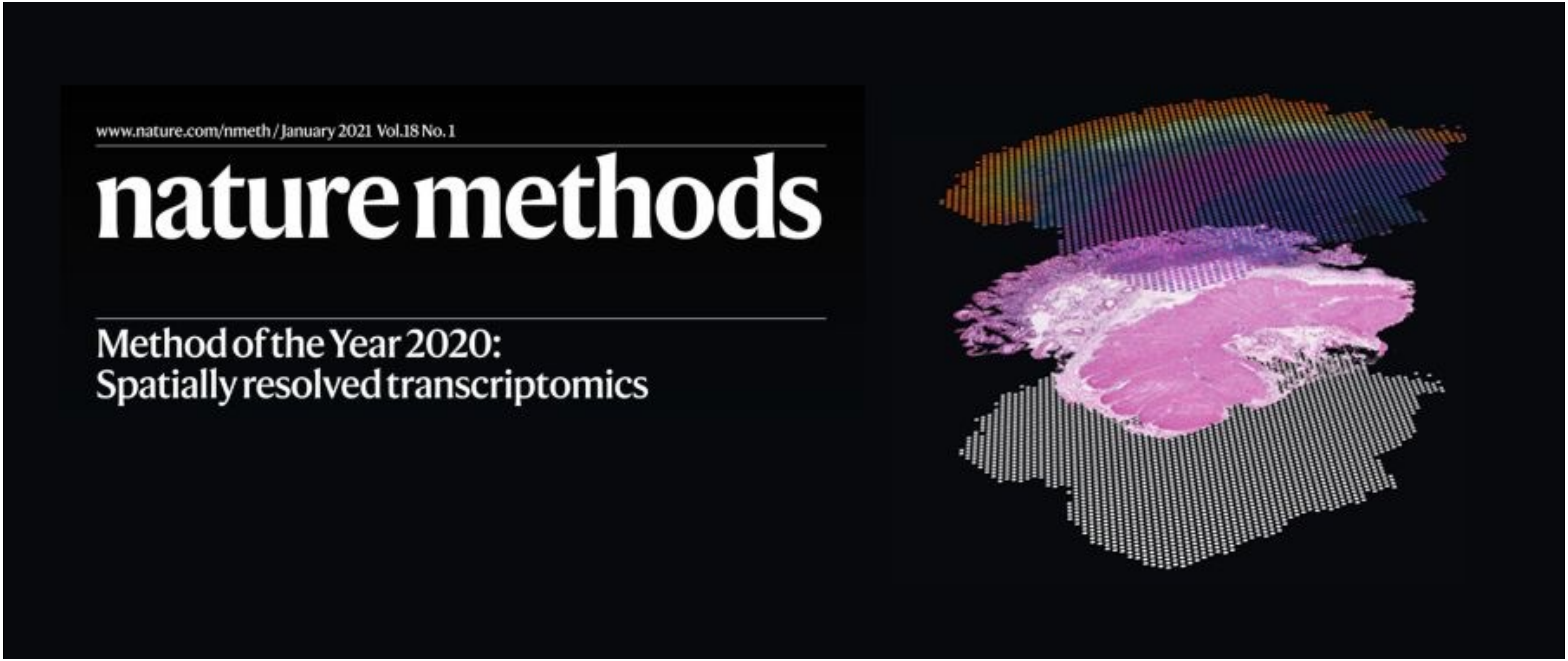
UNIT 3 TRANSCRIPTOMICS

Spatial Transcriptomics

April 28, 2022

Outline

- Spatial Transcriptomics
  - Sequencing based techniques
    - 10X Visium
  - Imaging based techniques
    - MERFISH
- Encoding of sequence data
  - Hemming code
  - One Hot
  - Simplex encoding



www.nature.com/nmeth / January 2021 Vol.18 No. 1

nature methods

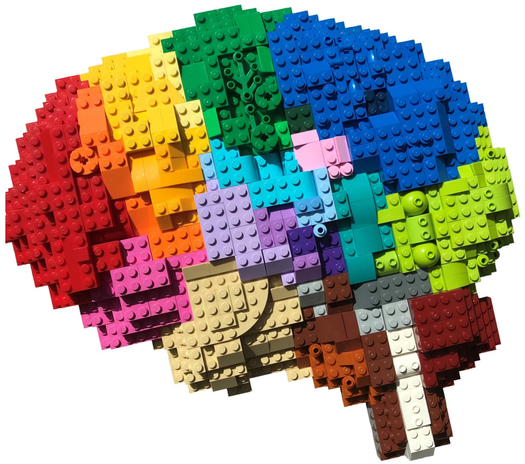
Method of the Year 2020:
Spatially resolved transcriptomics

Single-cell and Spatial Transcriptomics

Bulk transcriptomics

Single-cell transcriptomics

Spatial transcriptomics



Physiological reconstruction

Image credit: Bo Xia @BoXia7

Dimensionalities in transcriptomes

- Samples
- Transcripts / Genes
- Cells / Nucleus
- Spatial Locations
- Time / Differentiation stage

Spatial transcriptomics technologies

- **Sequencing based**

- Major steps
  - – 1. Dissection, capturing
  - – 2. Barcoding, sequencing

- Examples
  - – 10X Visium
  - – Slide-seq
  - – Nanostring GeoMx

- **Imaging based**

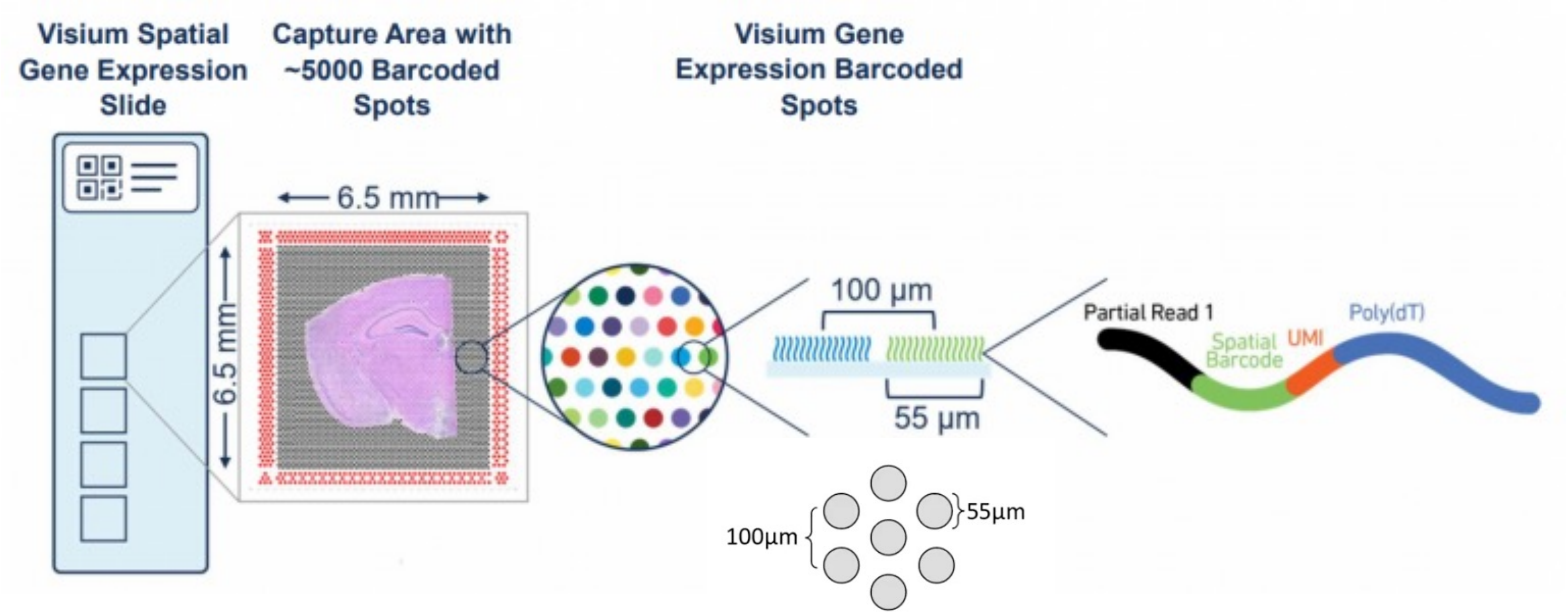
- Major steps
  - – 1. Target and probe design
  - – 2. Fluorescence in situ hybridization (FISH)

