuous**
    - Pseudotime trajectories

⚠️ Not all variation is good variation!

- Mitochondrial / chloroplast RNA
- Doublets
- Dropouts
- Ambient RNA
- Cell cycle

Clustering

- Leiden algorithm
  - Graph traversal of k-nearest neighbors
  - Identifies well-connected communities
  - Higher resolution → more clusters

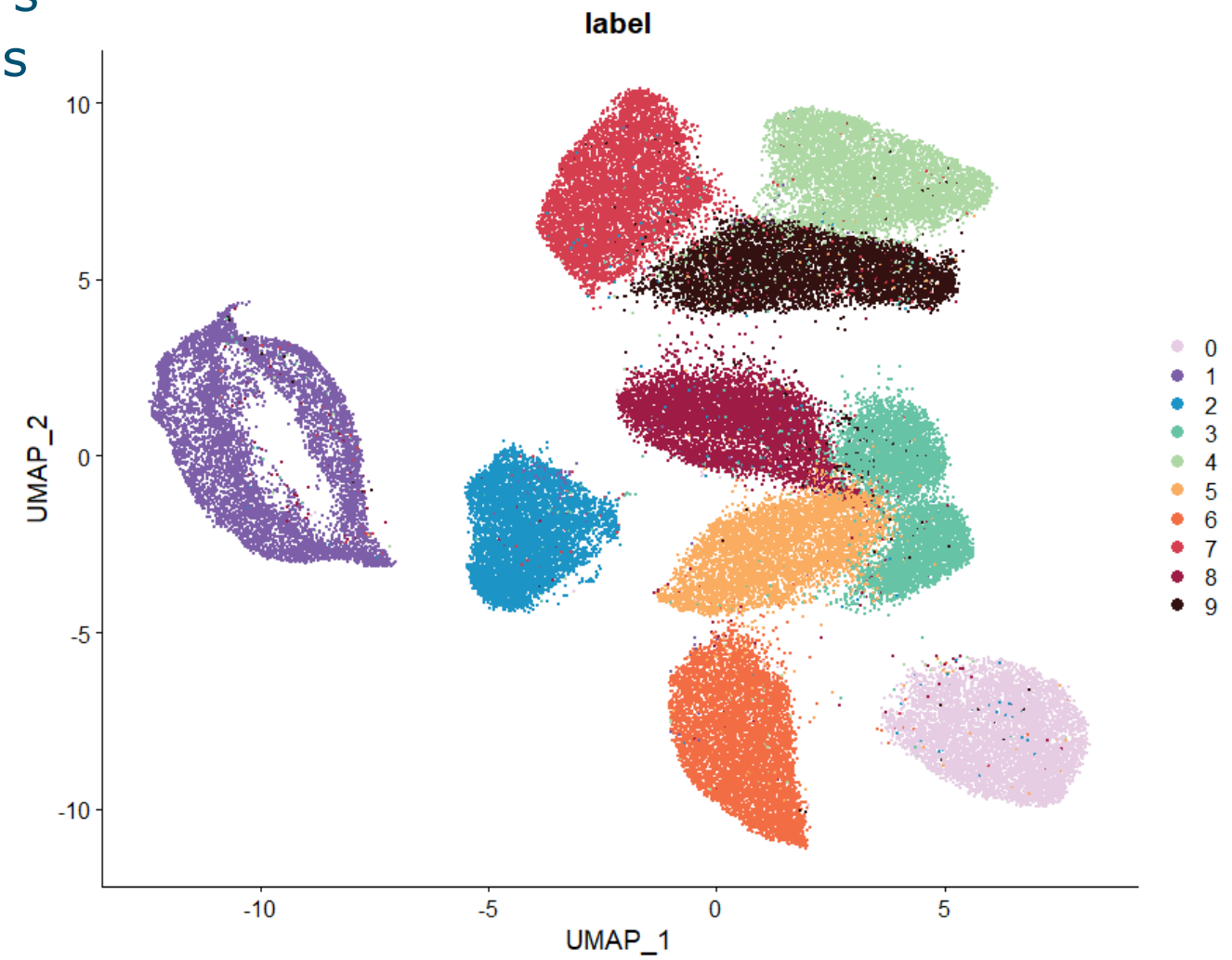


WAGENINGEN
UNIVERSITY & RESEARCH

Traag et al. 2019 *Sci Rep*

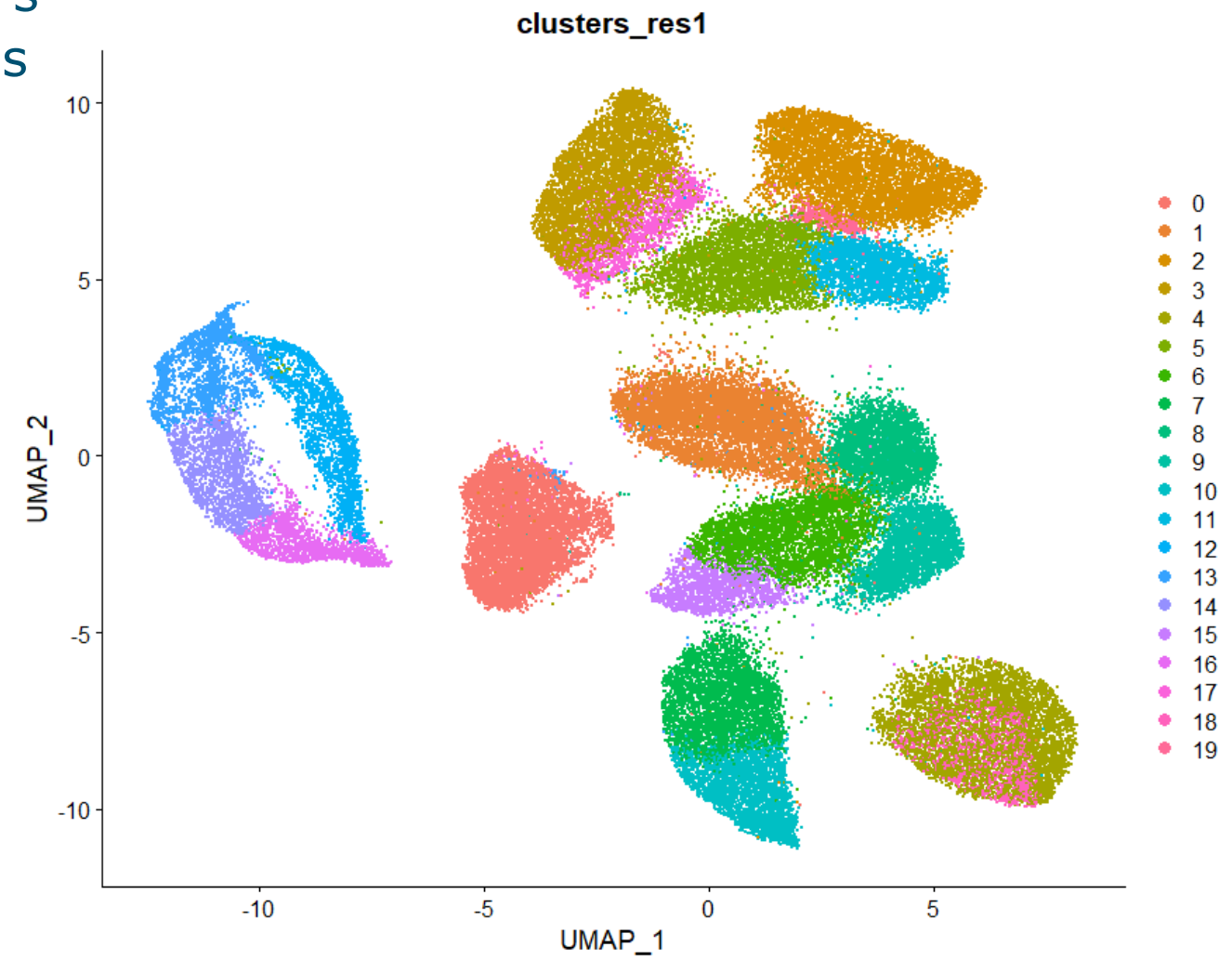
Clustering

- Leiden algorithm
  - Graph traversal of k-nearest neighbors
  - Identifies well-connected communities
  - Higher resolution → more clusters



Clustering

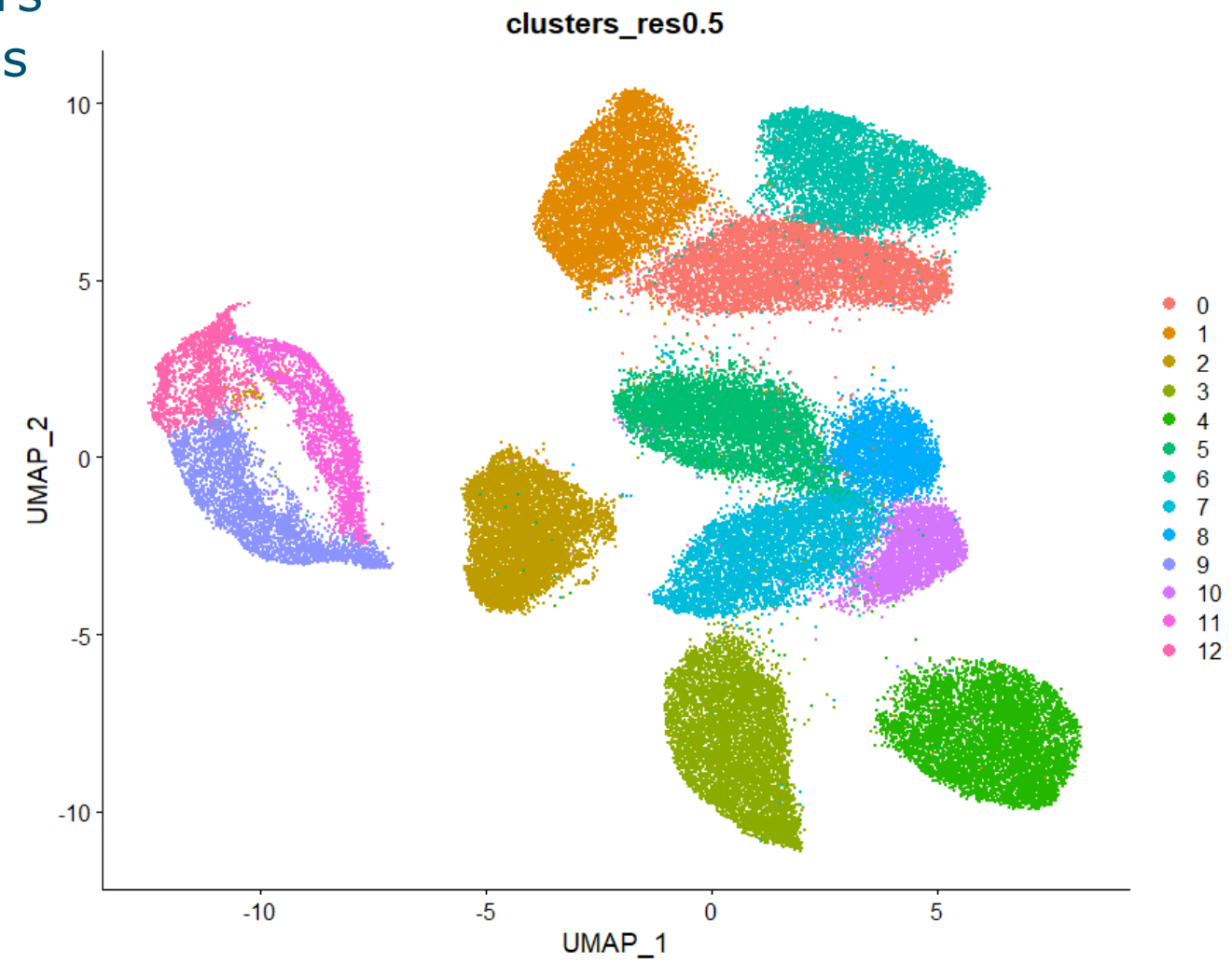
- Leiden algorithm
  - Graph traversal of k-nearest neighbors
  - Identifies well-connected communities
  - Higher resolution → more clusters



WAGENINGEN
UNIVERSITY & RESEARCH

Clustering

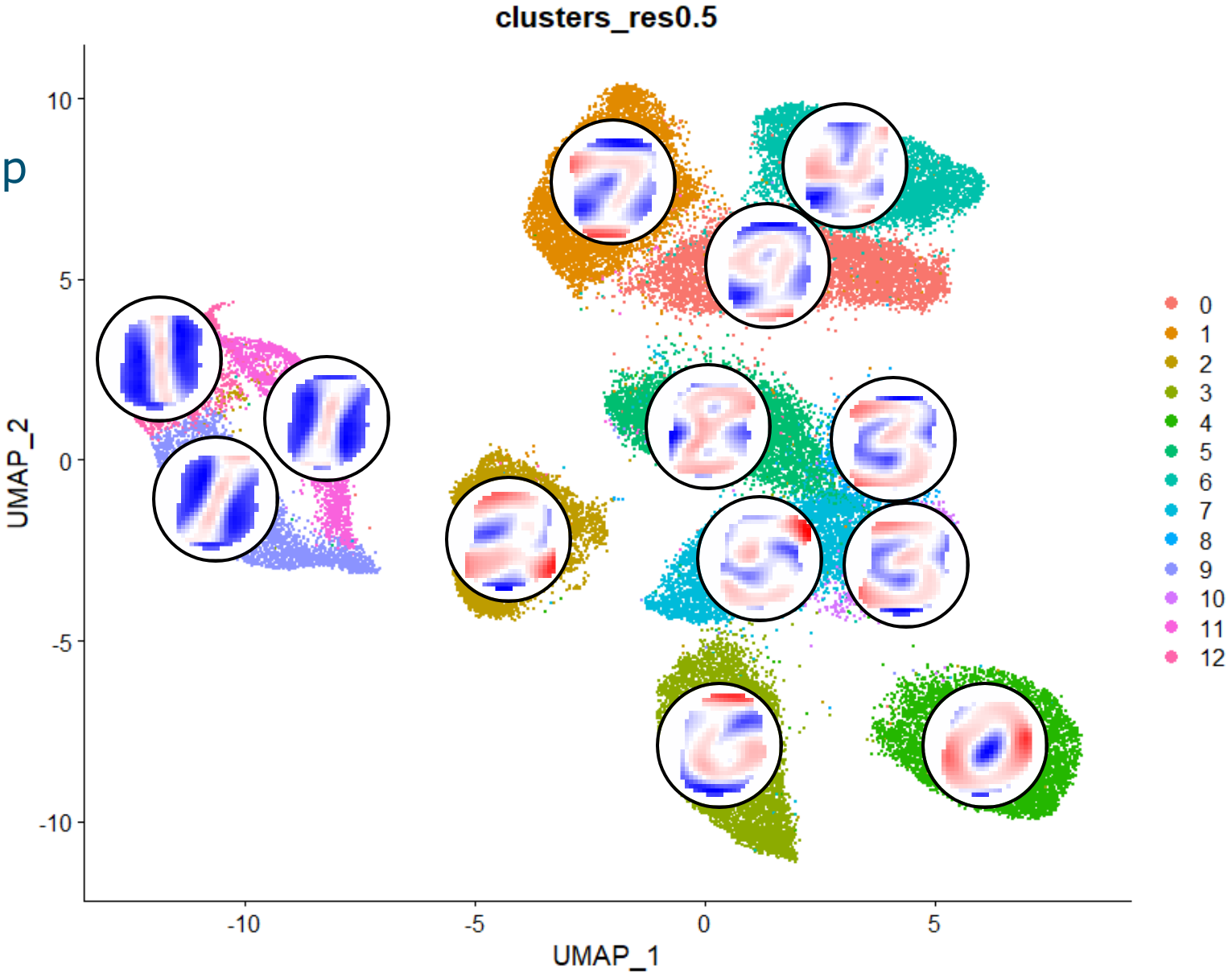
- Leiden algorithm
  - Graph traversal of k-nearest neighbors
  - Identifies well-connected communities
  - Higher resolution → more clusters



