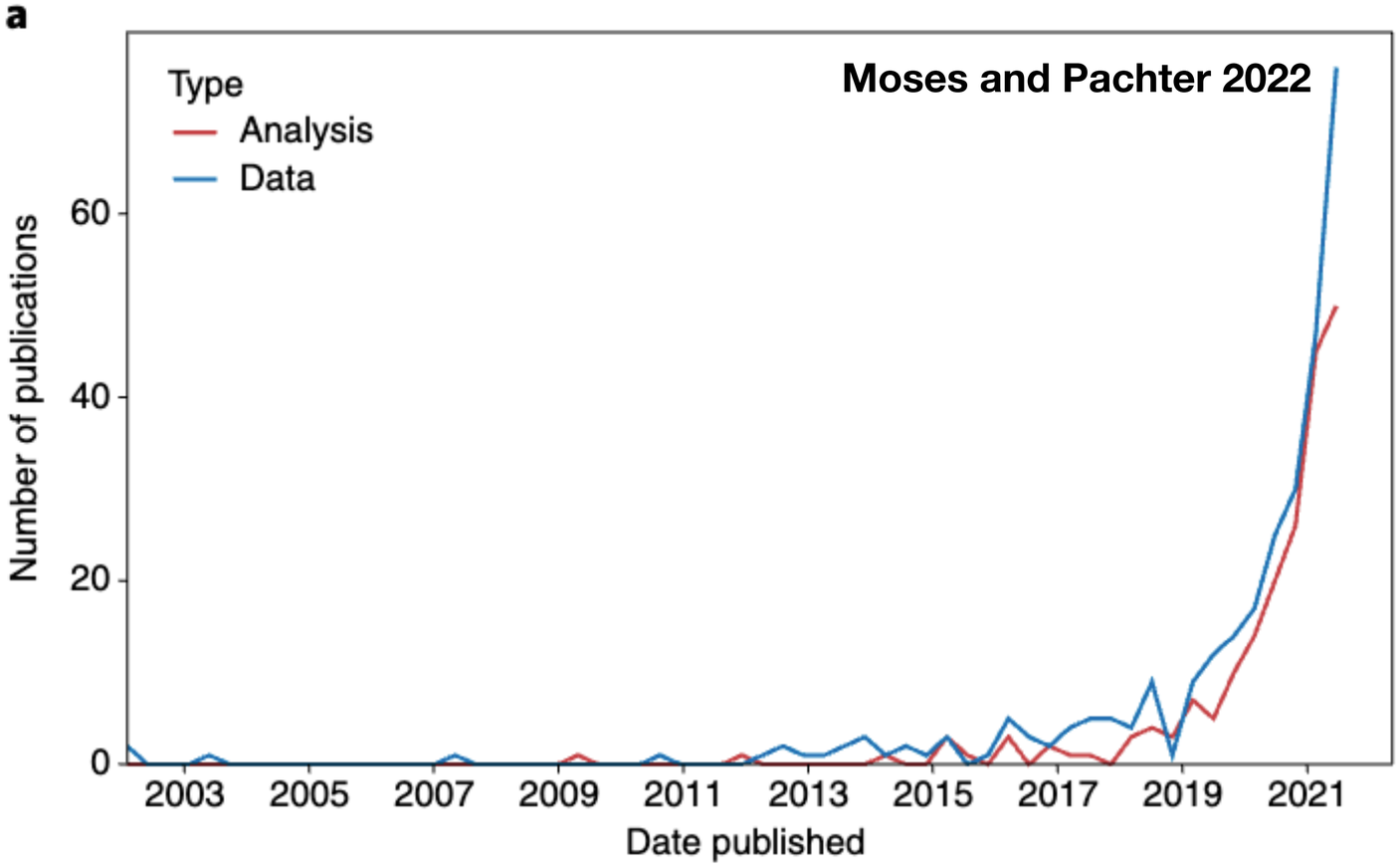
UCLA Collaboratory workshop (W31)

Spatial Transcriptomics

March 15, 2023 (Day 1)

Fangming Xie

An ~~introduction~~ invitation to spatial transcriptomics (ST)



- A rapidly developing field started from ~2015.
- Everything learned here will soon be obsolete?
- *“go for the messes — that's where the action is.”*

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

Published: 27 November 2003

Scientist

Four golden lessons

Steven Weinberg

Nature **426**, 389 (2003) | Cite this arti

Day 1: Overview

- **Talk**: background and concepts
- **Reading sessions**: reading and discussing research papers
- **Mini-quiz**: for fun and discussion

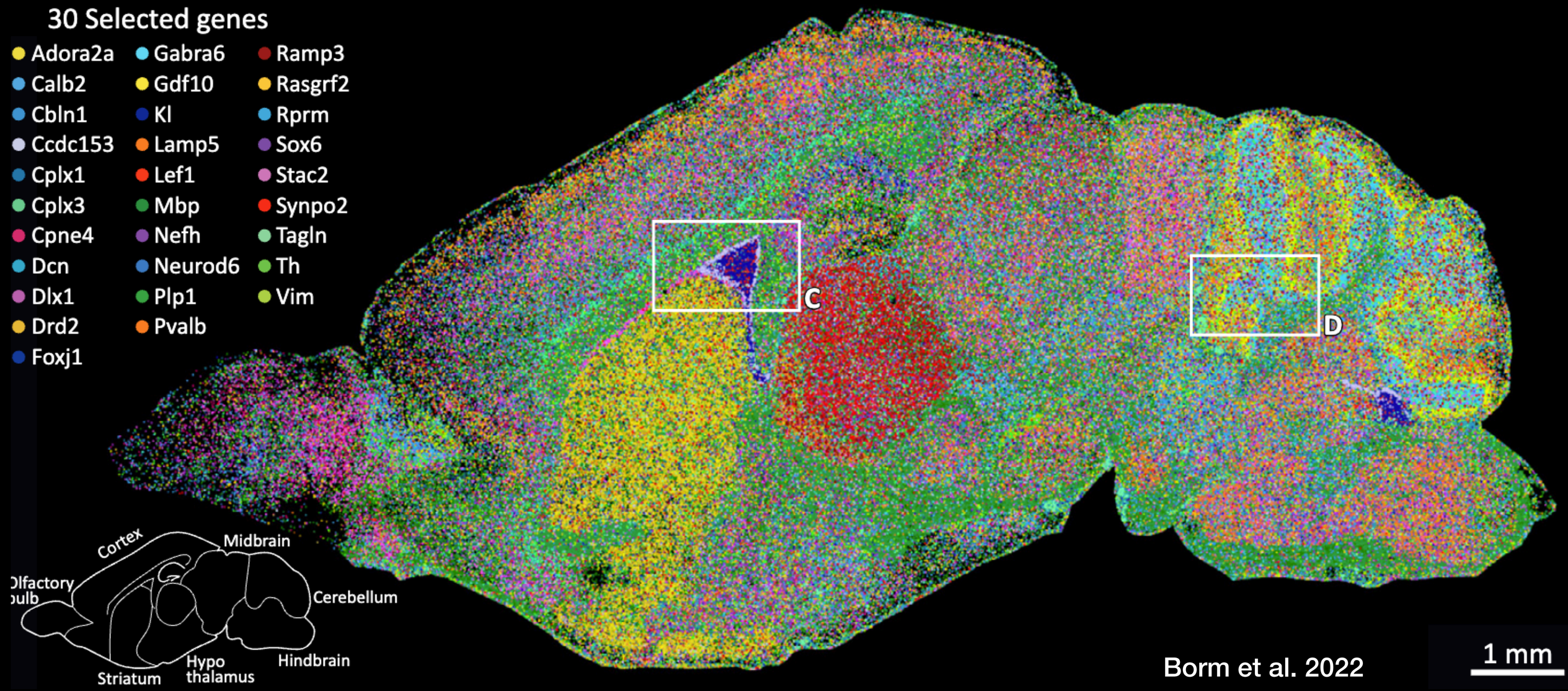
Day 2: Coding bootcamp

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

- **Grading**: Attendance, Take-home quiz and Homework (due on Friday 03/24)
- **Asking questions** are strongly encouraged — about anything at any time
- **A QCBio Collaboratory T-shirt** will be distributed tomorrow!

What is spatial transcriptomics?

- Technologies that make *transcripts (RNA)* seen, while preserving their spatial information in the cell or tissue, with high-throughput (many cells and many genes).



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?

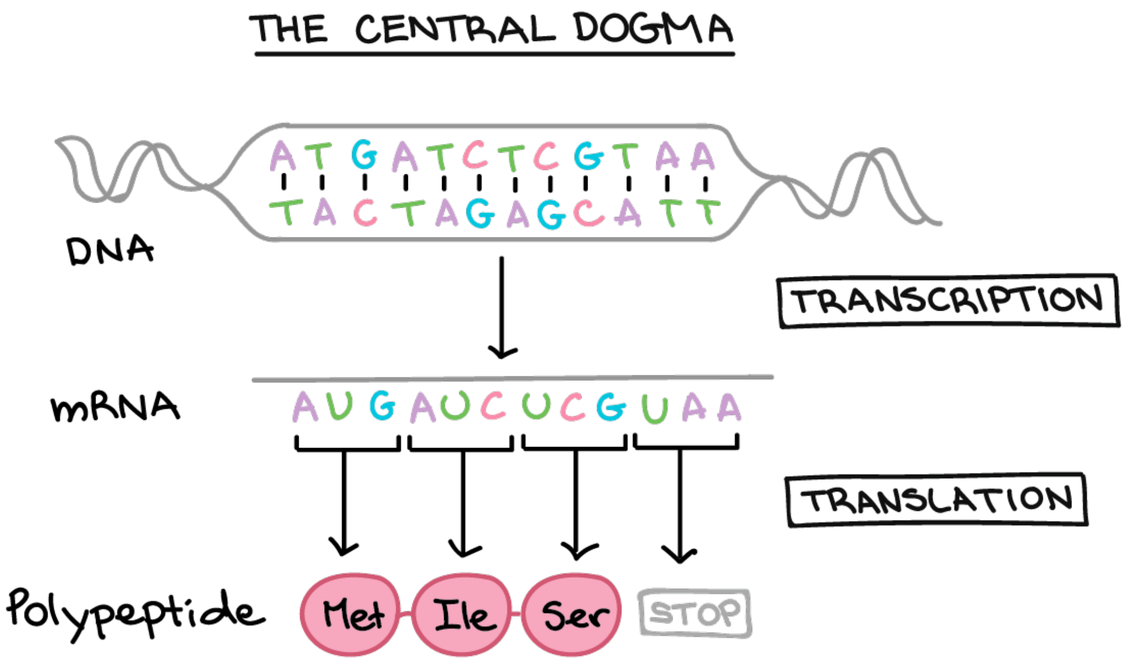
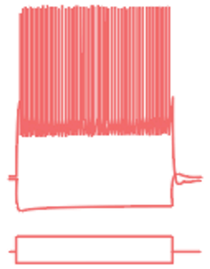
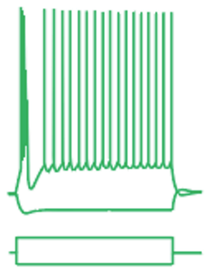


Figure from Khan Academy

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

Drawing of cortical neurons, Cajal, 1899

- Golgi's methods randomly stain brain cells, revealed diverse morphologies but with no molecular specificity.
- Molecular diversity underlies the morphological, anatomical, and functional diversity of brain cells.

	Morphology	Physiology	Cell body location in cortex	Connectivity	Molecular signature
PVALB+ Basket cell			All layers	Perisomatic projection	PVALB+ LHX6+ ...
RBP4+ Thick-tufted cell			Layer 5b	Sub-cortical projection	RBP4+ SLC17A7+ ...

Zane and Sanes, 2017

- **How do molecules give rise to diverse cell types that organize themselves to fulfill biologic functions?**
  - *Neuroscience*: what is the molecular logic underlying neural circuits?
  - *Developmental biology*: how cells divide and self-organize to form tissue?
  - *Cancer biology*: what are the biomarkers of tumor heterogeneity?
- It is not enough to understand molecules themselves, we need to understand how they ***organize*** to serve functions.
  - "What I cannot ***see***, I do not understand."

Species	*Escherichia coli*	*Gallus gallus*	*Homo sapiens*	*Daphnia pulex*	*Oryza sativa*
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

Does the genome specify everything?

— an information-theory perspective

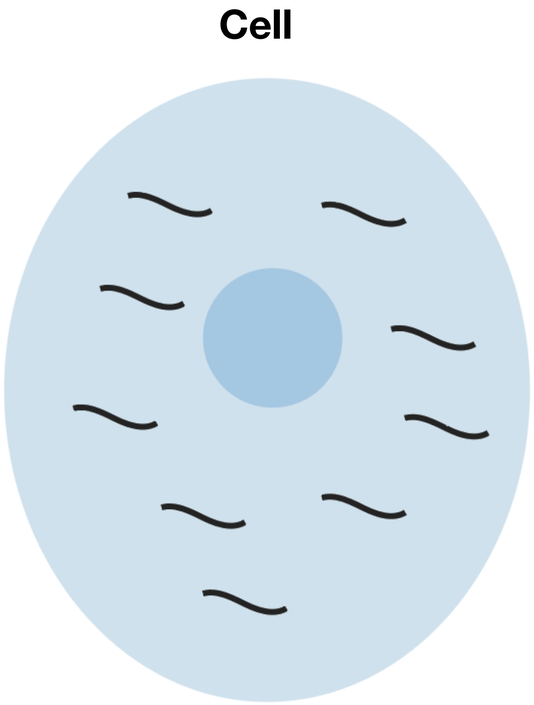
The wiring problem and the “genomic bottleneck”

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

We need ways to see how organisms organize their molecules!

Zador 2019, *Nat. Commun.* **Koulakov, Shuvaev, Lachi & Zador 2022, *bioRxiv***

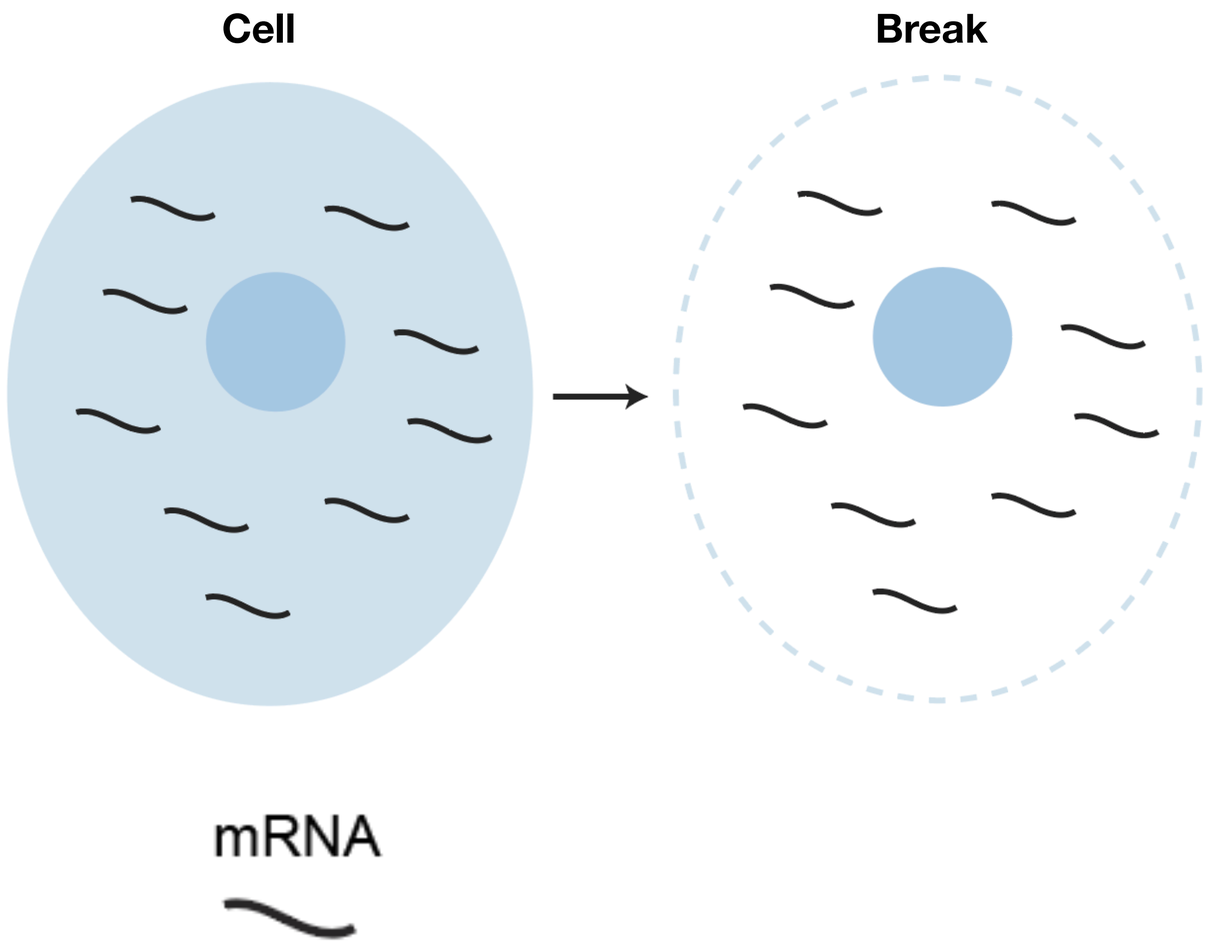
How can mRNA be *seen*?



mRNA

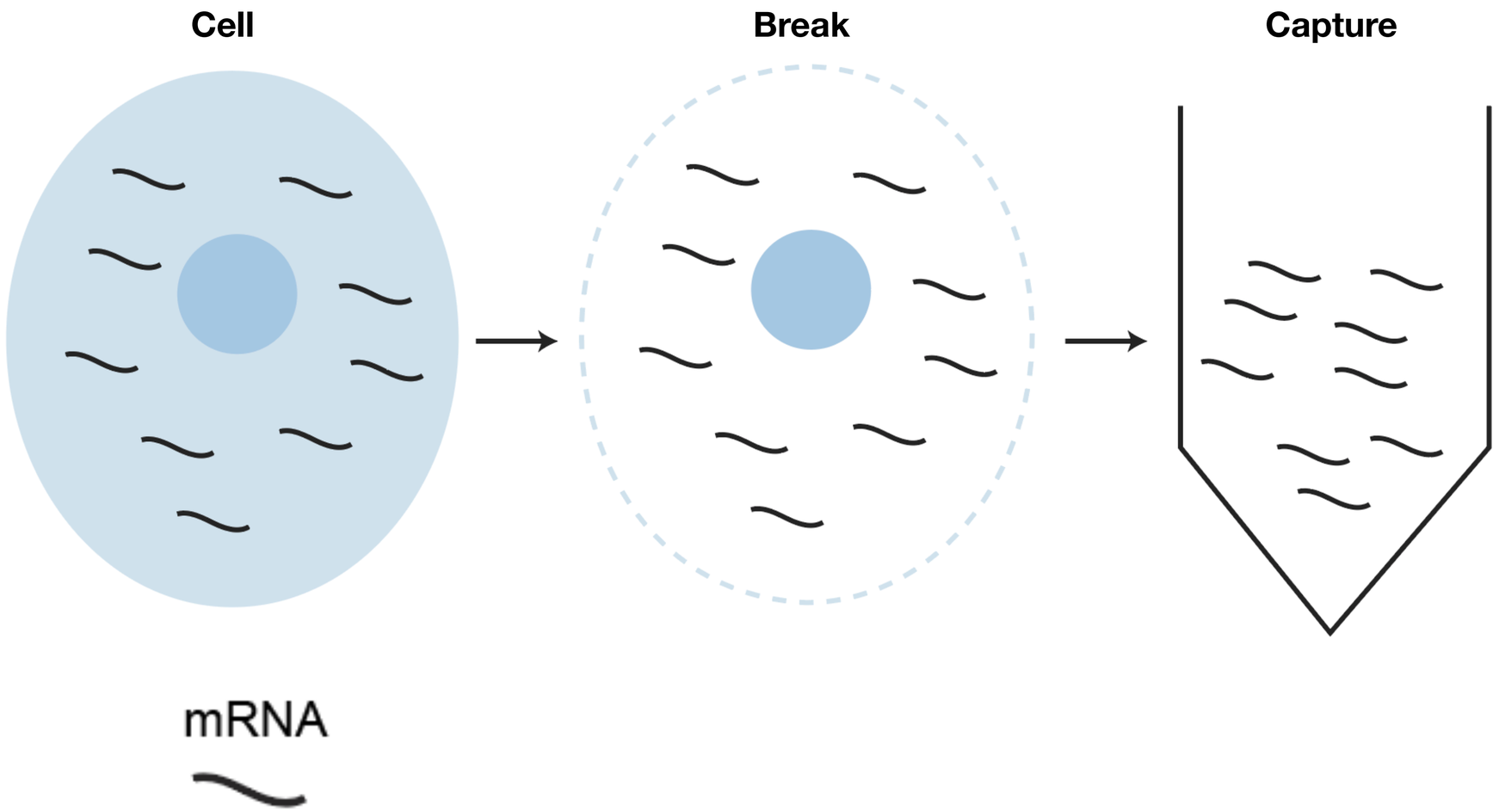
How can mRNA be *seen*?

Method 1: Sequencing



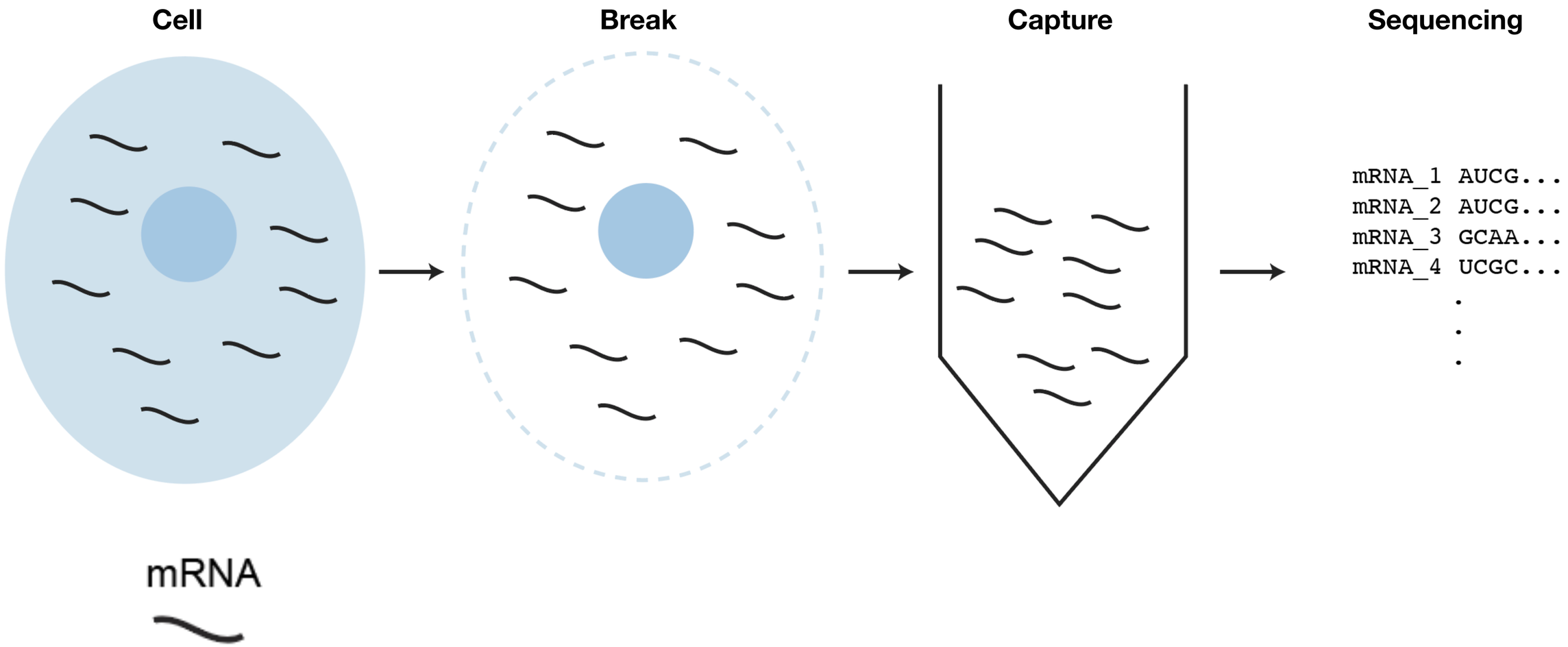
How can mRNA be *seen*?

Method 1: Sequencing



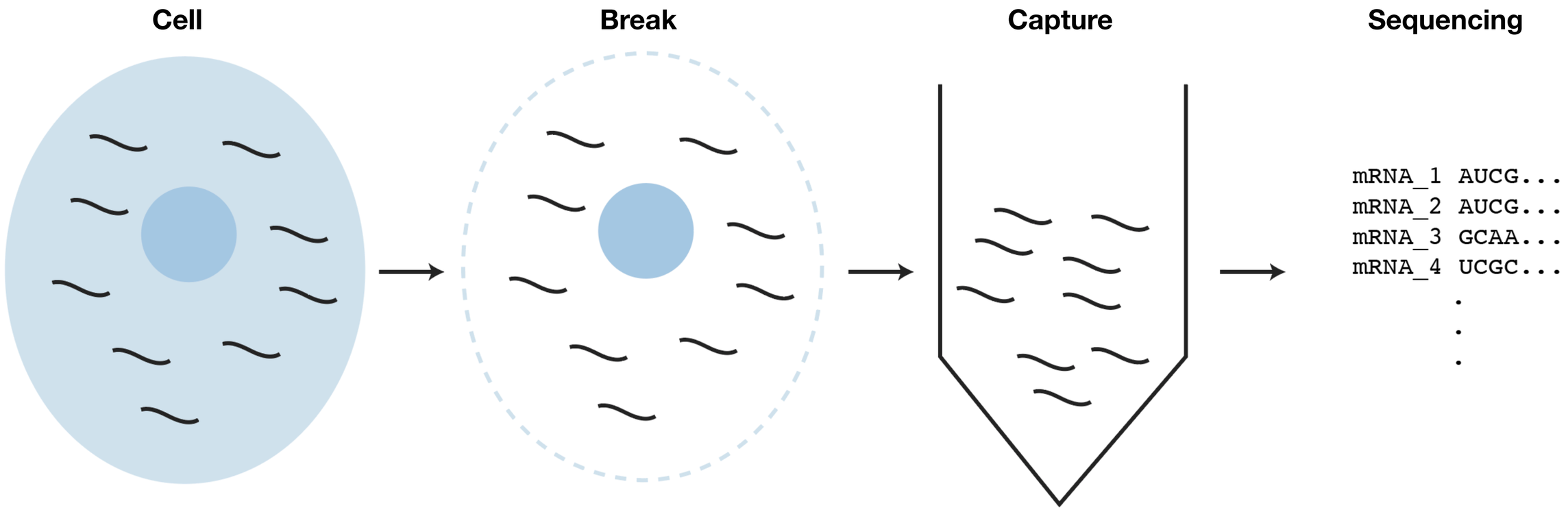
How can mRNA be *seen*?

Method 1: Sequencing



How can mRNA be *seen*?