- Examples
  - – MERFISH
  - – seqFISH

Spatial transcriptomics technologies

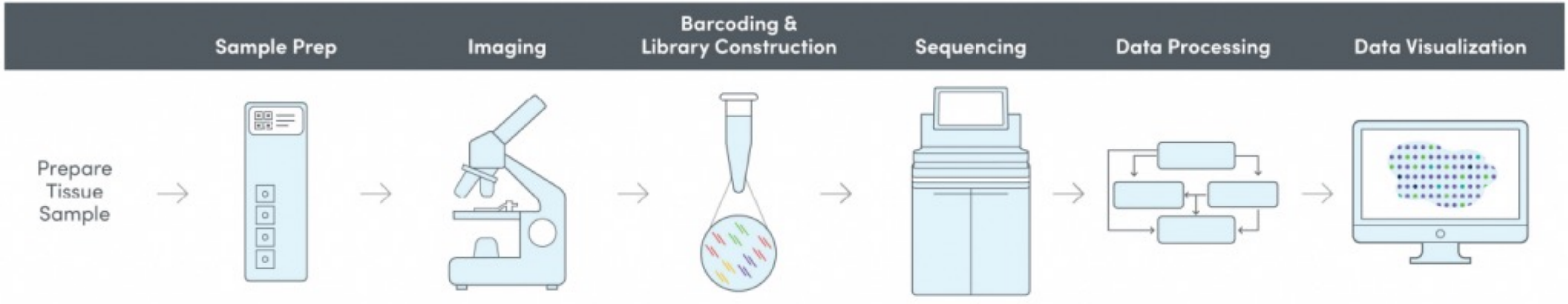
	Sequencing based	**Imaging based**
Pros	• Transcriptome-wide coverage • Easy scale-up • Sequencing data analysis	• Single-cell/single-molecule • High spatial resolution (<1*μ*m) • Continuous spatial locations
Cons	• Fixed spatial dissection • Low spatial resolution (~100*μ*m) • Not single-cell	• Coverage restricted to probes • More difficult experiments • Challenging data analysis

10X Genomics - Visium


Visium Spatial Gene Expression Slide
Capture Area with ~5000 Barcoded Spots
Visium Gene Expression Barcoded Spots
6.5 mm
6.5 mm
100 µm
55 µm
Partial Read 1
Spatial Barcode
UMI
Poly(dT)
100µm
55µm

10X Genomics

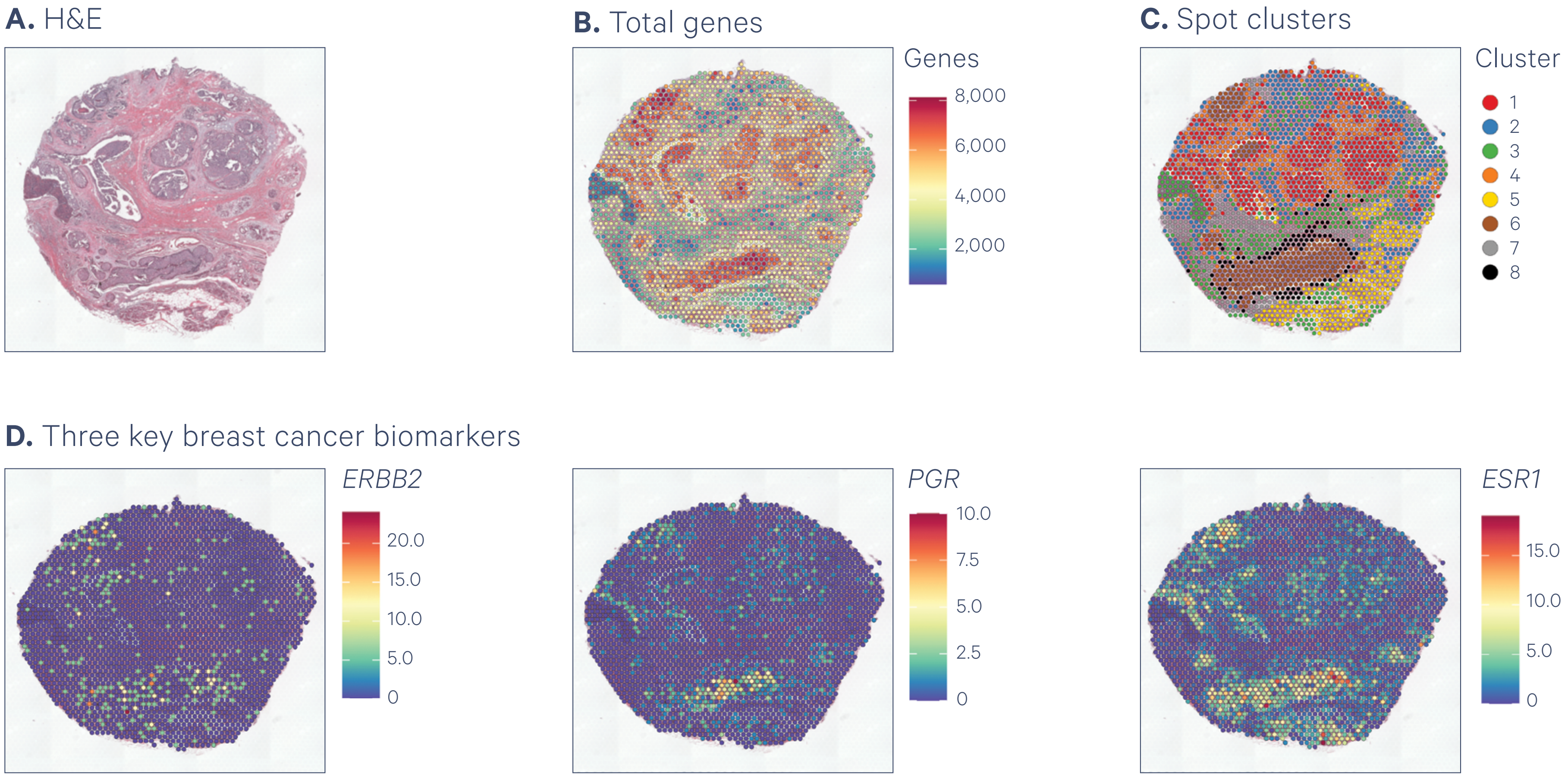
10X Genomics - Visium



10X Genomics

10X Genomics - Visium

Interrogation of ~18,000 genes in a human breast ductal carcinoma in situ FFPE sample



