

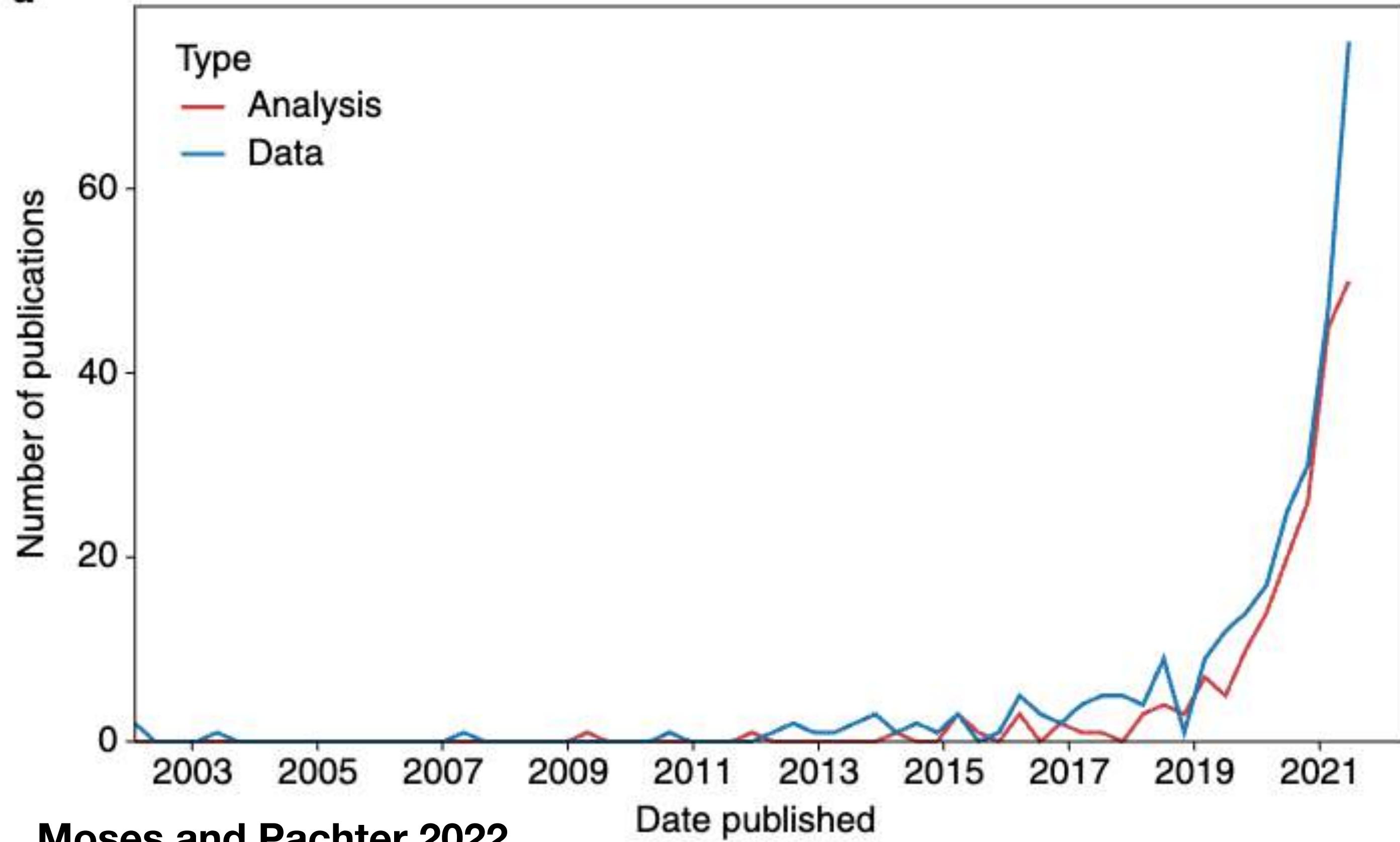
UCLA Collaboratory workshop (W31)

Spatial Transcriptomics

Oct 28, 2025 (Day 1)
Fangming Xie

An introduction *invitation* to spatial transcriptomics

a



- A rapidly growing field started from ~2015.
- Everything learned here will soon be obsolete?
- “*go for the messes – that's where the action is.*”
- “*what is known for certain is dull*”

[Published: 27 November 2003](#)

Scientist

Four golden lessons

[Steven Weinberg](#)

[Nature 426, 389 \(2003\)](#) | [Cite this article](#)

- After this workshop, you will NOT become an expert, but
- you will have jumped into the deep water, swimming (and/or sinking).

Day 1: Overview

- **Talk:** background and concepts
 - **Reading sessions:** reading and discussing research papers
 - **Mini-quiz:** for fun and discussion
-
- **Grading:** Attendance, Take-home quiz and Homework (due next Friday)
 - **Asking any questions at any time:** there is no silly question, no bad time asking it.

Day 2 and 3: Coding bootcamp

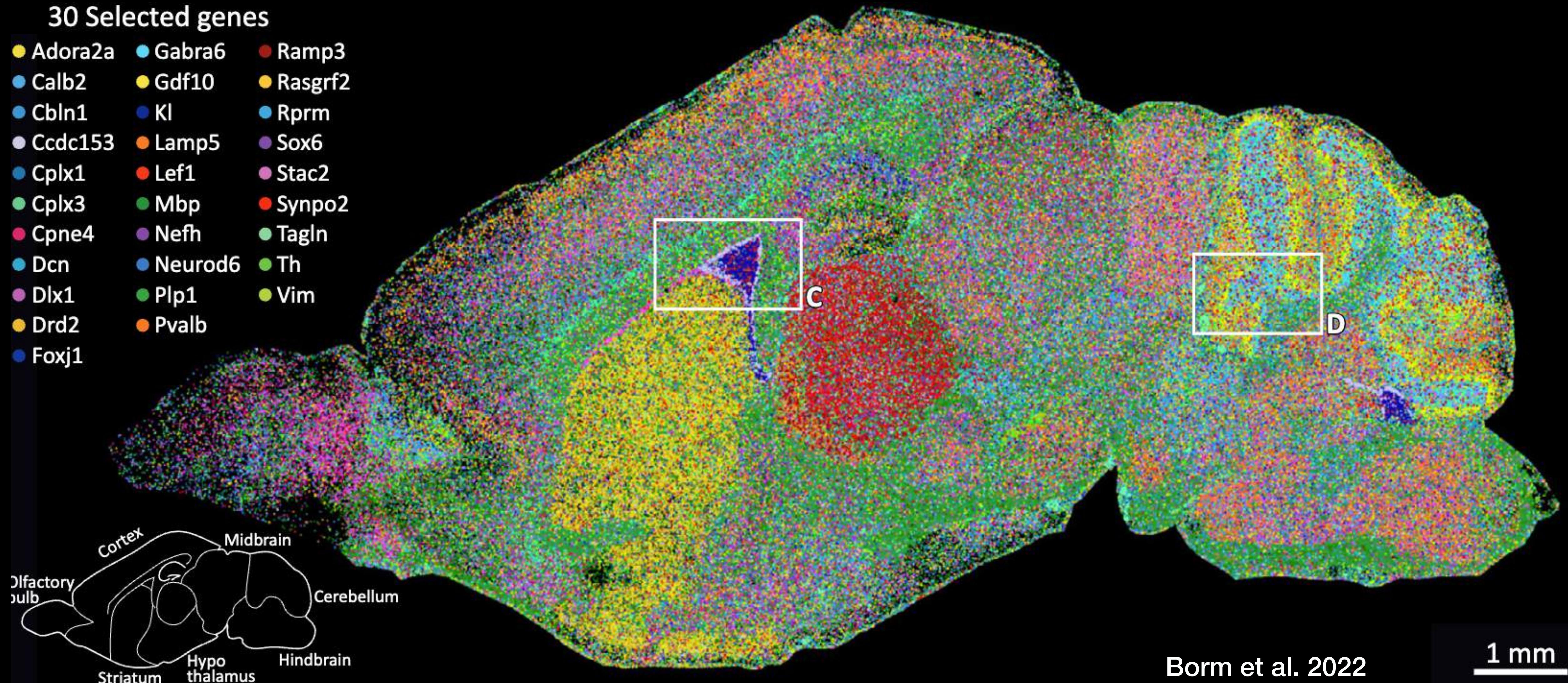
- **Talk and demo:** analysis techniques
- **Coding sessions:** hands-on practice in data analysis

What is spatial transcriptomics?

- Technologies that make *transcripts (RNA)* seen, while preserving their spatial locations in cells or tissues (*in-situ*), with high-throughput (*many cells and many genes simultaneously*).

30 Selected genes

● Adora2a	● Gabra6	● Ramp3
● Calb2	● Gdf10	● Rasgrf2
● Cbln1	● Kl	● Rprm
● Ccdc153	● Lamp5	● Sox6
● Cplx1	● Lef1	● Stac2
● Cplx3	● Mbp	● Synpo2
● Cpne4	● Nefh	● Tagln
● Dcn	● Neurod6	● Th
● Dlx1	● Plp1	● Vim
● Drd2	● Pvalb	
● Foxj1		

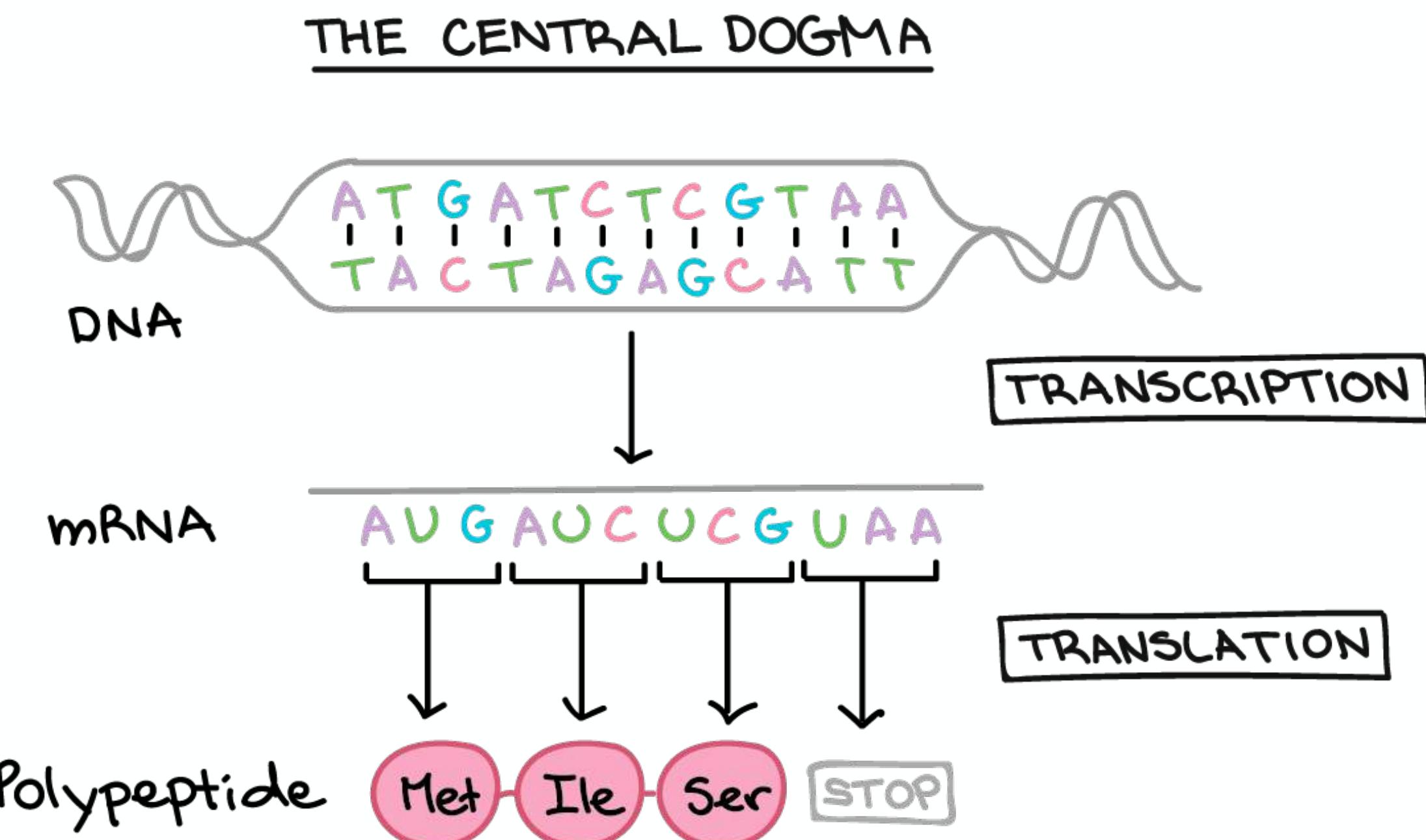


Borm et al. 2022

1 mm

Why do we want to see *mRNAs* in cells and tissues?

- Why not water and lipids?
- Why not DNA?
- Why not proteins?
- Why in cells and tissues (vs in tubes)?
- Why high-throughput?
 - Genes – how many genes in the genome?
 - Cells – how many cells in the brain?



Opinion

The cell as a bag of RNA

Stephen R. Quake^{1,*}

December 2021

Opinion

The cell as a bag of RNA

Stephen R. Quake^{1,*}

December 2021

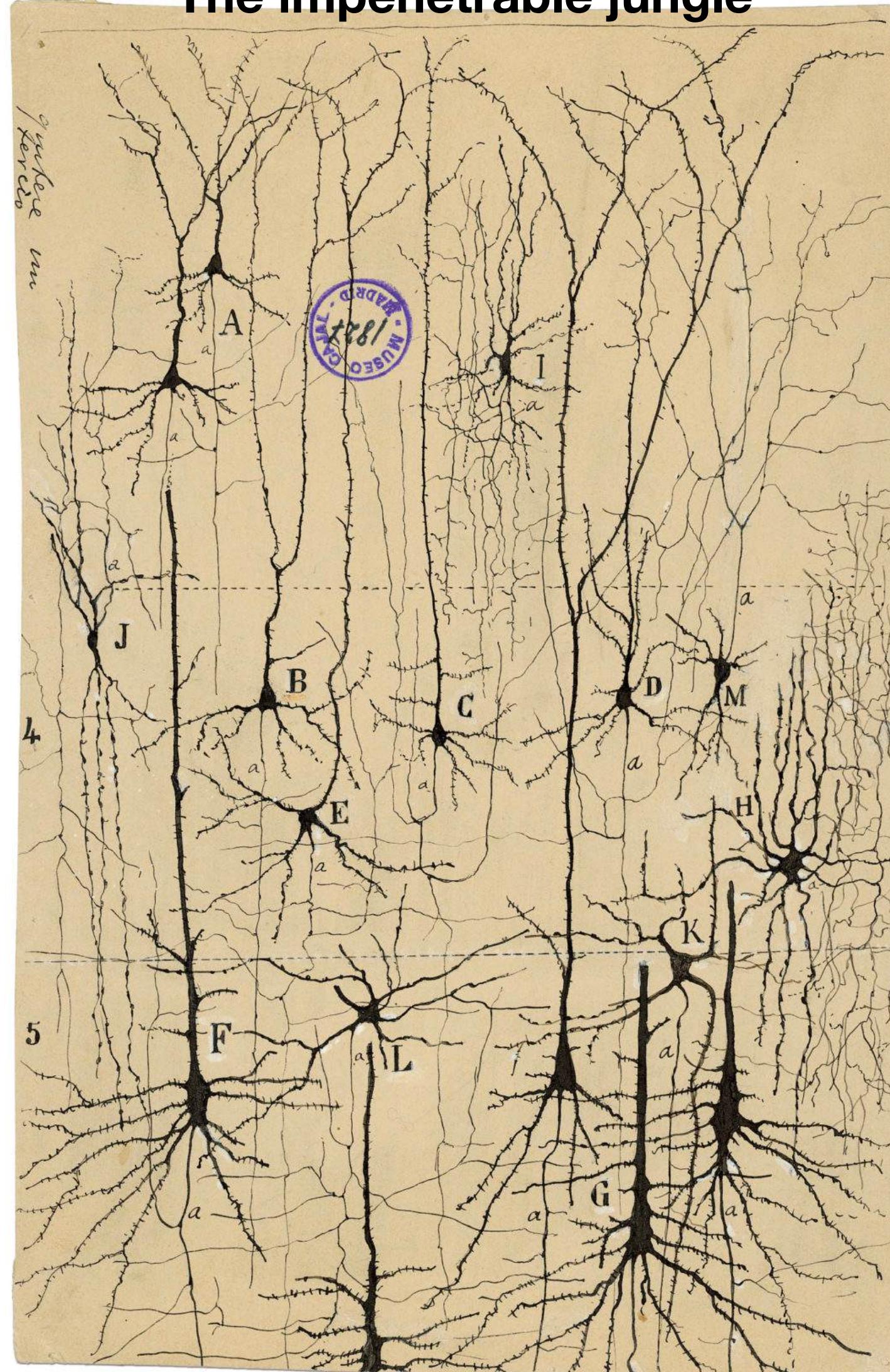
	Cell A	Cell B	Cell C
RNA 1	10	0	100
RNA 2	3	100	3
RNA 3	20	5	10,000
...

A simple but powerful abstraction

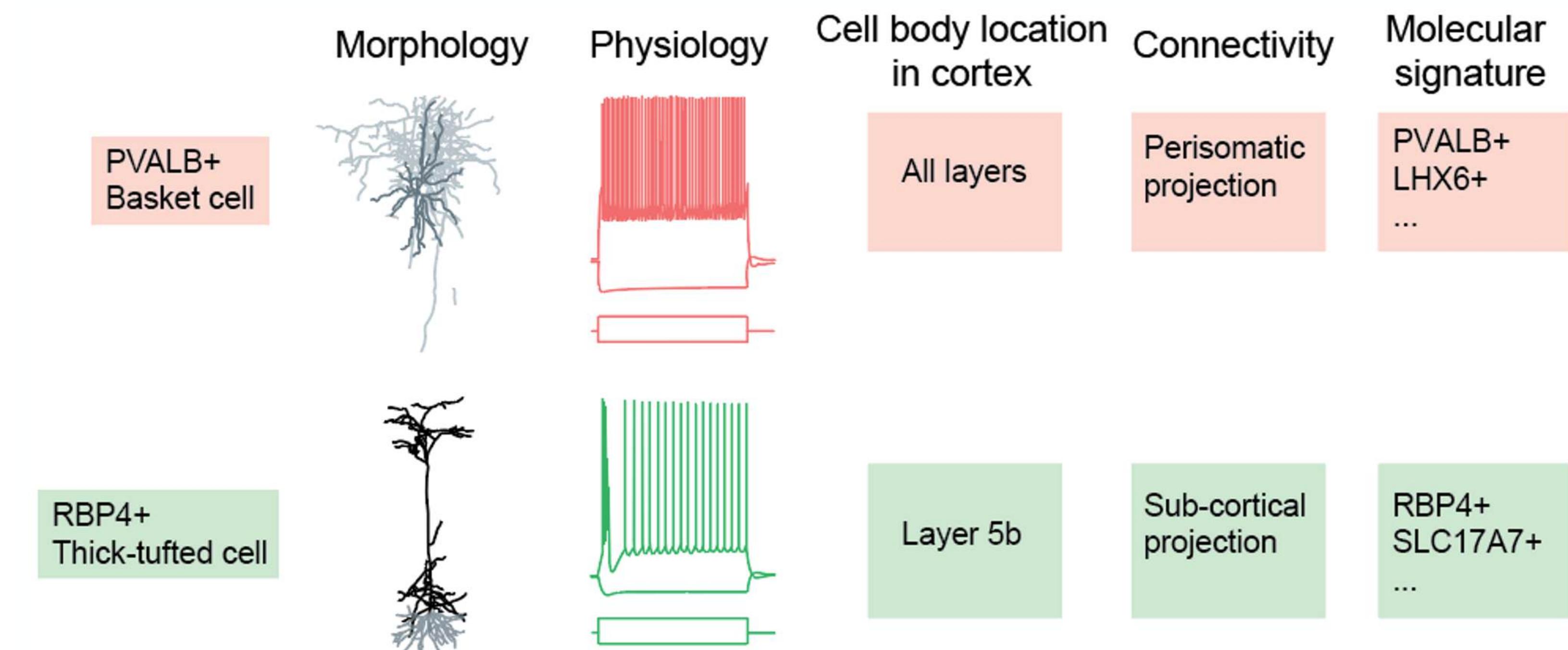
- Enabled by single-cell genomics
- Transformed biology into a quantitative and computational era

Example: diverse cell types and intricate organizations are defining features of the brain

“The impenetrable jungle”



- Golgi's methods randomly stained brain cells, revealing their elaborate shapes.
- Molecular signatures underlies other aspects of cellular properties.



- How do cells grow into different types and organize themselves to fulfill biologic functions?
 - Neuroscience: how does the brain (neural circuits) give rise to the mind?
 - Developmental biology: how do cells divide, differentiate, and organize themselves to form tissue?
 - Cancer biology: what are the biomarkers of tumors and tumor heterogeneity?
- It is not enough to understand molecules (parts), we need to understand how they **organize** into organisms.
- All living organisms are alike – created from the same simple parts. “What I cannot create, I do not understand.”

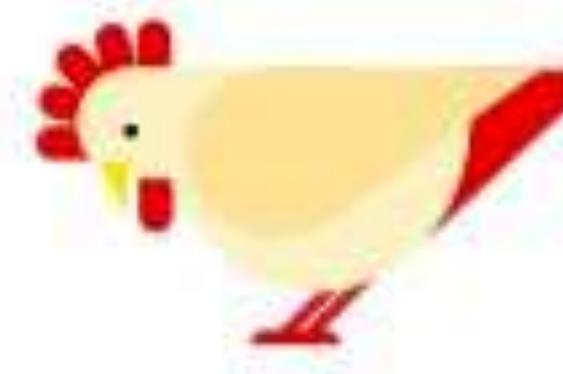
Species	<i>Escherichia coli</i>	<i>Gallus gallus</i>	<i>Homo sapiens</i>	<i>Daphnia pulex</i>	<i>Oryza sativa</i>
Number of Genes	~4,200	~17,000	~21,000	~31,000	~38,000
Common Name					
Bacteria	Chicken	Human	Water flea	Rice	

Image from Bio Ninja

“The genomic bottleneck” of wiring

— an information-theory perspective

- What is the size (number of base pairs) of the human genome? $n \sim 3 \text{ billion}$
- How much information is there in the human genome? $2 \text{ bits}^n \sim 6 \text{ billion bits} < 1 \text{ GB}$
- How many neurons does a human have? $N \sim 100 \text{ billion}$
- How much information is needed to specify the wiring of neurons? $N^2 \sim kN \log N$

The genome is strikingly small for what it does.

How much space is needed to store the information below?

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181, 6765, 10946, 17711, 28657, 46368, 75025, 121393, 196418, 317811, 514229, 832040, 1346269, 2178309, 3524578, 5702887, 9227465, 14930352, 24157817, 39088169, 63245986, 102334155, 165580141, 267914296, 433494437, 701408733, 1134903170, 1836311903, 2971215073, 4807526976, 7778742049, 12586269025, 20365011074, 32951280099, 53316291173, 86267571272, 139583862445, 225851433717, 365435296162, 591286729879, 956722026041, 1548008755920, 2504730781961, 4052739537881, 6557470319842, 10610209857723, 17167680177565, 27777890035288, 44945570212853, 72723460248141, 117669030460994, 190392490709135, 308061521170129, 498454011879264, 806515533049393, 1304969544928657, 2111485077978050, 3416454622906707, 5527939700884757, 8944394323791464, 14472334024676221, 23416728348467685, 37889062373143906, 61305790721611591, 99194853094755497, 160500643816367088, 259695496911122585, 420196140727489673, 679891637638612258, 110008778366101931, 1779979416004714189, 2880067194370816120, 4660046610375530309, 7540113804746346429, 12200160415121876738, 19740274219868223167, 31940434634990099905, 51680708854858323072, 83621143489848422977, 135301852344706746049, 218922995834555169026, 354224848179261915075, 573147844013817084101, 927372692193078999176,

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181, 6765, 10946, 17711, 28657, 46368, 75025, 121393, 196418, 317811, 514229, 832040, 1346269, 2178309, 3524578, 5702887, 9227465, 14930352, 24157817, 39088169, 63245986, 102334155, 165580141, 267914296, 433494437, 701408733, 1134903170, 1836311903, 2971215073, 4807526976, 7778742049, 12586269025, 20365011074, 32951280099, 53316291173, 86267571272, 139583862445, 225851433717, 365435296162, 591286729879, 956722026041, 1548008755920, 2504730781961, 4052739537881, 6557470319842, 10610209857723, 17167680177565, 27777890035288, 44945570212853, 72723460248141, 117669030460994, 190392490709135, 308061521170129, 498454011879264, 806515533049393, 1304969544928657, 2111485077978050, 3416454622906707, 5527939700884757, 8944394323791464, 14472334024676221, 23416728348467685, 37889062373143906, 61305790721611591, 99194853094755497, 160500643816367088, 259695496911122585, 420196140727489673, 679891637638612258, 1100087778366101931, 1779979416004714189, 2880067194370816120, 4660046610375530309, 7540113804746346429, 12200160415121876738, 19740274219868223167, 31940434634990099905, 51680708854858323072, 83621143489848422977, 135301852344706746049, 218922995834555169026, 354224848179261915075, 573147844013817084101, 927372692193078999176, ...

$$\begin{aligned}a_0 &= 1, \\a_1 &= 1, \\a_{n+1} &= a_n + a_{n-1} \dots\end{aligned}$$

- Complex patterns can be generated by compact rules
- Complex biological organizations have to be generated by much simpler developmental rules, given the limited space in DNA
- Many of these rules are unknown.

Opinion

The cell as a bag of RNA

Stephen R. Quake^{1,*}

December 2021

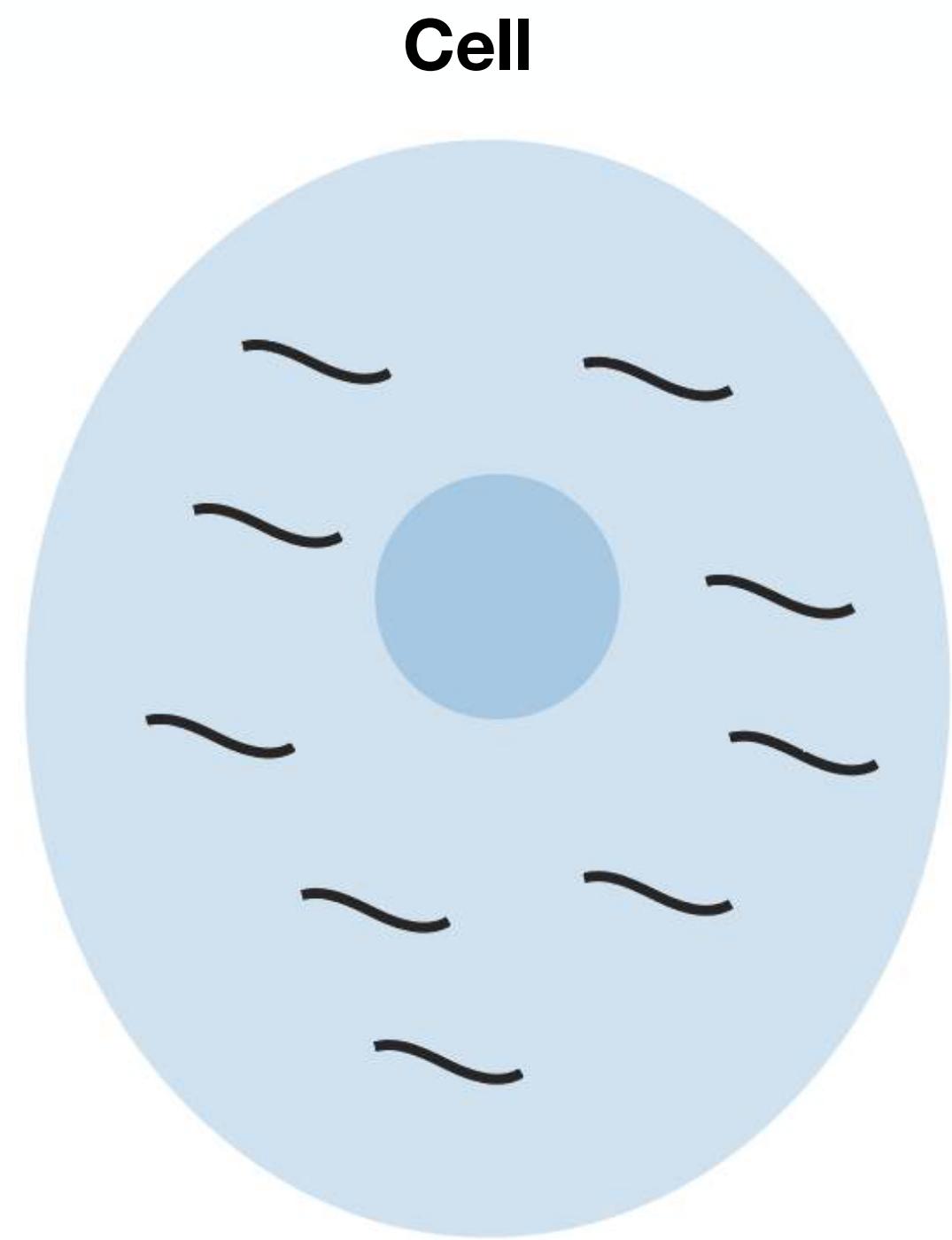
	Cell A	Cell B	Cell C
RNA 1	10	0	100
RNA 2	3	100	3
RNA 3	20	5	10,000
...

How can we measure RNAs in a cell?

How many types of RNAs are there in a cell?

How many copies of each type are there in a cell?

How can mRNA be seen?

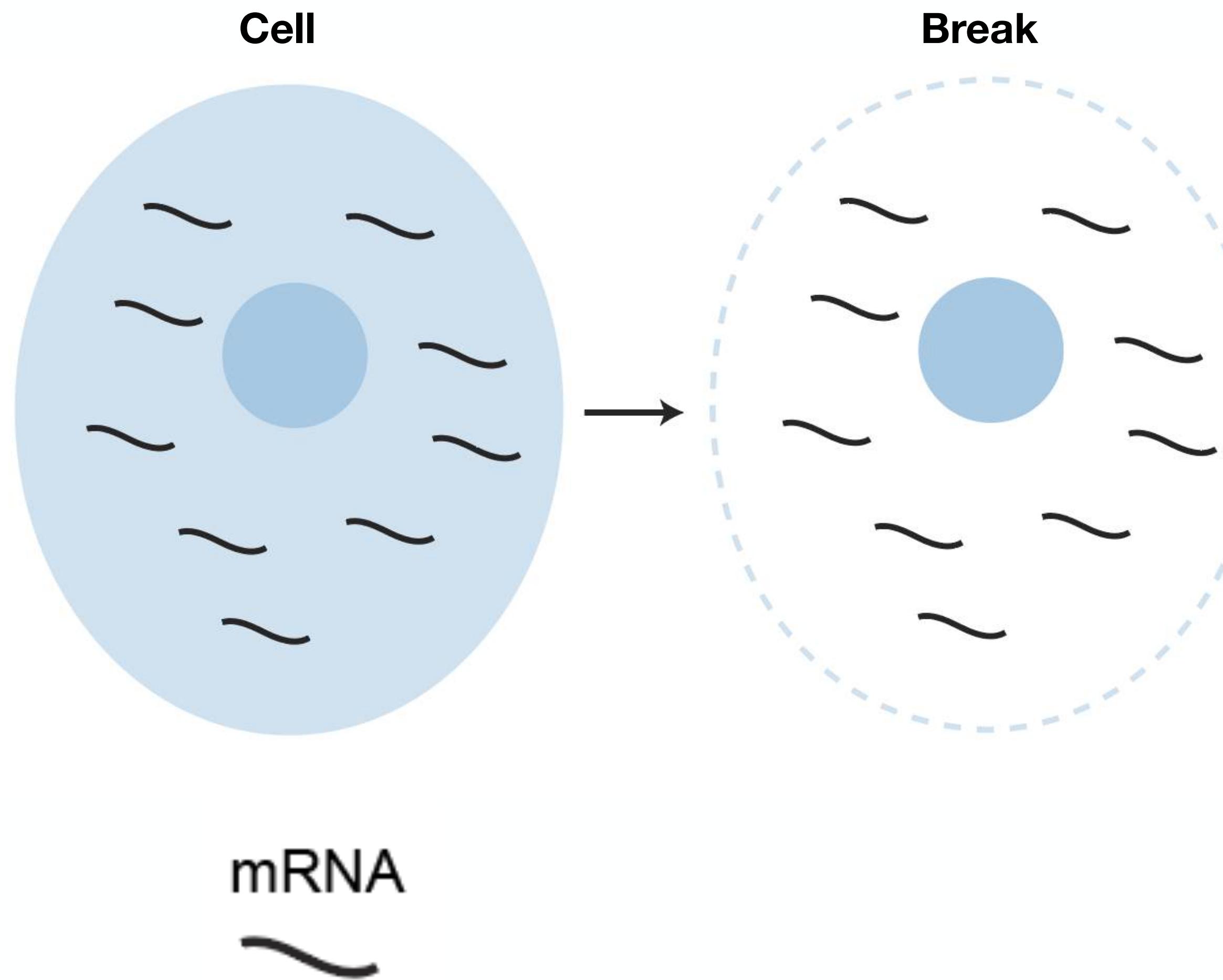


mRNA



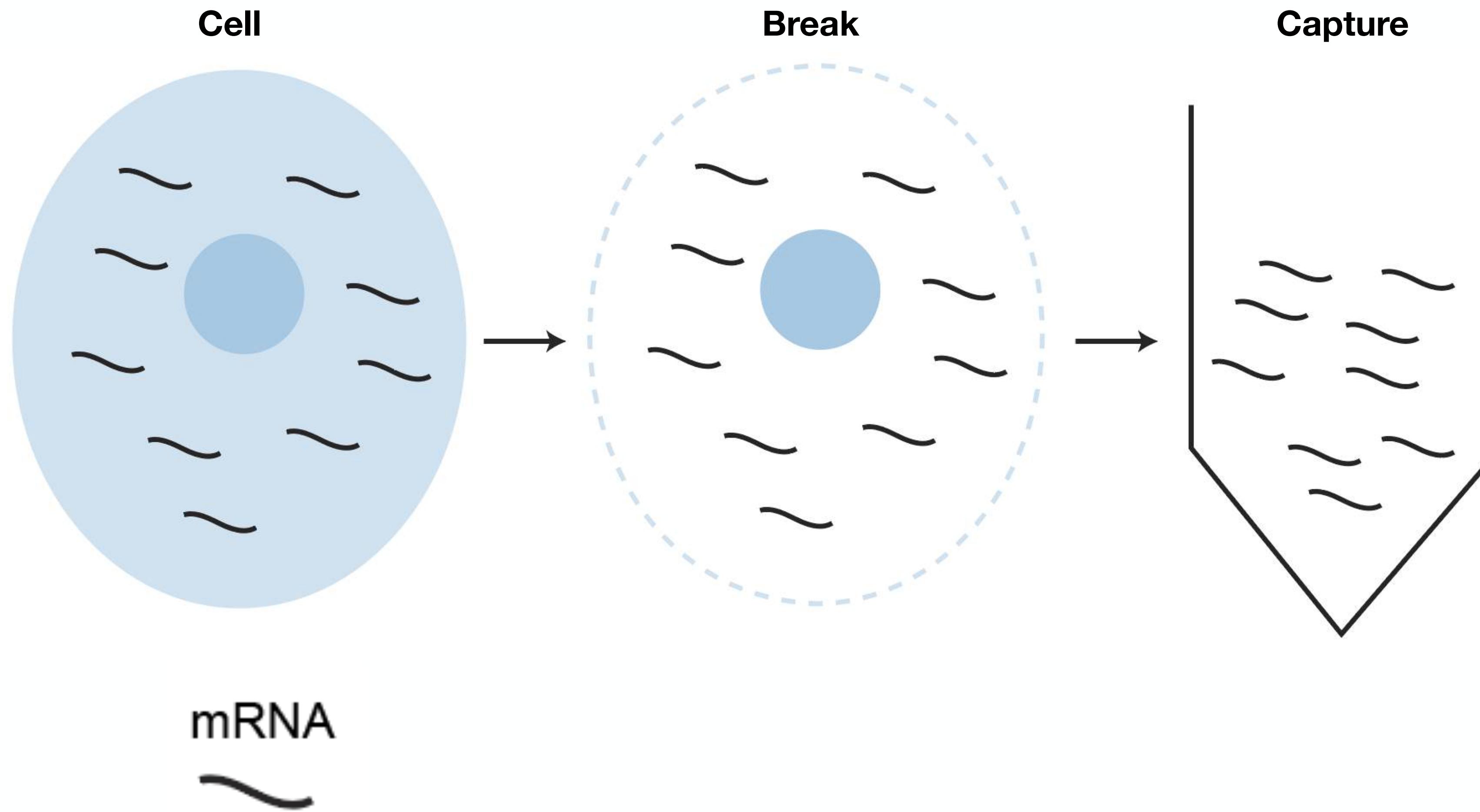
How can mRNA be seen?

Method 1: Sequencing



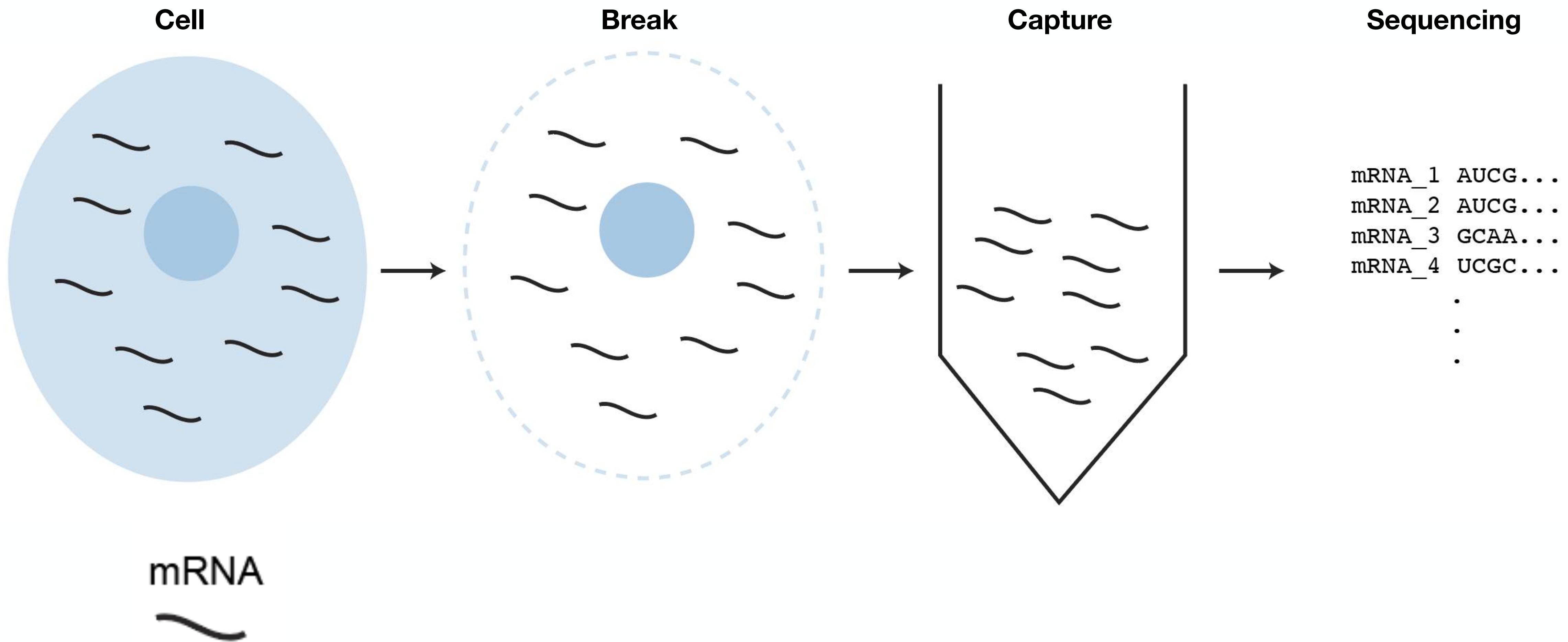
How can mRNA be seen?

Method 1: Sequencing



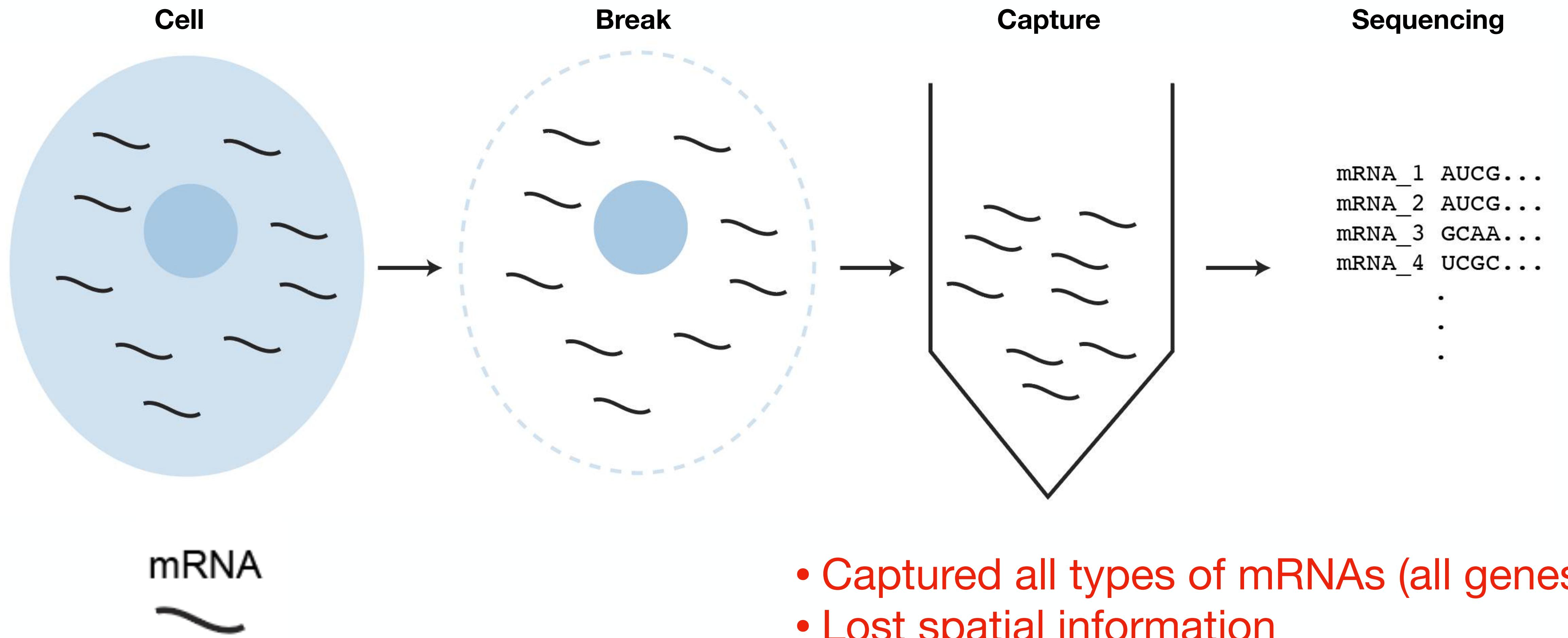
How can mRNA be seen?