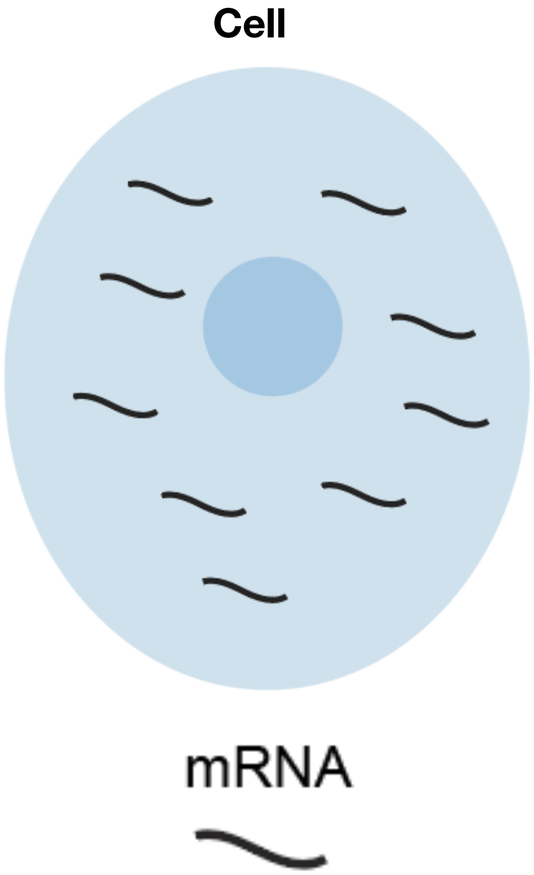
Method 1: Sequencing



- Captured all types of mRNAs (all genes)
- Lost spatial information

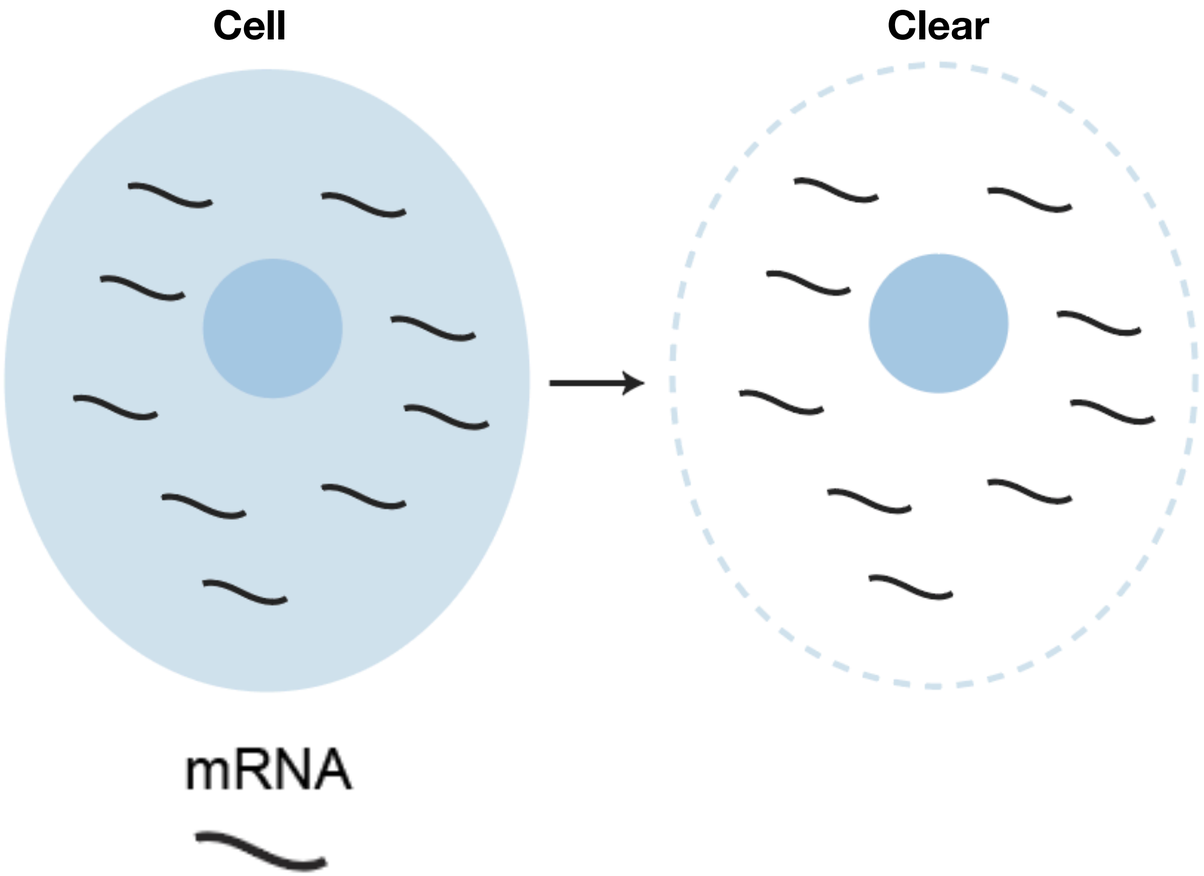
How can mRNA be *seen*?

Method 2: Staining and Imaging



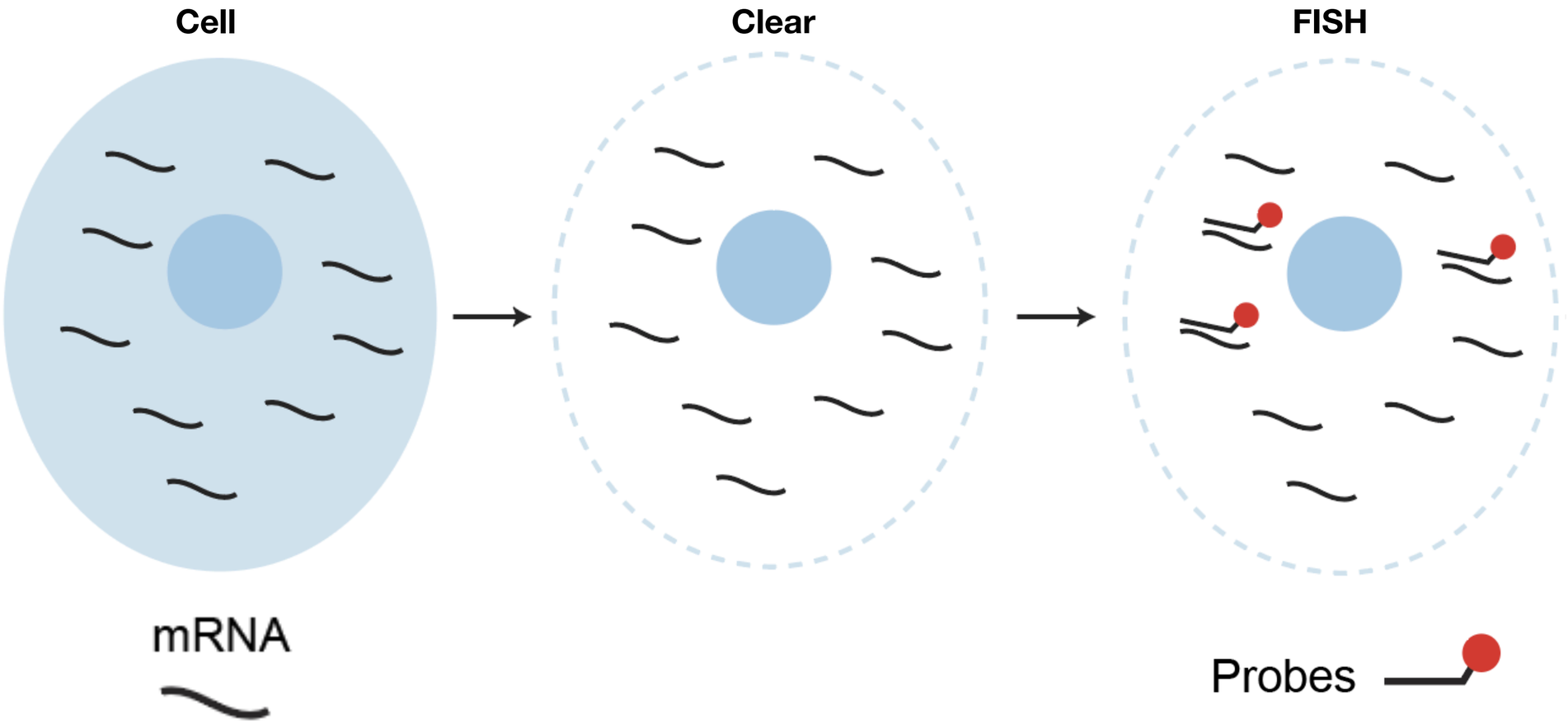
How can mRNA be *seen*?

Method 2: Staining and Imaging



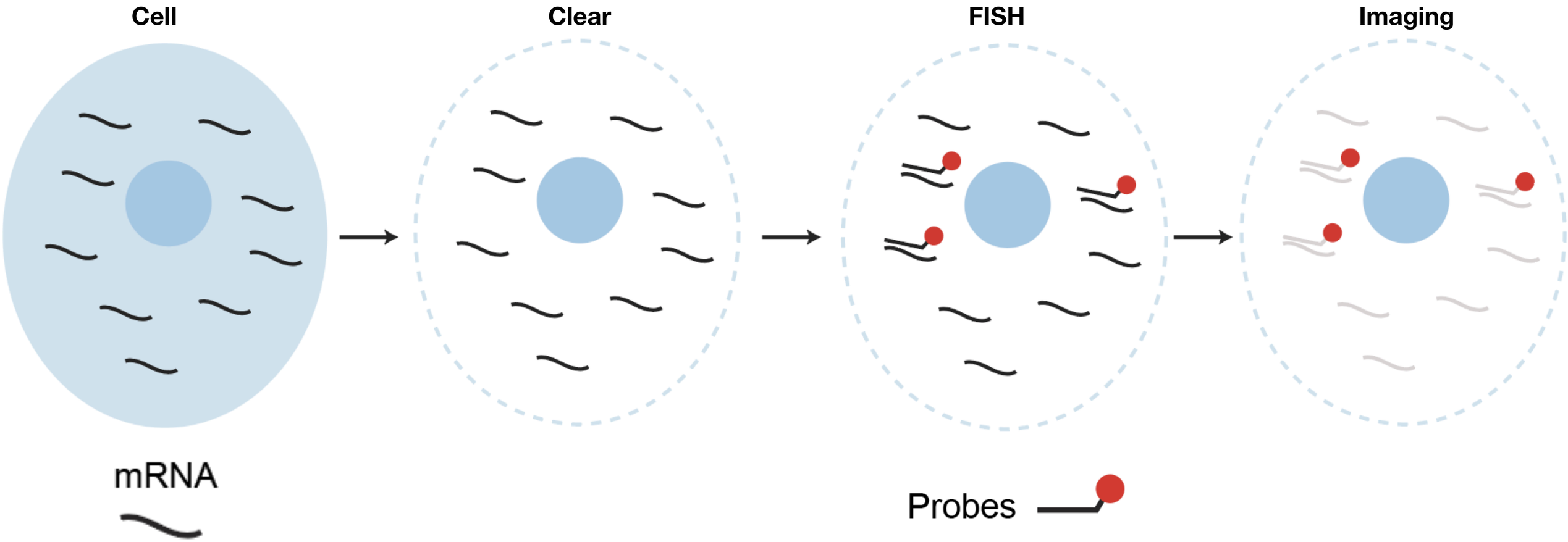
How can mRNA be *seen*?

Method 2: Staining and Imaging



How can mRNA be *seen*?

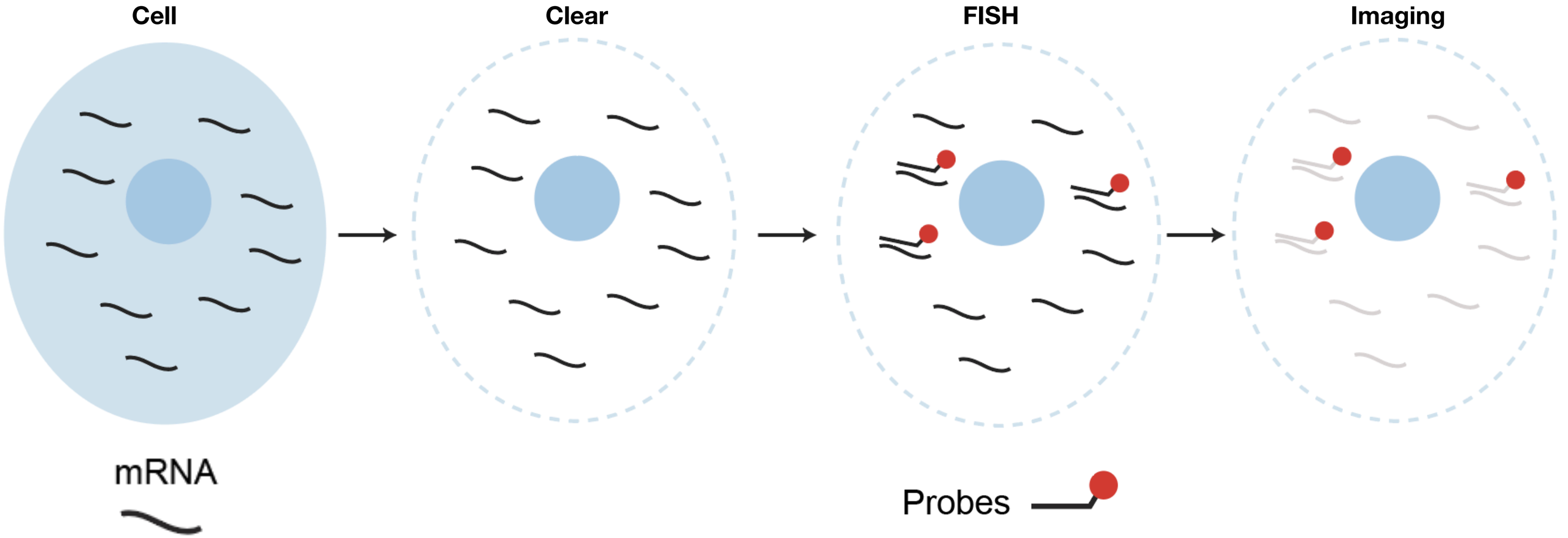
Method 2: Staining and Imaging



How can mRNA be *seen*?

Method 2: Staining and Imaging

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

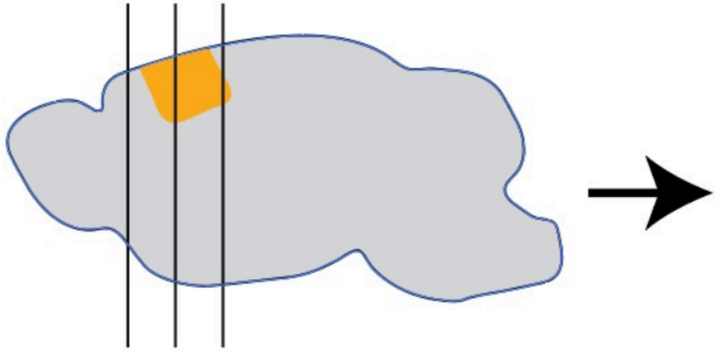


Precursors of spatial transcriptomics

- Single-cell transcriptome profiling (scRNA-seq):
  - All (most) genes, no spatial
- In-situ staining (smFISH):
  - One (few) gene, spatial

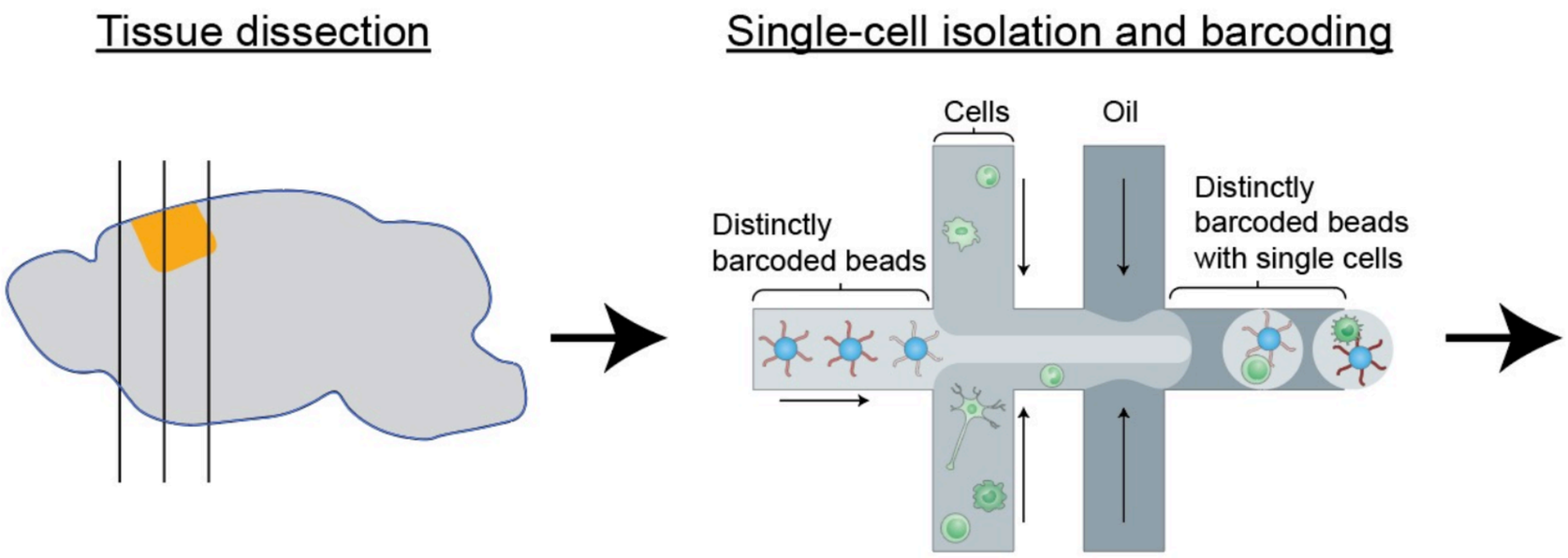
Single-cell RNA-sequencing (scRNA-seq)

Tissue dissection



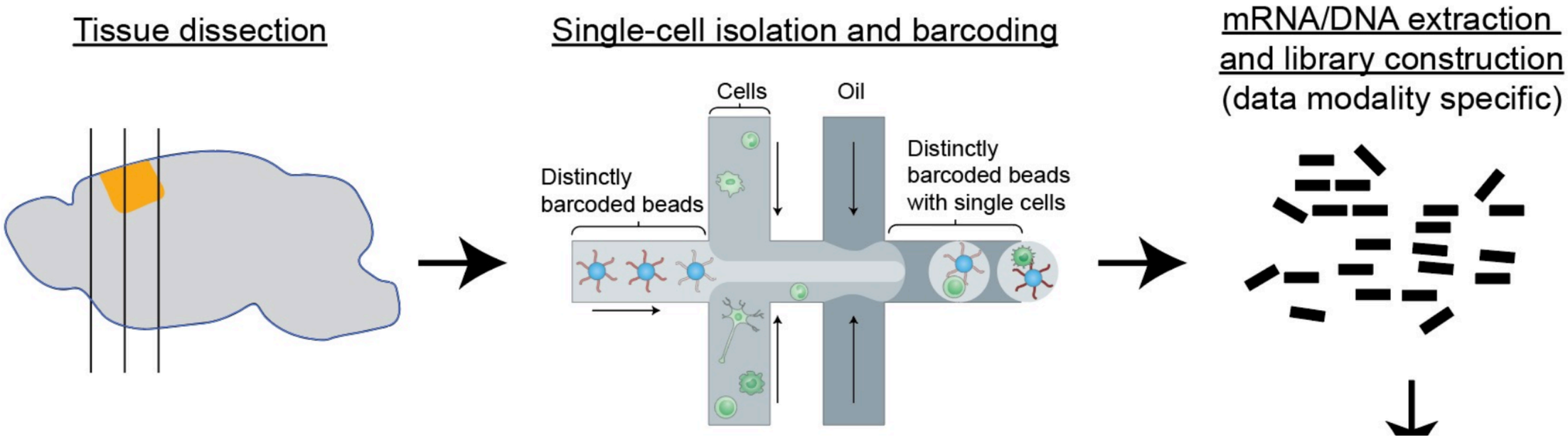
Adapted in part from Zane and Sane, 2017

Single-cell RNA-sequencing (scRNA-seq)



Adapted in part from Zane and Sane, 2017

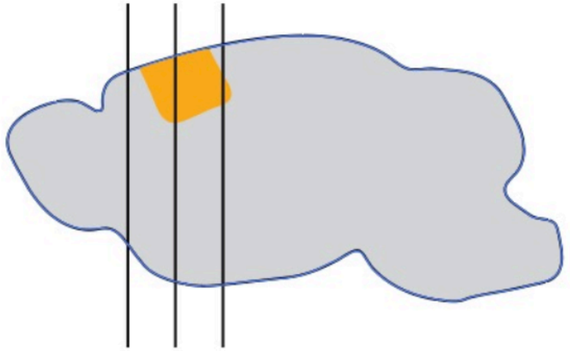
Single-cell RNA-sequencing (scRNA-seq)



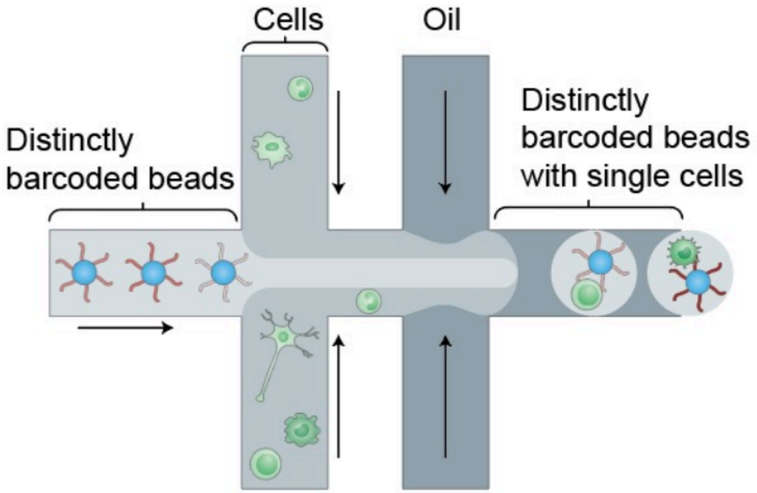
Adapted in part from Zane and Sane, 2017

Single-cell RNA-sequencing (scRNA-seq)

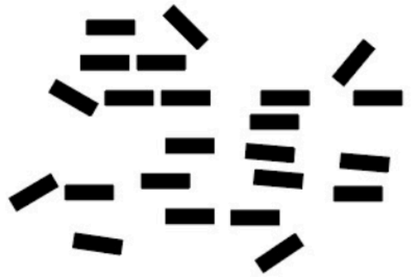
Tissue dissection



Single-cell isolation and barcoding



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



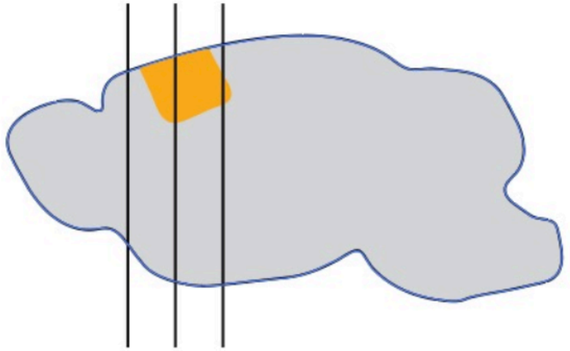
Sequencing

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

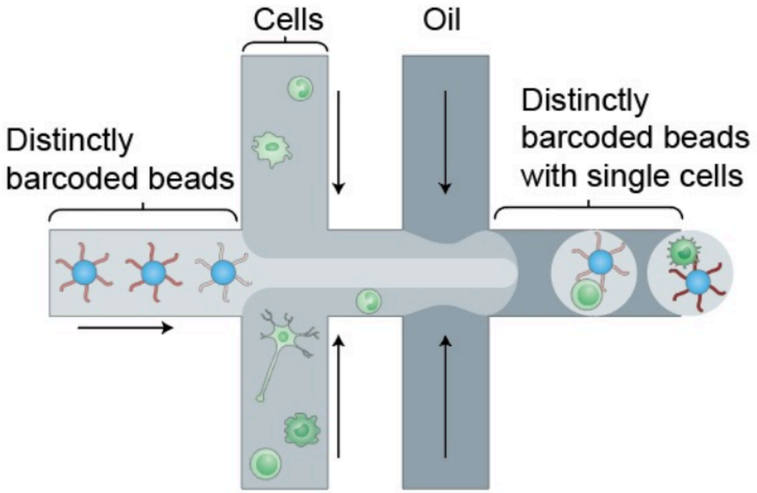
Adapted in part from Zane and Sane, 2017

Single-cell RNA-sequencing (scRNA-seq)

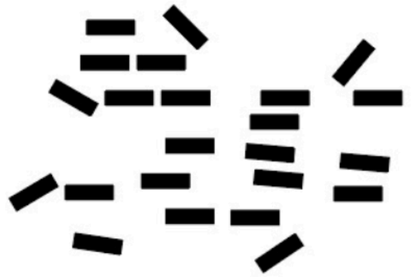
Tissue dissection



Single-cell isolation and barcoding



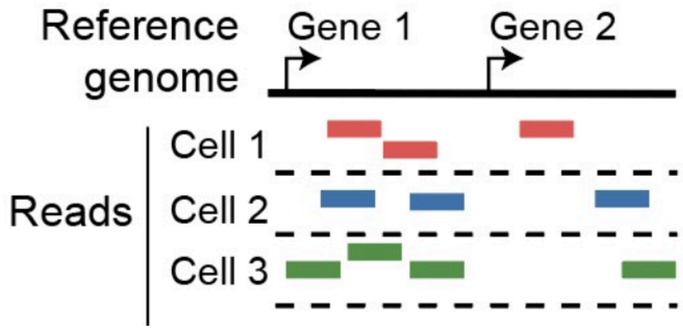
mRNA/DNA extraction and library construction (data modality specific)



Sequencing

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

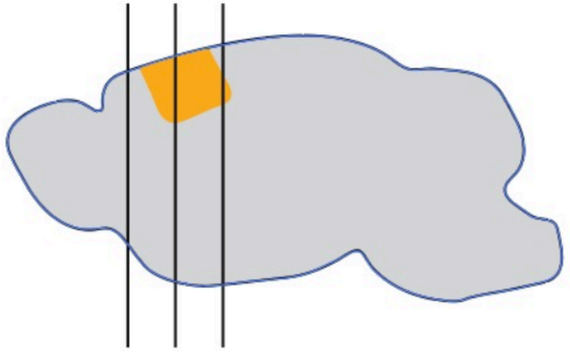
Mapping and Demultiplexing



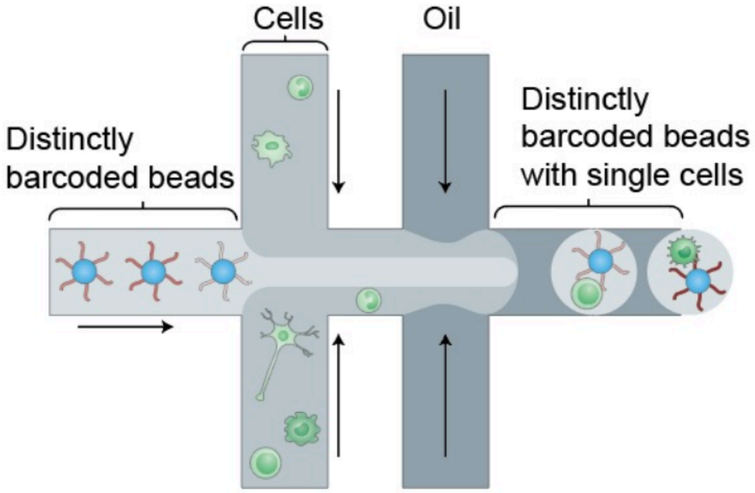
Adapted in part from Zane and Sane, 2017

Single-cell RNA-sequencing (scRNA-seq)

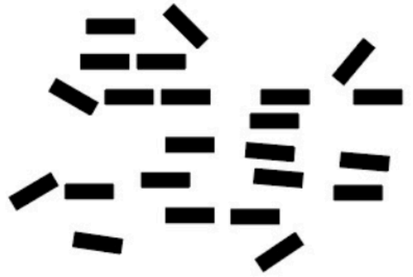
Tissue dissection



Single-cell isolation and barcoding



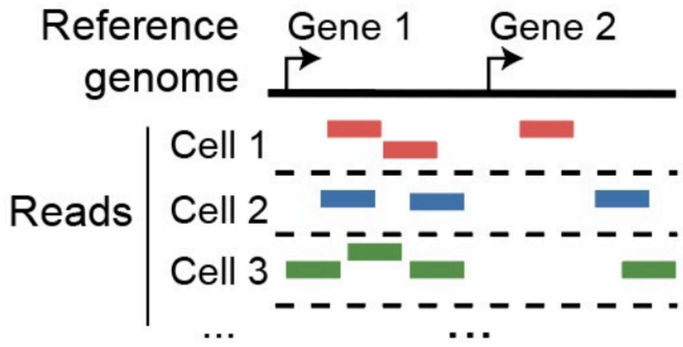
mRNA/DNA extraction and library construction (data modality specific)



