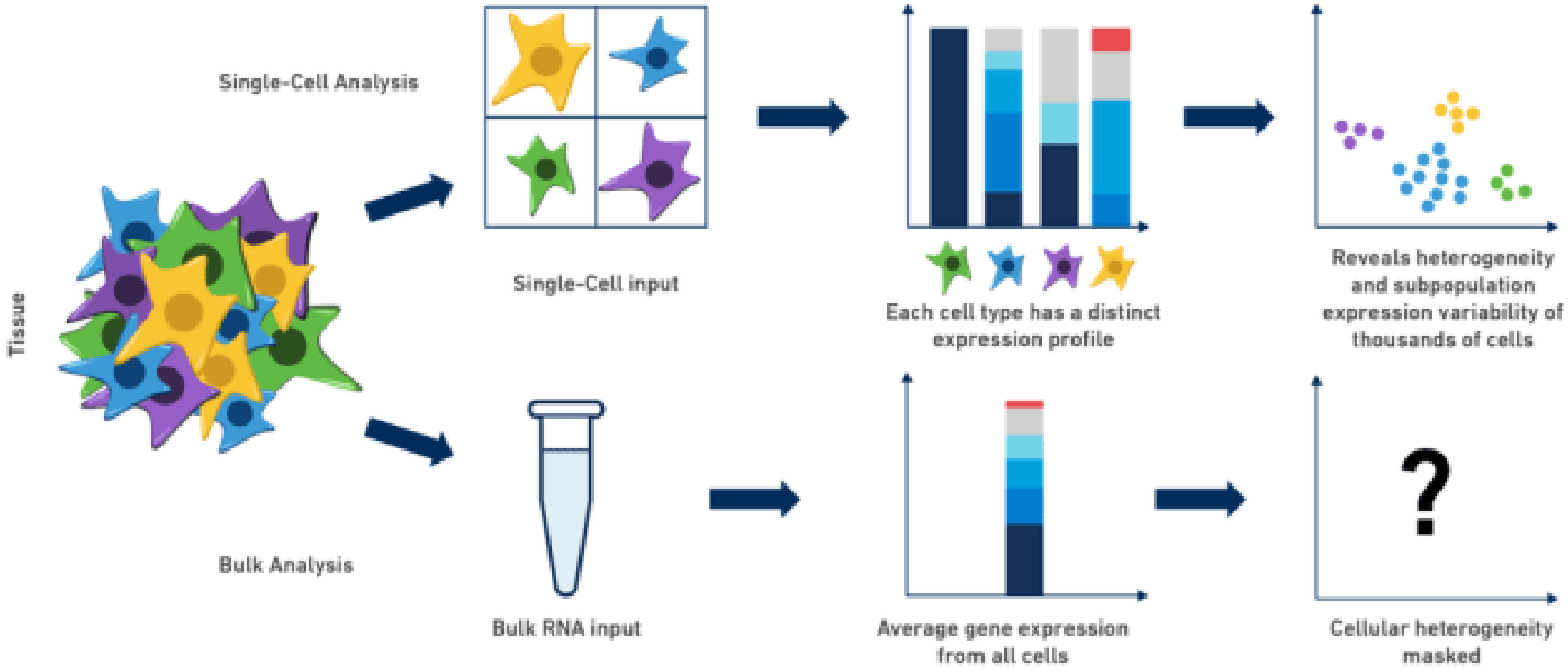
Single-cell RNA-sequencing

- Why single-cell ?

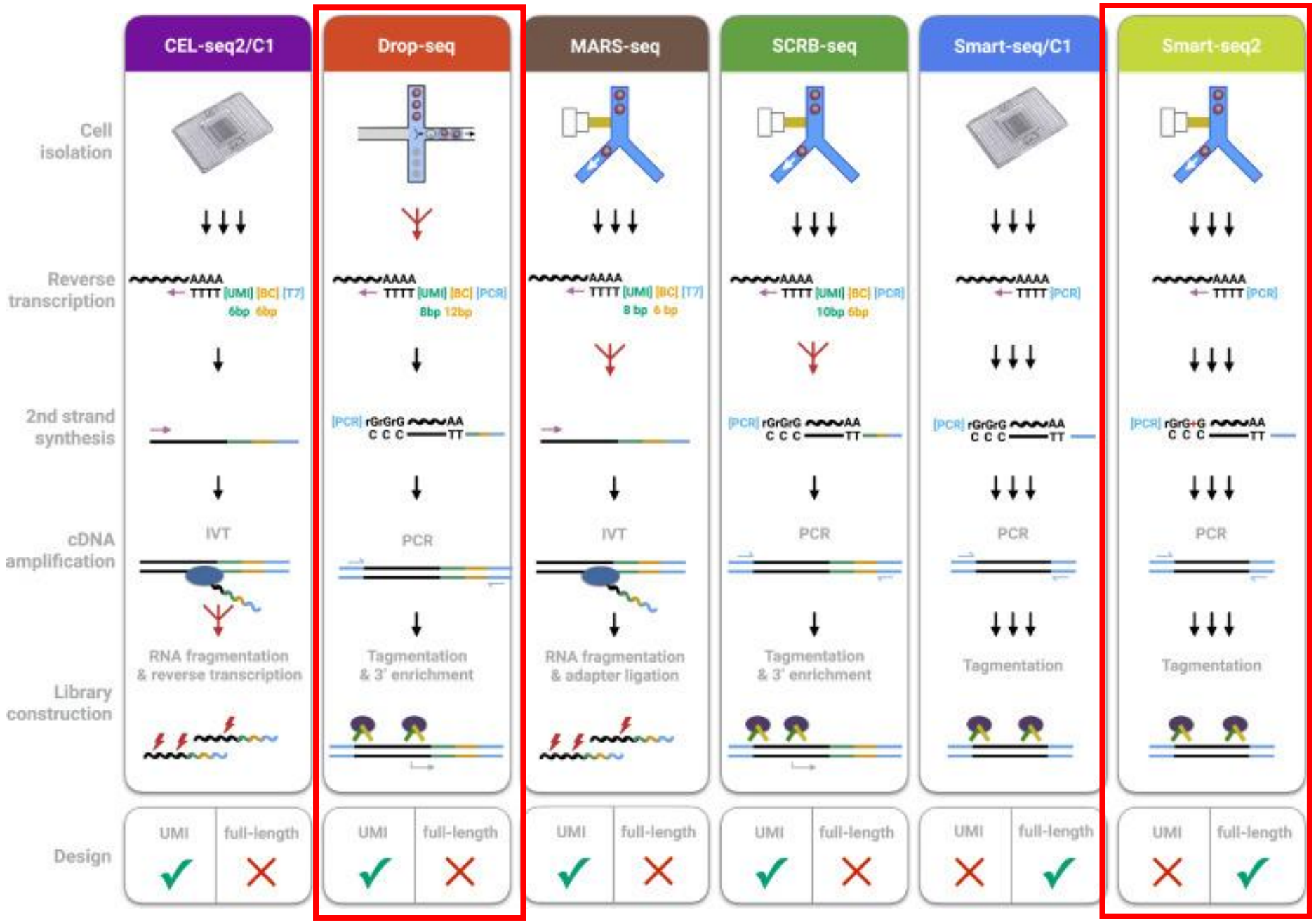
Bulk

Single Cell

- Why single-cell ?



- scRNA-seq technology

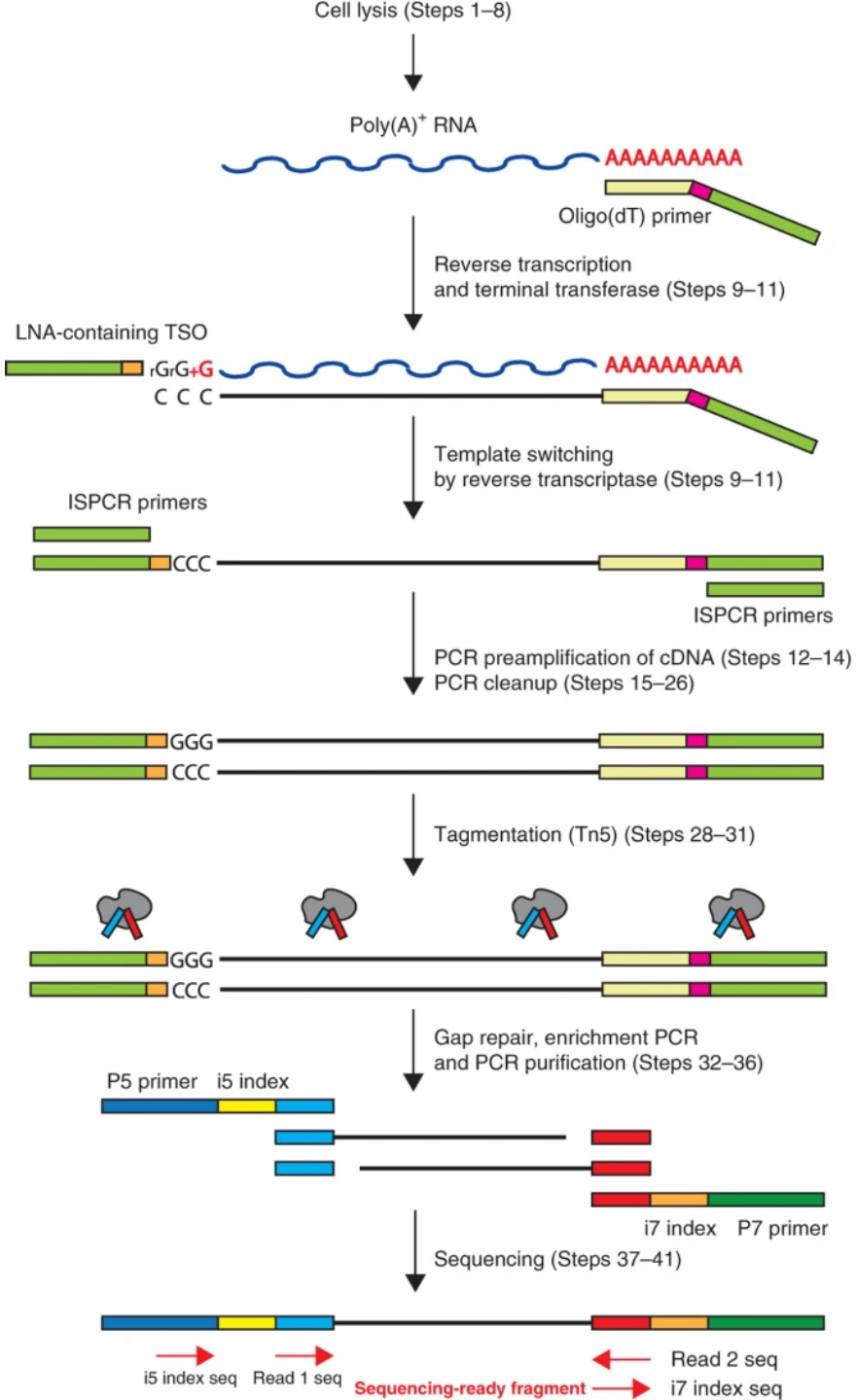


Comparative Analysis of Single-Cell RNA Sequencing Methods

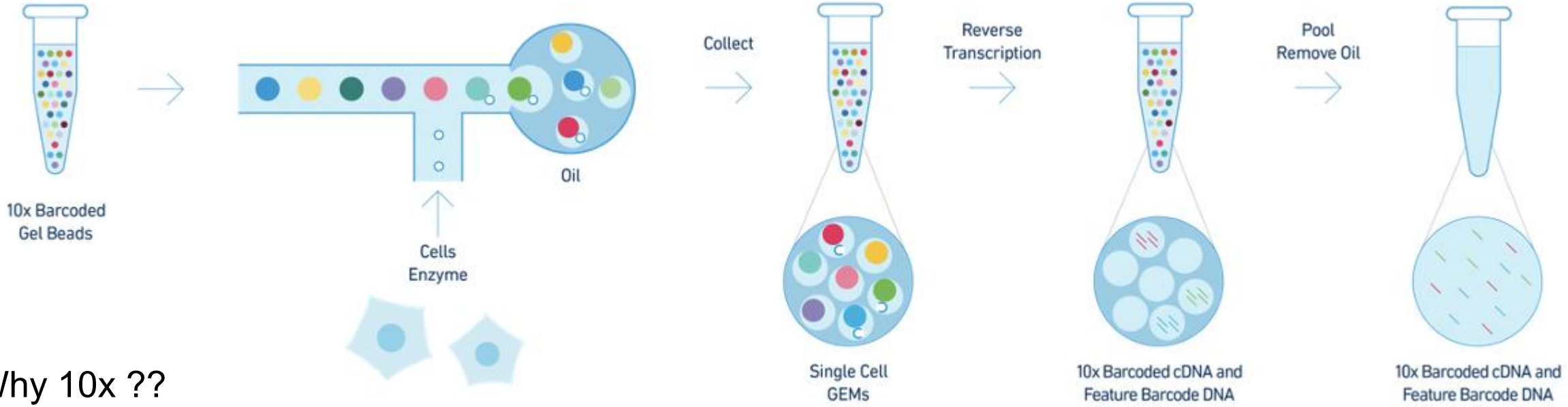
- Smart-seq2

-FACS sorting → 96-well PCR plates
-Cell Lysis
-Reverse transcription → cDNA …
-Tn5 tagmentation (chopping and tagging)
-Full-length sequencing
-Each Fastq file corresponds to each cell

1	SRR11548680	SAMN14600765	150	185.21 M	66.39 Mb	SRX8118737	GSM4477083	NextSeq 500
2	SRR11548681	SAMN14600764	150	254.55 M	91.26 Mb	SRX8118738	GSM4477084	NextSeq 500
3	SRR11548682	SAMN14600763	150	222.46 M	80.23 Mb	SRX8118739	GSM4477085	NextSeq 500
4	SRR11548683	SAMN14600762	150	224.80 M	81.26 Mb	SRX8118740	GSM4477086	NextSeq 500
5	SRR11548684	SAMN14600761	150	197.79 M	71.15 Mb	SRX8118879	GSM4477087	NextSeq 500
6	SRR11548685	SAMN14600760	150	217.61 M	77.94 Mb	SRX8118880	GSM4477088	NextSeq 500
7	SRR11548686	SAMN14600759	150	285.22 M	101.53 Mb	SRX8118881	GSM4477089	NextSeq 500
8	SRR11548687	SAMN14600758	150	263.85 M	96.43 Mb	SRX8118882	GSM4477090	NextSeq 500
9	SRR11548688	SAMN14600757	150	28.65 k	95.19 kb	SRX8118883	GSM4477091	NextSeq 500
10	SRR11548689	SAMN14600756	150	274.38 M	98.97 Mb	SRX8118884	GSM4477092	NextSeq 500
11	SRR11548690	SAMN14600755	150	214.02 M	77.81 Mb	SRX8118885	GSM4477093	NextSeq 500
12	SRR11548691	SAMN14600754	150	258.18 M	93.43 Mb	SRX8118886	GSM4477094	NextSeq 500
13	SRR11548692	SAMN14600753	150	193.40 M	69.72 Mb	SRX8118887	GSM4477095	NextSeq 500
14	SRR11548693	SAMN14600752	150	104.79 M	37.42 Mb	SRX8118888	GSM4477096	NextSeq 500
15	SRR11548694	SAMN14600751	150	152.84 M	55.17 Mb	SRX8118889	GSM4477097	NextSeq 500



- 10x technology (drop-seq based)



Why 10x ??
→ High throughput (for cell point of view)
& cheaper per cell

-Drawback: since droplet size is fixed and the cell must go inside the droplet, there is a physical restriction for a cell to enter the droplet (Big cell cannot go inside)

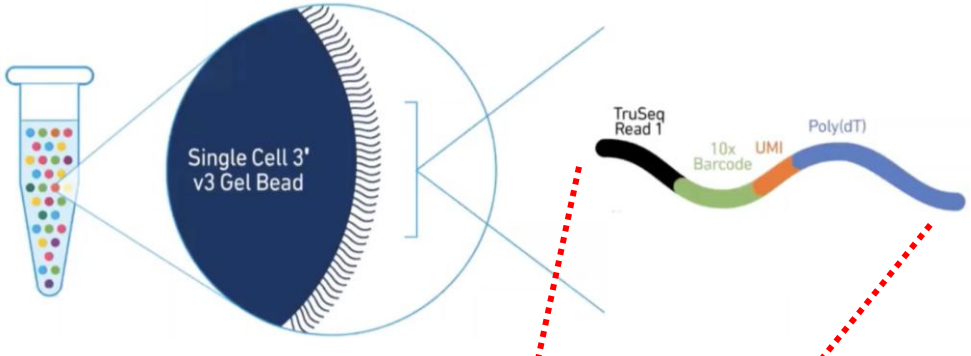
-Cell dissociation stress: hard to observe epithelial cells or stroma cells
Bias in liquid cell: Immune cells