A. H&E

B. Total genes

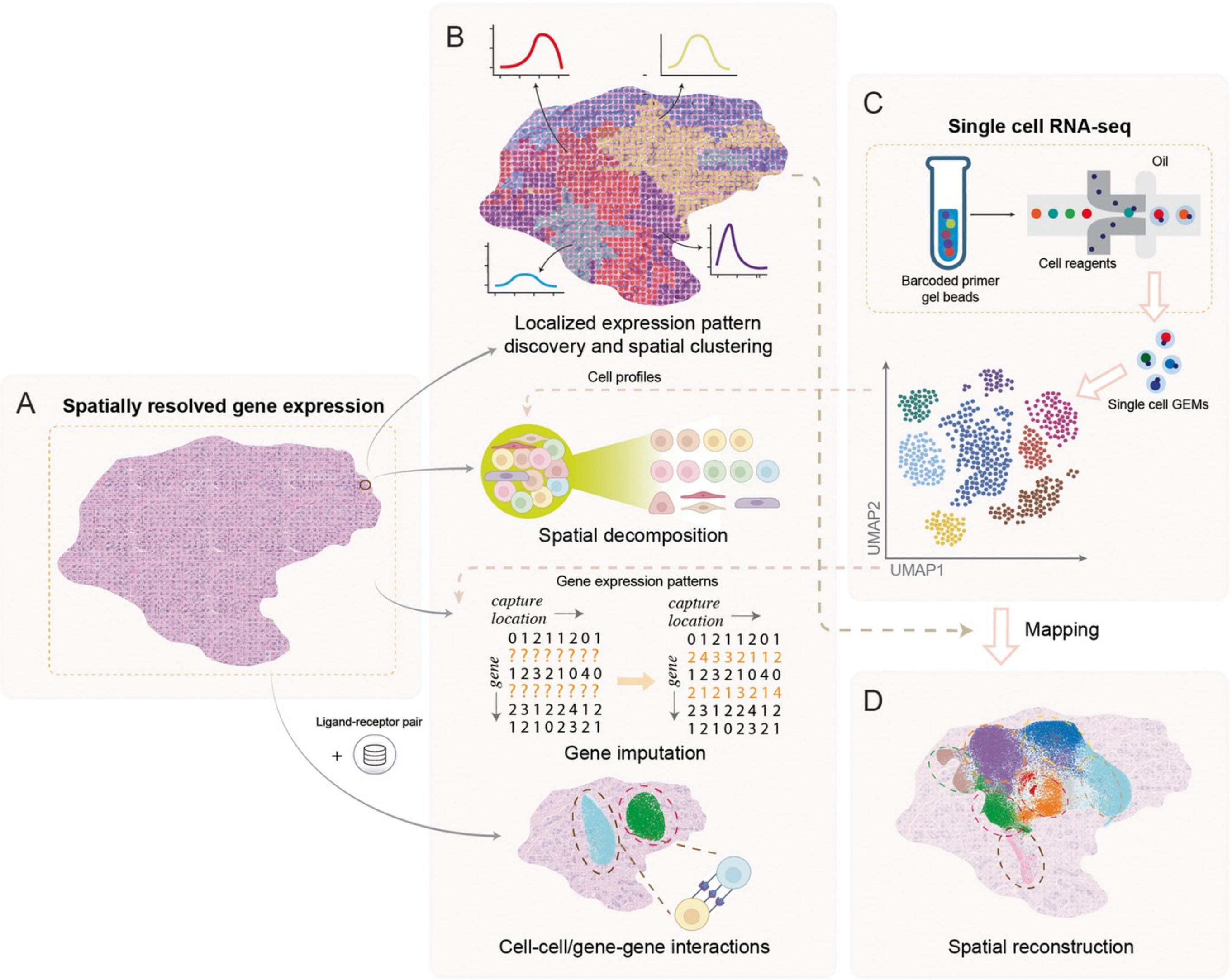
C. Spot clusters

D. Three key breast cancer biomarkers

10X Genomics

Computational Problems

- Localized gene expression profiling
- Spatial clustering
- Spatial decomposition and gene imputation
- Spatial location reconstruction for scRNA-seq
- Cellular interaction or gene interaction inference

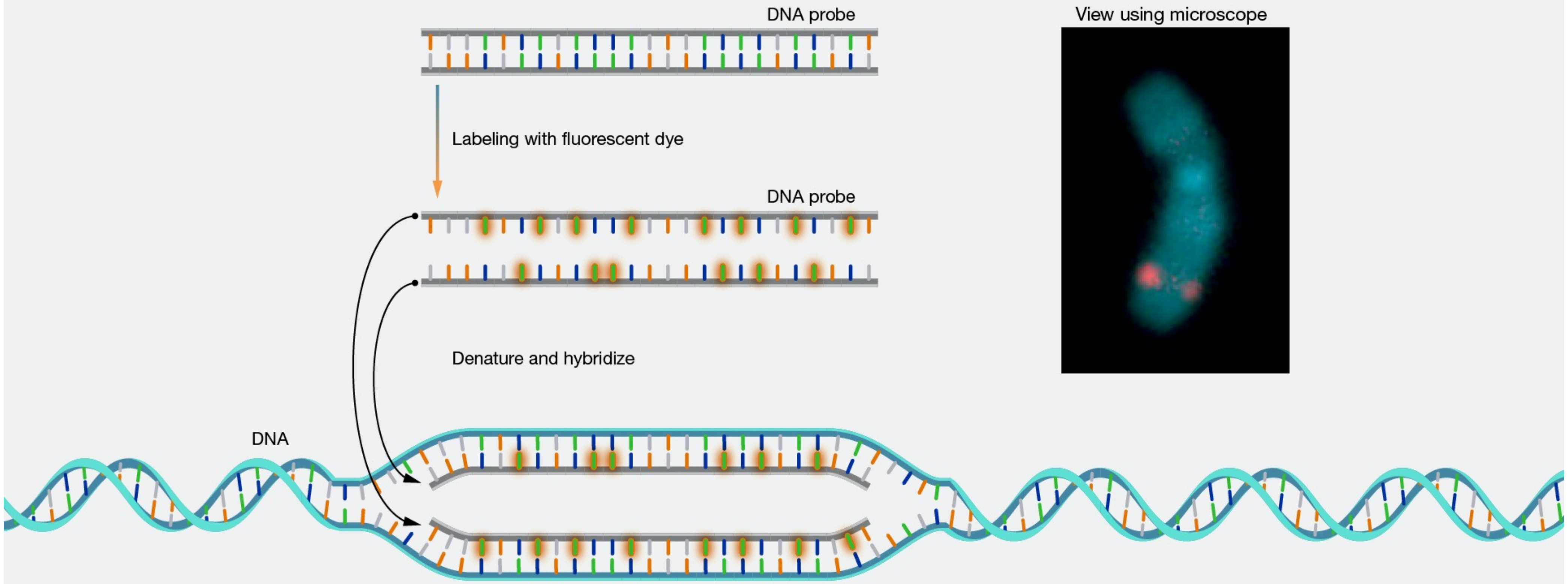


Zeng et al. Genome Bio 2022

MERFISH

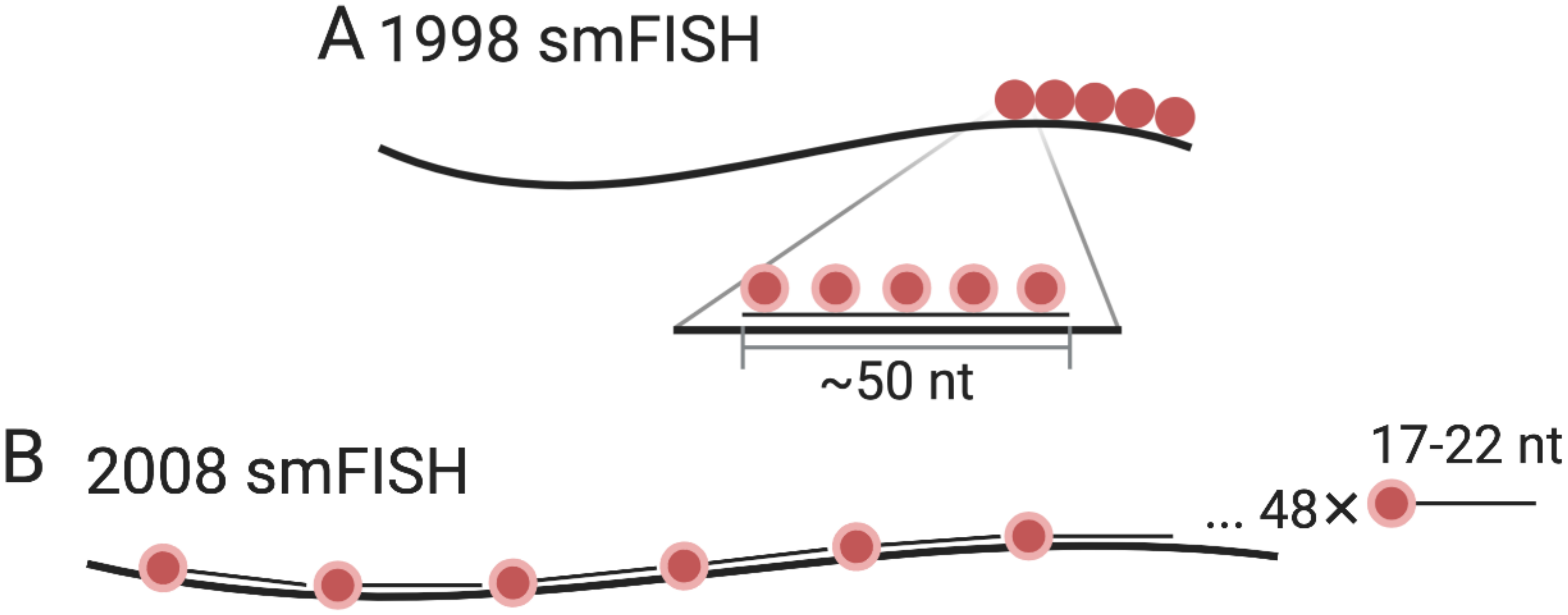
Multiplexed Error-Robust Fluorescence In Situ Hybridization

FISH



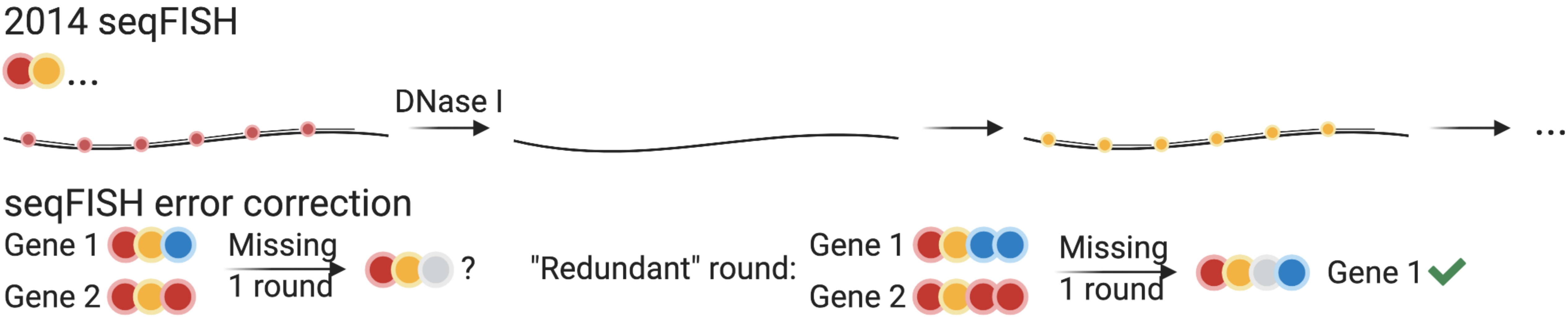
genome.gov

Single-Molecule FISH (smFISH)



Lior Pachter Lab

seqFISH



2014 seqFISH

…

DNase I

…

seqFISH error correction

Gene 1

Gene 2

Missing 1 round

?