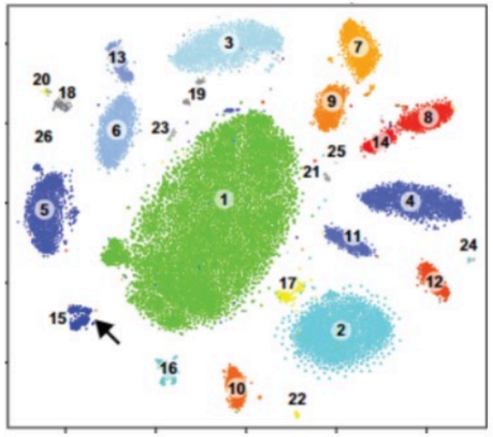
Sequencing

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

Mapping and Demultiplexing

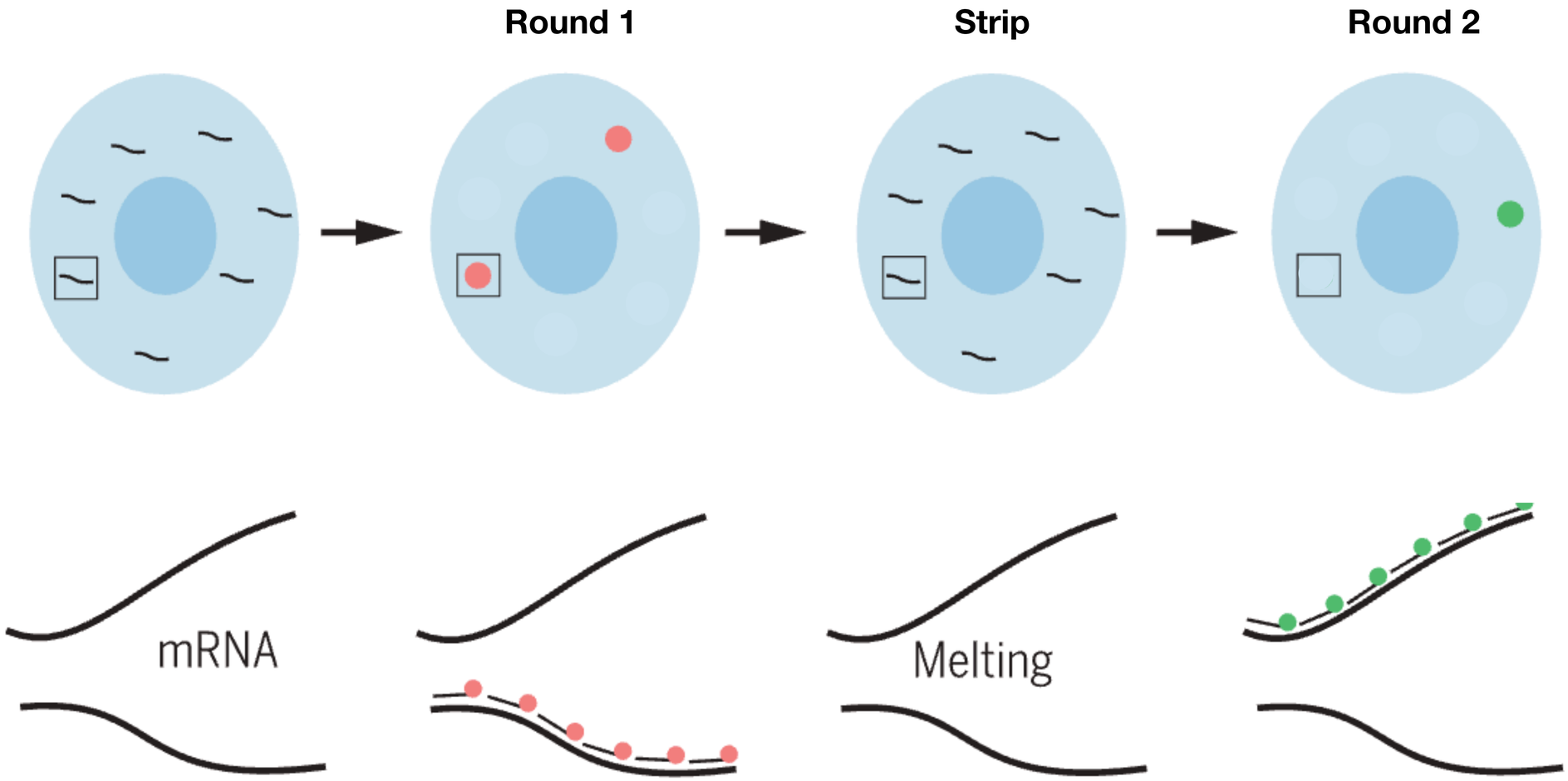


Downstream analysis
Clustering and visualization



Adapted in part from Zane and Sane, 2017

Single molecule Fluorescence In-Situ Hybridization (smFISH)



Lein et al. 2017

Spatial Transcriptomics bridges histology and genomics

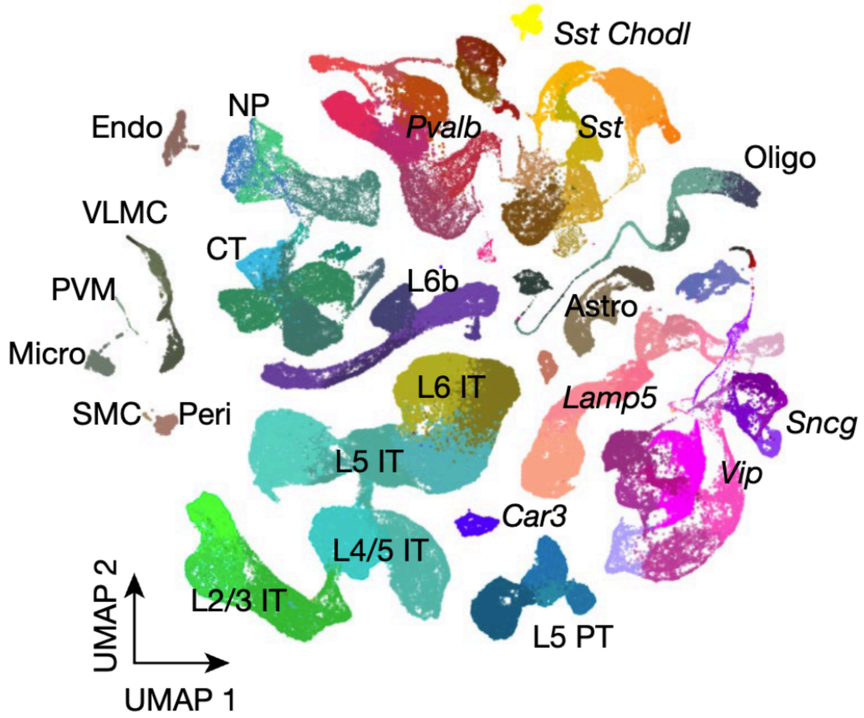
- Making histology profiling (staining) molecularly specific.
- Making genomics profiling (sequencing) spatially resolved.

Golgi Staining

Ramon y Cajal, 1899

+

scRNA-seq



Yao et al. 2021

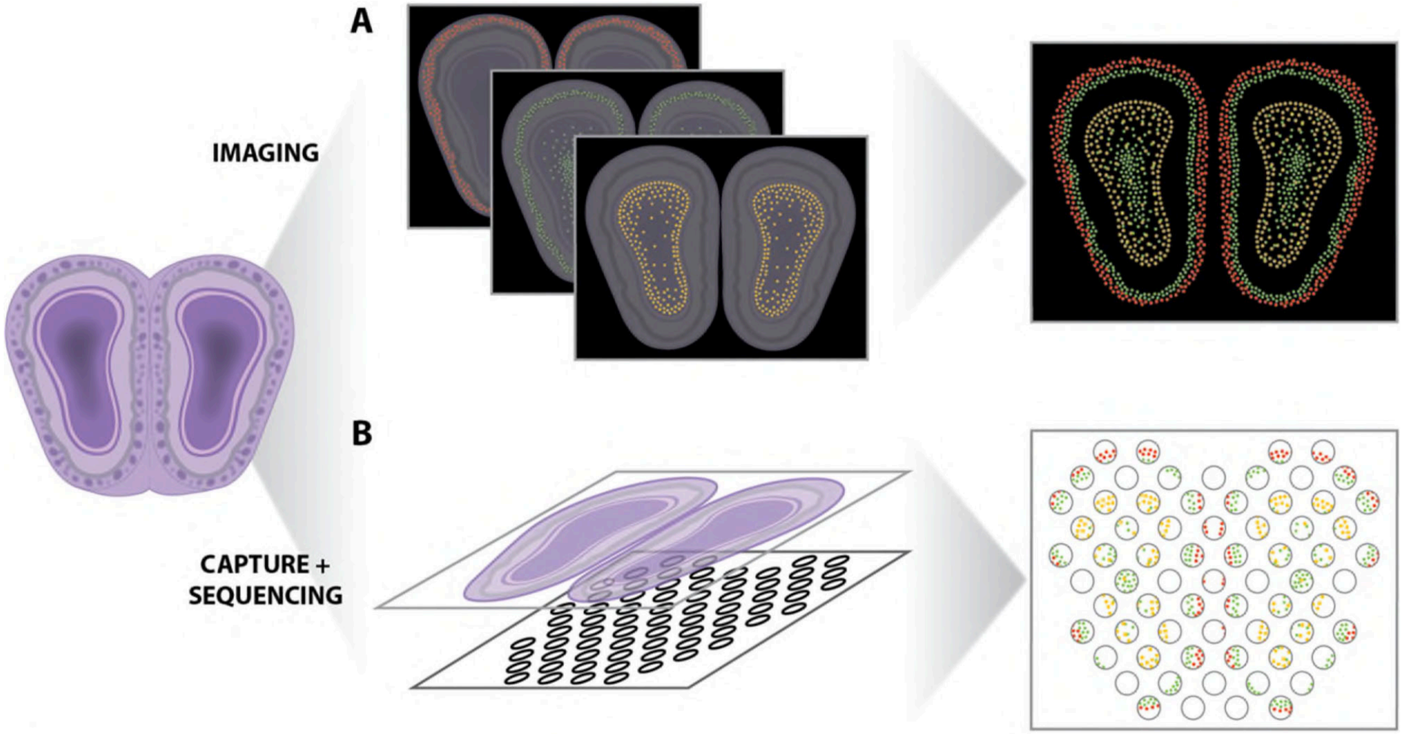
→

Spatial Transcriptomics

Michael Nunn, 2020

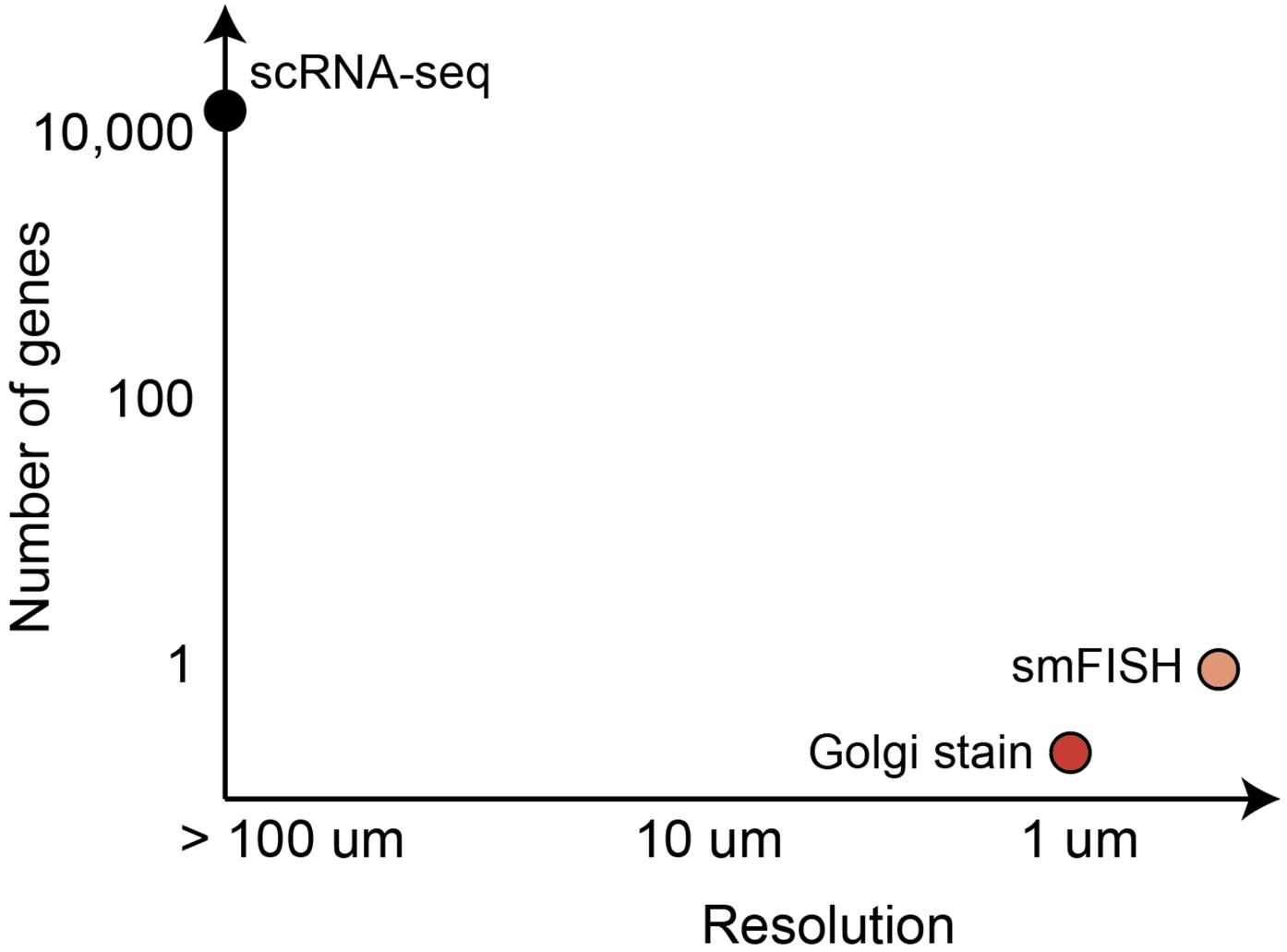
Two major branches of technologies: sequencing-based vs imaging-based assays

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

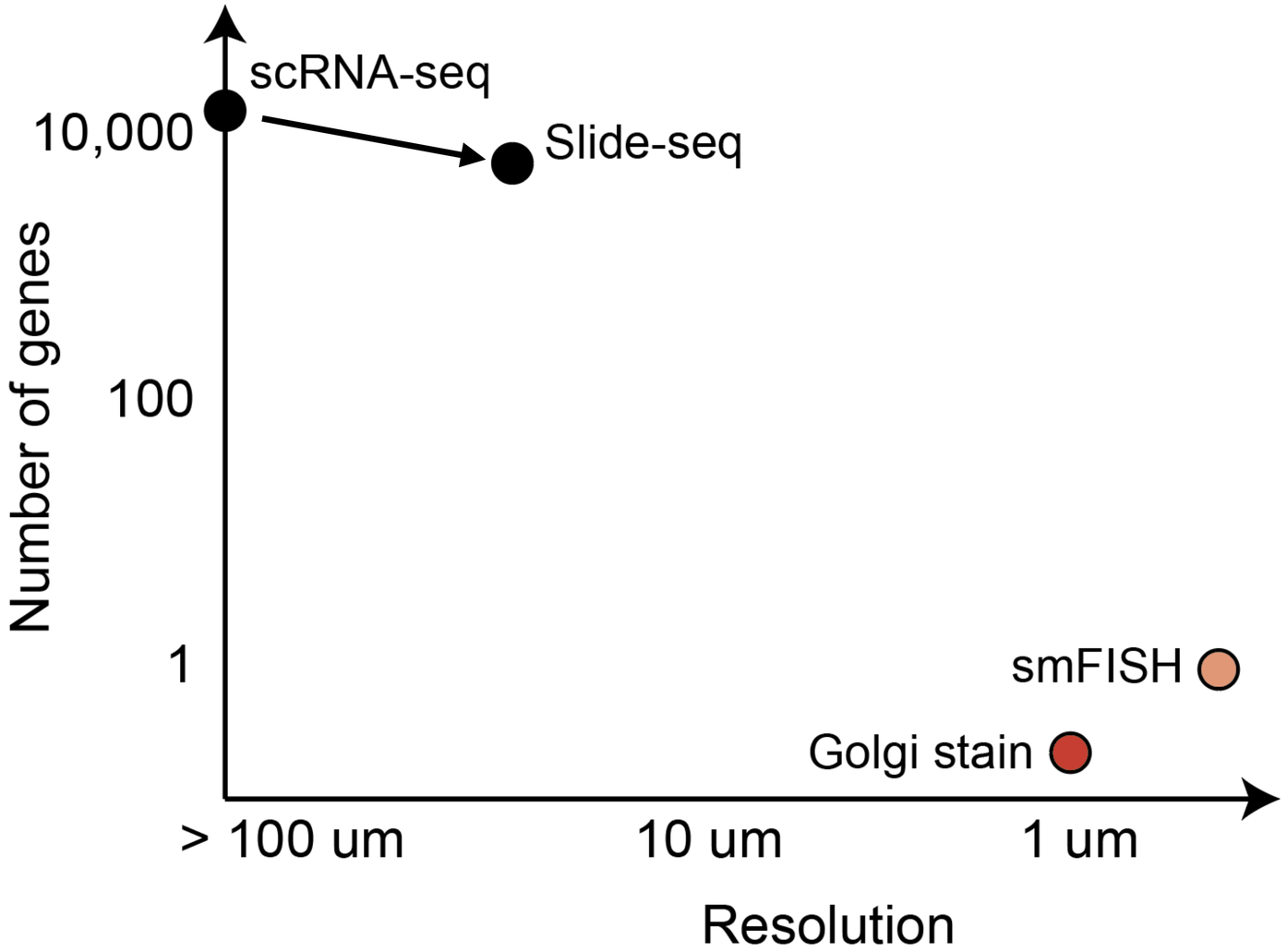


Atta and Fan, 2021, Nat Comm

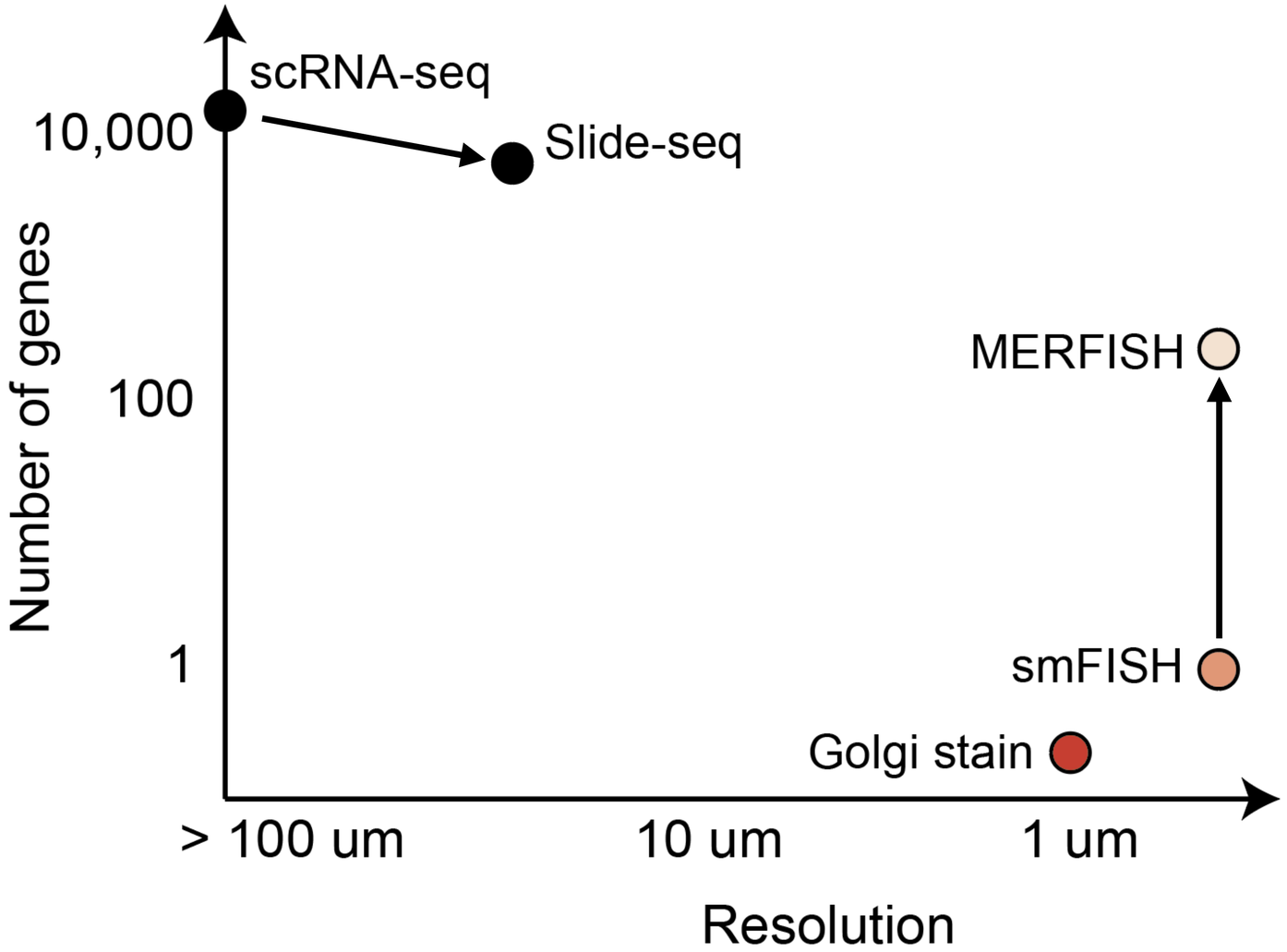
sequencing-based vs imaging-based assays



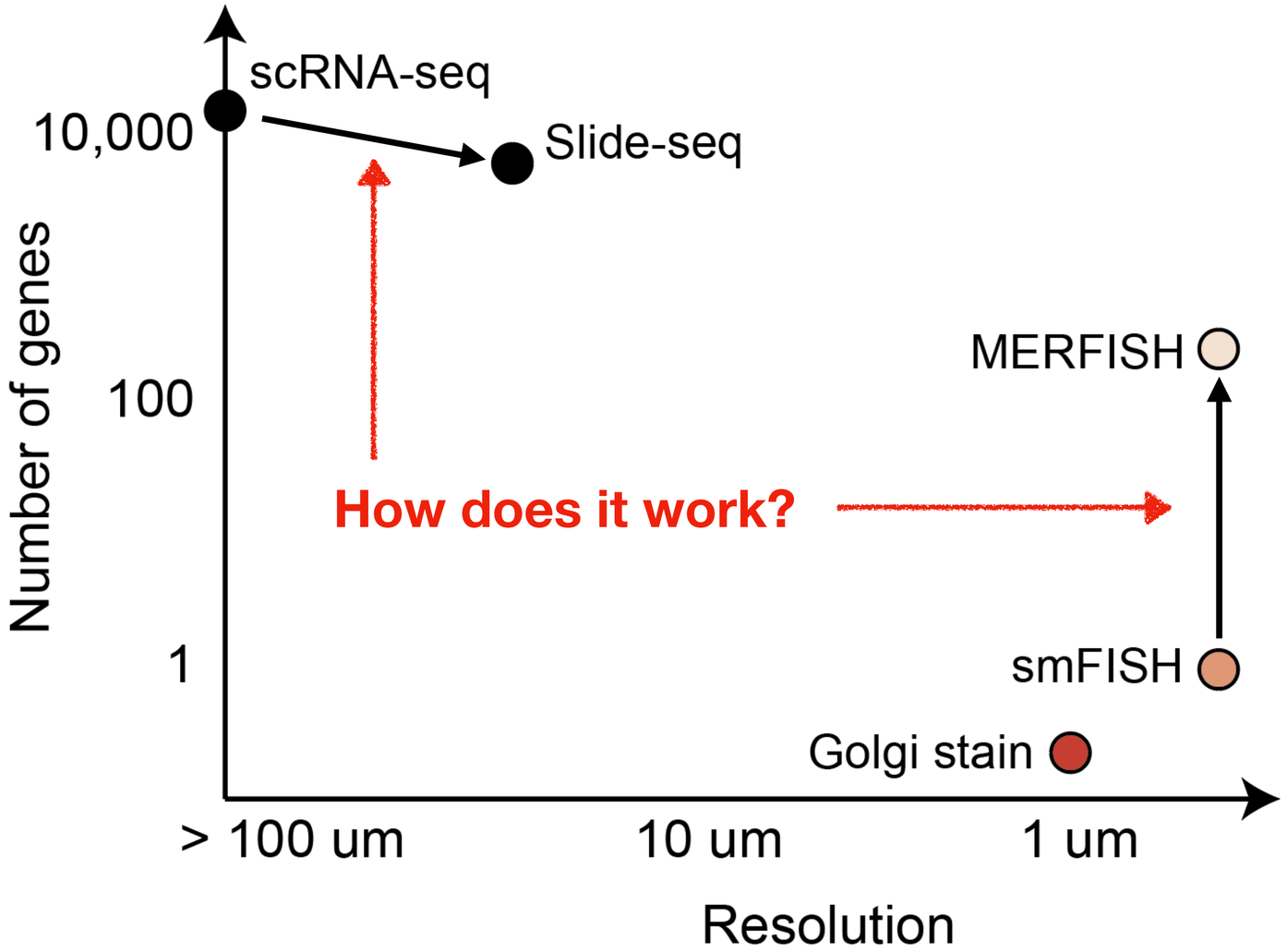
sequencing-based vs imaging-based assays



sequencing-based vs imaging-based assays



sequencing-based vs imaging-based assays

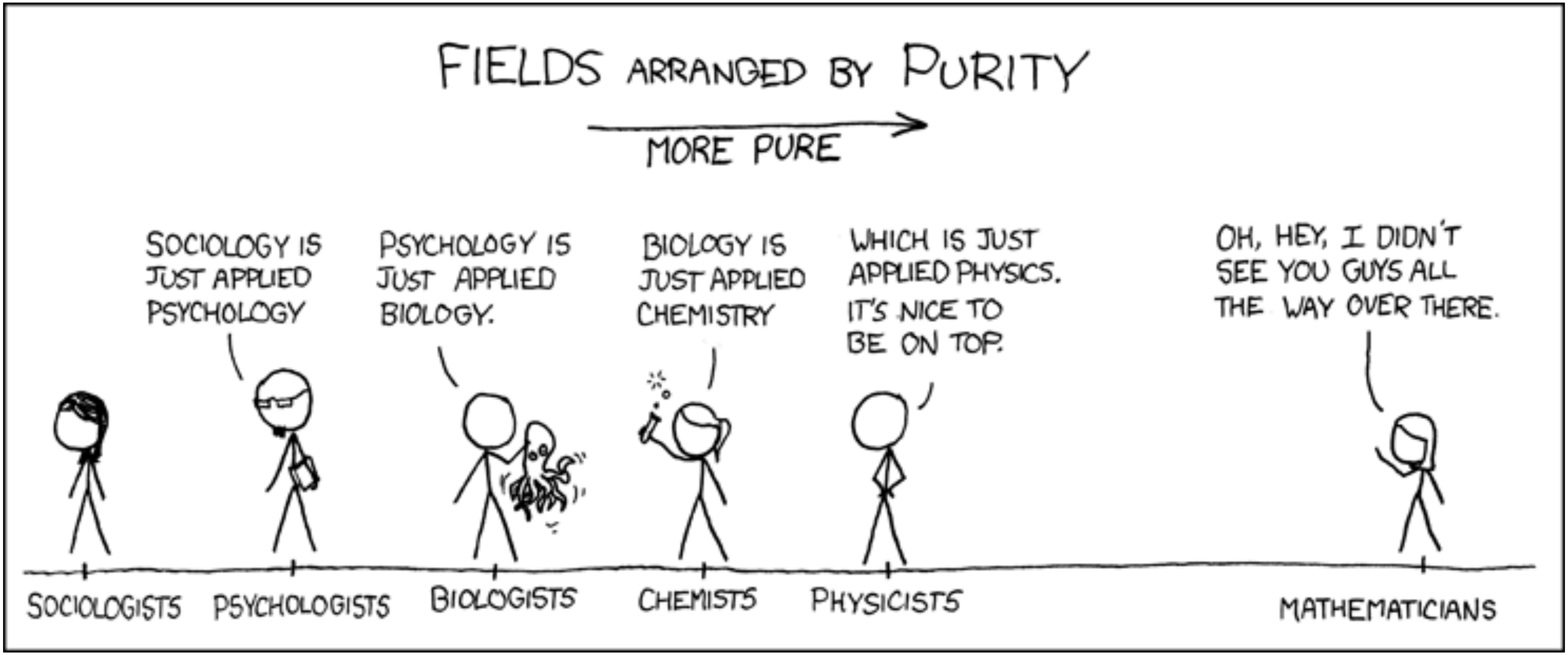


- Tradeoffs:
  - Multiplexing
  - Resolution
  - Throughput
  - Sensitivity

It’s break time!!!