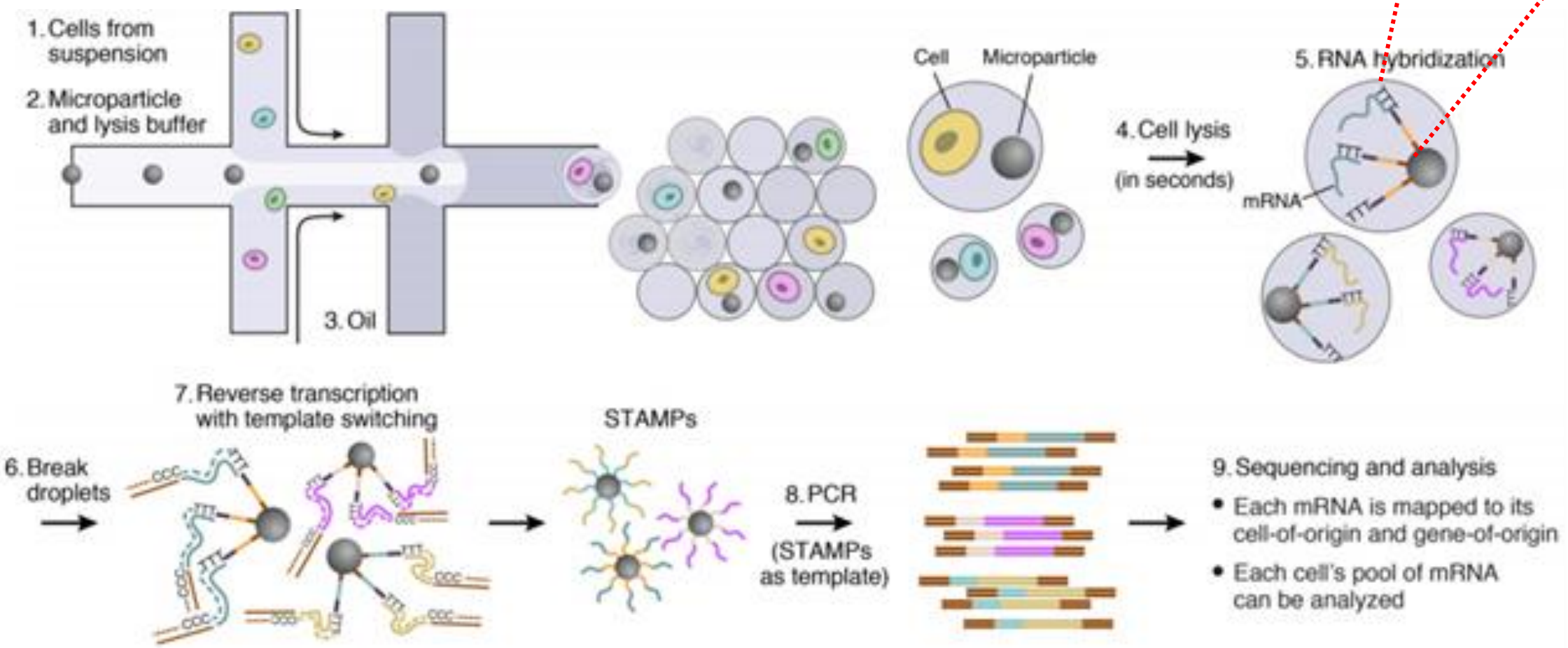
	Run	BioSample	Bases	Bytes	egfr_status
1	SRR11040644	SAMN14057011	21.96 G	14.67 Gb	wild type
2	SRR11040645	SAMN14057010	20.42 G	14.40 Gb	mutation
3	SRR11040646	SAMN14057009	21.22 G	13.77 Gb	mutation
4	SRR11040647	SAMN14057008	20.82 G	13.43 Gb	wild type
5	SRR11040648	SAMN14057007	24.77 G	16.94 Gb	mutation
6	SRR11040649	SAMN14057006	23.37 G	16.38 Gb	wild type
7	SRR11040650	SAMN14057005	24.45 G	17.20 Gb	mutation
8	SRR11040651	SAMN14057004	26.17 G	18.69 Gb	mutation
9	SRR11040652	SAMN14057003	22.82 G	15.15 Gb	wild type
10	SRR11040653	SAMN14057002	22.43 G	14.91 Gb	wild type

1 fastq file (result of single experiment)
→ ~3k ~ 10K cells!

- 10x technology

-10x barcode → distinguish each cell
-UMI: avoid PCR amplification bias → transcript counting
-Poly(dT): captures poly A tail from mRNA





- Raw data

-Of course, FASTQ file

Read 1

```
@D00547:1132:HMV3MBCXY:1:1106:2669:1969 1:N:0:CATTAGCG
NTCACACTCTTCGAGATCAACTGAGC
+
#<GGGGIIIIIIIGGGGGGIIIIIGI
@D00547:1132:HMV3MBCXY:1:1106:2763:1977 1:N:0:CATTAGCG
CCGGTAGTCCTTCAATACCTCACCCT
+
GGAAGGGGGGGIIIGGIIGGGGGGGG
@D00547:1132:HMV3MBCXY:1:1106:3377:1991 1:N:0:CATTAGCG
TGAAAGAGTATGCTTGGTCCGCGAGG
+
GGGGGGIGIGIIGIIIIIGIIIGGII
@D00547:1132:HMV3MBCXY:1:1106:4290:1957 1:N:0:CATTAGCG
NGGCTAGAGGATTCGGAGCGCAACGG
+
#<<GAGGIIIIIIGIGGIGGIIIGII
@D00547:1132:HMV3MBCXY:1:1106:4842:1958 1:N:0:CATTAGCG
NTTCGGGTCCCAAGATGGCTTACTAG
+
```

Read 2

```
@D00547:1132:HMV3MBCXY:1:1106:2669:1969 2:N:0:CATTAGCG
GCTAGACTGCTATGCACACAACGCCCACGCCCACGTTACCATTTTAAGATACTGTCAATGCTCAGTTAAAAATAAGACTTACTTAGGAGGAAAAAAAA
+
AGAGGIIIIIGGGGGIGGGGGIGIGIGGAGGGIIIGIIIGGIIIGIGIGGIIIIIIIIGIIIIIIIIIIIIIGGGIGIIIIIIIIIIIGGIIIIGIII
@D00547:1132:HMV3MBCXY:1:1106:2763:1977 2:N:0:CATTAGCG
CACTACCAGAAAAACACCTTGTGGTGAAGGTTCCAAGACCTGGGATCGATTCCAGATGAGGATCCACAAGCGACTCATTGATTTACATAGTCCTTCTG
+
AAGGGGIGIIIGGGGGAG..GGIIGGIIIGGGAGGGGGGG.AAGGGGGIIGIGGGIIGGAGAGGGGIGGGGGGIIIGGGGGAGGGGGGGGGGGGA<AG
@D00547:1132:HMV3MBCXY:1:1106:3377:1991 2:N:0:CATTAGCG
ACCAGCCCGCCCTGGGACCTCCACCTGAATGAACCTCTCAAGCCACTAGGCAGCTTTGTAACCGCCCTAGAGCCTCTGTCAAGTCTTGGACCAAGTAA
+
GGGA..<AA<<A....<<A<<.GA<..<GA.GG......<G.<.G.<G...G.<.....<G.....<...<<...<..<<.A.......<.<AG<<GG
@D00547:1132:HMV3MBCXY:1:1106:4290:1957 2:N:0:CATTAGCG
AACTGAGTTGTCCTACATACAAGTACATGTATTTAATGTTGTAAGAATTATGTACTGTTCCTATAAGTTTGCTATTAAAATACAAAAAACTATAAAAA
+
GGA..G<<.<<<GAGGGGGIG.<<GG.......<G.....<.<.....<G.......<..<AGG.GAG..<GG..<.<GGGI
@D00547:1132:HMV3MBCXY:1:1106:4842:1958 2:N:0:CATTAGCG
CCCAATGTTGTACGGCTGATGGATGTCTGTGCTACTTCCCGAACTGATCGGGACATCAAGGTCAACCTAGTCTTTGAGCACATAGACCAAGACCTGAG
```

Cell Barcode

UMI

mRNA

-Read1 (or I file): Cell barcode &UMI sequence

Variety of UMI → gene expression

-Read2: mRNA sequencing information → STAR → which gene?

But limited length; biased to 3’ region

Technically paired-end sequencing → Biologically single-end sequencing

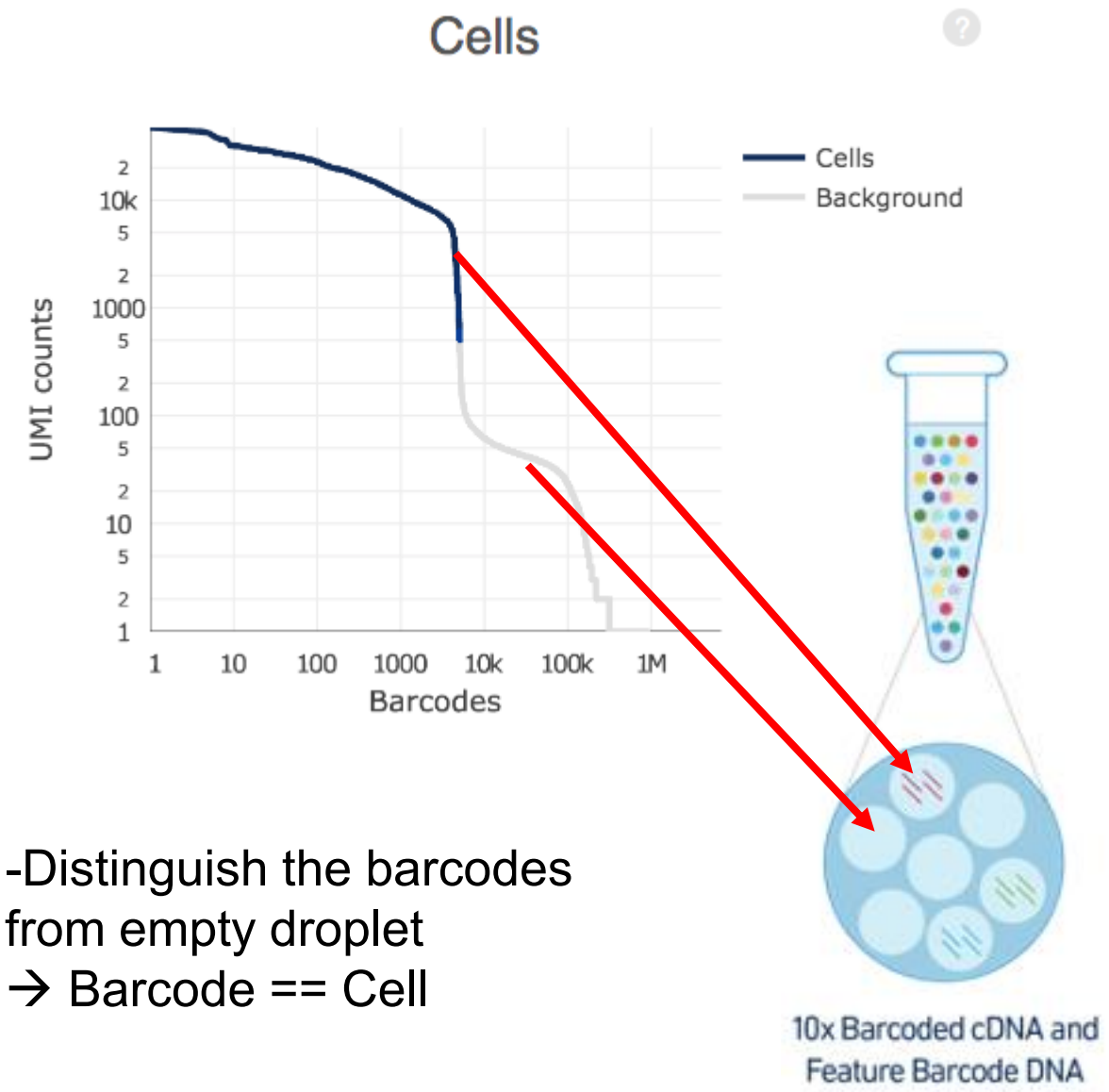
- Raw data

-Single cell → Transcript is scarce → low capture rate → high drop out (many zero count: 90 ~ 95 %)
+ short read (tend to be multi-mapped), single-end (low confidence
!! Hard to distinguish between real zero and drop out

	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0

	Cell1	Cell2	…	CellN
Gene1	3			
Gene2		3	.	
Gene3		14		18
…	.		.	.
…				
…	.	.	.	
GeneM	25	0		0

- Initial quality control



-Distinguish the barcodes from empty droplet
→ Barcode == Cell

Summary | Gene Expression | Antibody

Failed

125
Estimated Number of Cells

3,200
Mean Reads per Cell

13
Median Genes per Cell

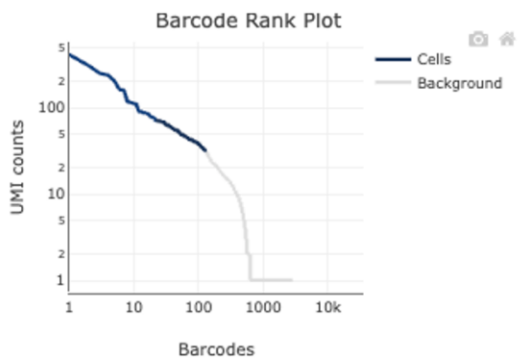
Sequencing

Number of Reads	400,000
Number of Short Reads Skipped	0
Valid Barcodes	94.0%
Valid UMIs	99.9%
Sequencing Saturation	75.0%
Q30 Bases in Barcode	96.4%
Q30 Bases in RNA Read	95.7%
Q30 Bases in UMI	96.3%

Mapping

Reads Mapped to Genome	100.0%
Reads Mapped Confidently to Genome	21.4%
Reads Mapped Confidently to Intergenic Regions	2.6%
Reads Mapped Confidently to Intronic Regions	12.5%
Reads Mapped Confidently to Exonic Regions	6.3%
Reads Mapped Confidently to Transcriptome	16.3%
Reads Mapped Antisense to Gene	2.0%

Cells

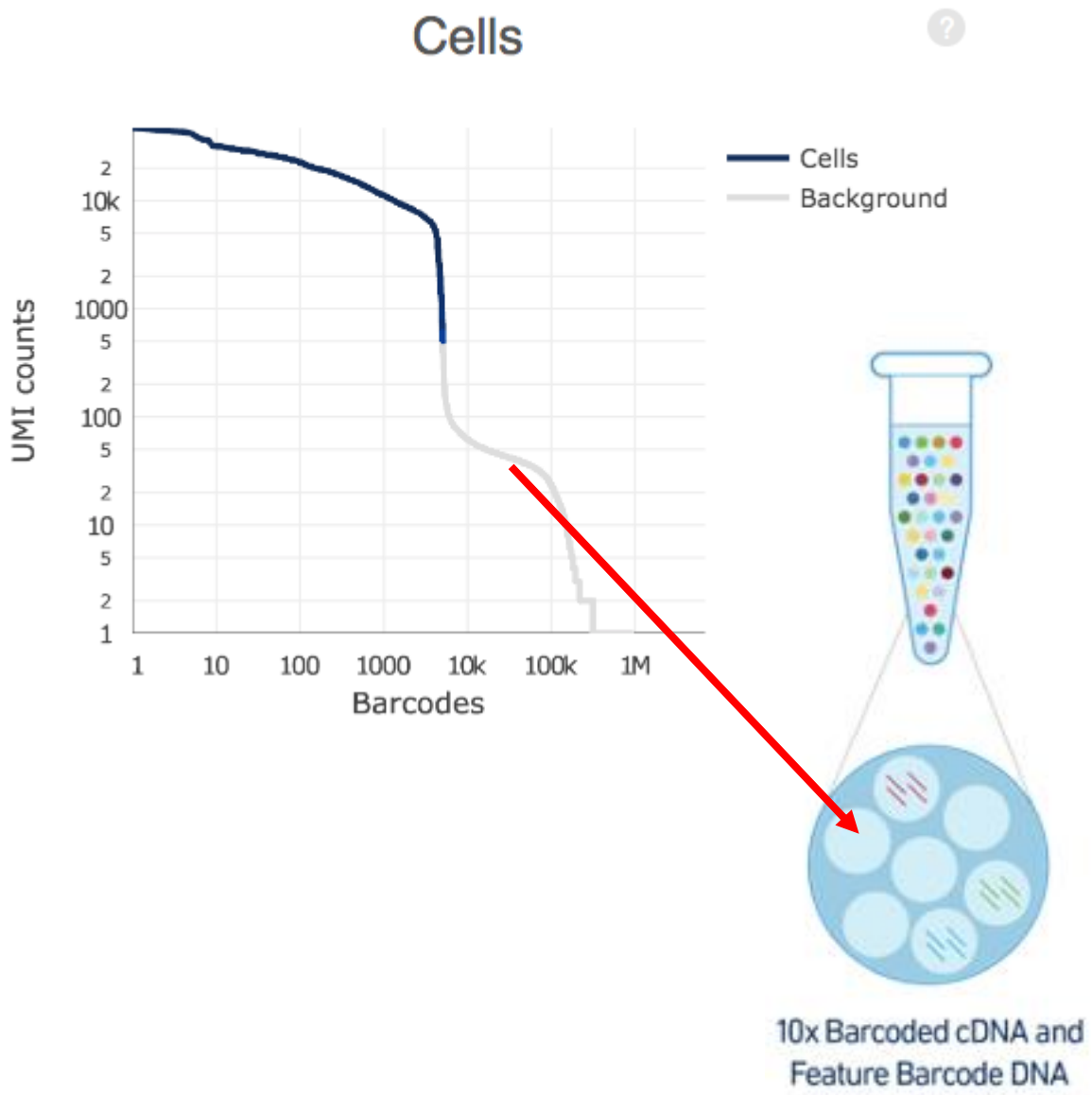


Estimated Number of Cells	125
Fraction Reads in Cells	49.1%
Mean Reads per Cell	3,200
Median UMI Counts per Cell	46
Median Genes per Cell	13
Total Genes Detected	78