WAGENINGEN
UNIVERSITY & RESEARCH

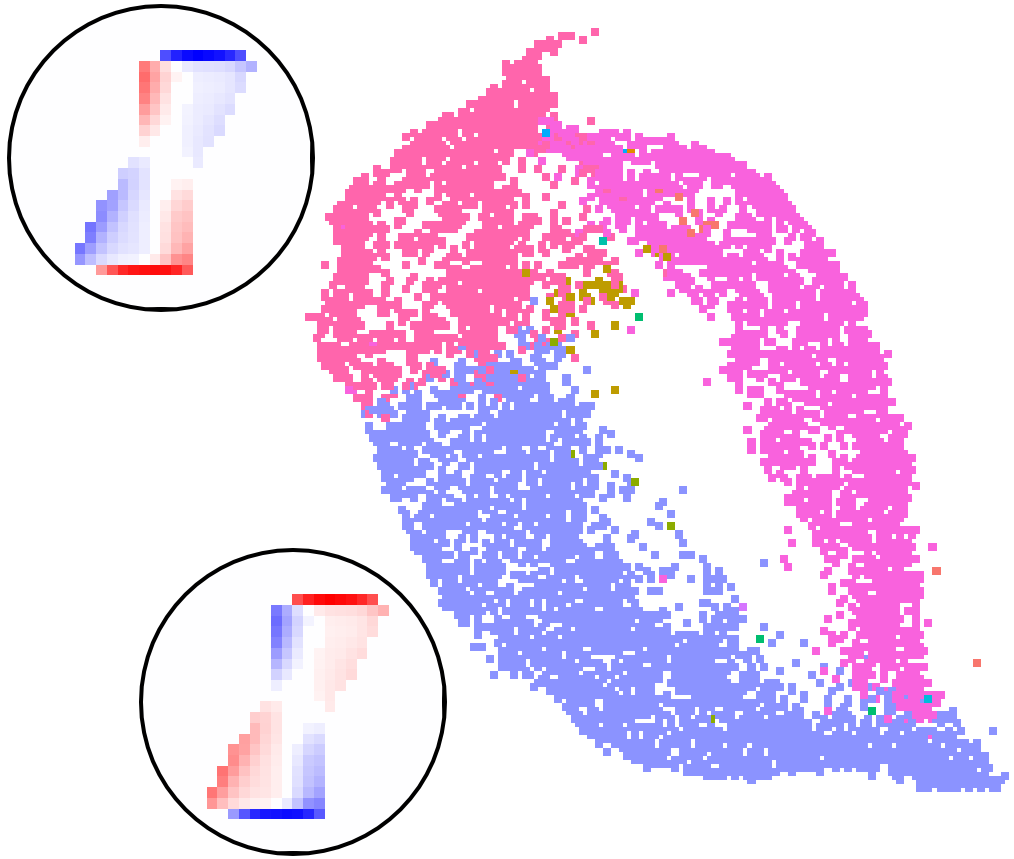
Marker genes

- One vs. all
  - Red – higher in cluster
  - Blue – higher in outgroup



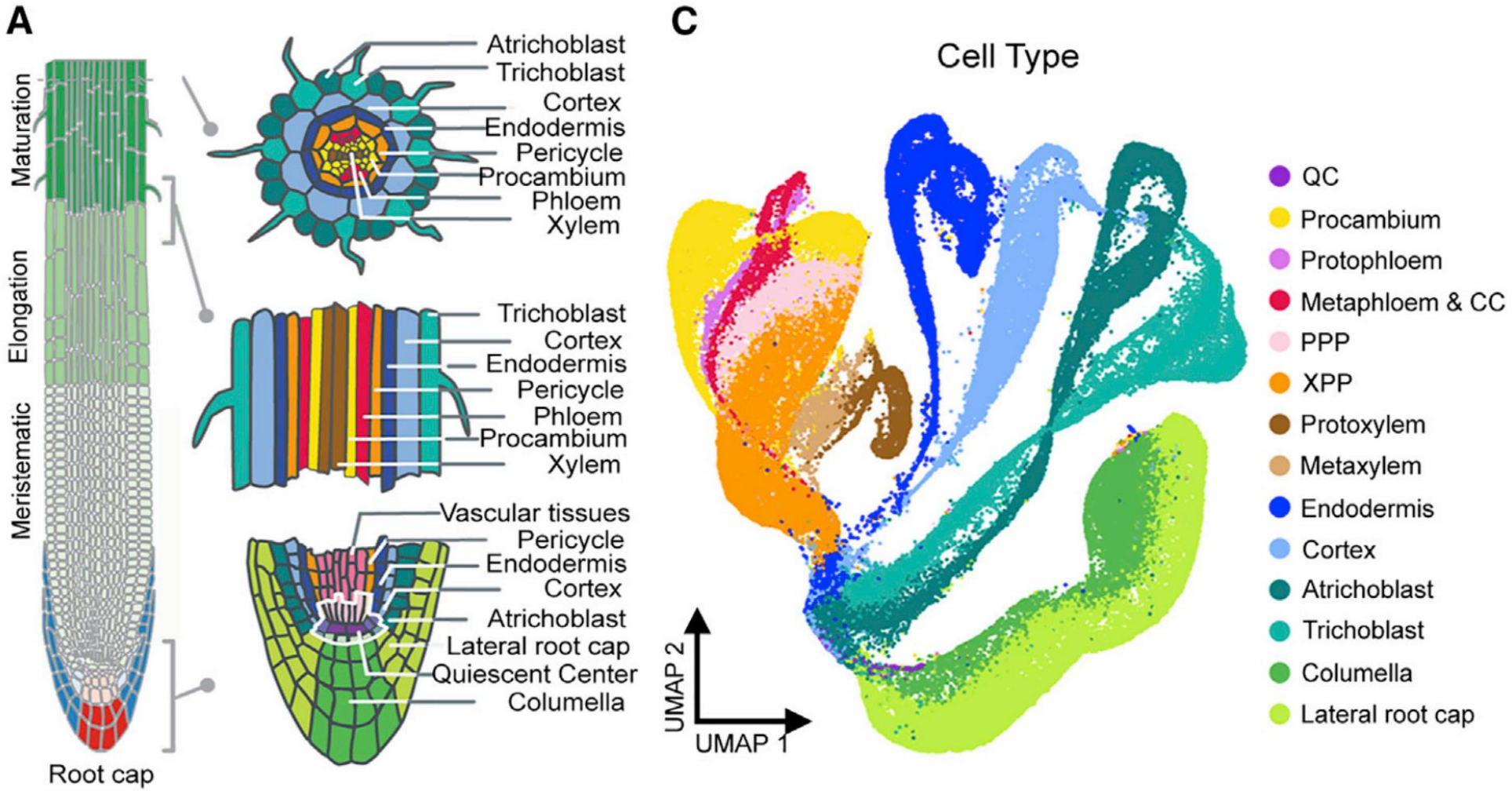
Marker genes

- One cluster vs. another
  - All other clusters are irrelevant
  - A vs. B = -(B vs. A)



Pseudotime trajectories

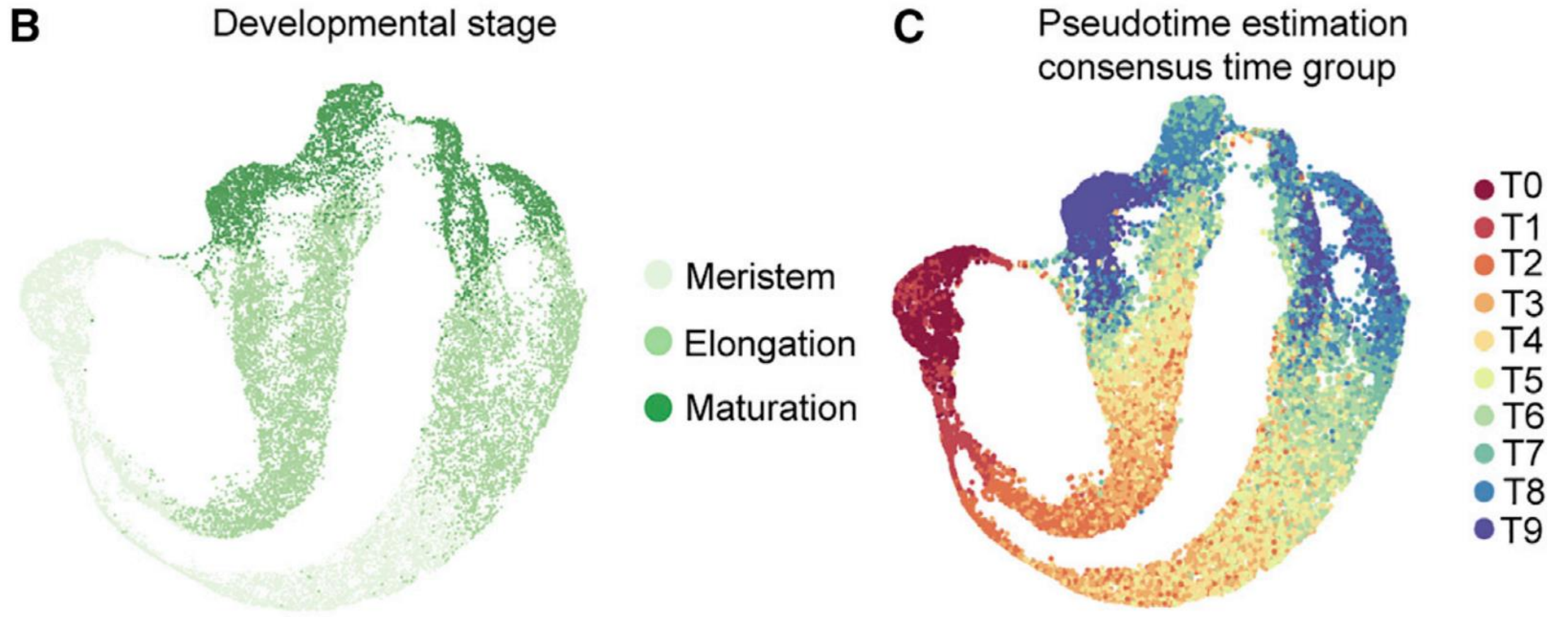
- Developing tissues show more **continuous** variation



WAGENINGEN
UNIVERSITY & RESEARCH

Shahan et al. 2022 *Dev Cell*

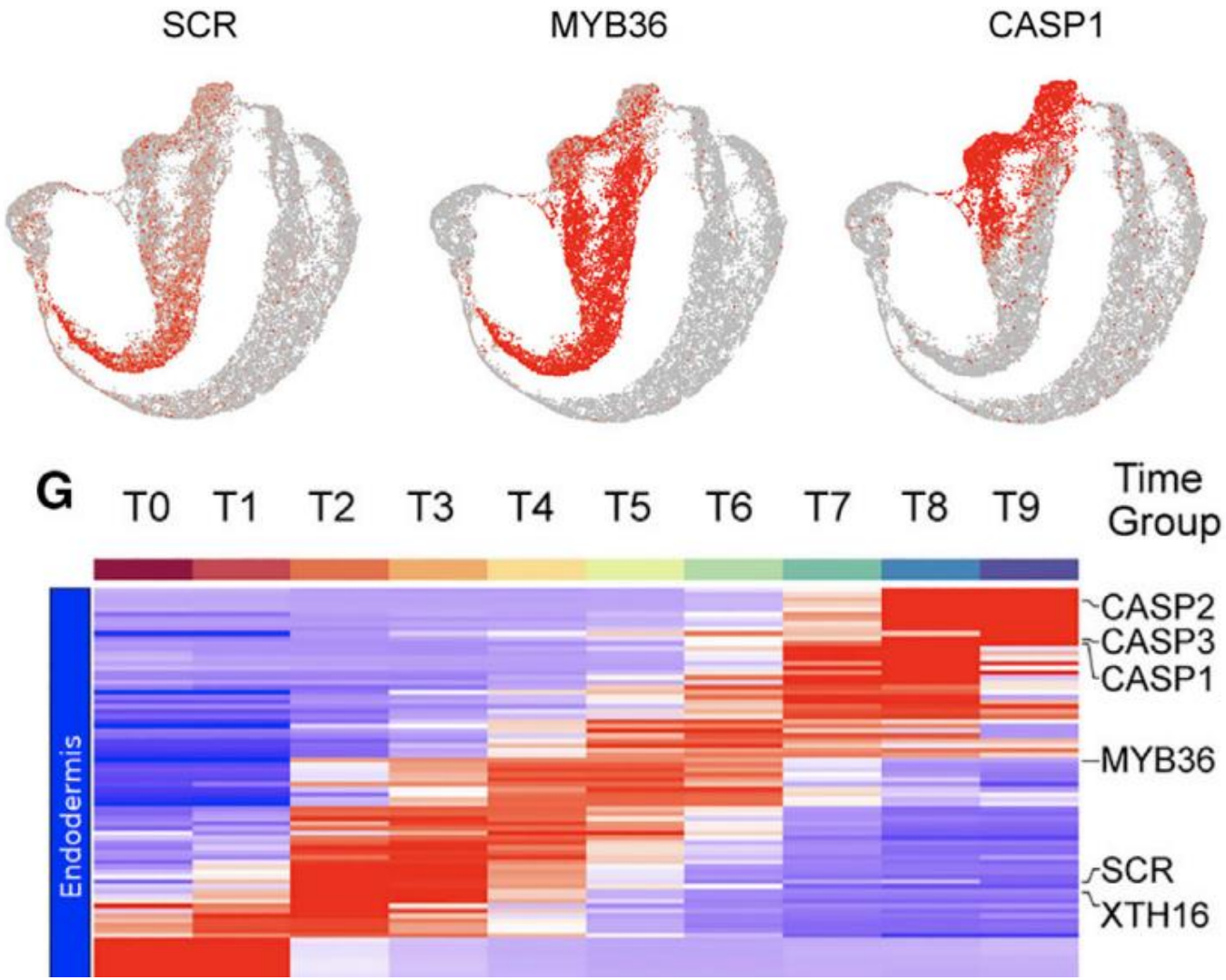
Pseudotime trajectories



WAGENINGEN
UNIVERSITY & RESEARCH

Shahan et al. 2022 *Dev Cell*

Pseudotime trajectories

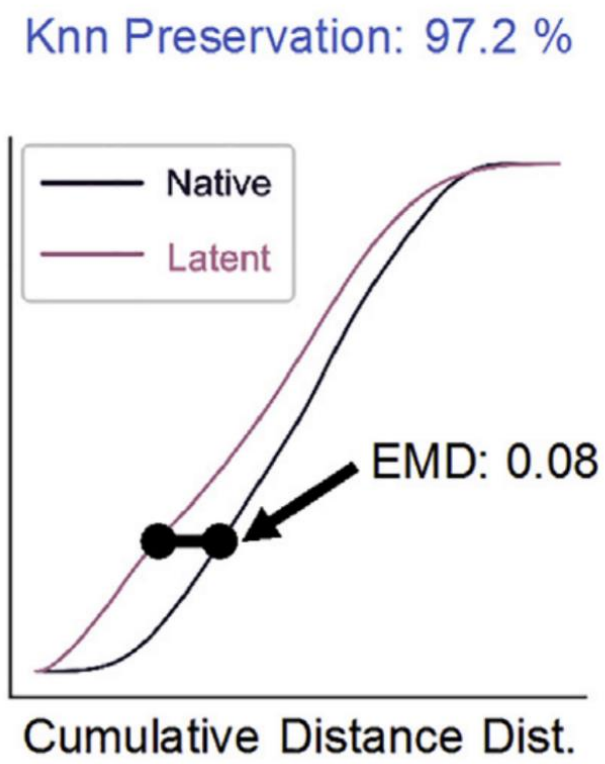


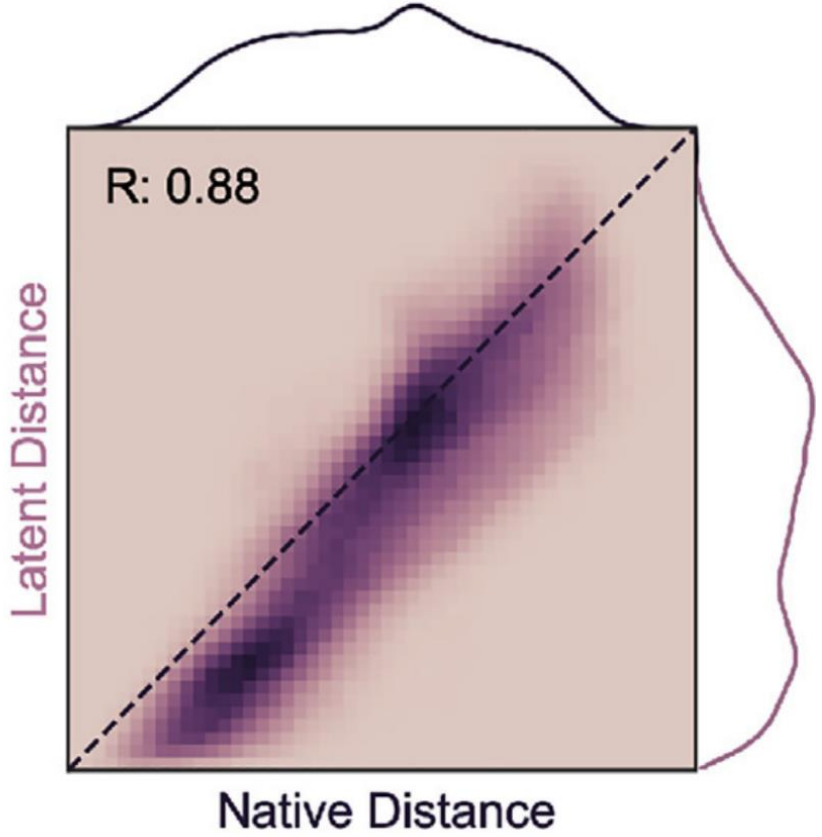
WAGENINGEN
UNIVERSITY & RESEARCH

Shahan et al. 2022 *Dev Cell*

Pseudotime trajectories

- But how can we know if a trajectory is accurate?
  - Calibrate against known marker genes
  - Measure distortions of "native" distances in the reduced space