KEEP
CALM
AND
ASK EASY
QUESTIONS

Google form for attendance
Day 1
https://forms.gle/MfuKUz5628JPryYR6



https://xkcd.com/435/

4 August 1972, Volume 177, Number 4047 SCIENCE

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.

Good reviews

- https://paperpile.com/shared/eGYOhc
  - 13 reviews so far
- Intro and major players in the field:
  - Methods of the year 2020 Nature Methods
- Computational challenges and current solutions:
  - Palla et al. 2022 Nature Biotech

Reading session 1: Chen et al. 2015 Science

- *Spatially resolved, highly multiplexed RNA profiling in single cells*
- How does MERFISH work? Read and try to understand **Figure 1** and **Figure 2A-D**.
- Questions to keep in mind:
  - What does the acronym MERFISH stands for?
  - How many genes can MERFISH measure?
  - MERFISH is the natural extension of which technology (mentioned in class)?
  - What allowed MERFISH to measure many genes?

[10 min] individual reading
[~15 min] open discussion; Q&A
[5 min] min-quiz — the winner gets no reward!

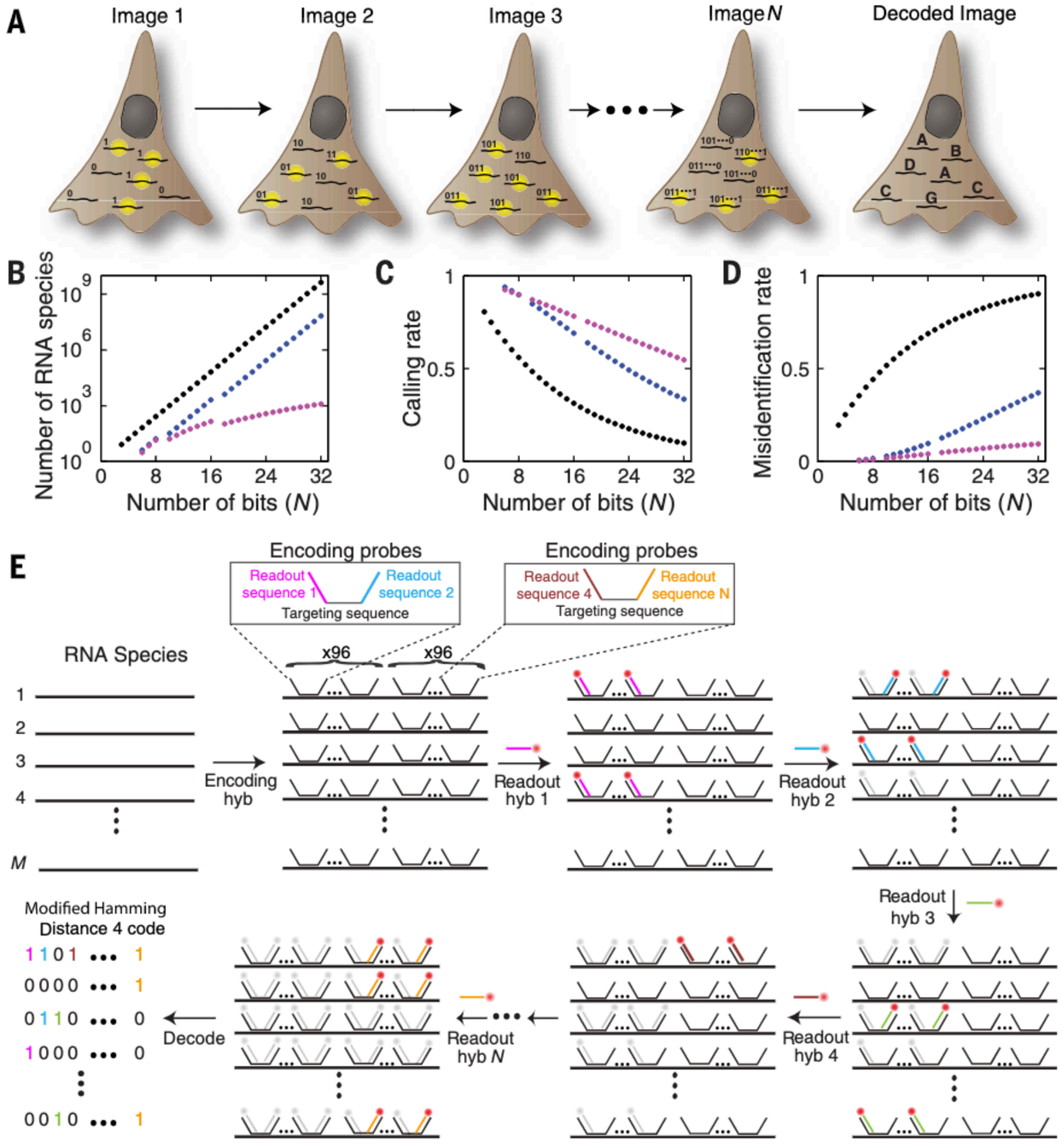


Fig. 1. MERFISH: A highly multiplexed smFISH approach enabled by combinatorial labeling and error-robust encoding. (**A**) Schematic depiction of the identification of multiple RNA species in *N* rounds of imaging. Each RNA species is encoded with a *N*-bit binary word, and during each round of imaging, 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.

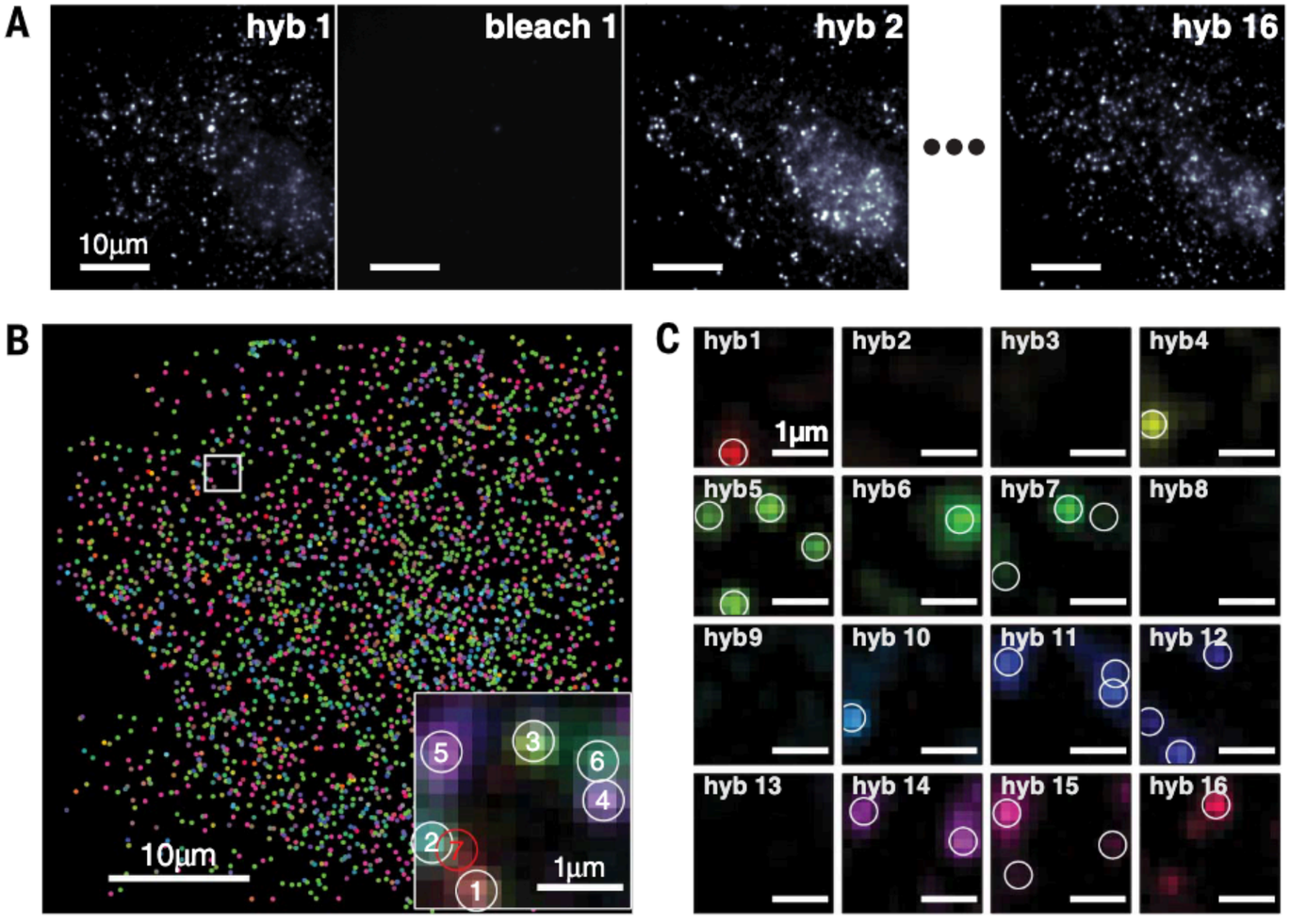


Fig. 2. Simultaneous measurement of 140 RNA species in single cells by use of MERFISH with a 16-bit MHD4 code. (**A**) Images of RNA molecules in an IMR90 cell after each hybridization round (hyb 1 to hyb 16). The images after photobleaching (for example, bleach 1) demonstrate efficient removal of fluorescent 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

Mini quiz

- Go to kahoot.it

Reading session 2: Rodriques et al. 2019 Science

- “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution”
- How does Slide-seq work? Read 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?

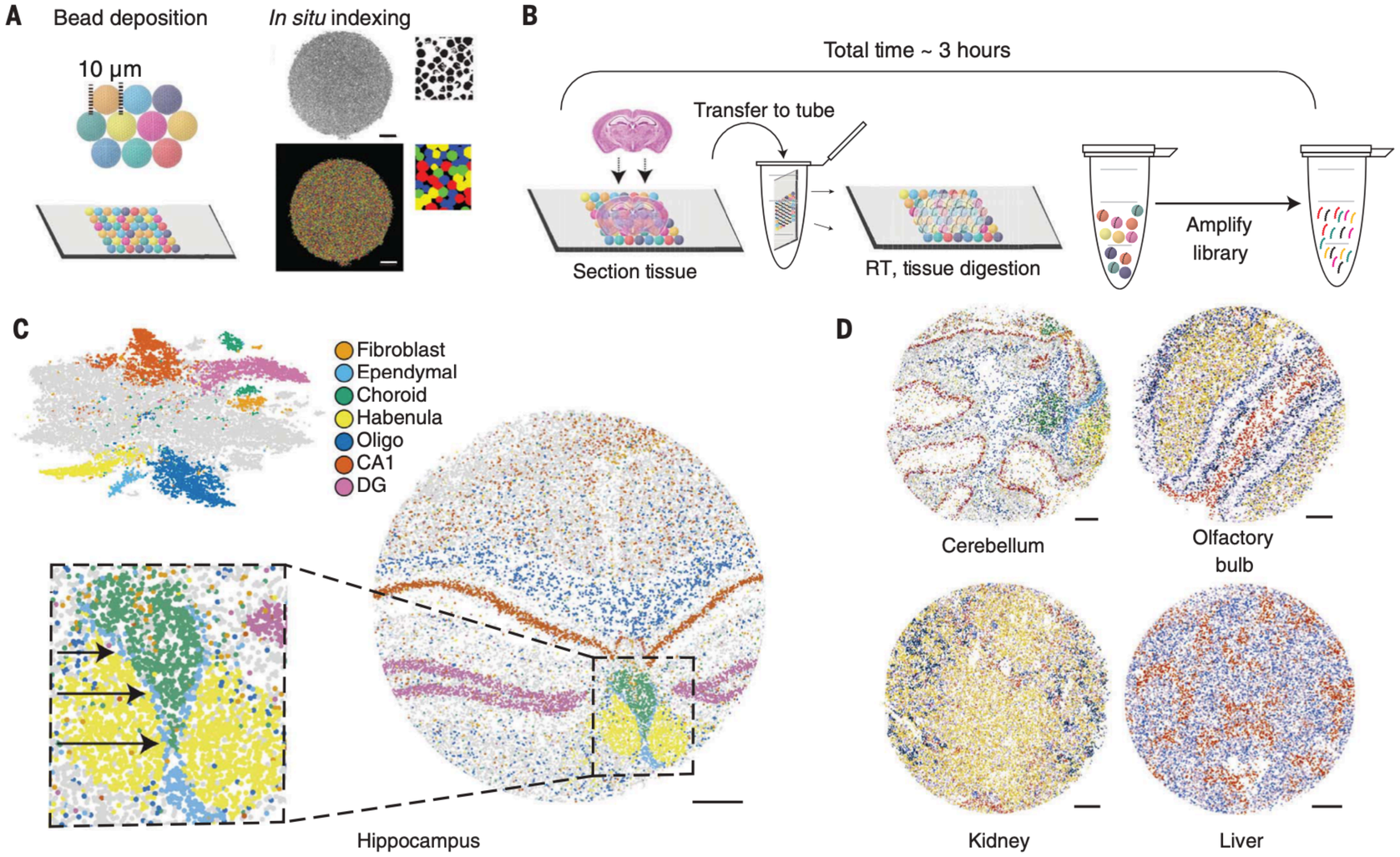


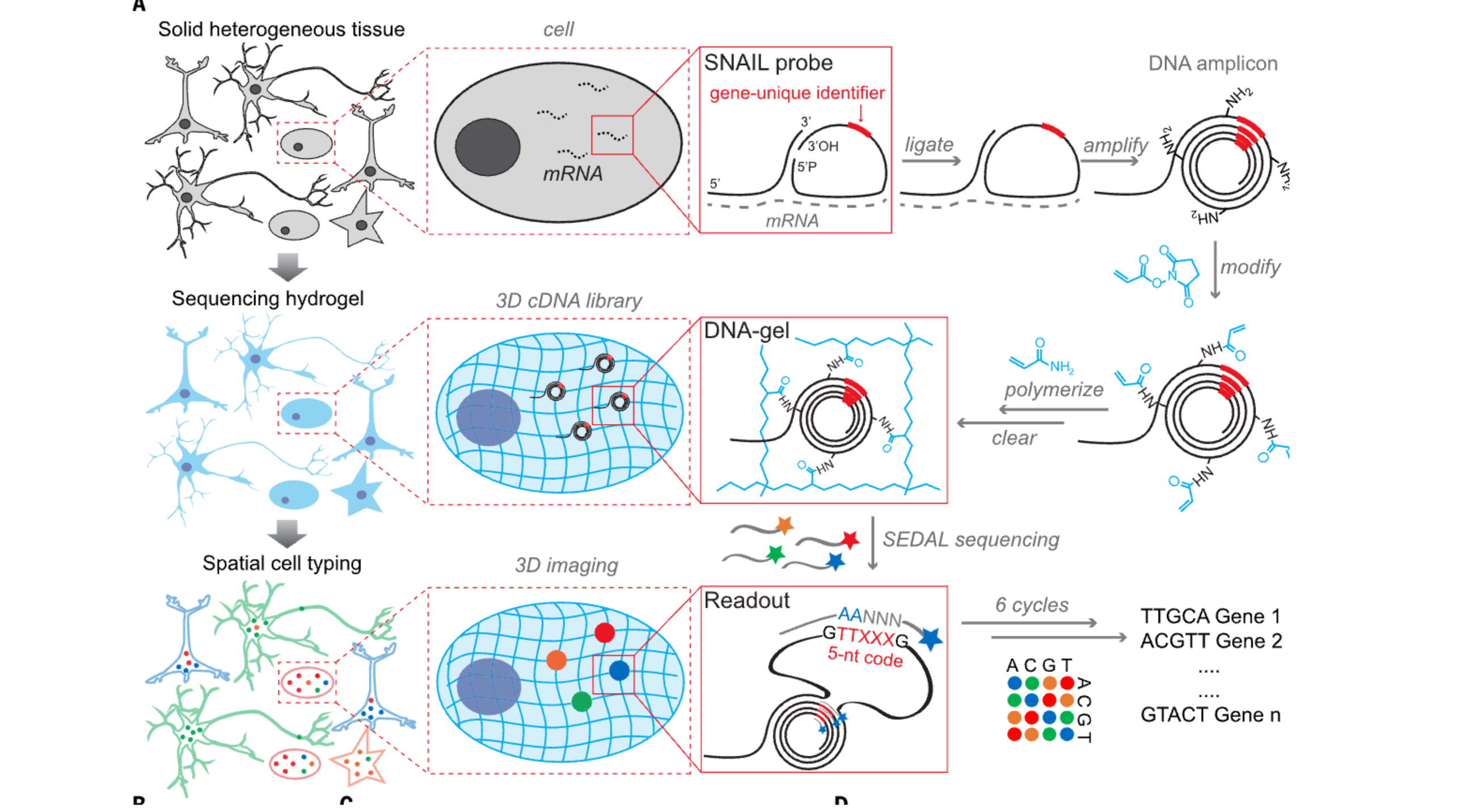
Fig. 1. High-resolution RNA capture from tissue by Slide-seq. (**A**) (Left) Schematic of array generation. A monolayer of randomly deposited, DNA barcoded beads (a "puck") is spatially indexed by SOLiD sequencing. (Top right) Representative puck with sequenced barcodes shown in black. (Bottom right) 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.

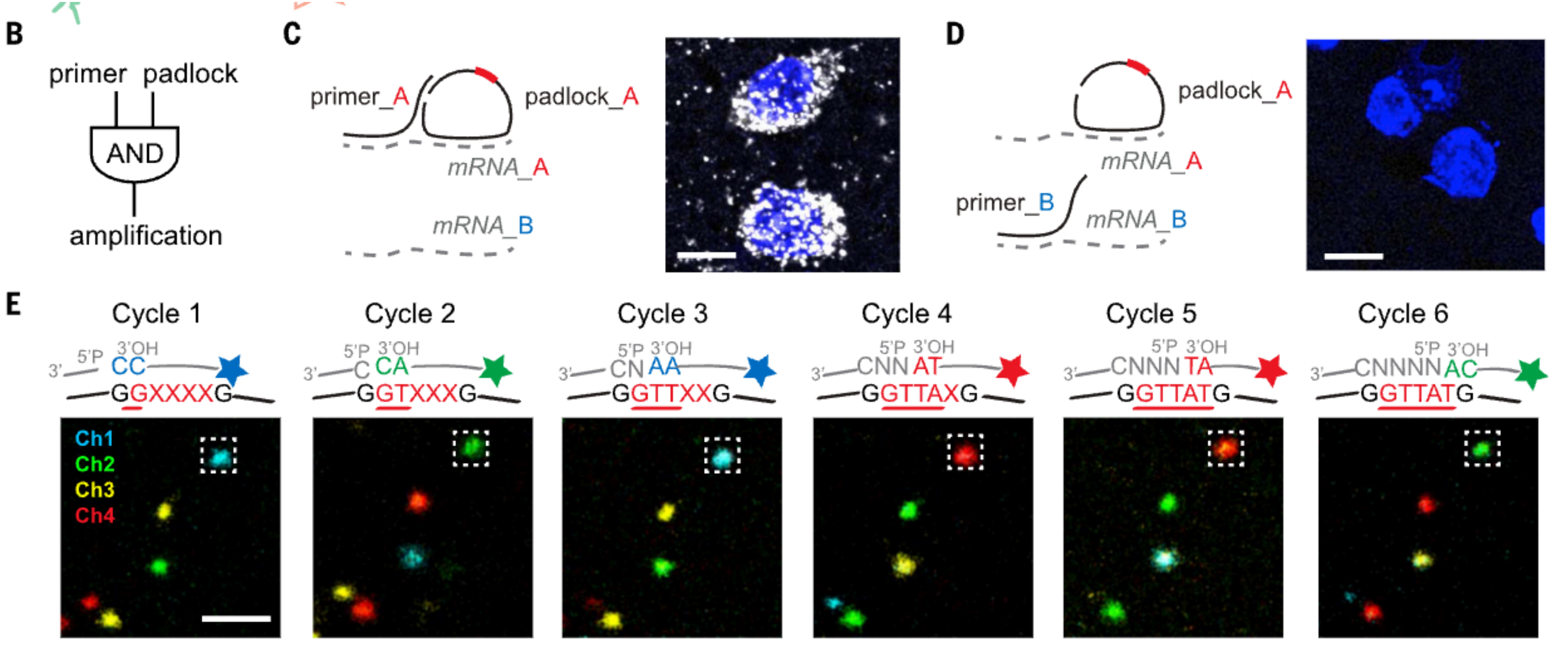
Mini quiz

- 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?


A
Solid heterogeneous tissue
cell
mRNA
SNAIL probe
gene-unique identifier
3'
3'OH
5'P
5'
mRNA
ligate
amplify
DNA amplicon
modify
Sequencing hydrogel
3D cDNA library
DNA-gel
polymerize
clear
SEDAL sequencing
Spatial cell typing
3D imaging
Readout
AANNN
GTTXXXG
5-nt code
6 cycles
A C G T
TTGCA Gene 1
ACGTT Gene 2
....
....
GTACT Gene n

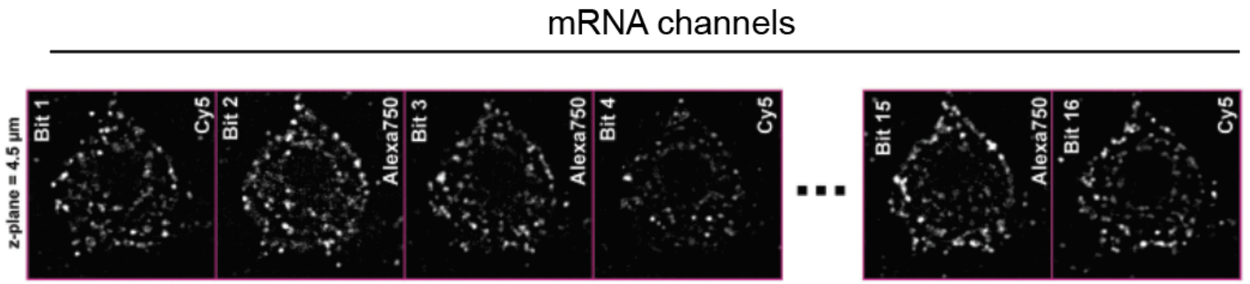

B
primer
padlock
AND
amplification
C
primer_A
padlock_A
mRNA_A
mRNA_B
D
padlock_A
mRNA_A
primer_B
mRNA_B
E
Cycle 1
Cycle 2
Cycle 3
Cycle 4
Cycle 5
Cycle 6
5'P 3'OH
3'
CC
GGXXXXG
CA
GGTXXXG
CN AA
GGTTXXG
CNN AT
GGTTAXG
CNNN TA
GGTTATG
CNNNN AC
GGTTATG
Ch1
Ch2
Ch3
Ch4