Method 1: Sequencing



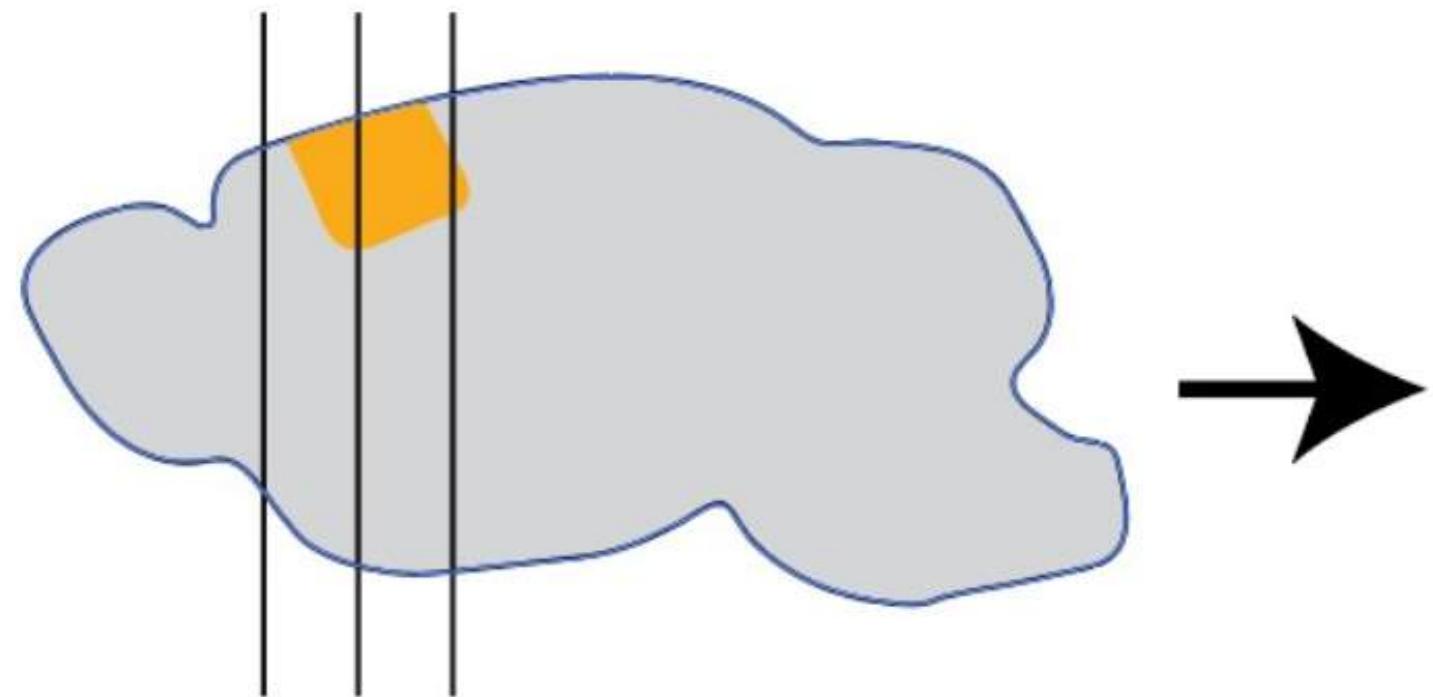
How can mRNA be seen?

Method 1: Sequencing



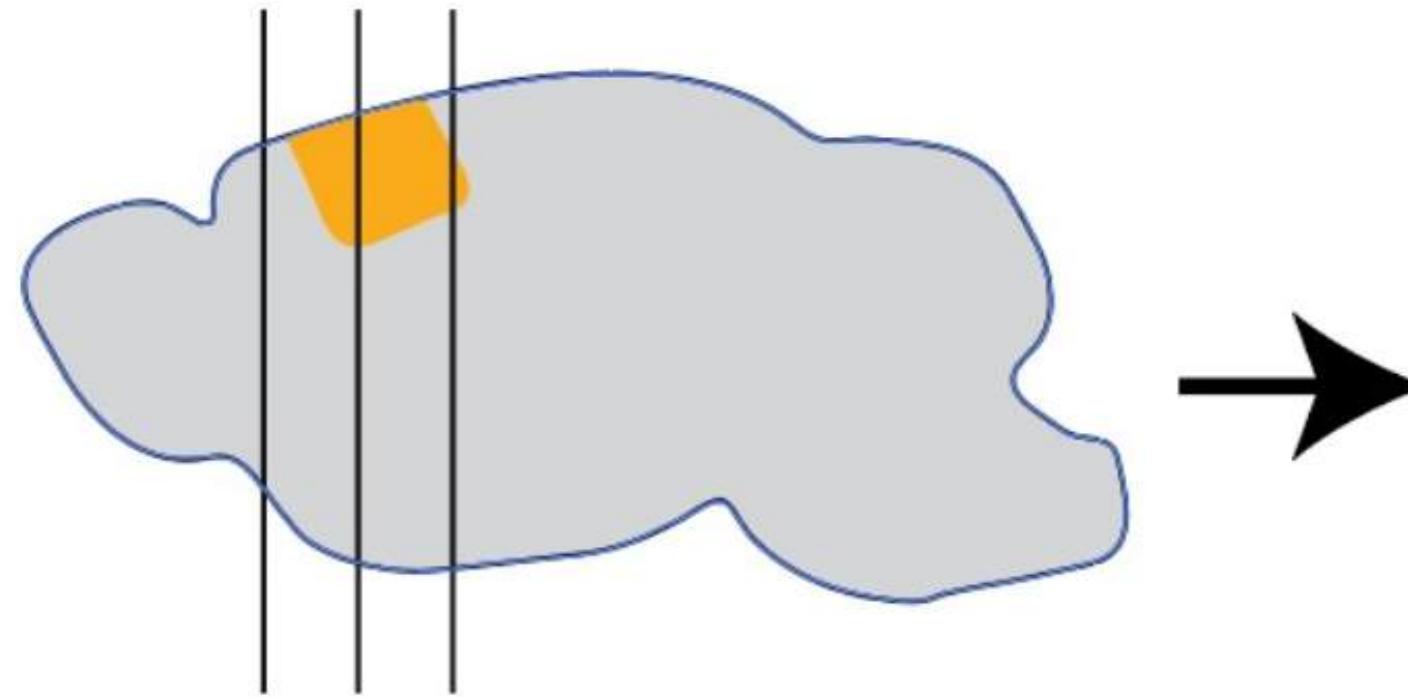
Single-cell RNA-sequencing (scRNA-seq)

Tissue dissection

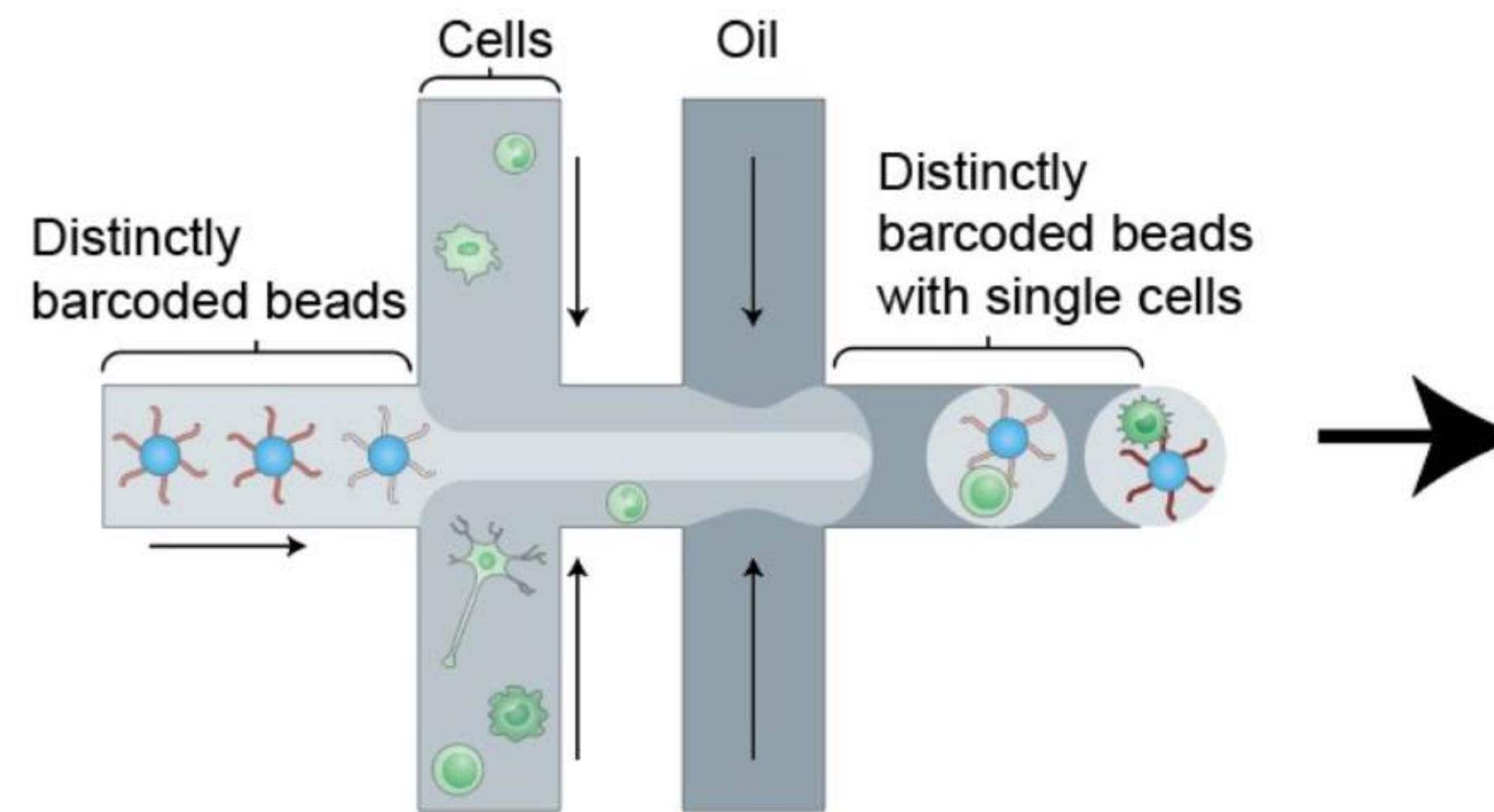


Single-cell RNA-sequencing (scRNA-seq)

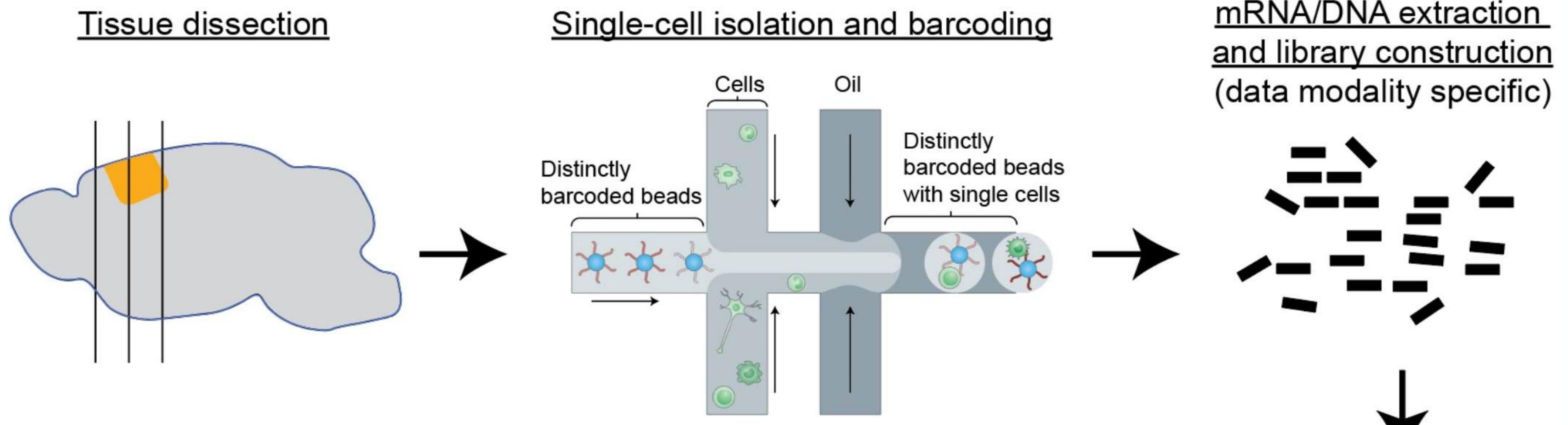
Tissue dissection



Single-cell isolation and barcoding

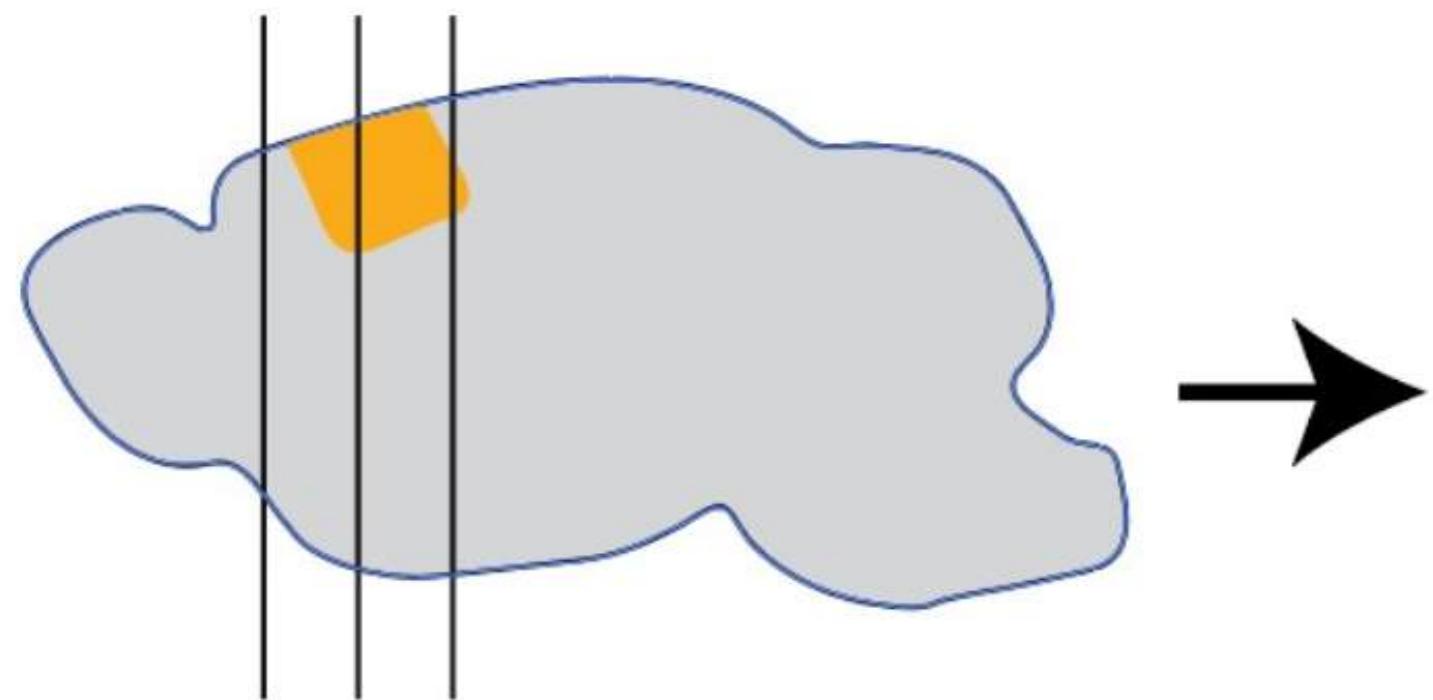


Single-cell RNA-sequencing (scRNA-seq)

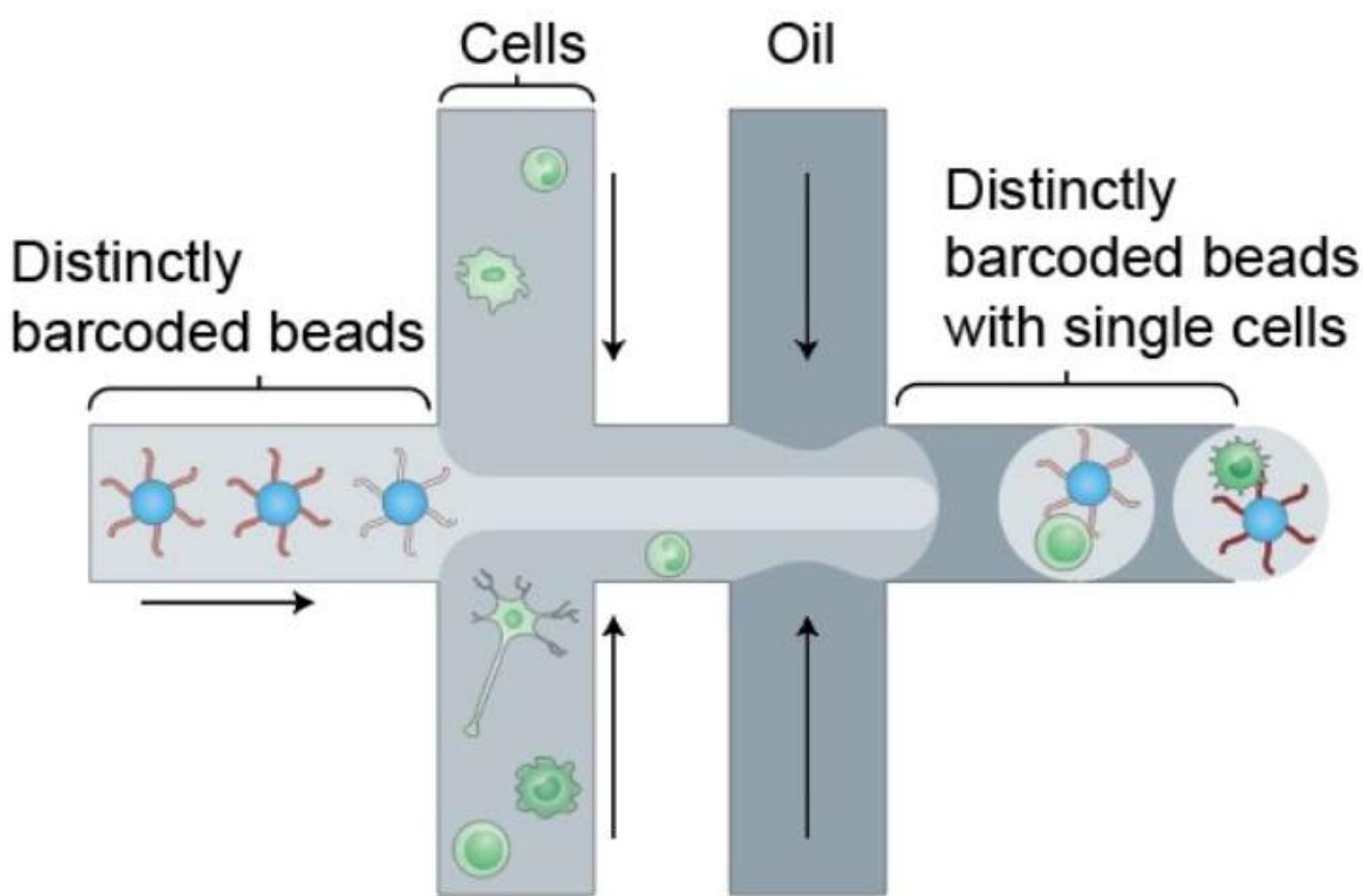


Single-cell RNA-sequencing (scRNA-seq)

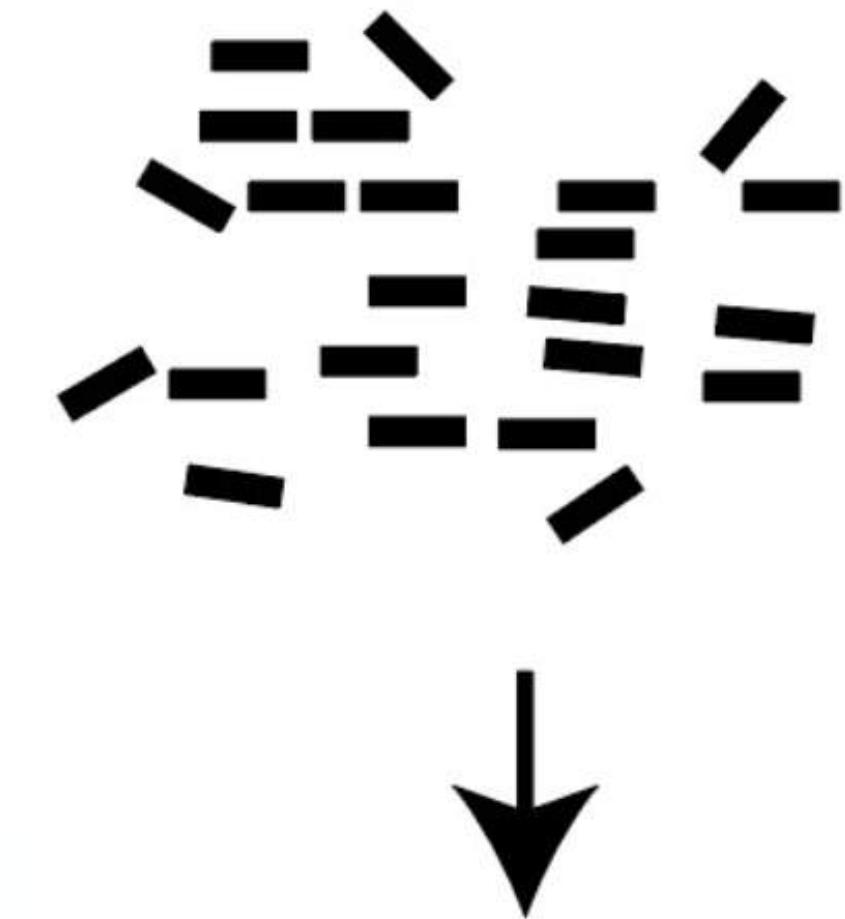
Tissue dissection



Single-cell isolation and barcoding



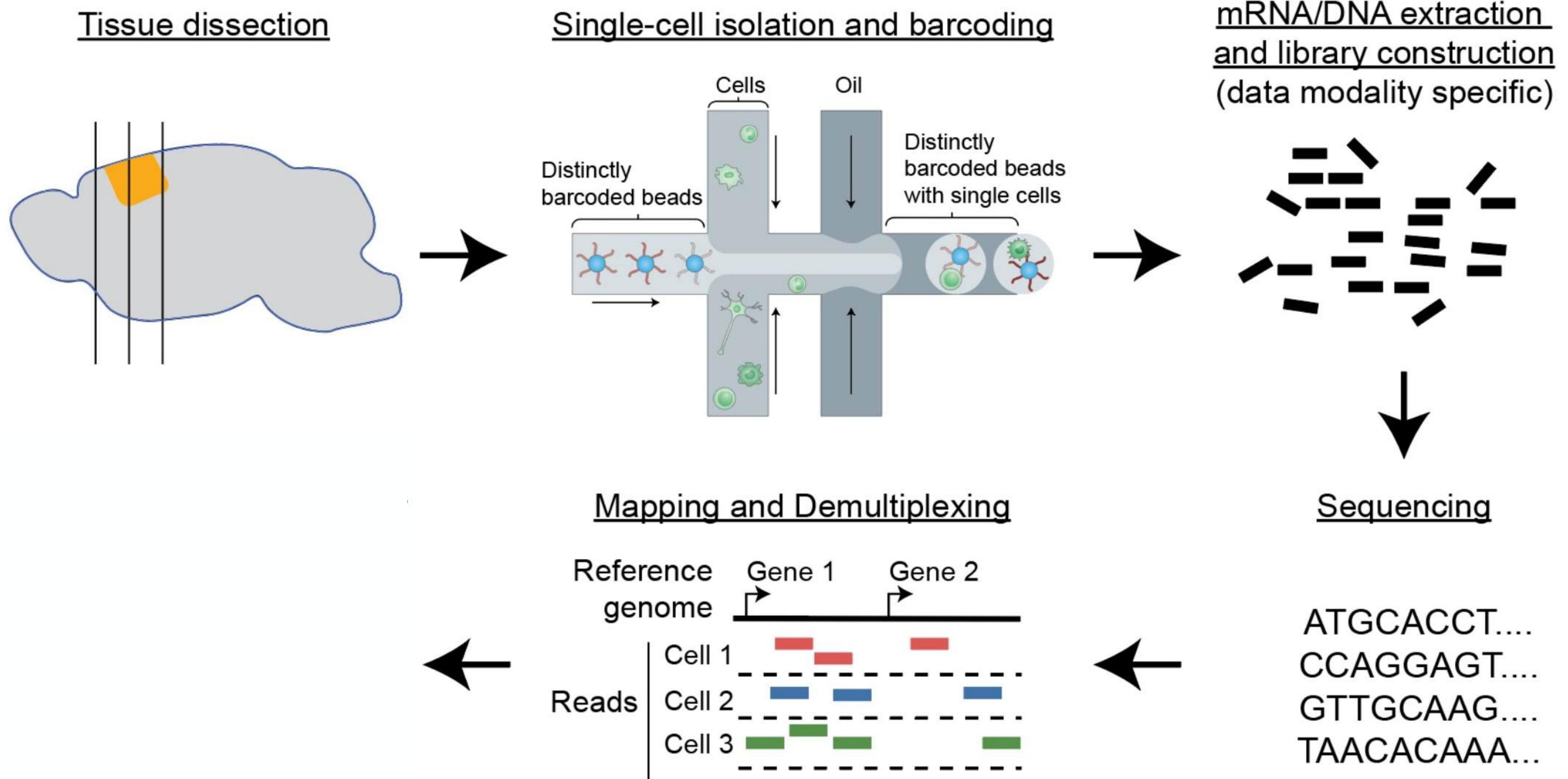
mRNA/DNA extraction and library construction (data modality specific)



Sequencing

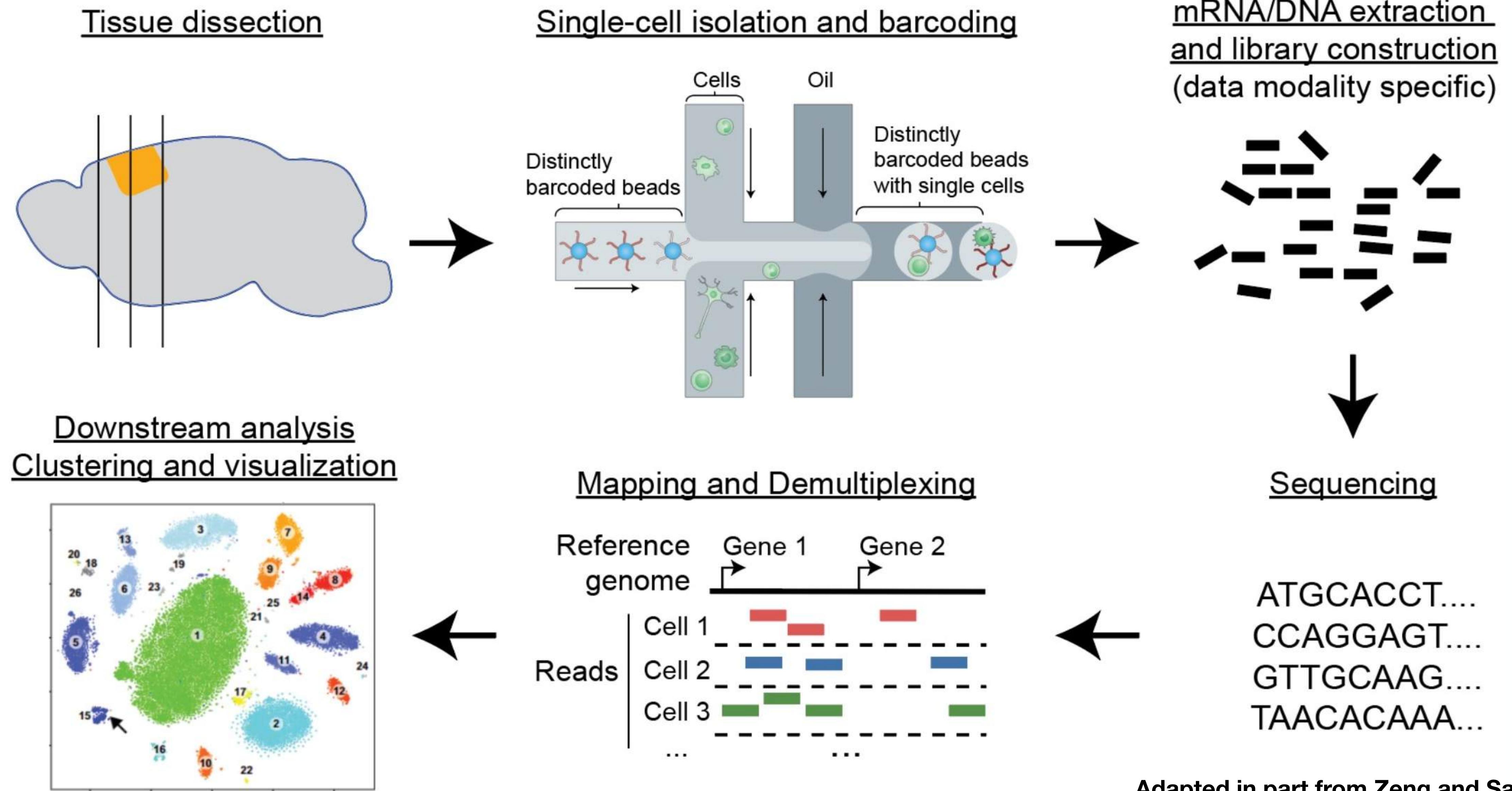
ATGCACCT....
CCAGGAGT....
GTTGCAAG....
TAACACAAA...

Single-cell RNA-sequencing (scRNA-seq)



Adapted in part from Zeng and Sane, 2017

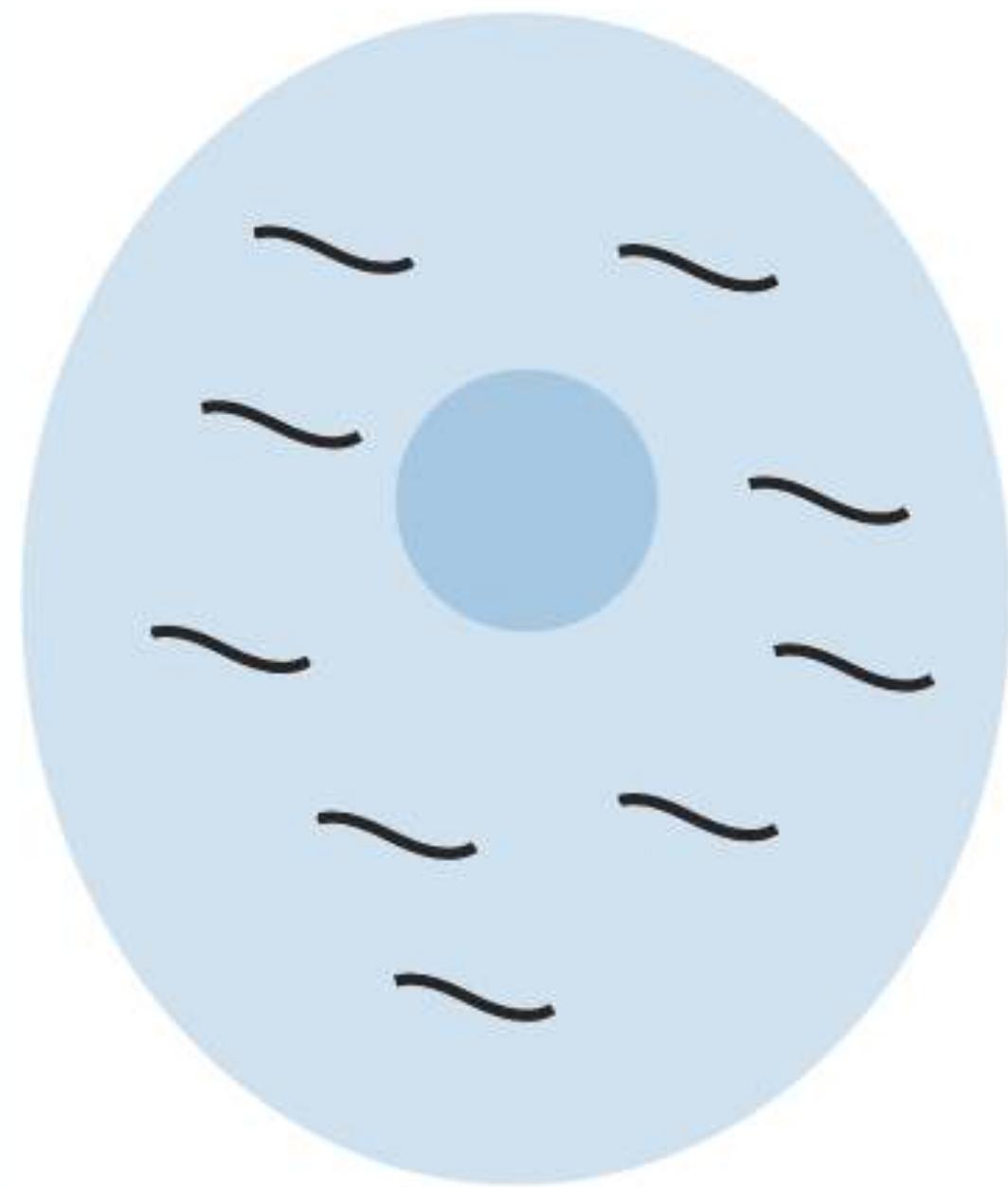
Single-cell RNA-sequencing (scRNA-seq)



How can mRNA be seen?

Method 2: Staining and Imaging

Cell

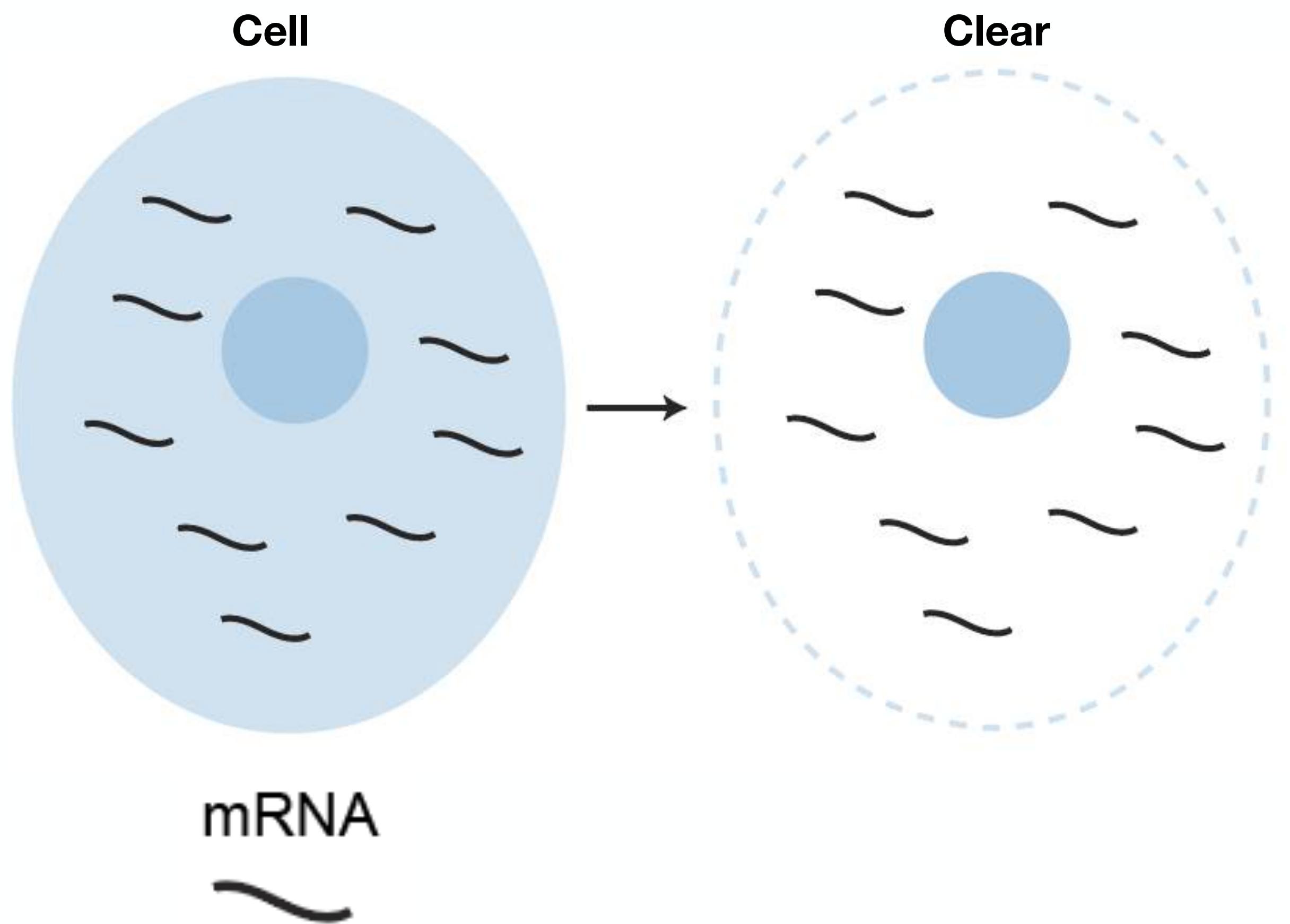


mRNA



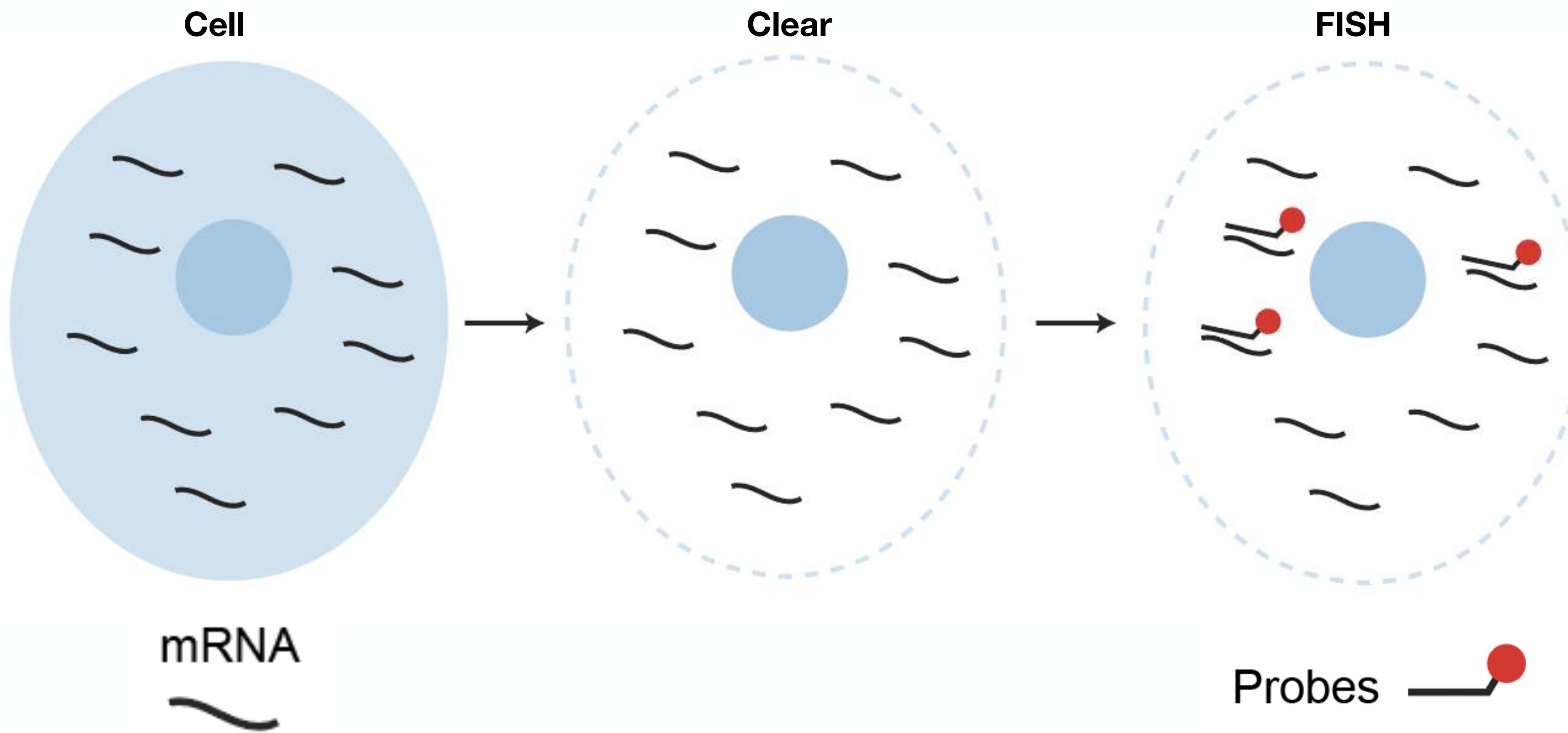
How can mRNA be seen?

Method 2: Staining and Imaging



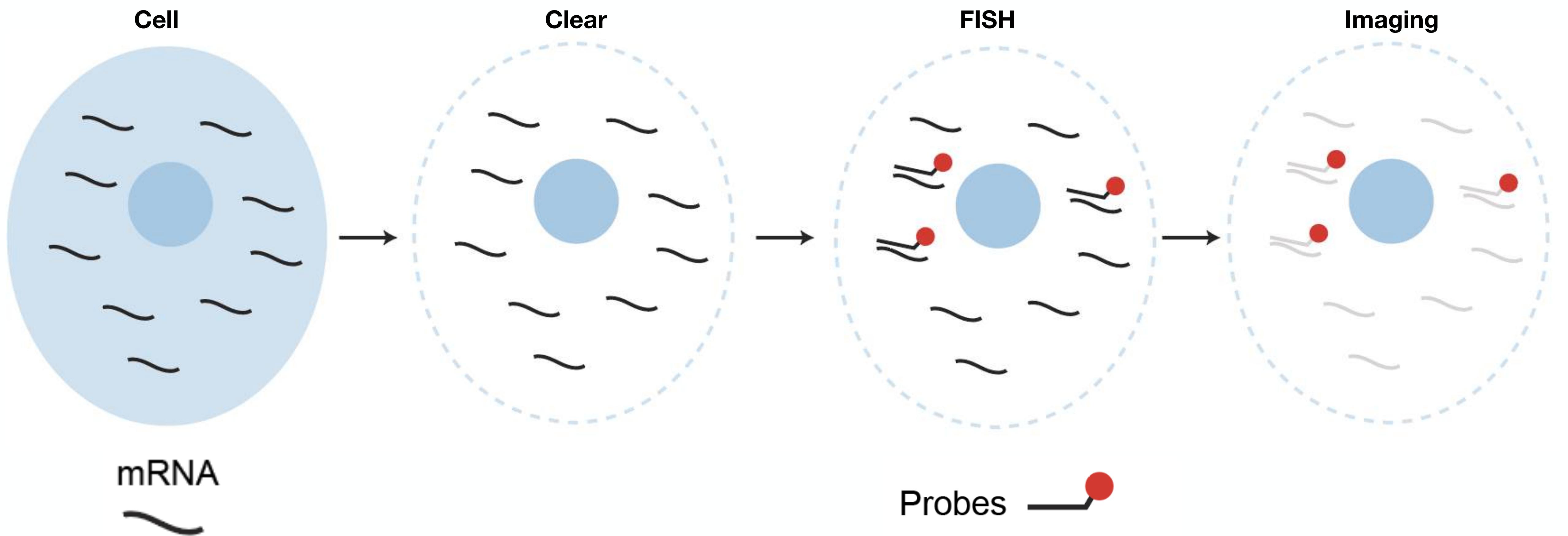
How can mRNA be seen?

Method 2: Staining and Imaging



How can mRNA be seen?