Sample

Sample ID	78388_chr21_400K
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	…refdata-cellranger-chr21-3.0.0
Transcriptome	GRCh38_chr21-3.0.0
Pipeline Version	7.0.0

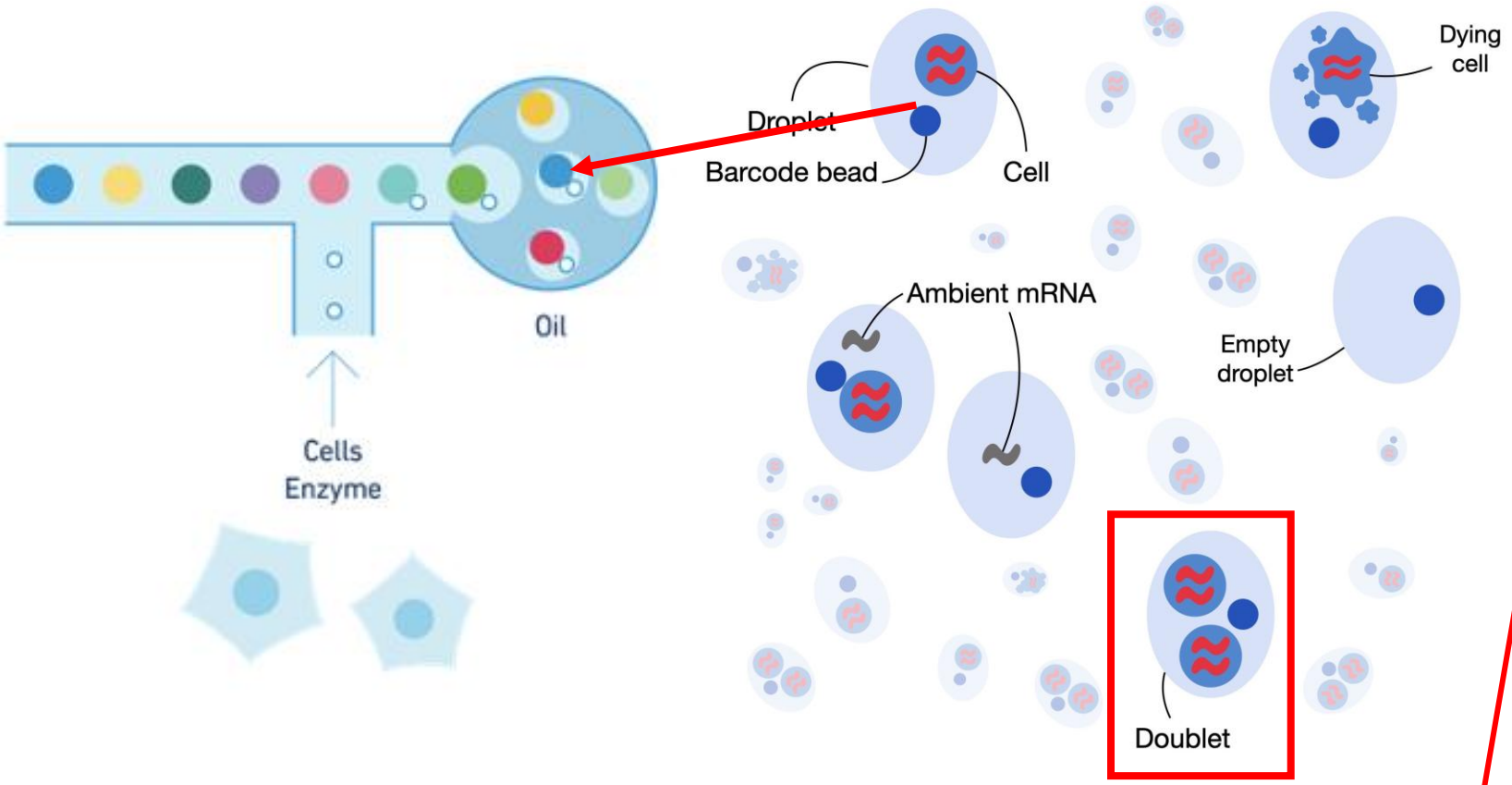
- Ambient RNA



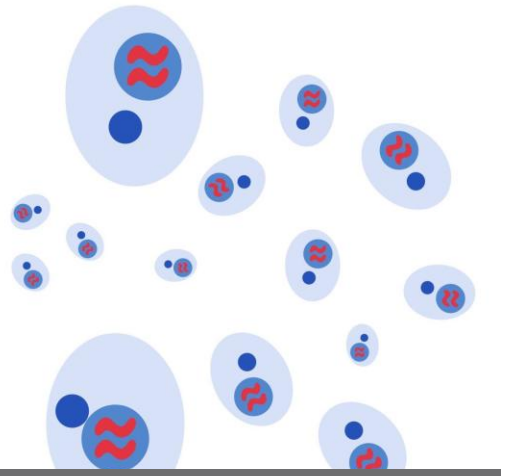
-There is still some RNA detection from the empty droplet
-This ambient RNA might be universal for all droplets, even those that contain a cell
-Cellbender (at the sequence level),
SoupX (at the count level)

-Training data set: empty droplet
→ Adjust the ambient RNA distribution to real cells
→ Adjust the expression values for each cell

- Doublet (multiplet) detection



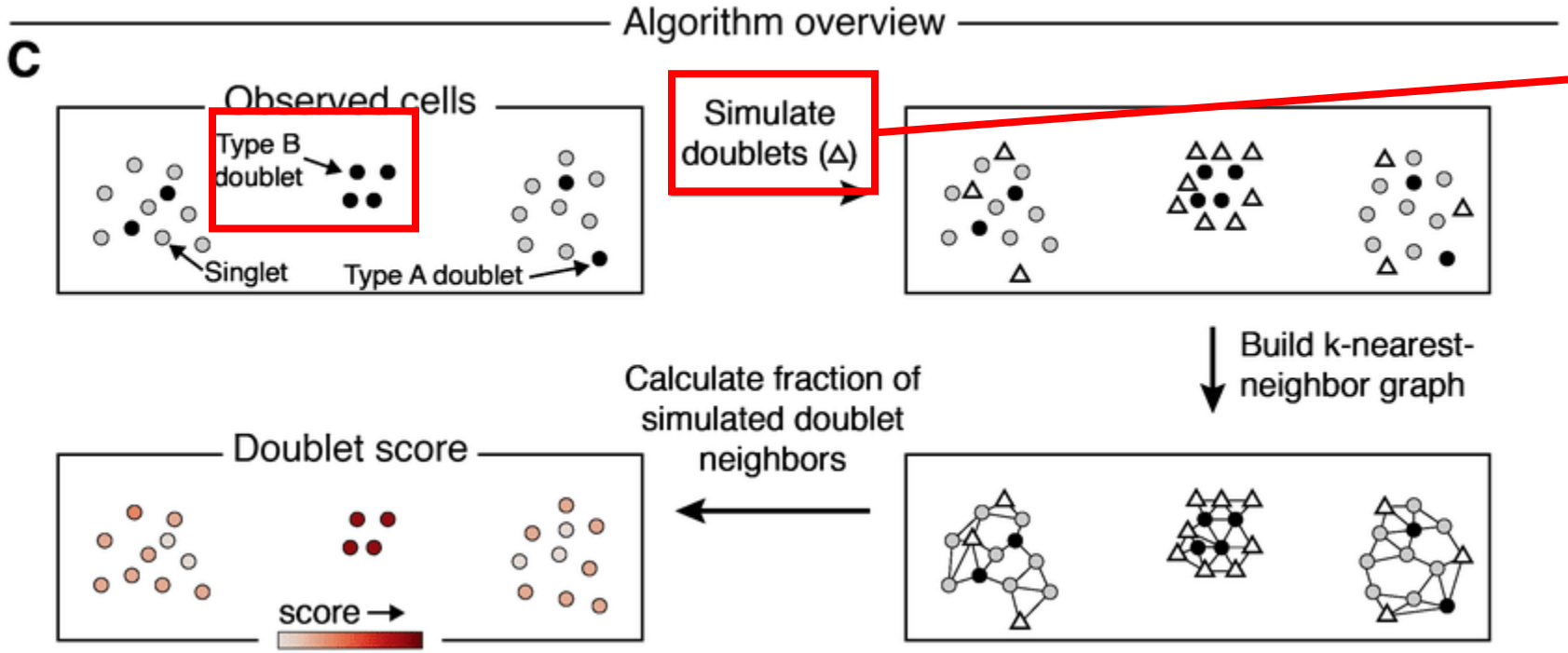
Quality Control



Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~825	~500
~0.8%	~1,650	~1,000
~1.6%	~3,300	~2,000
~2.4%	~4,950	~3,000
~3.2%	~6,600	~4,000
~4.0%	~8,250	~5,000
~4.8%	~9,900	~6,000
~5.6%	~11,550	~7,000
~6.4%	~13,200	~8,000
~7.2%	~14,850	~9,000
~8.0%	~16,500	~10,000

-10x can profile many cells,
but sometimes it is not perfect
→ Multiple cells in the droplet
→ Should be done before any bad cell remover:
doublet can be formed with bad cells, too

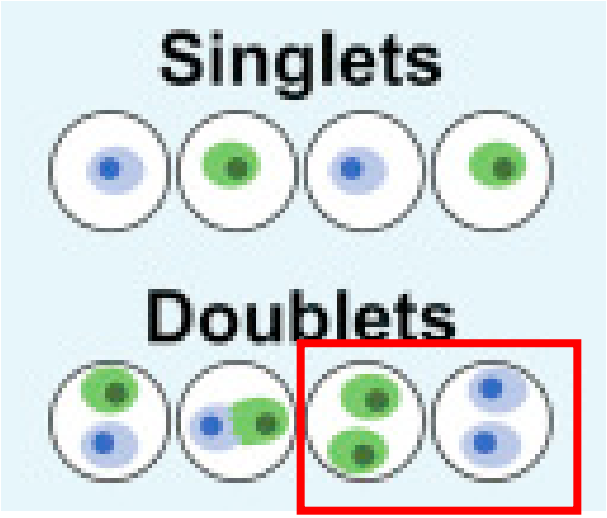
- Doublet (multiplet) detection



-make a synthetic doublets by merging two cells

-compare those synthetic doublets with Singlets and Doublets

DoubletFinder, Scrublet, scDblFinder, …



But it is still hard to distinguish the homotropic doublets

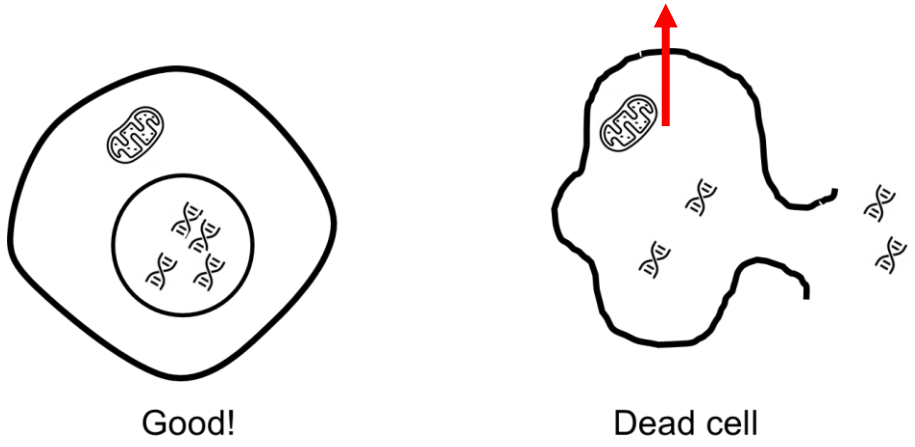
- # Quality control for bad cells

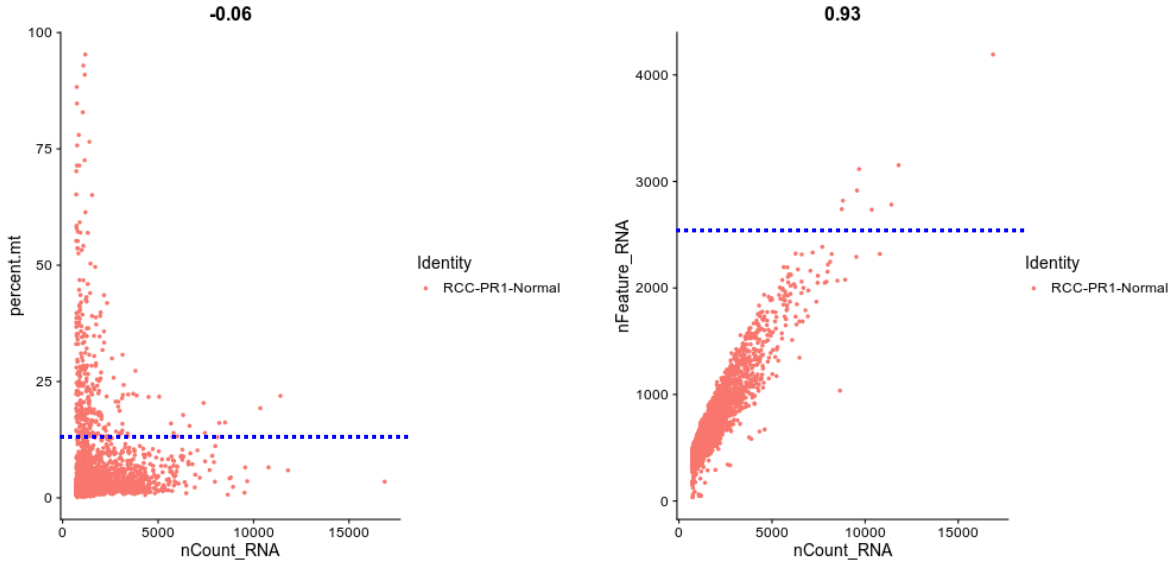
-Quality control

Why? There might be some bad cells collected (reason: cell stress during processing, high drop-out)

Low nCount or nFeature → empty droplet
nFeature > 150 ~ 200 (There should be certain number of gene detected: Housekeeping gene)

High nCount or nFeature → doublet
High MT % → dead cell





- # Quality control for bad cells

-Quality control

Oops!
1) Neutrophils (or other granulocytes): relatively low RNA content and relatively high levels of RNases and other inhibitory compounds, resulting in fewer transcripts detected
→ They need to secrete cytolytic enzymes, no time for other gene expression

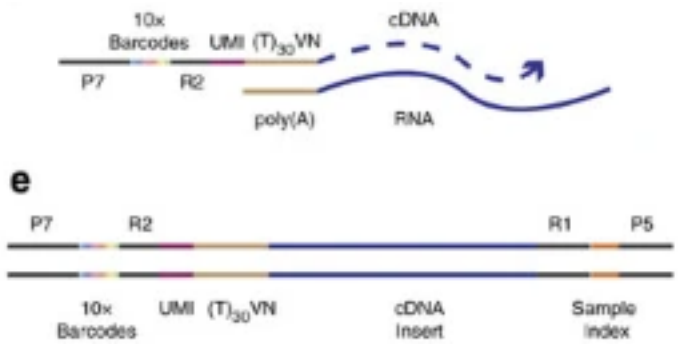
2) Plasma cell: relatively low RNA content
→ They need to secrete antibodies!