Mini quiz

- Go to kahoot.it

Data-processing workflow — image-based

Stacked raw images

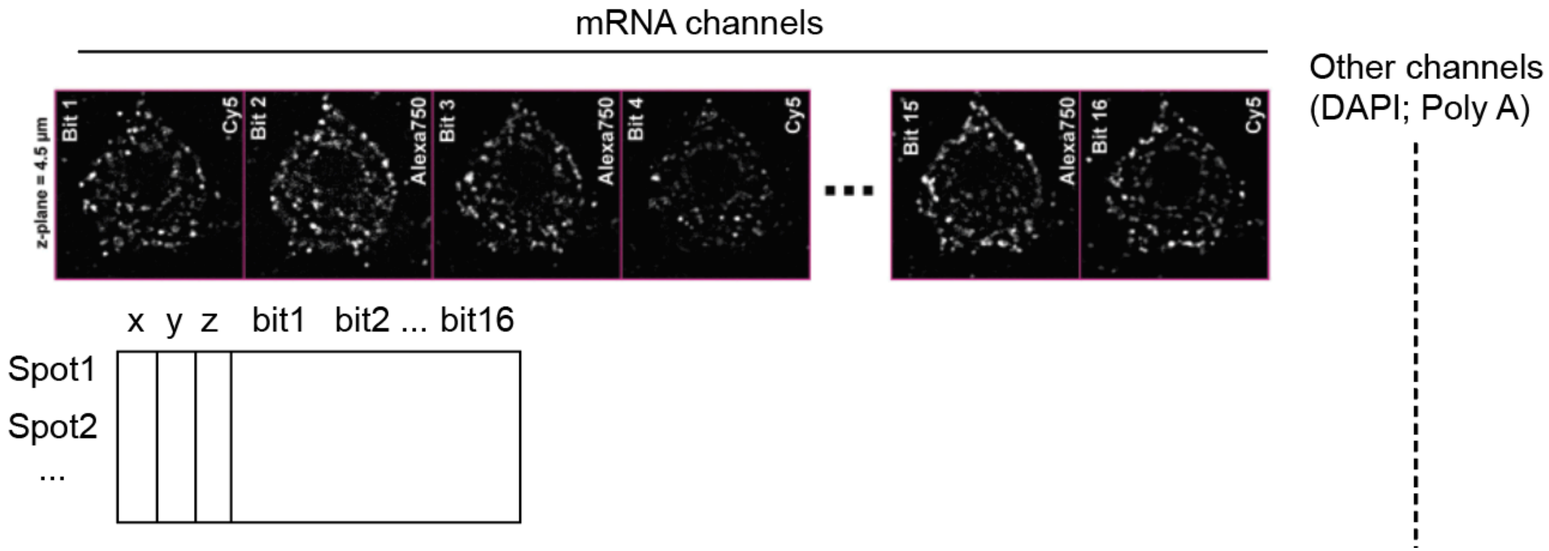


Data-processing workflow — image-based

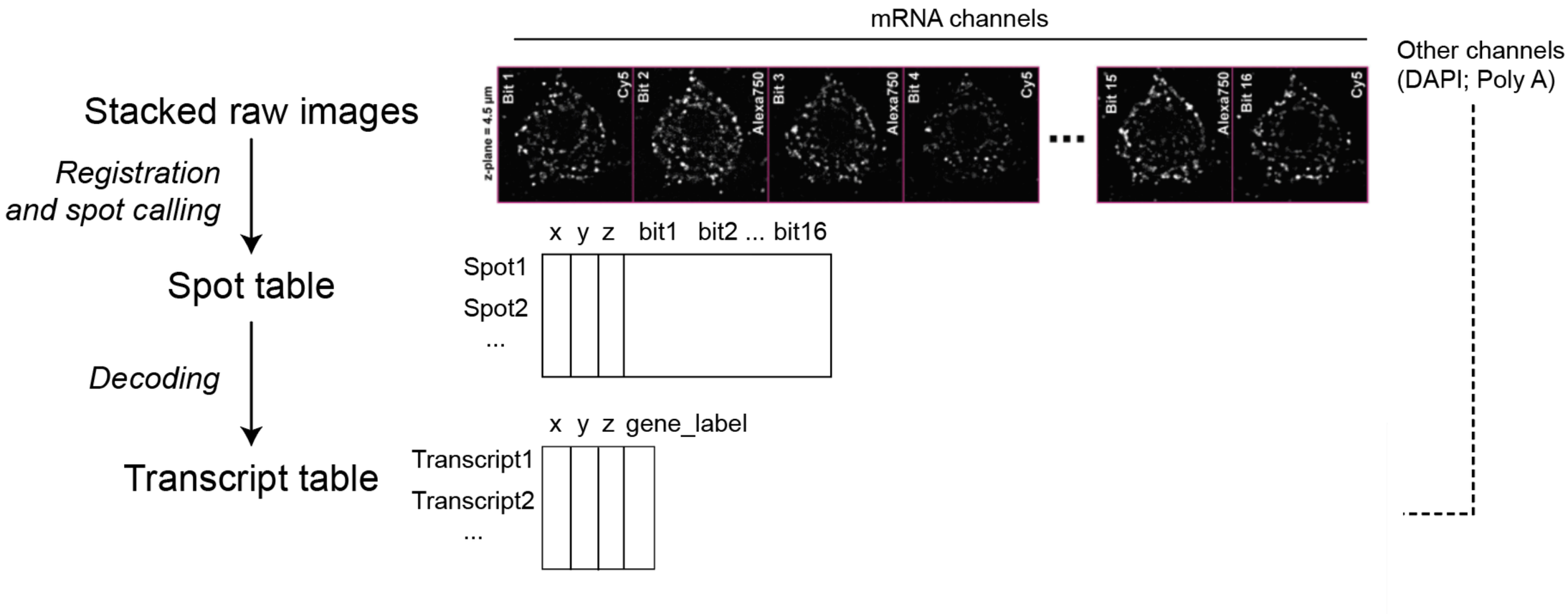
Stacked raw images

Registration and spot calling

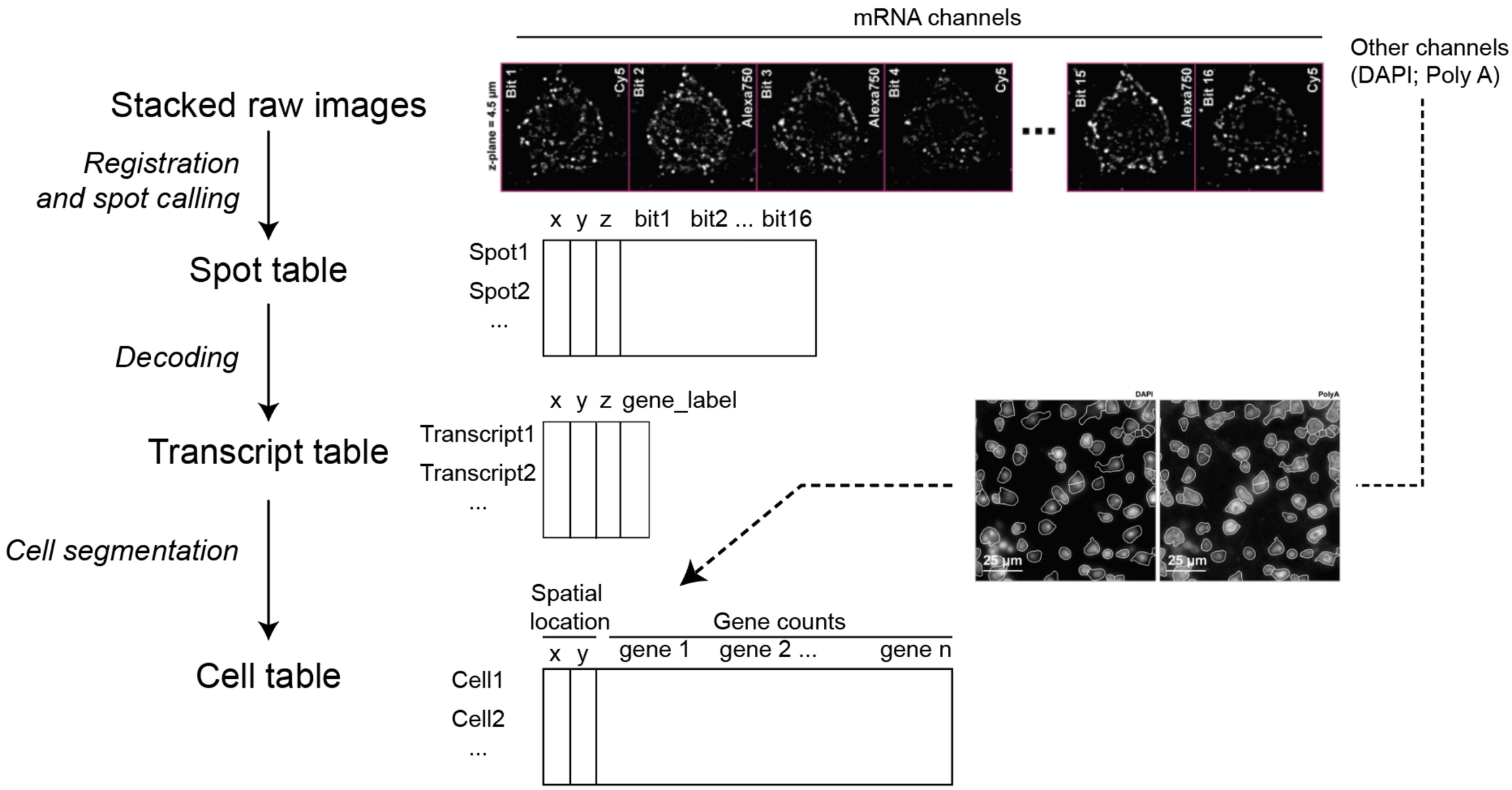
Spot table



Data-processing workflow — image-based



Data-processing workflow — image-based



Data-processing workflow — sequencing-based

Raw sequencing reads

```
Read1 ATGCAAC……TCGT
Read2 CATGGCA……CCGG
Read3 GTGCGAG……ACAG
……
```

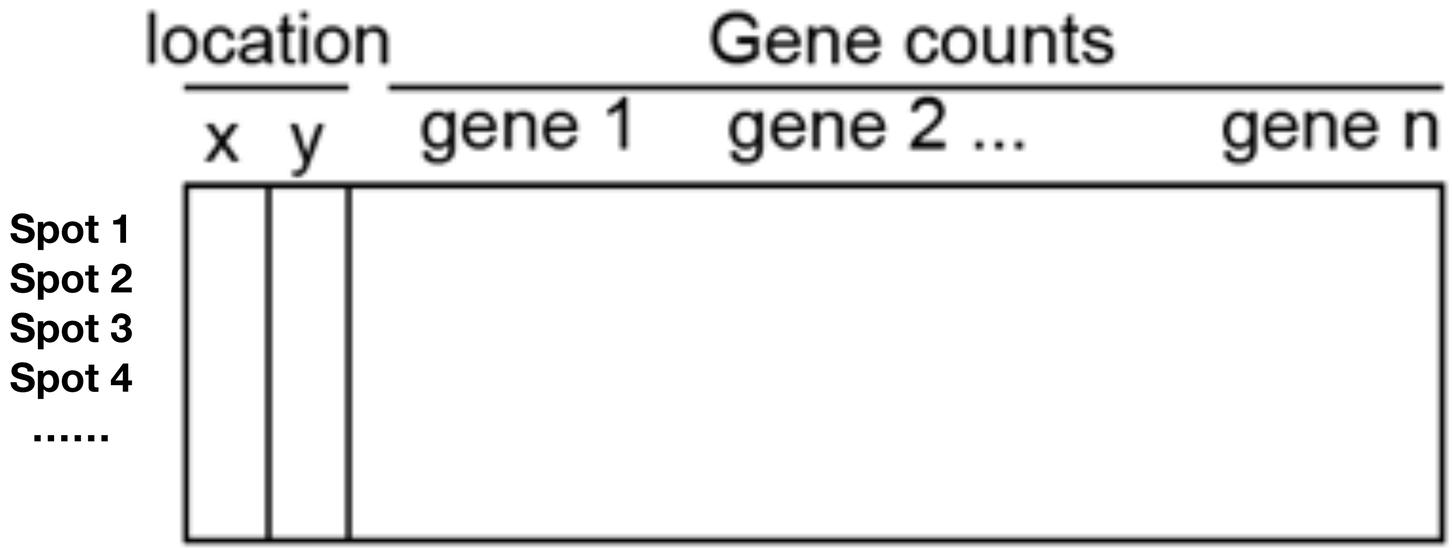
Demultiplexing

Section I **Section II**

ATGCAAC……TCGT
CATGGCA……CCGG
GTGCGAG……ACAG
……

Map to the genome
(Which gene?)

Look up spatial coordinates
(Which spot?)



Thank you and see you tomorrow!

UCLA Collaboratory workshop (W31)

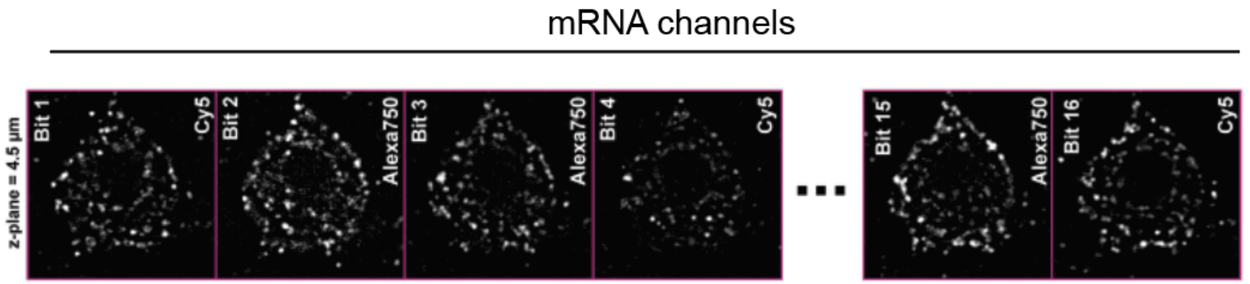
Spatial Transcriptomics

March 16, 2023 (Day 2)

Fangming Xie

Data-processing workflow — image-based

Stacked raw images

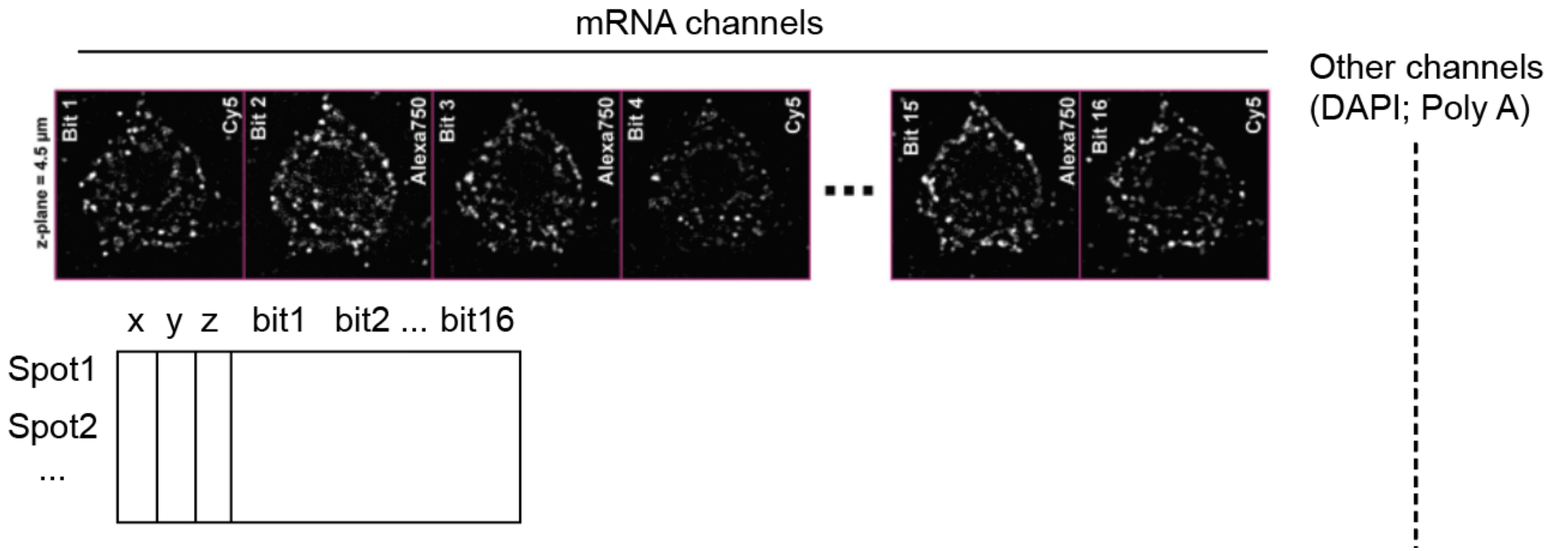


Data-processing workflow — image-based

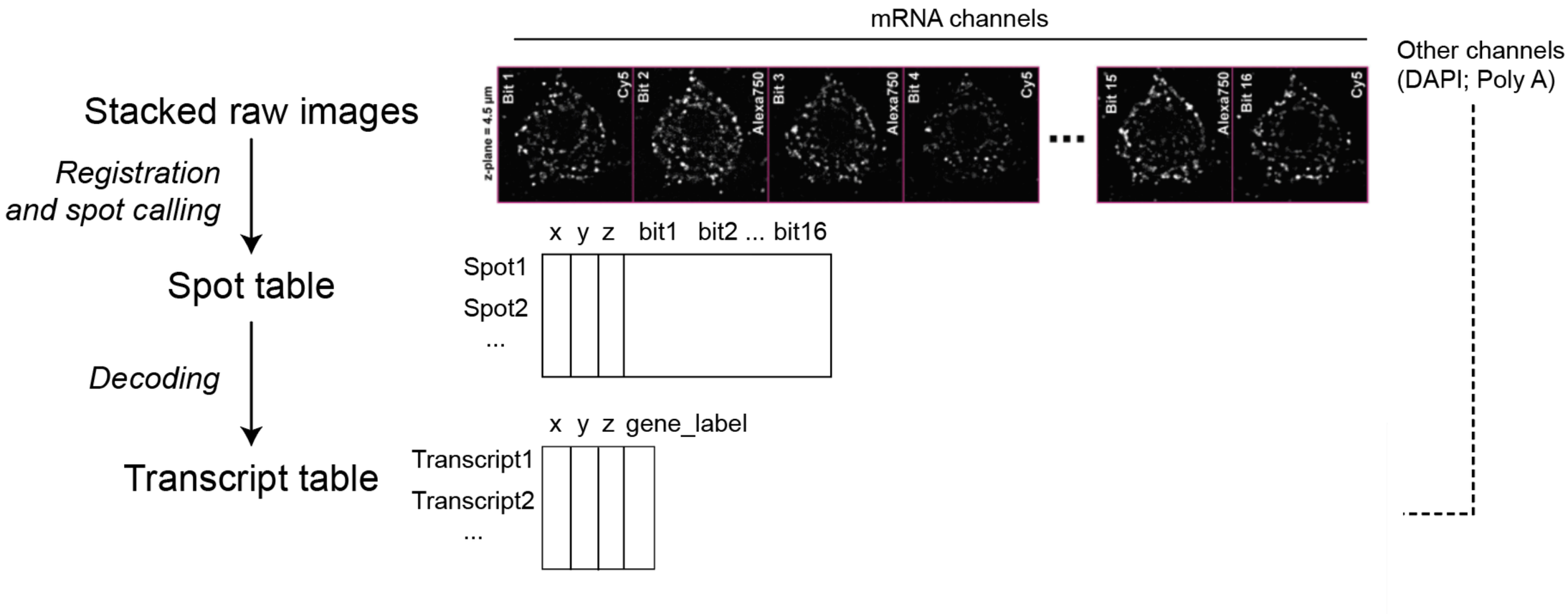
Stacked raw images

Registration and spot calling

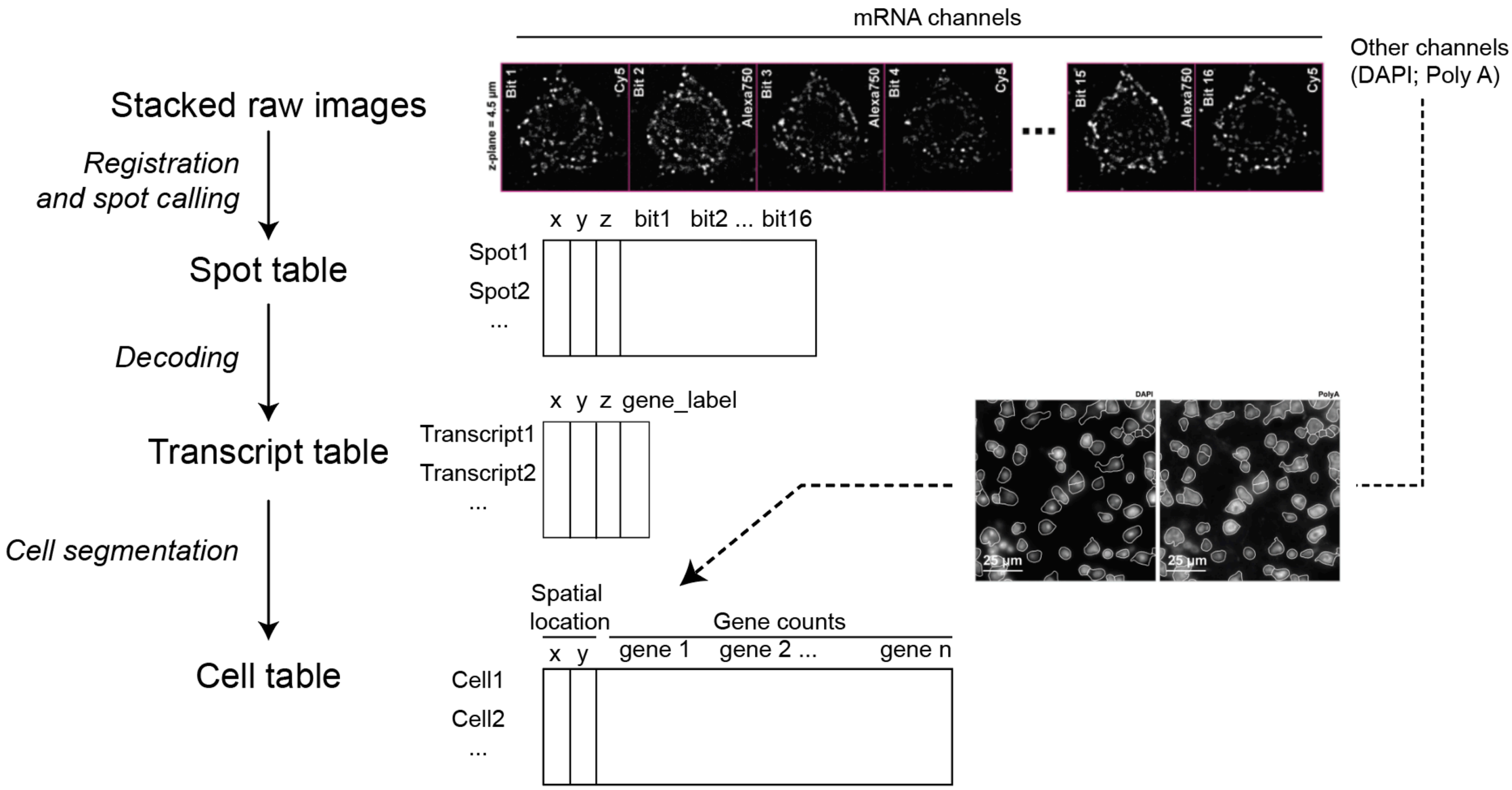
Spot table



Data-processing workflow — image-based



Data-processing workflow — image-based



Data-processing workflow — sequencing-based

Raw sequencing reads

```
Read1 ATGCAAC……TCGT
Read2 CATGGCA……CCGG
Read3 GTGCGAG……ACAG
……
```

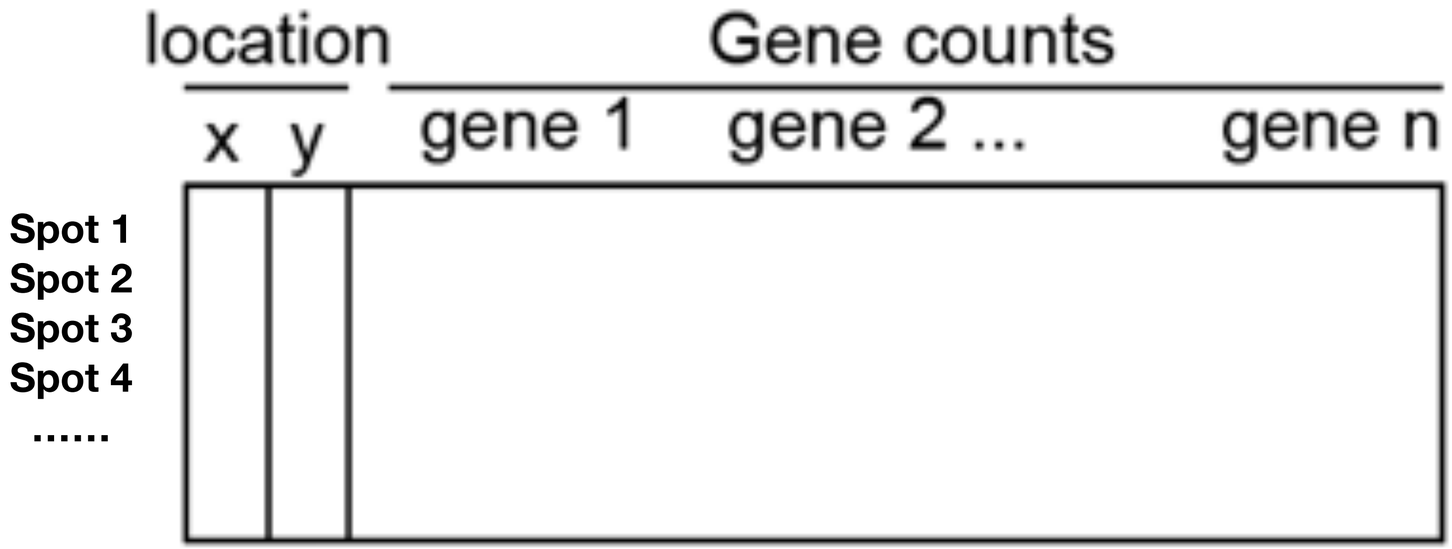
Demultiplexing

Section I **Section II**

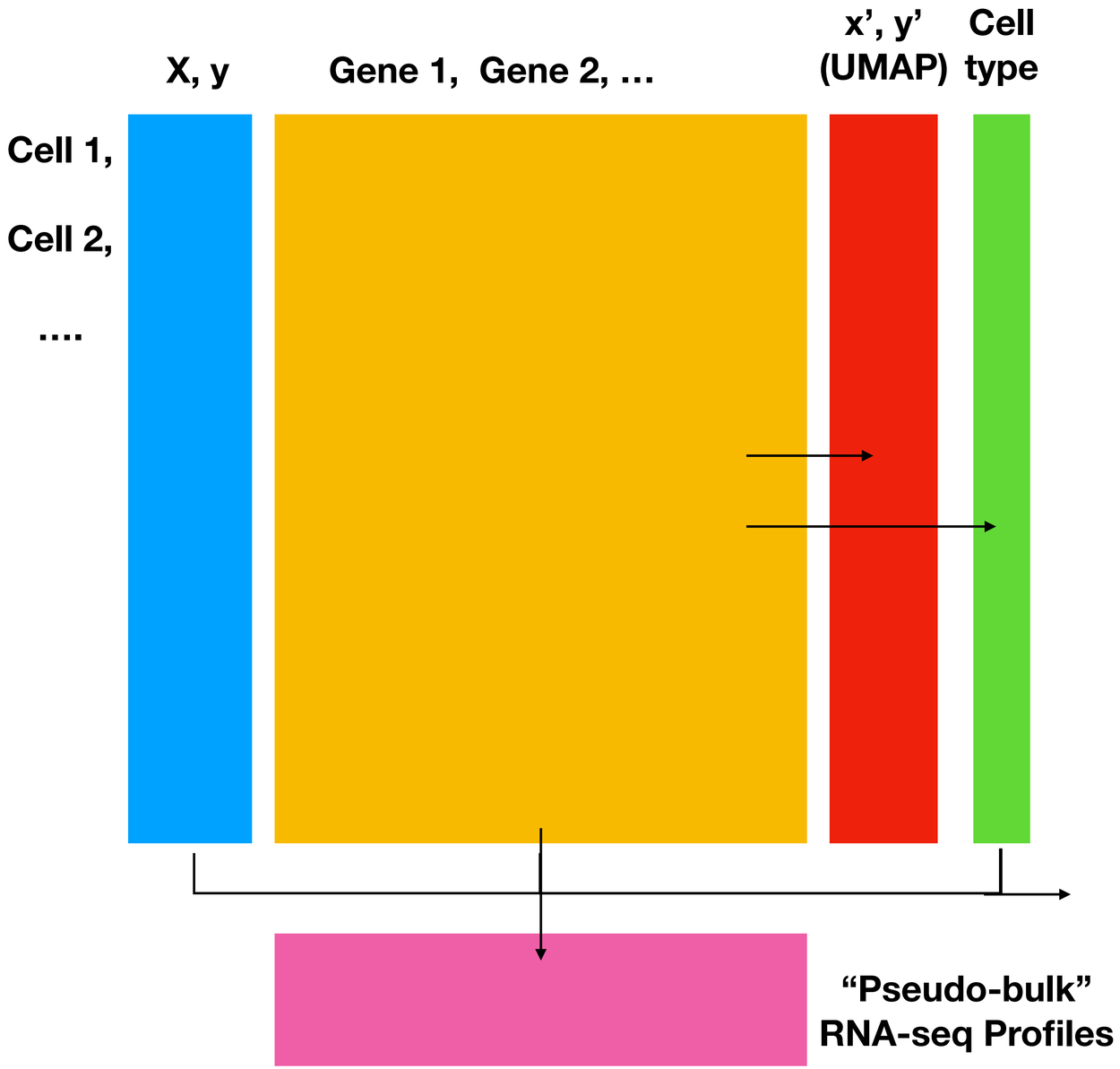
ATGCAAC……TCGT
CATGGCA……CCGG
GTGCGAG……ACAG
……

Map to the genome
(Which gene?)

Look up spatial coordinates
(Which spot?)



Analysis overview

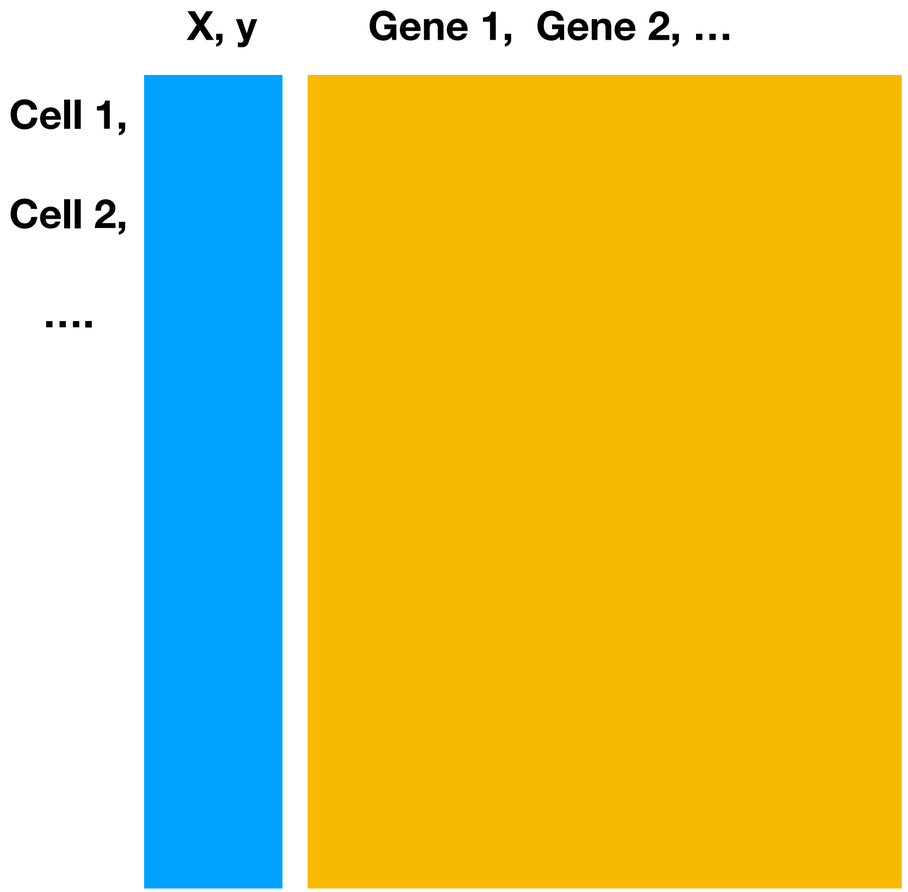


- [x, y, one gene at a time] — spatial distribution of gene expression
- [All genes] — dimensionality reduction (PCA, UMAP)
- [All genes] — clustering/cell typing (k-means, hierarchical, Leiden)
- [x, y, cell type] — spatial proximity, interaction, and spatial enrichment of cell types.