"Redundant" round:

Gene 1

Gene 2

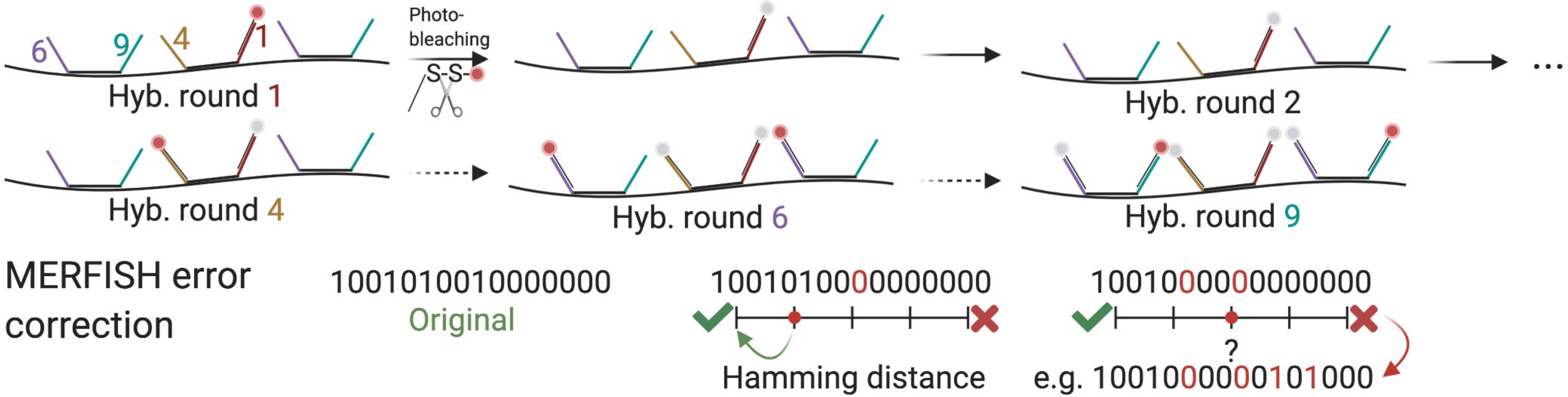
Missing 1 round

Gene 1 ✓

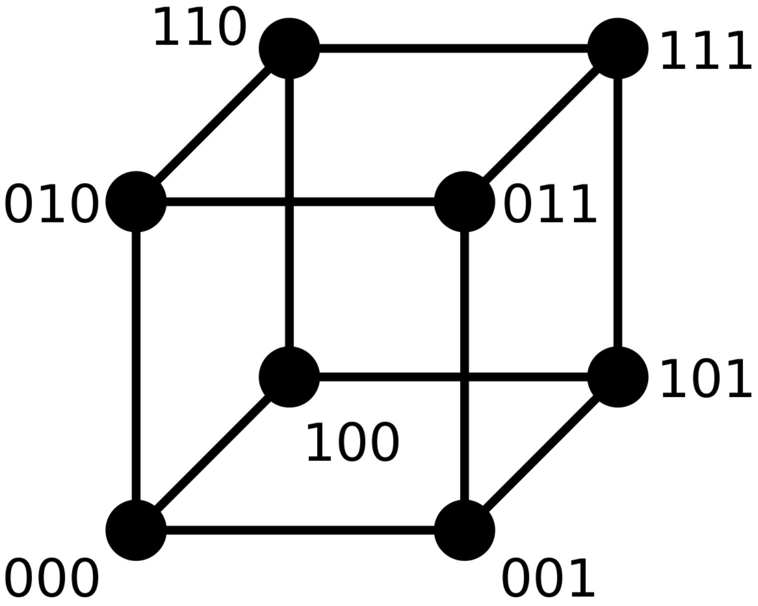
MERFISH

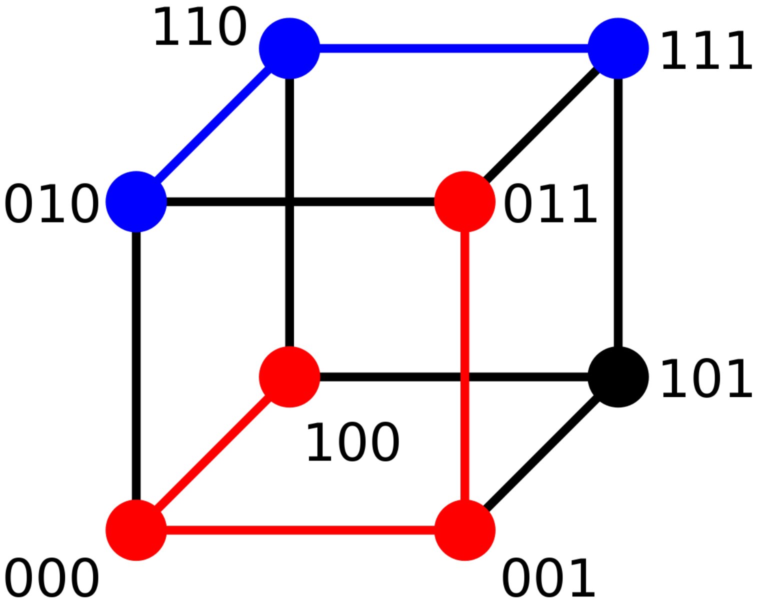
2015 MERFISH

1001010010000000, first two rounds



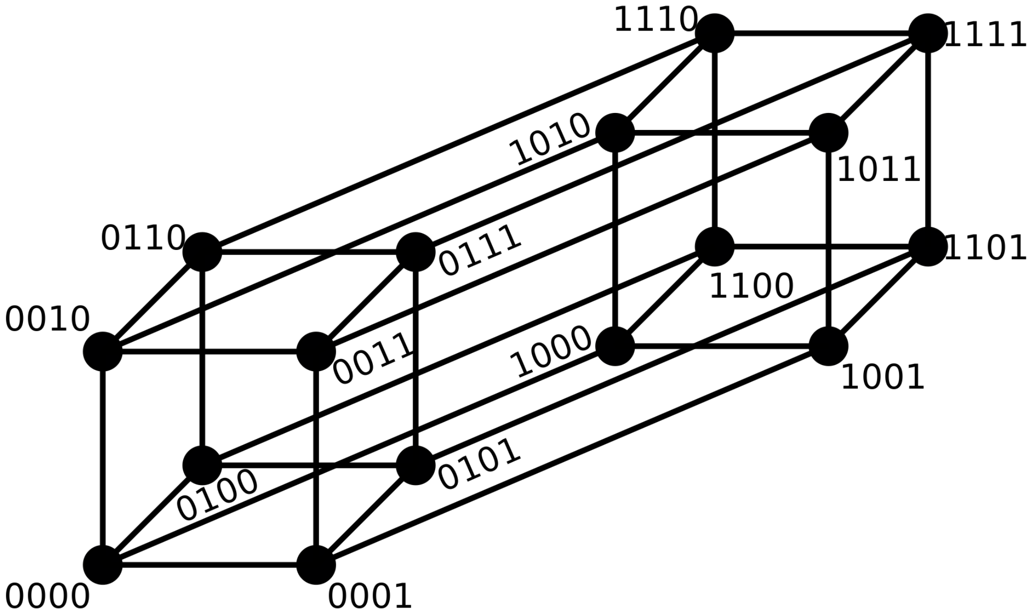