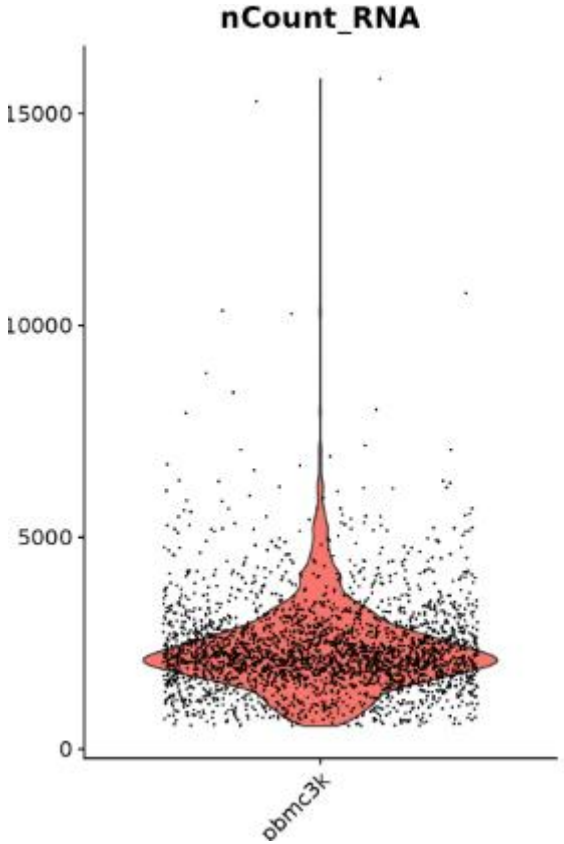
3) Red blood cell: low RNA content and nFeatures
→ It has no nucleus; no transcription (But we usually don’t analyze it)

• Normalization and Scaling

-Total read count normalization: Adjust read-depth between each cell
Log-transformation: adjust the variance of the gene expression matrix
Scaling: gives equal weight in downstream analyses, so that highly-expressed genes do not dominate

!! No gene length normalization
→ Only captures a short region of 3' end → no bias for gene length since every gene has only one 3' end



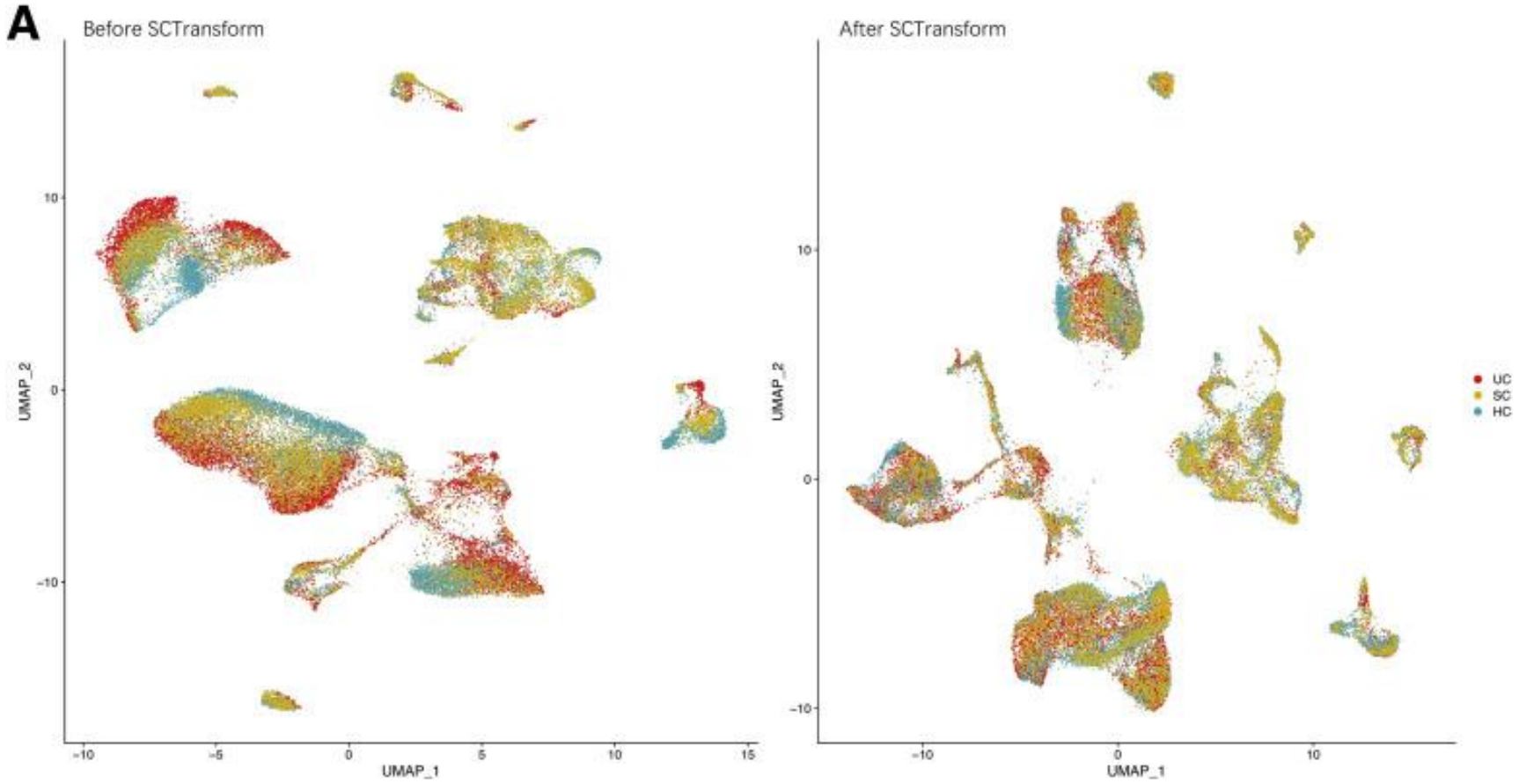


Broad range of read depth for each cell
→ Different distribution of gene expression for each cell

- sctransform normalization

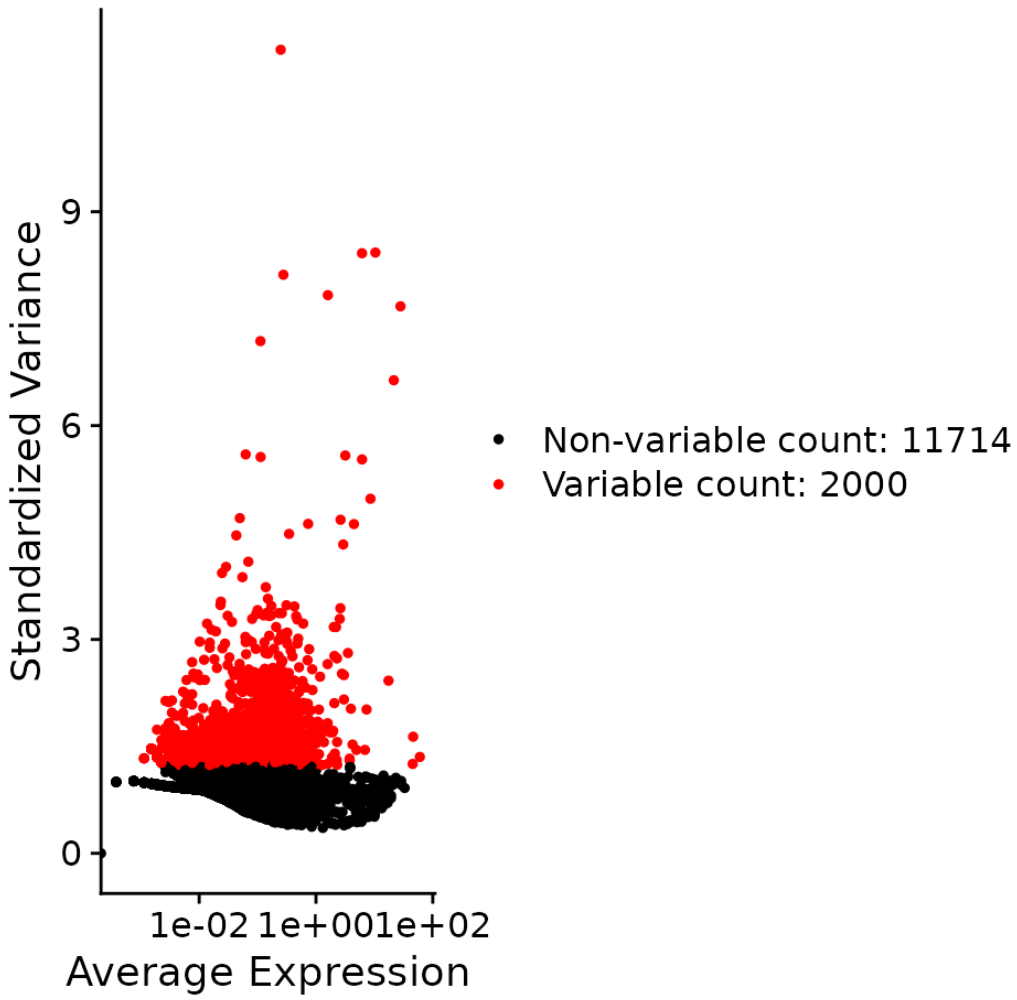
- Each cell is heterogeneous
→ Confounded by technical factors (sequencing depth)
→ Cell cycling phase

- sctransform regresses out those confounding effects (using NB-GLM)



• Feature selection

- scRNA-seq tends to have a lot of drop-out
→ Not all the genes are informative → select informative genes and reduce noise
→ Highly variable genes (across different cells)



-In bioinformatics (or data analysis point of view), Variance can reflect the amount of information

-It is likely that highly expressed genes (high mean value) have high variance
→ therefore, we must adjust variance by mean value

-Common red flag hkg genes
→ HLA, TCR/BCR (individual diversity), cell cycling (we don't want cell cycling phase affect the cell annotation)

!! We always need to consider individual (batch) specific genes (to remove)

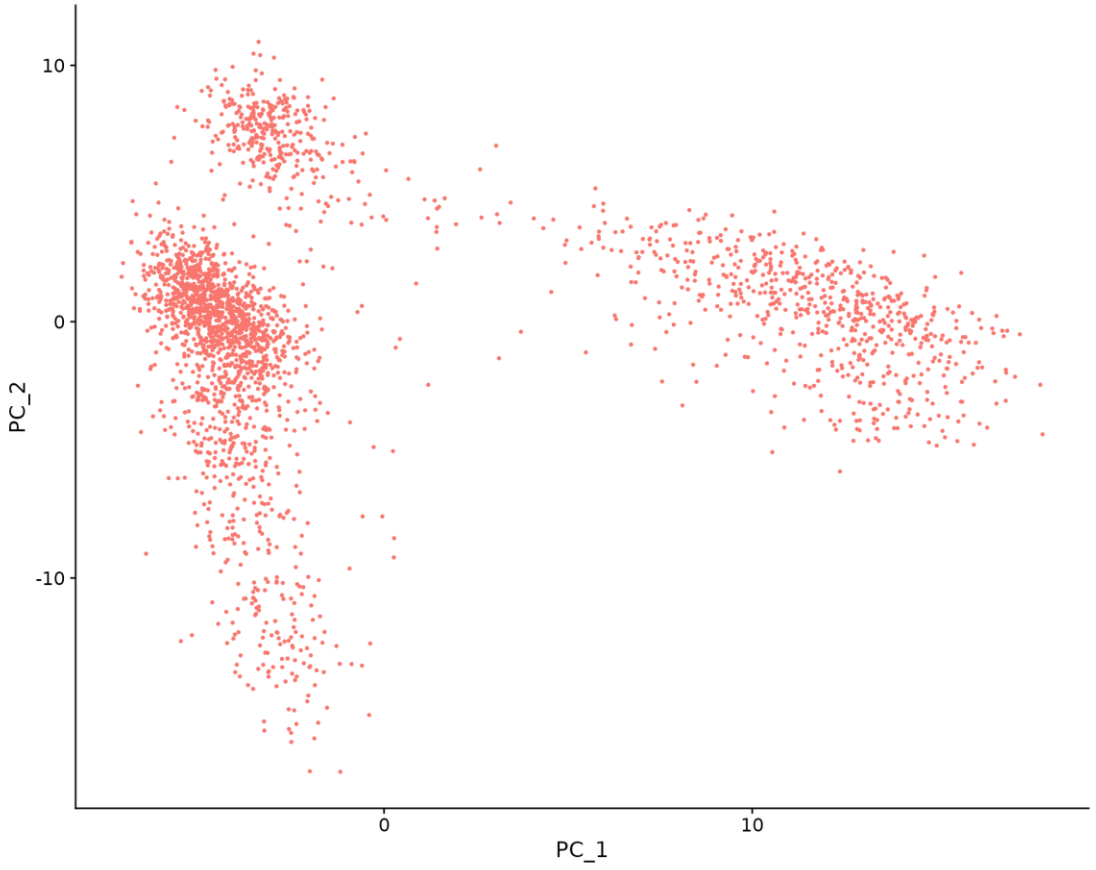
• Dimension reduction (PCA & UMAP)

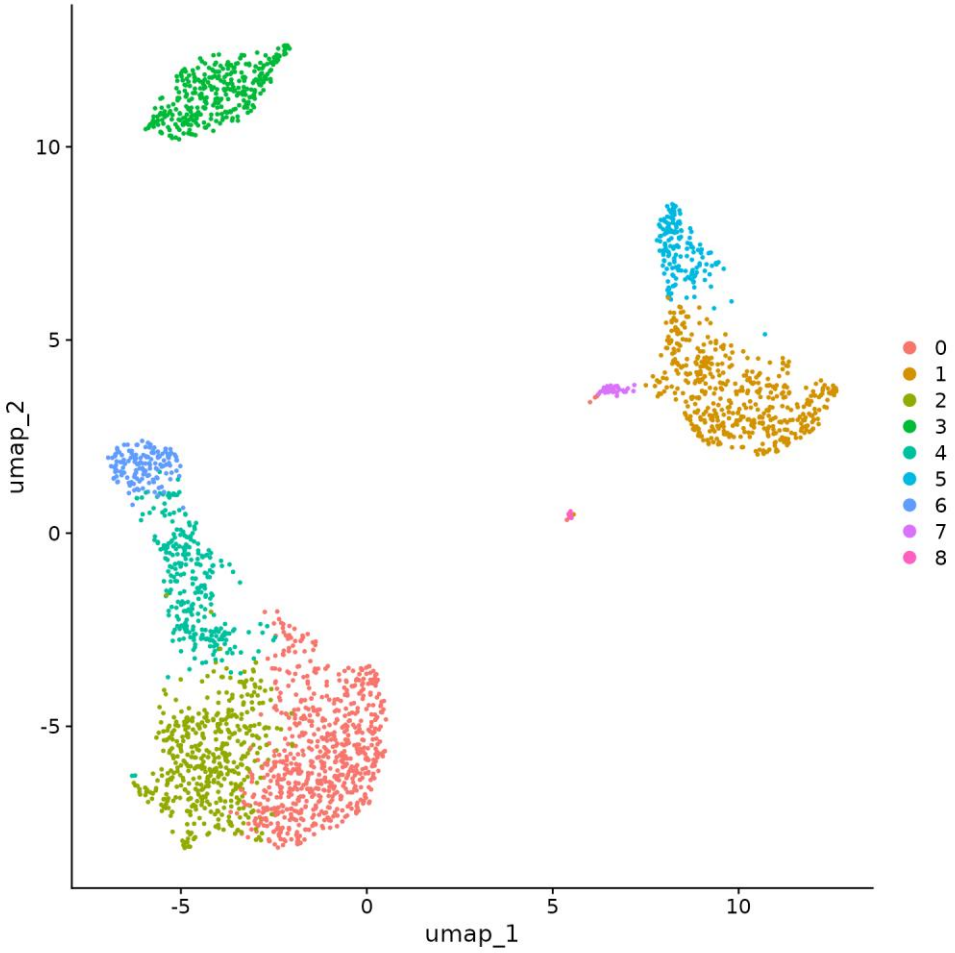
-Too many features (genes) → hard to interpret

→ Dimension reduction: abstract of many features!