The coding bootcamp

- For 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
- Stop me anytime for questions

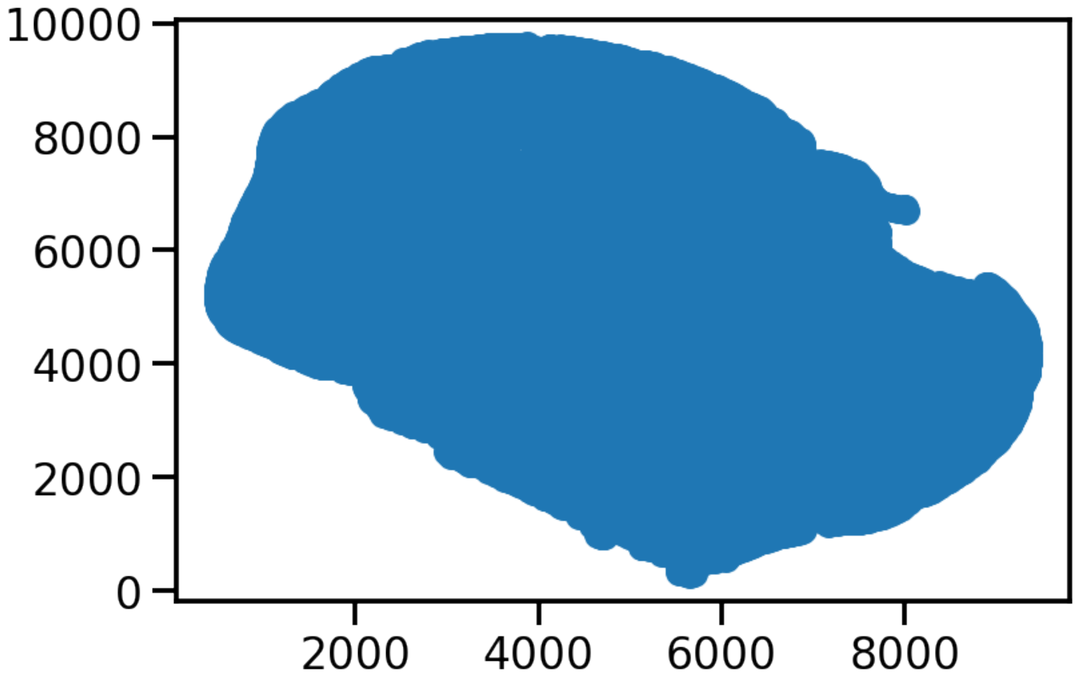
Coding exercise 1:
visualize the spatial distribution of gene expression



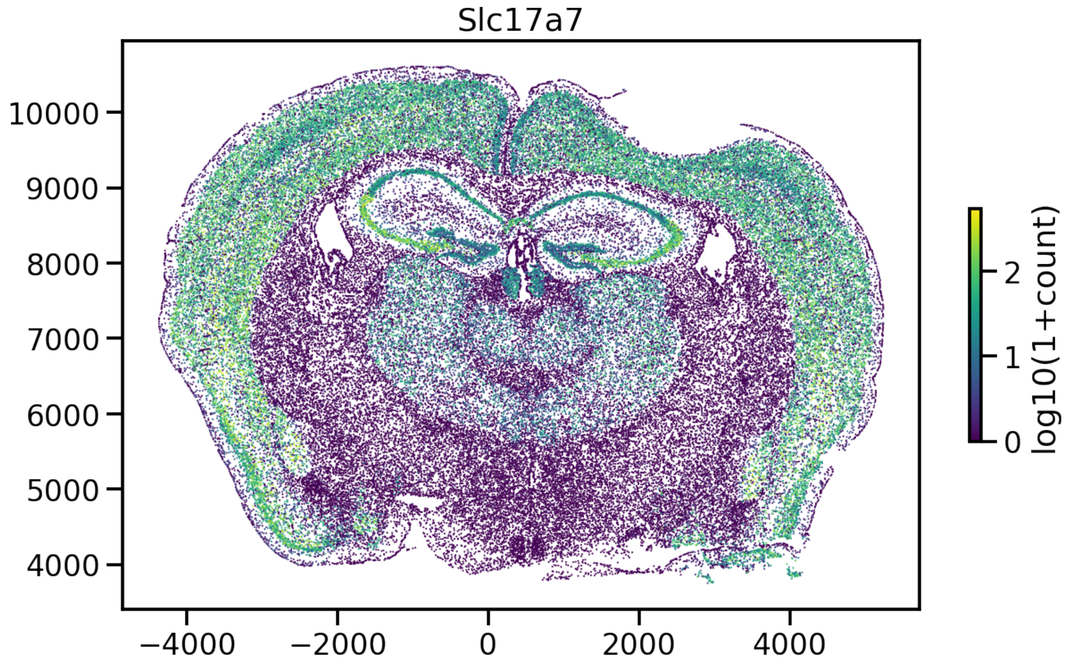
- Practicing data visualization on a public MERFISH mouse brain dataset from Vizgen.
- If you are experienced in coding, you can skip this and choose to work on bonus exercise 1.

Coding exercise 1:
Making a scatter plot from scratch

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



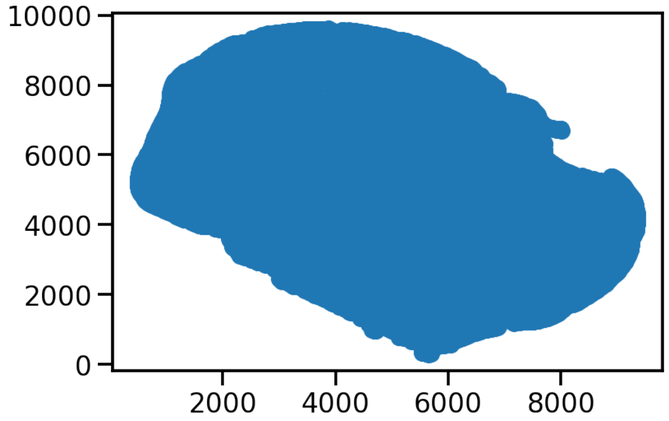
- End here:

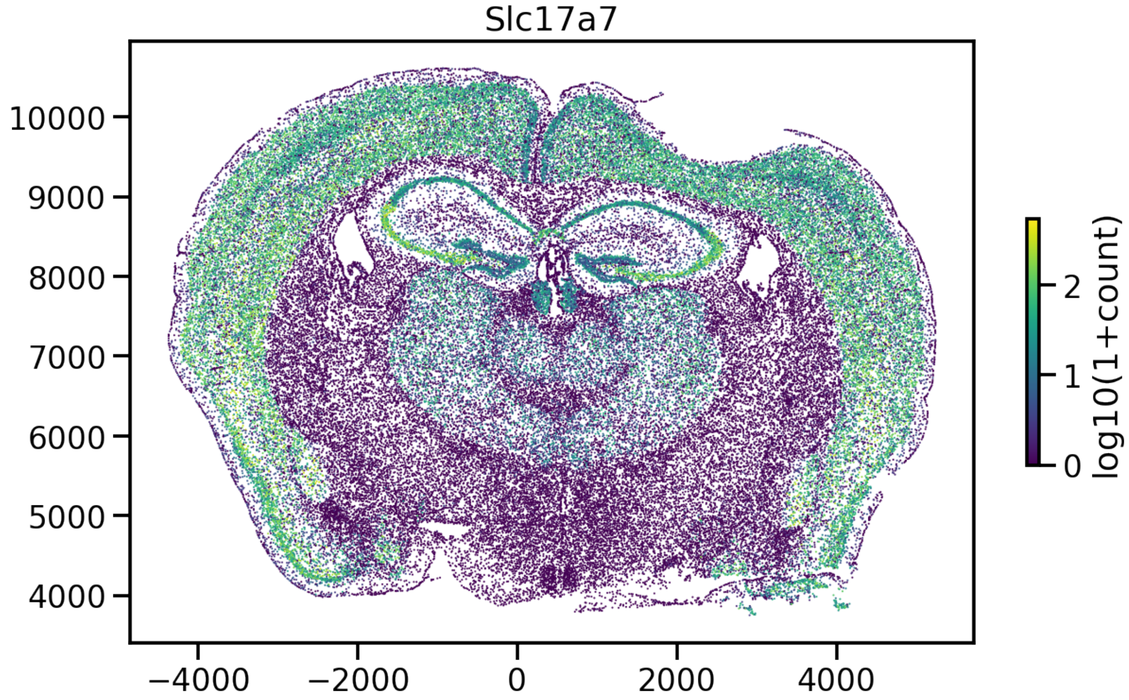


The devil is in the details

Problems we will encounter:

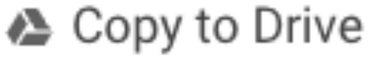
- How to get cell coordinates and gene expression from data?
- How to make a scatter plot colored by gene expression?
- How to add a color bar?
- Why the plot doesn’t look good?
- How to deal with crowding?
- How to deal with outliers?
- How to preserve aspect ratio?
- How to rotate spatial coordinates?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



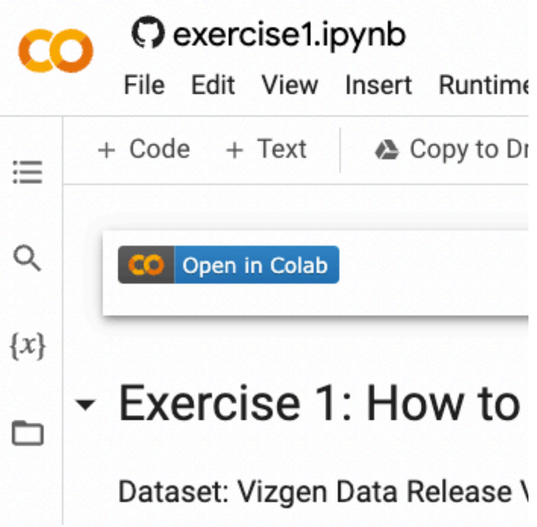


Setting up the coding environment

- 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: 

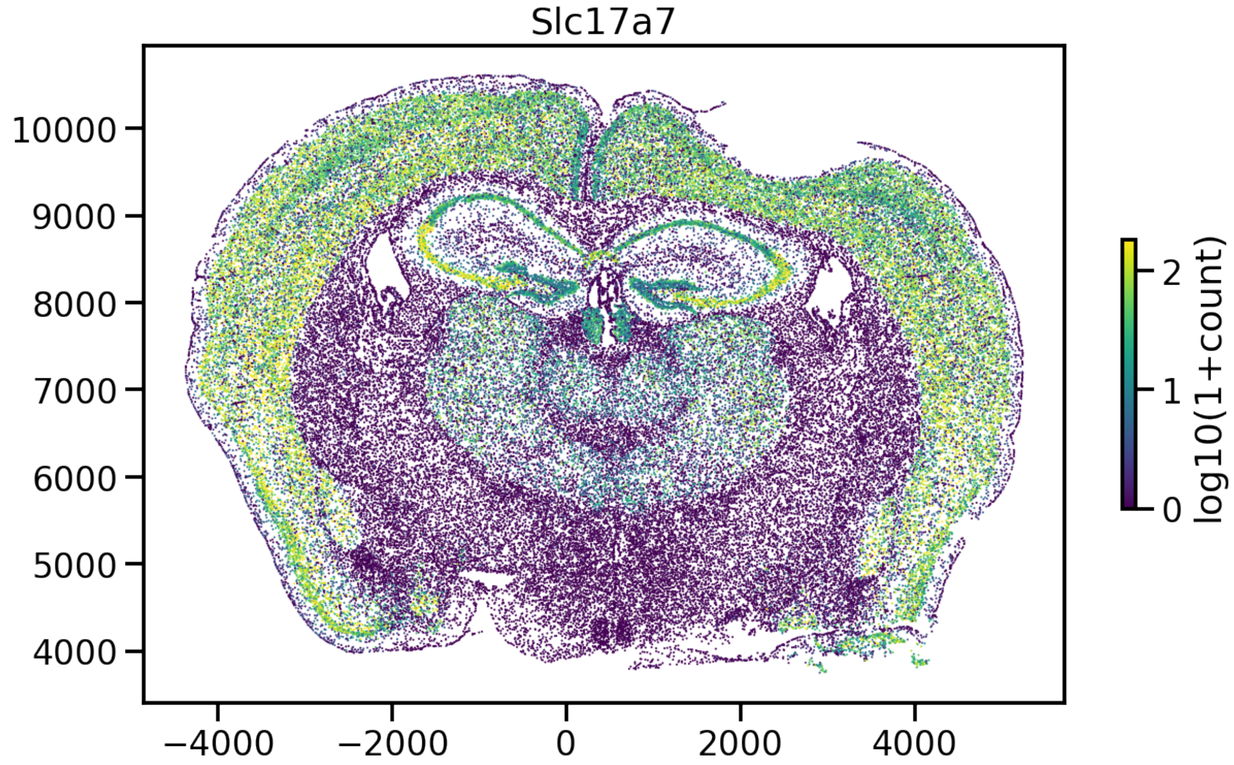
- Share your results (screenshots) here.



Data is adapted from *Vizgen Data Release V1.0. May 2021*

Review: Visualizing gene expression in-situ

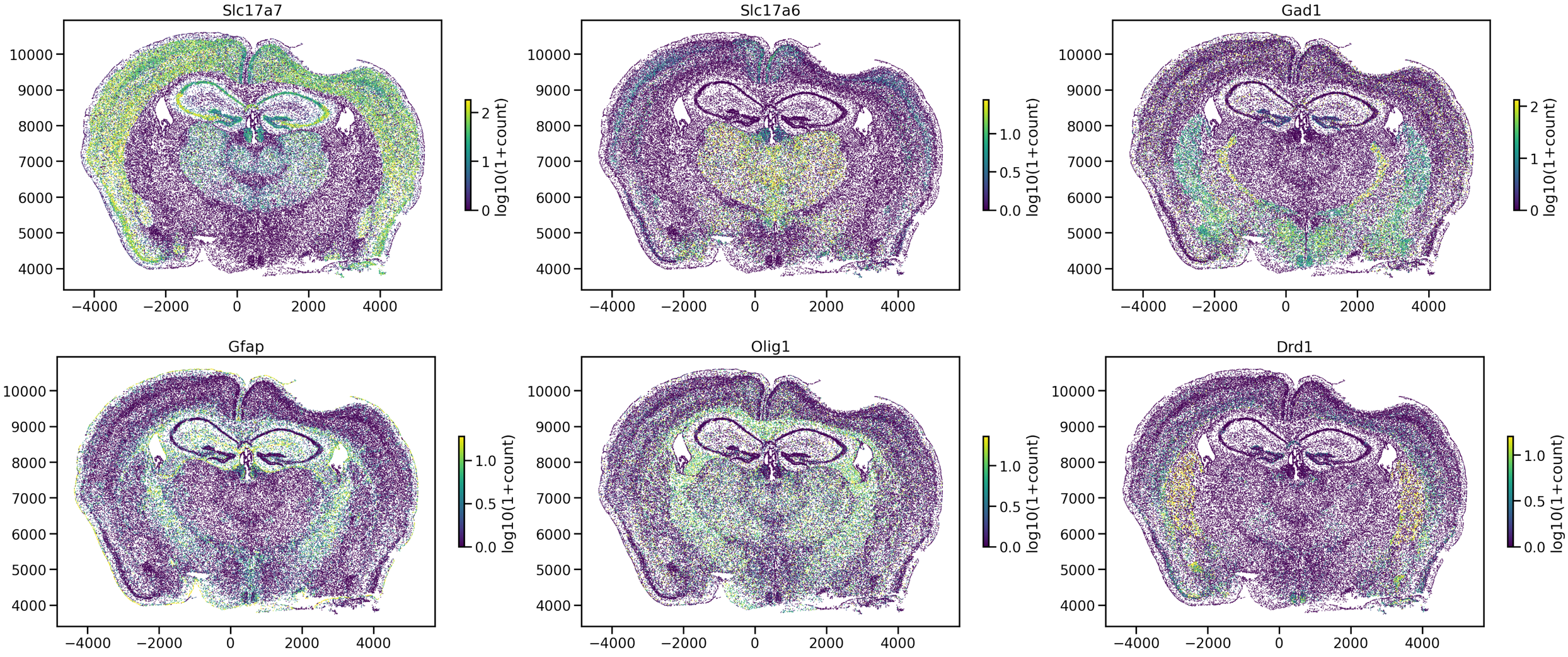
- x, y, gene expression
- Be mindful of data-point crowding
- Caps at ~95%
- Normalization



- 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 reveals new problem.

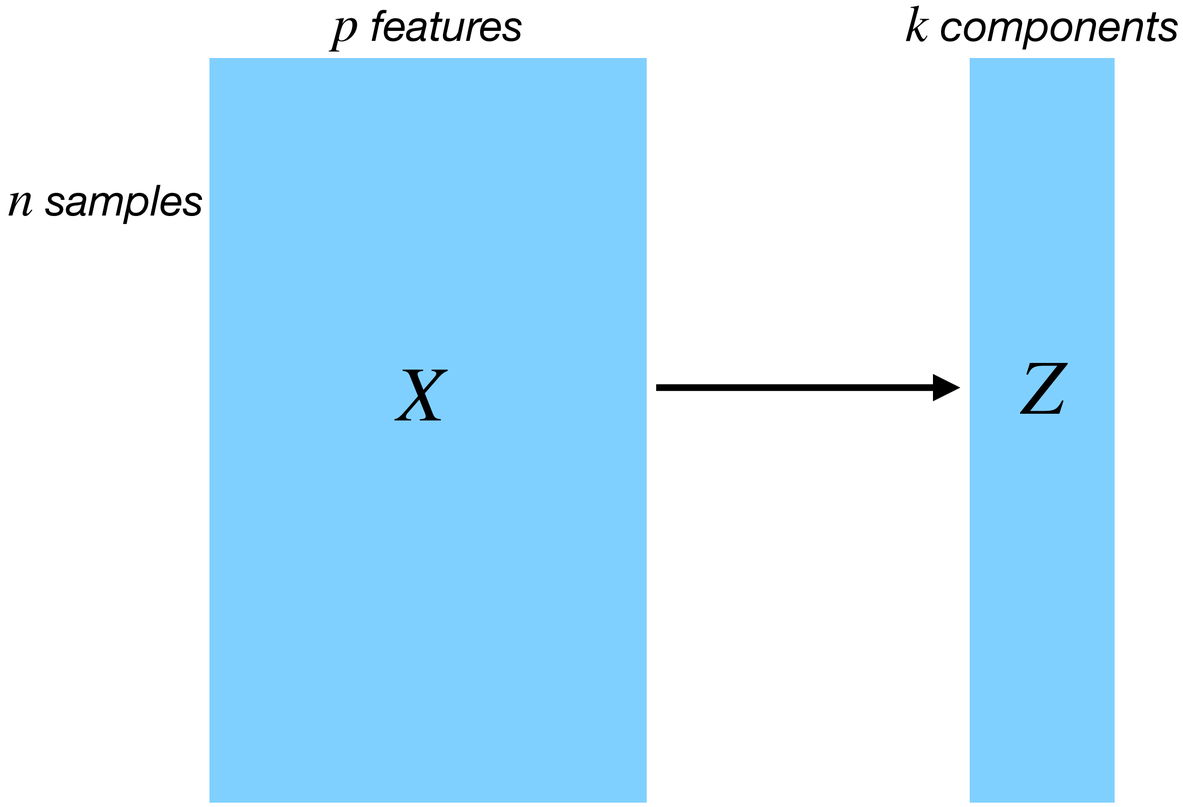


Part 2. Dimensionality reduction

- **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

The magic of dimensionality reduction

$$X \in R^{n \times p} \rightarrow Z \in R^{n \times k}$$



$$(k \ll p < n)$$

scRNA-seq:

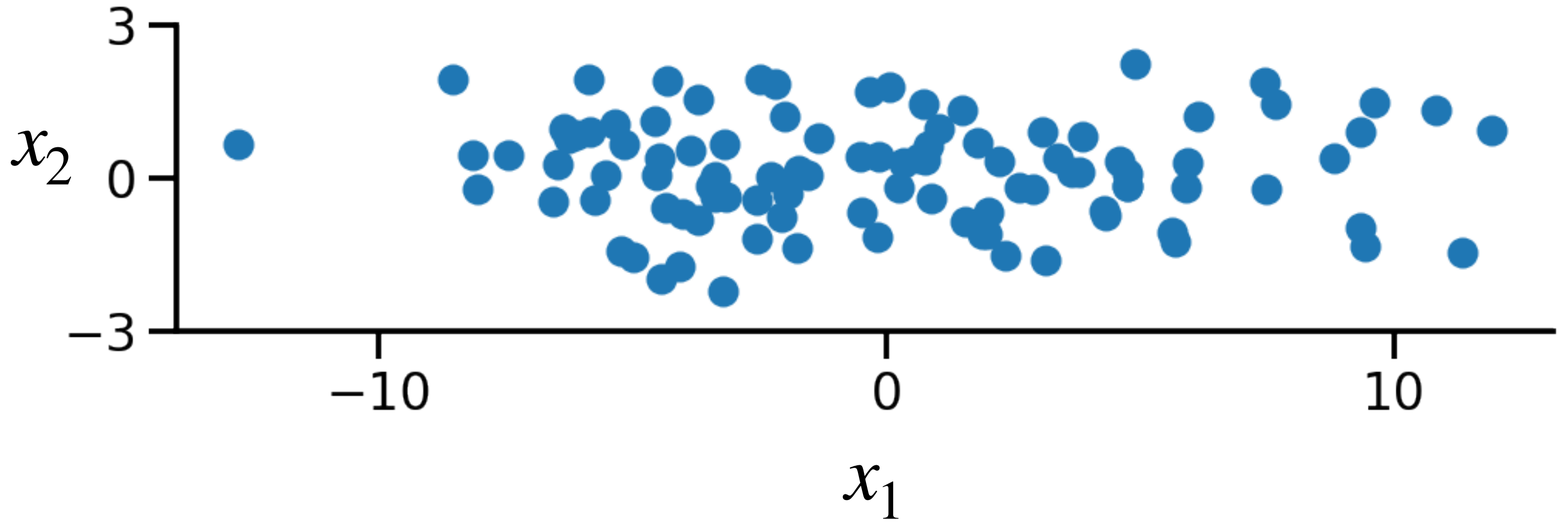
- n: 100k ~ 1million cells
- p: 10k ~ 50k genes
- k: 10 ~ 100 components