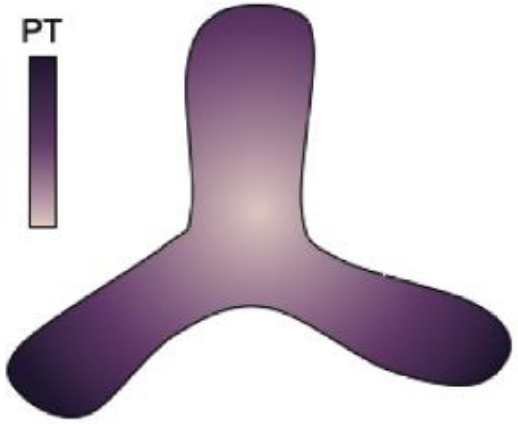


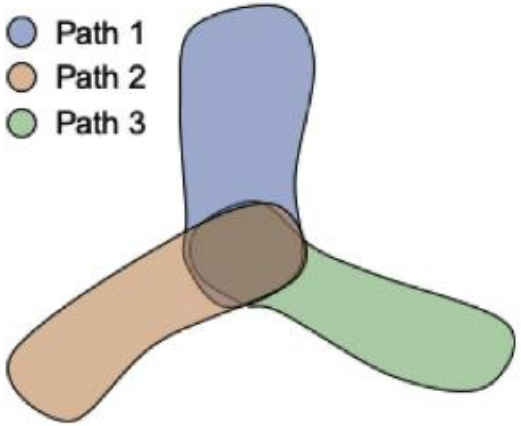
WAGENINGEN
UNIVERSITY & RESEARCH

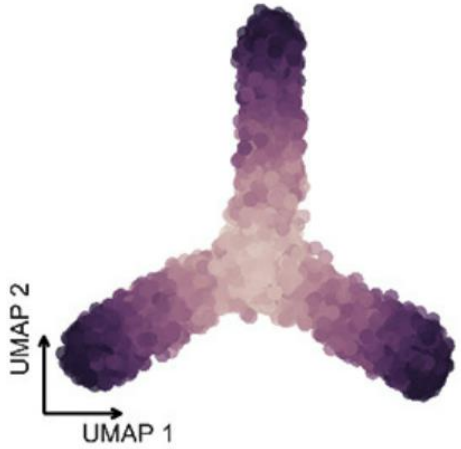
Heiser & Lau 2020 *Cell Reports*

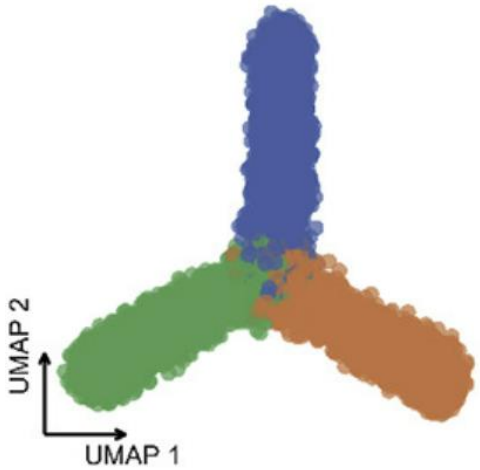
Pseudotime trajectories

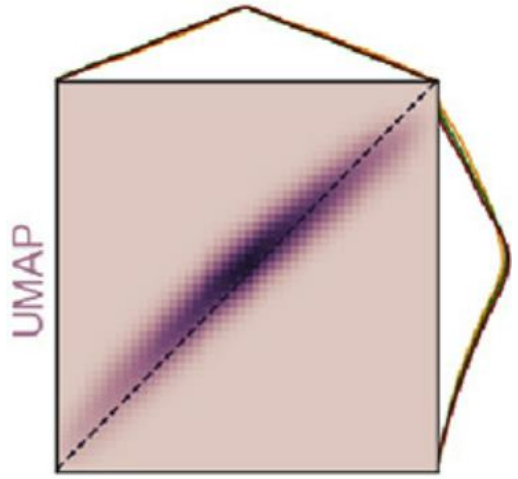
- UMAP excels when data are continuous and uniform









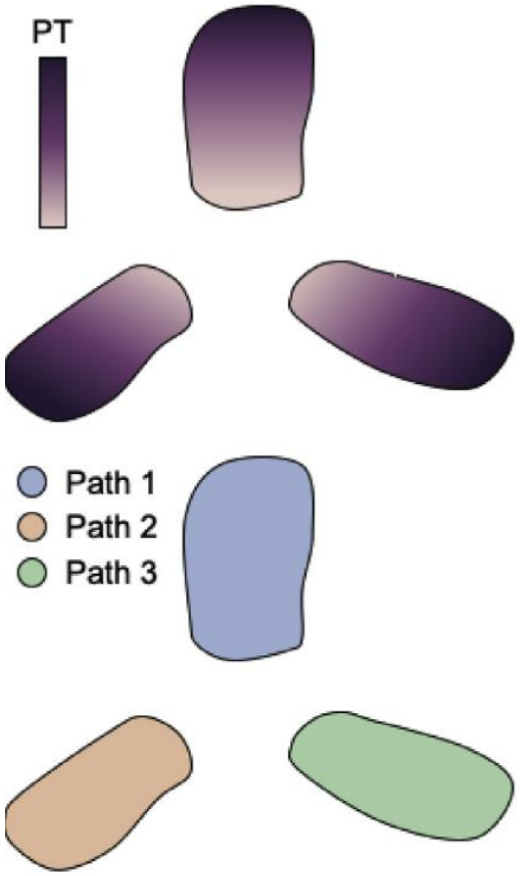


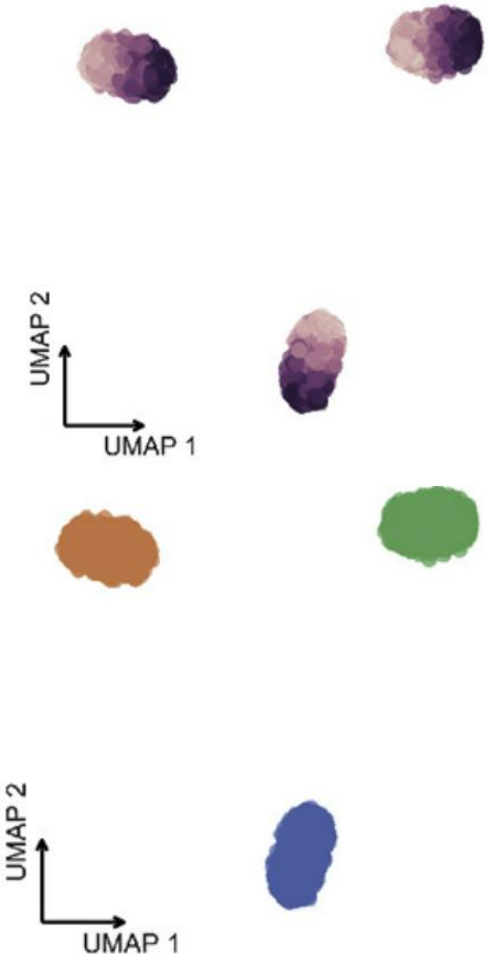
WAGENINGEN
UNIVERSITY & RESEARCH

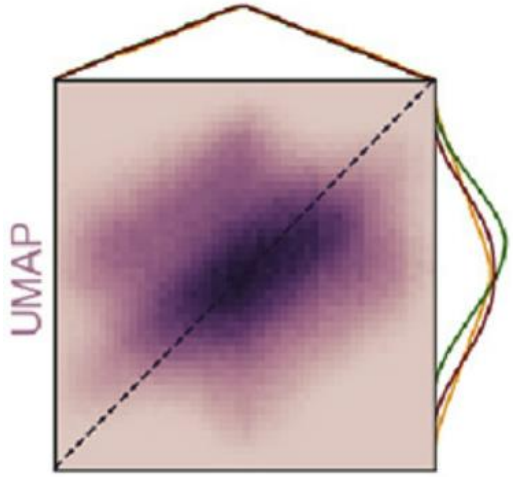
Heiser & Lau 2020 *Cell Reports*

Pseudotime trajectories

⚠ UMAP fails to globally orient discontinuous data







Heiser & Lau 2020 *Cell Reports*

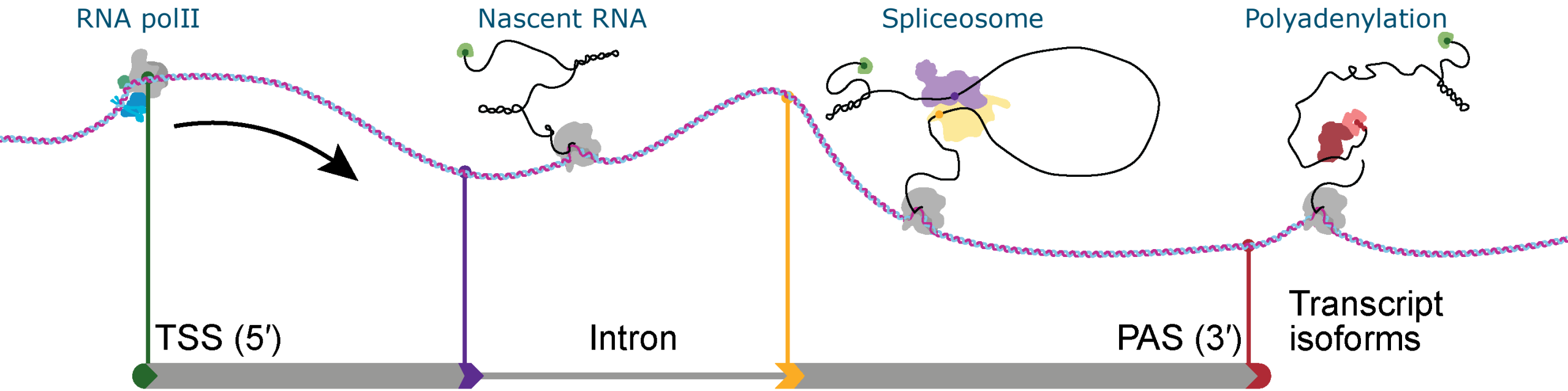
Pseudotime trajectories

⚠ Assumes a continuous path between cell types
- Incomplete data creates bad embeddings
- Not all heterogeneity is developmental!

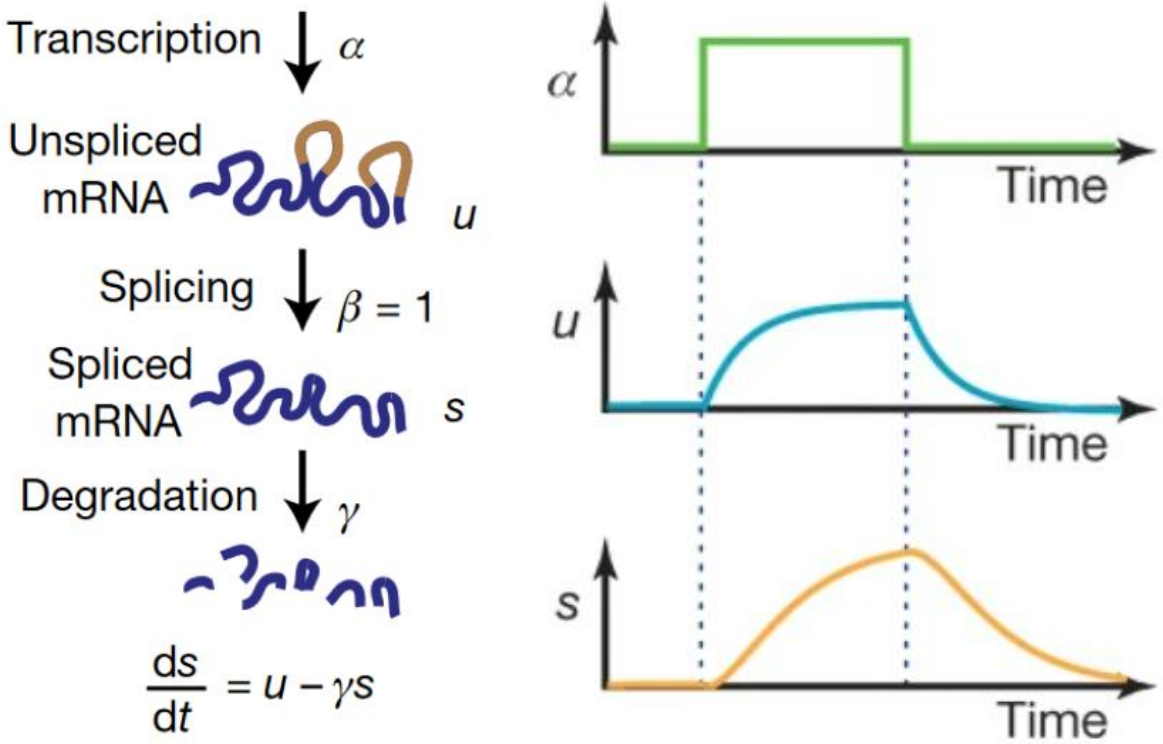
⚠ Operates in reduced dimensions
- 2D projections can have massive distortion!
- Need to be “anchored” in reality
- DR methods (UMAP, t-SNE) are not globally coherent

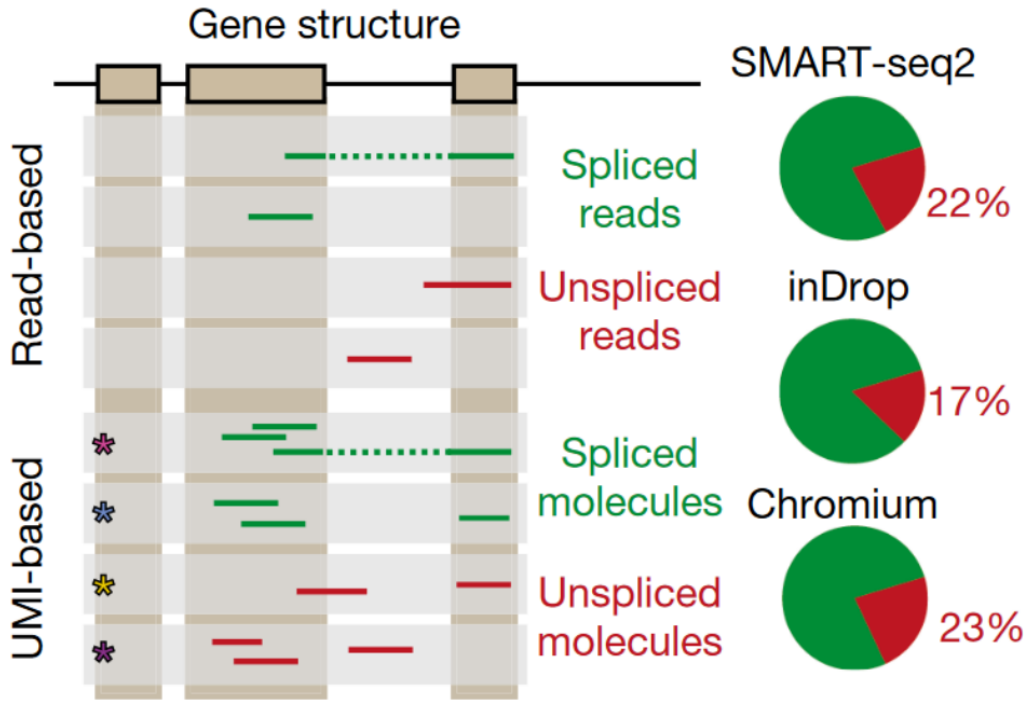
What else can we do with single-cell RNA?