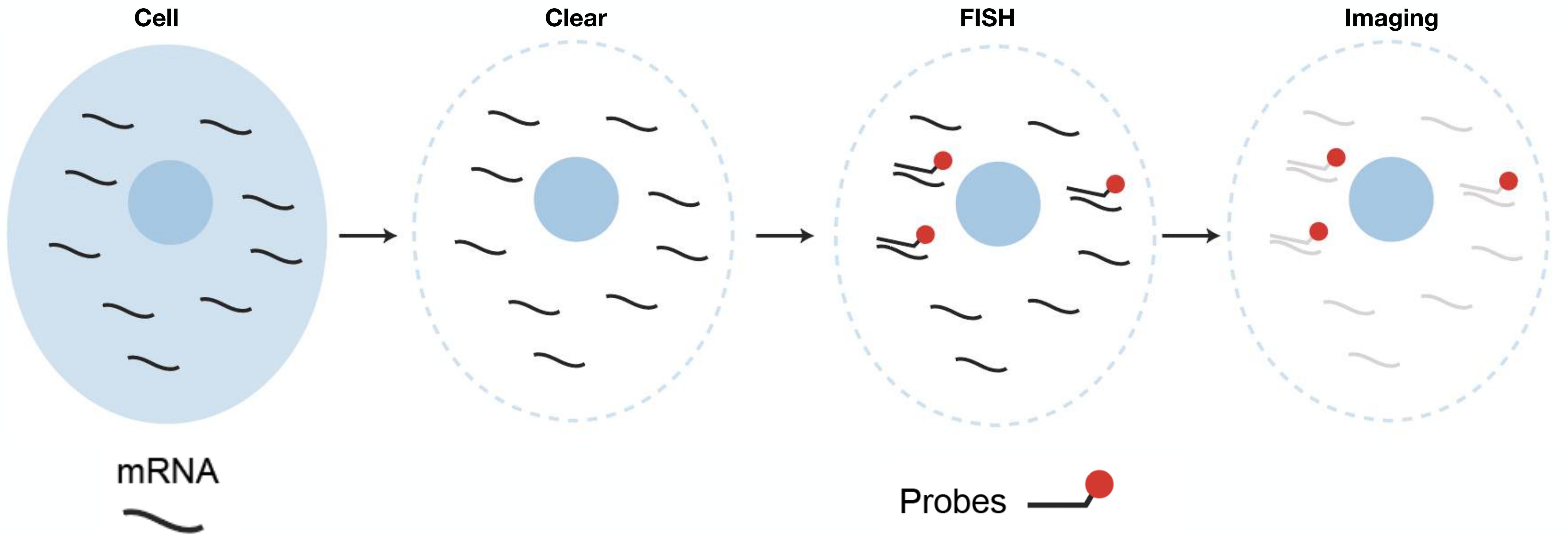
Method 2: Staining and Imaging



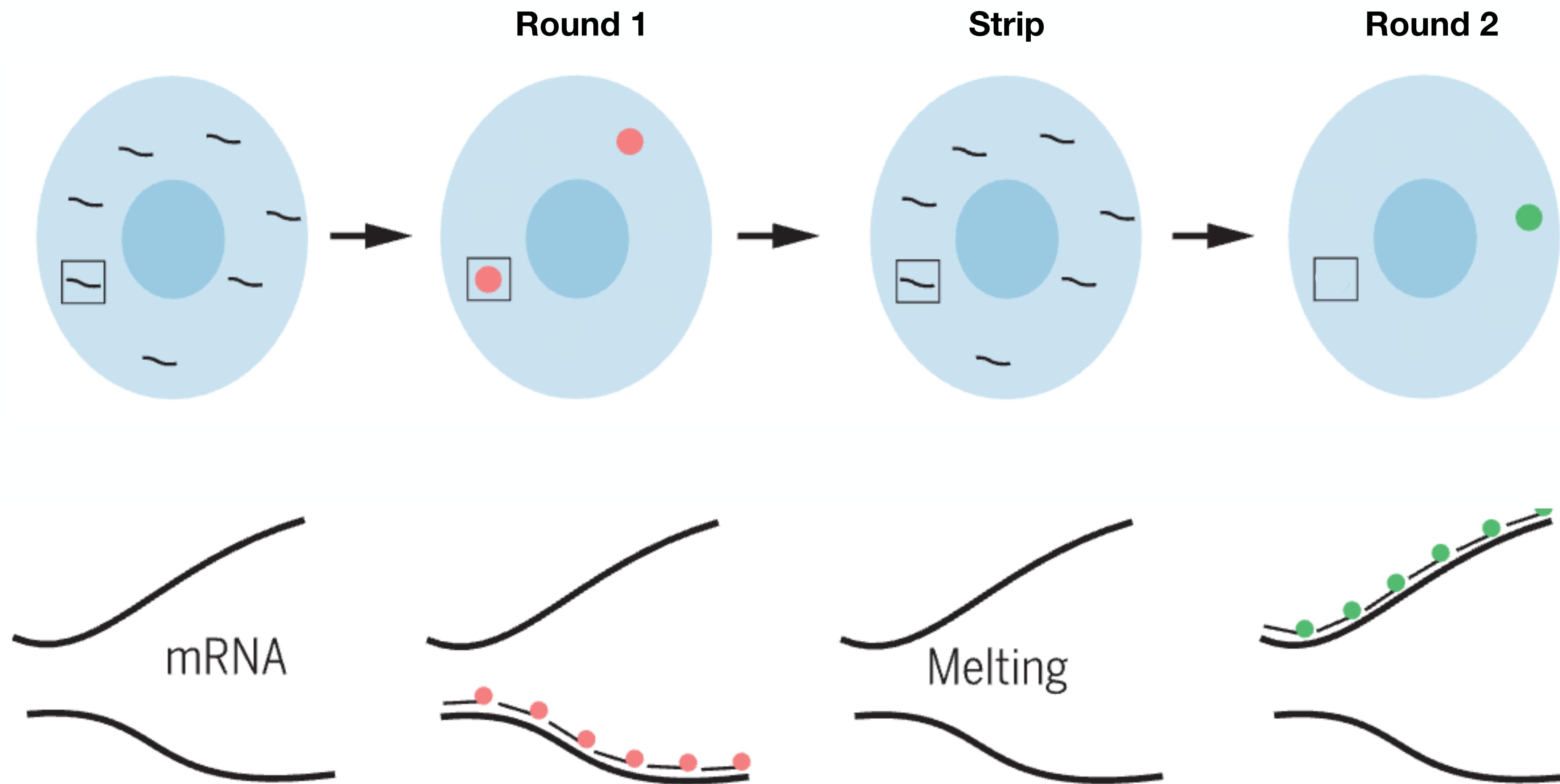
How can mRNA be seen?

Method 2: Staining and Imaging

- Captured 1 type of mRNA (1 gene)
- Preserved spatial information

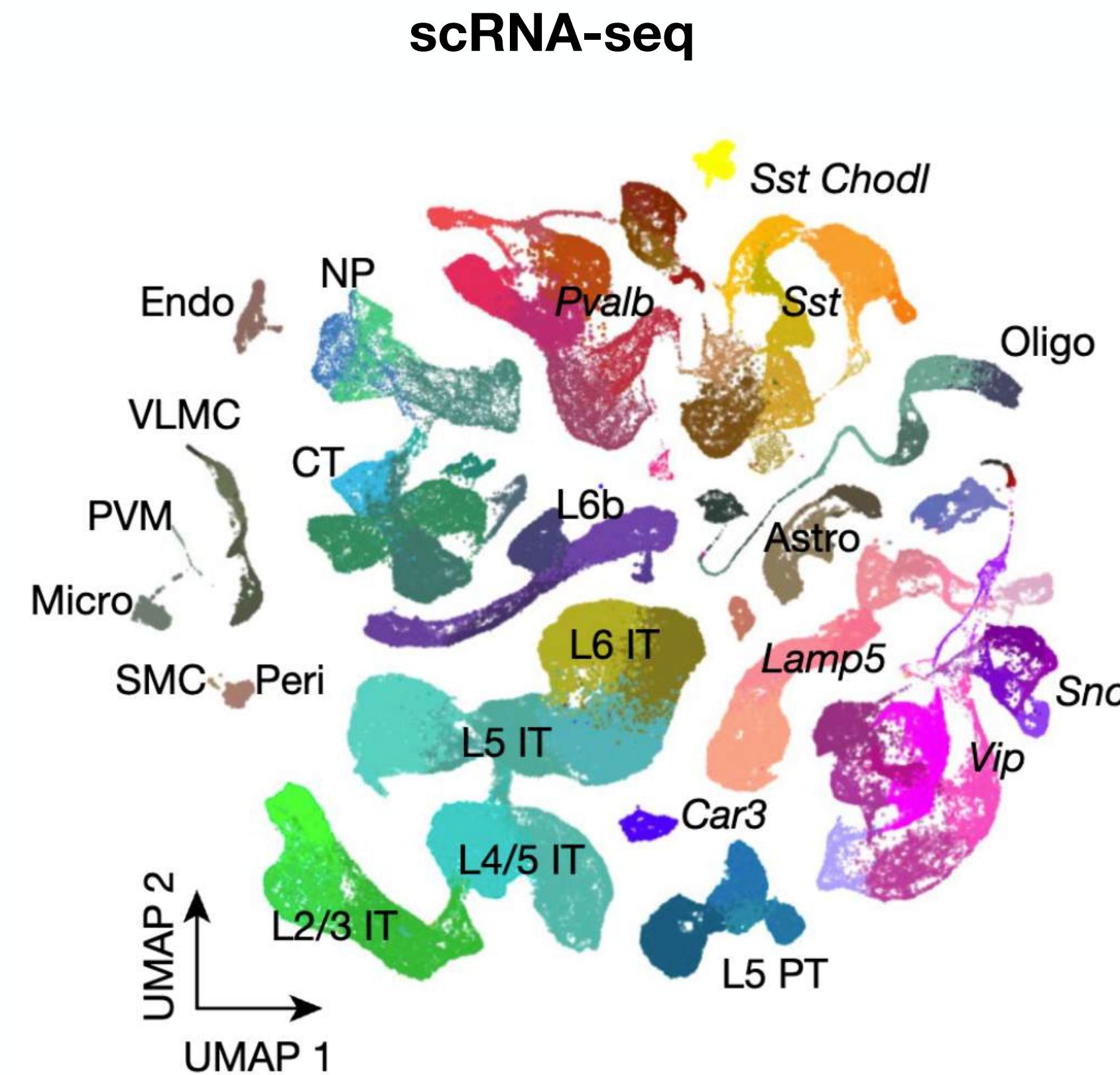
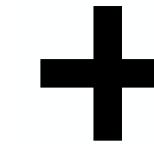
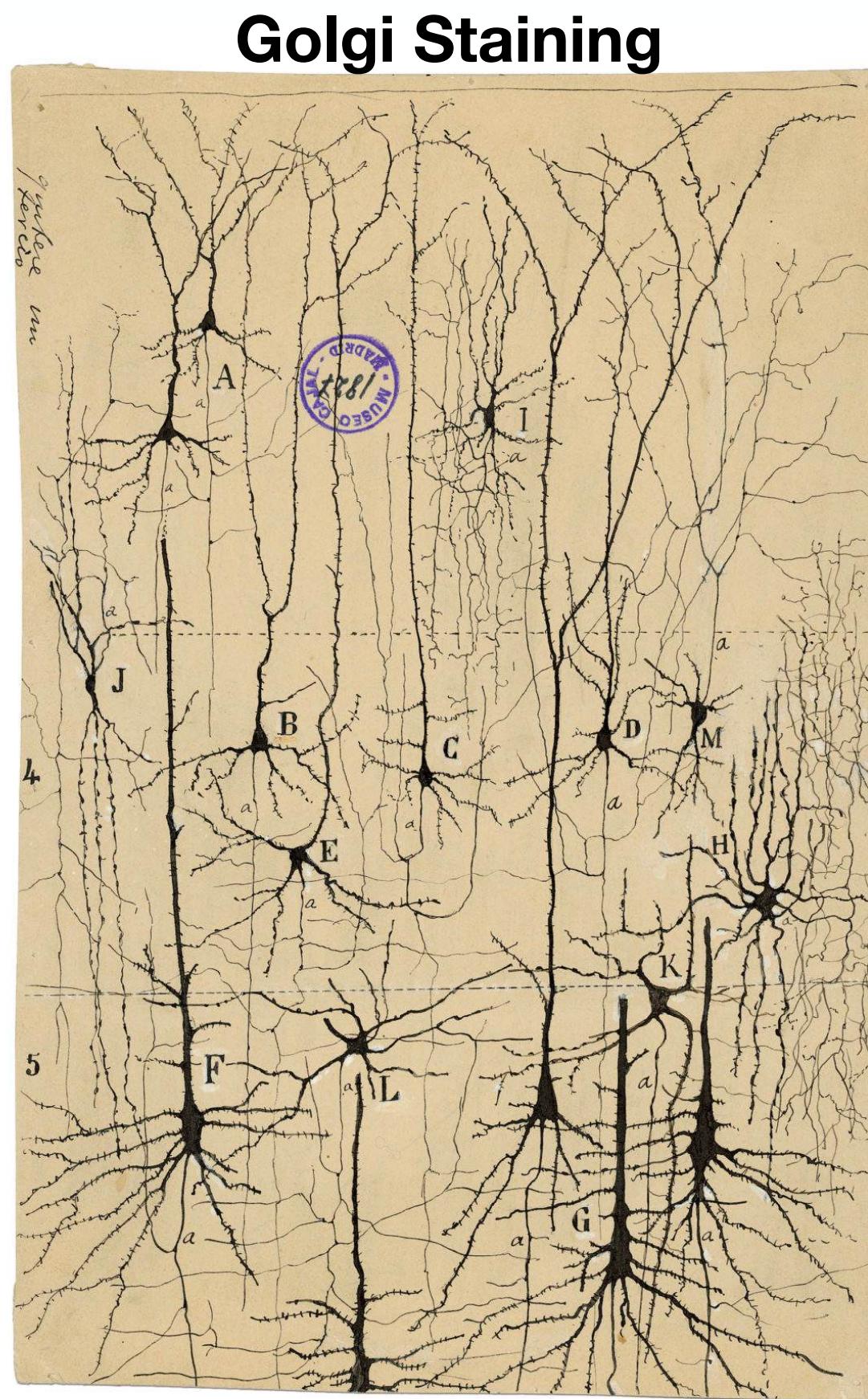


Single molecule Fluorescence In-Situ Hybridization (smFISH)

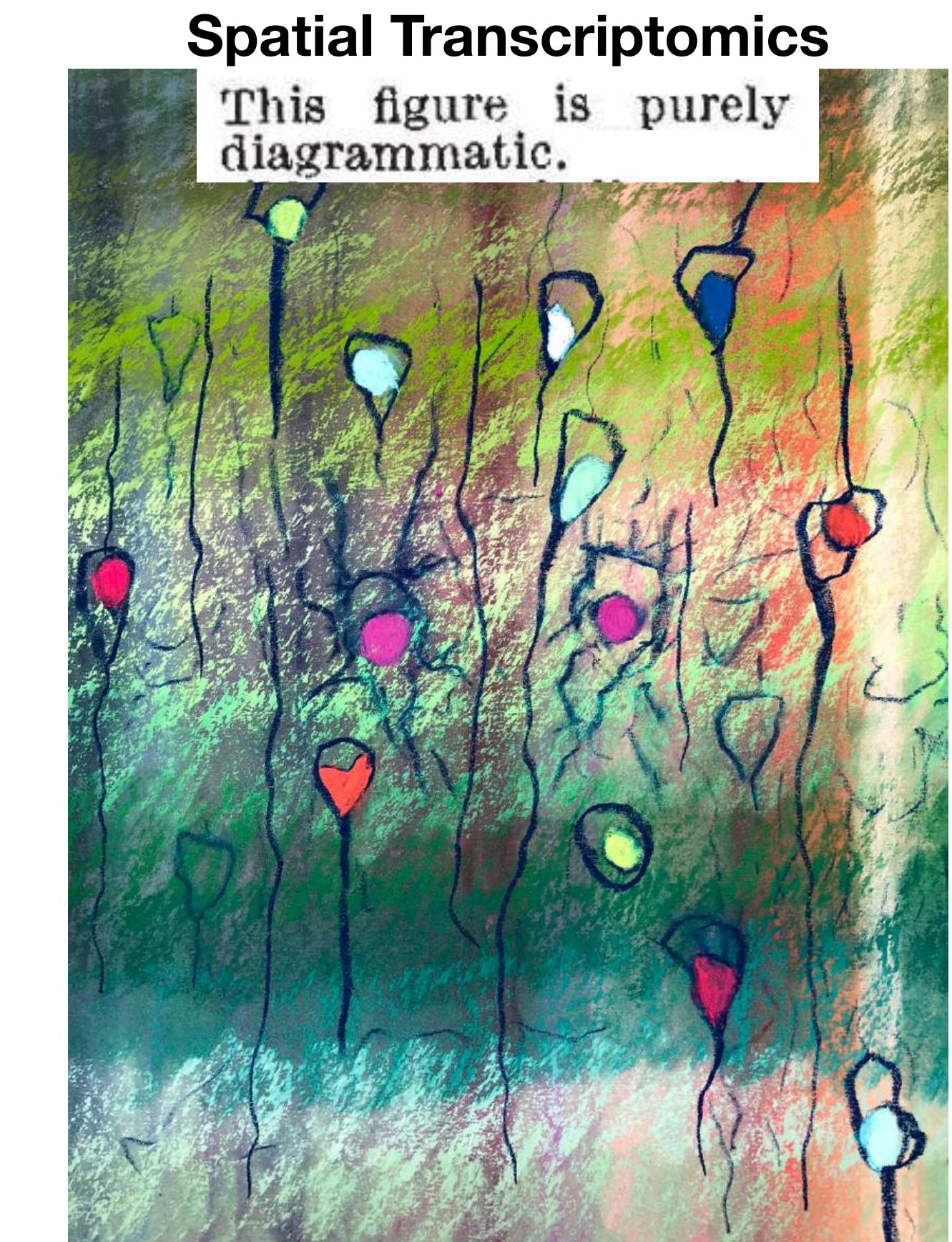


Spatial Transcriptomics bridges histology and genomics

- Making histology (staining) molecularly specific.
- Making genomics (sequencing) spatially resolved.



Yao et al. 2021

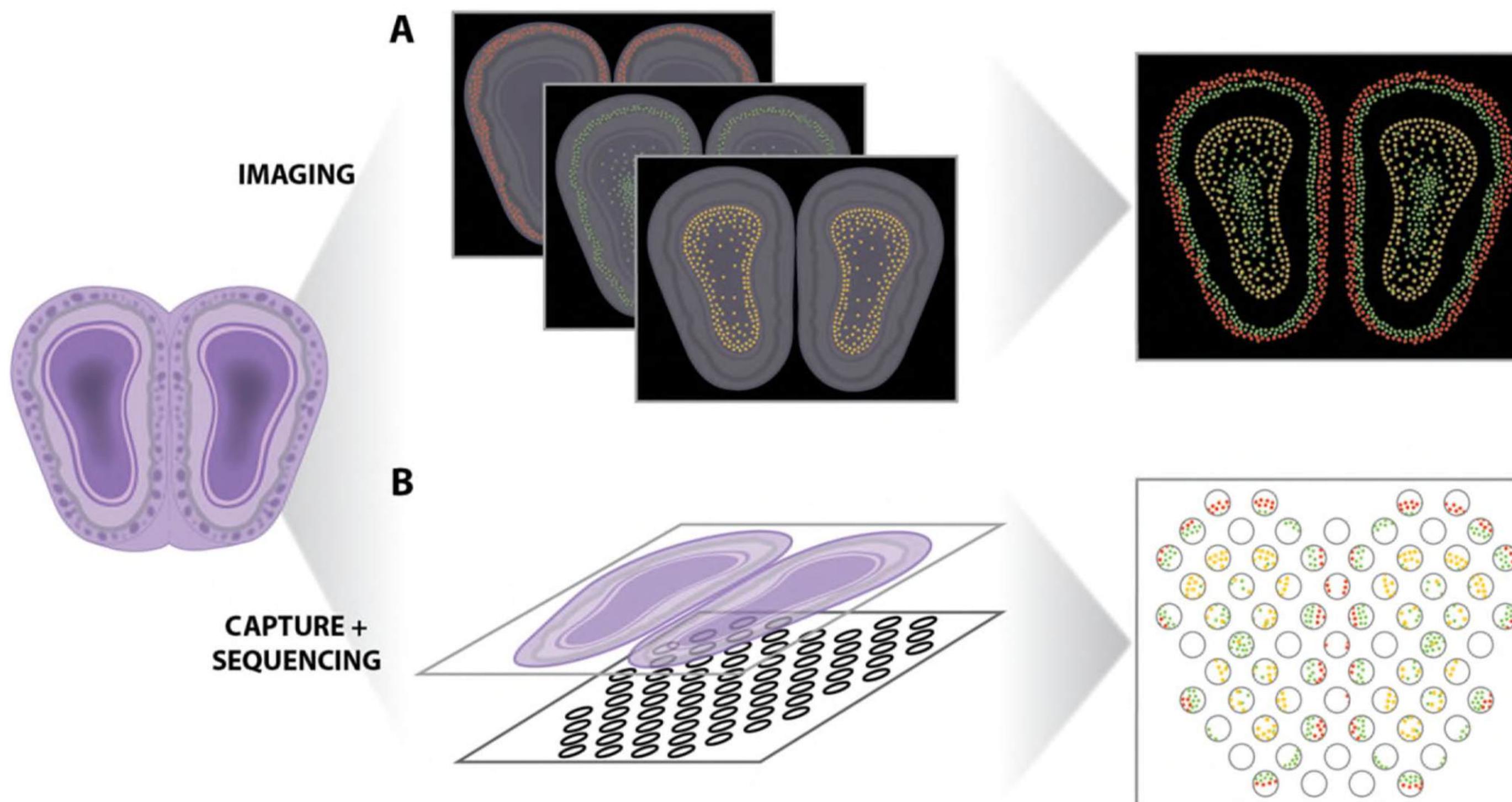


Ramon y Cajal, 1899

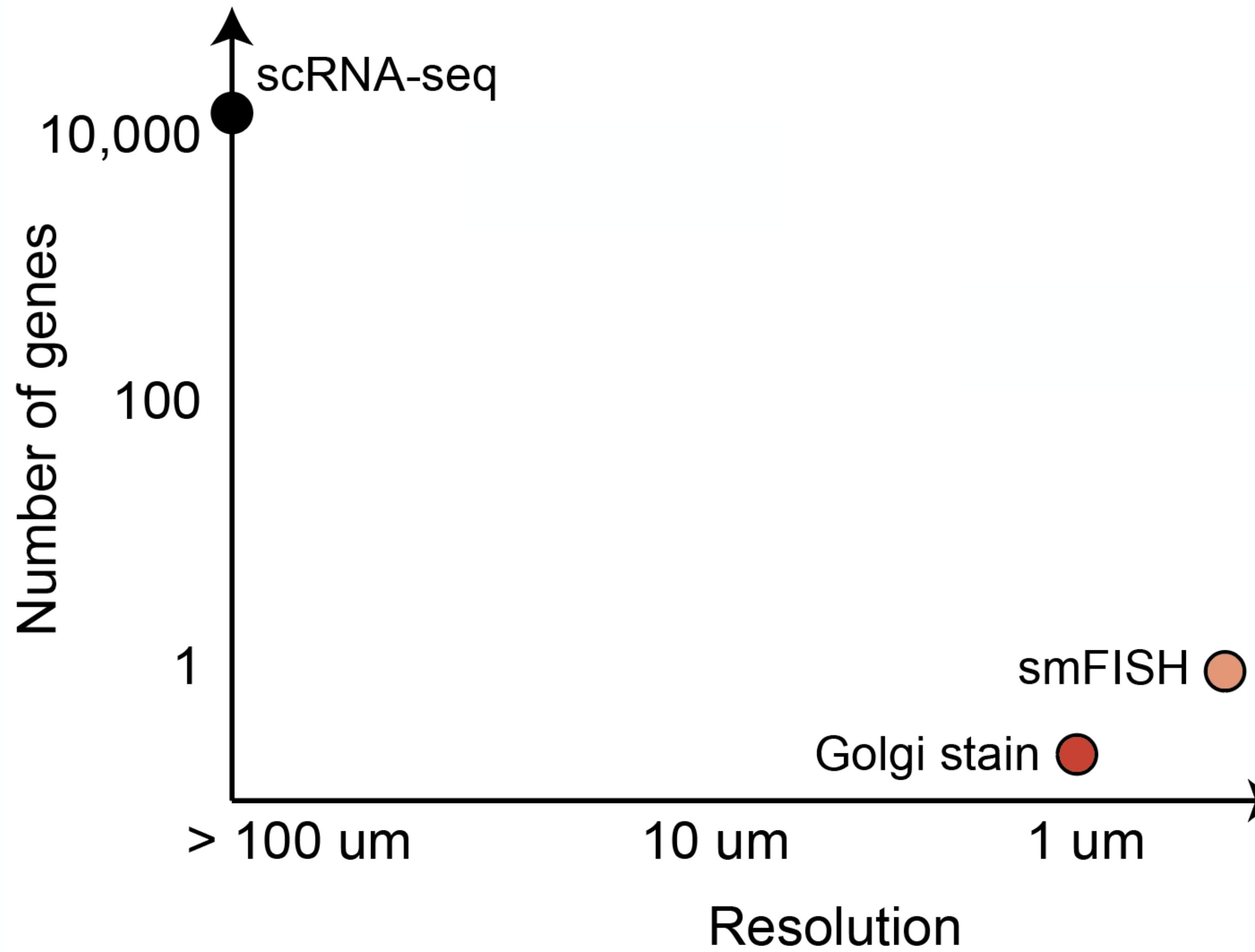
Michael Nunn, 2020

Two major branches of technologies: sequencing-based vs imaging-based assays

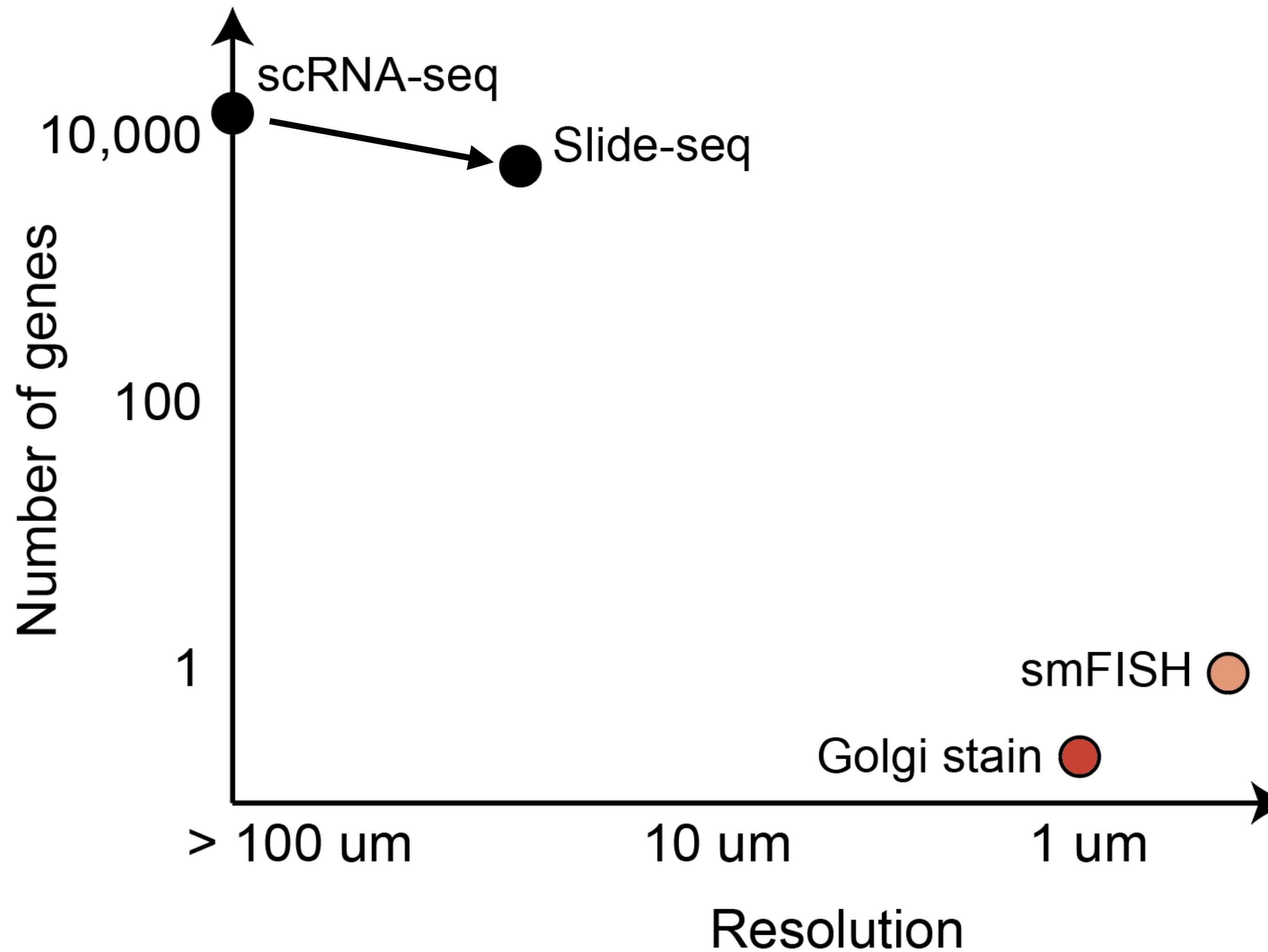
- Imaging-based: MERFISH, seqFISH, STARmap, ...
- Sequencing-based: Slide-seq, 10X Visium, ...



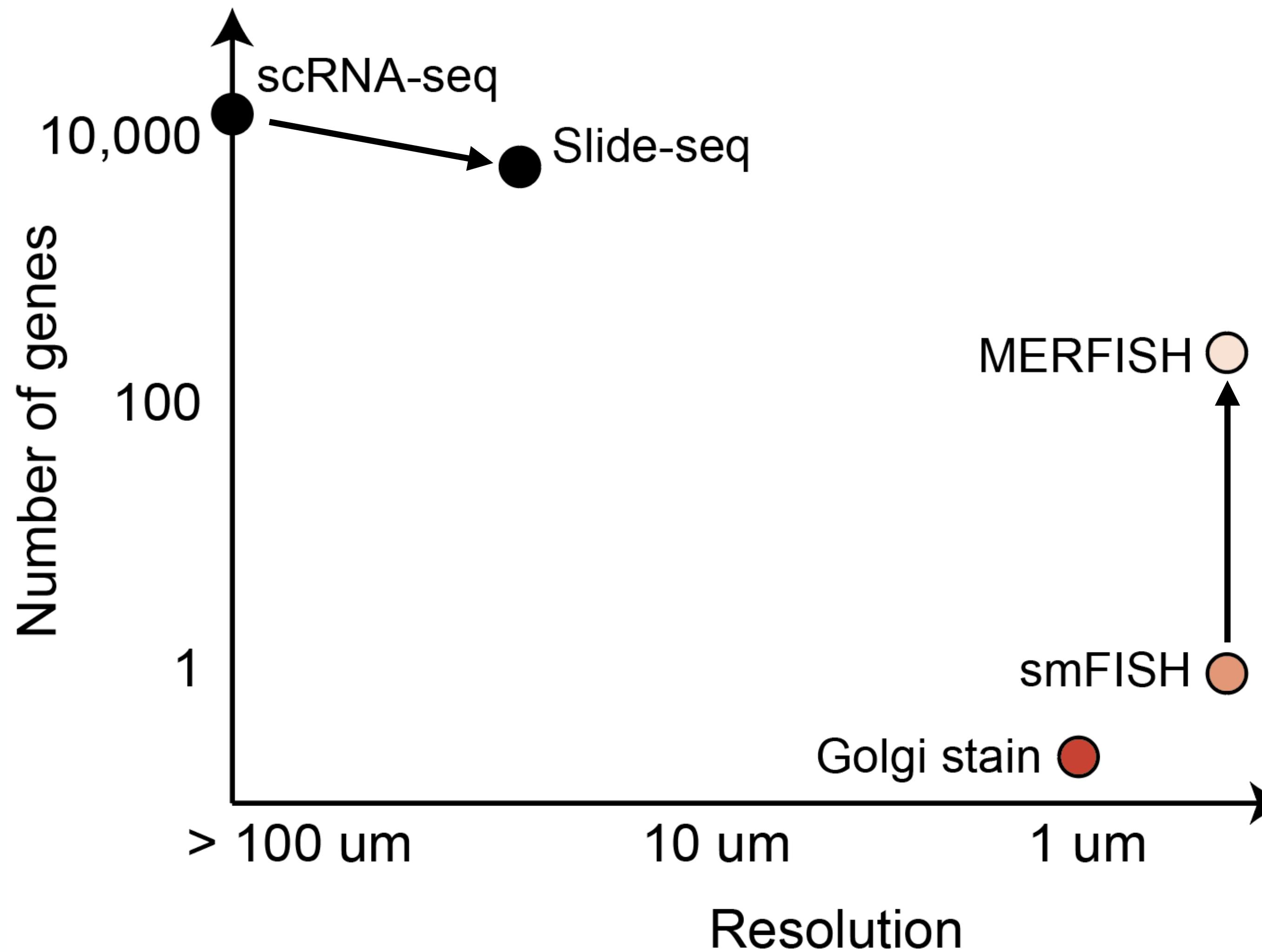
sequencing-based vs imaging-based assays



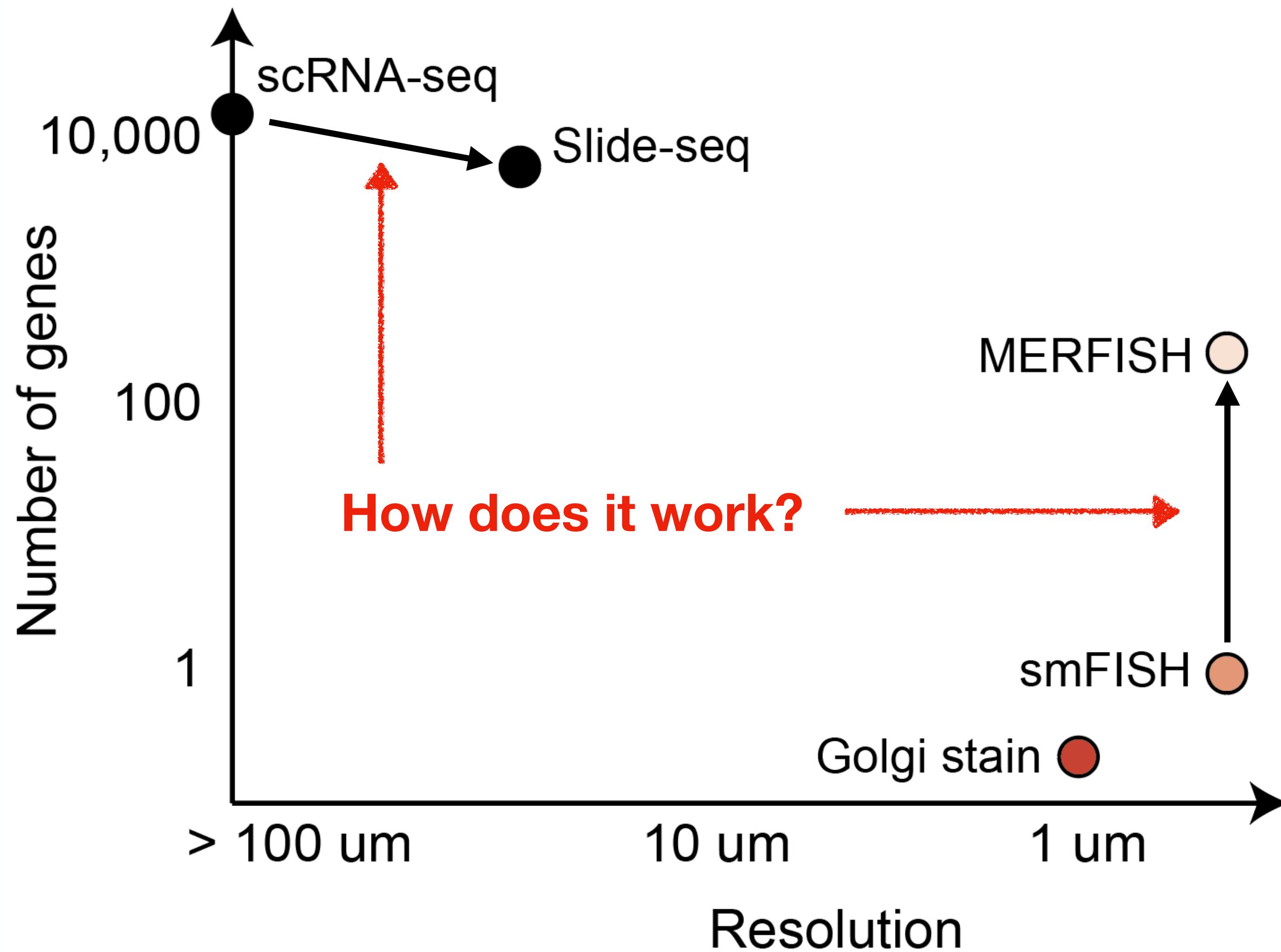
sequencing-based vs imaging-based assays



sequencing-based vs imaging-based assays



sequencing-based vs imaging-based assays



- Tradeoffs:
 - Multiplexing
 - Resolution
 - Throughput
 - Sensitivity

It's break time!!!

FIELDS ARRANGED BY PURITY

— MORE PURE —→

SOCIOLOGY IS
JUST APPLIED
PSYCHOLOGY

PSYCHOLOGY IS
JUST APPLIED
BIOLOGY.

BIOLOGY IS
JUST APPLIED
CHEMISTRY

WHICH IS JUST
APPLIED PHYSICS.
IT'S NICE TO
BE ON TOP.

OH, HEY, I DIDN'T
SEE YOU GUYS ALL
THE WAY OVER THERE.



SOCIOLOGISTS



PSYCHOLOGISTS



BIOLOGISTS



CHEMISTS



PHYSICISTS



MATHEMATICIANS

In studying life, you keep diving from higher levels to lower ones until somewhere along the way life fades out, leaving you empty-handed.

— Albert Szent-Györgyi (taken from the biography of Katalin Kariko)

More Is Different

Broken symmetry and the nature of the hierarchical structure of science.

P. W. Anderson

The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted

planation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid state physics, plasma physics, and perhaps

less relevance they seem to have to the very real problems of the rest of science, much less to those of society.

The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other. That is, it seems to me that one may array the sciences roughly linearly in a hierarchy, according to the idea: The elementary entities of science X obey the laws of science Y.

**Reading sessions /
Journal club**

Reading session plan

- [10 min] Reading. Take notes on where you are confused about.
- [5 min] Pair and share. Come up with at least one question.
- [? min] Discussion (Q & A) to address your questions.
- [5 min] Mini-quiz. (Not graded)

Reading session 1: Chen et al. 2015 Science

- *Spatially resolved, highly multiplexed RNA profiling in single cells*
- How does MERFISH work? Read the **first page** and try to understand **Figure 1**.
- Some questions to consider:
 - What does MERFISH stands for?
 - How many genes can MERFISH measure?
 - MERFISH is the extension of which technology (mentioned in class)?
 - What allowed MERFISH to measure many genes?

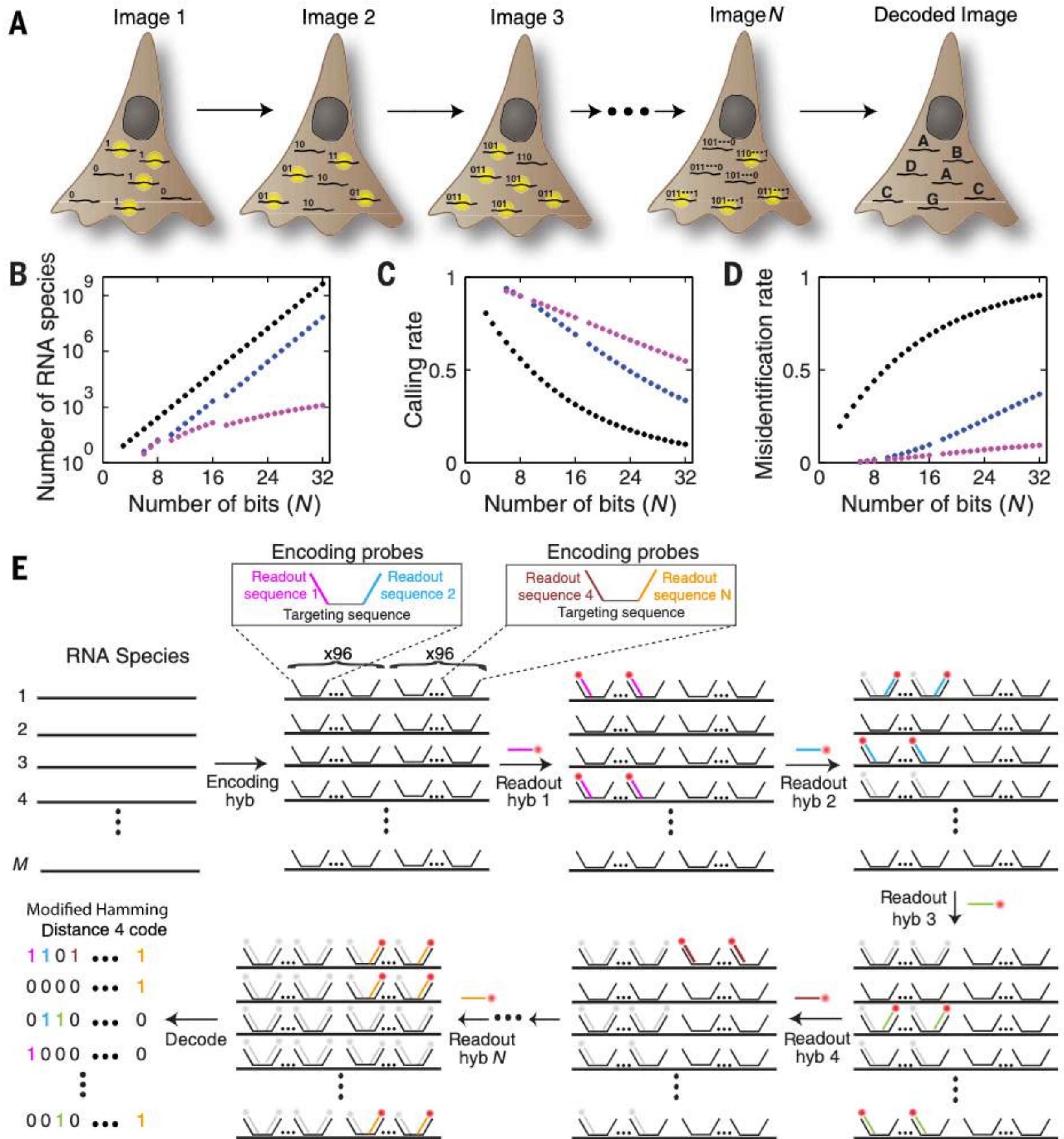


Fig. 1. MERFISH: A highly multiplexed smFISH approach enabled by combinatorial labeling and error-robust encoding. (A) Schematic depiction of the identification of multiple RNA species in N rounds of imaging. Each RNA species is encoded with a N -bit binary word, and during each round of imaging, only the subset of RNAs that should read 1 in the corresponding bit emit signal. (B to D) The number of addressable RNA species (B); the rate at which these RNAs are properly identified—the “calling rate” (C); and the rate at which RNAs are incorrectly identified as a different RNA species—the “misidentification rate” (D); plotted as a function of the number of bits (N) in the binary words encoding RNA. Black indicates a simple binary code that includes all 2^N-1 possible binary words. Blue indicates the HD4 code in which the Hamming distance separating words is 4. Purple indicates a modified HD4 (MHD4) code where the number of 1 bits are kept at four. The calling and misidentification rates are calculated with per-bit error rates of 10% for the 1→0 error and 4% for the 0→1 error. (E) Schematic diagram of the implementation of a MHD4 code for RNA identification. Each RNA species is first labeled with ~192 encoding probes that convert the RNA into a specific combination of readout sequences (Encoding hyb). These encoding probes each contain a central RNA-targeting region flanked by two readout sequences, drawn from a pool of N different sequences, each associated with a specific hybridization round. Encoding probes for a specific RNA species contain a particular combination of four of the N readout sequences, which correspond to the four hybridization rounds in which this RNA should read 1. N subsequent rounds of hybridization with the fluorescent readout probes are used to probe the readout sequences (hyb 1, hyb 2, ..., hyb N). The bound probes are inactivated by photobleaching between successive rounds of hybridization. For clarity, only one possible pairing of the readout sequences is depicted for the encoding probes; however, all possible pairs of the four readout sequences are used at the same frequency and distributed randomly along each cellular RNA in the actual experiments.