PCA: commonly used in bulk RNA-seq data → insufficient for scRNA-seq

UMAP: adjusted for scRNA-seq (similar cells to be close and different cells to be far away)





- # Batch correction

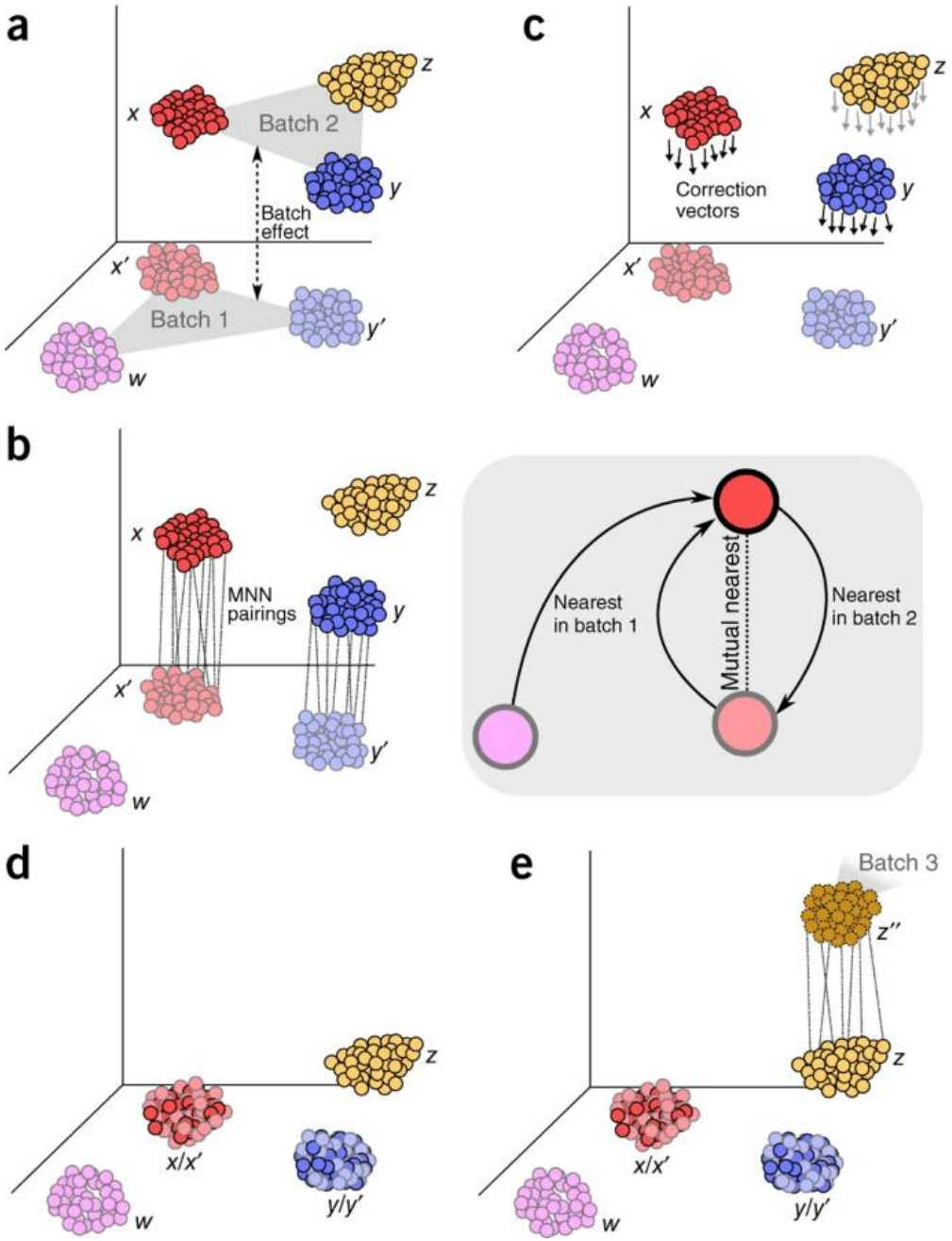
-Reason → To remove technical variation or confounding effect between samples

-Assessment
Batch: good mixing
Cell type (or cluster): separated

-Entropy: $S = -\sum_{i=1}^{n} p_i \ln p_i = \ln n$ (High: mixing)

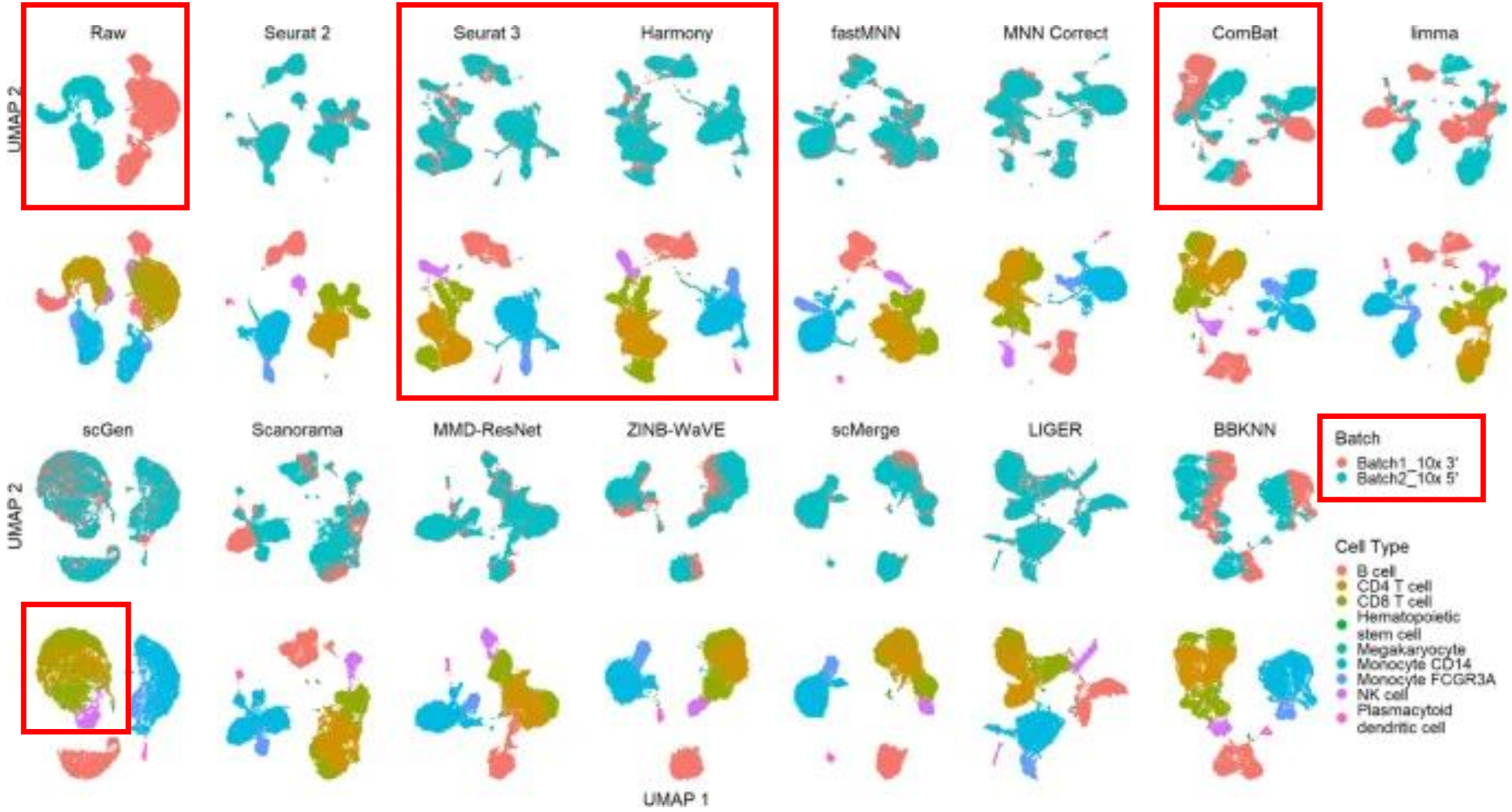
-Silhouette Coefficient: $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$

b: distance to closest neighbors
a: distance to self cluster
s → high: far from neighbors → separated



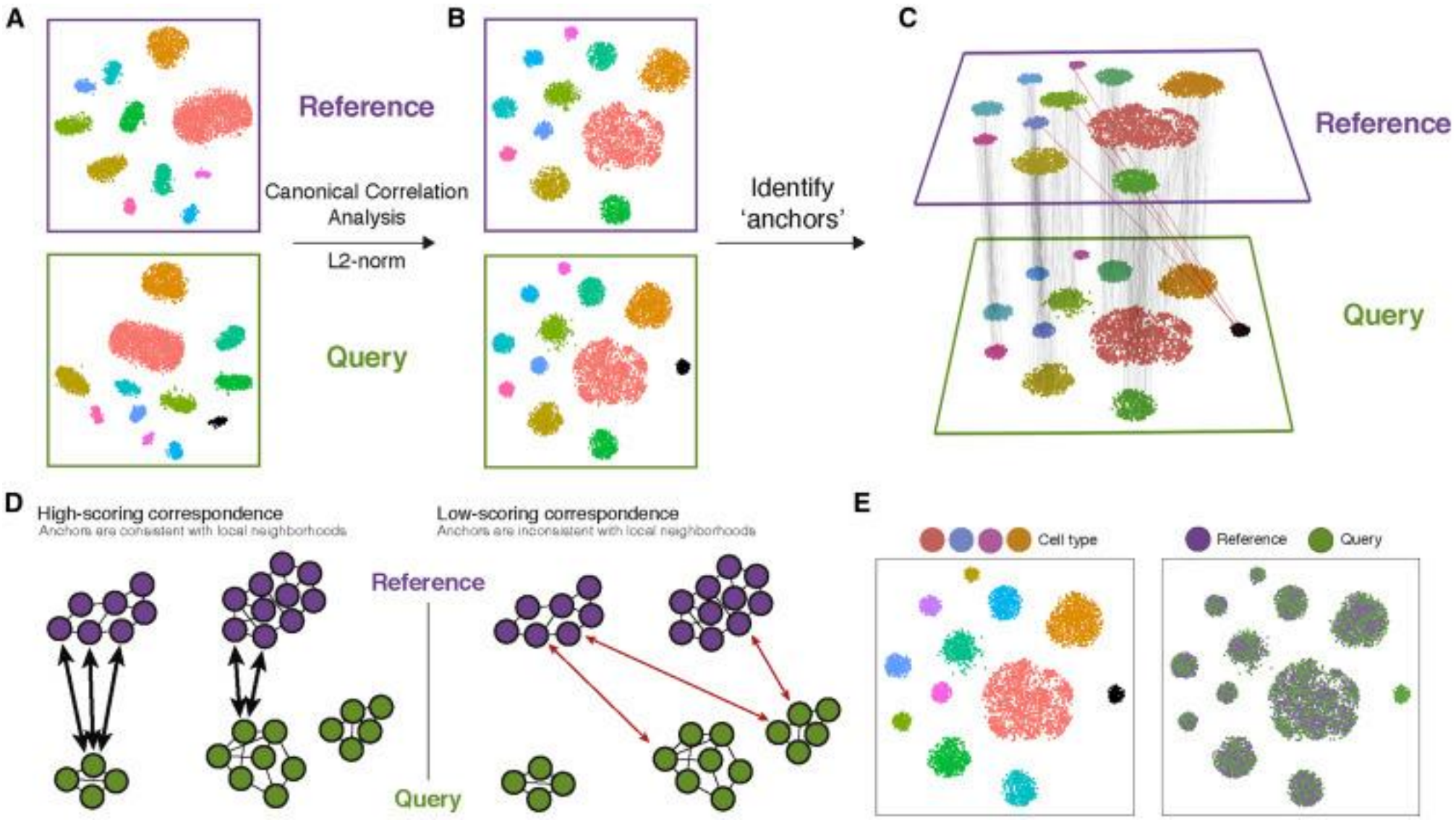
Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors

- # Batch correction

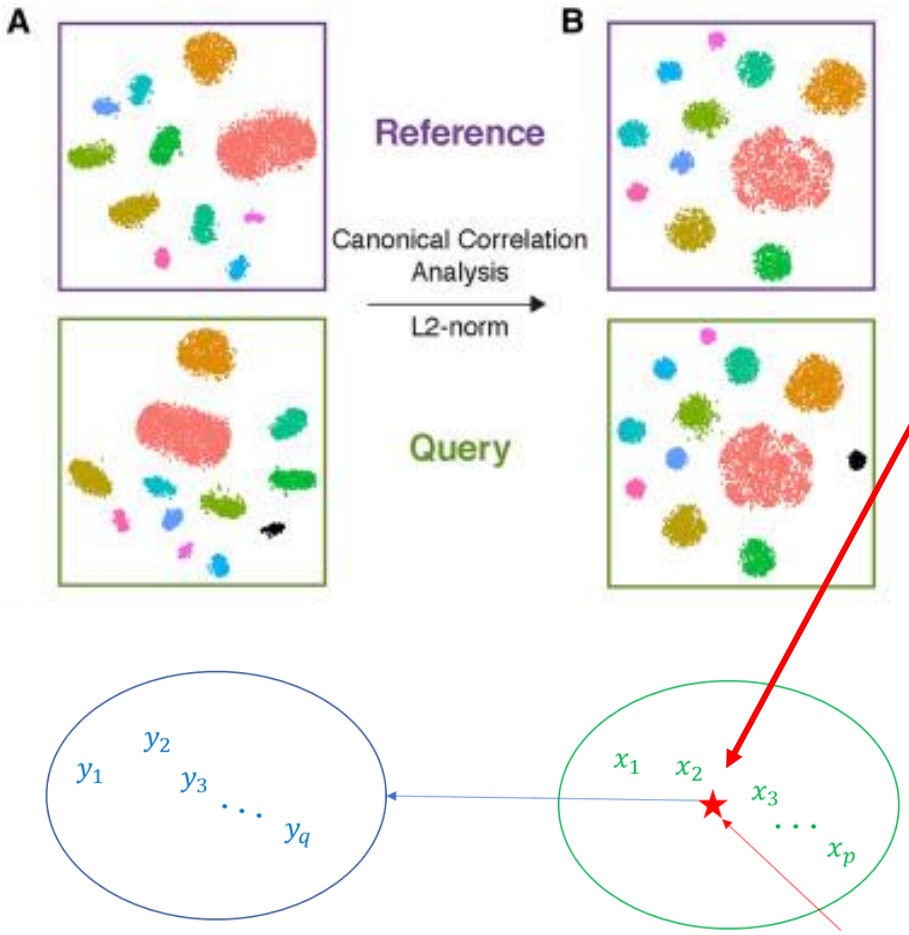


A benchmark of batch-effect correction methods for single-cell RNA sequencing data

- Batch correction (Seurat)



- Batch correction (Seurat)



*Canonical Correlation Analysis (CCA)

$$\bar{x} = \sum a_i x_i$$

$$\bar{y} = \sum b_j y_j$$

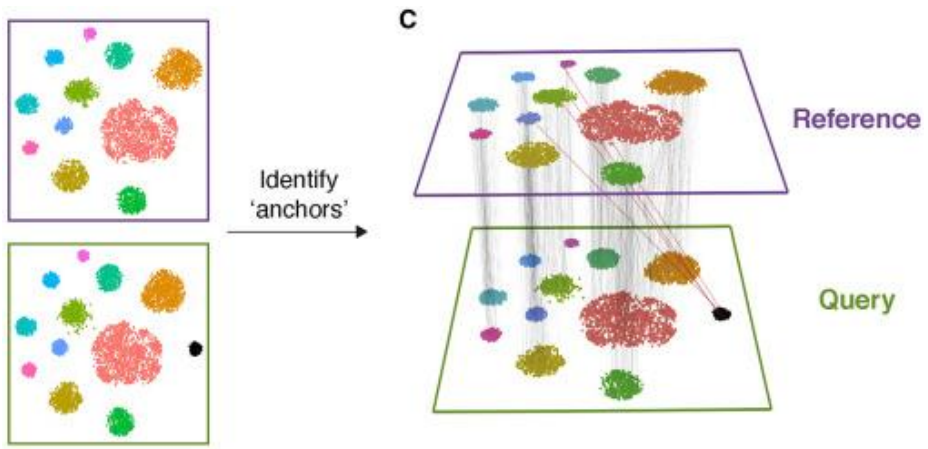
- x,y: gene expression for each group
- Linear combination for each group

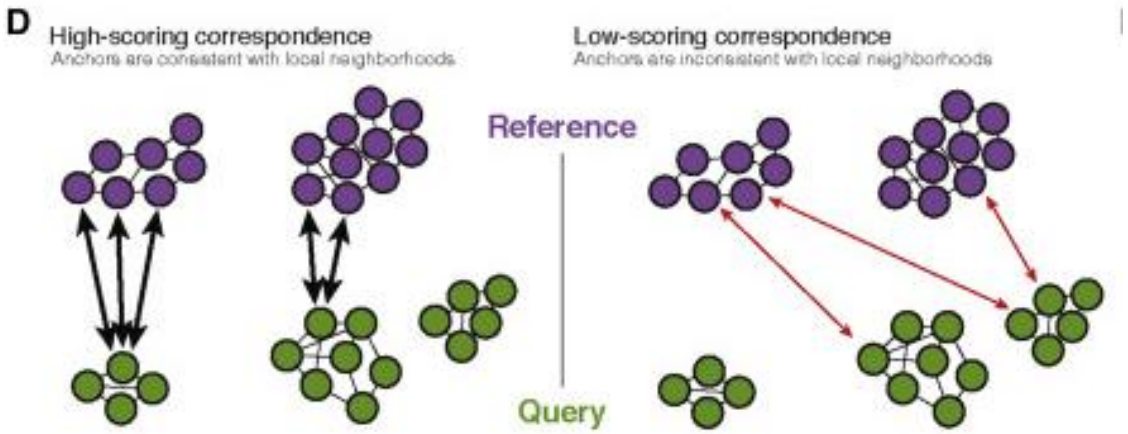
→ Maximize the correlation between $\bar{x}$ & $\bar{y}$

→ L2-normalization: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{d} |x_i|^2}$

Normalize each canonical vector

- Batch correction (Seurat)





*Find anchor by MNN (mutual nearest neighbor)
- Batch1, sample1 → KNN (k-nearest neighbor) from batch2
- See if there is a pair of samples by KNN
- Go back to raw gene expression (top 200 genes from CCA) → KNN (200) for anchor cells in the ref
→ See if query exists in 200 neighbors

- SSN (shared nearest neighbor): see if the neighbor of ref-anchors has a similar neighbors of query-anchor
→ SSN will be used to weigh each anchor

- # Batch correction (Seurat)