Measuring RNA birth and death



RNA “velocity”: predicting the future?

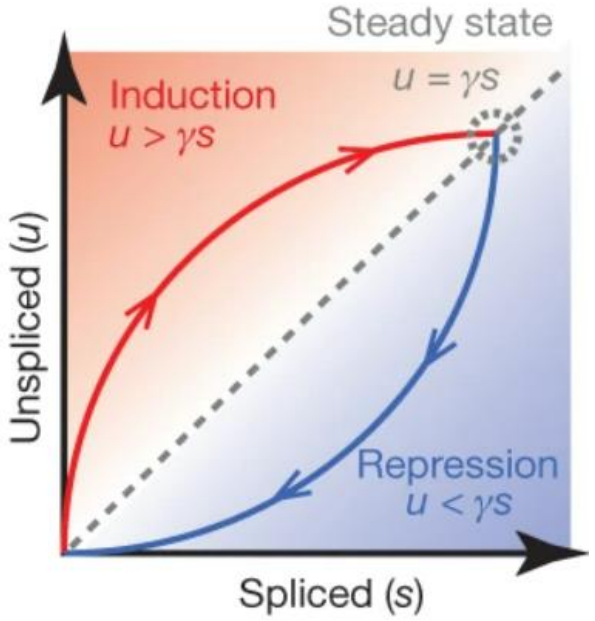




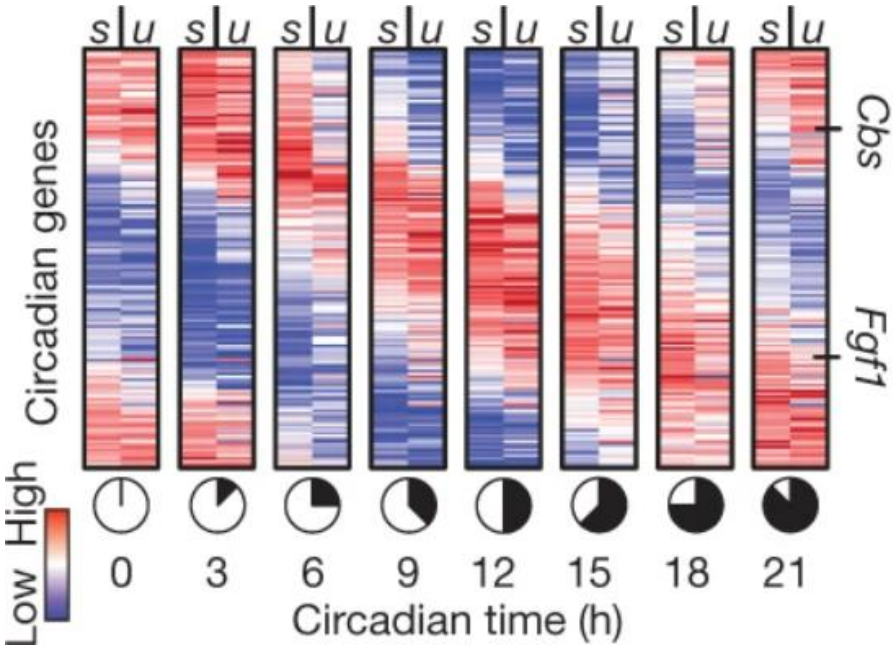
La Manno et al. 2018 *Nature*

RNA “velocity”: predicting the future?

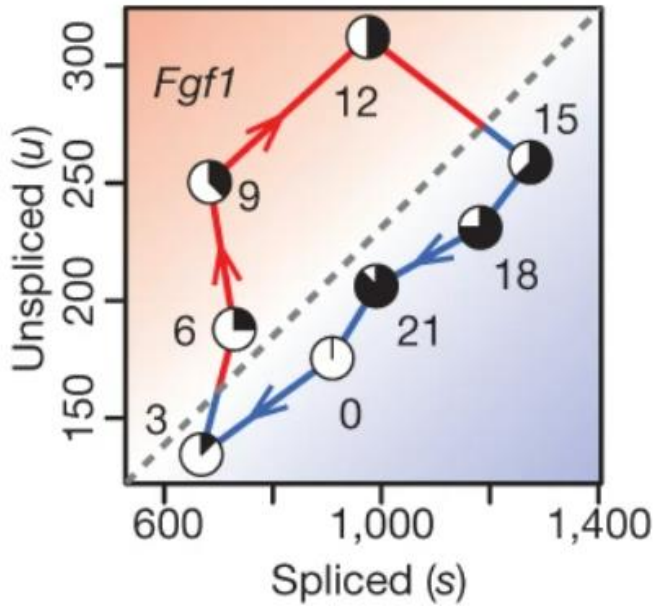
Model RNA kinetics



Unspliced RNA precedes spliced RNA



Kinetics of a single circadian gene

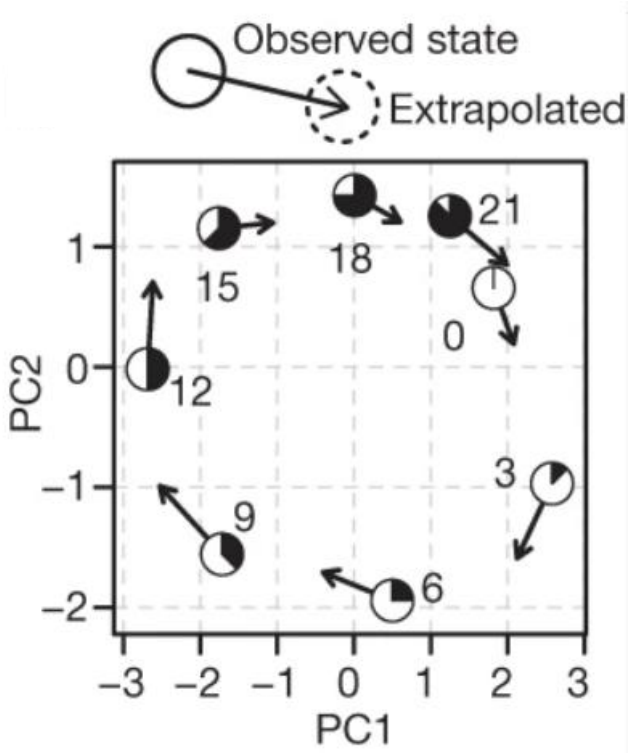


WAGENINGEN
UNIVERSITY & RESEARCH

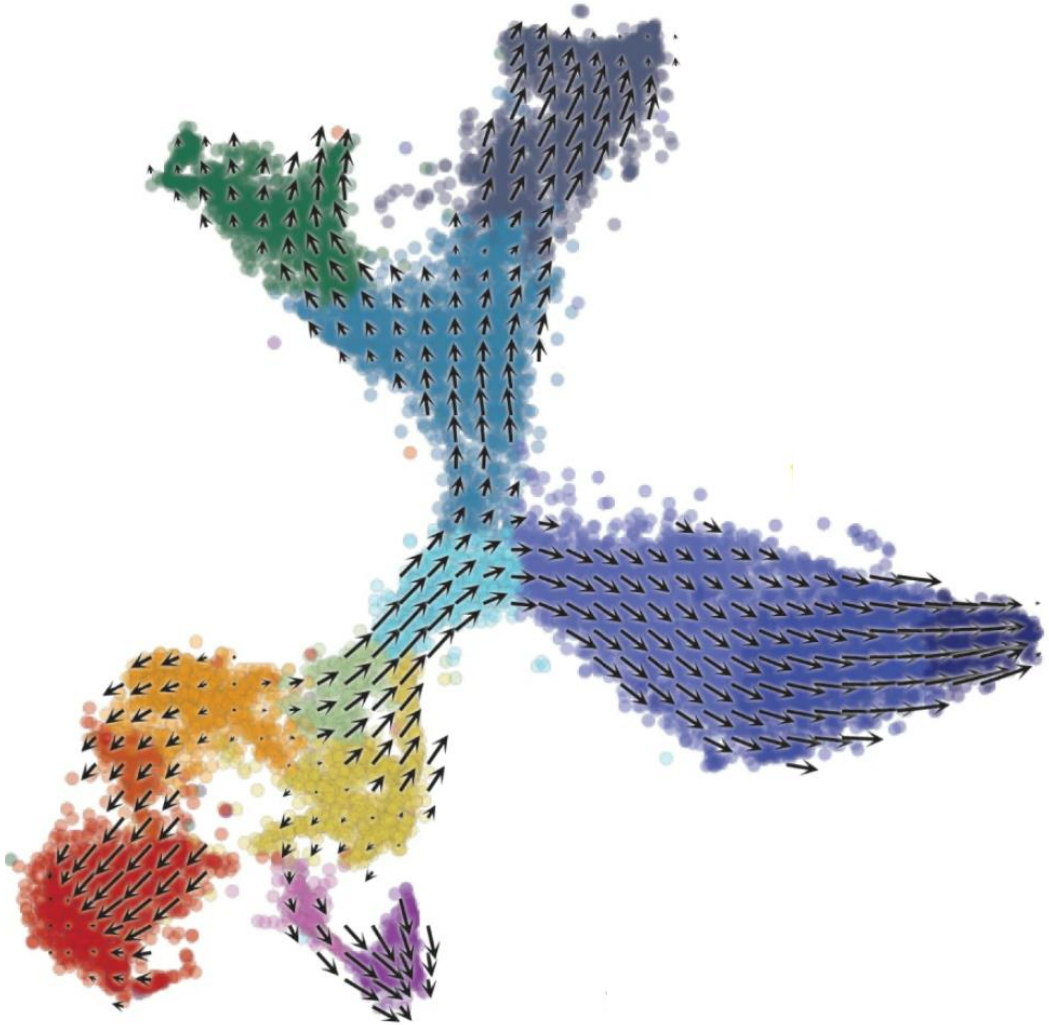
La Manno et al. 2018 *Nature*

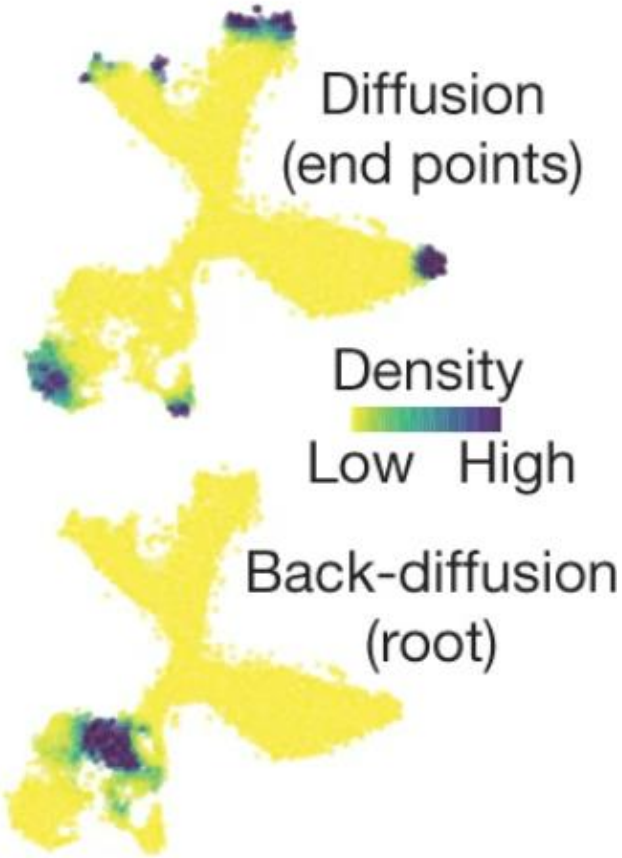
RNA “velocity”: predicting the future?

Circadian cycle



Developing hippocampus cells



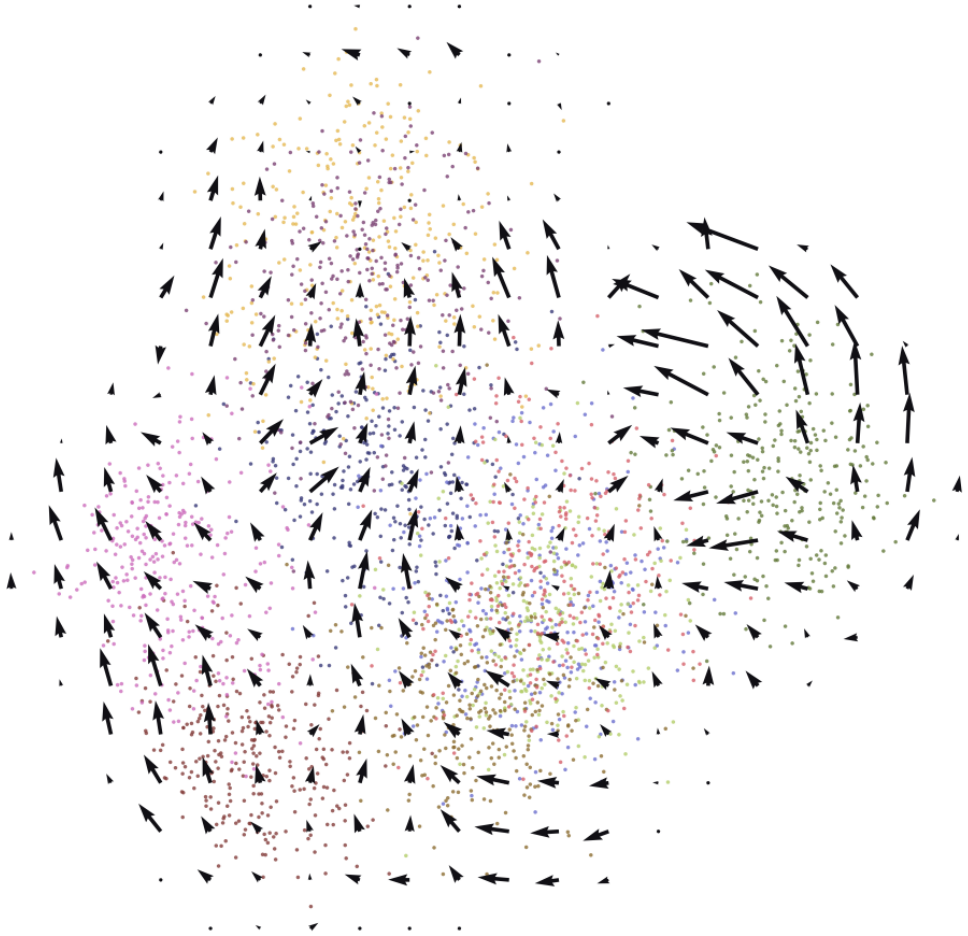


La Manno et al. 2018 *Nature*

RNA “velocity”: predicting the future?