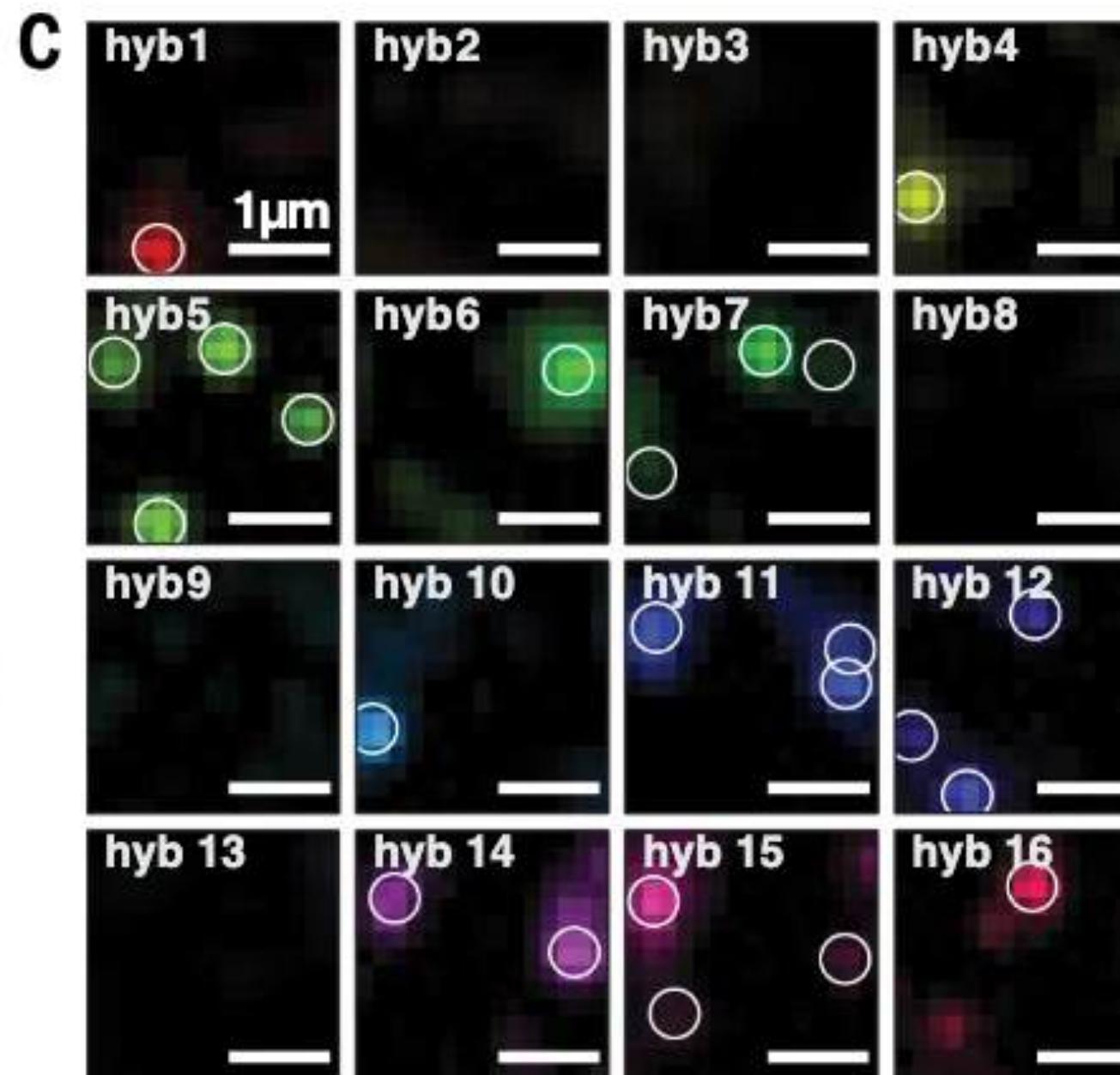
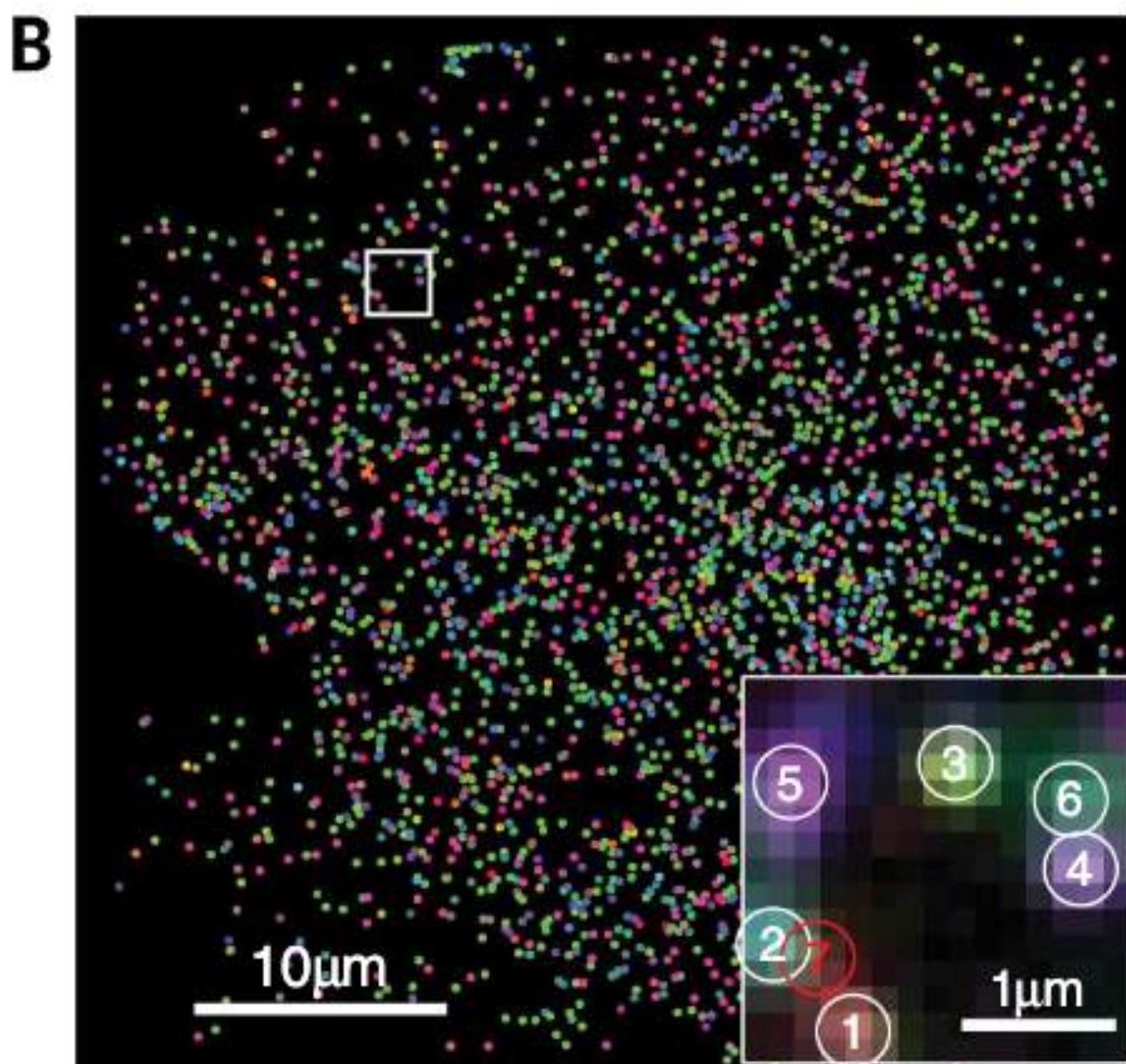
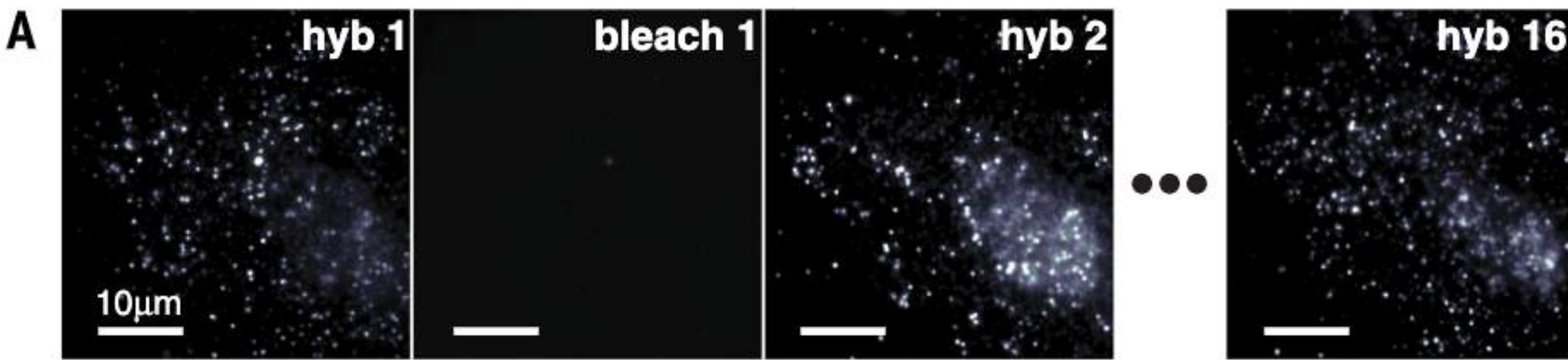
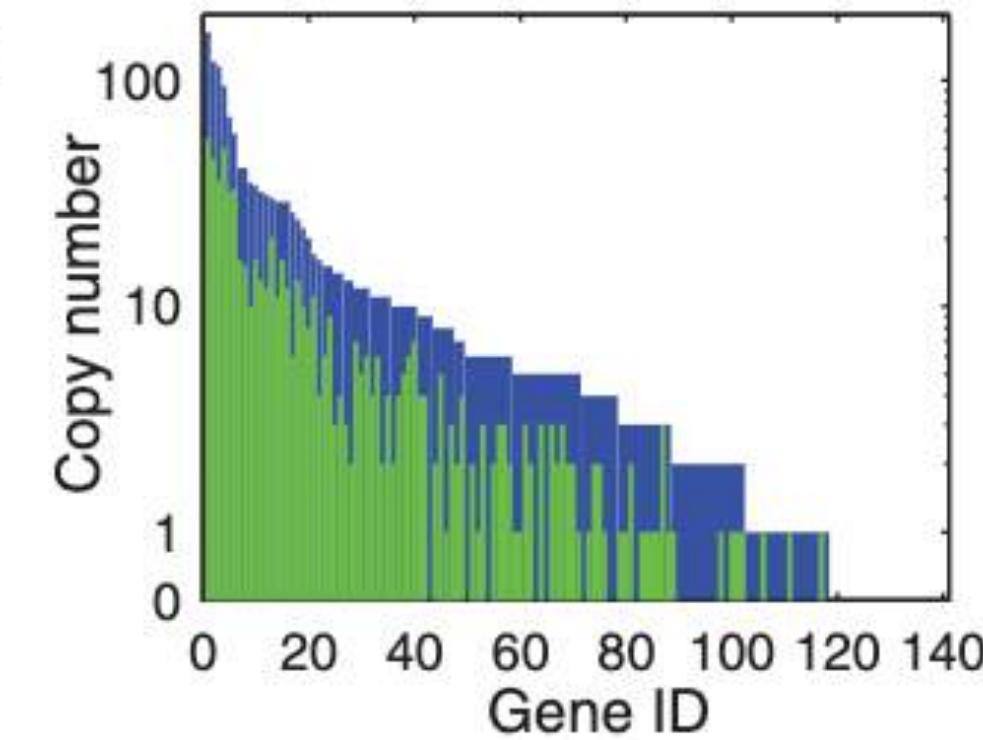


Fig. 2. Simultaneous measurement of 140 RNA species in single cells by use of MERFISH with a 16-bit MHD4 code. (A) Images of RNA molecules in an IMR90 cell after each hybridization round (hyb 1 to hyb 16). The images after photobleaching (for example, bleach 1) demonstrate efficient removal of fluorescent signals between hybridizations. (B) The localizations of all detected single molecules in this cell colored according to their measured binary words. (Inset) The composite, false-colored fluorescent image of the 16 hybridization rounds for the boxed subregion with numbered circles indicating potential RNA molecules. A red circle indicates an unidentifiable molecule, the binary word of which does not match any of the 16-bit MHD4 code words even after error correction. (C) Fluorescent images from each round of hybridization for the boxed subregion in (B), with circles indicating potential RNA molecules. (D) Corresponding words for the spots identified in (C). Red crosses represent the corrected bits. (E) The RNA

D

Hybridization round

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
2	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0
3	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1
4	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
5	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0
6	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0



Mini quiz

- Go to kahoot.it

Reading session 2: Rodrigues et al. 2019 Science

- *Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution*
- How does Slide-seq work? Read the **first 2 paragraphs** and try to understand **Figure 1**.
- Questions to keep in mind:
 - Slide-seq is the natural extension of which technology?
 - What's the spatial resolution of Slide-seq?
 - How many genes can Slide-seq measure?
 - What allowed Slide-seq to retain spatial information of RNAs using sequencing?

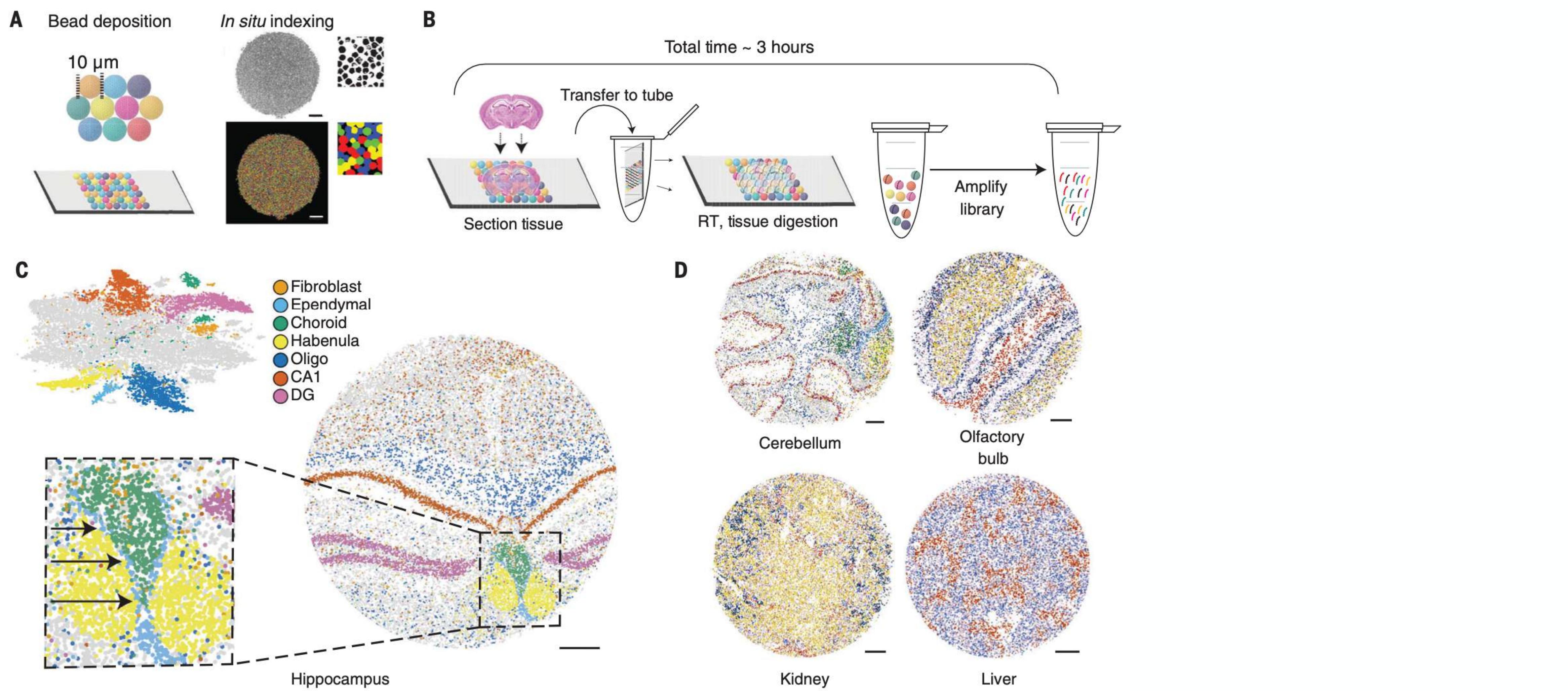


Fig. 1. High-resolution RNA capture from tissue by Slide-seq.

(A) (Left) Schematic of array generation. A monolayer of randomly deposited, DNA barcoded beads (a “puck”) is spatially indexed by SOLiD sequencing. (Top right) Representative puck with sequenced barcodes shown in black. (Bottom right) Composite image of the same puck colored by the base calls for a single base of SOLiD sequencing. **(B)** Schematic of the sample preparation procedure. RT, reverse transcription. **(C)** (Top left) t-distributed

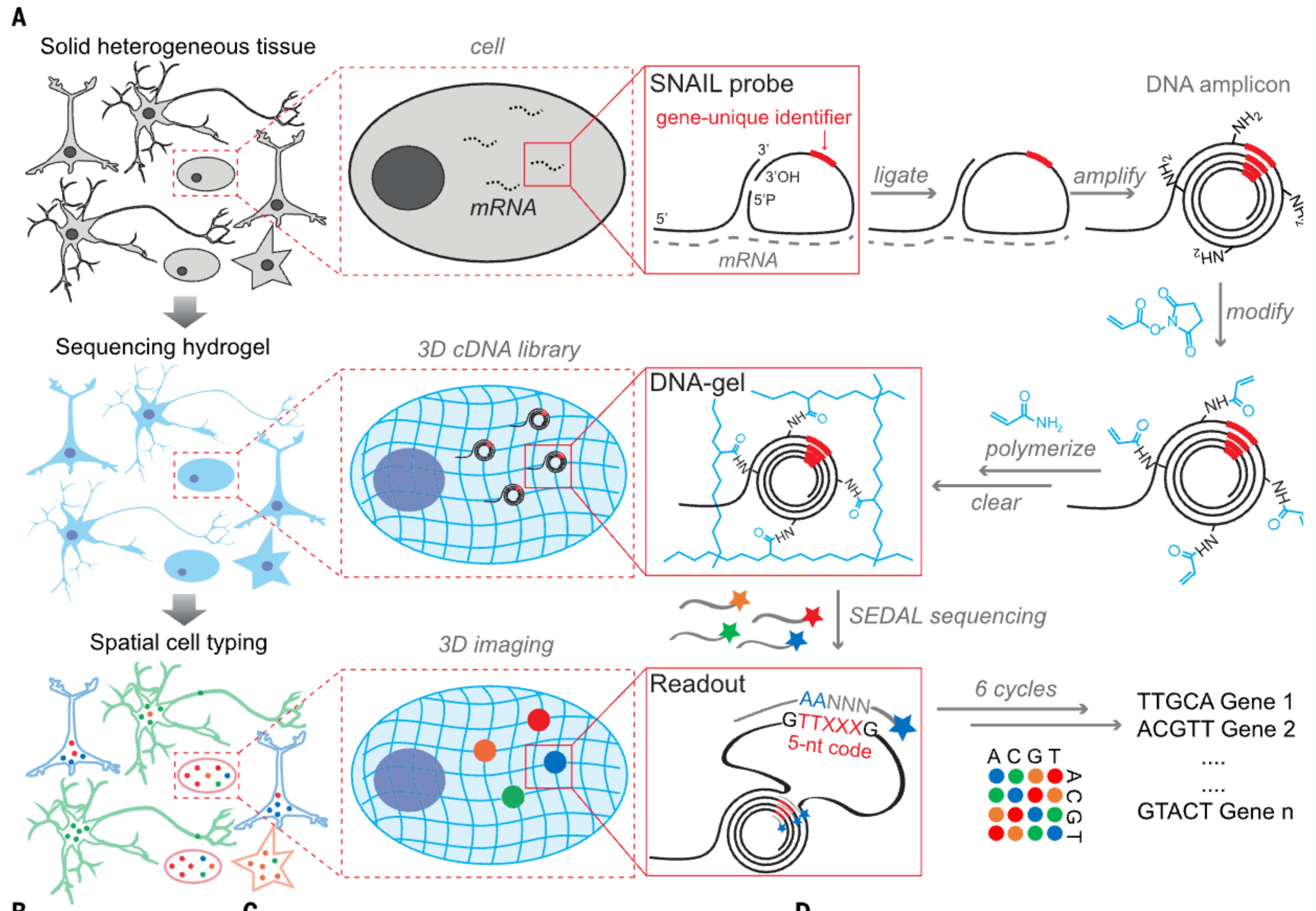
stochastic neighbor embedding (tSNE) representation of Slide-seq beads from a coronal mouse hippocampus slice with colors indicating clusters. GD, dentate gyrus. (Right) The spatial position of each bead is shown, colored by the cluster assignments shown in the tSNE. (Bottom left) Inset indicating the position of a single-cell-thick ependymal cell layer (arrows). **(D)** As in (C), but for the indicated tissue type (see fig. S2 for clustering and cluster identities). All scale bars: 500 μm .

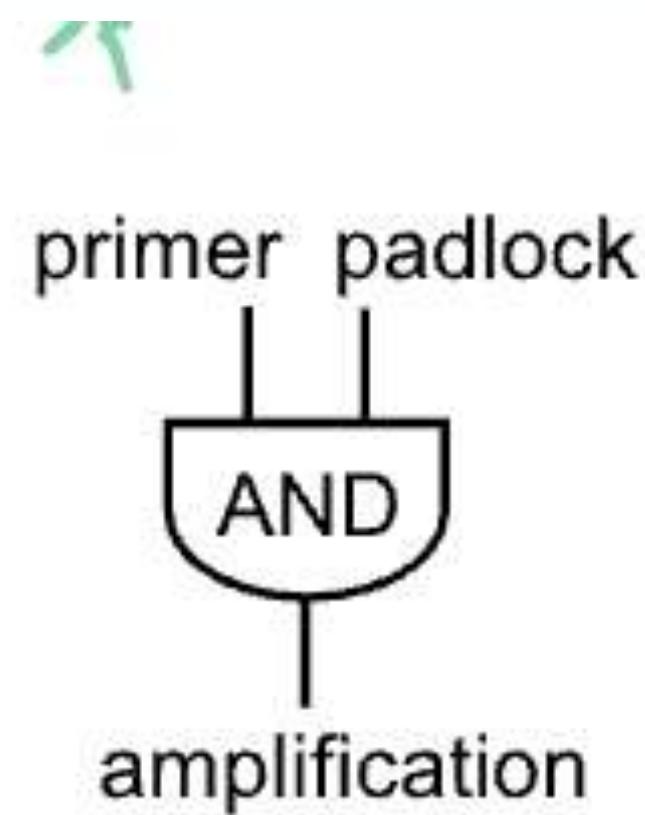
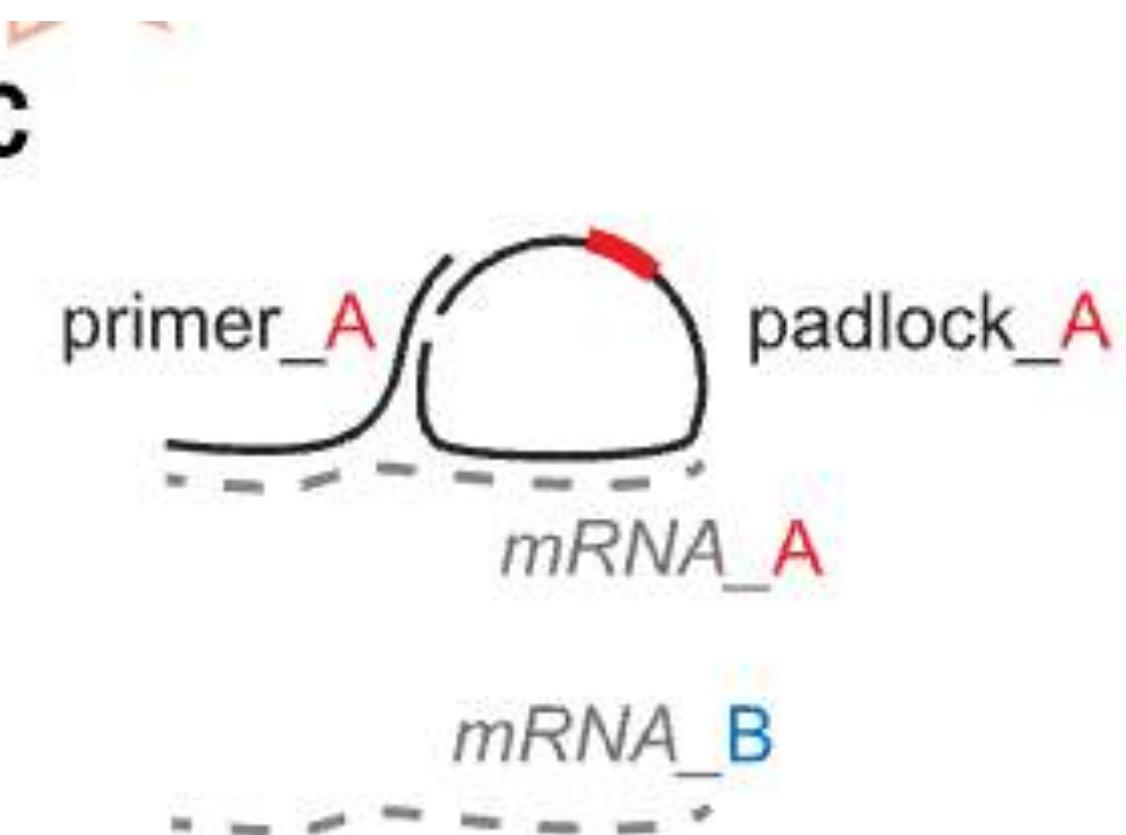
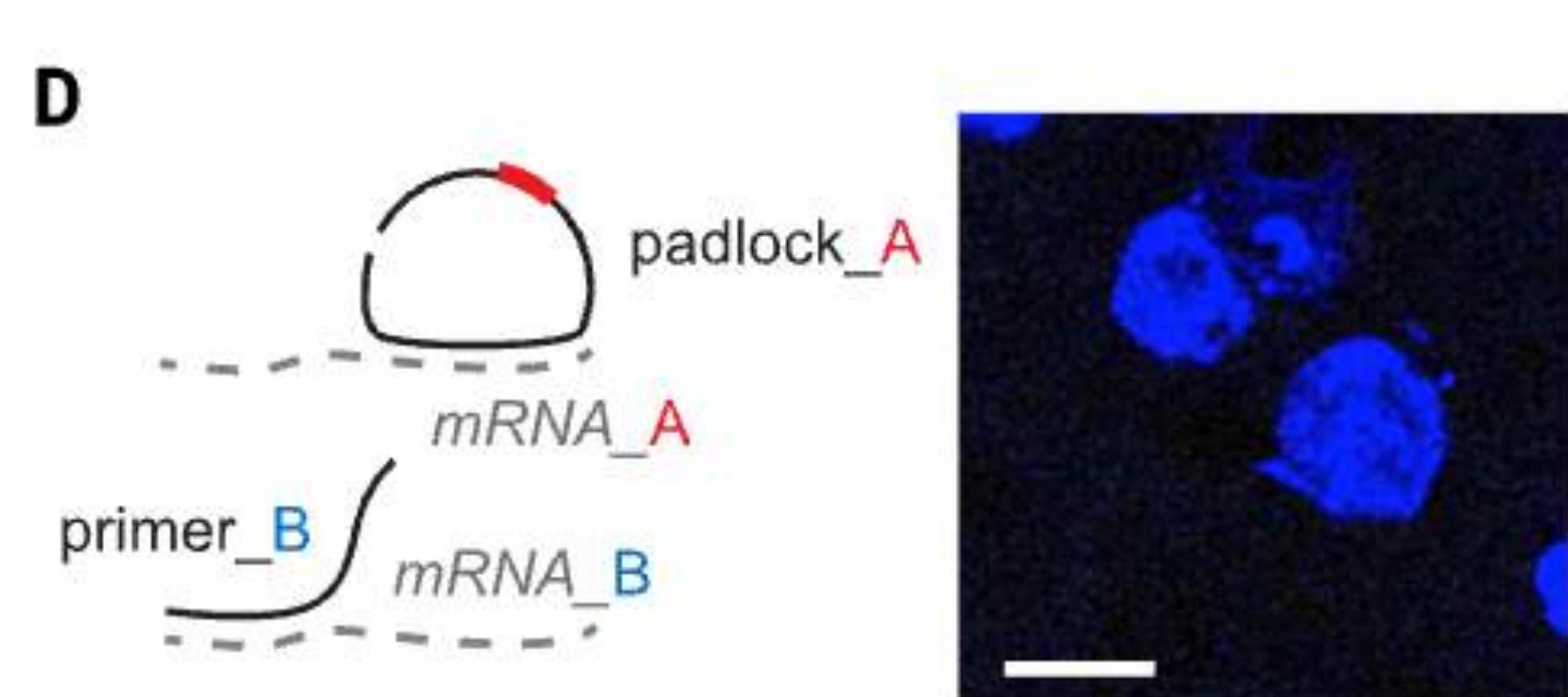
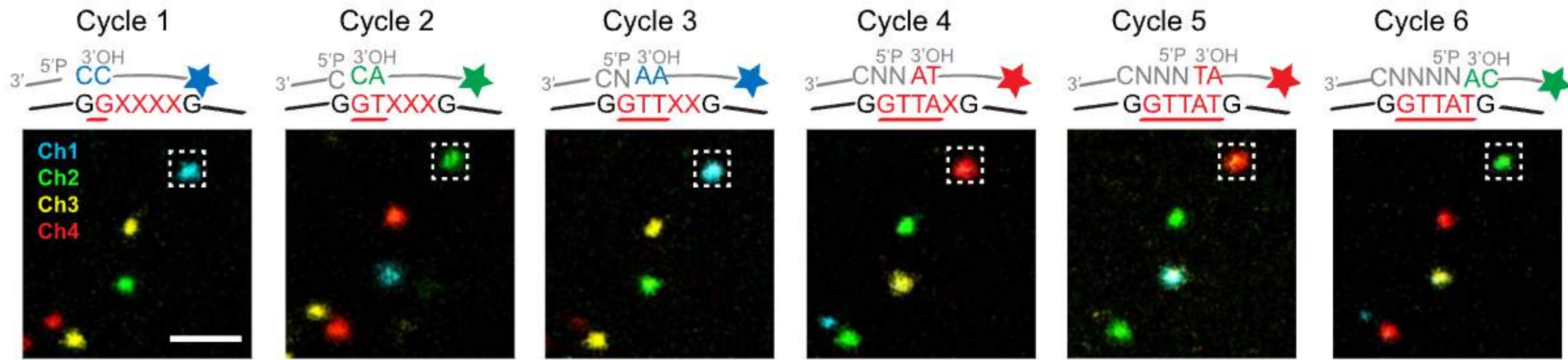
Mini quiz

- Go to kahoot.it

Reading section 3: Wang et al. 2018 Science

- *Three-dimensional intact-tissue sequencing of single-cell transcriptional states*
- How does STARmap work? Read and try to understand **Figure 1**.
- Questions to keep in mind:
 - Is STARmap sequencing-based or imaging-based spatial transcriptomics?
 - What allowed STARmap to achieve high *specificity* of gene detection?
 - What allowed STARmap to achieve high *sensitivity* of gene detection?



B**C****D****E**

Mini quiz

- Go to kahoot.it

UCLA Collaboratory workshop (W31)

Spatial Transcriptomics

Oct 29, 2025 (Day 2)
Fangming Xie

Opinion

The cell as a bag of RNA

Transcriptomes

	Gene 1	Gene 2	Gene 3
Cell 1	10	0	100
Cell 2	3	100	3
Cell 3	20	5	10,000
...

Opinion

The cell as a bag of RNA **in a spatial context**

Transcriptomes

	Gene 1	Gene 2	Gene 3
Cell 1	10	0	100
Cell 2	3	100	3
Cell 3	20	5	10,000
...

Spatial location

Location (x)	Location (y)
5.2	2.1
3.4	18.2
2.0	9.0
...	...

Opinion

The cell as a bag of RNA **in a spatial context**

Transcriptomes

	Gene 1	Gene 2	Gene 3
Cell 1	10	0	100
Cell 2	3	100	3
Cell 3	20	5	10,000
...

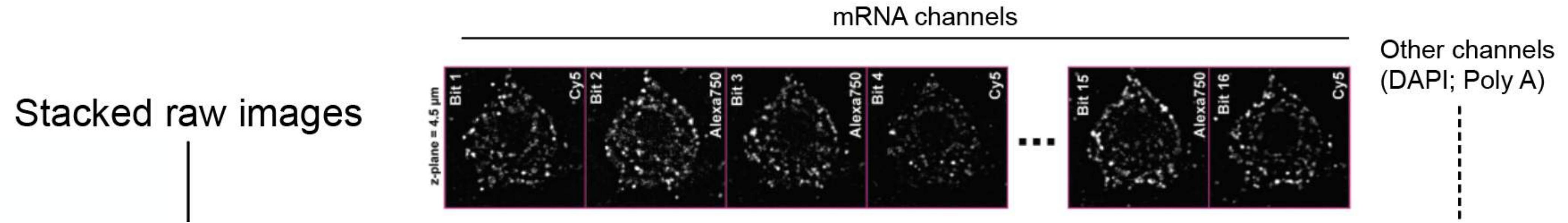
Spatial location

Location (x)	Location (y)
5.2	2.1
3.4	18.2
2.0	9.0
...	...

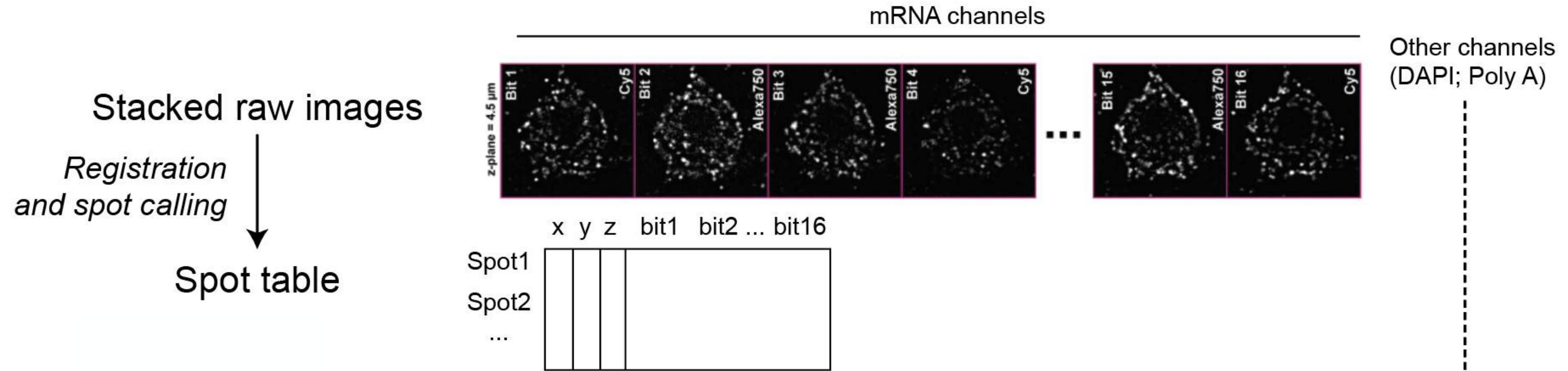
How to get this table?

What to do with it?

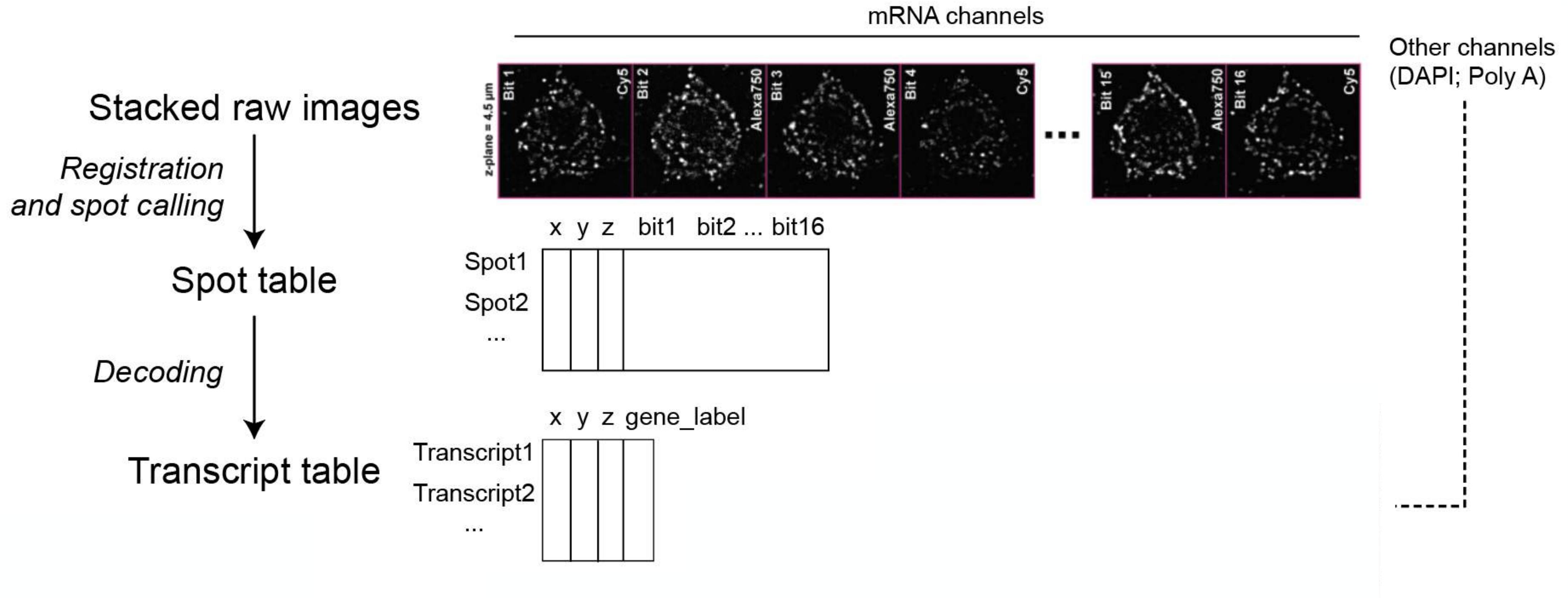
Data-processing workflow – image-based (MERFISH)



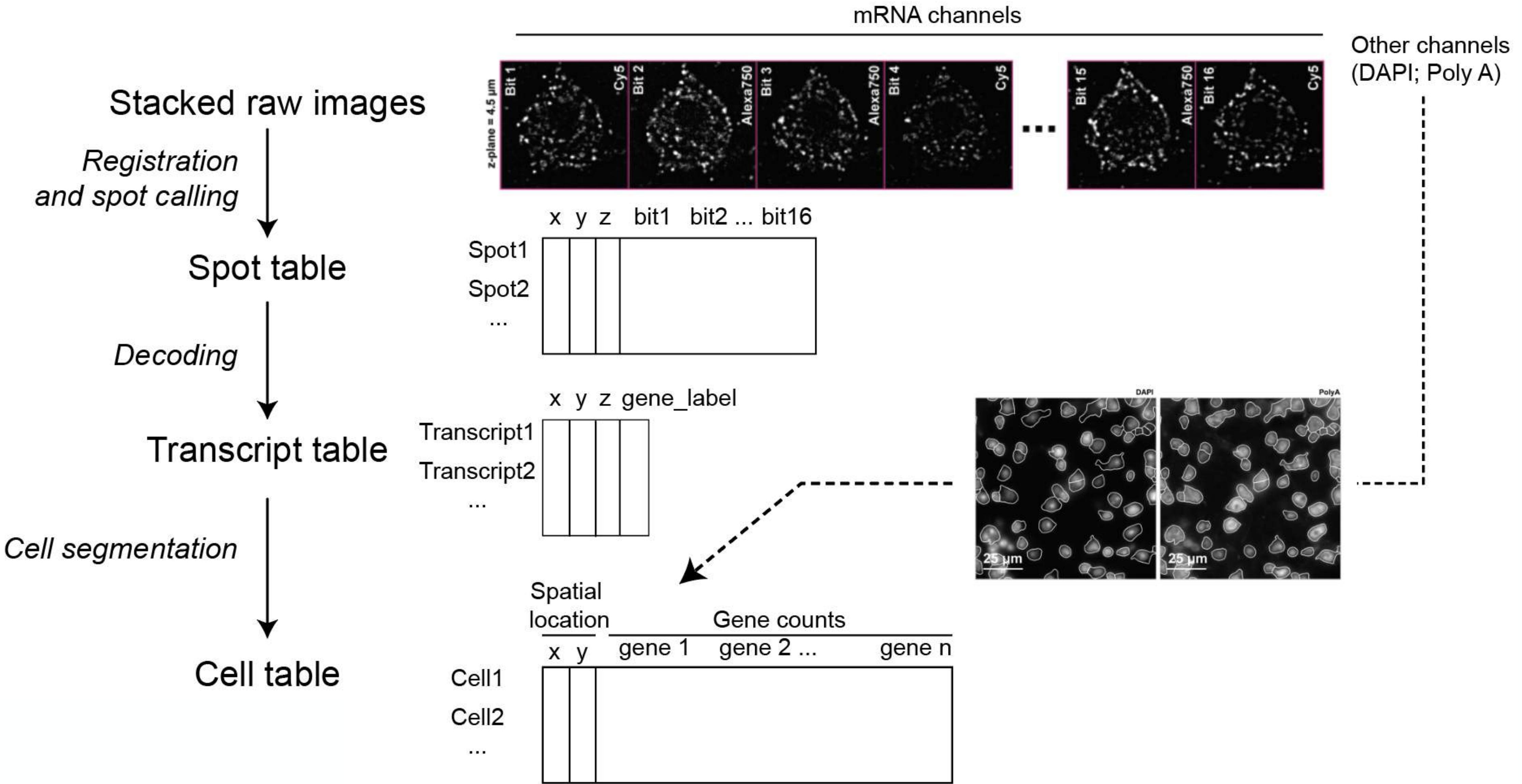
Data-processing workflow – image-based (MERFISH)