Hamming Distance

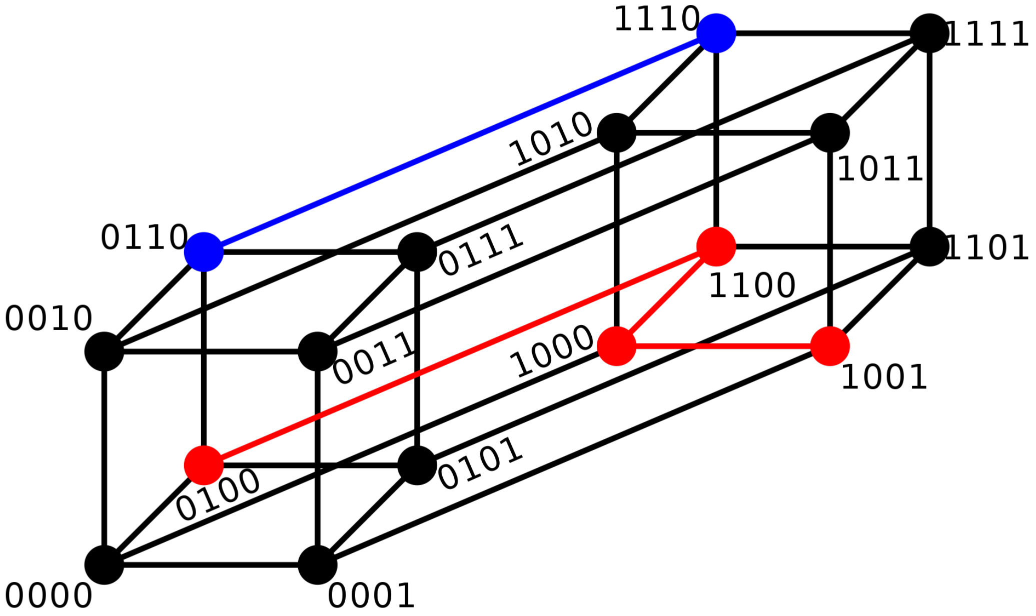




100→011 has Hamming distance 3
010→111 has Hamming distance 2

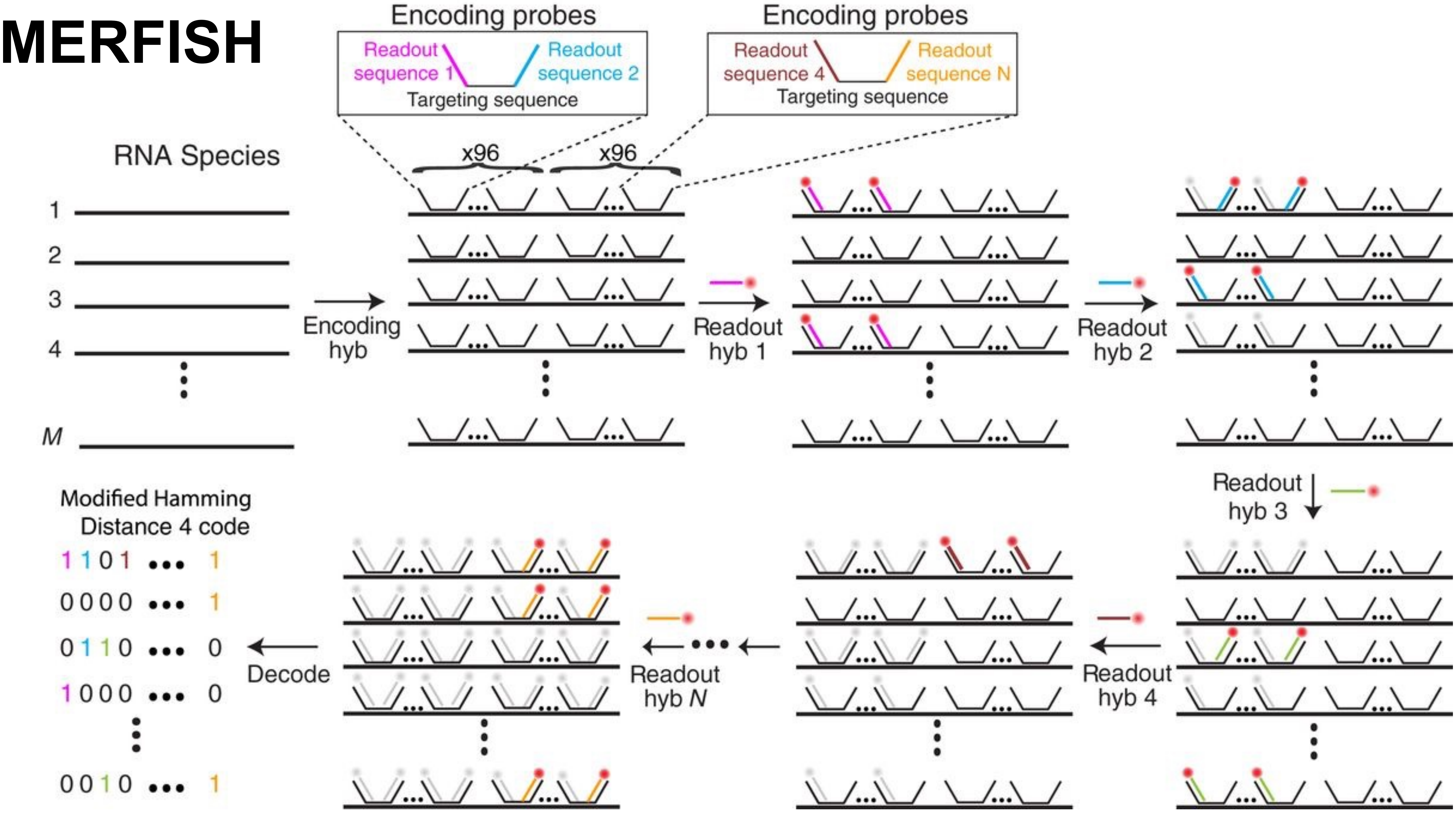
Hamming Distance





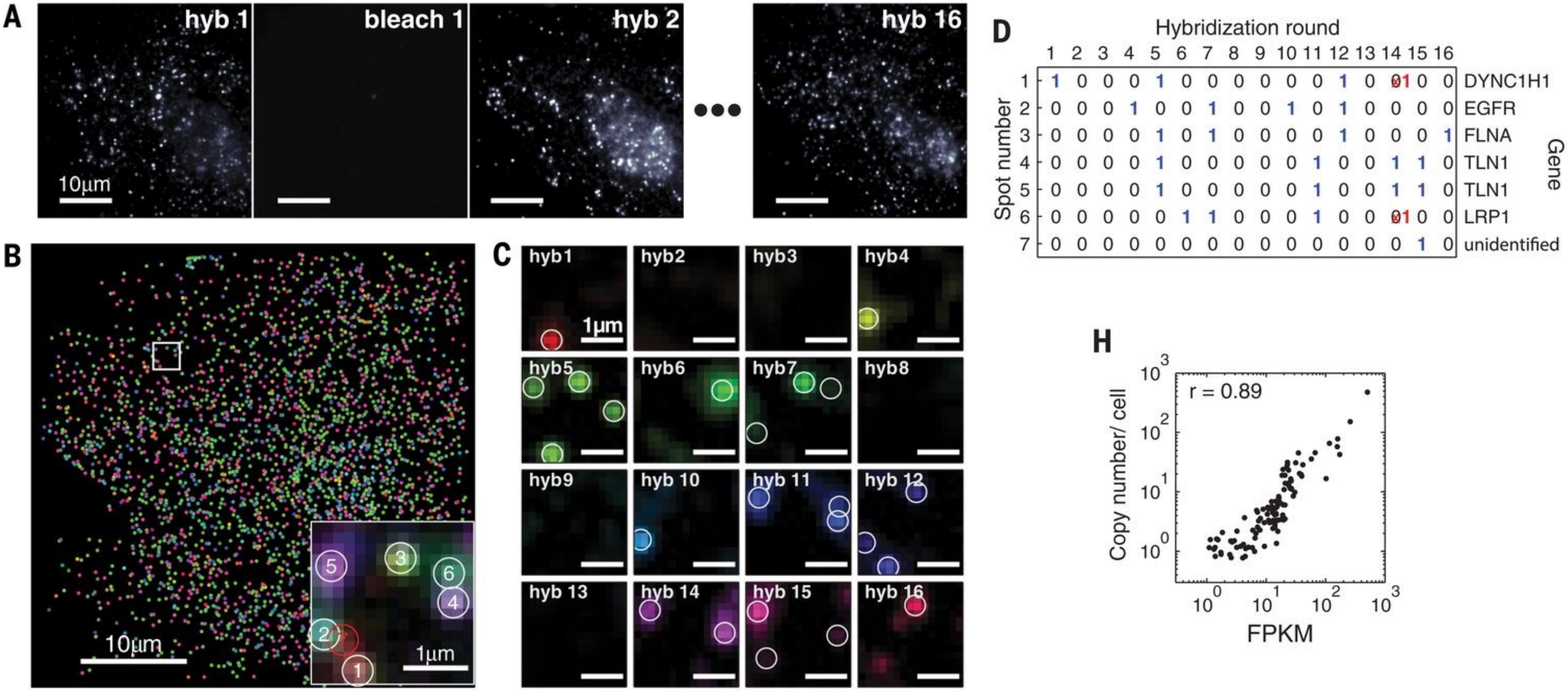
0100→1001 has Hamming distance 3
0110→1110 has Hamming distance 1