⚠ Vectors will always be found, even when none exist

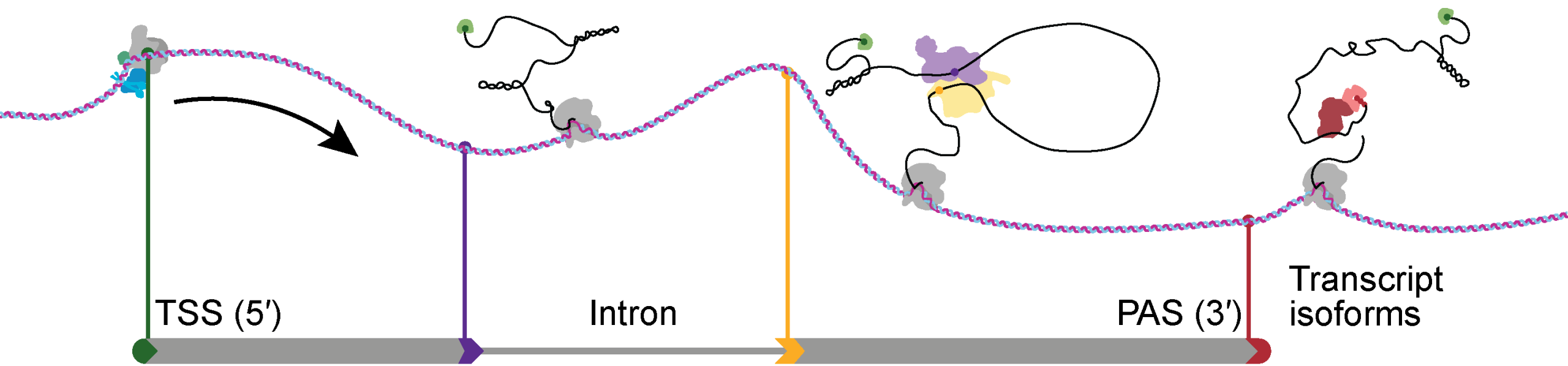
Noisy simulation of 10 discrete cell types



WAGENINGEN
UNIVERSITY & RESEARCH

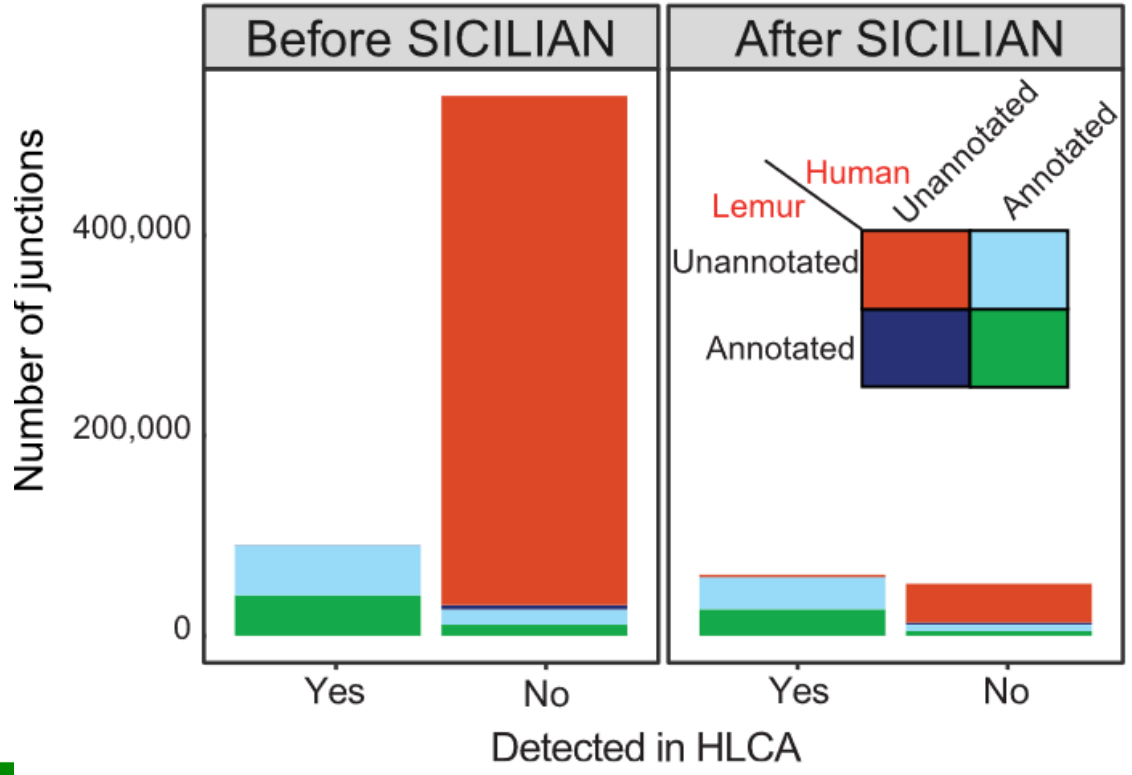
Gorin et al. 2022 *PLoS Comp Biol*

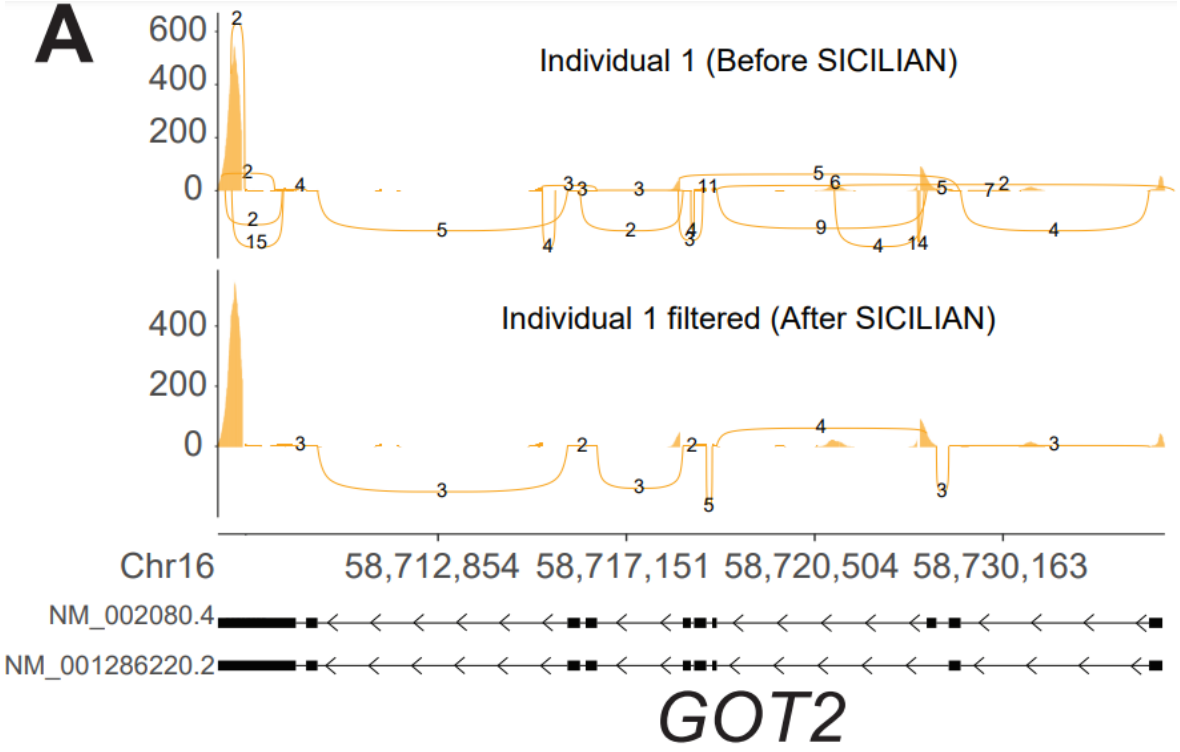
Genes → isoforms?



Limits to isoform identification

- ⚠ Single-cell sequencing is noisy
  - Gapped alignments yield more false splice junctions than true ones
  - SICILIAN filters splice junctions by sequence entropy
- ⚠ Most scRNA-seq data is heavily biased to 3’ ends (10X Genomics)



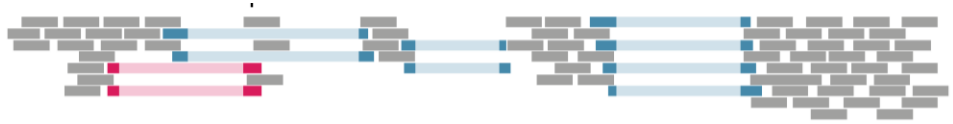


WAGENINGEN
UNIVERSITY & RESEARCH

Dehghannasiri et al. 2021 *Genome Biology*

Single-cell isoform assembly?

- Full-length single cell sequencing? (Smart-seq)
  - Transcript assemblers have abysmal precision (4 false transcripts per true transcript)



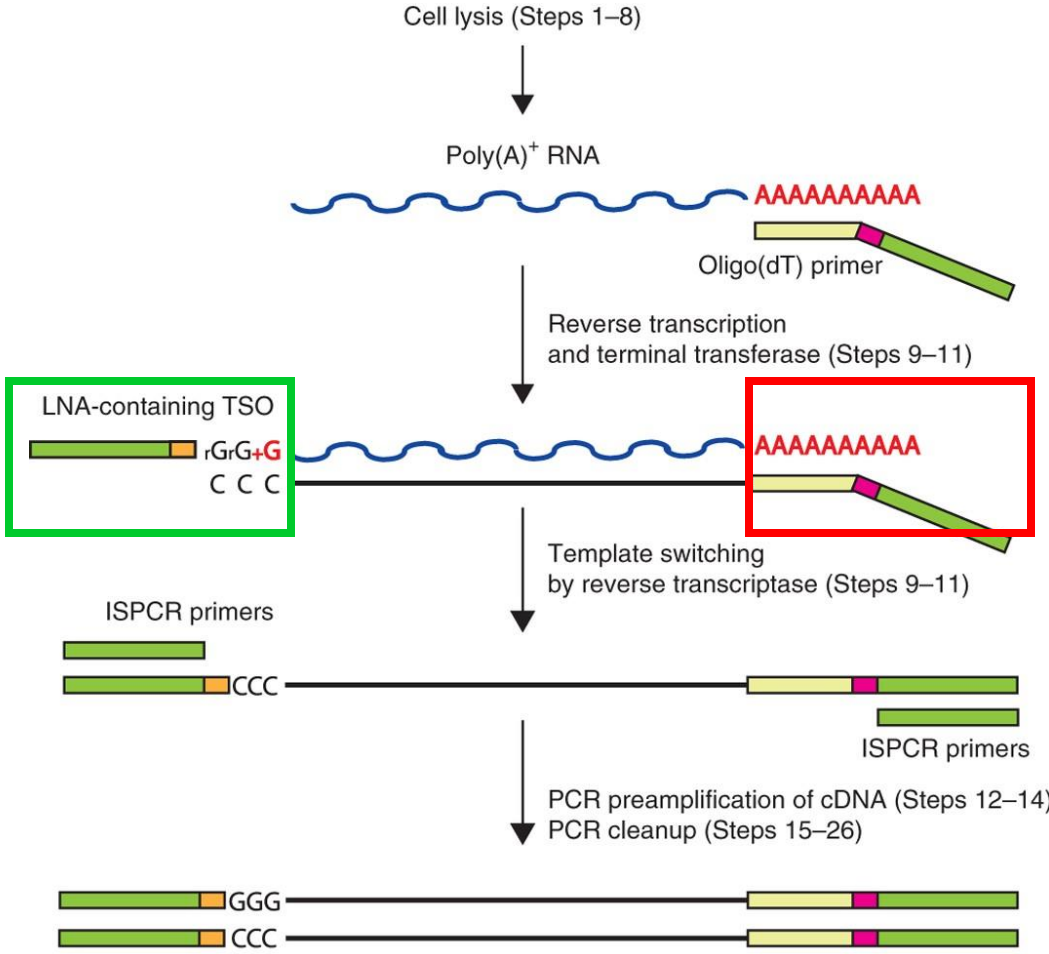
WAGENINGEN UNIVERSITY & RESEARCH

Schon et al. 2022 *Genome Biology*

Something is missing...

Smart-seq2 protocol

Template-switching oligo (TSO)



Poly(A) tail

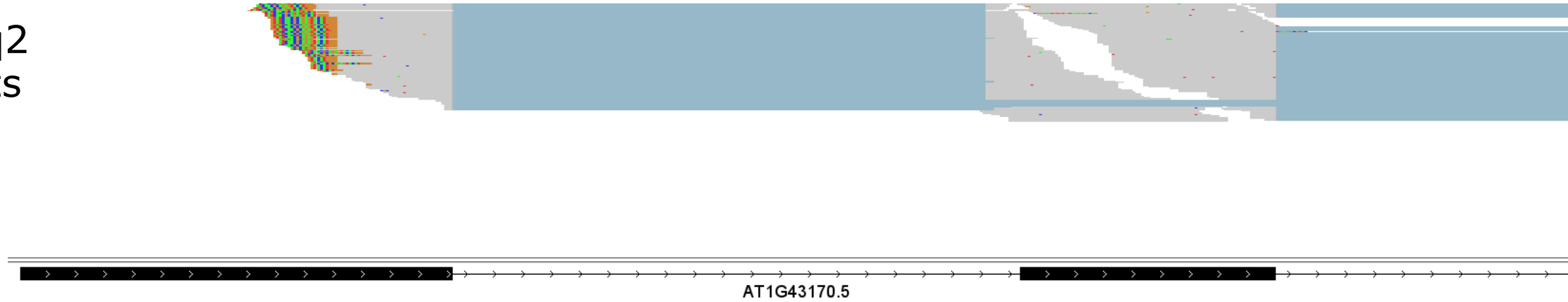
WAGENINGEN
UNIVERSITY & RESEARCH

Picelli et al. 2014 *Nature Protocols*

Something is missing...

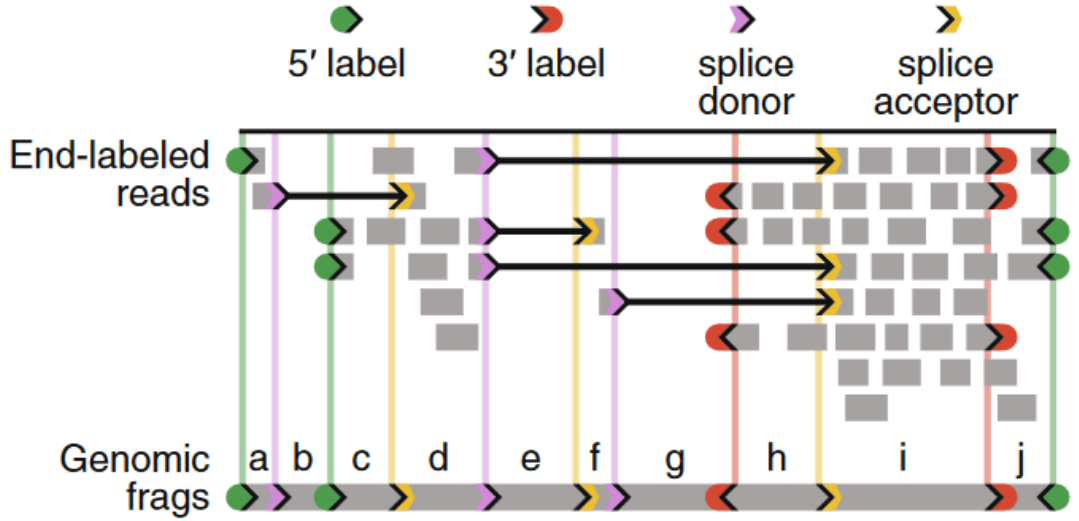
TSO positions

Smart-seq2 alignments



- Could RNA end labels improve transcript assembly?