*Obtaining batch corrected expression

$$B = Y[,a] - X[,a]$$

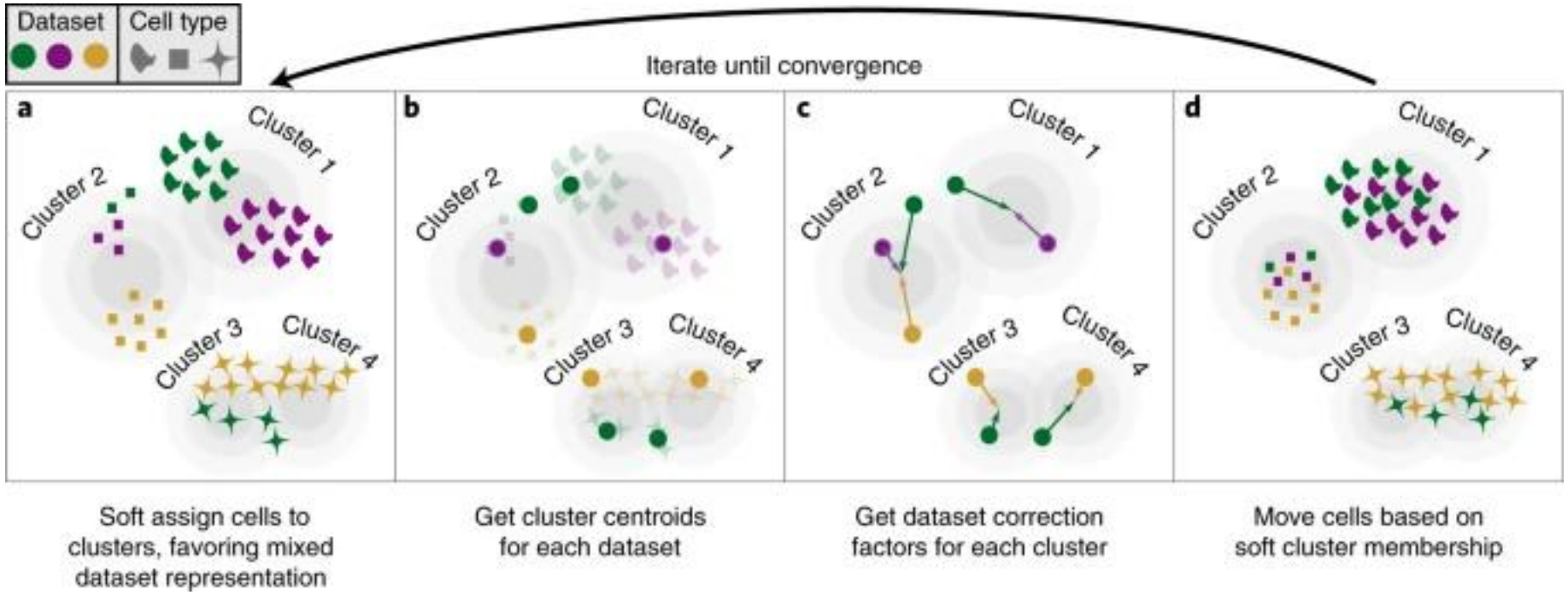
$$C = BW^T$$

$$\widehat{Y} = Y - C$$

B: batch effect (X,Y: gene expression space from each batch)
C: correction
W: weight matrix (from anchor)
$\hat{Y}$: corrected gene expression

*Multiple data integration: pairwise integration from the closest pair first

- Batch correction (Harmony)

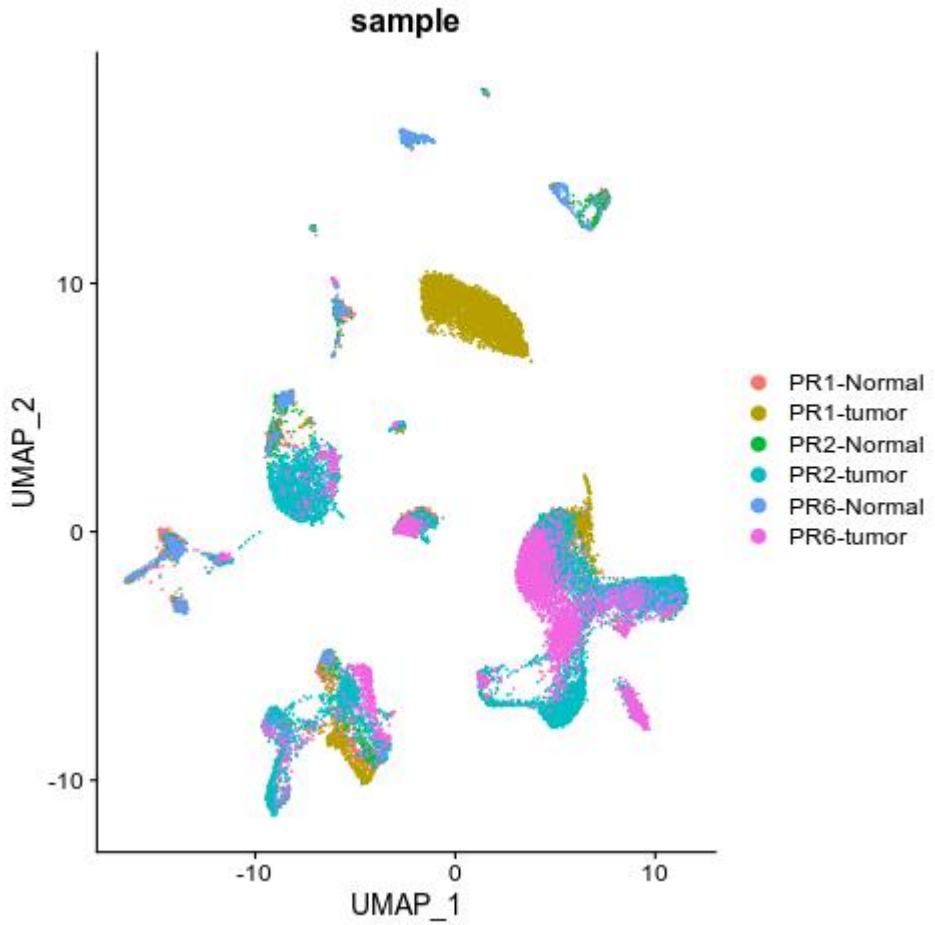


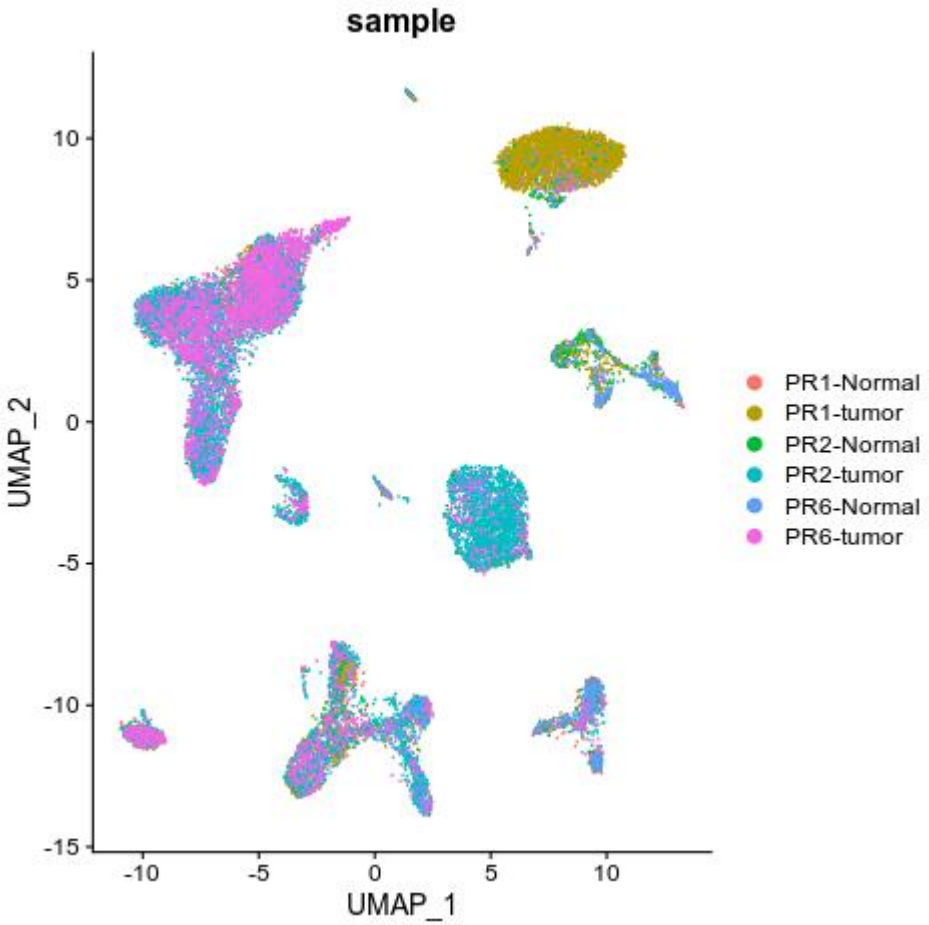
Correcting the PCA embedding into a batch corrected embedding (harmony space)

Soft k-mean clustering: k-mean clustering + entropy regularization (of each cluster membership)
Correction: batch diversity regularization
Iterate until convergence

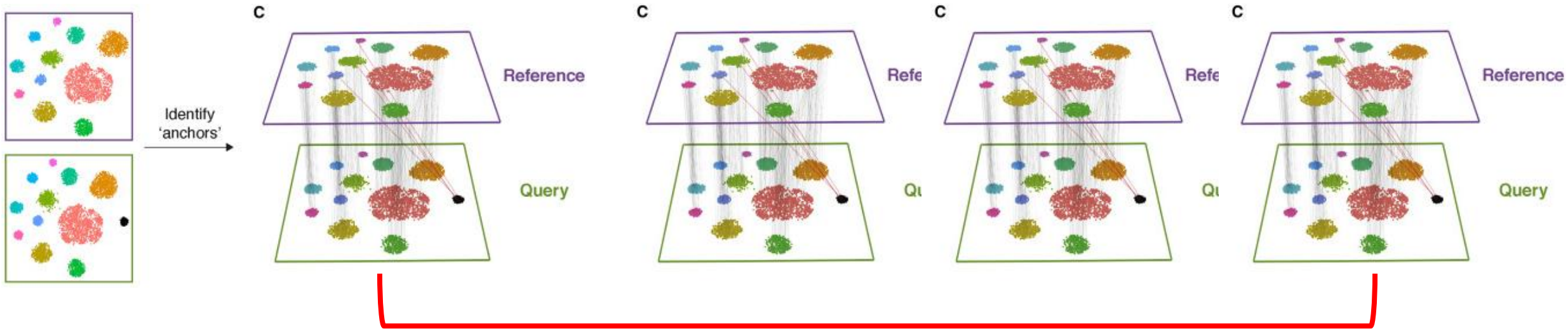
- Batch correction (Harmony)

- Sample level batch correction





- Label Transfer



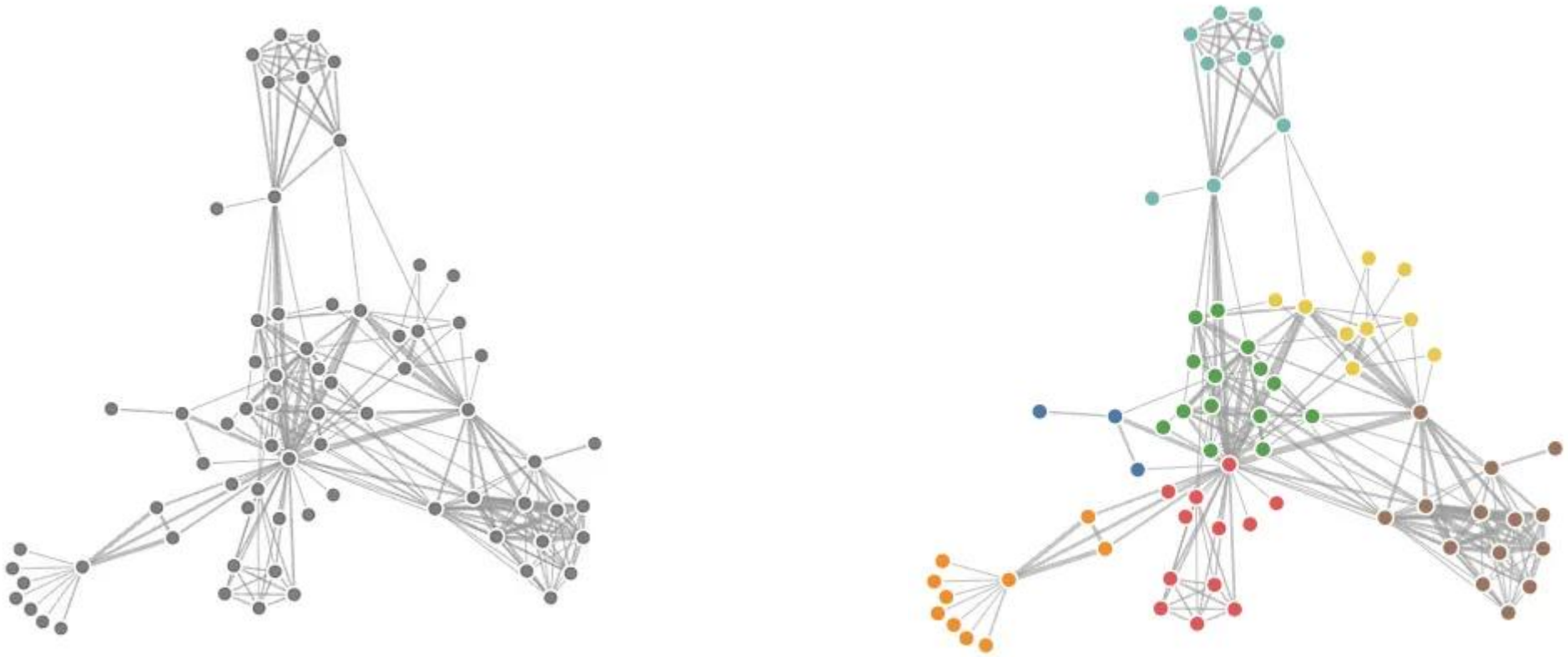
-Reference → query1, query2, query3 … (independently)
When? Large, comprehensive, and reliable reference data exists!

→ No need to celltype annotation, etc

- # Clustering

- Louvain clustering: considering the modularity of the (cell) graph

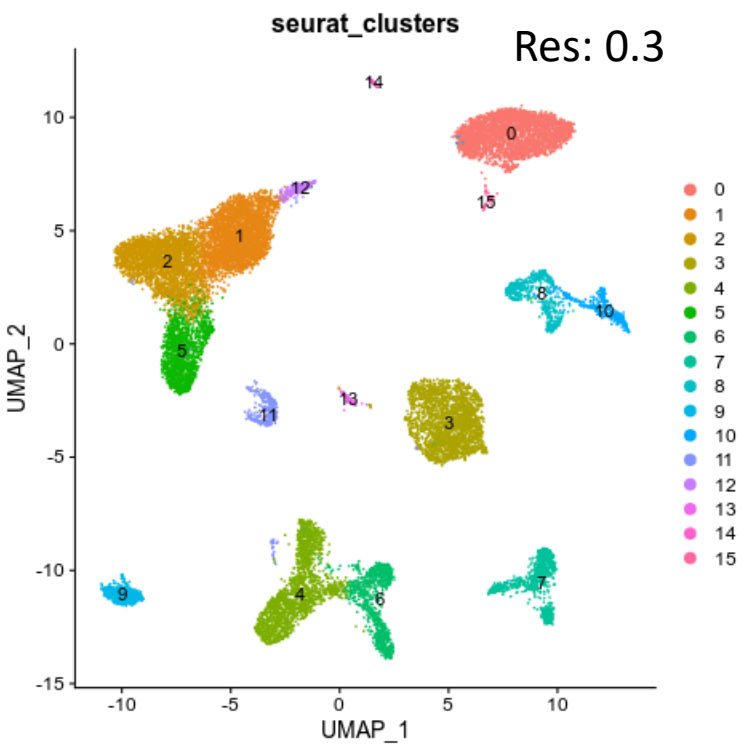
Before performing the clustering, Seurat package obtained SNN (shared-nearest neighbor) graph