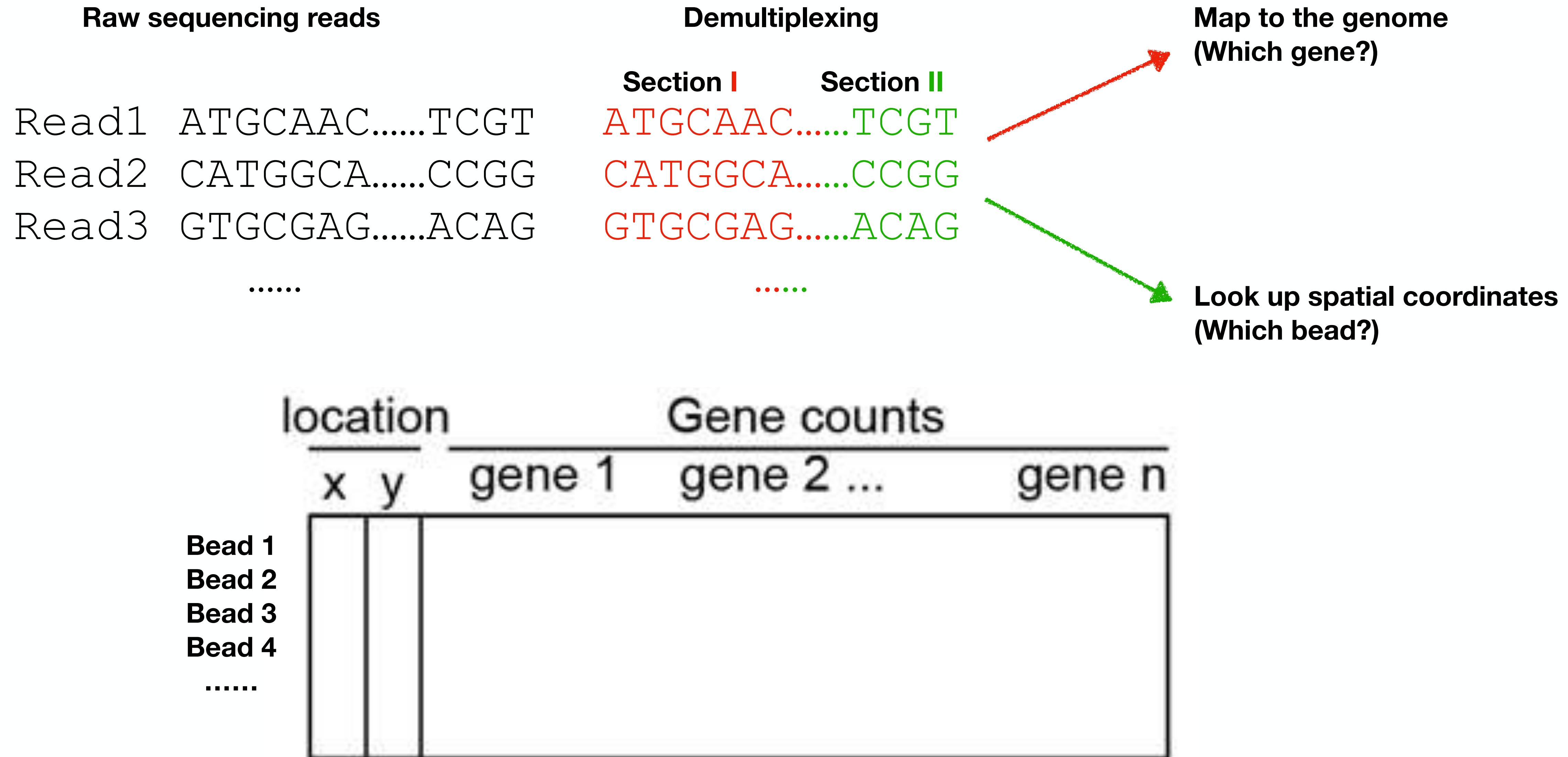
Data-processing workflow – image-based (MERFISH)



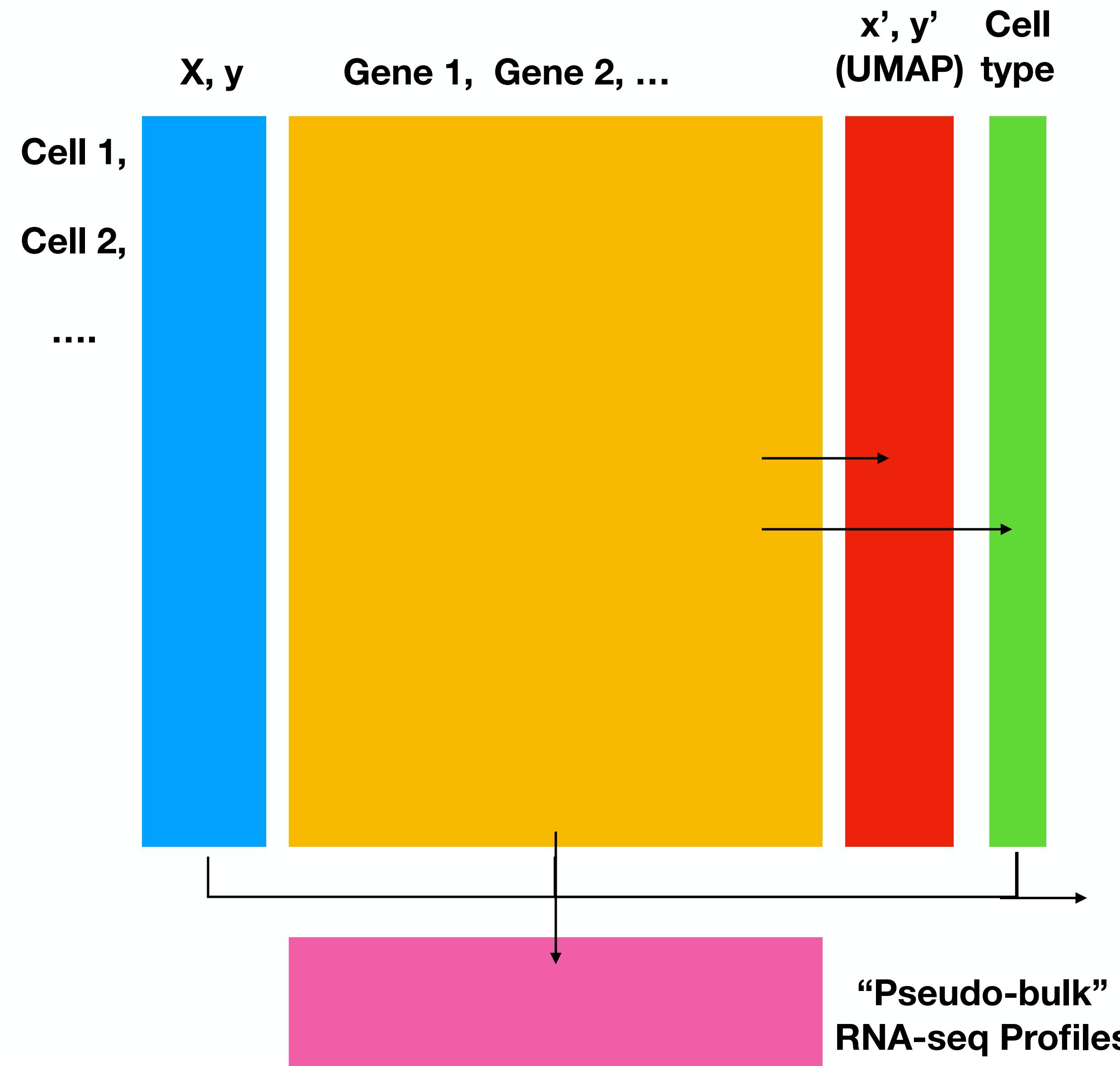
Data-processing workflow – image-based (MERFISH)



Data-processing workflow – sequencing-based (Slide-seq)



Analysis overview



- Where (and in which cells) is gene X expressed? [Visualization]
- Can we capture the main components of transcriptomic diversity between cells? [Dimensionality reduction]
- Are there different cell types? [Clustering]
- Which cell types interact with which others? [spatial enrichment of cell types; ligand-receptor pairings...]

The coding bootcamp

- For exercise 1 to 5:
 - Talk and demo — me; on simulated data
 - Coding exercise — you; on real data
 - Show — post your results (screenshots) in the shared Google Slides
- Pause anytime for questions

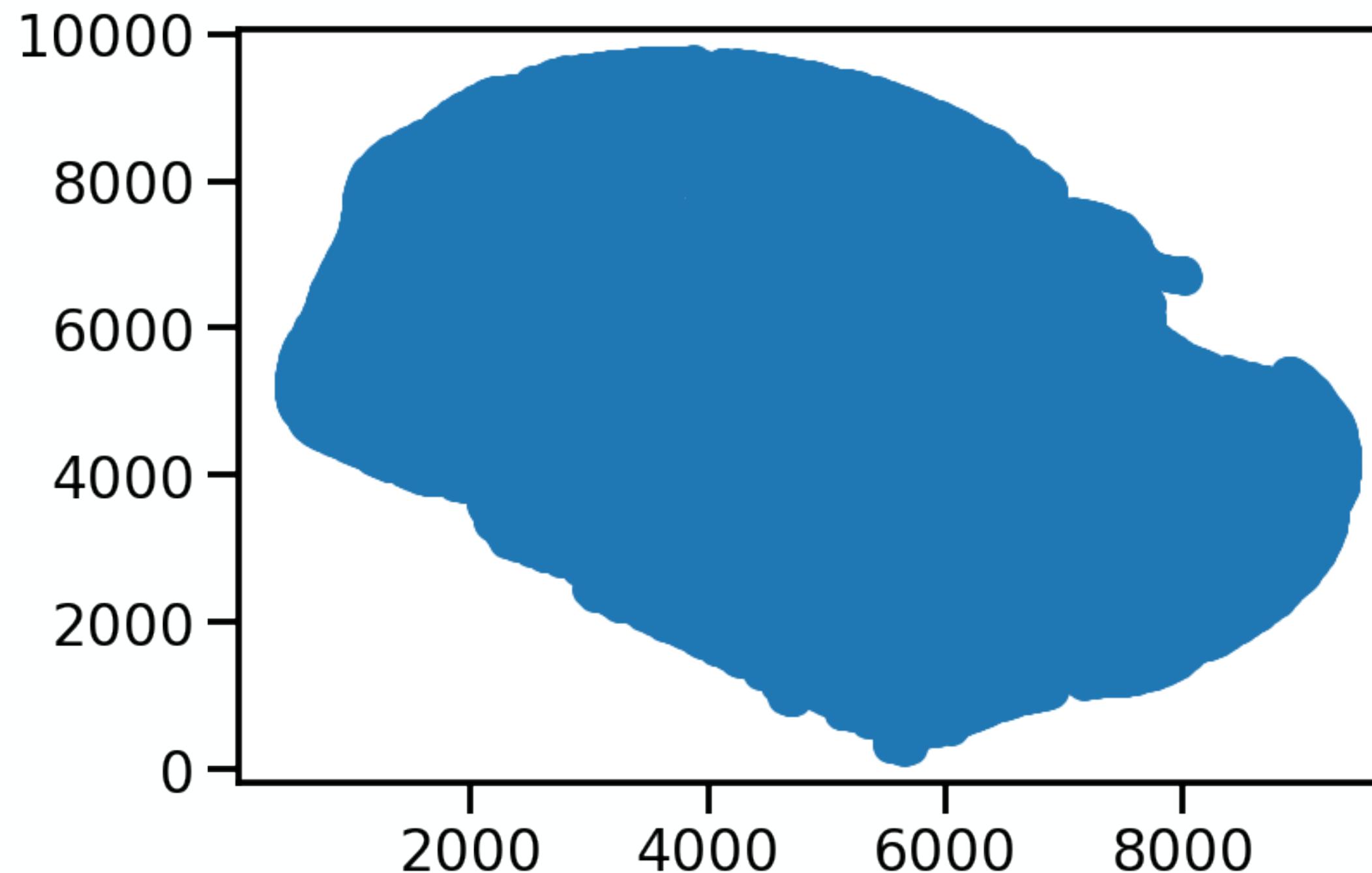
Coding exercise 1: visualize the spatial distribution of gene expression

	x, y	Gene 1, Gene 2, ...
Cell 1,		
Cell 2,		
....		

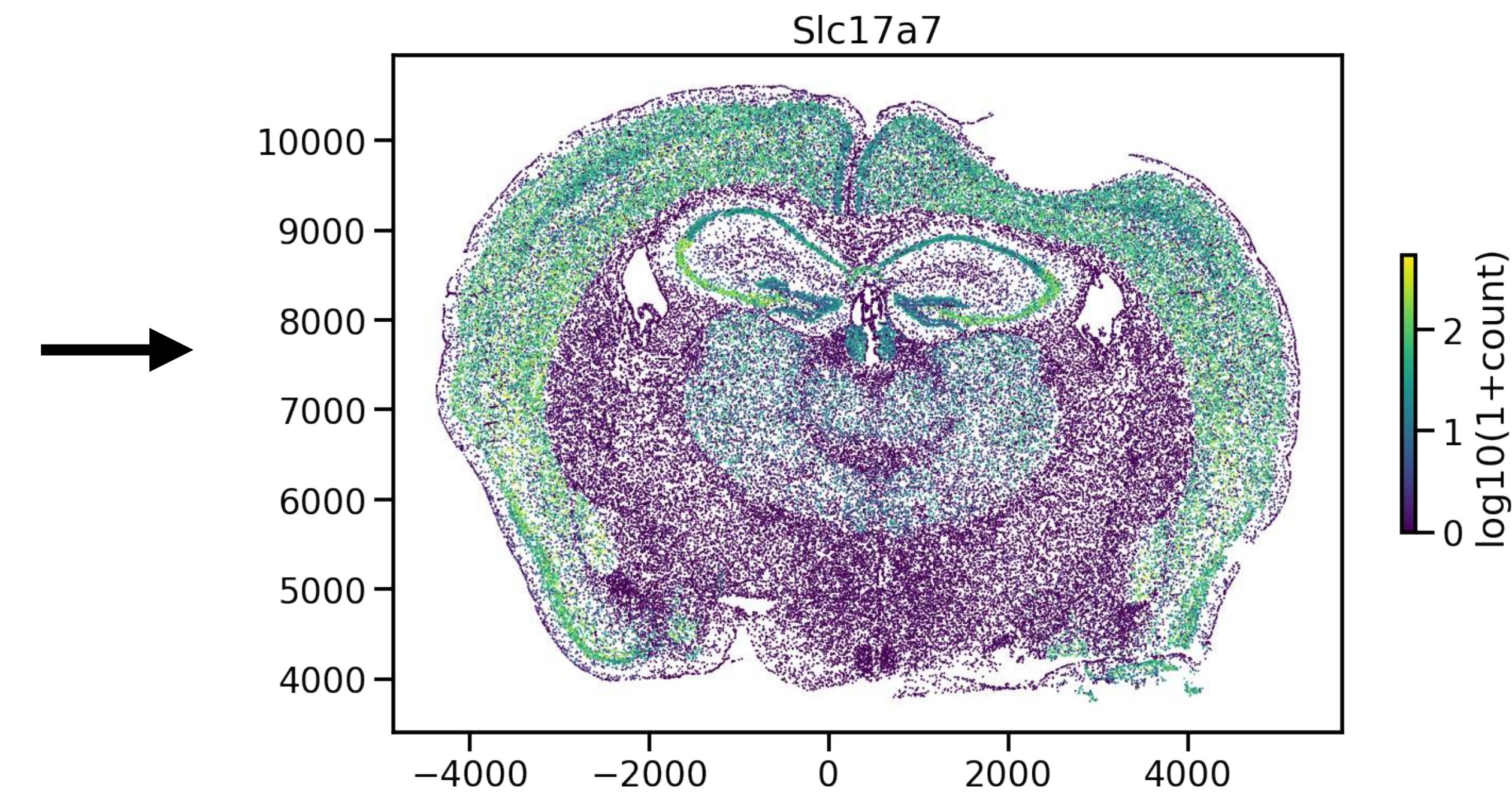
- Practicing data visualization on a public MERFISH mouse brain dataset from Vizgen.
- If you are experienced in coding, you can skip this and choose to work on [bonus exercise 1](#).

Coding exercise 1: Making a scatter plot from scratch

- Start from here: `plt.scatter(x,y)`

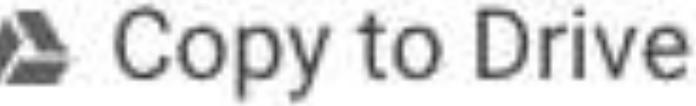


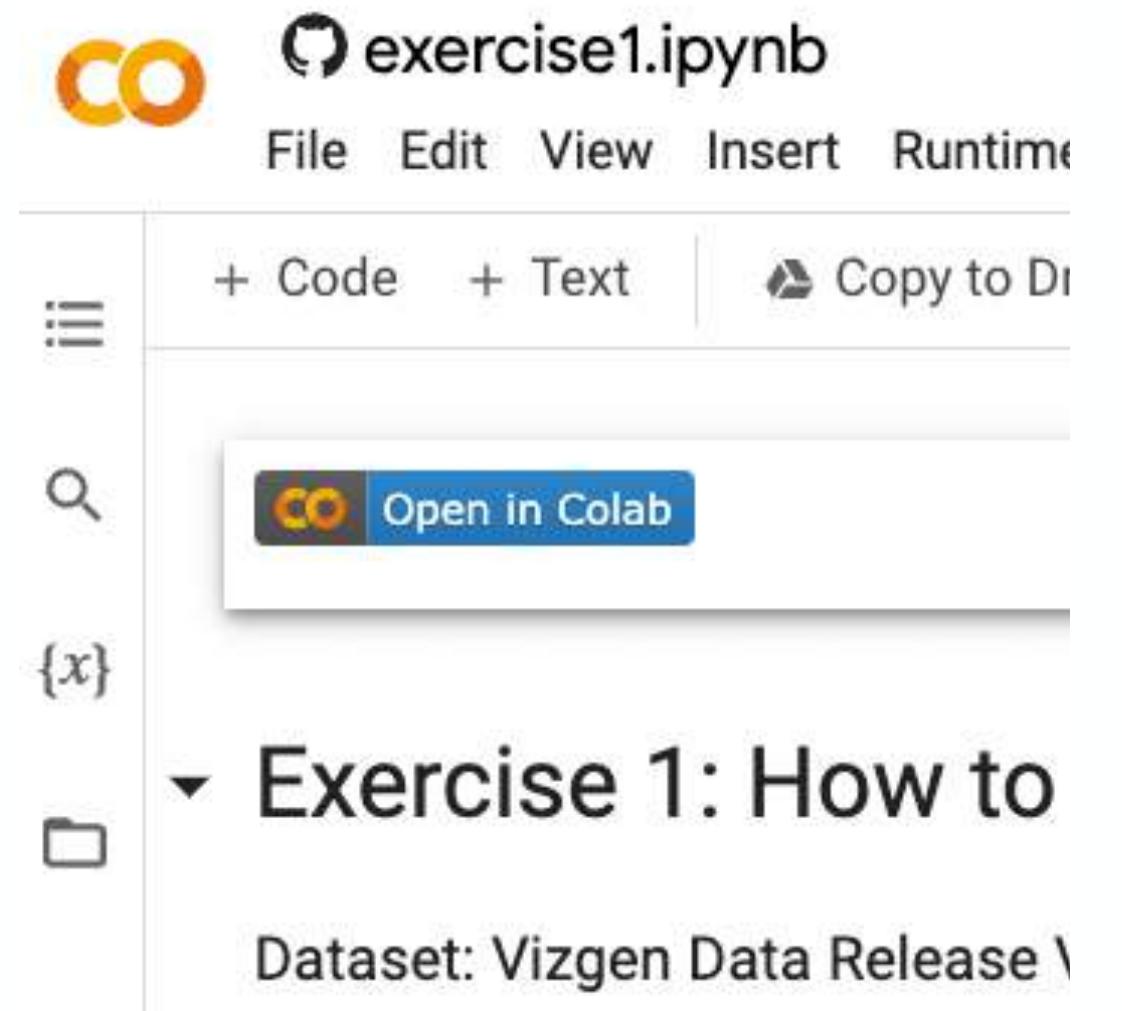
- End here:



Setting up the coding environment

- Go to the GitHub page:
 - <https://github.com/FangmingXie/collab-workshop-st>
- Open workbook/exercise1.ipynb and click the icon:

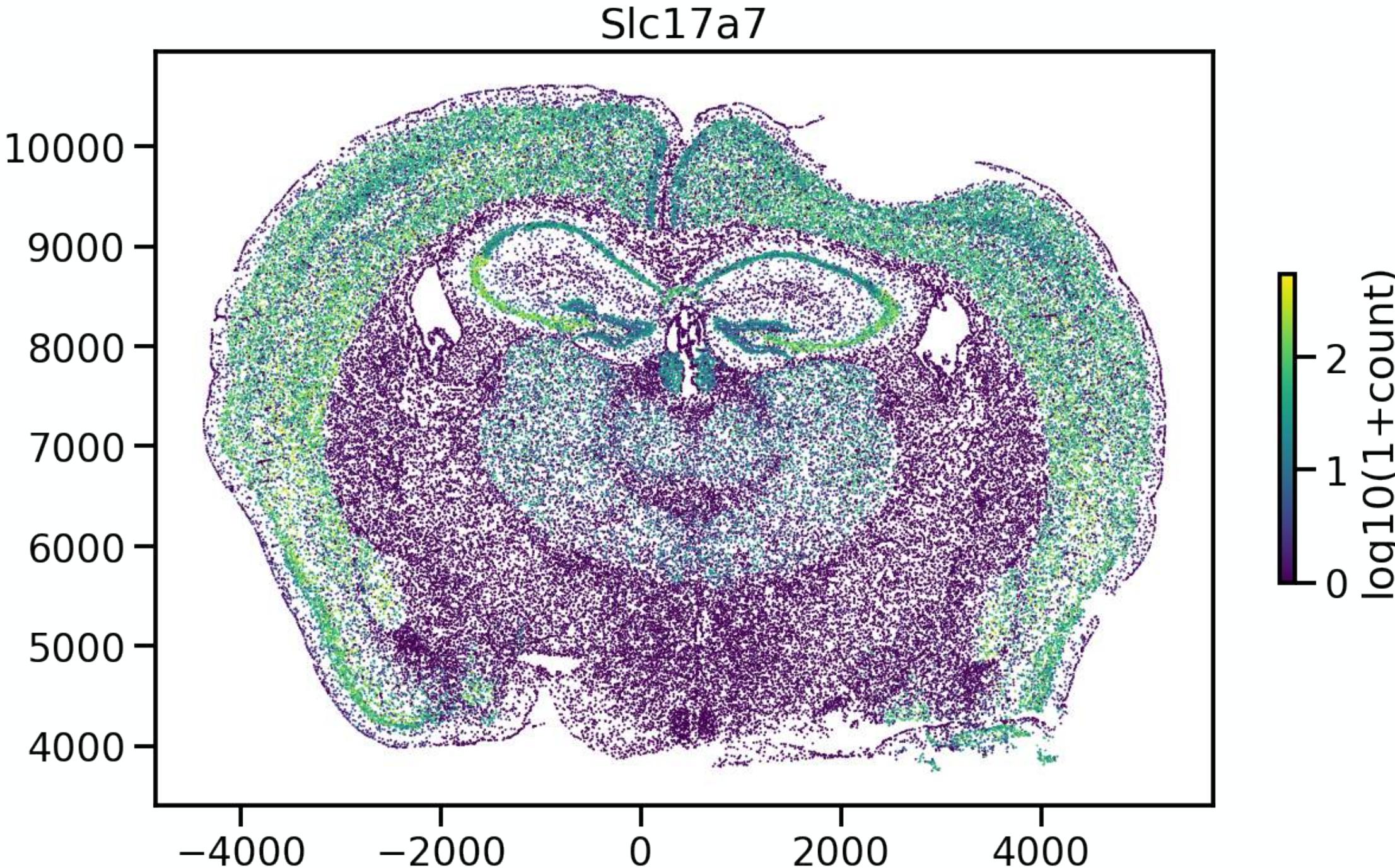
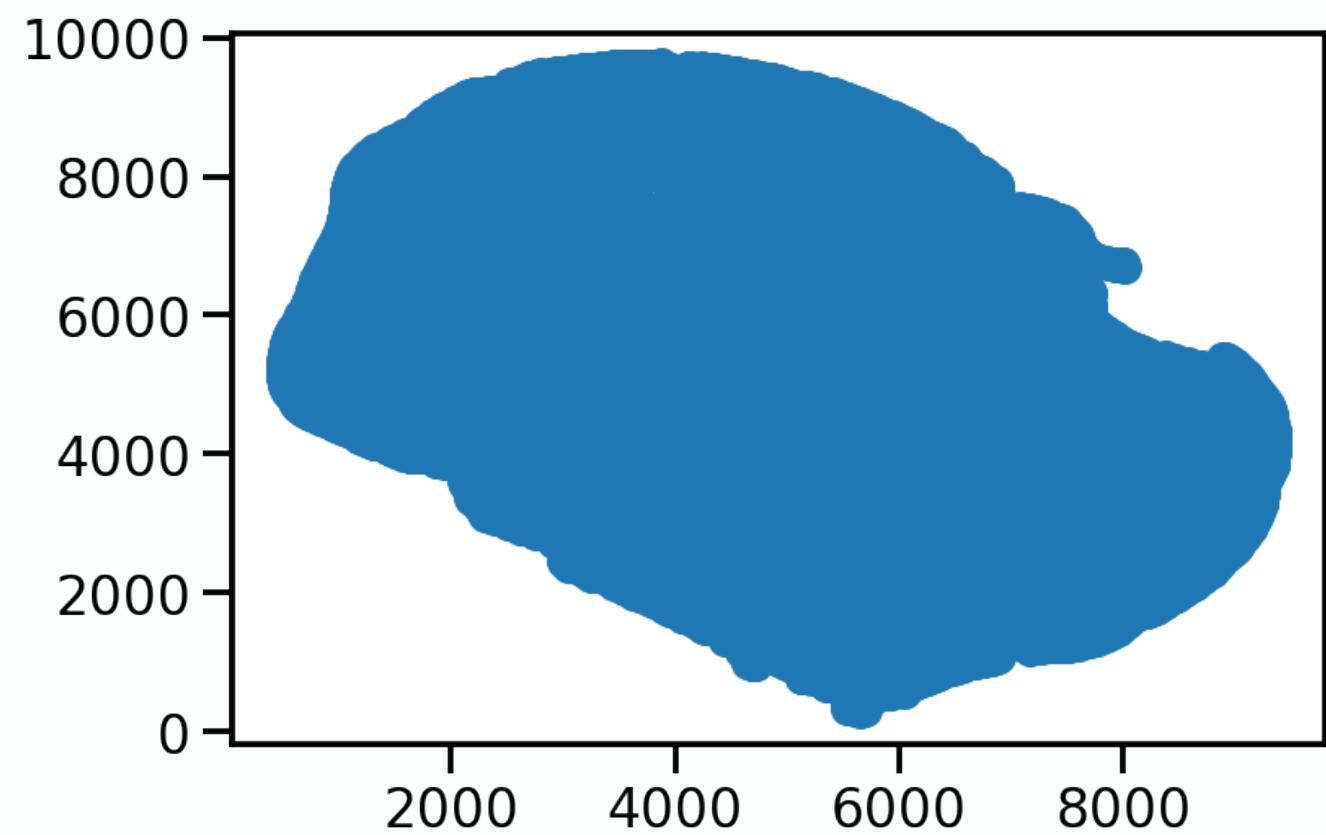
- To save your progress, copy this notebook by clicking:

- Share your results (screenshots) here.



The devil is in the details

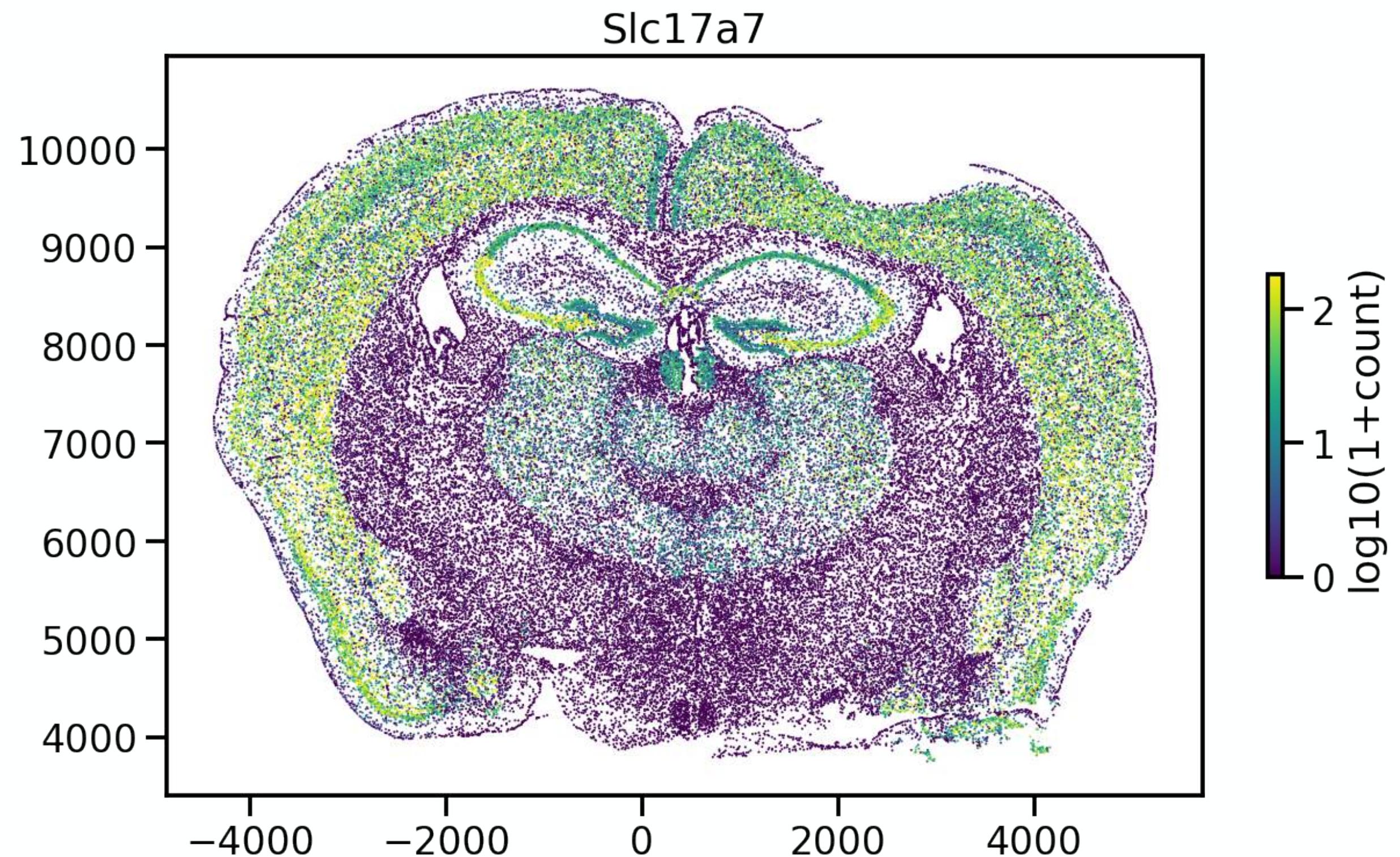
Problems we will encounter:

- How to get cell coordinates and gene expression from data?
- How to make a scatter plot colored by gene expression?
- How to add a color bar?
- Why the plot doesn't look good?
- How to deal with crowding?
- How to deal with outliers?
- How to preserve aspect ratio?
- How to rotate spatial coordinates?
- How to fine tune elements in figures so they look good for readers?



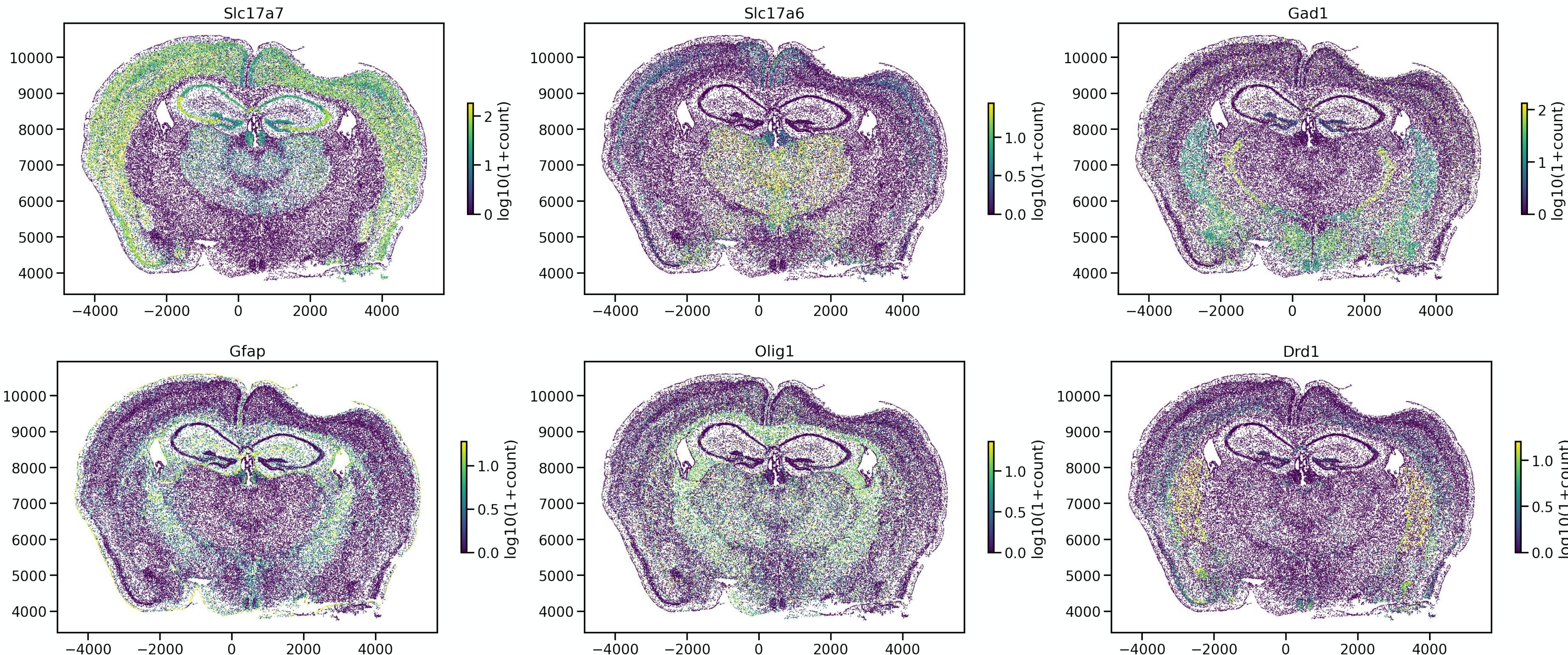
Review: Visualizing gene expression in-situ

- x, y, gene expression
 - Be mindful of data-point crowding
 - Caps at ~95%
 - Normalization
-
- It may take you long to get it once, but once you get it, it is *effortless* to reproduce and reuse it.

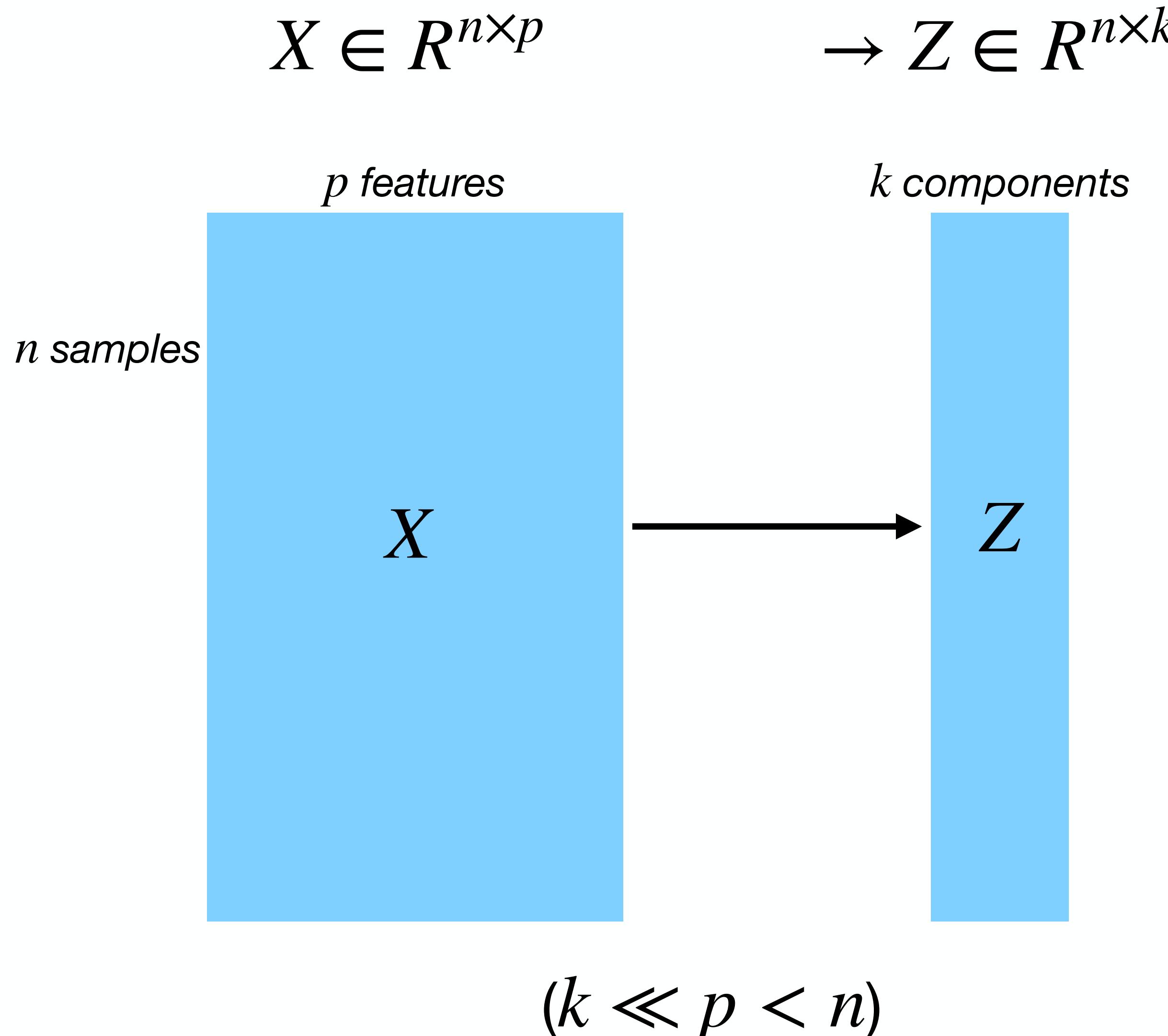


Gallery – what works for one should work for all

- Yet, applying your code to a broader range of cases may reveal new problems.



The magic of dimensionality reduction



scRNA-seq:

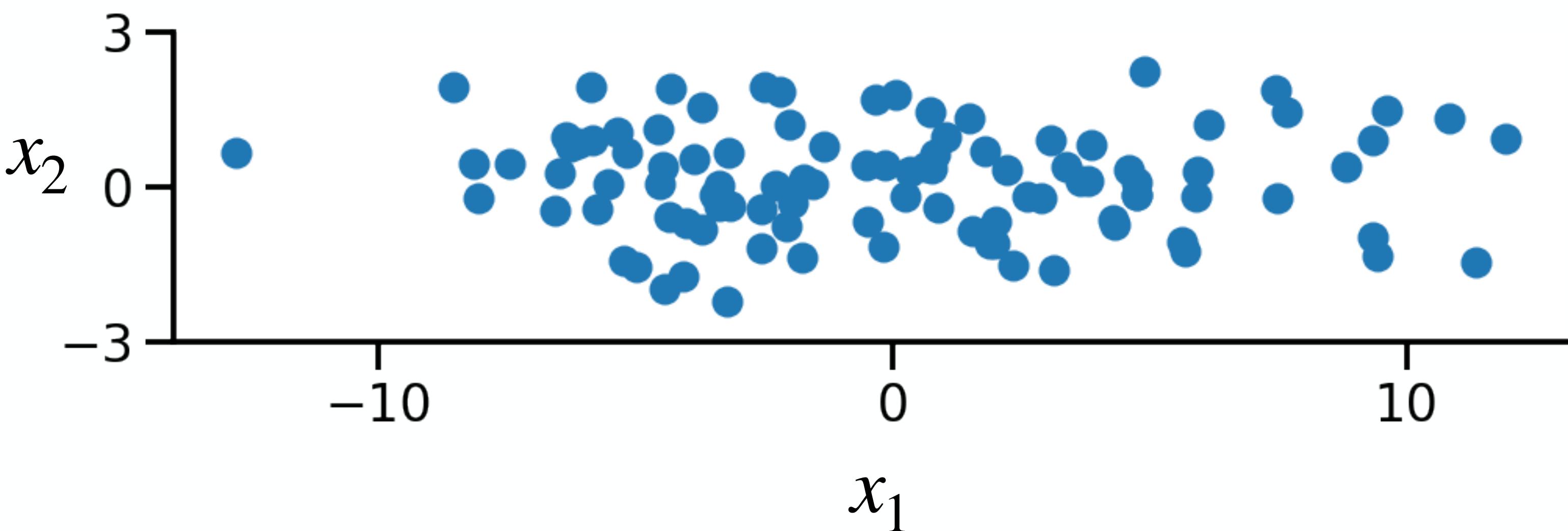
- n : 100k ~ 1 million cells
- p : 10k ~ 50k genes
- k : 10 ~ 100 components

Key assumption / rationale:

- Biology has structure. Genes co-express to form functional modules.
- Data often lives in low-dimensions embedded in high-dimensional space.

A toy example

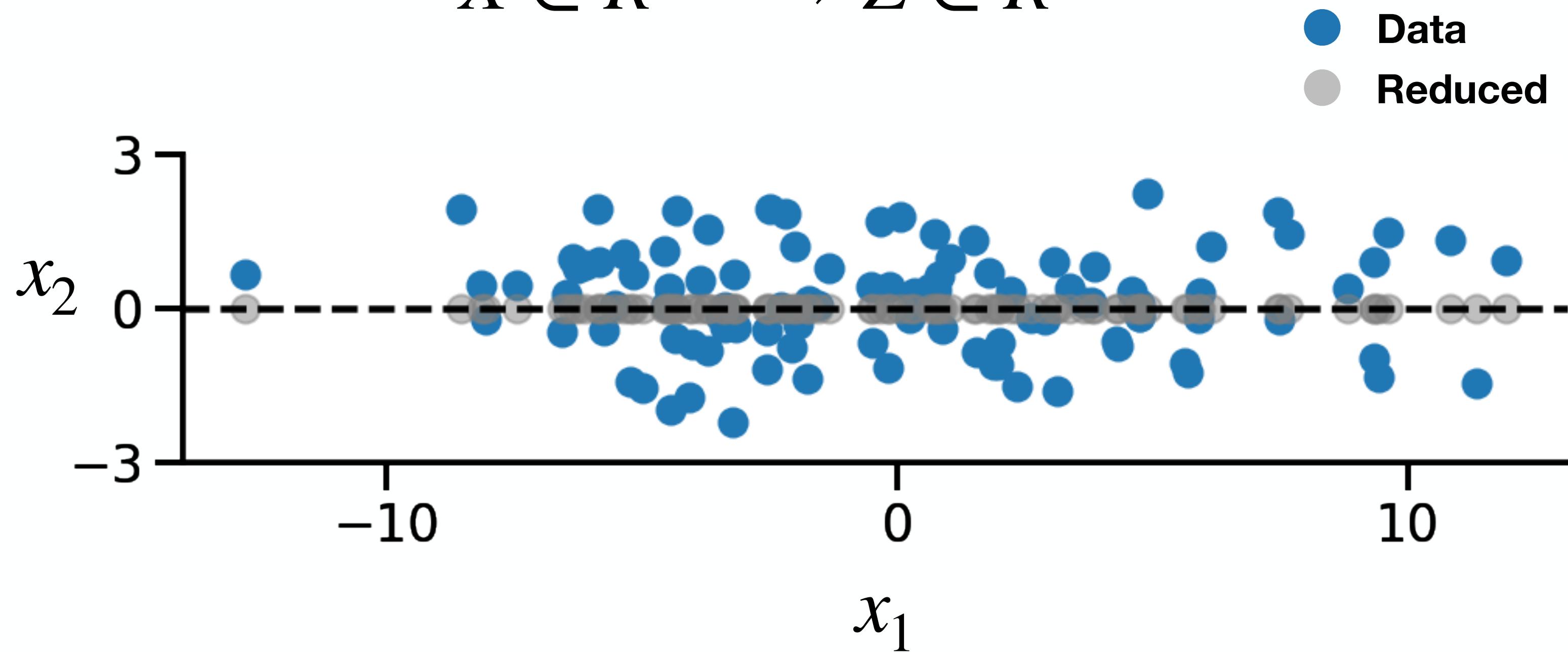
x_1	x_2
[[8.82,	0.4],
[4.89,	2.24],
[9.34,	-0.98],
[4.75,	-0.15],
[-0.52,	0.41],
[0.72,	1.45],
[3.81,	0.12],
[2.22,	0.33],
[7.47,	-0.21],
[1.57,	-0.85],
[-12.76,	0.65],
[4.32,	-0.74],
[11.35,	-1.45],
[0.23,	-0.19],
[7.66,	1.47],
[0.77,	0.38],
[-4.44,	-1.98],
[-1.74,	0.16],
[-6.15,	1.2],



Strategy: remove the low-variance column

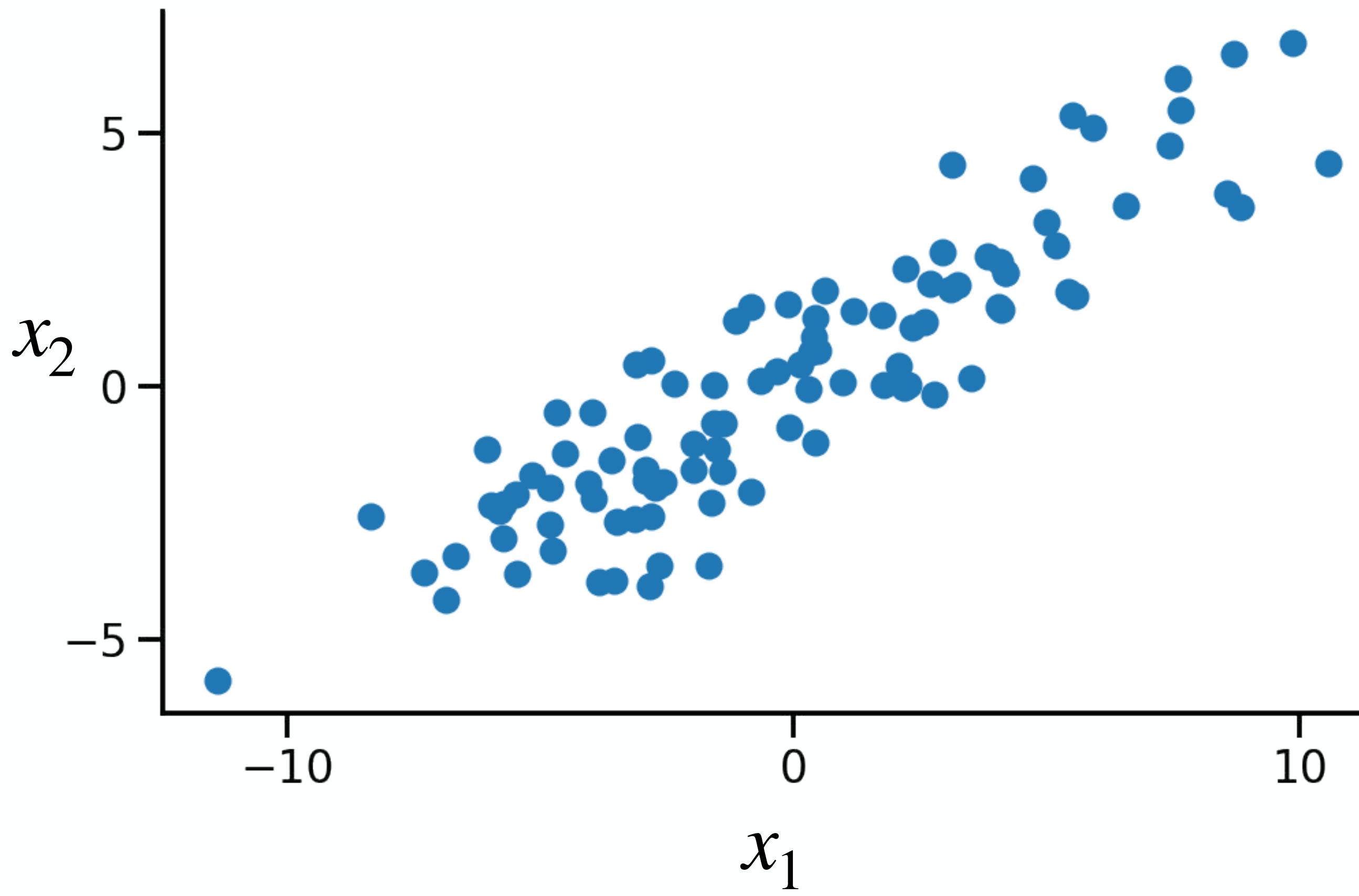
x_1	x_2
[[8.82,	0.4],
[4.89,	2.24],
[9.34,	-0.98],
[4.75,	-0.15],
[-0.52,	0.41],
[0.72,	1.45],
[3.81,	0.12],
[2.22,	0.33],
[7.47,	-0.21],
[1.57,	-0.85],
[-12.76,	0.65],
[4.32,	-0.74],
[11.35,	-1.45],
[0.23,	-0.19],
[7.66,	1.47],
[0.77,	0.38],
[-4.44,	-1.98],
[-1.74,	0.16],
[6.15,	1.2],

$$X \in R^{n \times 2} \rightarrow Z \in R^{n \times 1}$$



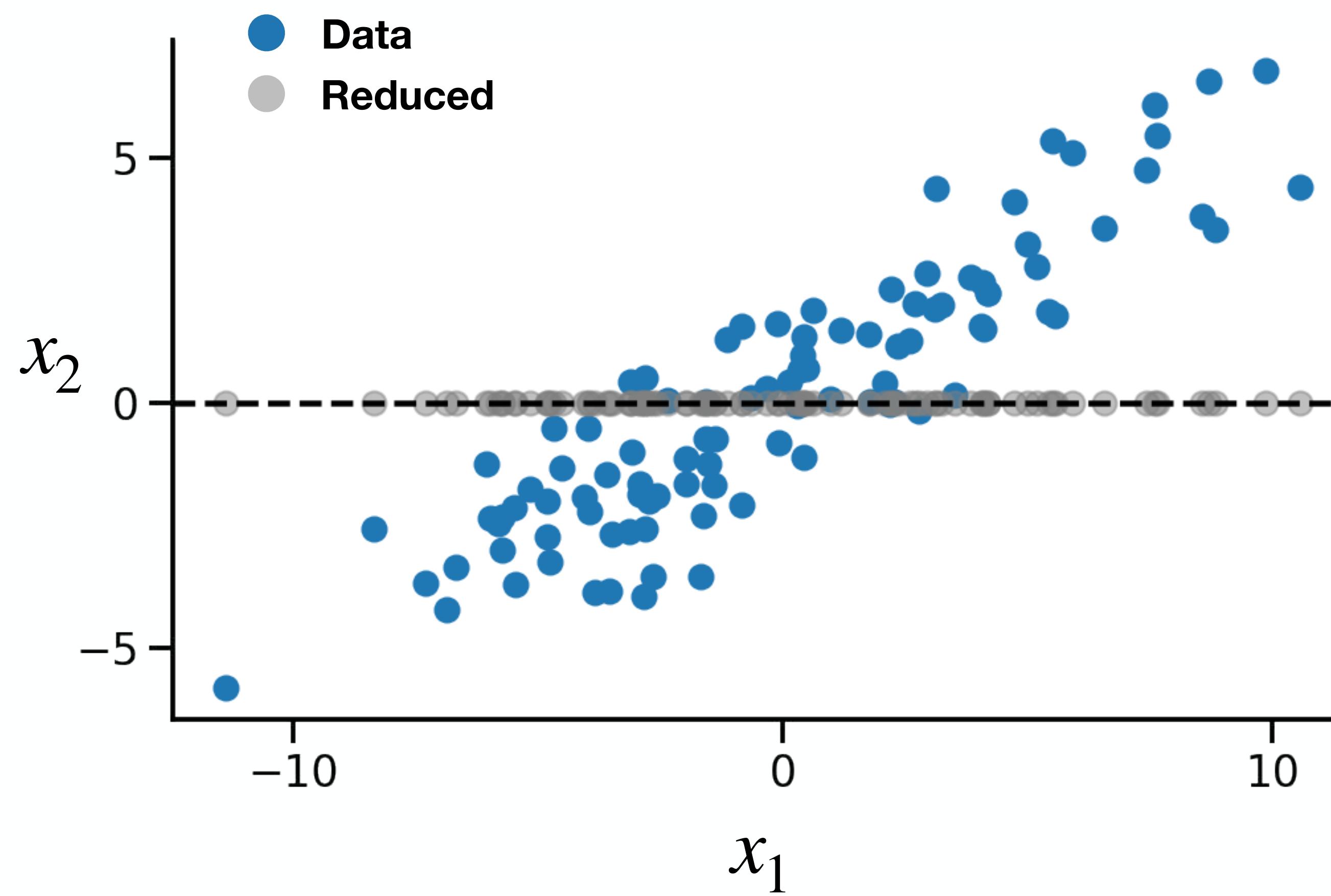
How about now?

x_1	x_2
[[7.44,	4.76],
[3.12,	4.39],
[8.58,	3.82],
[4.19,	2.24],
[-0.65,	0.1],
[-0.1 ,	1.62],
[3.23,	2.01],
[1.76,	1.4],
[6.57,	3.56],
[1.78,	0.04],
[-11.38,	-5.82],
[4.11,	1.52],
[10.56,	4.41],
[0.29,	-0.05],
[5.9 ,	5.1],
[0.48,	0.71],
[-2.85,	-3.93],
[-1.58,	-0.73],
[4.73,	4.12]



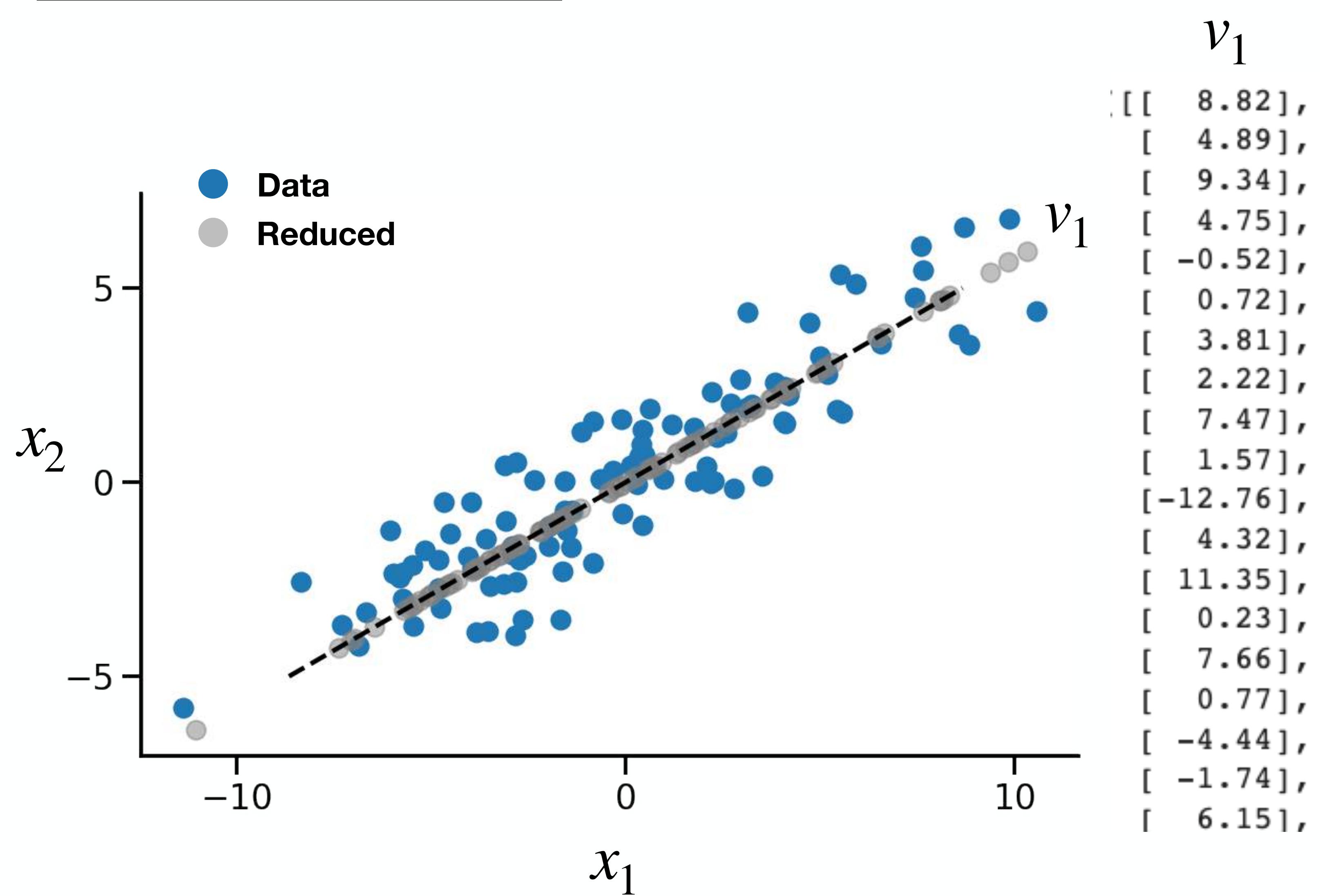
Still removing the low-variance column?

	x_1	x_2
[[7.44,	4.76],
[3.12,	4.39],
[8.58,	3.82],
[4.19,	2.24],
[-0.65,	0.1],
[-0.1 ,	1.62],
[3.23,	2.01],
[1.76,	1.4],
[6.57,	3.56],
[1.78,	0.04],
[-11.38,	-5.82],
[4.11,	1.52],
[10.56,	4.41],
[0.29,	-0.05],
[5.9 ,	5.1],
[0.48,	0.71],
[-2.85,	-3.93],
[-1.58,	-0.73],
[4.73,	4.12]



A better strategy: remove the low-variance column after a rotation

	\hat{x}_1	\hat{x}_2
[[7.44,	4.76],
[3.12,	4.39],
[8.58,	3.82],
[4.19,	2.24],
[-0.65,	0.1],
[-0.1 ,	1.62],
[3.23,	2.01],
[1.76,	1.4],
[6.57,	3.56],
[1.78,	0.04],
[-11.38,	-5.82],
[4.11,	1.52],
[10.56,	4.41],
[0.29,	-0.05],
[5.9 ,	5.1],
[0.48,	0.71],
[-2.85,	-3.93],
[-1.58,	-0.73],
[4.73,	4.12]



Principal Component Analysis (PCA) formalizes the intuition

- Reduce data dimensions by removing low-variance basis after a rotation.
- Identify the “best” basis in the feature space to project data points.
- Amounts to a Singular Value Decomposition (SVD):

$$X = USV^T = \underbrace{u_1 s_1 \cdot v_1^T}_{\text{PC 1}} + \underbrace{u_2 s_2 \cdot v_2^T}_{\text{PC 2}} + \dots + \underbrace{u_r s_r \cdot v_r^T}_{\text{PC r}}$$

Principal Component Analysis

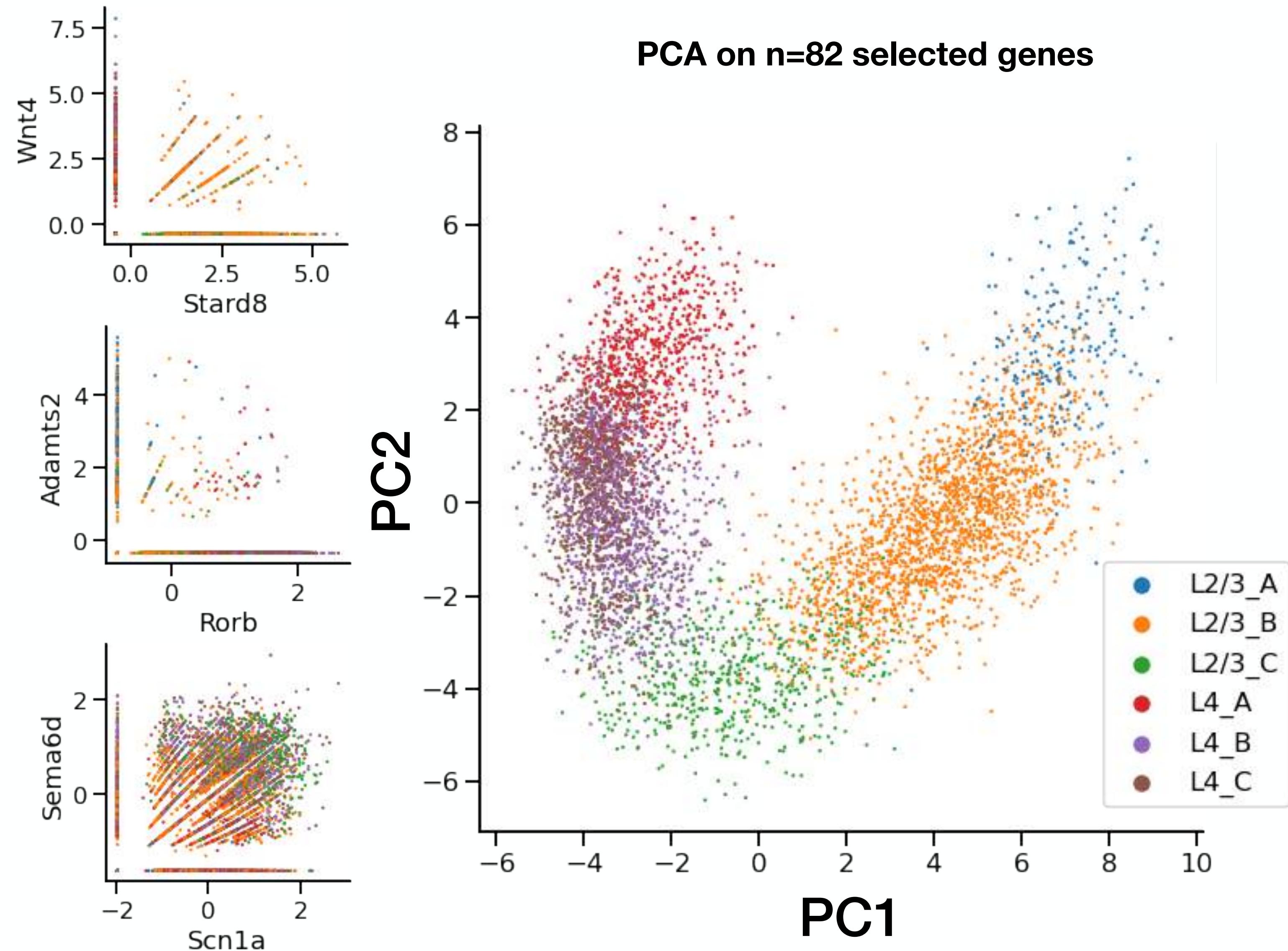
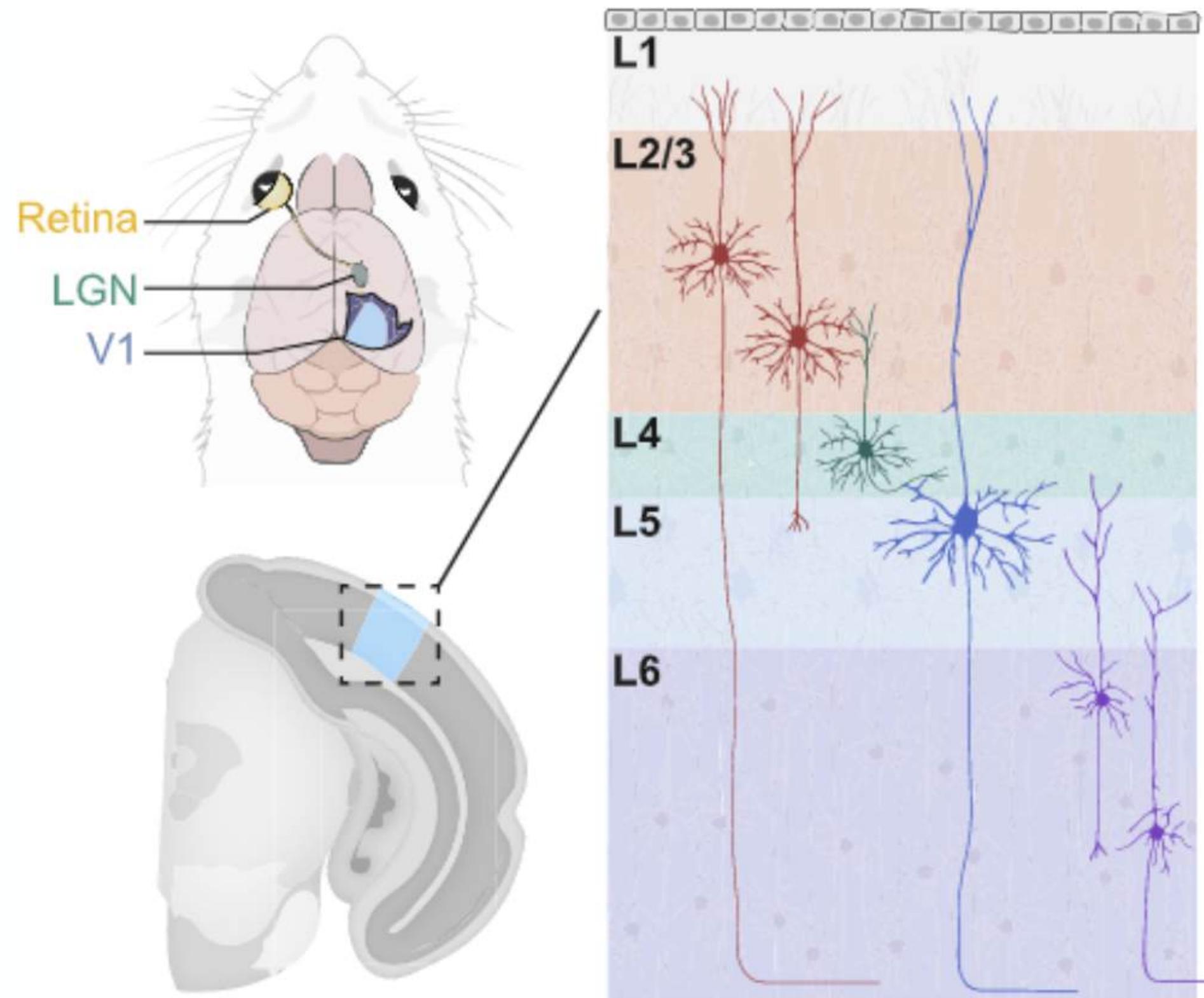
- Reduce data dimensions by removing low-variance basis after a rotation.
- Identify the “best” basis in the feature space to project data points.
- Amounts to a Singular Value Decomposition (SVD):

$$X = USV^T = \underbrace{u_1 s_1}_{\text{scores}} \cdot \underbrace{v_1^T}_{\text{basis}} + \underbrace{u_2 s_2}_{\text{scores}} \cdot \underbrace{v_2^T}_{\text{basis}} + \dots + \underbrace{u_r s_r}_{\text{scores}} \cdot \underbrace{v_r^T}_{\text{basis}}$$

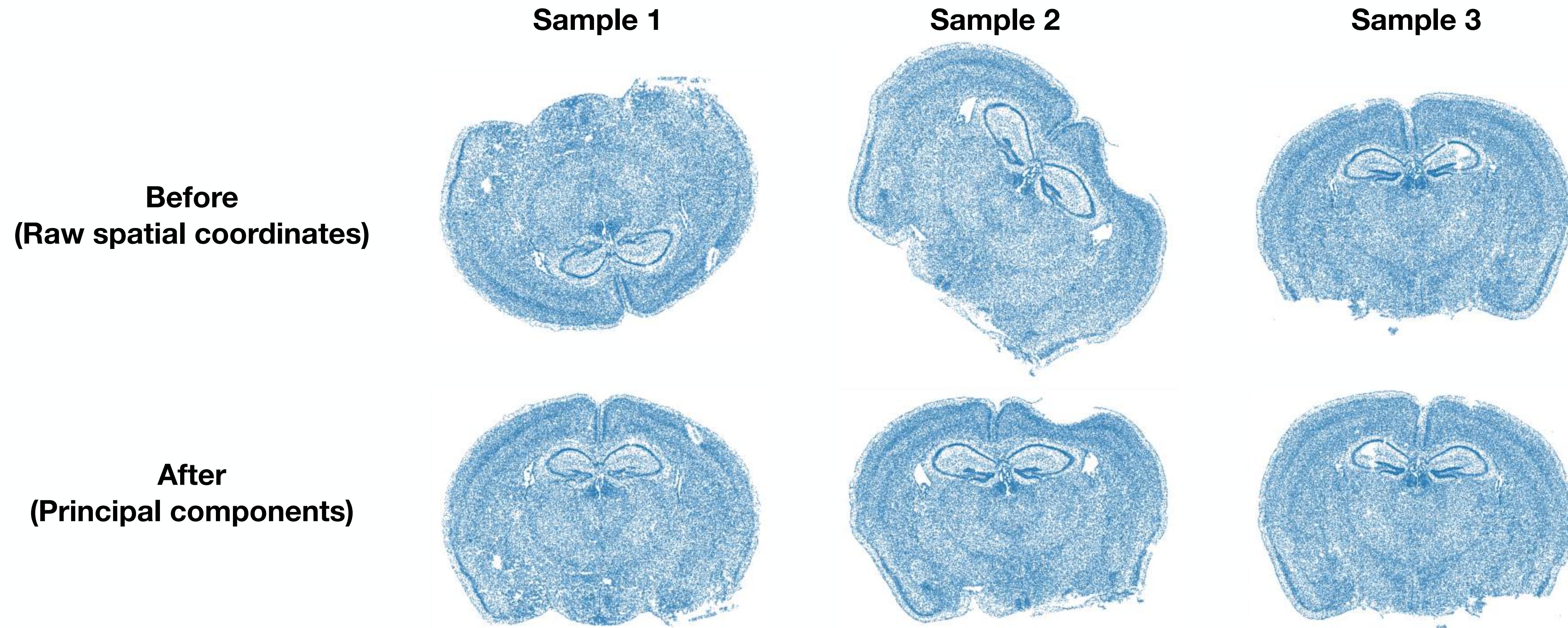
standard deviation

Works in more than 2D!

Application: Transcriptional state of cortical neurons



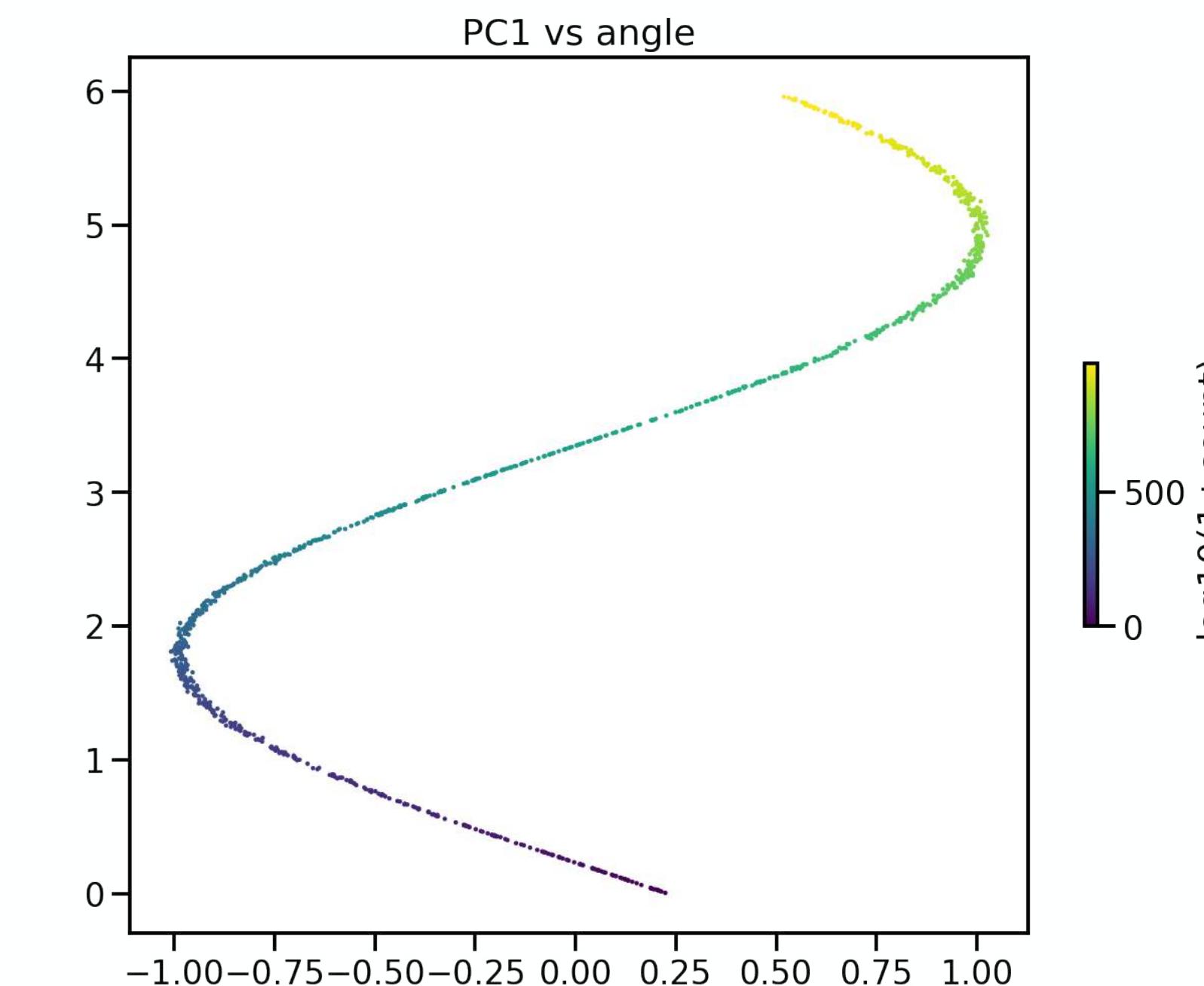
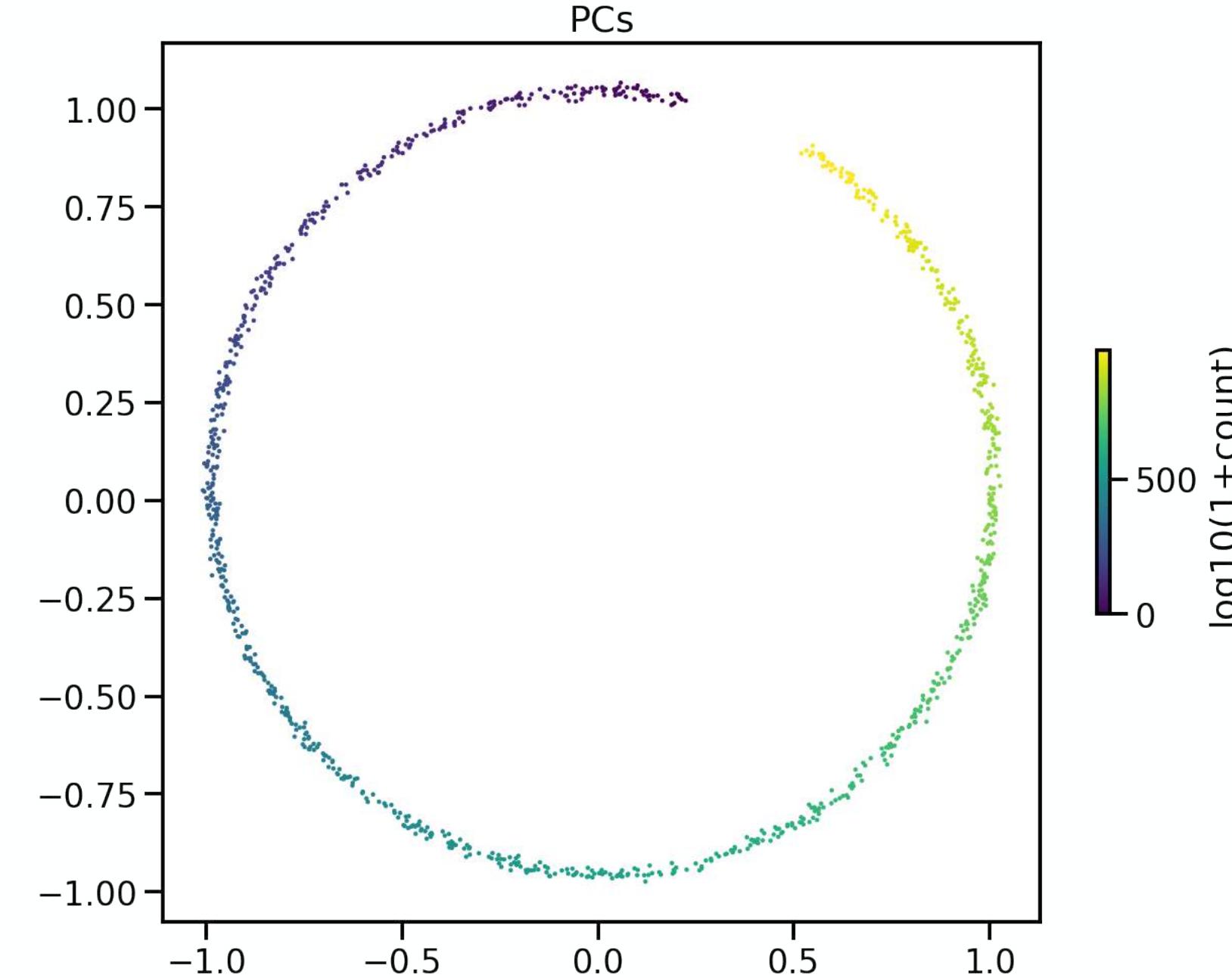
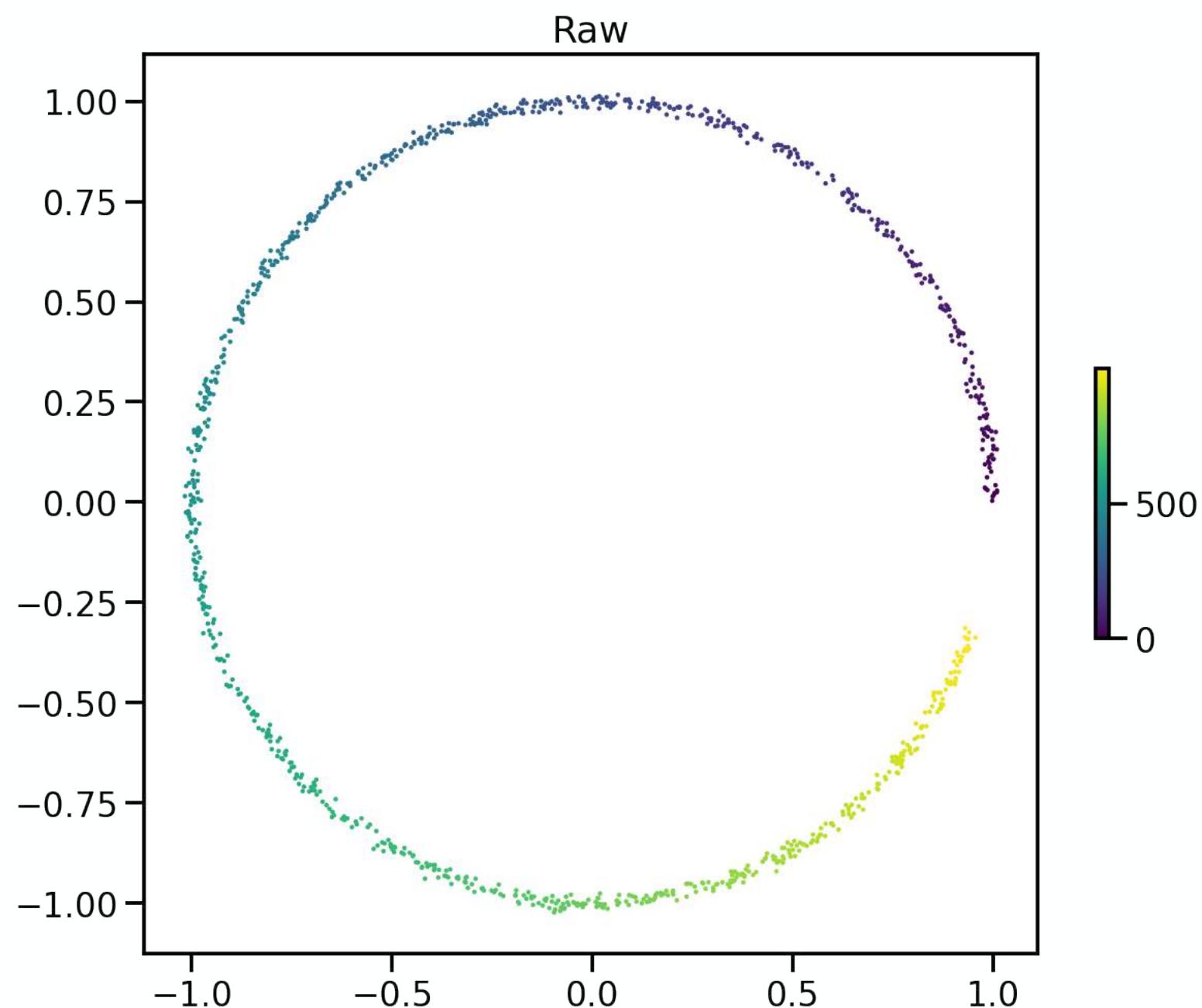
Application: rotate the mouse brain



Data from Vizgen: <https://info.vizgen.com/mouse-brain-map?submissionGuid=d92e5a3b-77dc-4e68-8bf5-93e6d071a0e5>

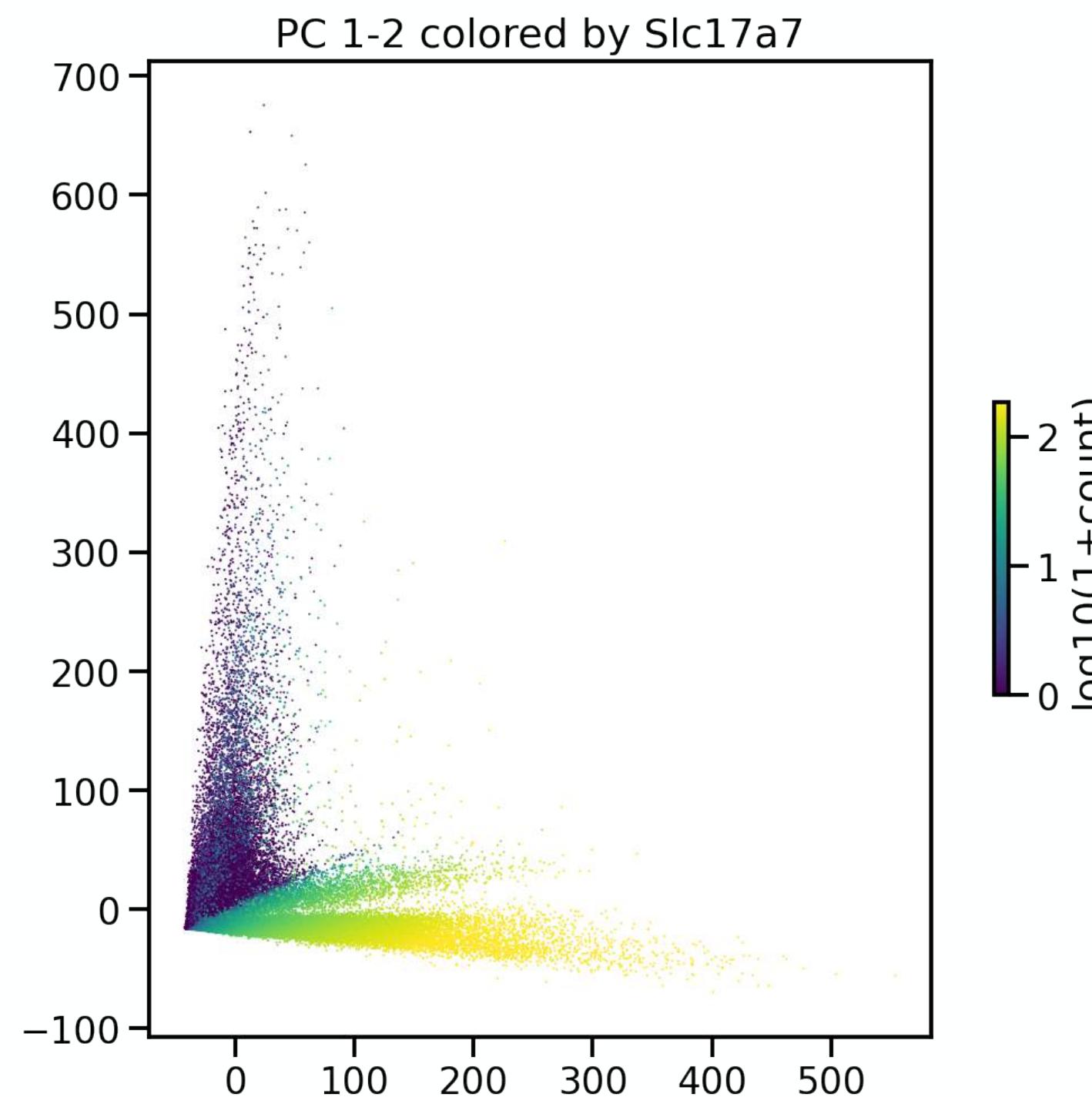
Demo: applying PCA to a toy model data

- How does this toy model relate everything we learned about PCA and UMAP?



Exercise 2a: apply PCA to real data (Vizgen)

- Apply PCA to reduce the dimension from all genes ($n=483$) to 20 PCs
- Plot the first 2 PCs and color it by the expression levels of *Slc17a7*. What is the pattern? What is your interpretation?
- How much variance does each of the 20 PCs accounts for? How much in total?



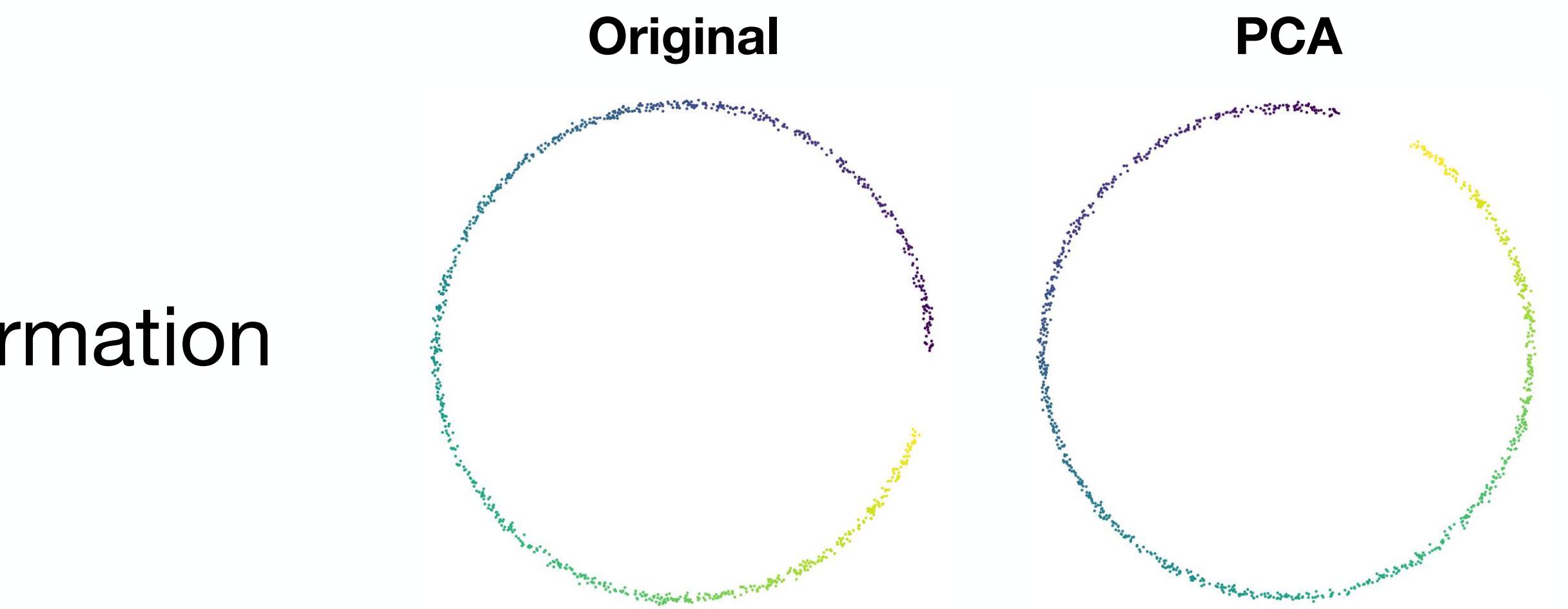
Discussion: when does PCA fail?

- Wrong answers only?

When does PCA fail?

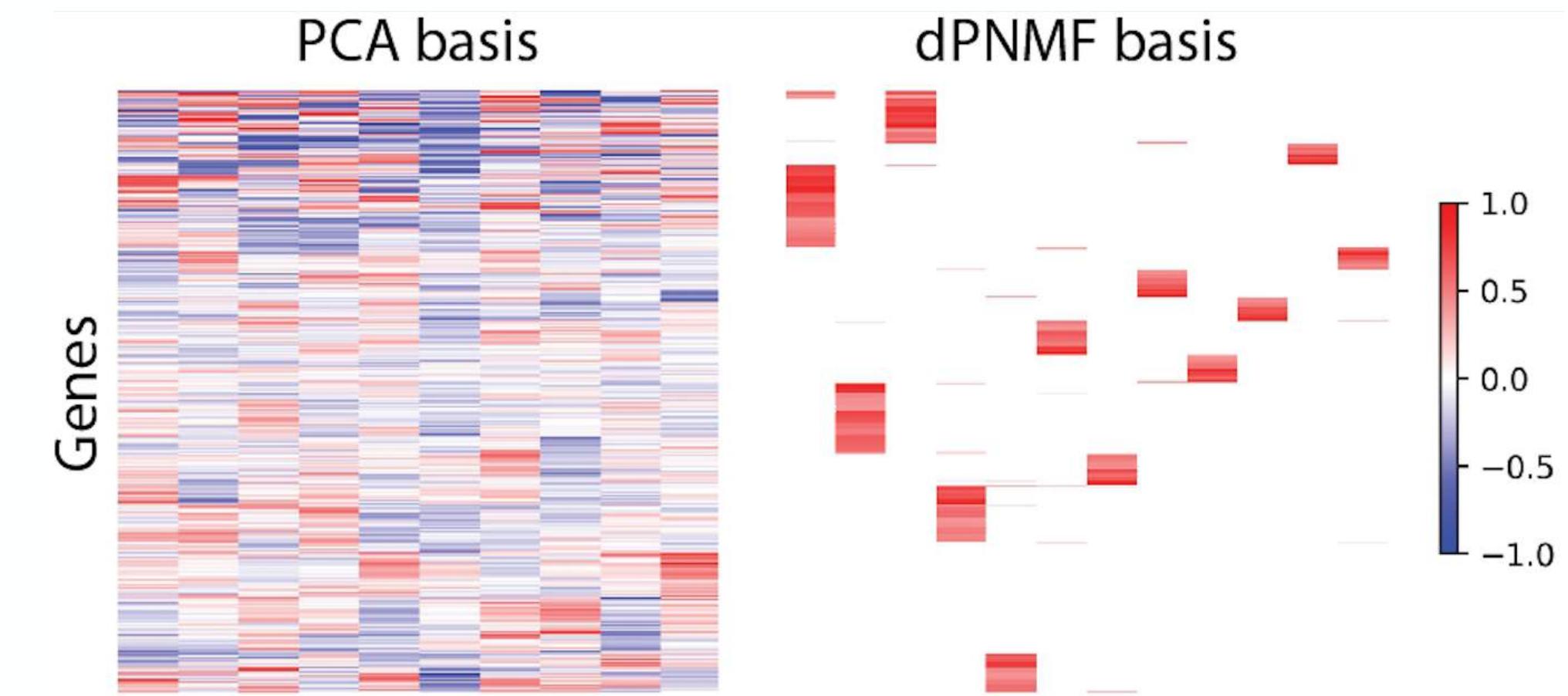
When does PCA fail?

- Linear rotation vs Non-linear transformation



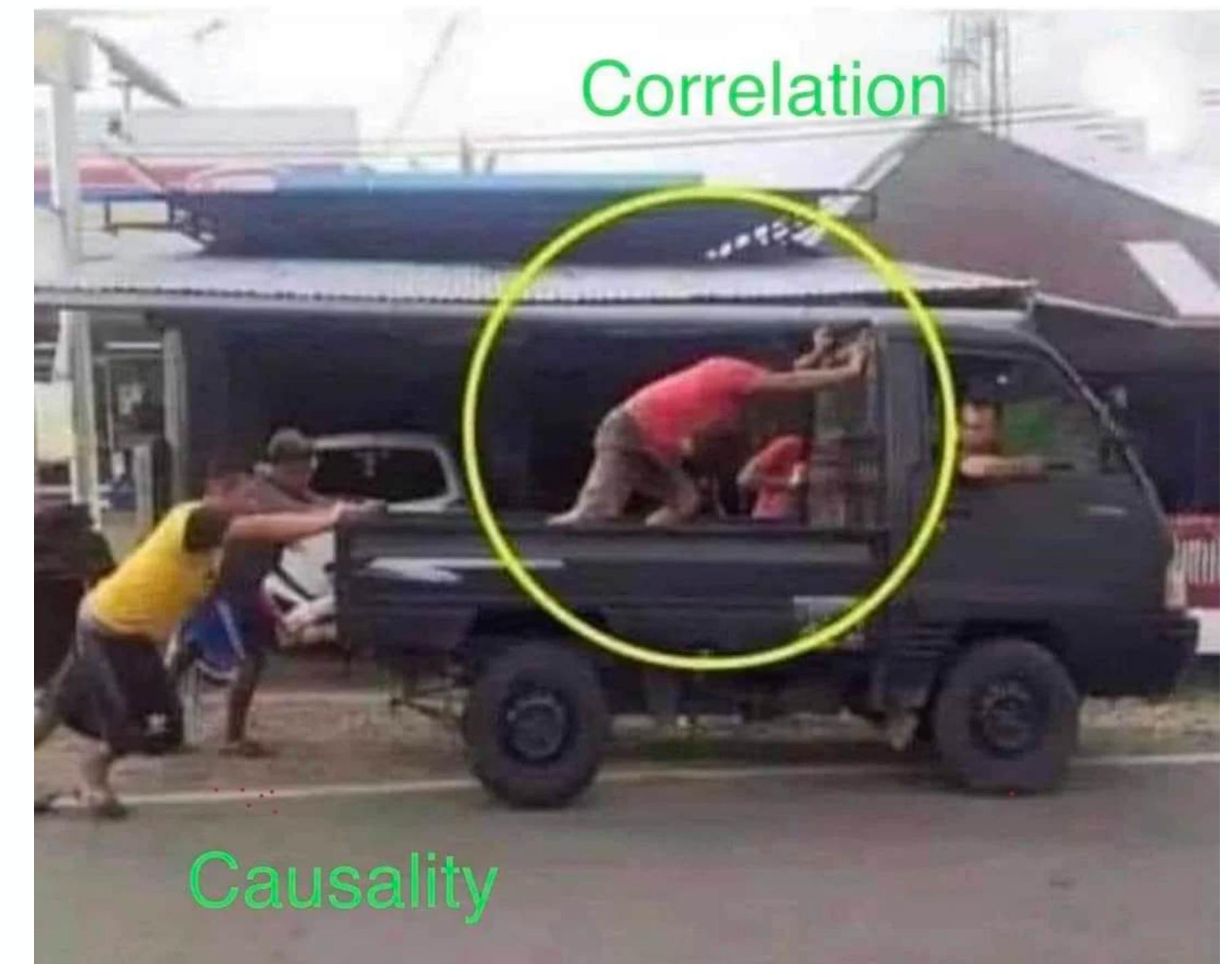
When does PCA fail?

- Linear rotation vs Non-linear transformation
- “dense” components vs “sparse” modules



When does PCA fail?

- Linear rotation vs Non-linear transformation
- “dense” components vs “sparse” modules
- Correlation vs causation
- Unbiased vs modeling-based

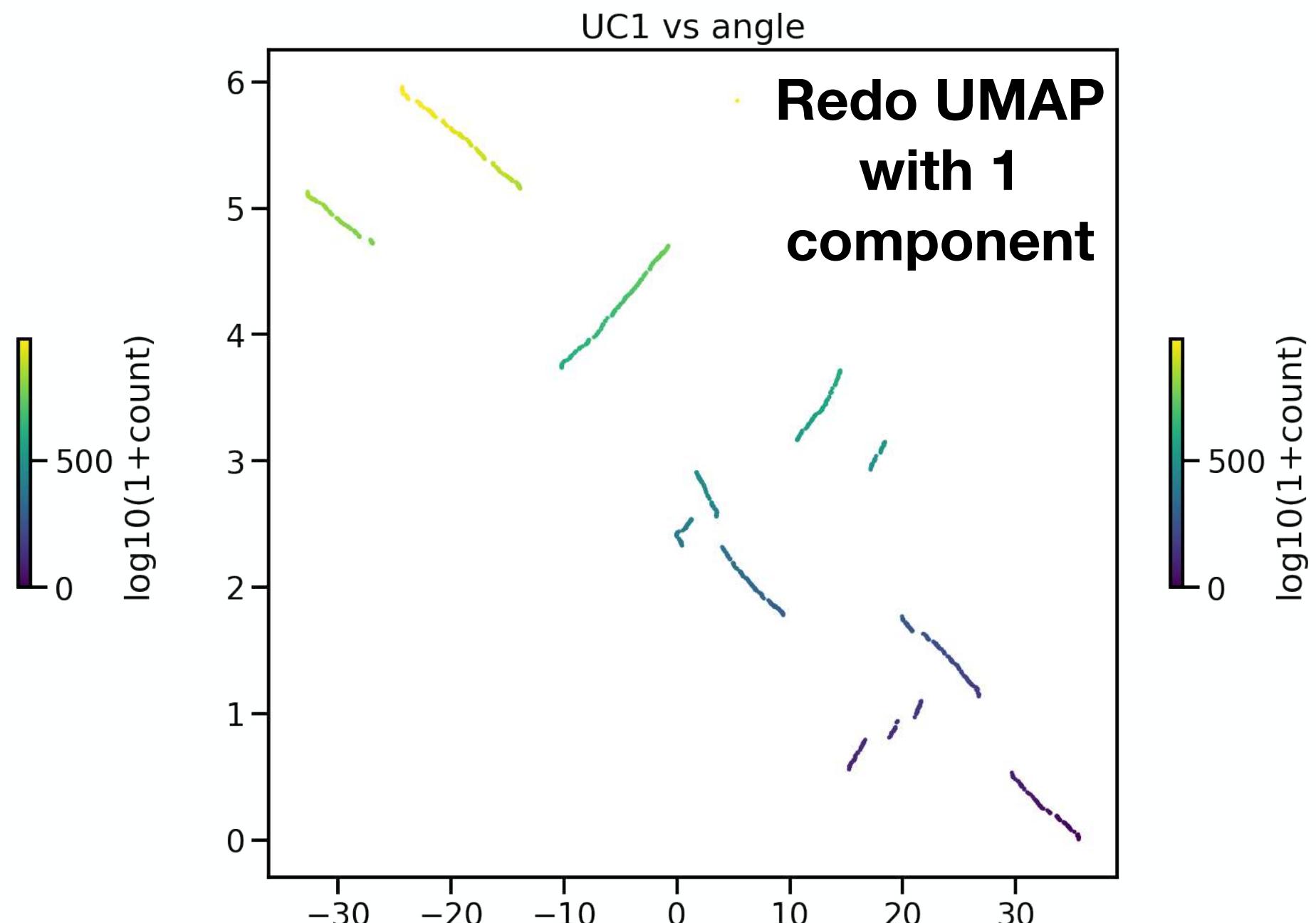
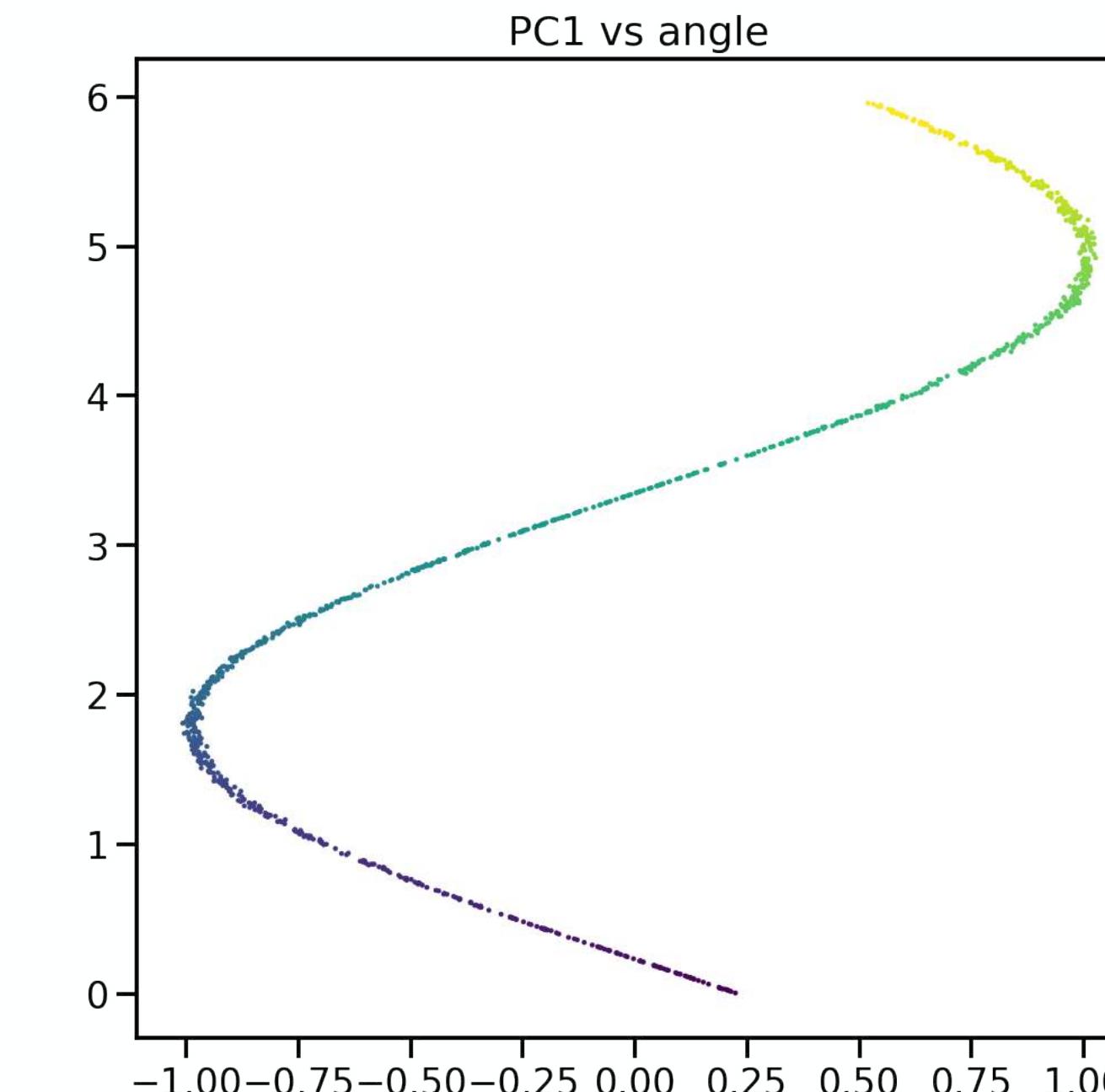
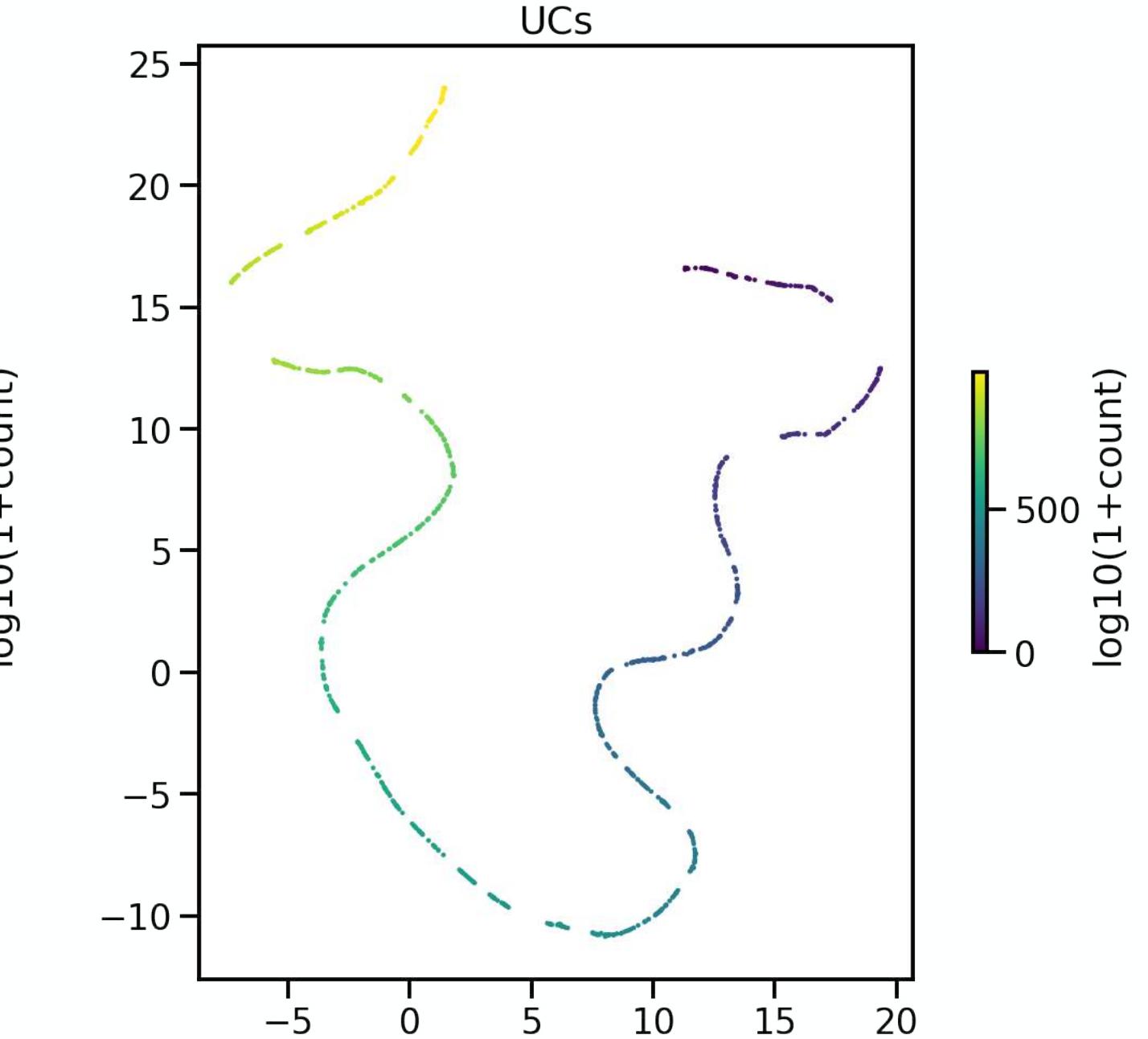
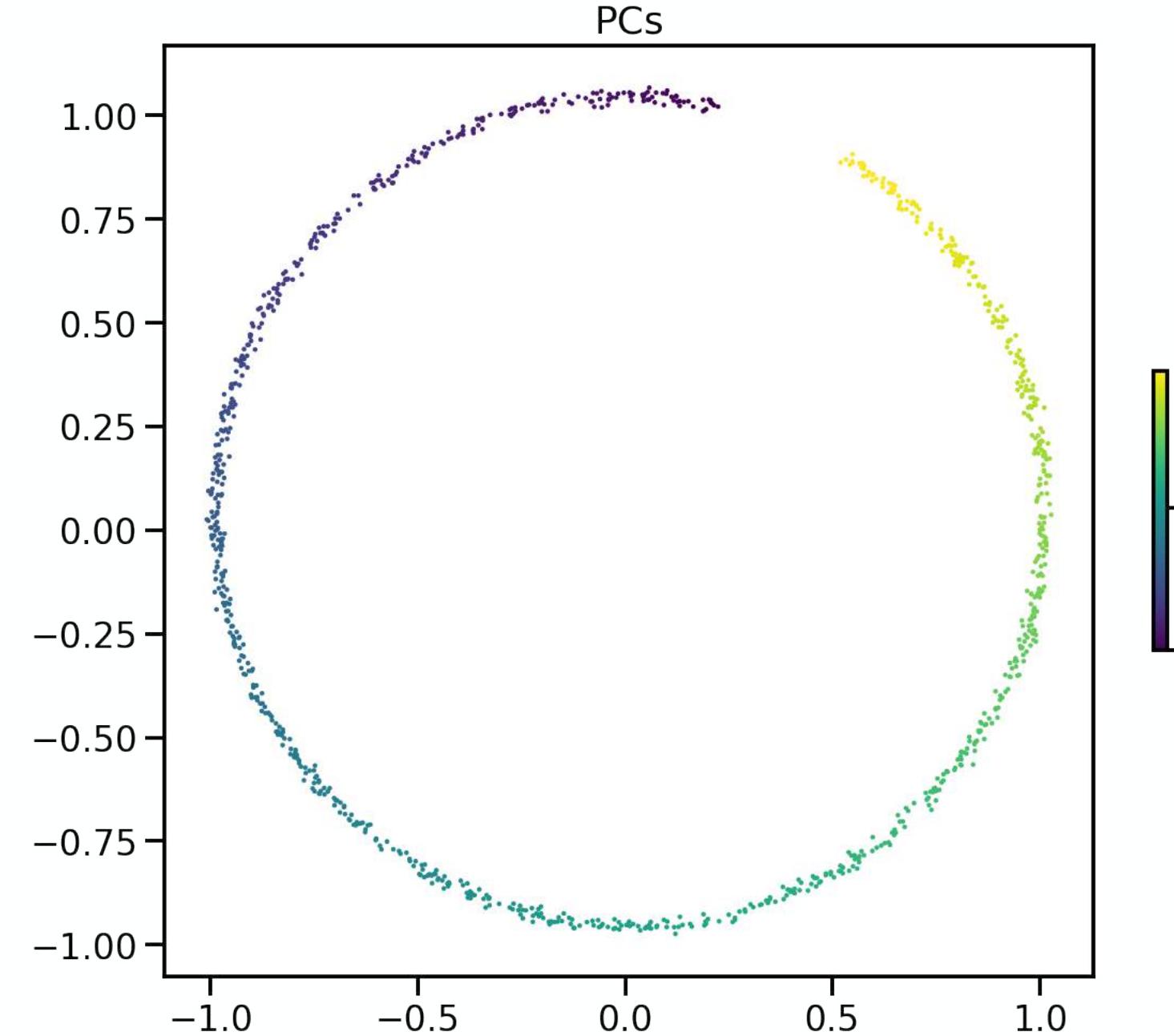
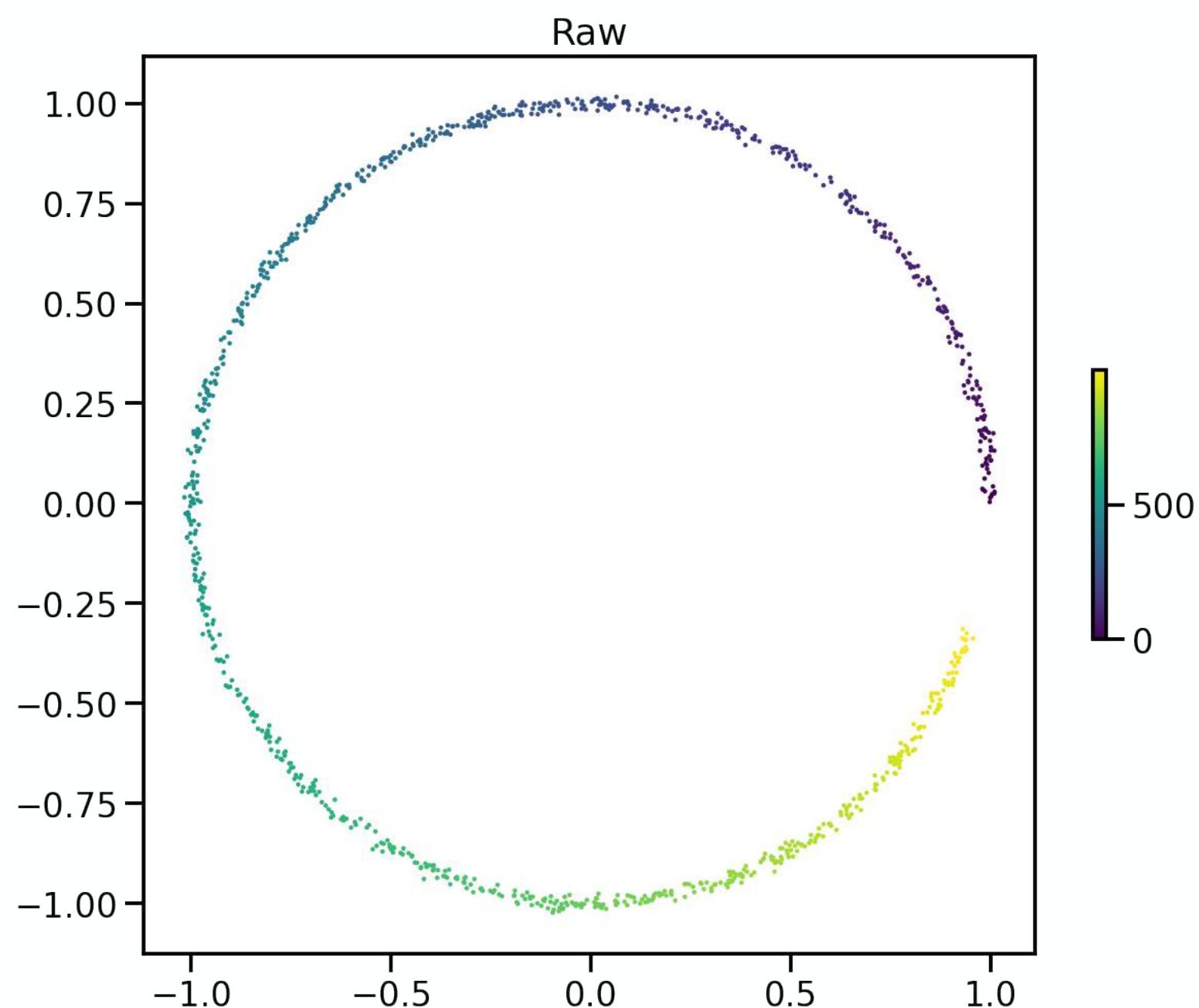


UMAP

- **UMAP:** uniform manifold approximation and projection for dimension reduction
 - Complex, non-linear, flexible, slower but reasonably fast
 - Preserves local structure (neighborhood) as much as possible
- Steps:
 - Represents high-dim data points by a neighborhood graph
 - Embed the graph to low-dim (most often 2) while preserving neighborhood
- How can we understand such a complex thing?

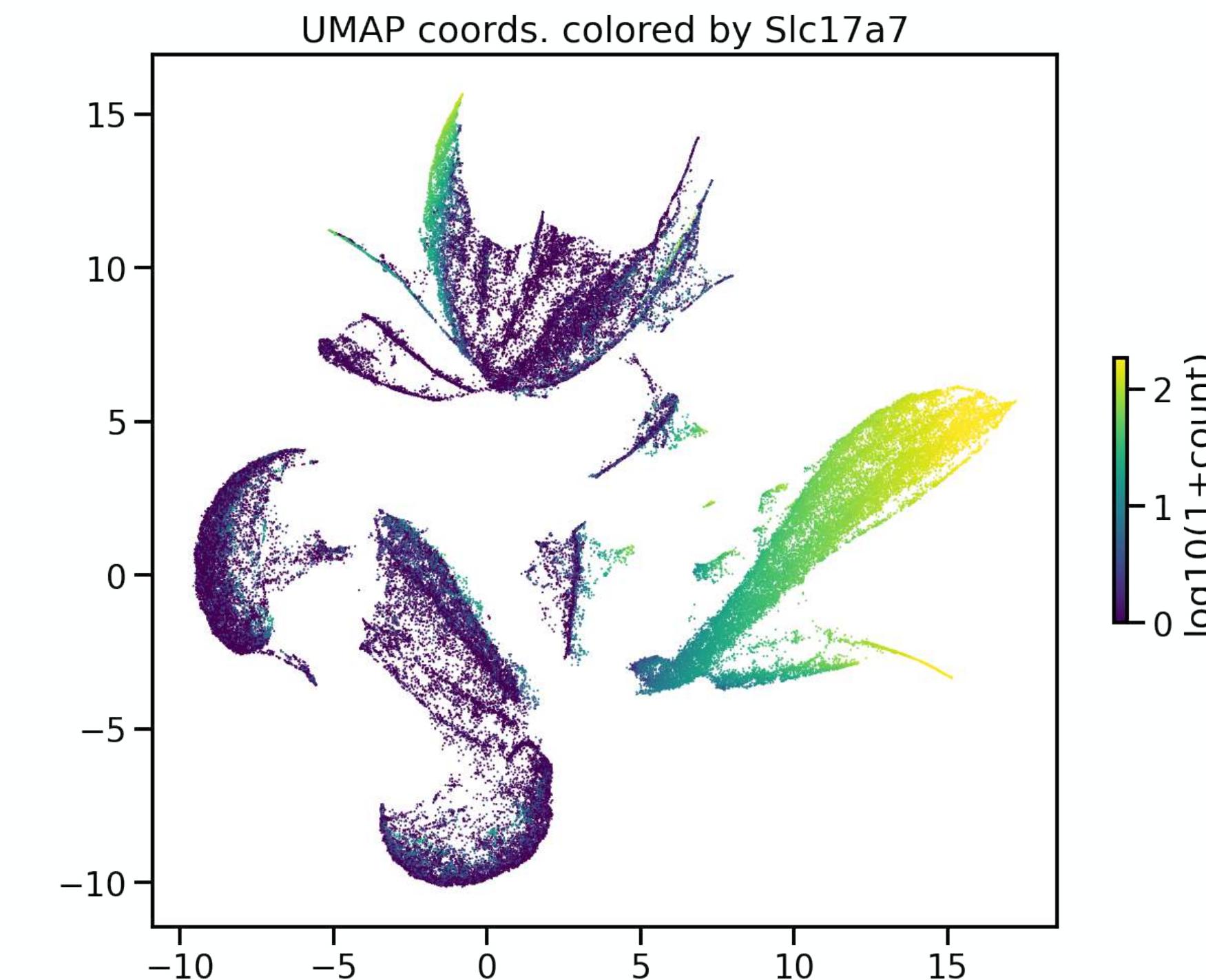
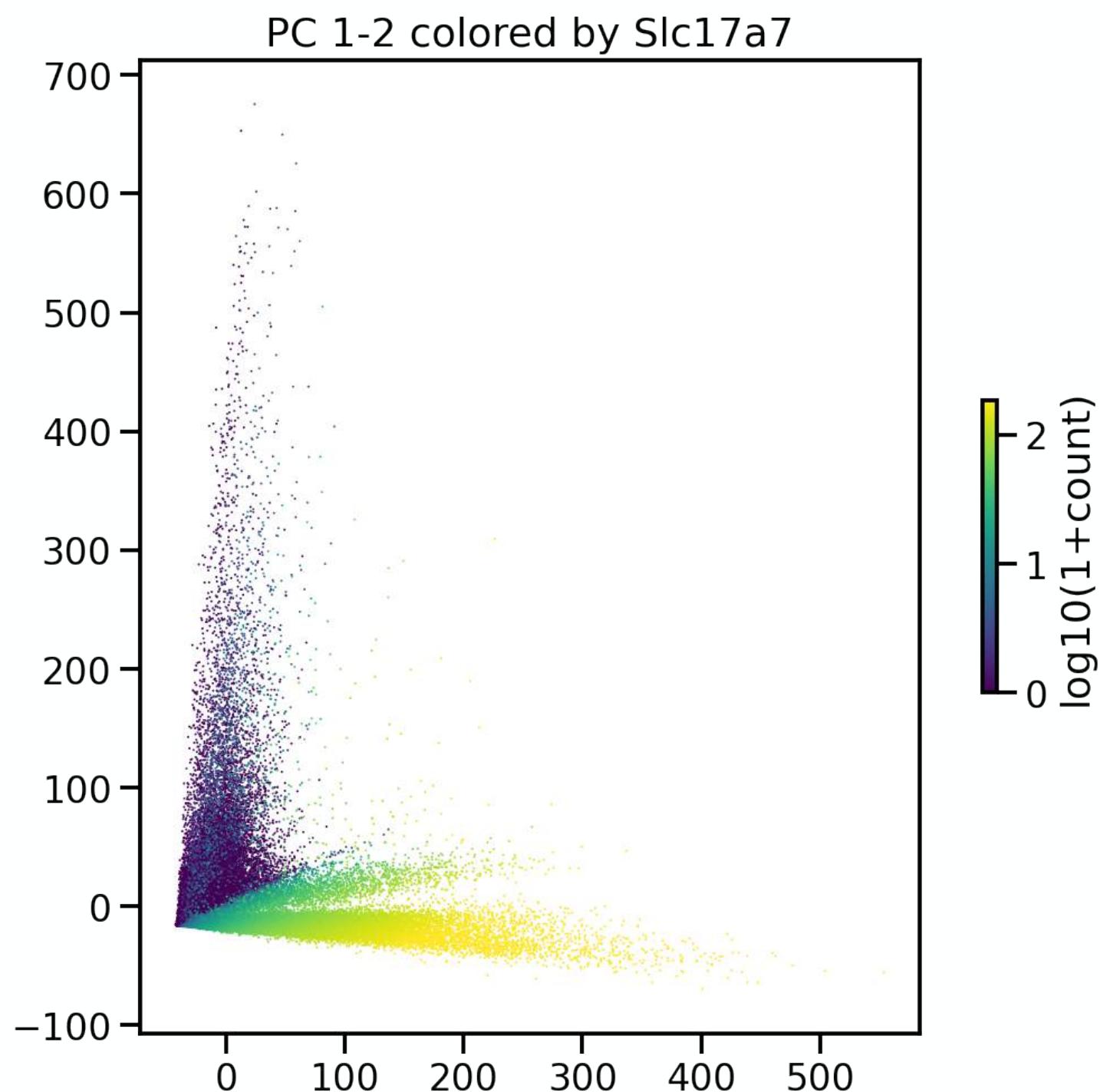
Probing PCA vs UMAP with a toy model

- How does this toy model relate everything we learned about PCA and UMAP?



Exercise 2b: apply PCA and UMAP to real data (Vizgen)

- Apply PCA to reduce the dimension from all genes (n=483) to 20 PCs
- Apply UMAP on the 20 PCs to further reduce the dimension from 20 to 2.



Two ends of dimensionality reduction

- **PCA:** principal component analysis
 - Simple, linear, clear, robust, fast
 - Preserves global structure (everything) as much as possible
- **UMAP:** uniform manifold approximation and projection for dimension reduction
 - Complex, non-linear, flexible, slower but reasonably fast
 - Preserves local structure (neighborhood) as much as possible

UCLA Collaboratory workshop (W31)

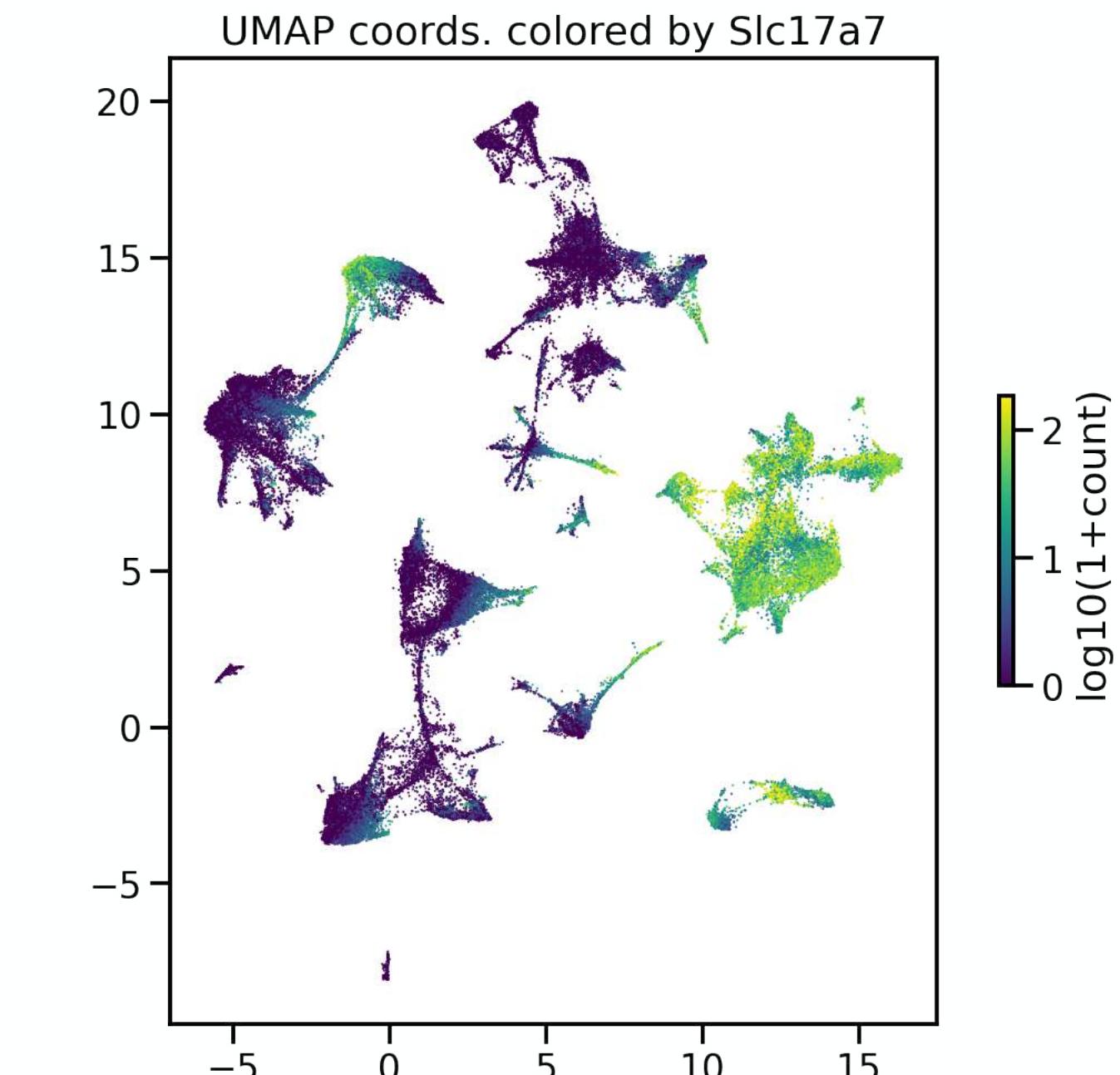
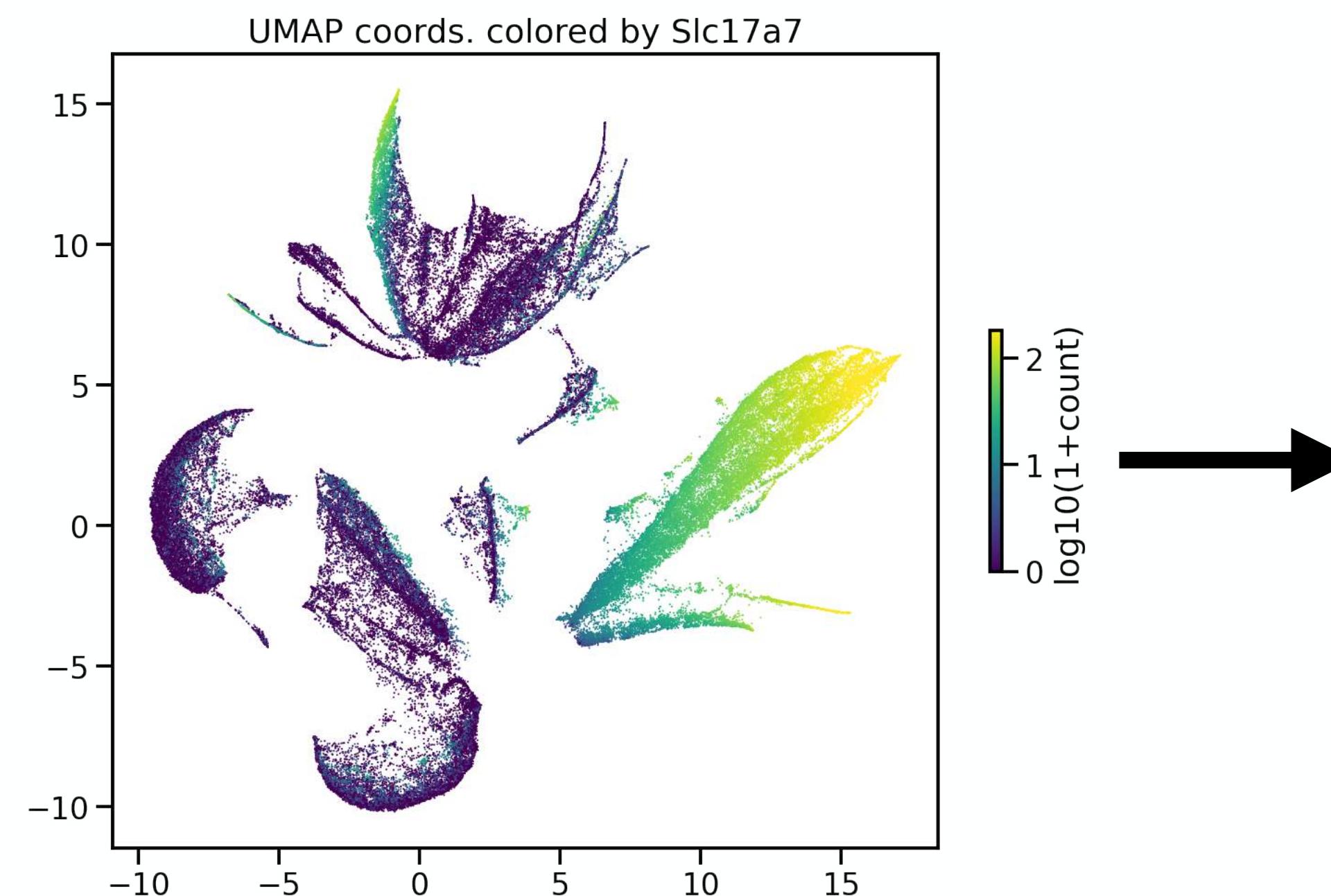
Spatial Transcriptomics

Oct 30, 2025 (Day 3)
Fangming Xie

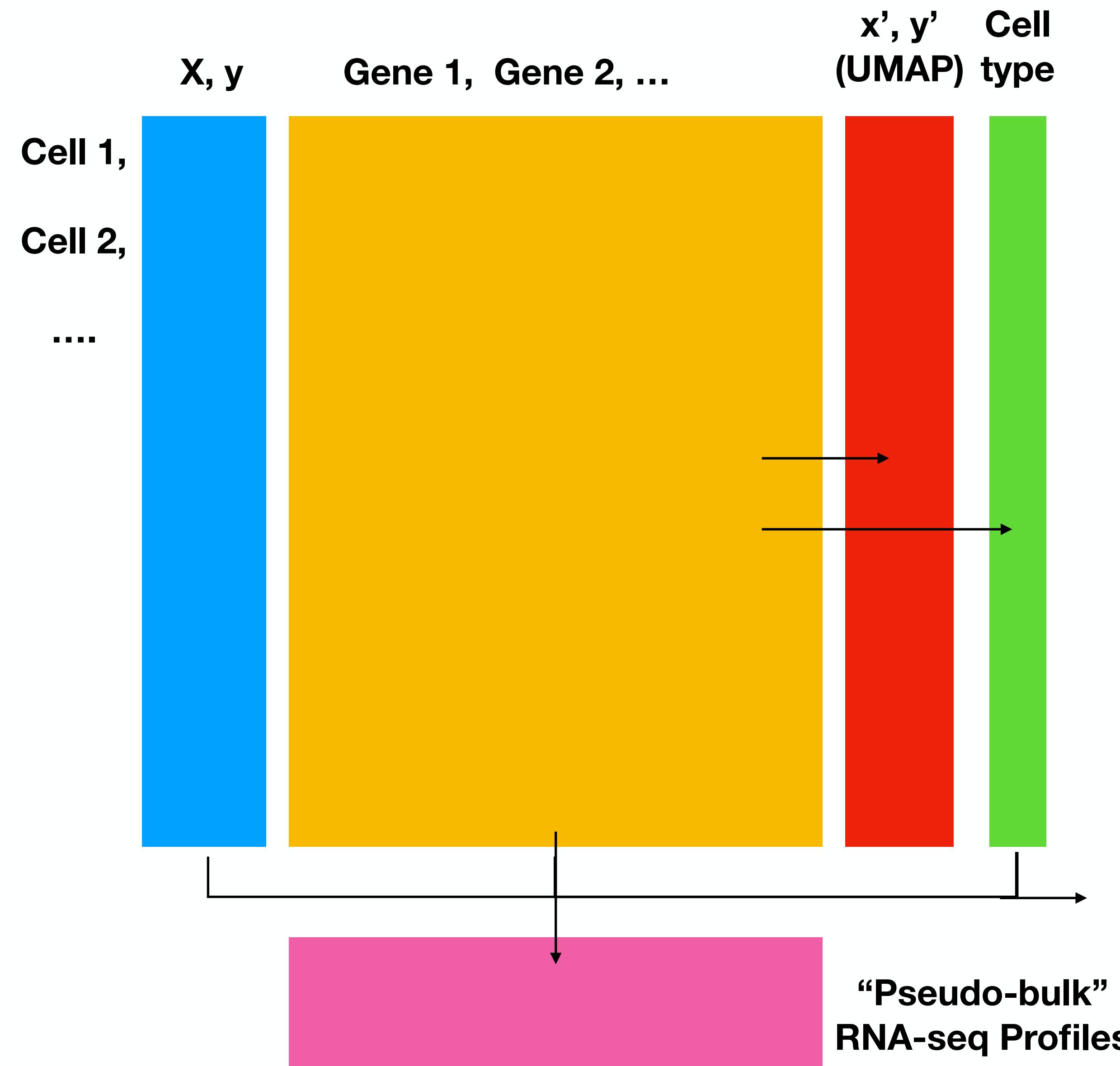
Exercise 3: Normalization

- Neither PCA nor UMAP are magic – they cannot separate technical noise from biological signals.
- We need to normalize the raw data to remove known technical artifacts.
- In single-cell transcriptomics, a rule of thumb is to normalize raw counts by cell size (or cell library size) and do log transformation. (why?)

Graded patterns are all over the place, suggesting continuous variations that are likely due to technical variation.



Analysis overview

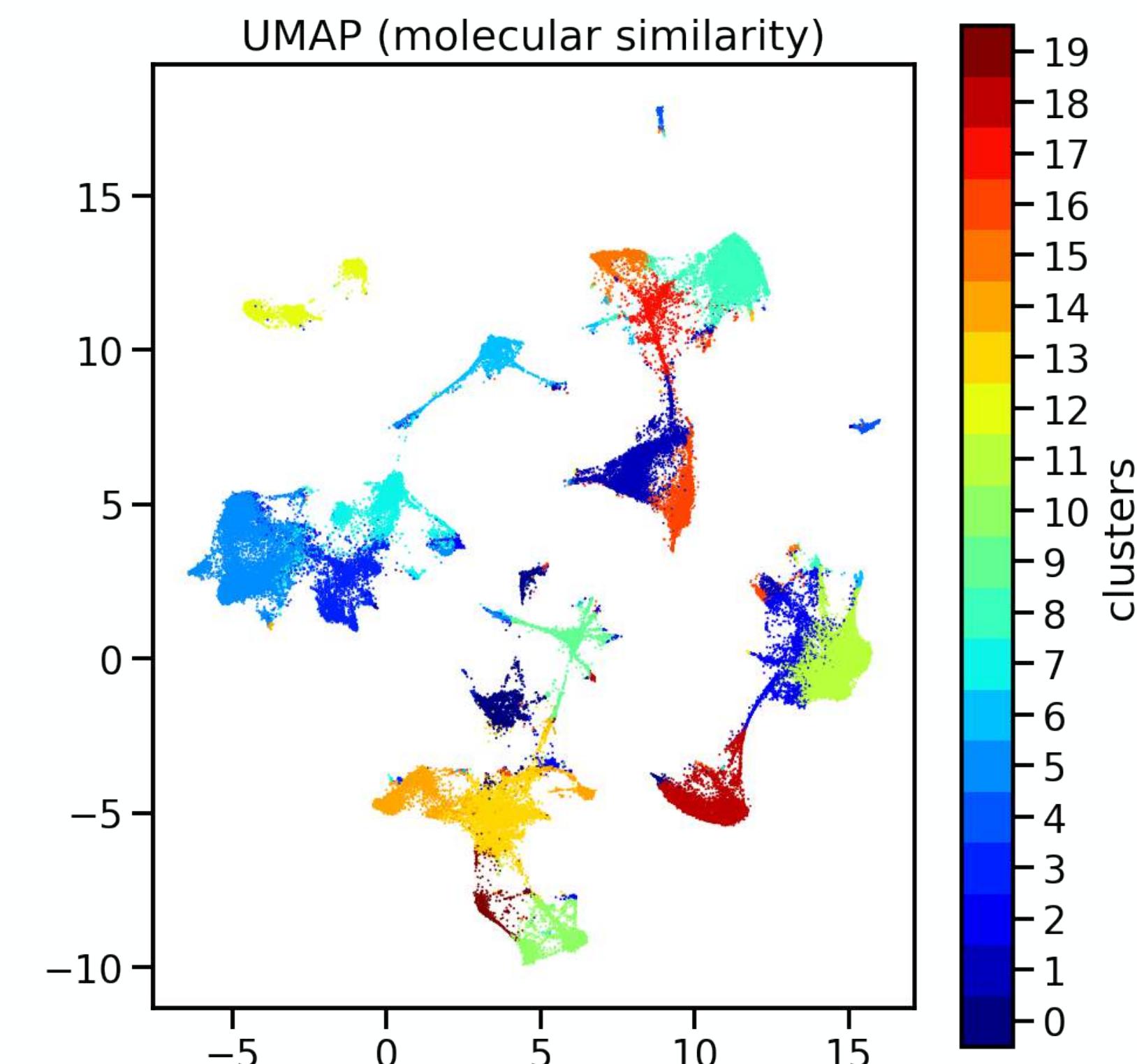
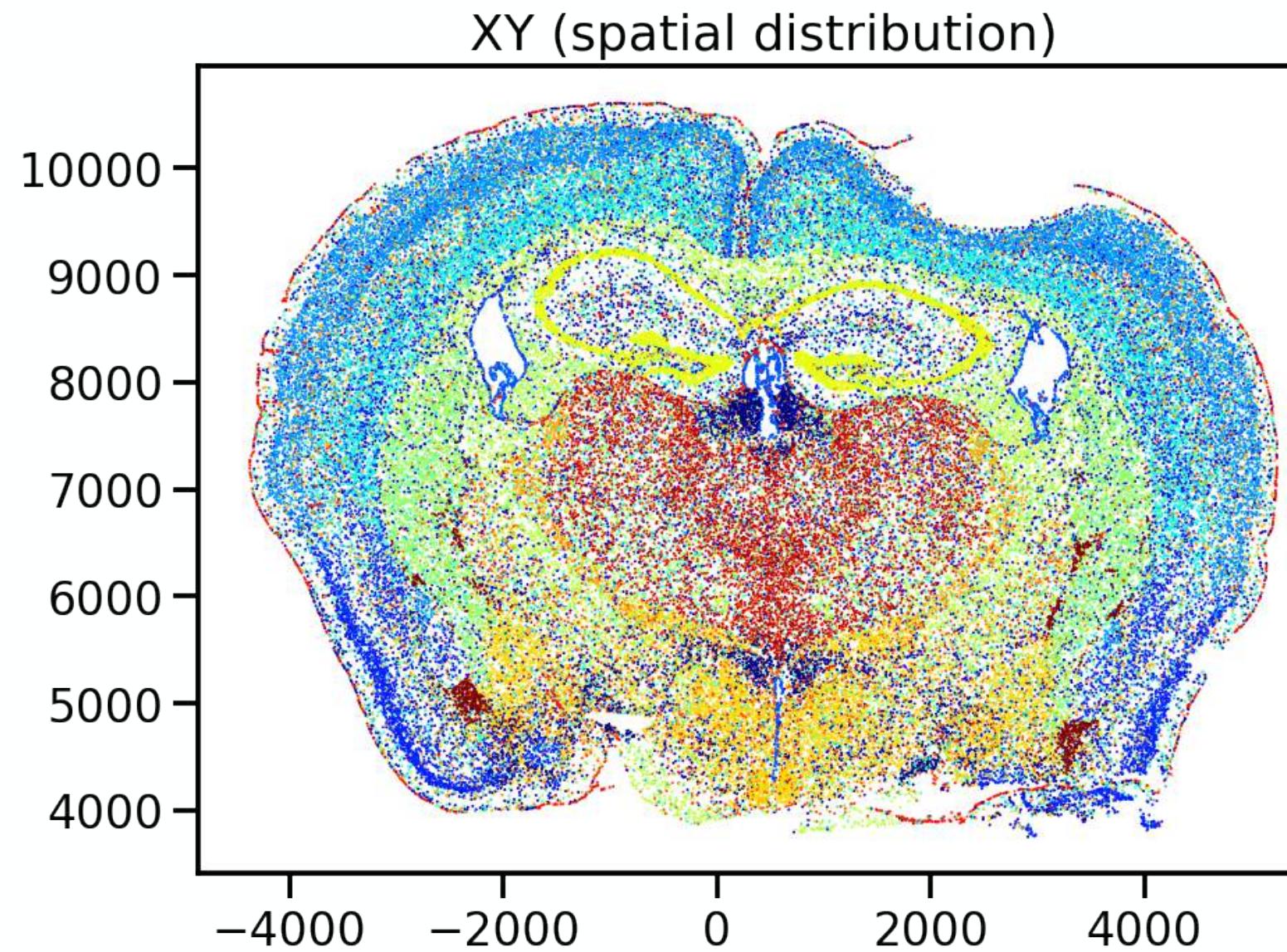


- Where (and in which cells) is gene X expressed? [Visualization]
- Can we capture the main components of transcriptomic diversity between cells? [Dimensionality reduction]
- Are there different cell types? [Clustering]
- Which cell types interact with which others? [spatial enrichment of cell types; ligand-receptor pairings...]

Exercise 4: Clustering - KMeans

<https://github.com/FangmingXie/collab-workshop-st>

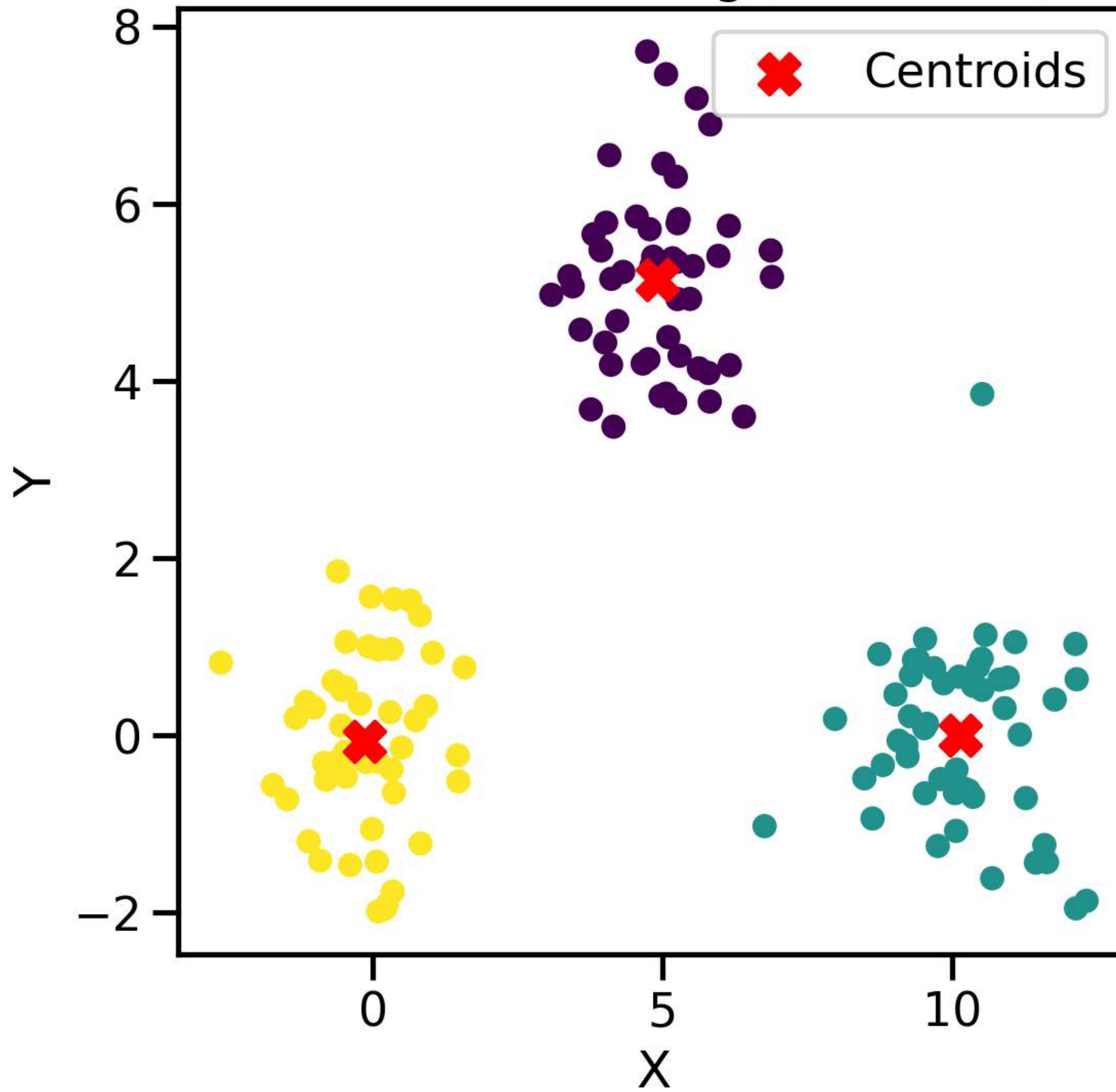
- Group cells by their similarity in gene expressions
- K = number of clusters (cell types) you aim to identify (based on prior knowledge or trial and error)
- Procedure: Repeat until convergence
 - 1. compute cluster centroids;
 - 2. assign cells to the cluster with the closest centroid.



A demo of K-Means on the board

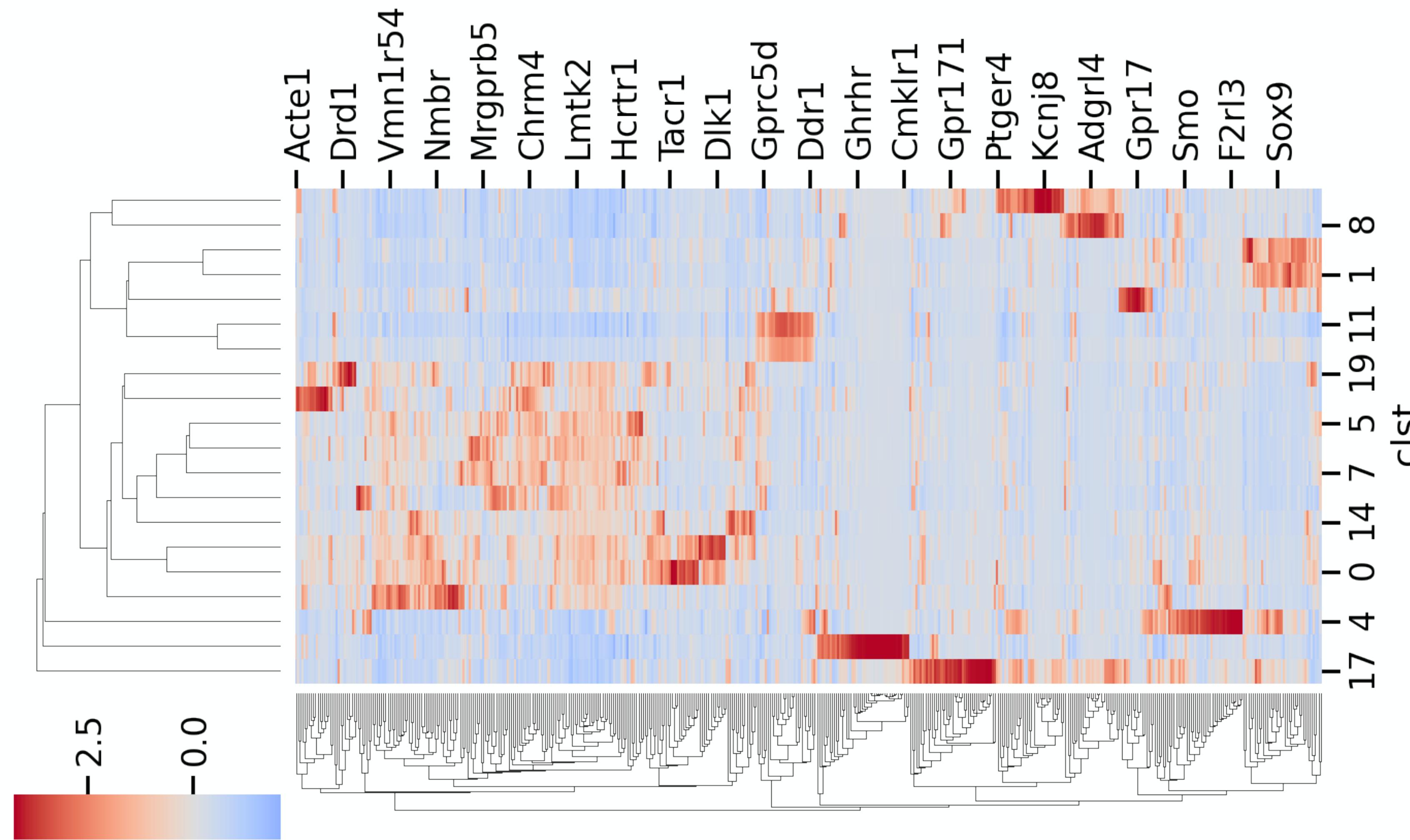
- Data points: [9, 10, 11; -1, 0, 1; -9, -10, -11]
- Apply K-means with K=3
- Random initiate the centroid locations (The month of your birthday from three of you)
- Iterate
 - Assigning data points to the centroid
 - Update centroid

K-Means Clustering (3 clusters)



Bonus exercise 1: Cell-type centroid gene expression profiles

- Often also called “pseudo-bulk” RNA-seq profiles due to historical reason.



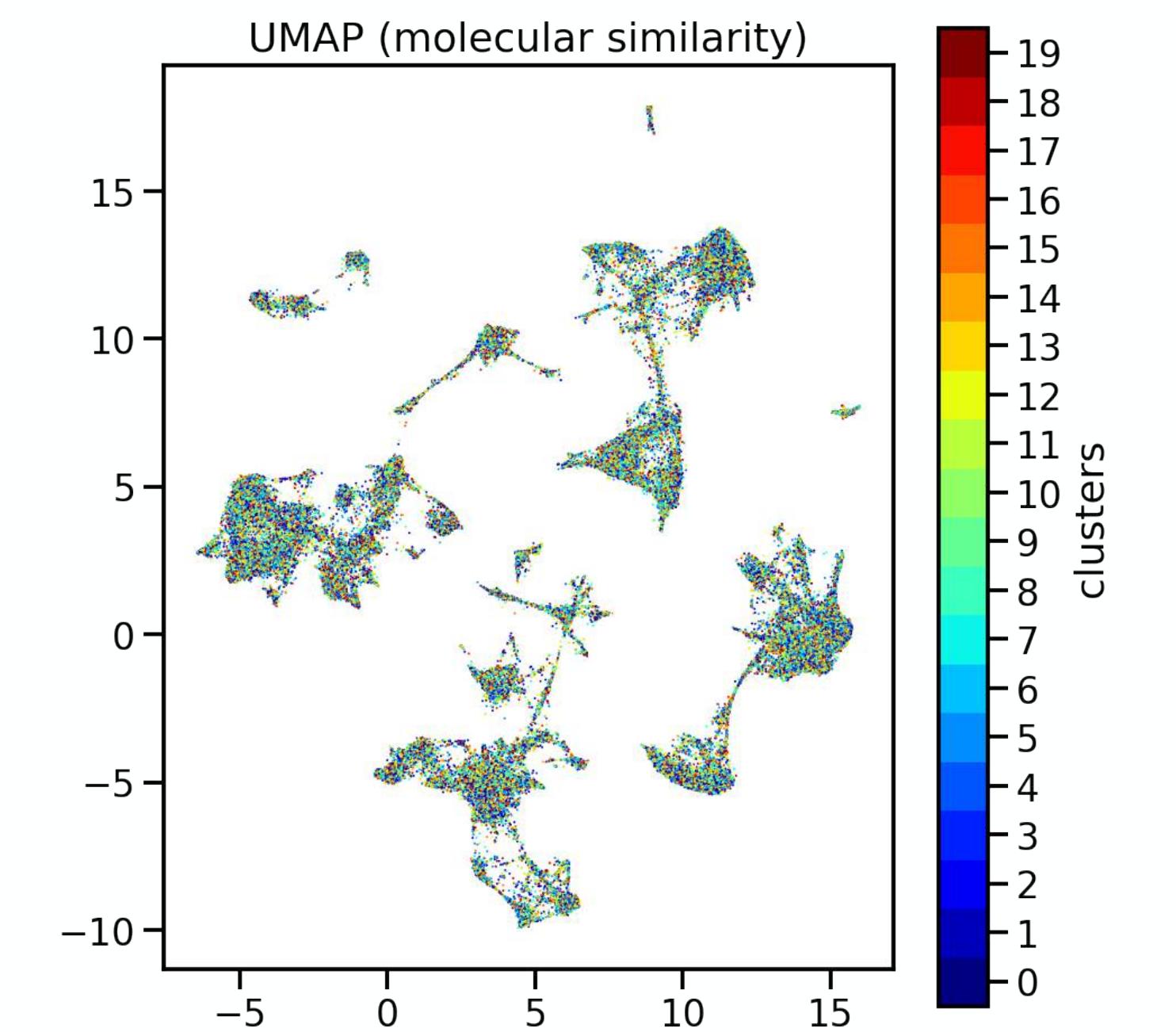
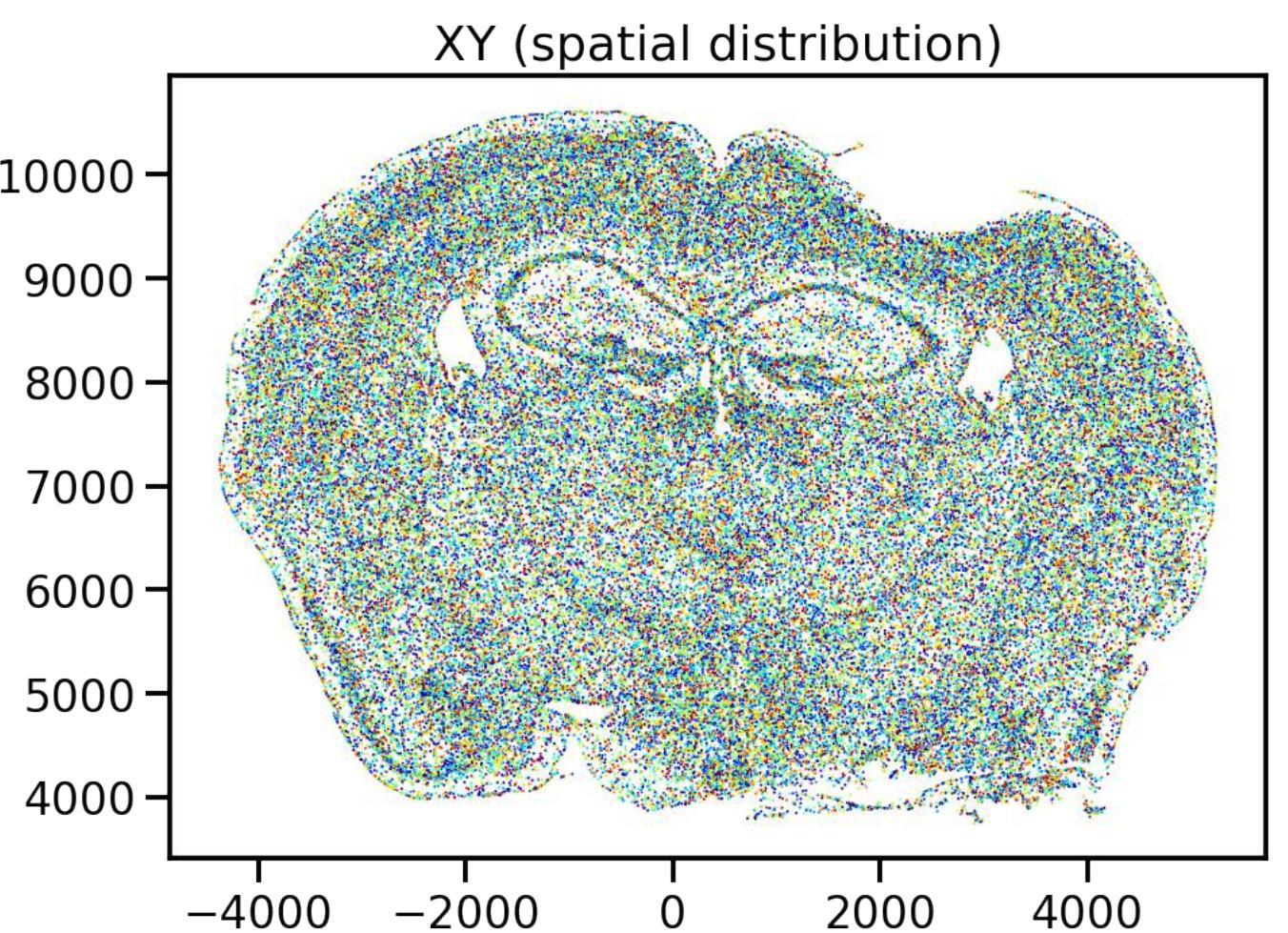
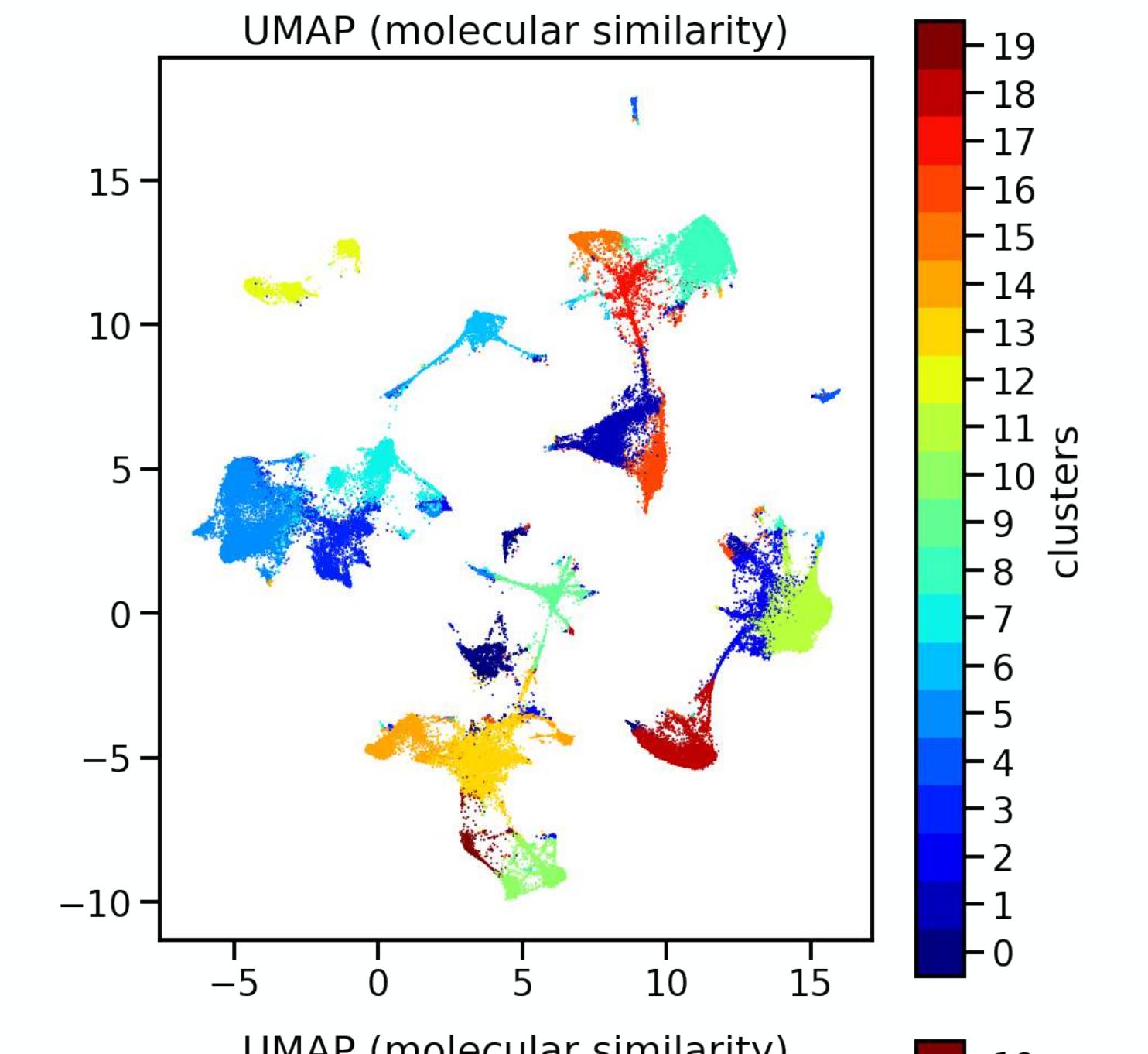
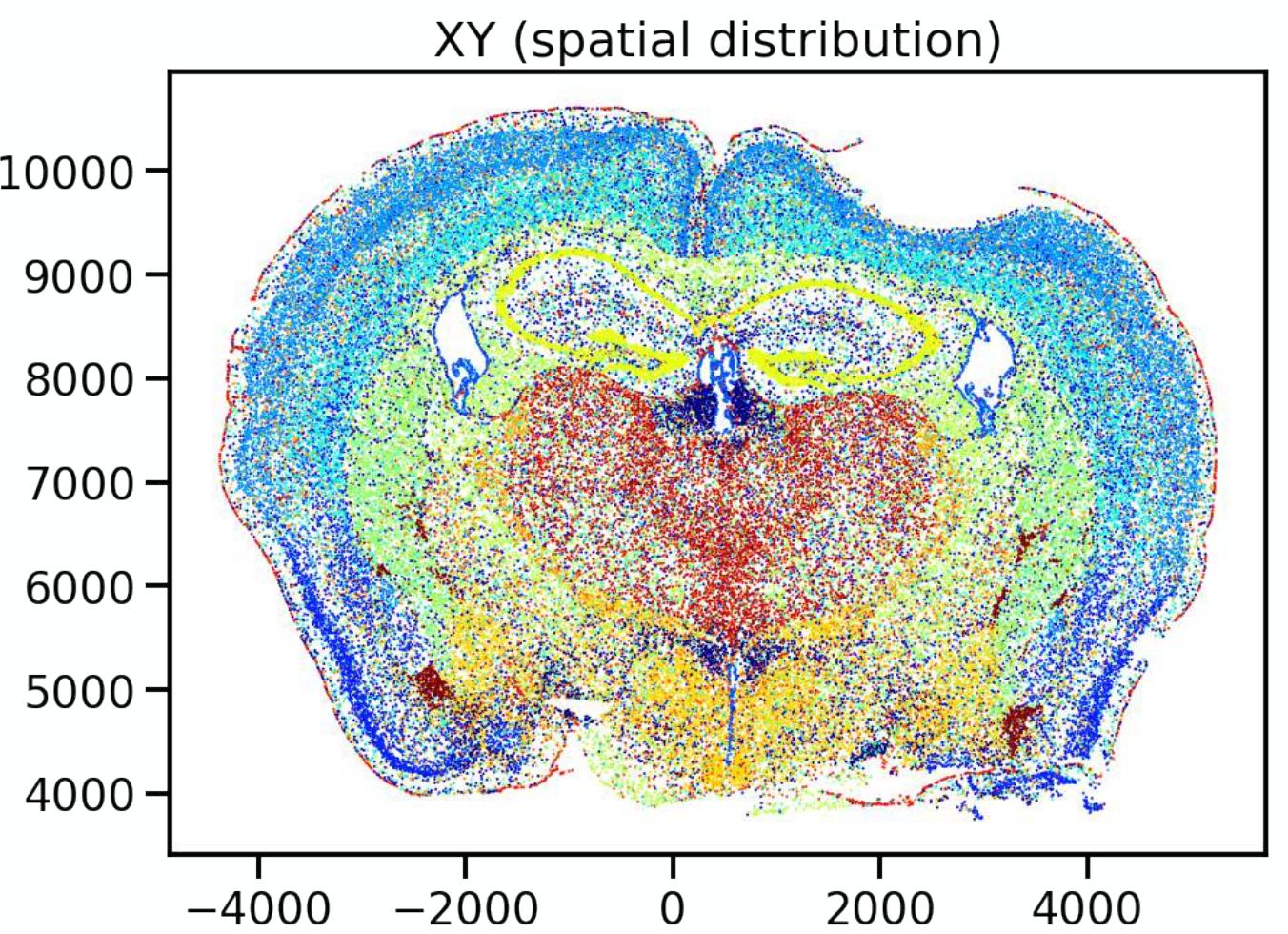
below is the class evaluation form

Exercise 5 (Bonus exercise 2): Spatial neighborhood enrichment analysis

- Are some cell types close to some other cell types? Or are cell types distributed randomly in space?
- How do we get at this question?
- Related paper: Fang et al. 2022 Science – Figure 4

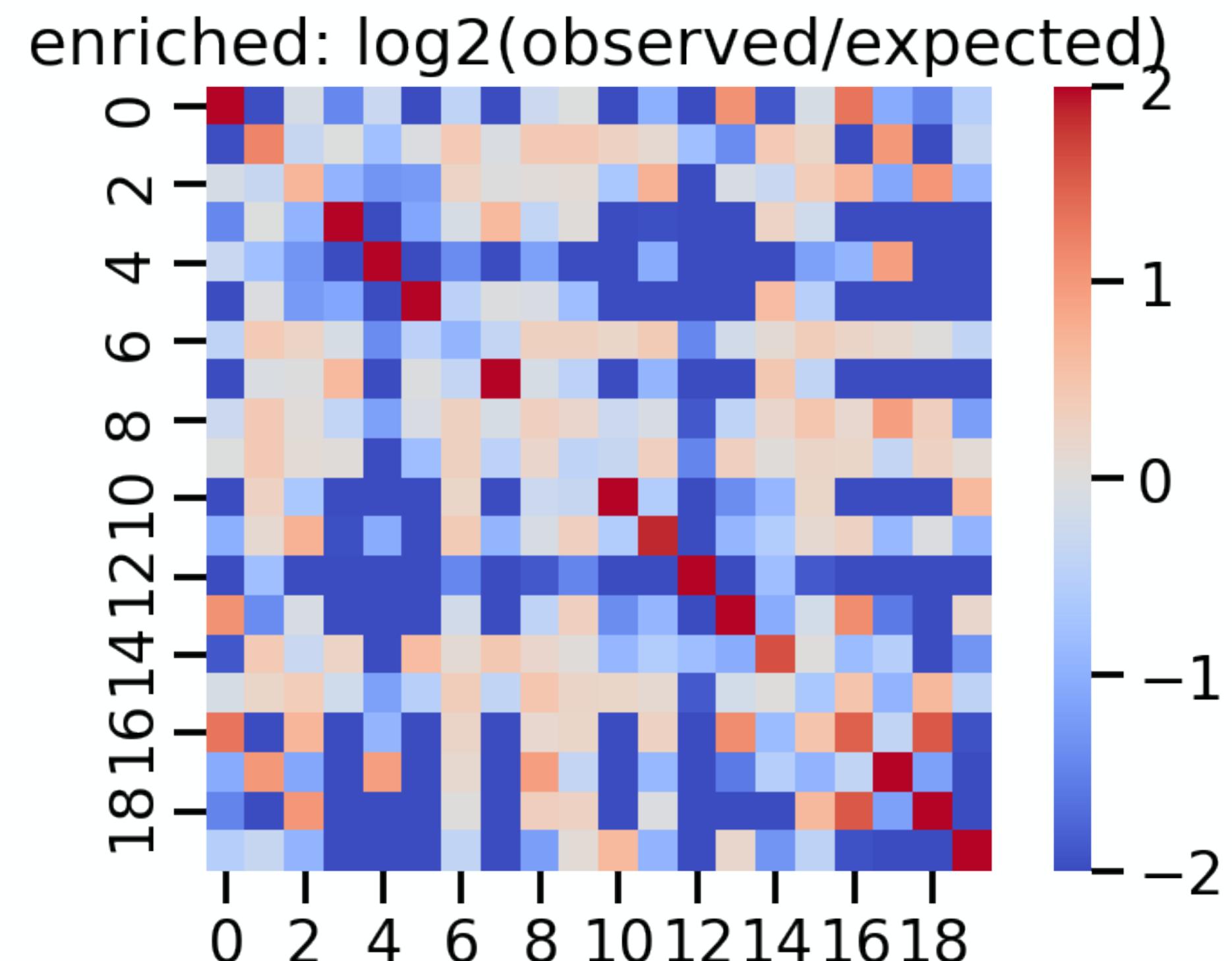
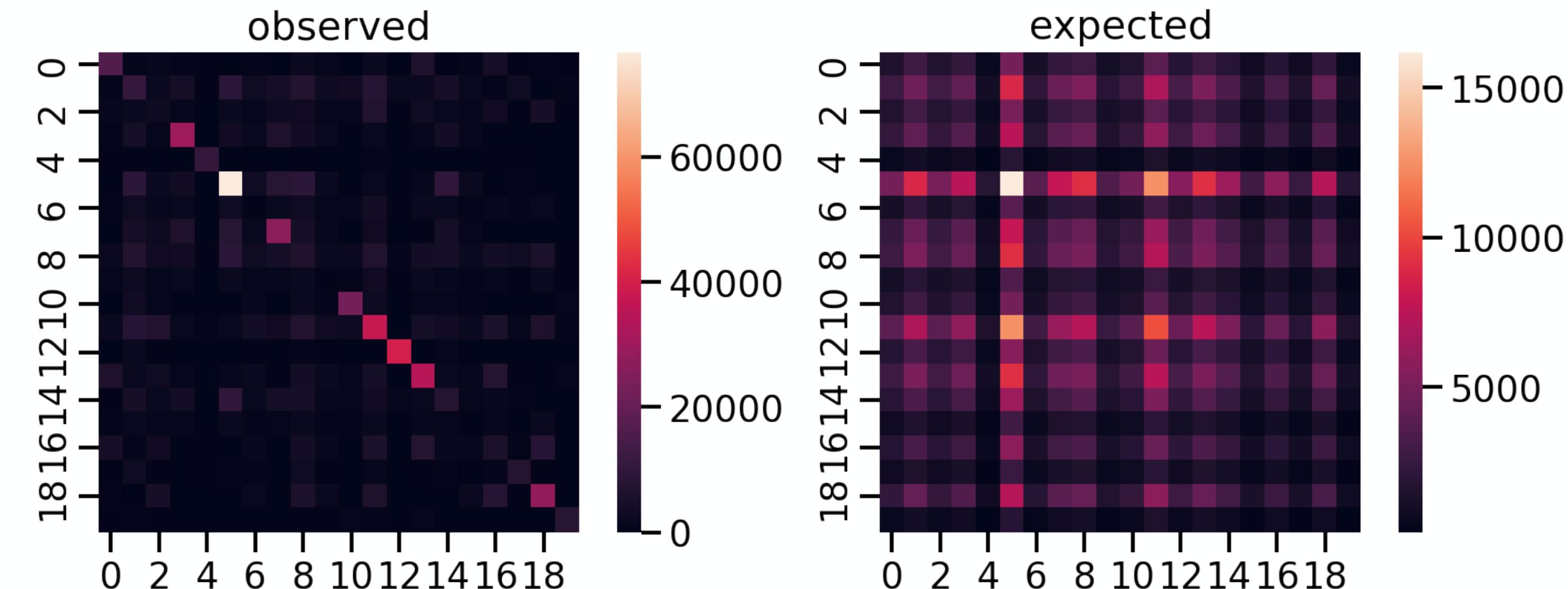
Data shuffling is a powerful trick to test hypothesis

- Count for each cell-cell spatial neighbor, which cell types do they belong to?
- Compare this with shuffled cell type labels
- This is a simple way to generate non-trivial null hypothesis



Data shuffling is a powerful trick to test hypothesis

- Count for each cell-type pair, how many neighbors
- Compare this with shuffled cell type labels
- This is a simple way to generate non-trivial null hypothesis



Exercise 6: Squidpy

- Standardized pipeline
- Well maintained
- Fast evolving
- Many tutorials
- <https://squidpy.readthedocs.io/en/stable/notebooks/tutorials/index.html>



Tutorials

This section contains various tutorials showcasing spatial molecular data analysis with `squidpy`.

Analysis of spatial datasets using `squidpy`

This section contains tutorials showcasing core Squidpy functionalities by applying them to a diverse set of different spatial datasets.

- Analyze Imaging Mass Cytometry data
- Analyze seqFISH data
- Analyze Slide-seqV2 data
- Analyze Visium fluorescence data
- Analyze Visium H&E data
- Analyze Merfish data
- Analyze MIBI-TOF image data
- Analyze 4i data
- Analyze Nanostring data
- Analyze Vizgen data
- Vizgen Mouse Liver Squidpy Vignette
- Analyze Xenium data

Thank you!