End-guided assembly

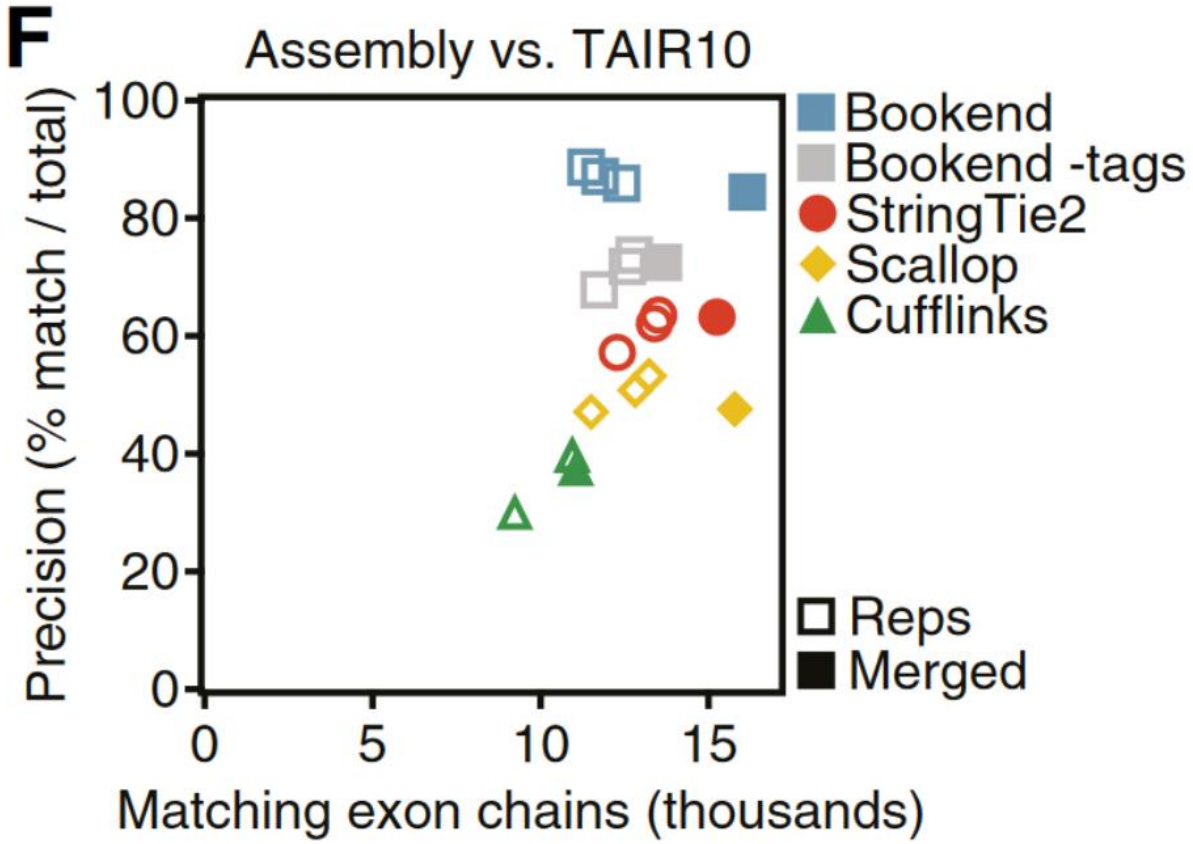


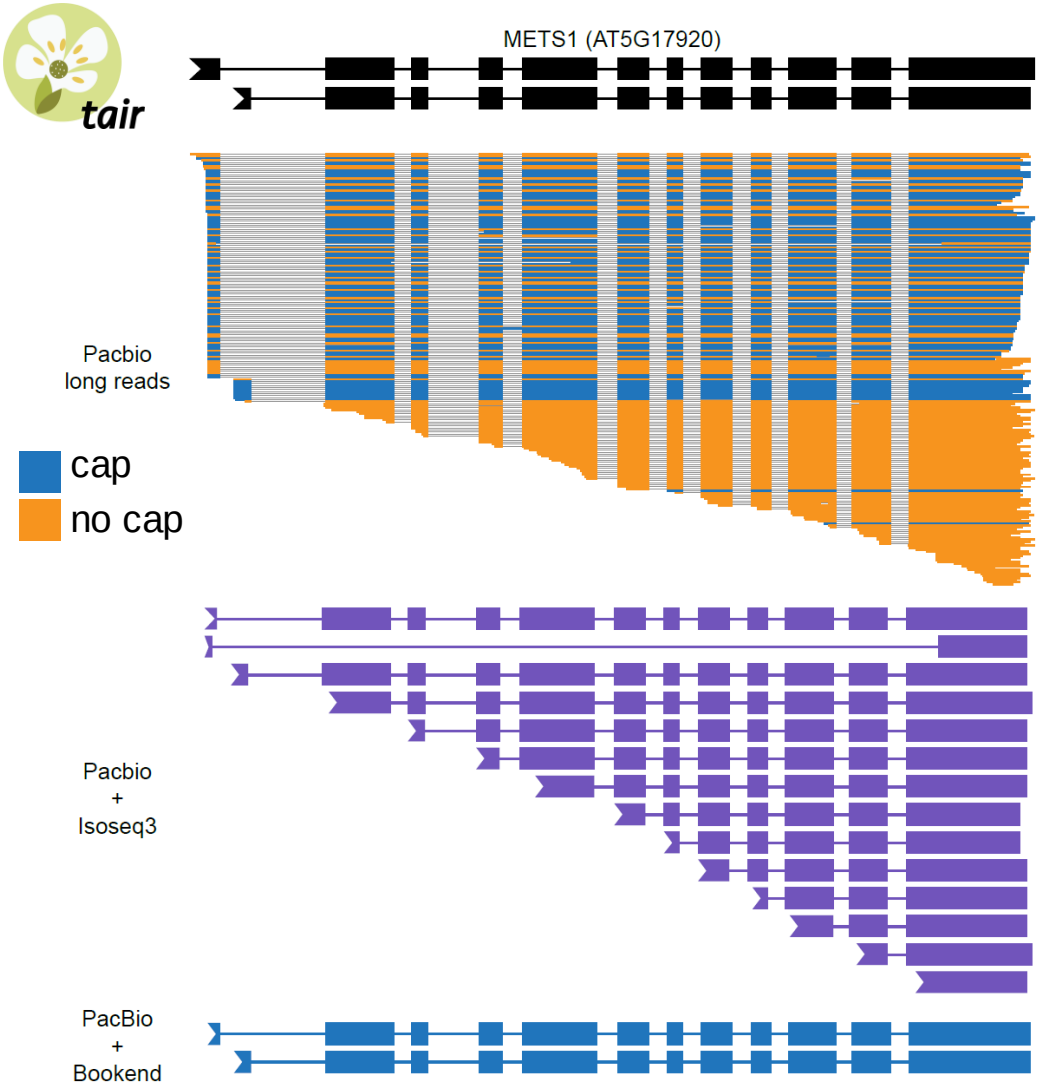
bookend

WAGENINGEN
UNIVERSITY & RESEARCH

Schon et al. 2022 *Genome Biology*

End-guided assembly



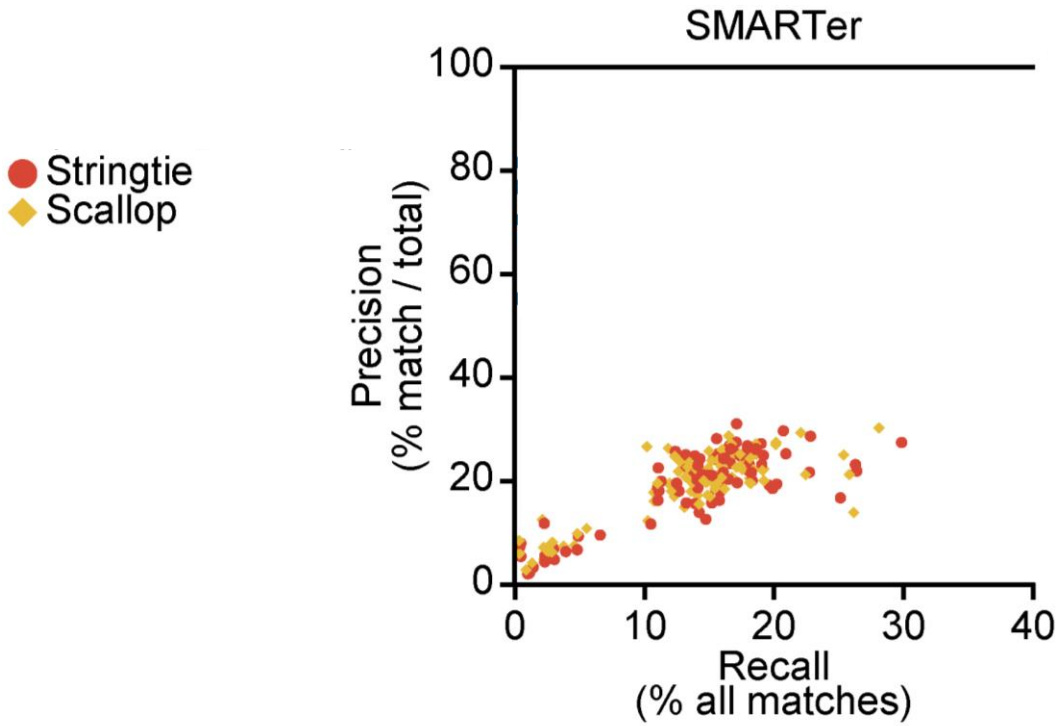


Schon et al. 2022 *Genome Biology*

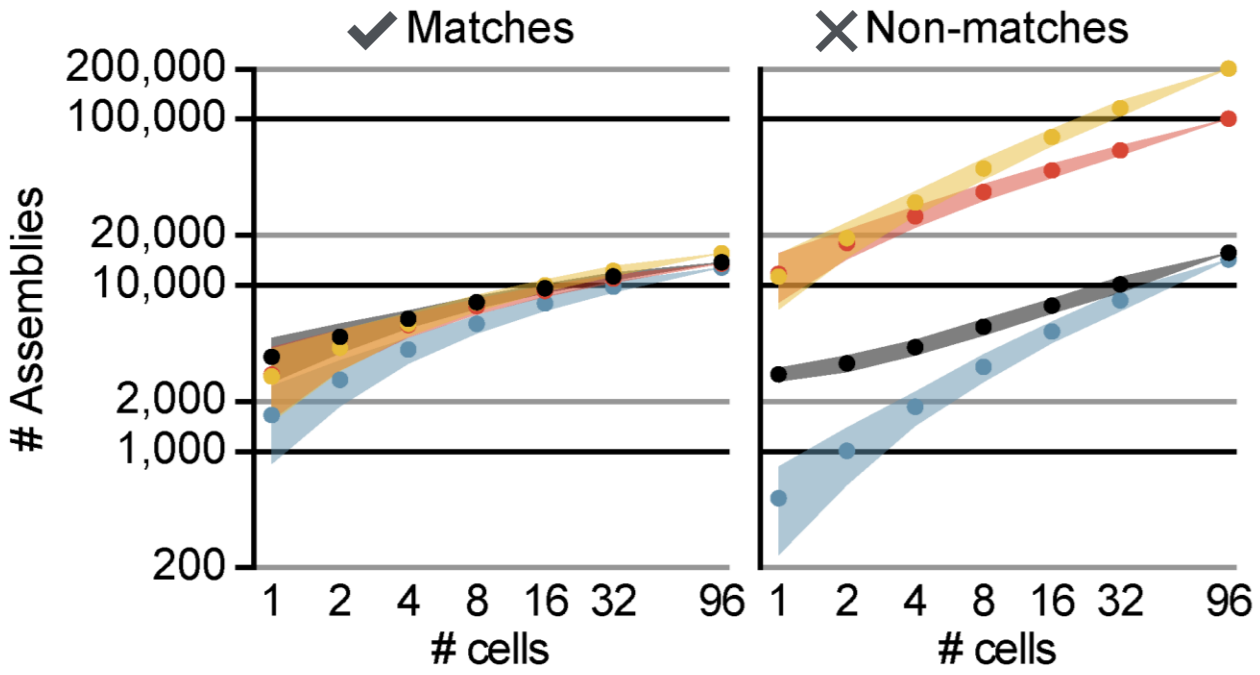
WAGENINGEN
UNIVERSITY & RESEARCH

End-guided assembly

Mouse embryonic stem cell assemblies
(96 single cells)



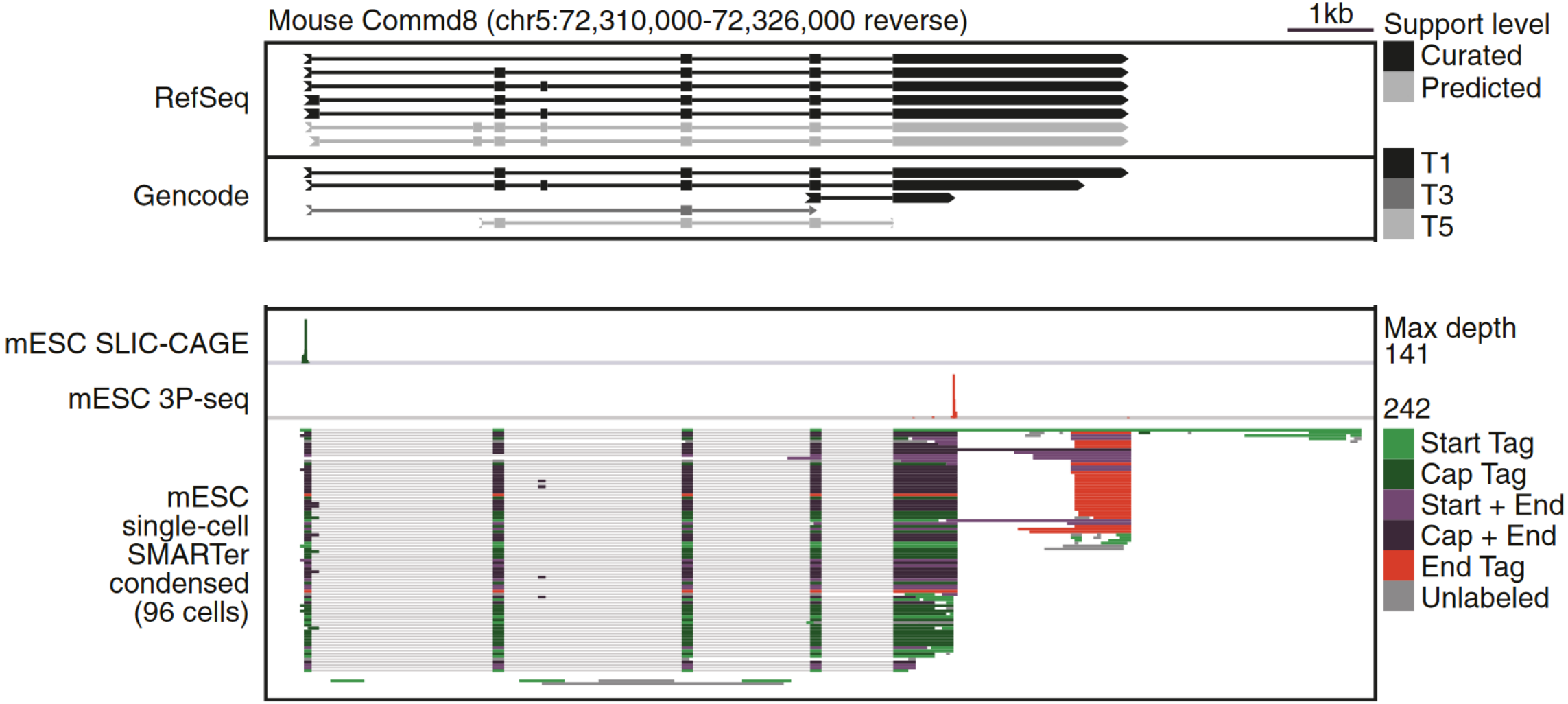
Multi-cell assemblies
vs. reference (RefSeq)