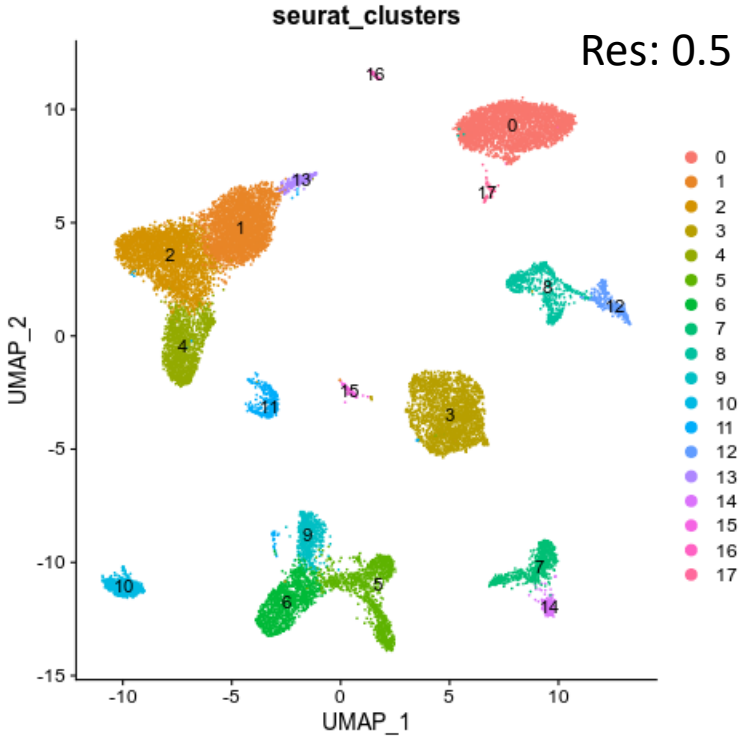


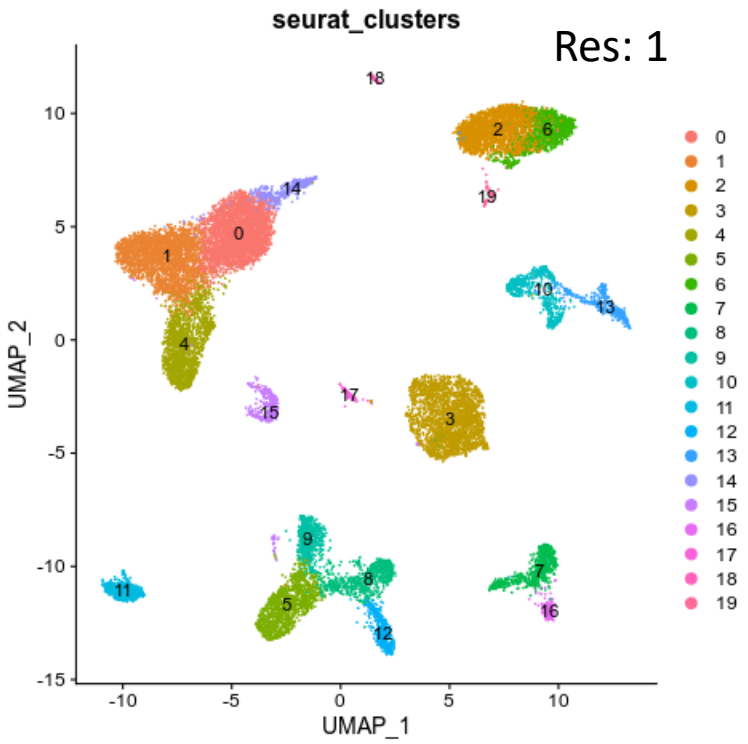
- # Clustering

- Louvain clustering

Resolution → controls the number of clusters







- # Celltype annotation

How? Distribution of marker gene expression
-Functional marker: CD3 for T cells
-Expression marker: MHC class2 for T cells

* Immune cell
Tcell: "CD3D", "CD3E"
CD4 T cell: "CD4"
CD8 T cell: "CD8A", "CD8B"
Treg: "FOXP3", "IL2RA"
NK cell: "KLRB1", "GNLY", "KLRD1", "NKG7"
B cell: "MS4A1", "CD79B"
Macrophage: "C1QA", "C1QB", "CD14", "CD68"
Monocyte: "FCN1","S100A8", "S100A9"
Mast: "TPSAB1", "CPA3“
Cycling: "MKI67", "TOP2A“

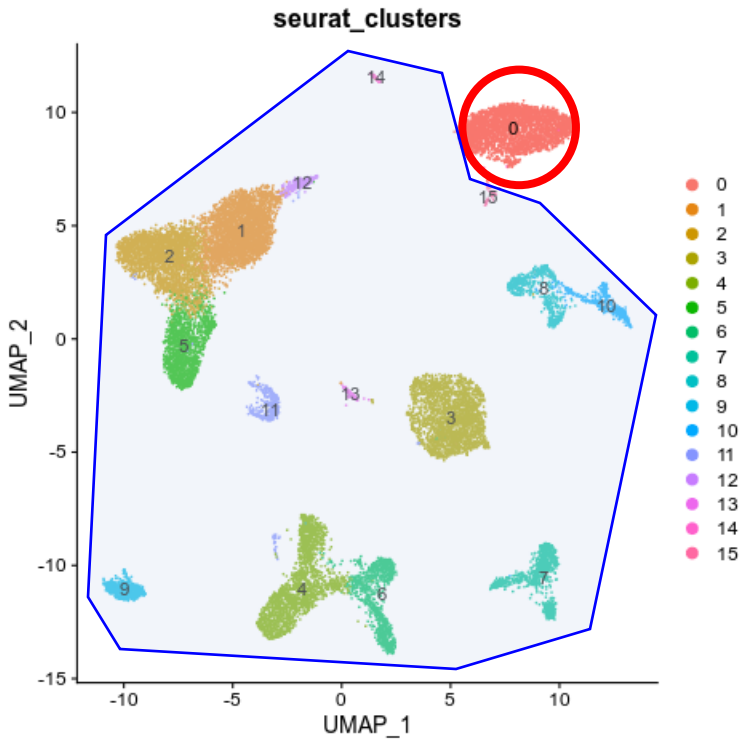
* Non-immune cell
Pericyte: "CSPG4", "MCAM", "MYH11"
Endothelial cell: "RAMP2", "RNASE1", "ENG", "EGFL7"
Cancer: "PAX8"
Epithelial cell: "SLC26A7", "EPCAM", "MUC1"

- # Celltype annotation

FindAllMarkers: This is not a “marker” but just DEG! Don’t confuse

Wilcoxon-rank sum test + Bonferroni correction

→ Nonparametric approach (does not require a specific distribution of data)

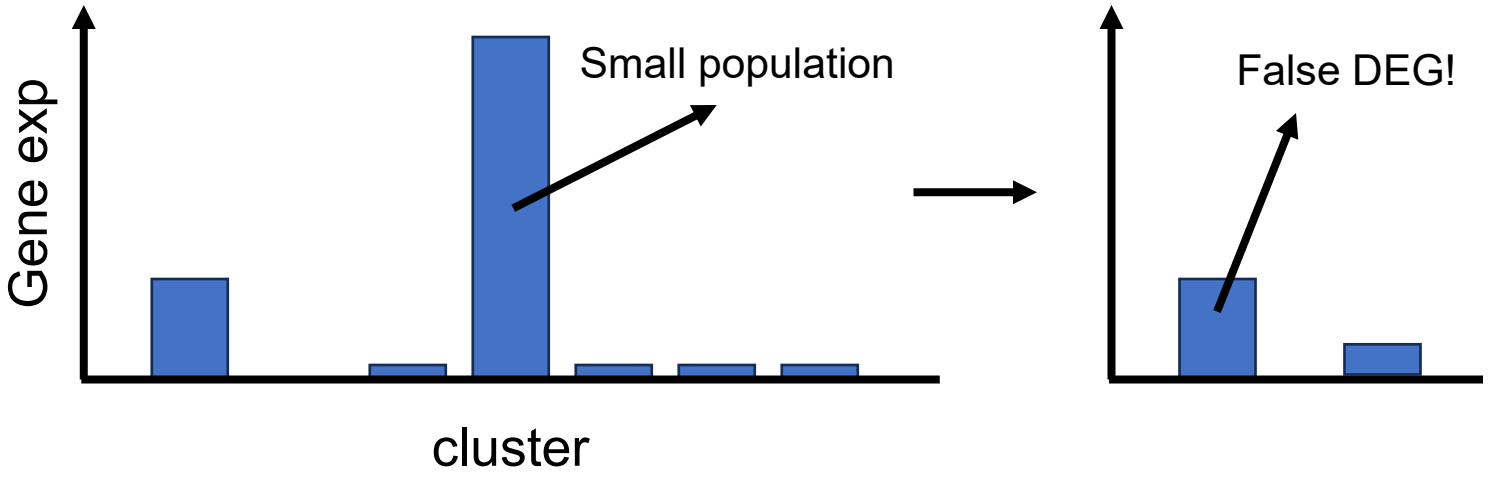


0 cluster vs the others (1~15)

Same for every cluster

Caveat!

→ Dilution effect → False positive



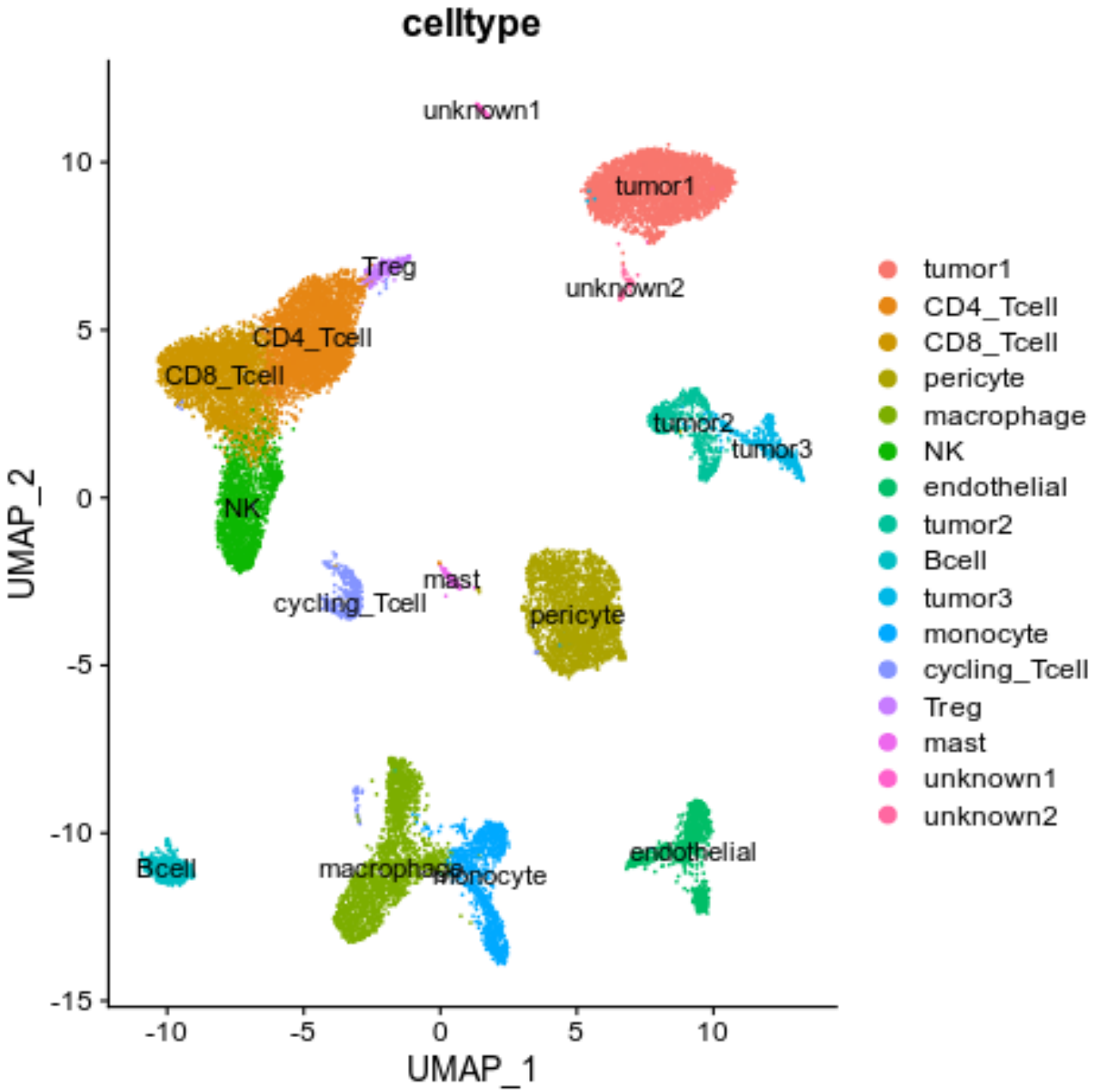
- # Celltype annotation

FindAllMarkers

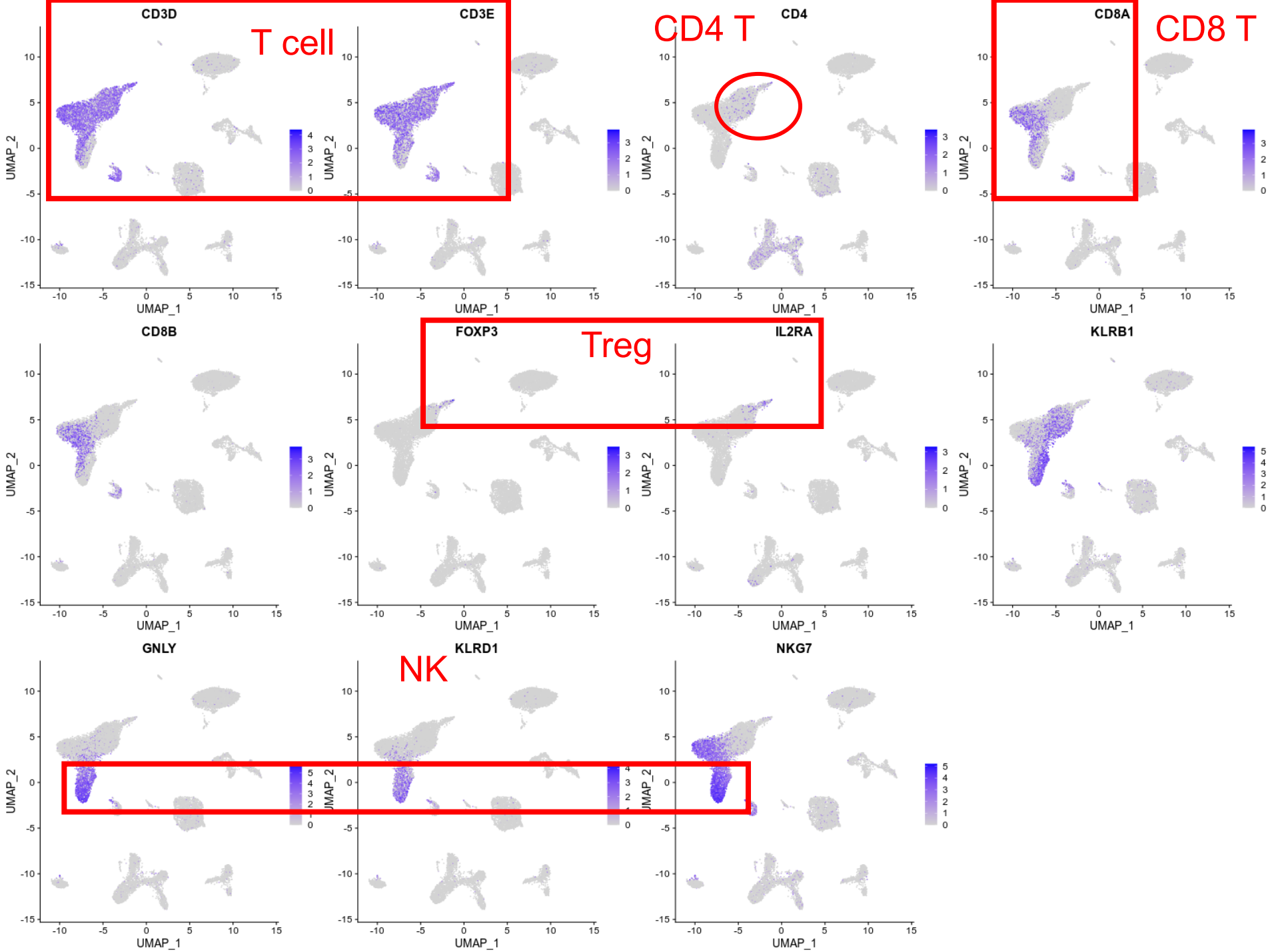
Adjusted p-value, average_Log2FC + expression cell ratio (pct.1, pct.2)

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
CRYAB	0	4.405264	0.996	0.158	0	0	CRYAB
RBP4	0	3.934706	0.880	0.032	0	0	RBP4
UGT2B7	0	3.844966	0.950	0.039	0	0	UGT2B7
SPP1	0	3.729567	0.924	0.081	0	0	SPP1
WFDC2	0	3.711430	0.954	0.081	0	0	WFDC2
CLU	0	3.401655	0.915	0.089	0	0	CLU
DEFB1	0	3.202296	0.934	0.110	0	0	DEFB1
TNFRSF12A	0	2.974357	0.872	0.094	0	0	TNFRSF12A
PDZK1IP1	0	2.943278	0.761	0.028	0	0	PDZK1IP1
KRT18	0	2.863520	0.847	0.068	0	0	KRT18
TFPI2	0	2.857835	0.692	0.022	0	0	TFPI2
AC073218.2	0	2.825047	0.808	0.029	0	0	AC073218.2
SOSTDC1	0	2.822084	0.679	0.018	0	0	SOSTDC1
IGFBP6	0	2.698095	0.809	0.064	0	0	IGFBP6
CXCL14	0	2.682915	0.777	0.053	0	0	CXCL14
TMEM176A	0	2.603952	0.756	0.043	0	0	TMEM176A
SLPI	0	2.561452	0.519	0.020	0	0	SLPI
TSPAN1	0	2.555455	0.718	0.028	0	0	TSPAN1

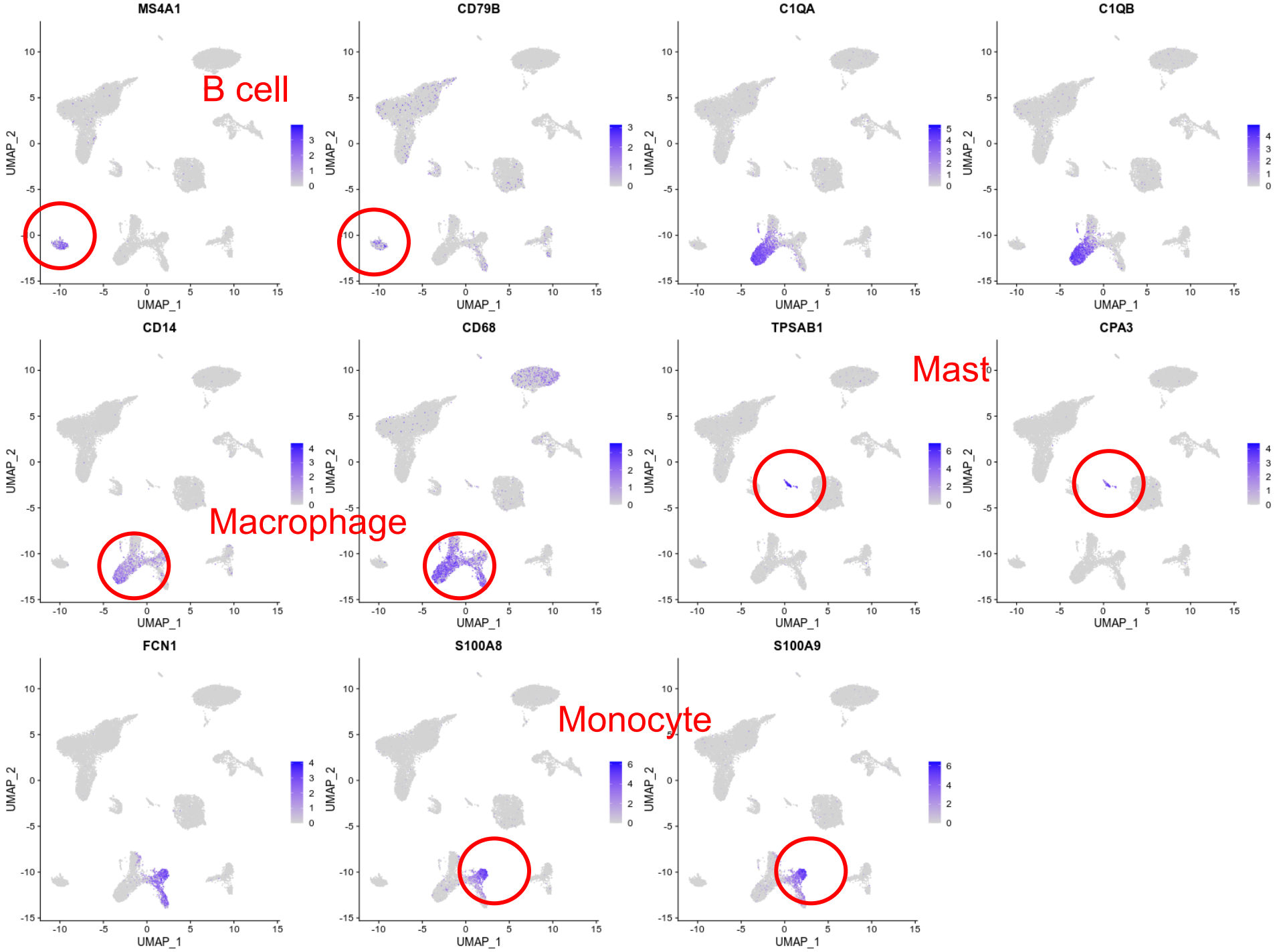
- Celltype annotation



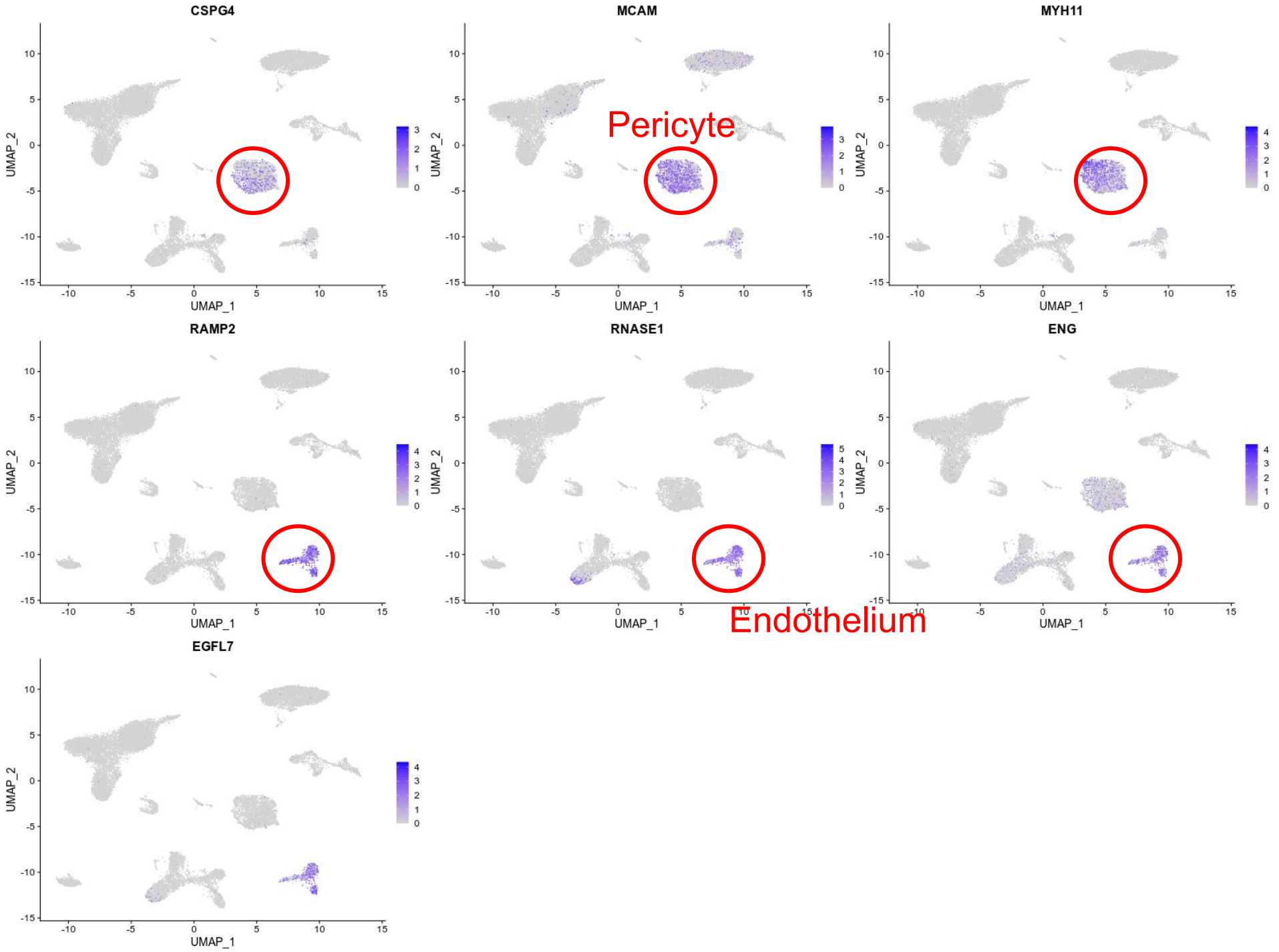
T cell and NK


CD3D
T cell
CD3E
CD4 T
CD4
CD8A
CD8 T
CD8B
FOXP3
Treg
IL2RA
KLRB1
GNLY
KLRD1
NK
NKG7
UMAP_1
UMAP_2

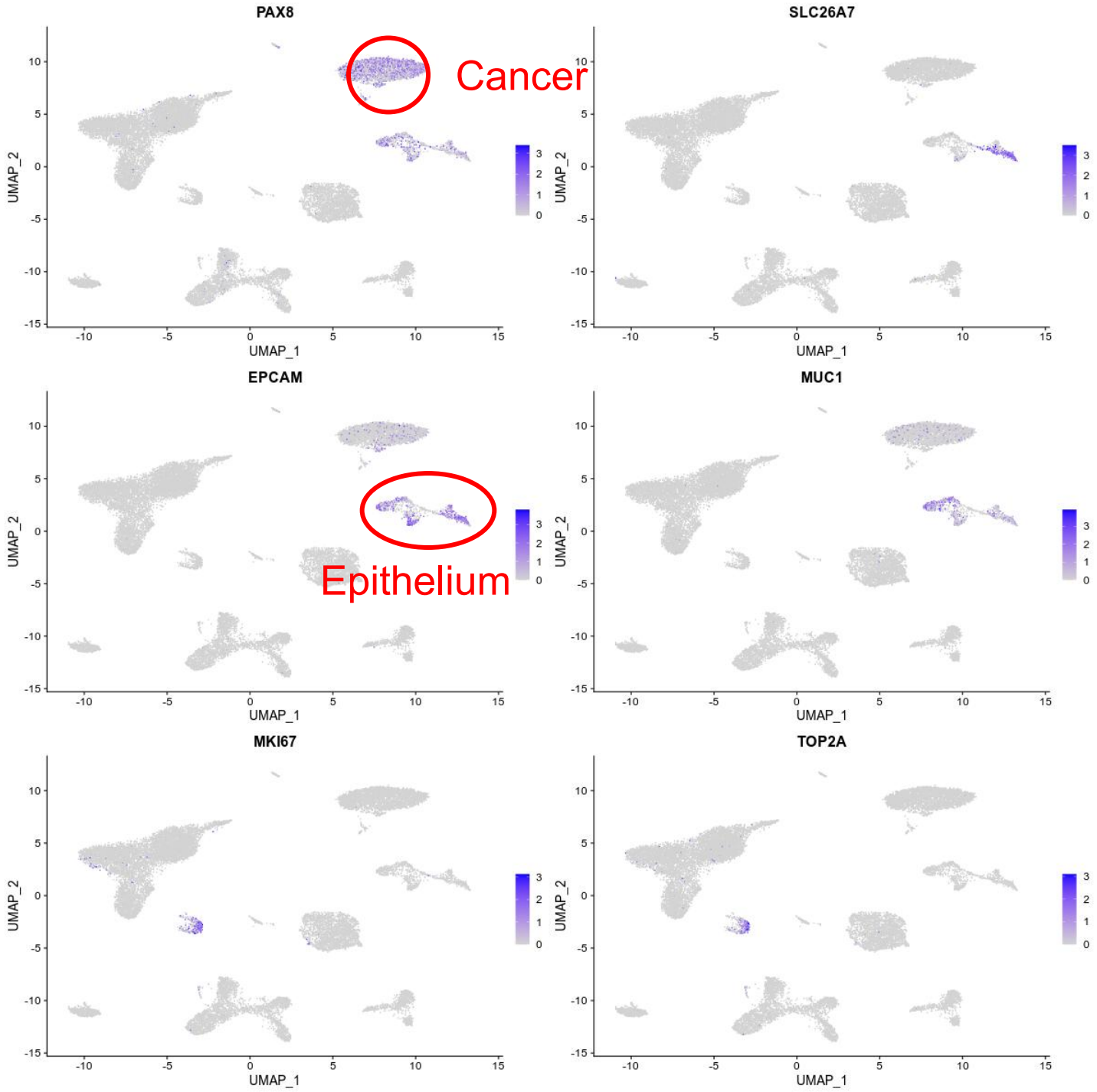
Bcell and myeloid



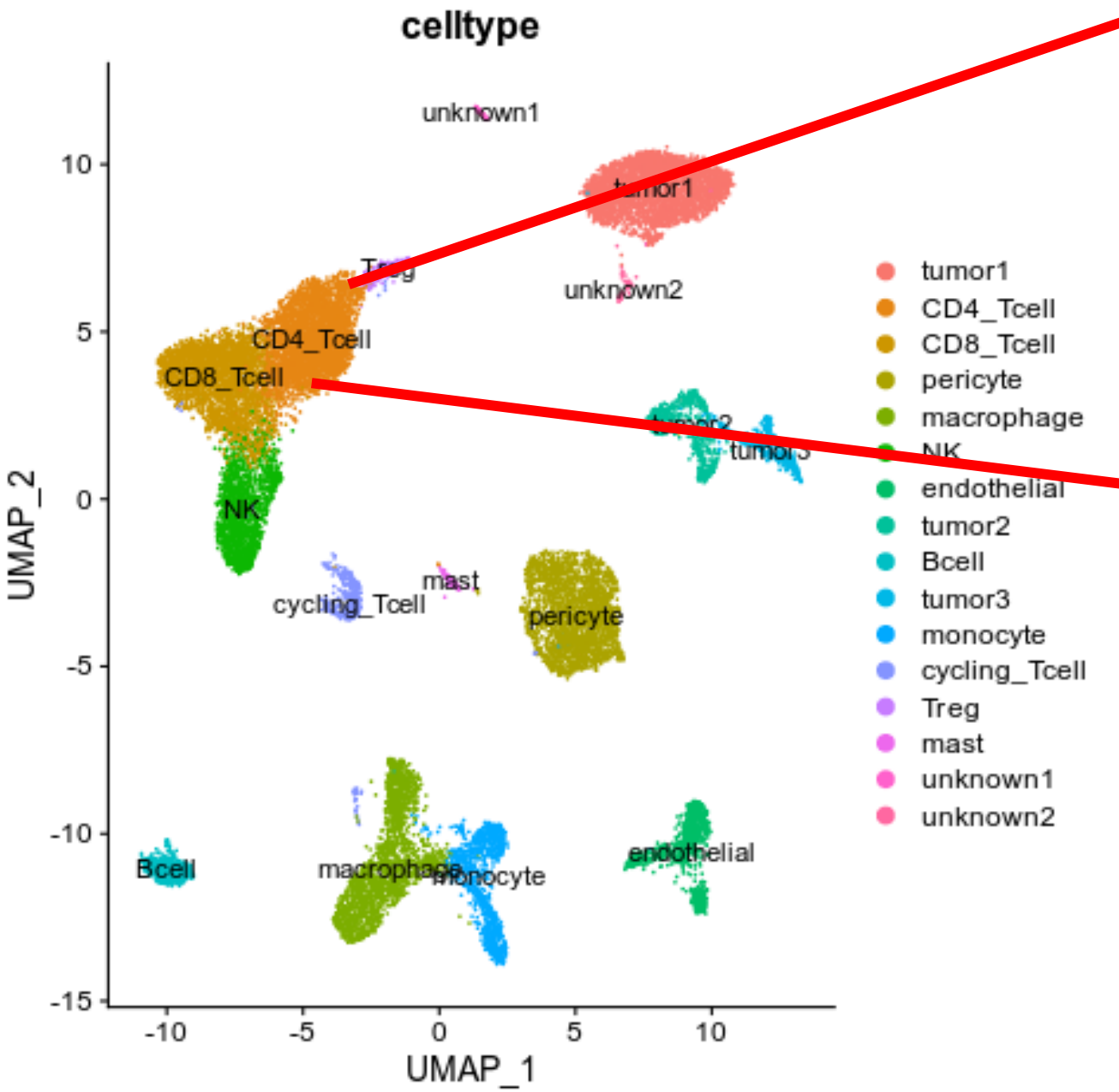
Stroma cell

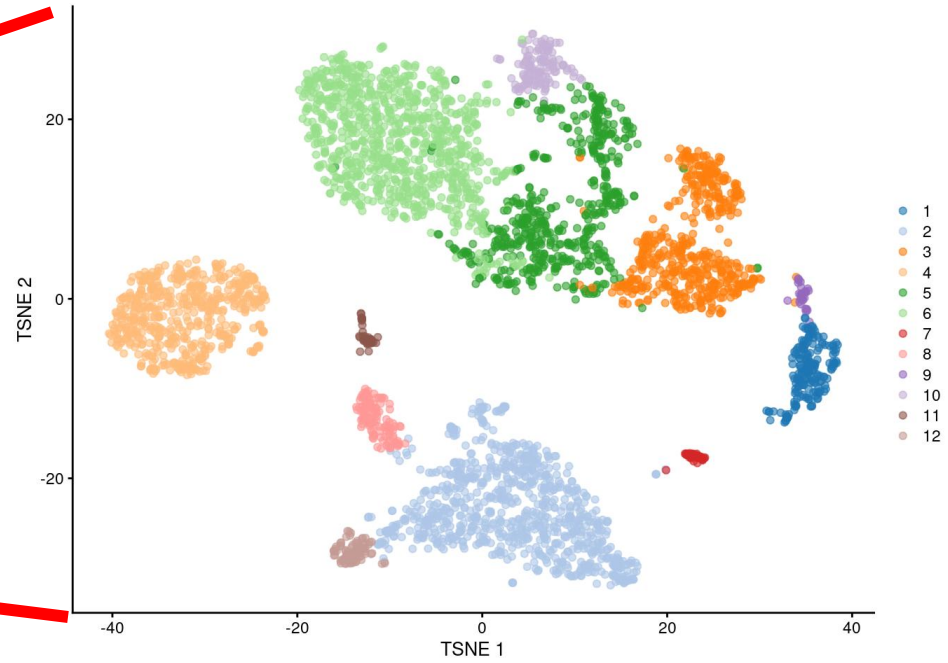

CSPG4
MCAM
MYH11
Pericyte
RAMP2
RNASE1
ENG
Endothelium
EGFL7
UMAP_1
UMAP_2

Epithelial and tumor



- Celltype annotation





Subsetting and reclustering → Annotation
- HVG can be reset: better for comparing within CD4_T cell subclusters