MERFISH



Chen et al. Science 2015

MERFISH



Hybridization round

Spot number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Gene
1	1	0	0	0	1	0	0	0	0	0	0	1	0	~~0~~1	0	0	DYNC1H1
2	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	EGFR
3	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	FLNA
4	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	TLN1
5	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	TLN1
6	0	0	0	0	0	1	1	0	0	0	1	0	0	~~0~~1	0	0	LRP1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	unidentified

Chen et al. Science 2015

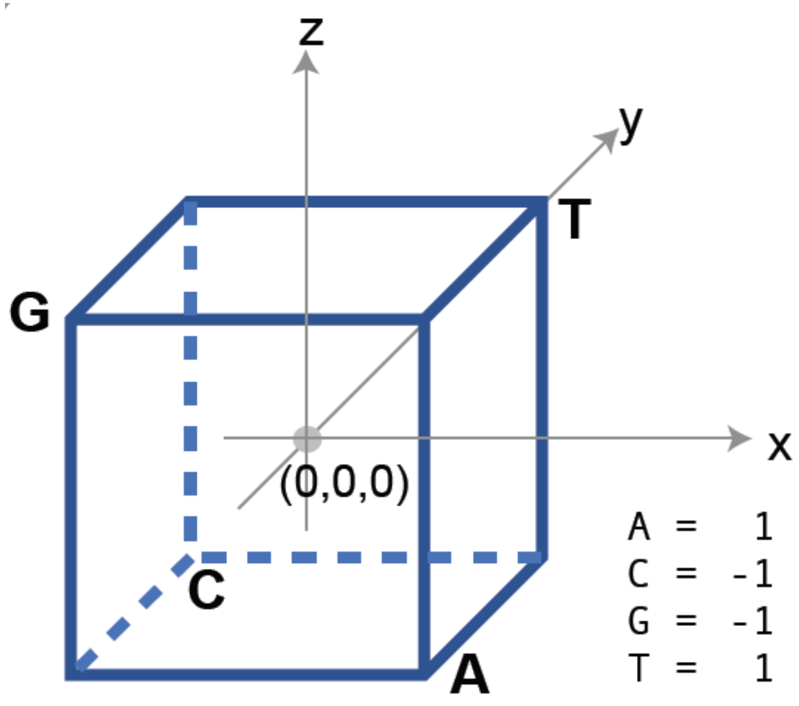
Gray Code and One-Hot Code

Decimal	Binary	Gray	Decimal of Gray	One-Hot
0	0000	0000	0	0000000000000001
1	0001	0001	1	0000000000000010
2	0010	0011	3	0000000000000100
3	0011	0010	2	0000000000001000
4	0100	0110	6	0000000000010000
5	0101	0111	7	0000000000100000
6	0110	0101	5	0000000001000000
7	0111	0100	4	0000000010000000
8	1000	1100	12	0000000100000000
9	1001	1101	13	0000001000000000
10	1010	1111	15	0000010000000000
11	1011	1110	14	0000100000000000
12	1100	1010	10	0001000000000000
13	1101	1011	11	0010000000000000
14	1110	1001	9	0100000000000000
15	1111	1000	8	1000000000000000

One-hot encoding for DNA sequences

	C	G	A	T	A	A	C	C	G	A	T	A	T
A	0	0	1	0	1	1	0	0	0	1	0	1	0
C	1	0	0	0	0	0	1	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0	0	0
T	0	0	0	1	0	0	0	0	0	0	1	0	1

Simplex Encoding (Hadamard)



A	=	1	-1	-1
C	=	-1	1	-1
G	=	-1	-1	1
T	=	1	1	1

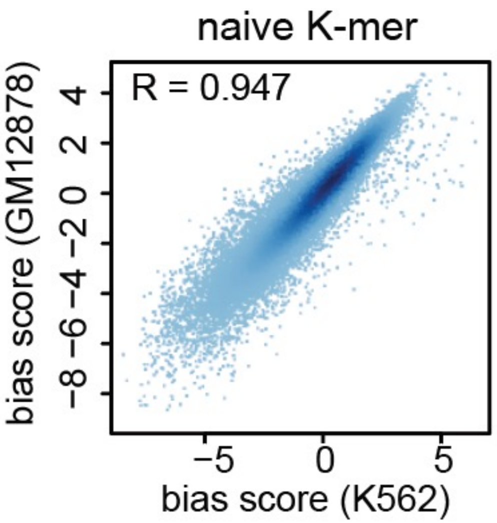
AA	=	A⊗A	=	1	-1	-1	-1	1	1	-1	1	1
AC	=	A⊗C	=	-1	1	-1	1	-1	1	1	-1	1
AG	=	A⊗G	=	-1	-1	1	1	1	-1	1	1	-1
AT	=	A⊗T	=	1	1	1	-1	-1	-1	-1	-1	-1
CA	=	C⊗A	=	-1	1	1	1	-1	-1	-1	1	1
CC	=	C⊗C	=	1	-1	1	-1	1	-1	1	-1	1
CG	=	C⊗G	=	1	1	-1	-1	-1	1	1	1	-1
CT	=	C⊗T	=	-1	-1	-1	1	1	1	-1	-1	-1
GA	=	G⊗A	=	-1	1	1	-1	1	1	1	-1	-1
GC	=	G⊗C	=	1	-1	1	1	-1	1	1	-1	1
GG	=	G⊗G	=	1	1	-1	1	1	-1	-1	-1	1
GT	=	G⊗T	=	-1	-1	-1	-1	-1	-1	1	1	1
TA	=	T⊗A	=	1	-1	-1	1	-1	-1	1	-1	-1
TC	=	T⊗C	=	-1	1	-1	-1	1	-1	-1	1	-1
TG	=	T⊗G	=	-1	-1	1	-1	-1	1	-1	-1	-1
TT	=	T⊗T	=	1	1	1	1	1	1	1	1	1

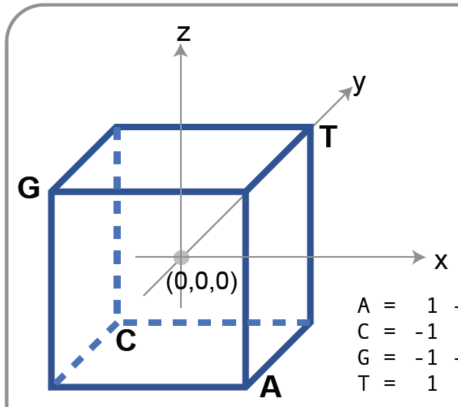
$$p = 1 + 3k + 9(k-1) = 12k - 8$$

Simplex encoding reduces dimensionality

k	naïve k-mer (4^k)	Simplex encoding (12k-8)
4	256	40
6	4096	64
8	65536	88
10	1048576	112

SELMA (Simplex-Encoded Linear Model for Accessible chromatin) improves cleavage bias estimation





```
A =   1 -1 -1
C =  -1  1 -1
G =  -1 -1  1
T =   1  1  1
```

```
AA = A⊗A =  1 -1 -1 -1  1  1 -1  1  1
AC = A⊗C = -1  1 -1  1 -1  1  1 -1  1
AG = A⊗G = -1 -1  1  1  1 -1  1  1 -1
AT = A⊗T =  1  1  1 -1 -1 -1 -1 -1 -1
CA = C⊗A = -1  1  1  1 -1 -1 -1  1  1
CC = C⊗C =  1 -1  1 -1  1 -1  1 -1  1
CG = C⊗G =  1  1 -1 -1 -1  1  1  1 -1
CT = C⊗T = -1 -1 -1  1  1  1 -1 -1 -1
GA = G⊗A = -1  1  1 -1  1  1  1 -1 -1
GC = G⊗C =  1 -1  1  1 -1  1  1 -1  1
GG = G⊗G =  1  1 -1  1  1 -1 -1 -1  1
GT = G⊗T = -1 -1 -1 -1 -1 -1  1  1  1
TA = T⊗A =  1 -1 -1  1 -1 -1  1 -1 -1
TC = T⊗C = -1  1 -1 -1  1 -1 -1  1 -1
TG = T⊗G = -1 -1  1 -1 -1  1 -1 -1 -1
TT = T⊗T =  1  1  1  1  1  1  1  1  1
```

$$p = 1 + 3k + 9(k-1) = 12k - 8$$

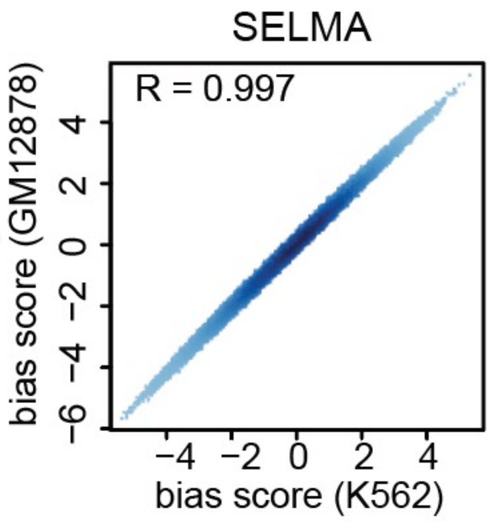
linear regression

$$\begin{array}{cc} \text{naive k-mer bias (y)} & & \text{encoding parameters (X)} \\ \begin{pmatrix} -0.04 \\ -1.01 \\ 0.16 \\ \vdots \\ -1.61 \end{pmatrix} & \sim & \begin{pmatrix} 1 & -1 & -1 & \cdots & 1 \\ 1 & -1 & 1 & \cdots & 1 \\ 1 & -1 & 1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \end{array}$$