WAGENINGEN
UNIVERSITY & RESEARCH

Schon et al. 2022 *Genome Biology*

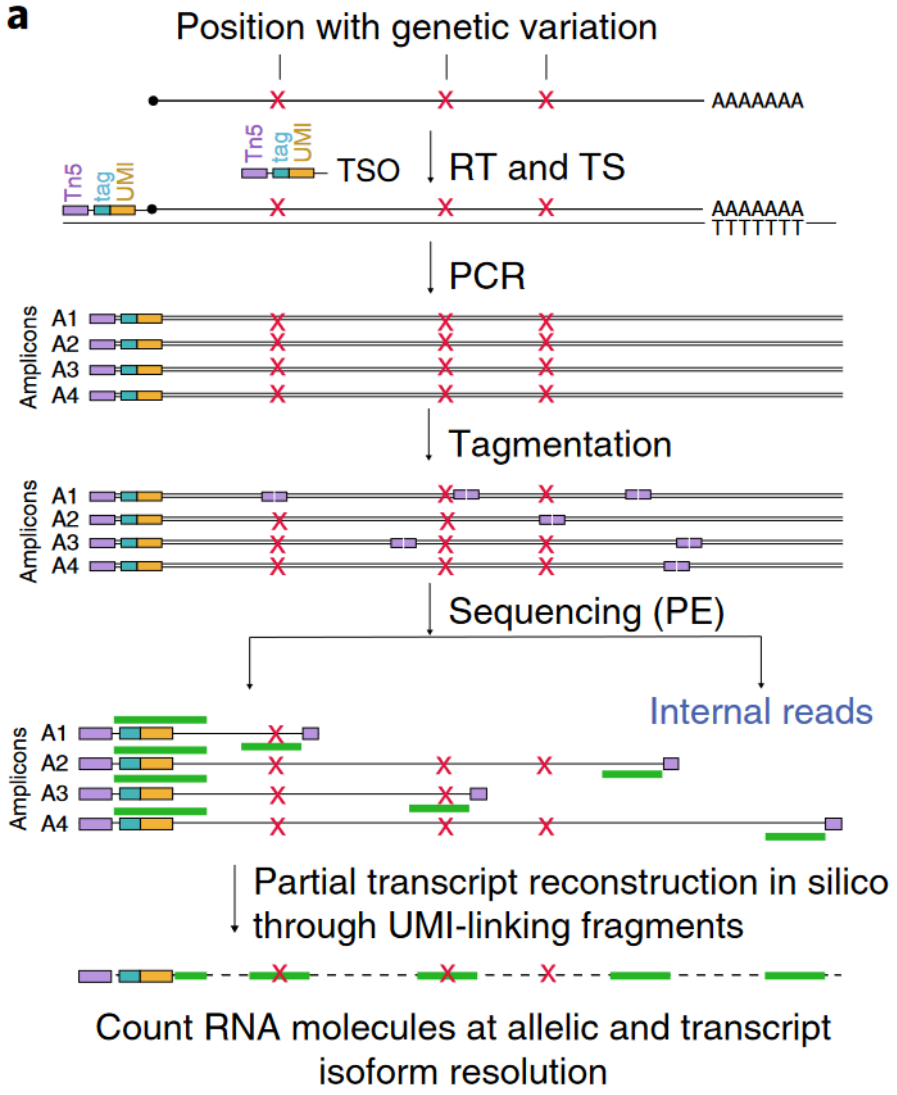
Meta-assembly: each cell is unique



WAGENINGEN
UNIVERSITY & RESEARCH

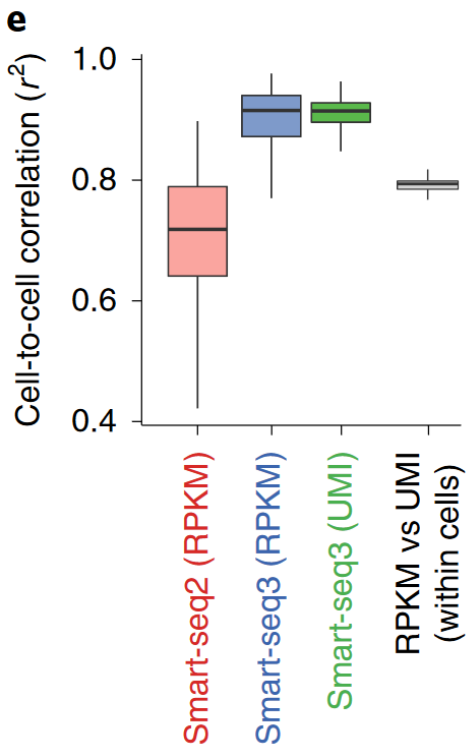
Schon et al. 2022 *Genome Biology*

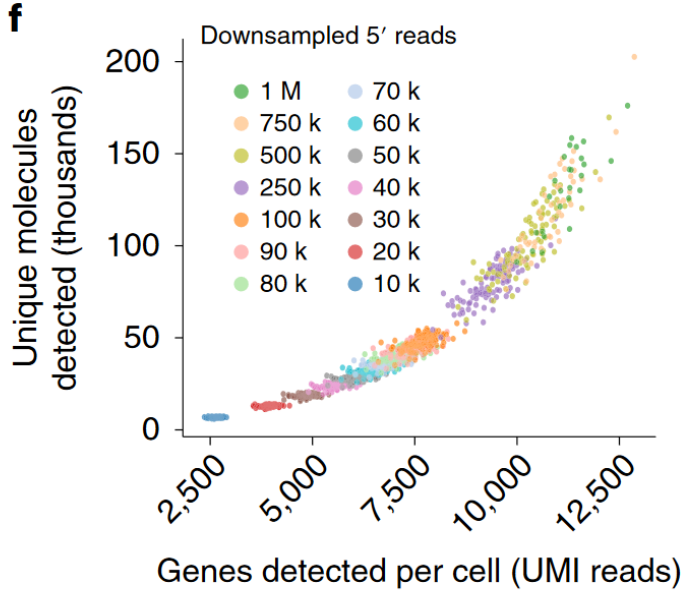
Partial molecule reconstruction



Smart-seq3

- Tn5-UMI-TSO → 20-30% 5’ PE reads
- Can capture >100,000 UMIs/cell
  ~25% of smFISH!

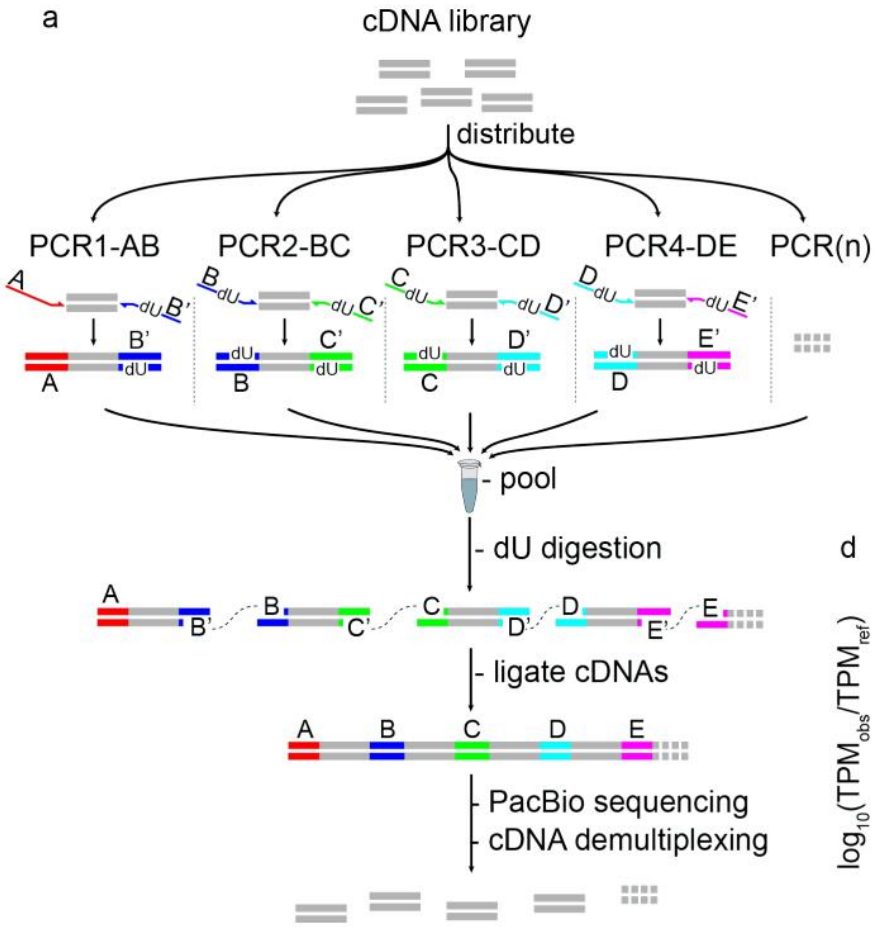


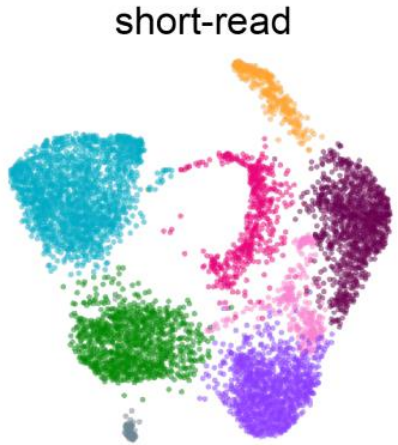


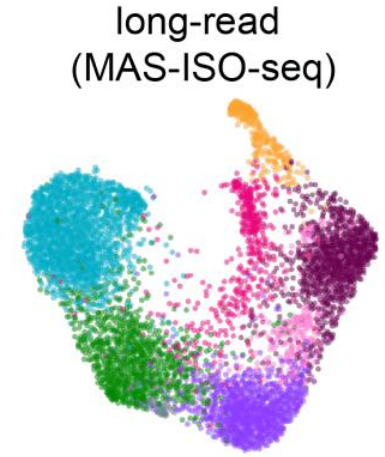
WAGENINGEN
UNIVERSITY & RESEARCH

Hagemann-Jensen et al. 2020 *Nature Biotechnology*

Long reads in single cells







d

$\log_{10}(TPM_{obs}/TPM_{ref})$

WAGENINGEN
UNIVERSITY & RESEARCH

Al'Khafaji et al. 2021 *bioRxiv*

Conclusions

- Development is a cellular process
- We now have many methods for sequencing many RNAs x many cells
- Downstream analysis is both art and science
  - Well-grounded in reality:
    marker genes
    timepoints
    validation
  - "Number of clusters" is (partly) subjective
  - Continuous methods are possible, but prone to distortions
- Not all data is created equal!
  - Make sure the protocol is sufficient to answer your question
  - Isoform analysis requires full-length sequencing:
    Smart-seq (2,3,xpress)
    MAS-ISO-SEQ

Acknowledgements

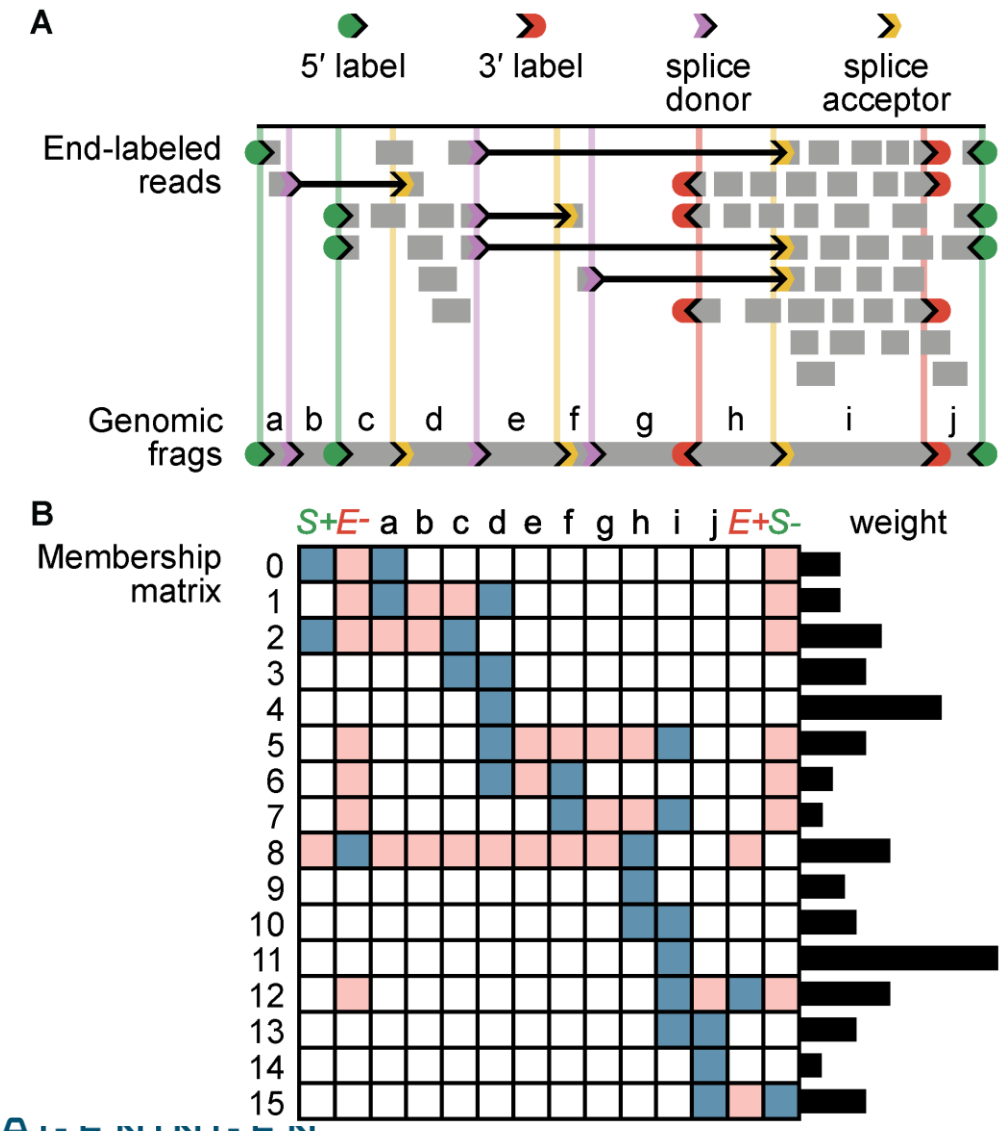
(former) Nodine Lab
Max Kellner
Alexandra Plotnikova
Falko Hofmann
Stefan Lutzmayer
Ping Kao
Ranjith Papareddy
Katalin Paldi
Magda Mosiolek
Balaji Enugutti

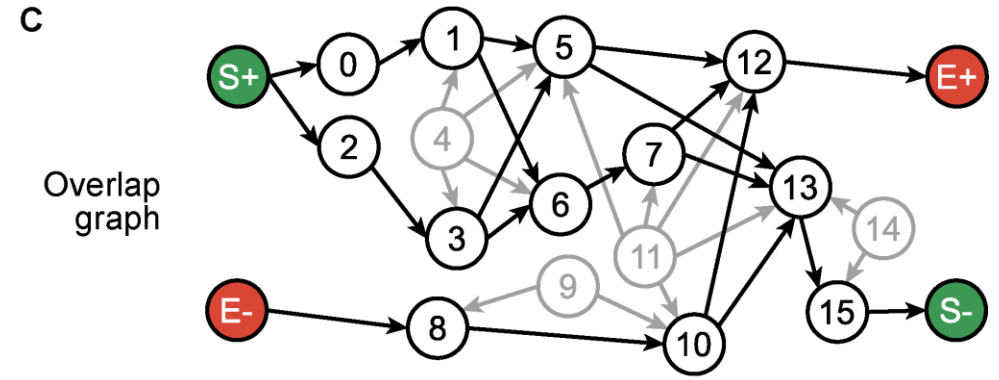
Thesis Advisory Committee
Stefan Ameres
Mike Axtell
Andrea Pauli

Thesis Reviewers
Kelly Swarts
Dolf Weijers

VBCF
Next Generation Sequencing
Andreas Sommer
Ido Tamir

The Bookend framework

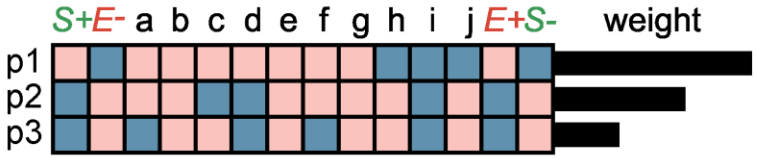




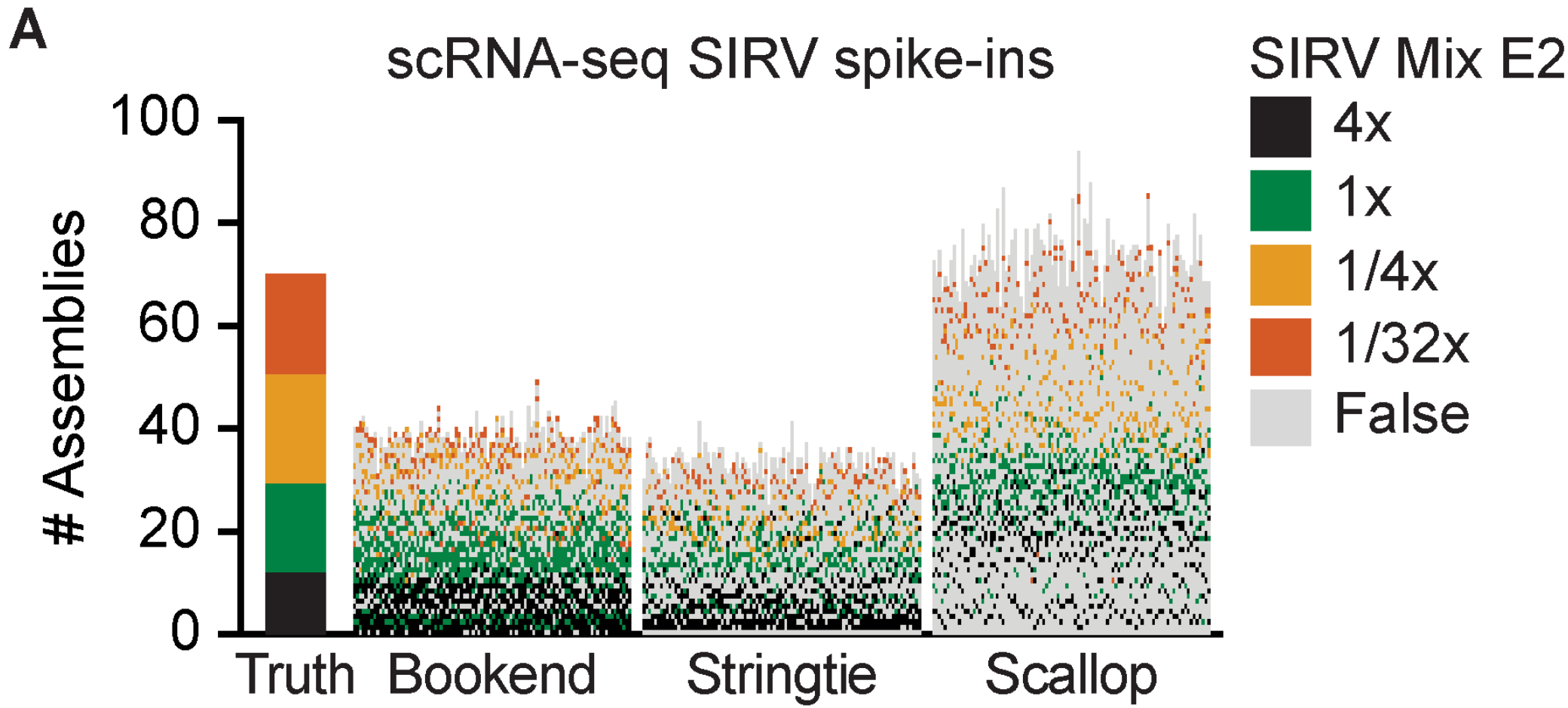
D

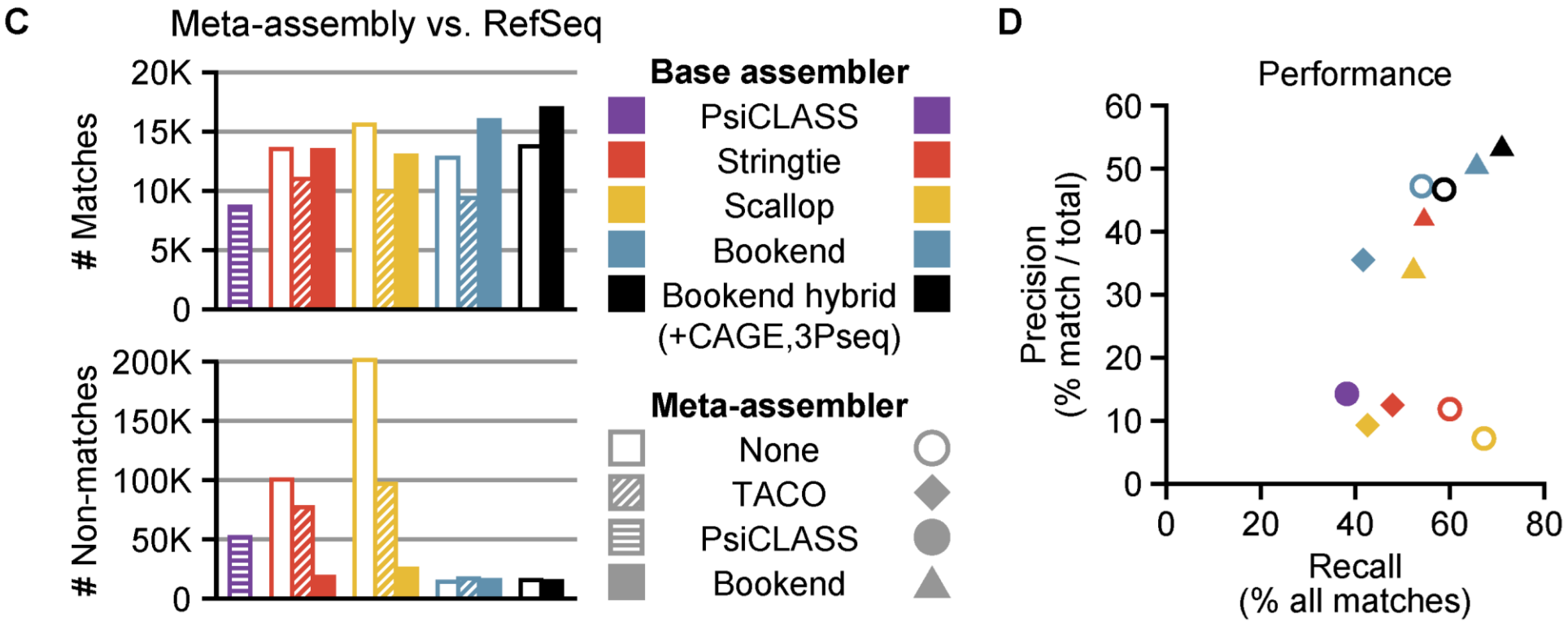
Greedy paths

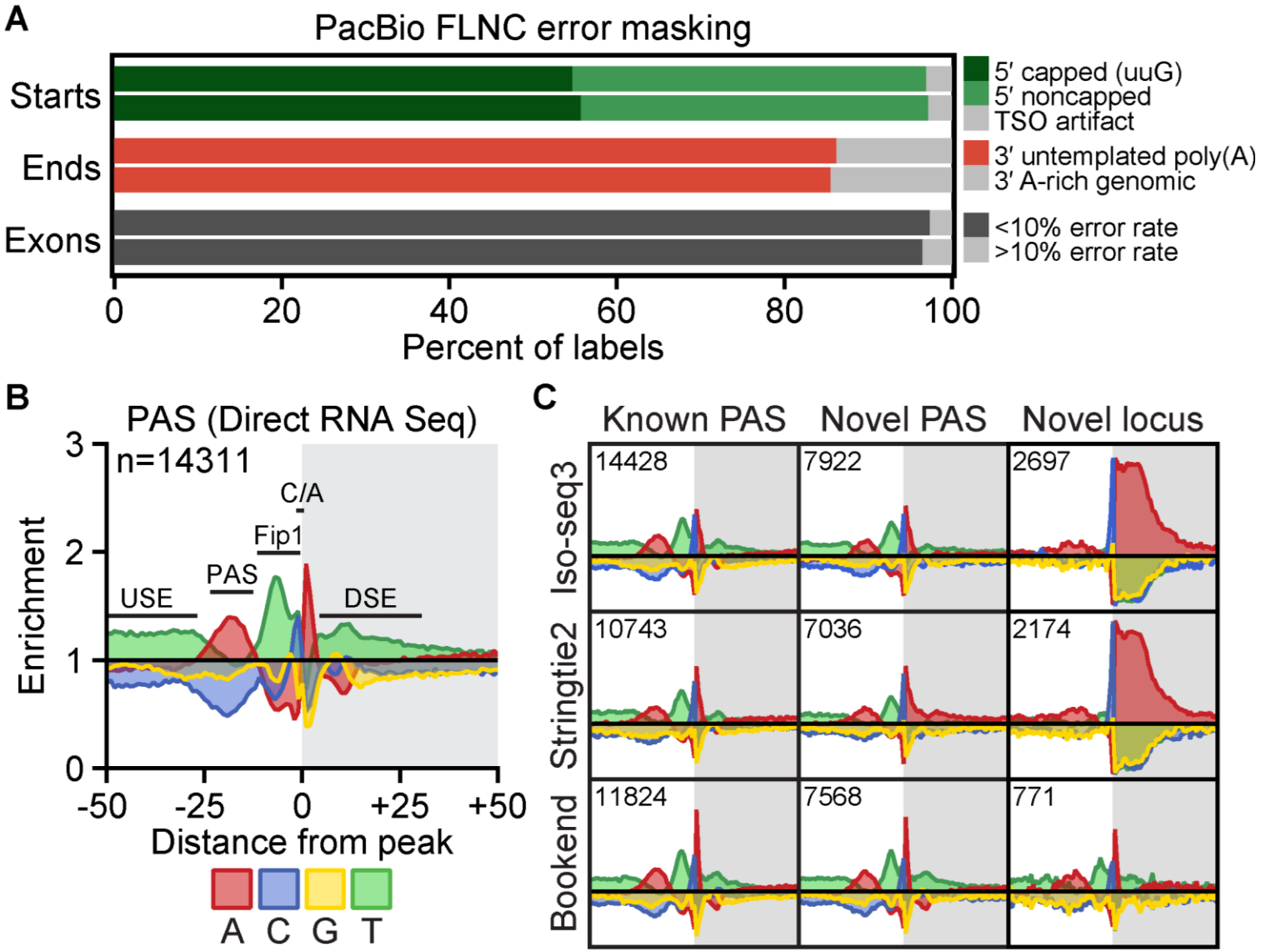
p1: *E*-→8/9→10/11→13/14→15→*S*-
p2: *S*+→2→3/4→5/11→12→*E*+
p3: *S*+→0→1/4→6→7/11→12→*E*+

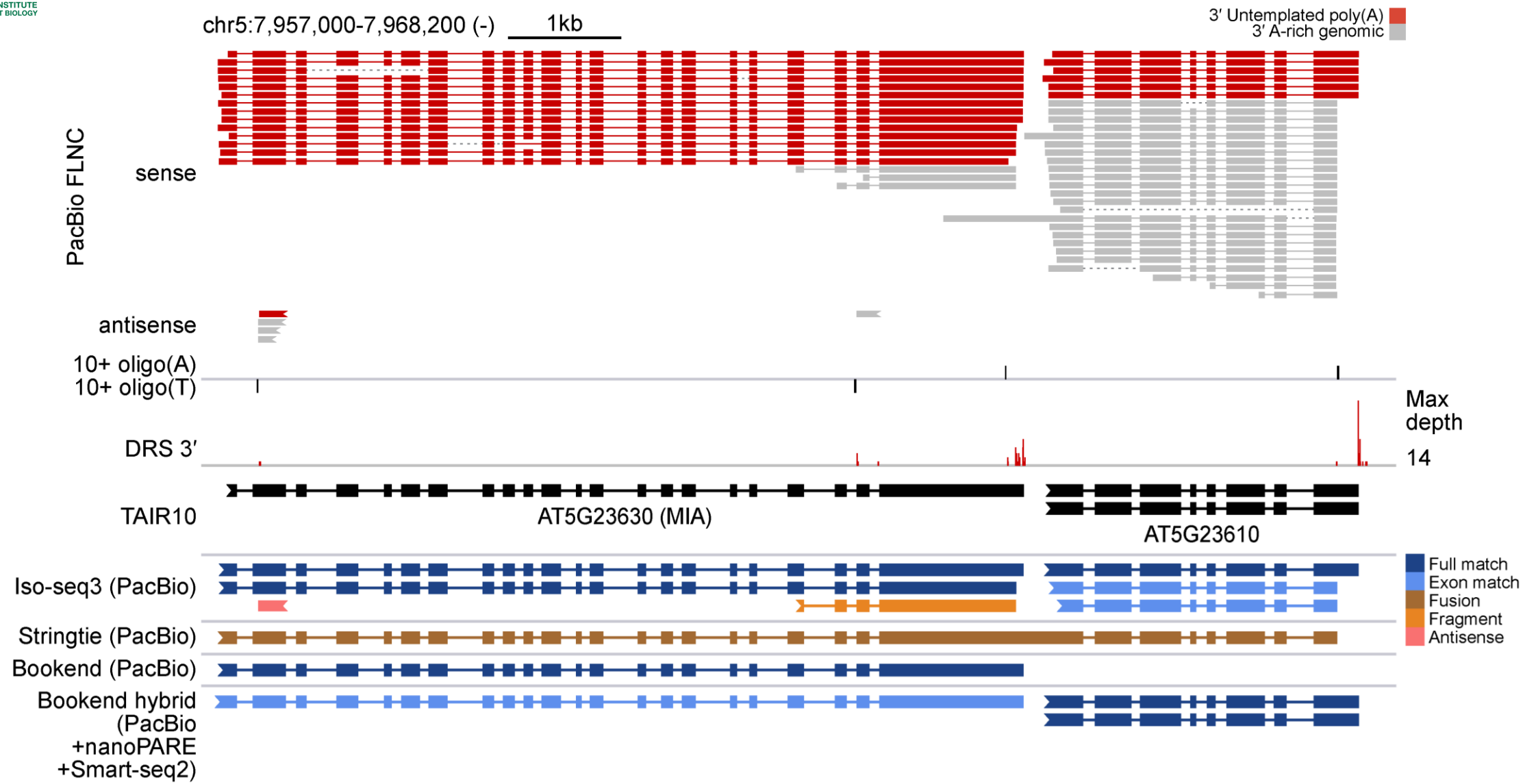


bookend


A
scRNA-seq SIRV spike-ins
Assemblies
100
80
60
40
20
0
Truth
Bookend
Stringtie
Scallop
SIRV Mix E2
4x
1x
1/4x
1/32x
False


C
Meta-assembly vs. RefSeq
Matches
20K
15K
10K
5K
0
Non-matches
200K
150K
100K
50K
0
Base assembler
PsiCLASS
Stringtie
Scallop
Bookend
Bookend hybrid
(+CAGE,3Pseq)
Meta-assembler
None
TACO
PsiCLASS
Bookend
D
Performance
Precision
(% match / total)
60
50
40
30
20
10
0
0
20
40
60
80
Recall
(% all matches)


A
PacBio FLNC error masking
Starts
Ends
Exons
0
20
40
60
80
100
Percent of labels
5′ capped (uuG)
5′ noncapped
TSO artifact
3′ untemplated poly(A)
3′ A-rich genomic
<10% error rate
>10% error rate
B
PAS (Direct RNA Seq)
n=14311
Enrichment
C/A
Fip1
PAS
USE
DSE
-50
-25
0
+25
+50
Distance from peak
A C G T
C
Known PAS
Novel PAS
Novel locus
Iso-seq3
Stringtie2
Bookend
14428
7922
2697
10743
7036
2174
11824
7568
771


chr5:7,957,000-7,968,200 (-)
1kb
3′ Untemplated poly(A)
3′ A-rich genomic
PacBio FLNC
sense
antisense
10+ oligo(A)
10+ oligo(T)
Max depth
14
DRS 3′
TAIR10
AT5G23630 (MIA)
AT5G23610
Iso-seq3 (PacBio)
Stringtie (PacBio)
Bookend (PacBio)
Bookend hybrid (PacBio +nanoPARE +Smart-seq2)
Full match
Exon match
Fusion
Fragment
Antisense