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

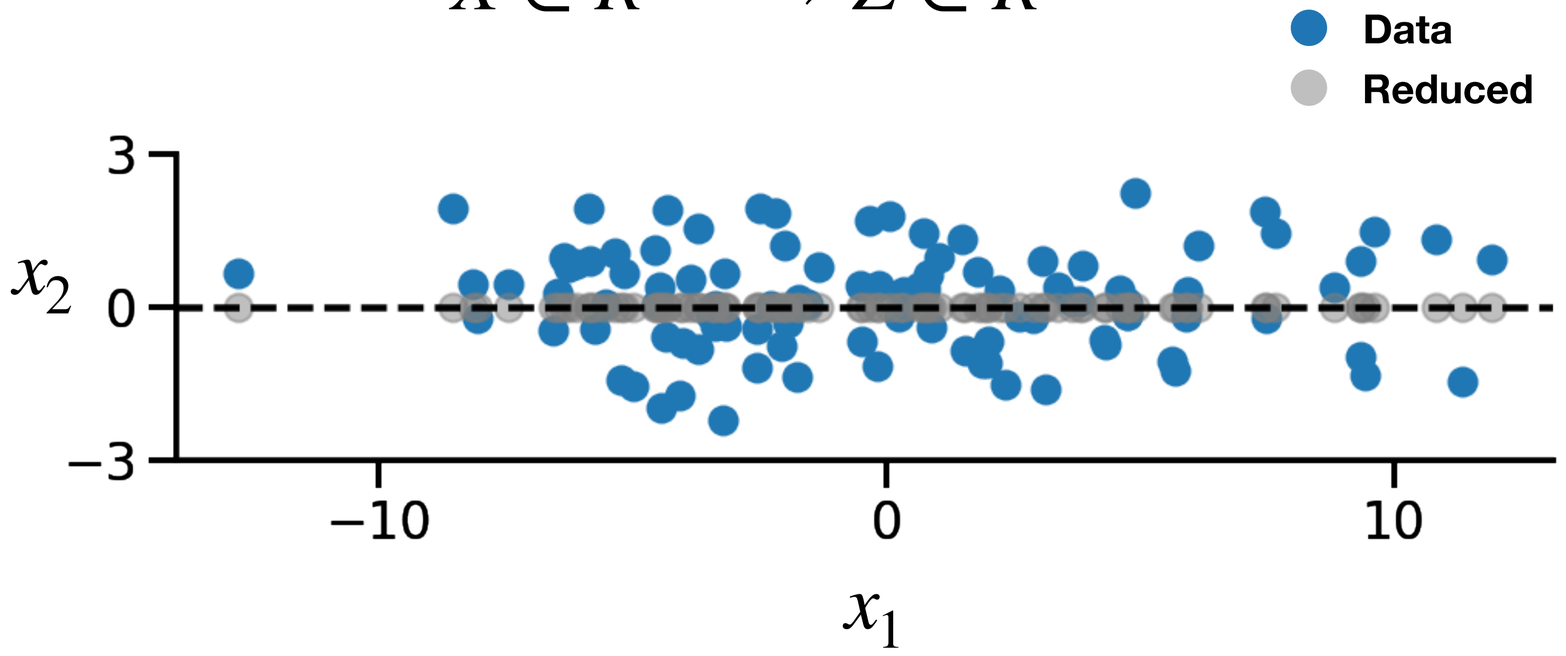


Strategy: remove the low-variance column

x_1 x_2

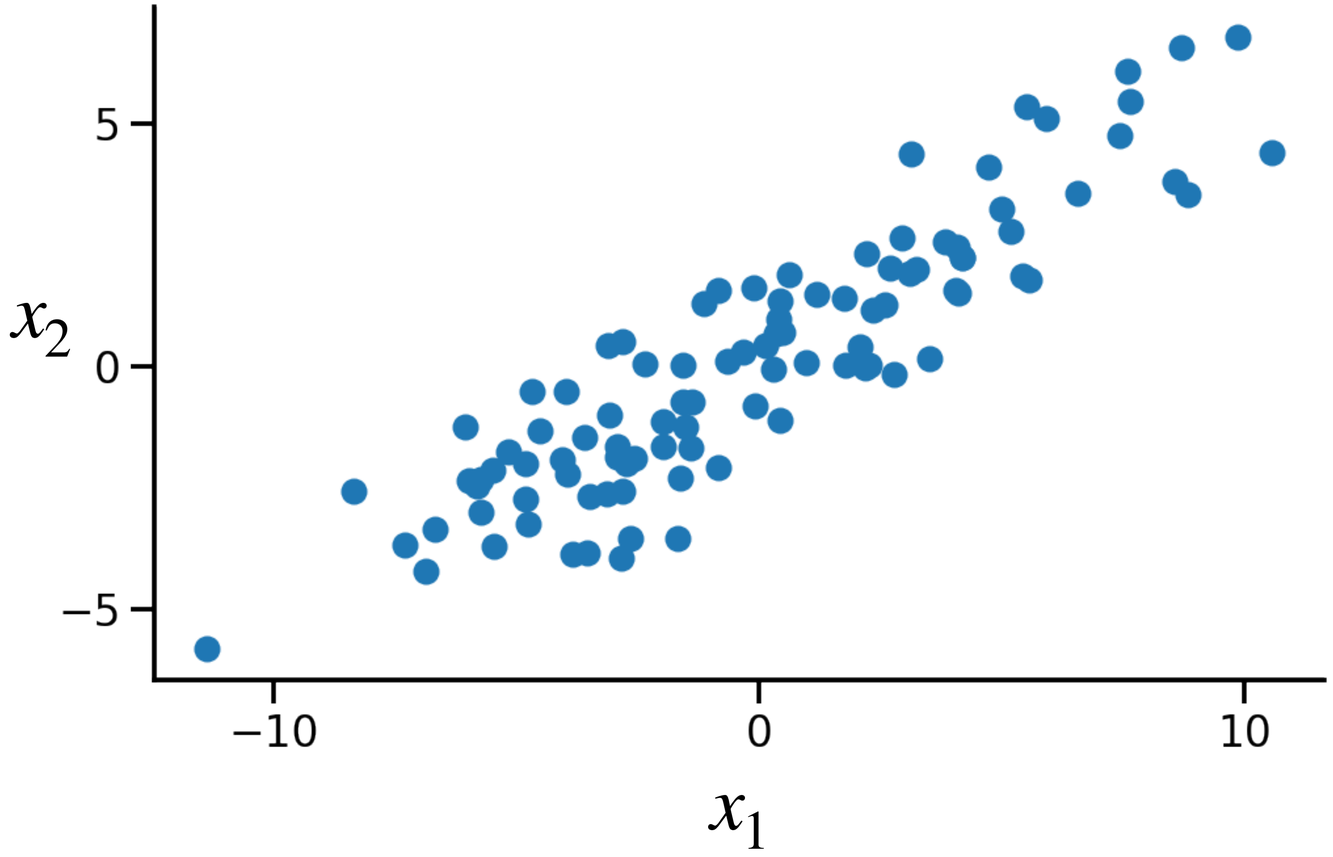
```
[[  8.82,   0.4 ],
 [  4.89,   2.24],
 [  9.34,  -0.98],
 [  4.75,  -0.15],
 [ -0.52,   0.41],
 [  0.72,   1.45],
 [  3.81,   0.12],
 [  2.22,   0.33],
 [  7.47,  -0.21],
 [  1.57,  -0.85],
 [-12.76,   0.65],
 [  4.32,  -0.74],
 [ 11.35,  -1.45],
 [  0.23,  -0.19],
 [  7.66,   1.47],
 [  0.77,  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

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



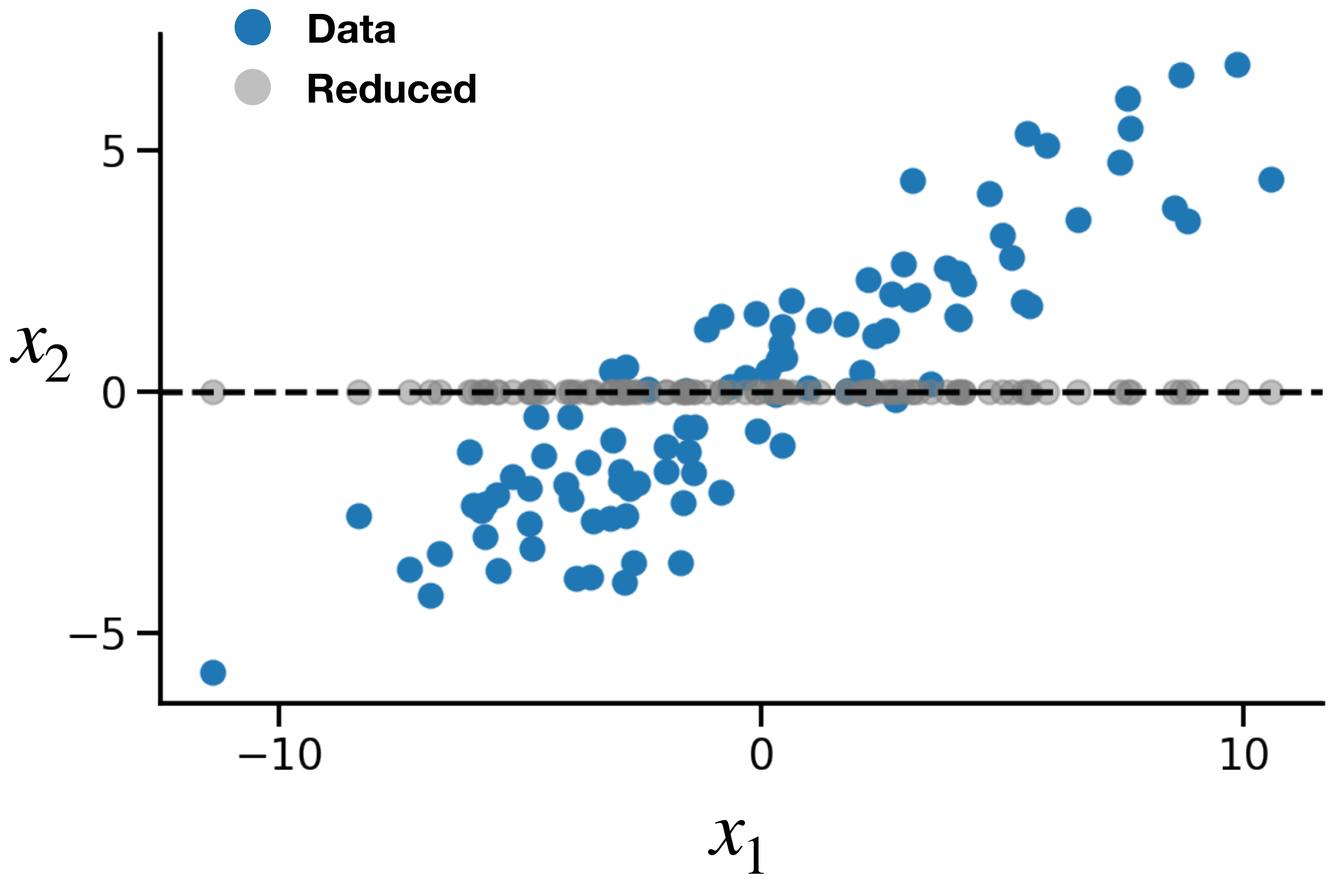
How about now?

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?

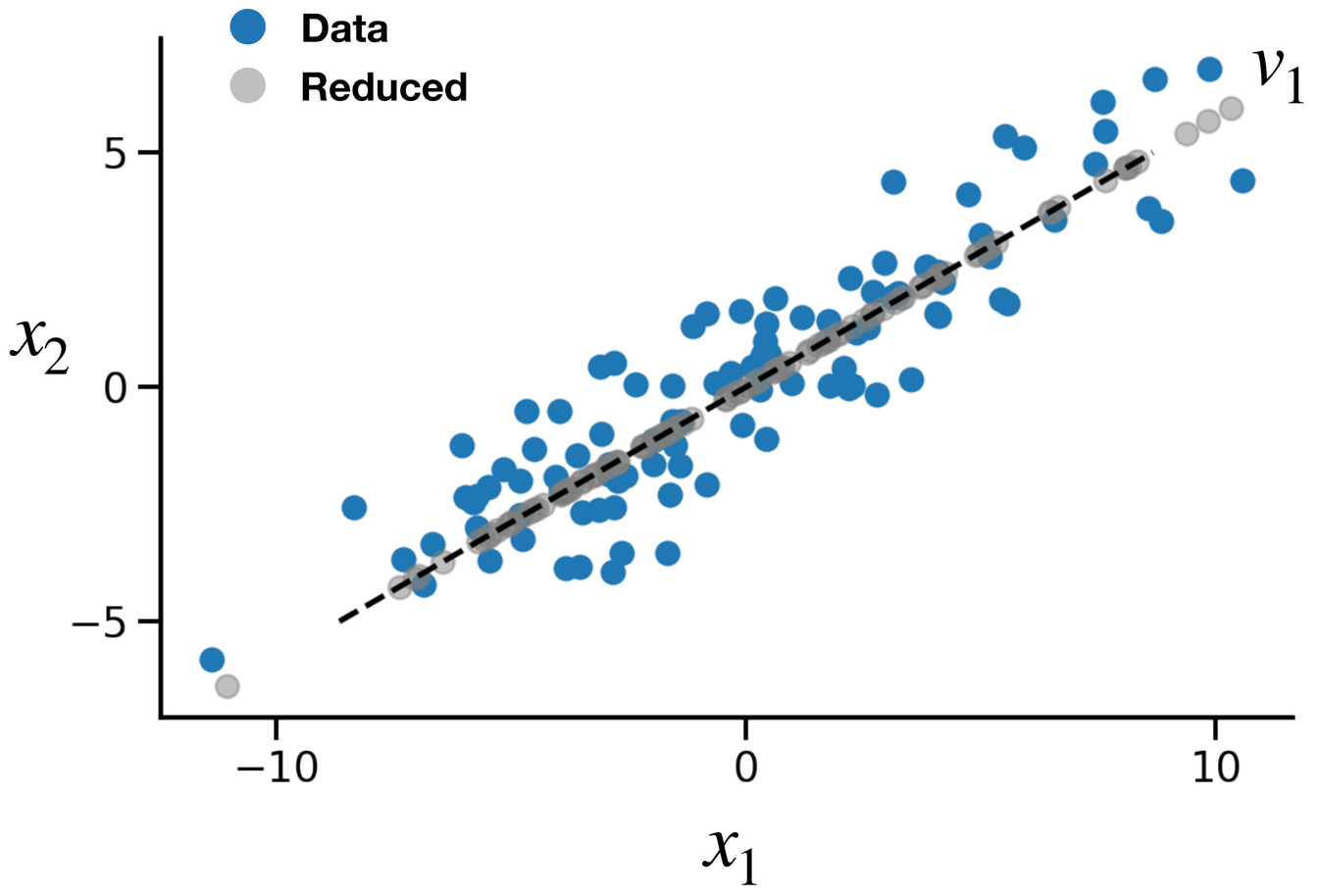
x_1	x_2
7.44	4.76
3.12	4.39
8.58	3.82
4.19	2.24
-0.65	0.1
-0.1	1.62
3.23	2.01
1.76	1.4
6.57	3.56
1.78	0.04
-11.38	-5.82
4.11	1.52
10.56	4.41
0.29	-0.05
5.9	5.1
0.48	0.71
-2.85	-3.93
-1.58	-0.73
4.73	4.12



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],
```



v_1

```
[[  8.82],
 [  4.89],
 [  9.34],
 [  4.75],
 [ -0.52],
 [  0.72],
 [  3.81],
 [  2.22],
 [  7.47],
 [  1.57],
 [-12.76],
 [  4.32],
 [ 11.35],
 [  0.23],
 [  7.66],
 [  0.77],
 [ -4.44],
 [ -1.74],
 [  6.15],
```

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):

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

Principal Component Analysis

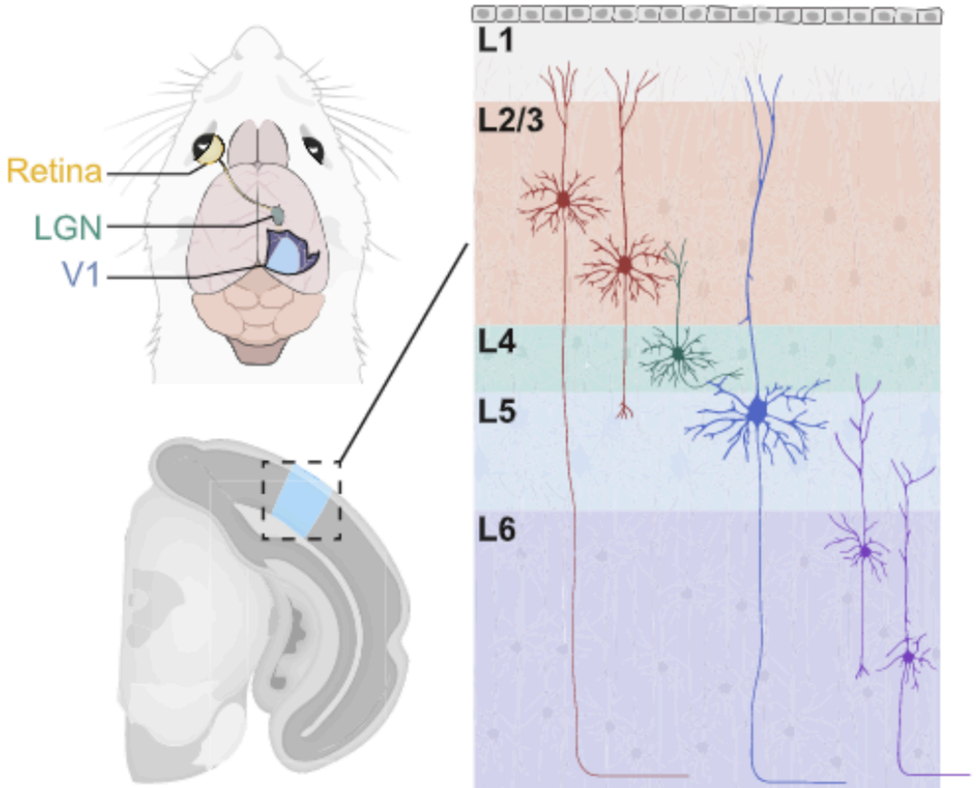
- Reduce data dimensions by removing low-variance basis after a rotation.
- Identify the “best” basis in the feature space to project data points.
- Amounts to a Singular Value Decomposition (SVD):

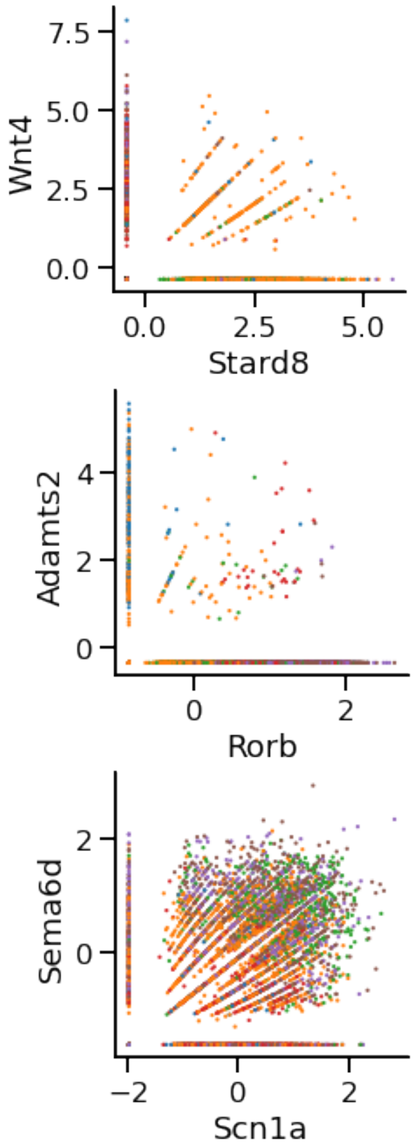
$$X = USV^T = \underbrace{u_1 s_1}_{\text{scores}} \cdot \underbrace{v_1^T}_{\text{basis}} + u_2 s_2 \cdot v_2^T + \ldots + u_r s_r \cdot v_r^T$$

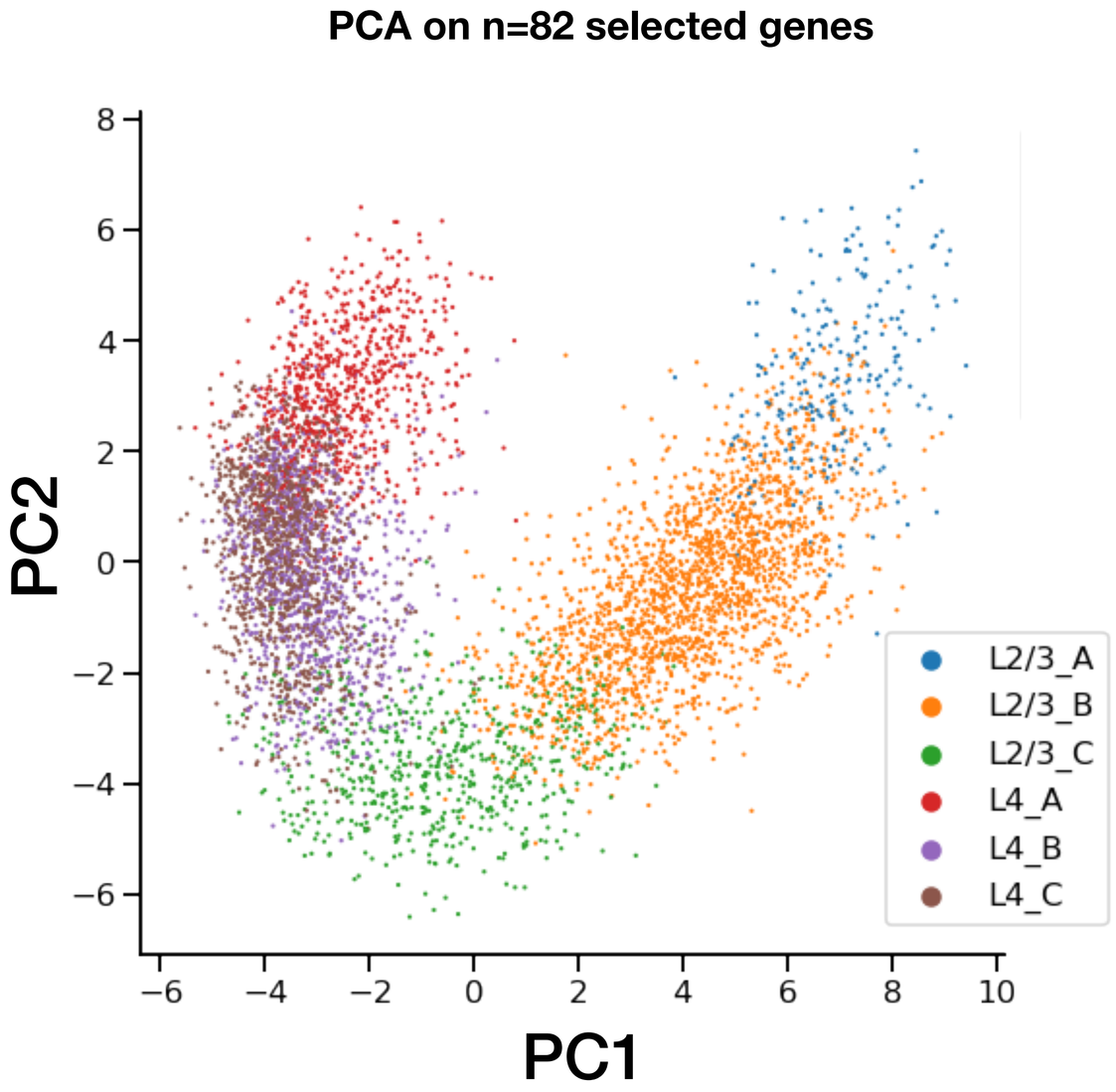
standard deviation

Works in more than 2D!

Application: Transcriptional state of cortical neurons

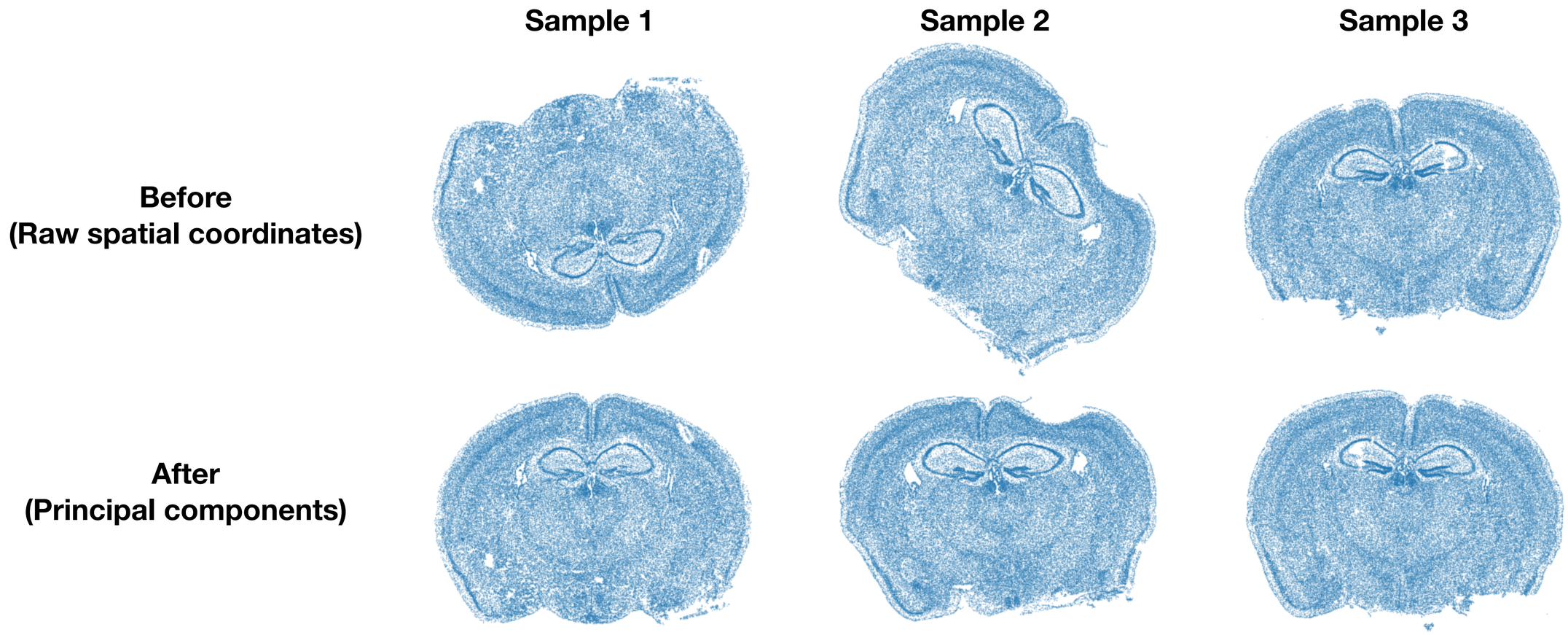






Data from Cheng et al. 2022 *Cell*; Zipursky lab, UCLA

Application: rotate the mouse brain



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

Discussion / Bonus point

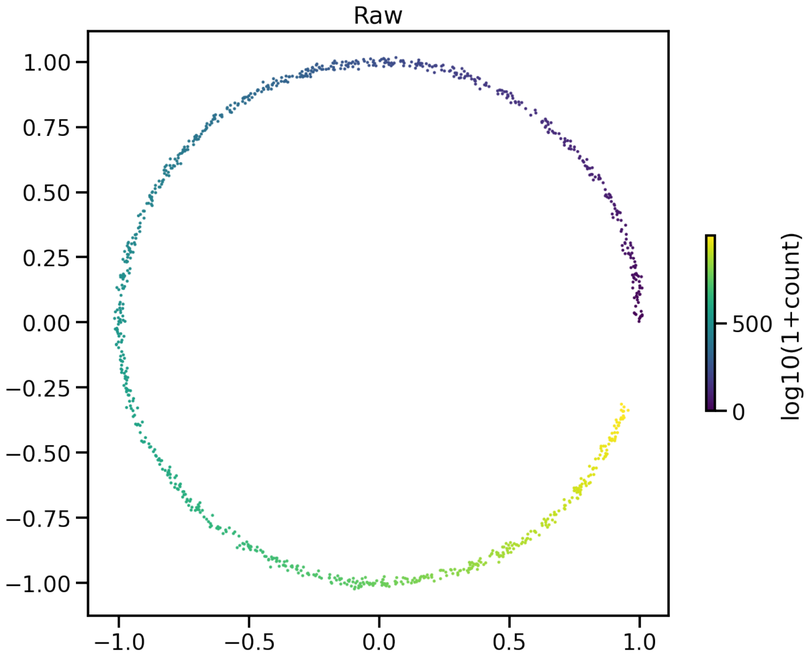
- When does PCA fail?

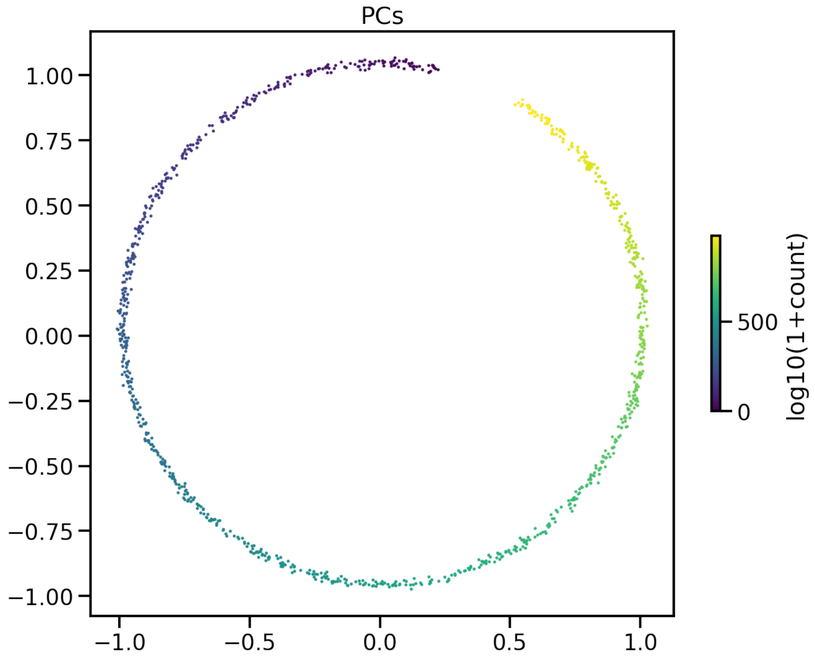
UMAP

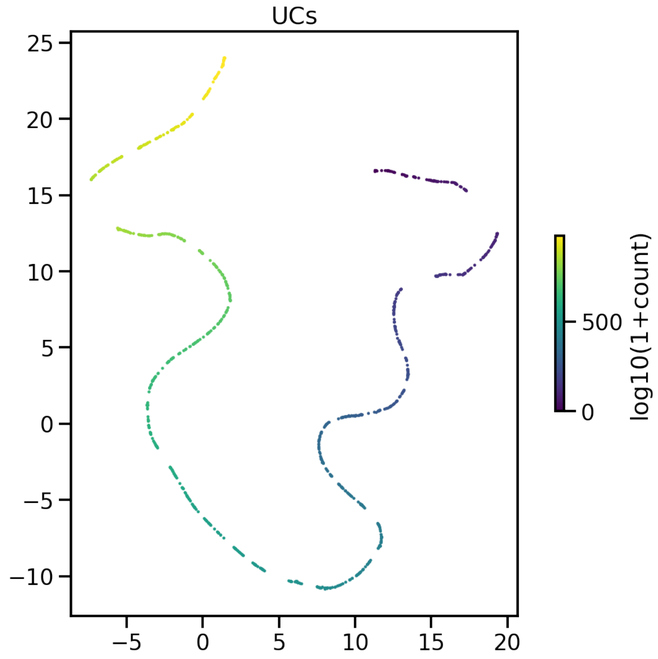
- **UMAP**: uniform manifold approximation and projection for dimension reduction
  - Complex, non-linear, flexible, slower but reasonably fast
  - Preserves local structure (neighborhood) as much as possible
- Steps:
  - Represents high-dim data points by a neighborhood graph
  - Embed the graph to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

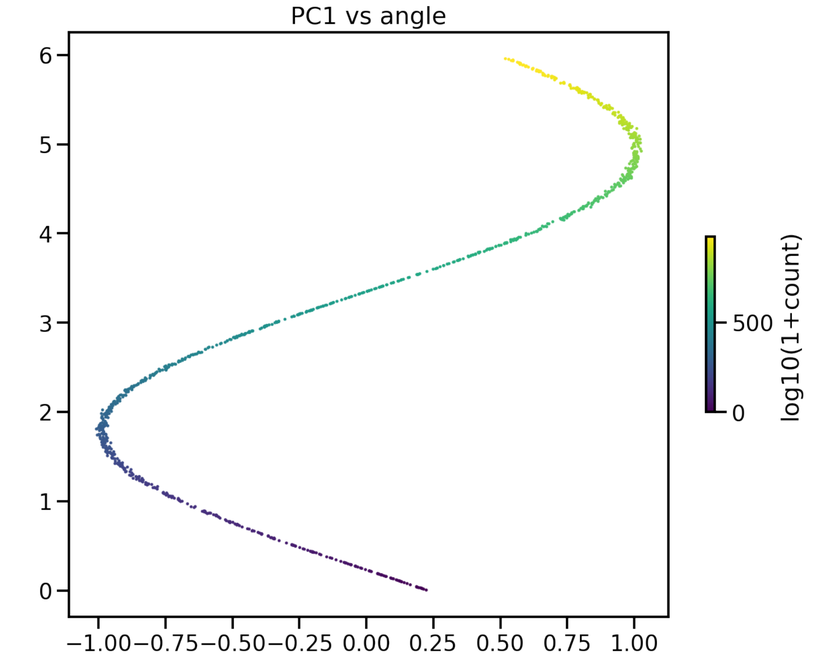
Revealing PCA vs UMAP with a toy model

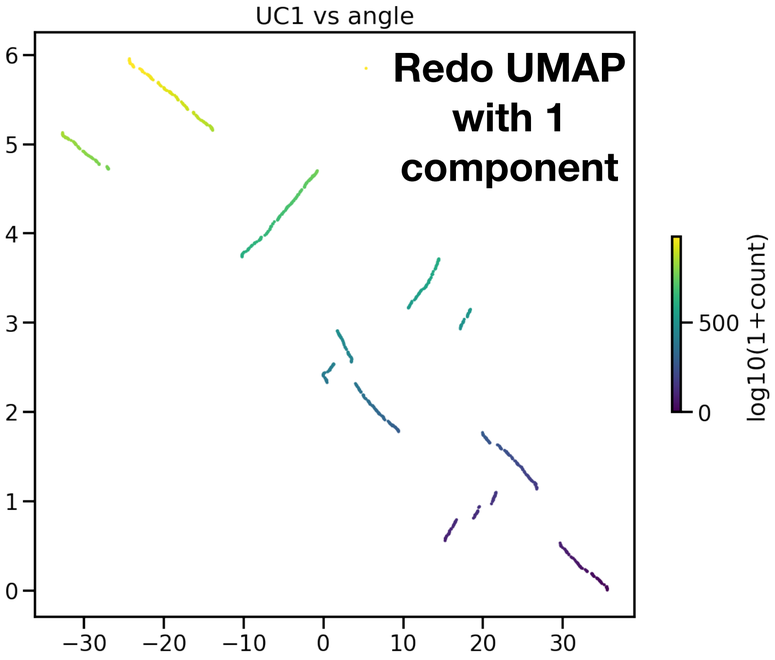
- How does this toy model relate everything we learned about PCA and UMAP?