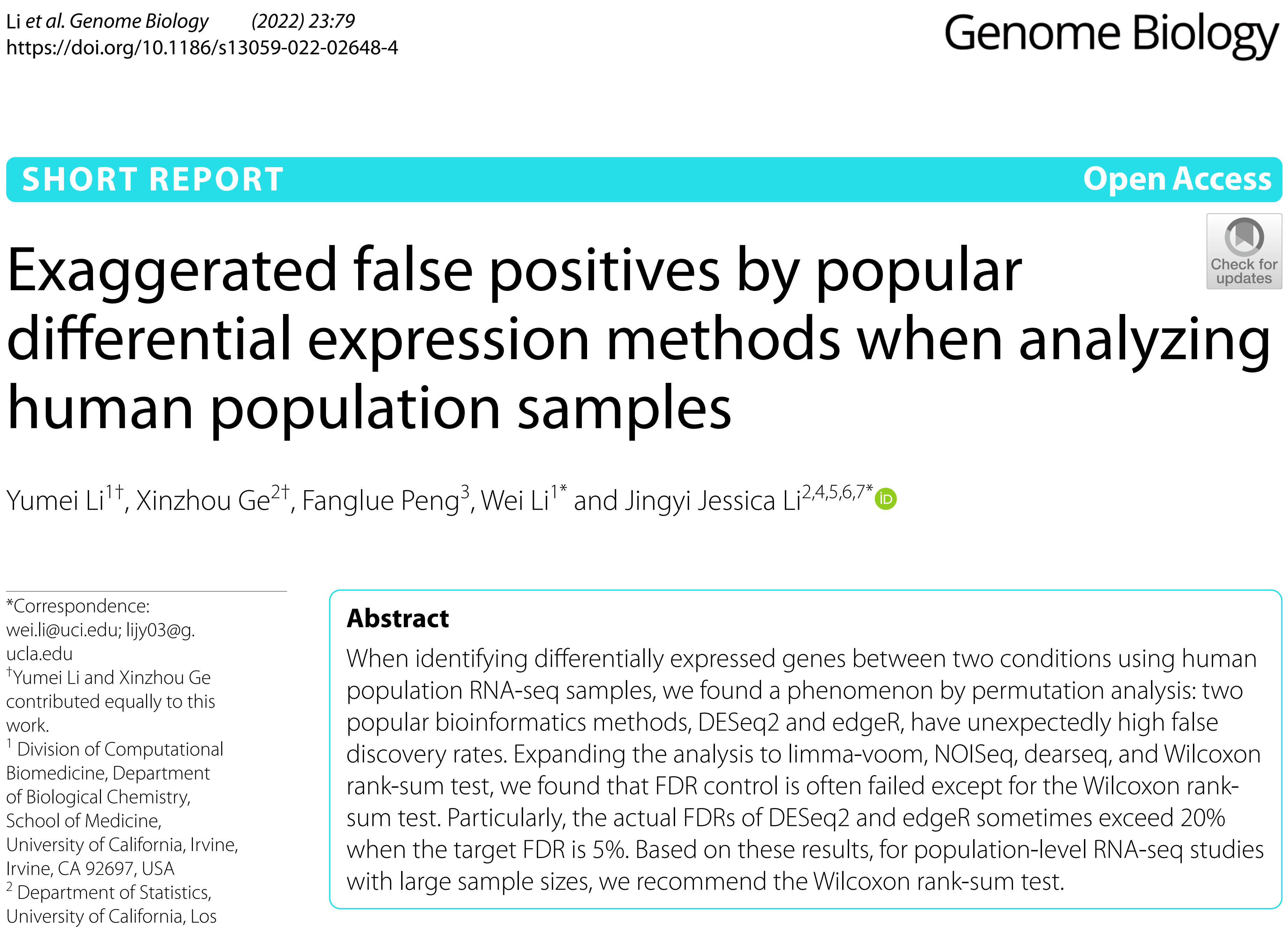
$$p = 12k - 8$$

$$y = \sum_{i=1}^{p} a_i X_i$$

SELMA cleavage bias
= ŷ (fitted value of the regression)



Hu *et al.*, *under review. bioRxiv* 2021



Li *et al. Genome Biology* (2022) 23:79
https://doi.org/10.1186/s13059-022-02648-4

Genome Biology

SHORT REPORT

Open Access

Exaggerated false positives by popular differential expression methods when analyzing human population samples

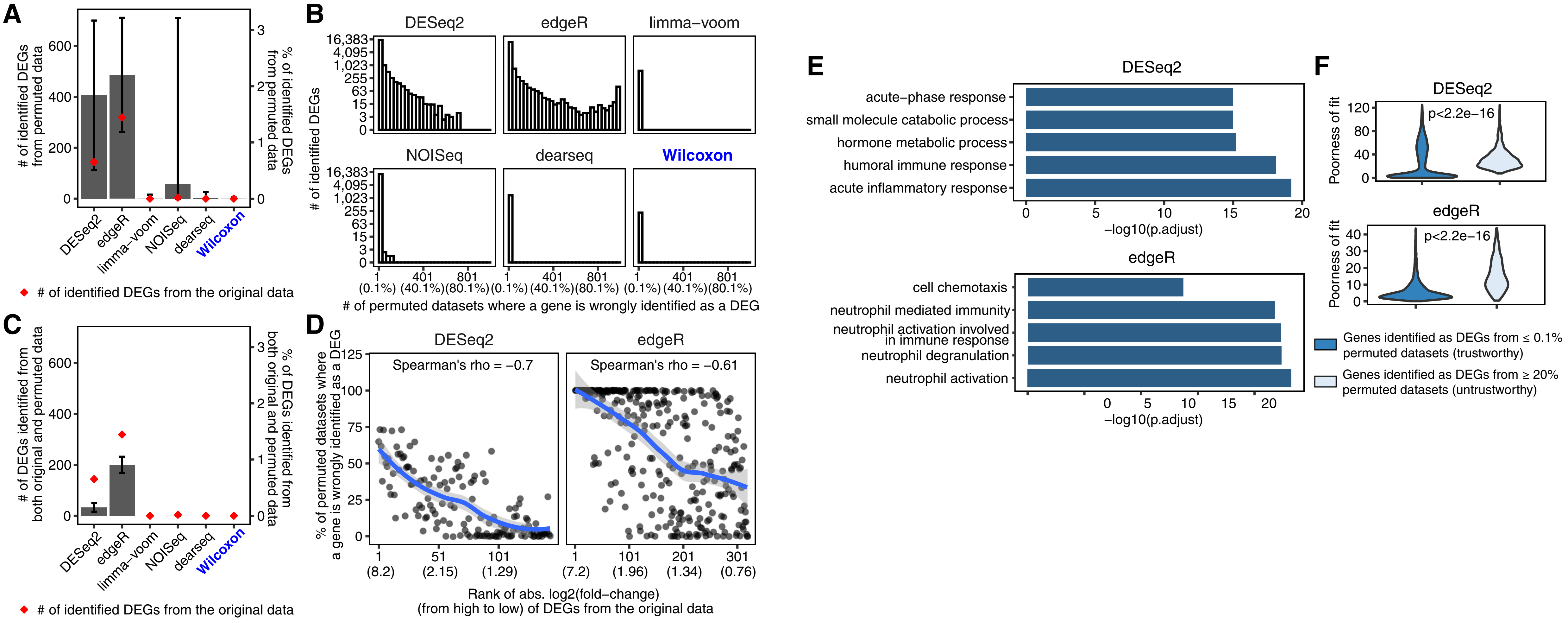
Yumei Li[1†], Xinzhou Ge[2†], Fanglue Peng[3], Wei Li[1*] and Jingyi Jessica Li[2,4,5,6,7*]

*Correspondence:
wei.li@uci.edu; lijy03@g.ucla.edu
†Yumei Li and Xinzhou Ge contributed equally to this work.
[1] Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA
[2] Department of Statistics, University of California, Los

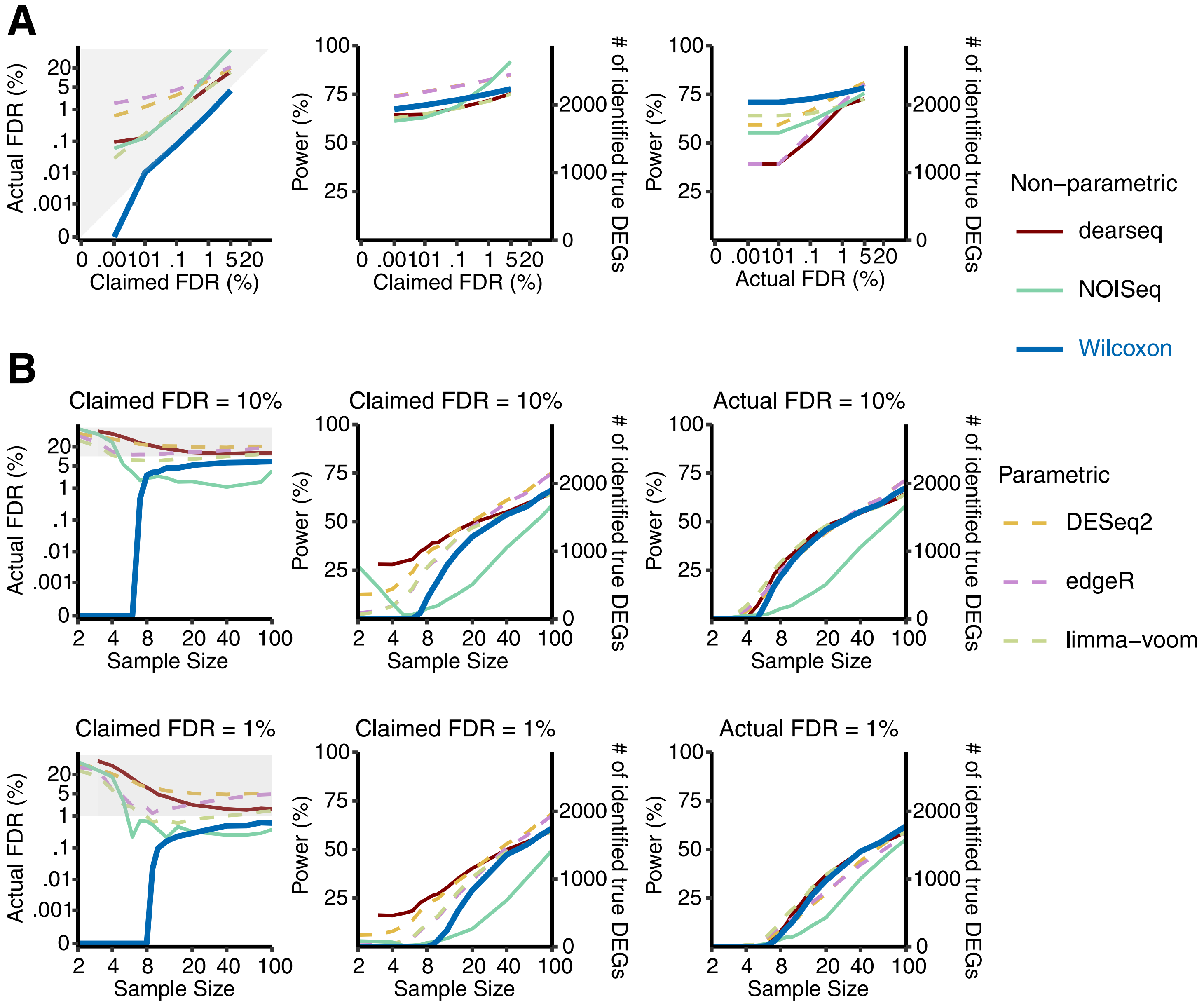
Abstract

When identifying differentially expressed genes between two conditions using human population RNA-seq samples, we found a phenomenon by permutation analysis: two popular bioinformatics methods, DESeq2 and edgeR, have unexpectedly high false discovery rates. Expanding the analysis to limma-voom, NOISeq, dearseq, and Wilcoxon rank-sum test, we found that FDR control is often failed except for the Wilcoxon rank-sum test. Particularly, the actual FDRs of DESeq2 and edgeR sometimes exceed 20% when the target FDR is 5%. Based on these results, for population-level RNA-seq studies with large sample sizes, we recommend the Wilcoxon rank-sum test.

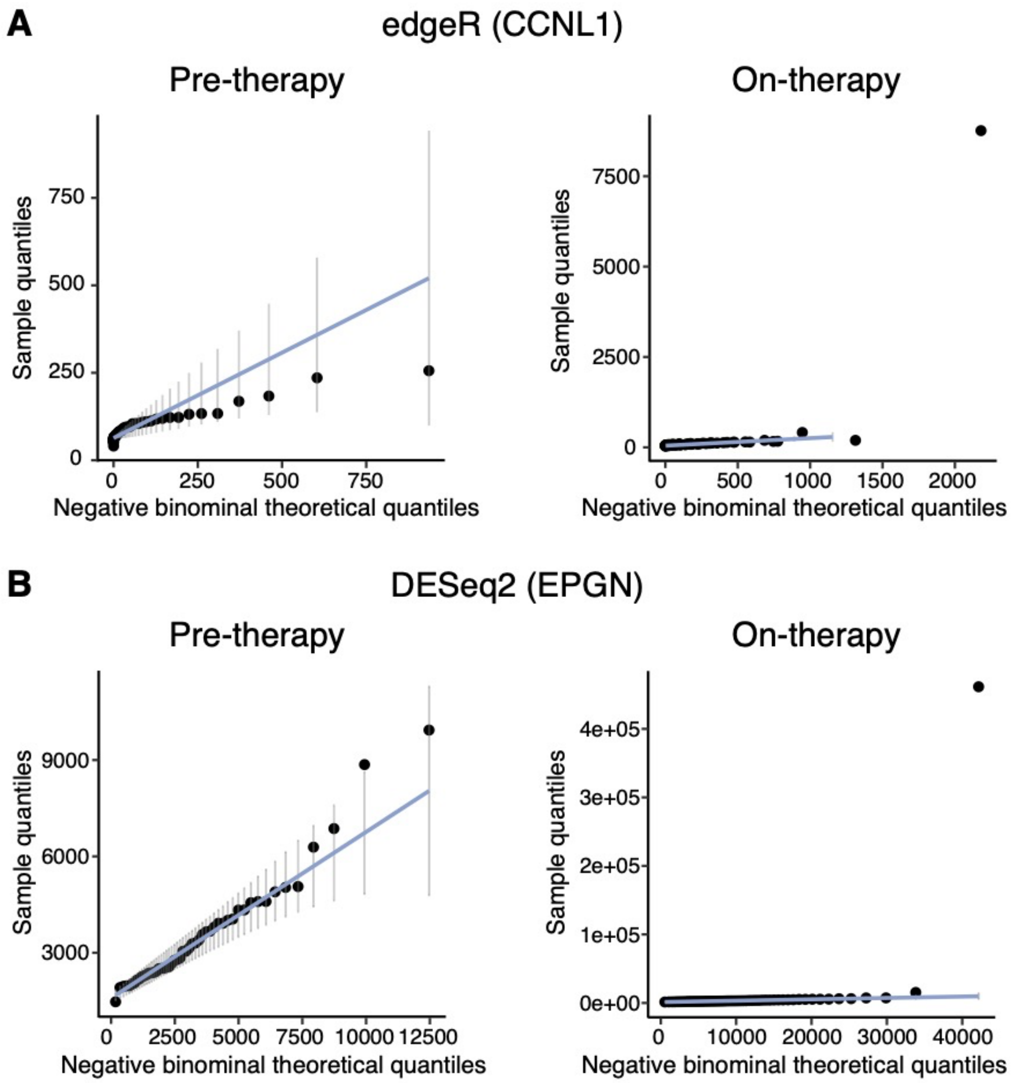
Exaggerated false DEGs can be identified by DESeq2 and edgeR from human samples



Wilcoxon rank-sum test is better when sample size > 8



Gene expression can deviate from NB distribution



Summary

- Spatial transcriptomics techniques
- Encoding strategies
- Differential gene expression