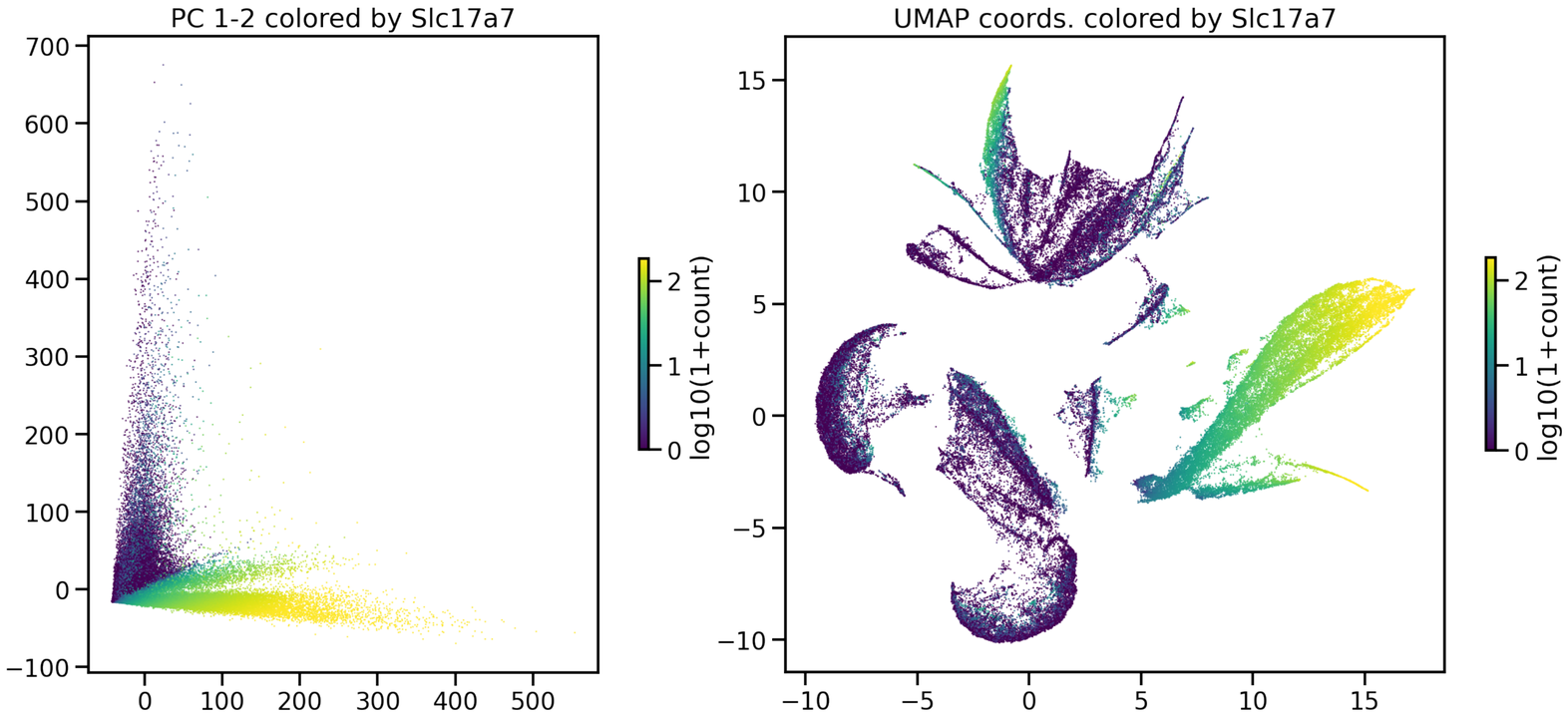






Exercise 2: apply PCA and UMAP to real data (Vizgen)

- 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.
- How much variance does each of the 20 PCs accounts for? How much in total?
- What seems to be special about the gene “Slc17a7”?

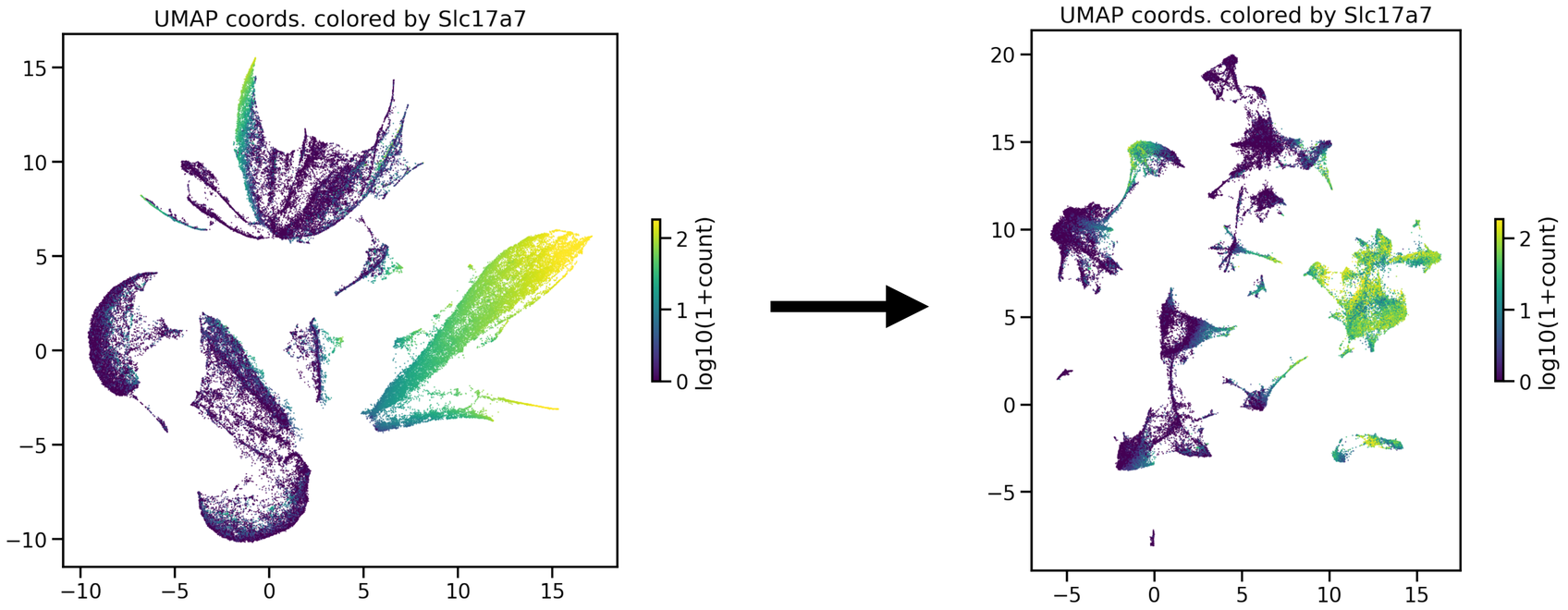


Something unsatisfying:

Graded patterns are all over the place, suggesting continuous changes that are likely technical variation.

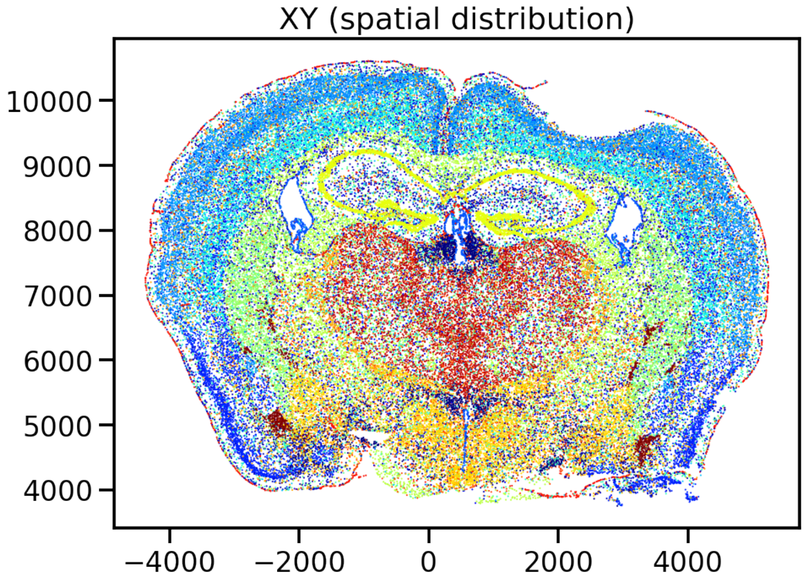
Exercise 3: Normalization

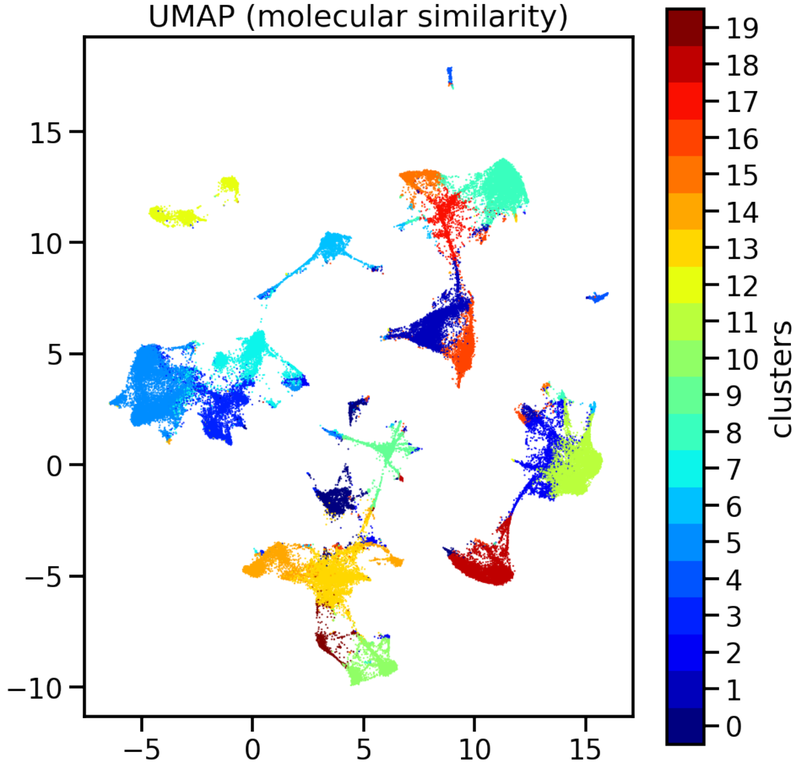
- Neither PCA nor UMAP are magic — they cannot separate technical noise from biological signals.
- We need to normalize the raw data to remove known technical artifacts.
- In single-cell transcriptomics, a rule of thumb is to normalize raw counts by cell size (or cell library size) and do log transformation.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



Exercise 4: Clustering - KMeans

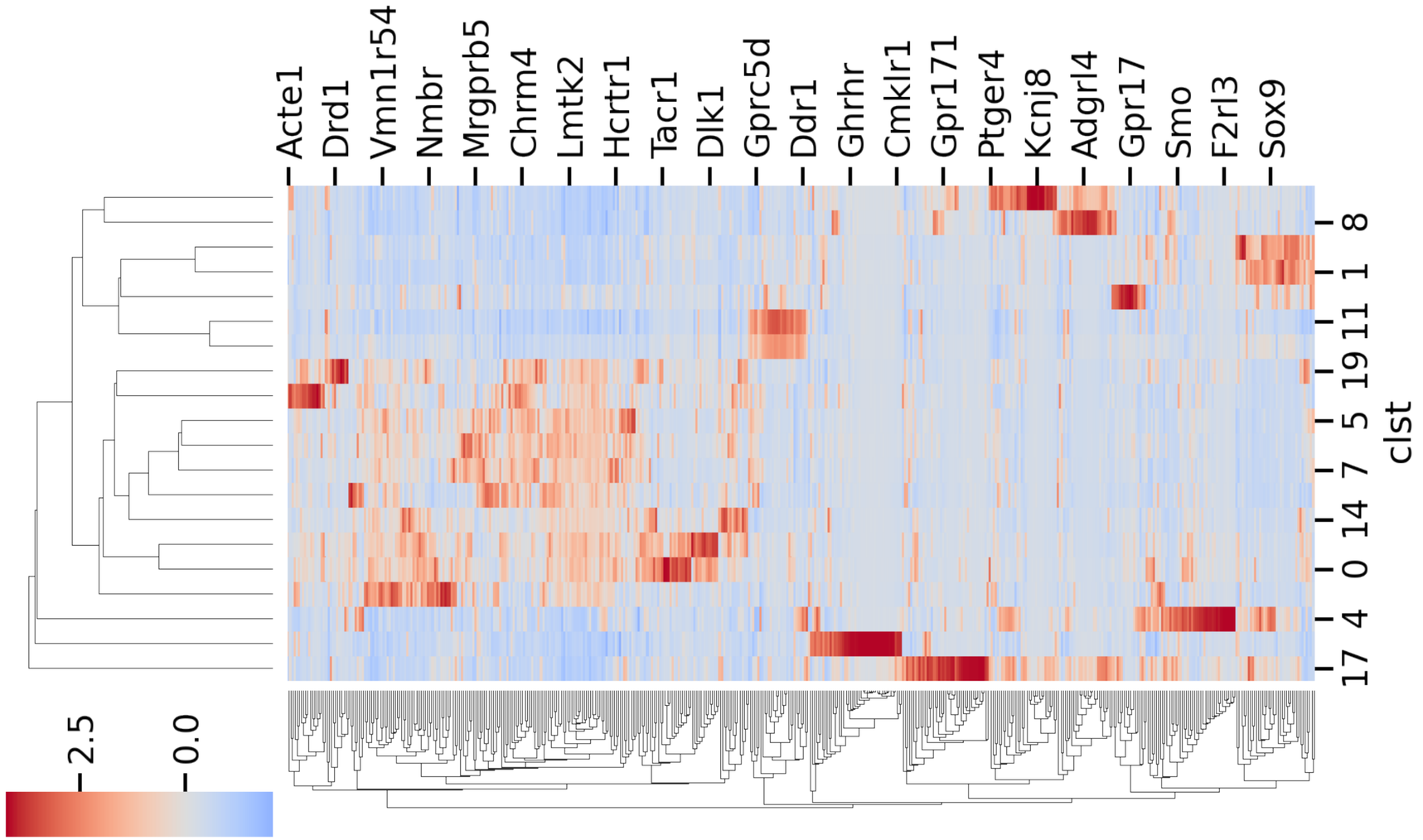
- Group cells by their similarity in gene expressions
- K = number of clusters (cell types) you aim to identify (based on prior knowledge or trial and error)
- Procedure: Repeat until convergence
  - 1. compute cluster centroids;
  - 2. assign cells to the cluster with the closest centroid.





Bonus exercise 1:
Cell-type centroid gene expression profiles

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

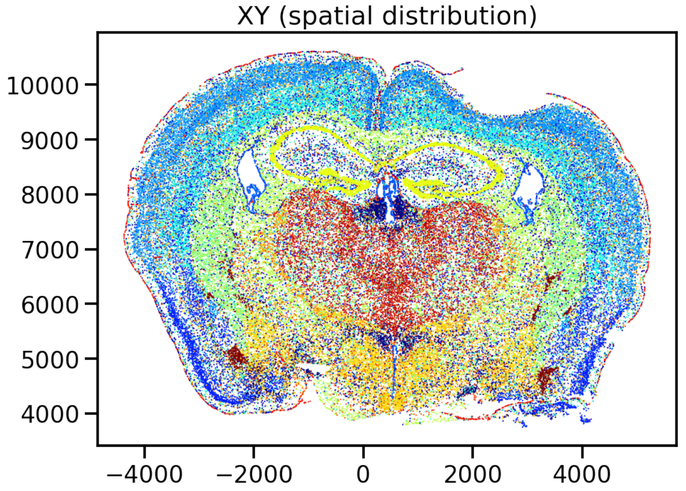


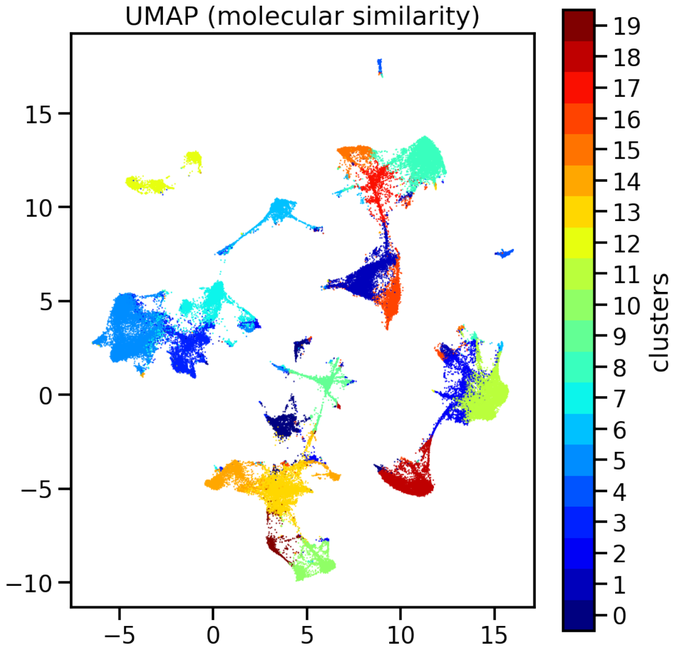
Bonus exercise 2:
Spatial neighborhood enrichment analysis

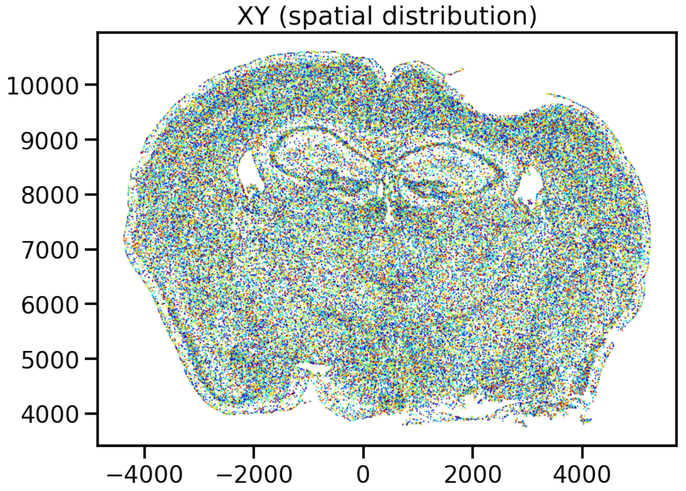
- Are some cell types close to some other cell types? Or are cell types distributed randomly in space?
- How do we get at this question?
- Related paper: Fang et al. 2022 Science — Figure 4

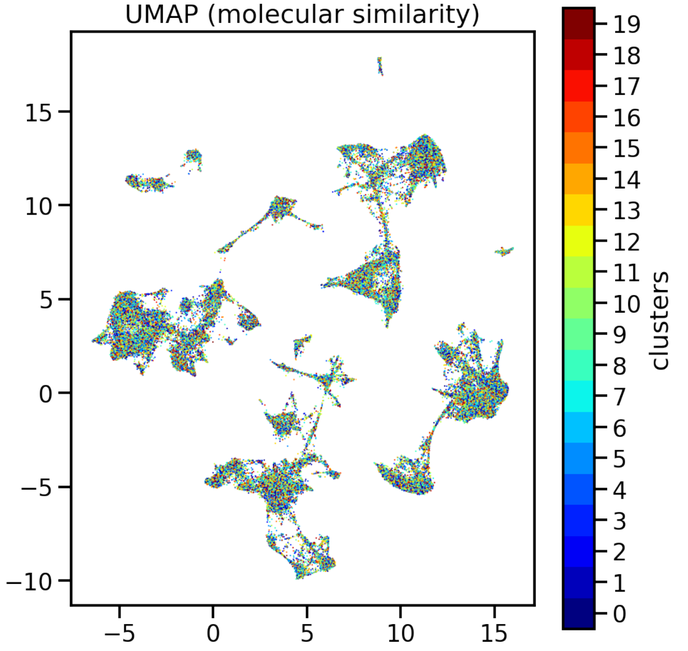
Data shuffling is a powerful trick to test hypothesis

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

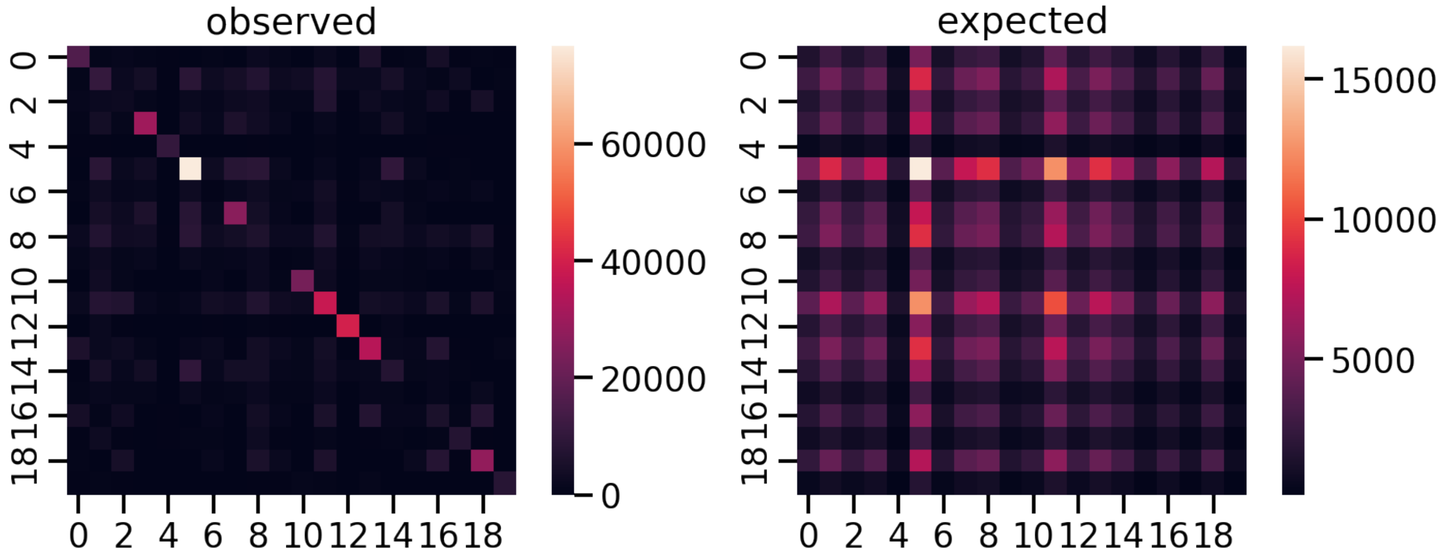




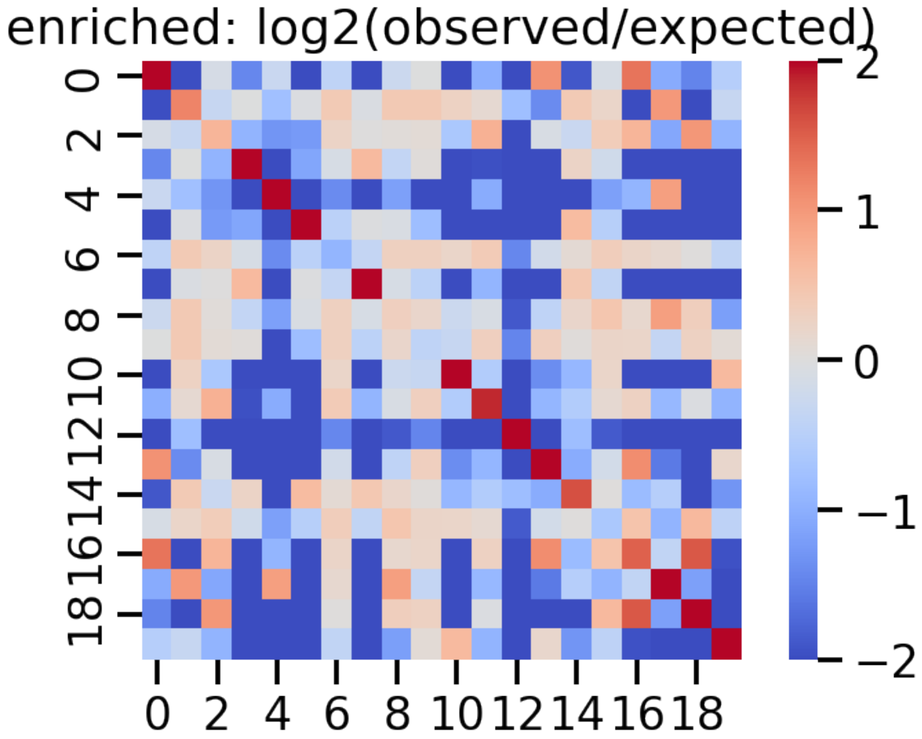




Data shuffling is a powerful trick to test hypothesis



- 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 hypothesis

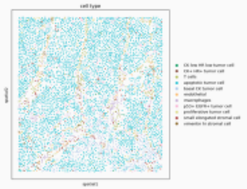


Canned analysis tool: Squidpy

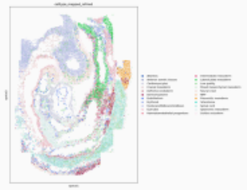
- https://squidpy.readthedocs.io/en/latest/tutorials.html
- Hardest step (for me) is to install the package.

Analysis of spatial datasets using `squidpy`

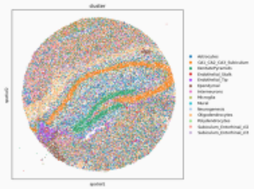
This section contains tutorials showcasing core Squidpy functionalities by applying them to a diverse set of different spatial datasets.



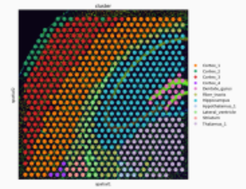
Analyze Imaging Mass Cytometry data



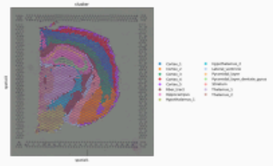
Analyze seqFISH data



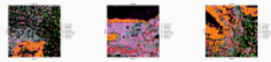
Analyze Slide-seqV2 data



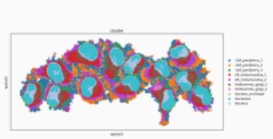
Analyze Visium fluorescence data



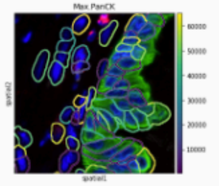
Analyze Visium H&E data



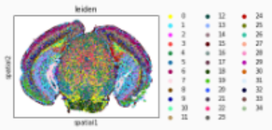
Analyze MIBI-TOF image data



Analyze 4i data



Analyze Nanostring data



Analyze Vizgen data

Thank you!

Everything should be made as simple as possible, but not simpler.

Day 1 email (skip)

Dear all,

Here are the materials for today's workshop (W31; spatial transcriptomics).

- Attendance form (only if you need to be graded): https://forms.gle/MfuKUz5628JPryYR6
- Papers to be discussed today (attached).

Please always let me know if you have questions!

Best,
Fangming

Day 2 email (skip)

Hi all,

Here are some links that will be used for today's workshop.
- Attendance form (only if you need to be graded): https://forms.gle/YWezo7bsjYJ4qQLV7
- The GitHub repo for this class: https://github.com/FangmingXie/collab-workshop-st

Best,
Fangming

Email after class

Hi all,

Thanks for coming to the workshop! This email contains additional information to wrap up this class.

For those who need to be graded, here are the assignments (due on March 24 Friday):
- Take-home quiz: https://forms.gle/xrGtSdJ9c9sDwbxp6
- Homework: https://github.com/FangmingXie/collab-workshop-st/blob/main/hw/hw.ipynb

I designed it with the hope that they should not take you more than 30 minutes to complete. But please let me know and ask for help if you spent more than 1 hour on it.

And here are the full solutions to the coding exercises:
- https://github.com/FangmingXie/collab-workshop-st/blob/main/trash/solution1.ipynb
- https://github.com/FangmingXie/collab-workshop-st/blob/main/trash/solution2-5.ipynb

Finally, below is the class evaluation QR code, where you can provide feedback anonymously about this class: