單細胞RNA定序資料之產生與數據分析簡介
Introduction on the generation and analysis of single-cell RNA sequencing data

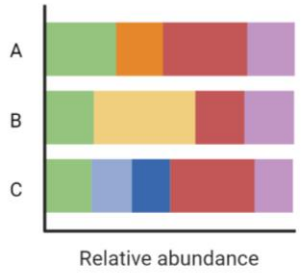
國家衛生研究院
癌症研究所
江士昇 副研究員

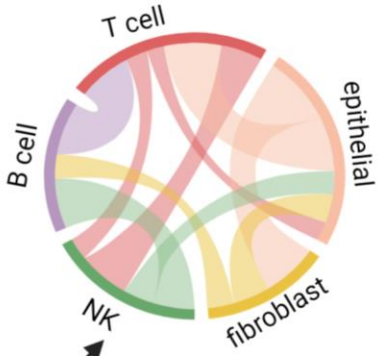
Outline

1. Single-cell RNA-sequencing (scRNA-seq) in cancer research.
2. Bulk-tissue *versus* single-cell RNA sequencing
3. Workflow of scRNA-seq experiment.
4. Barcoding
5. Data preprocessing: quality control
6. Data normalization
7. Dimension reduction, clustering and cell typing.
8. Quantitative analysis
9. Pathway analysis
10. Analytic tools and software
11. Challenges and limitations

Single-cell RNA sequencing analysis

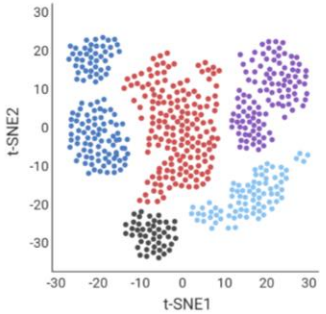
Cell type composition

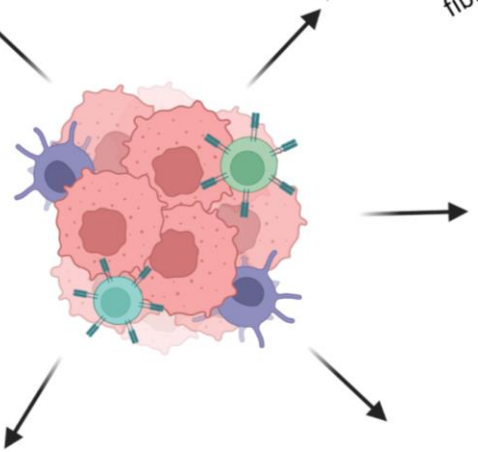




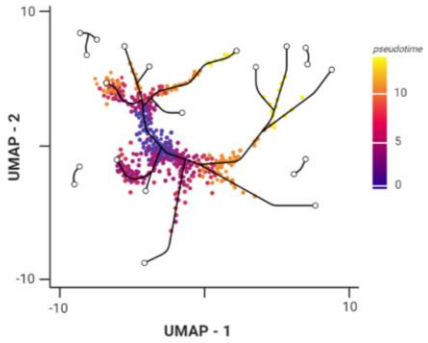
Cell-to-cell communication

Identify special cell populations

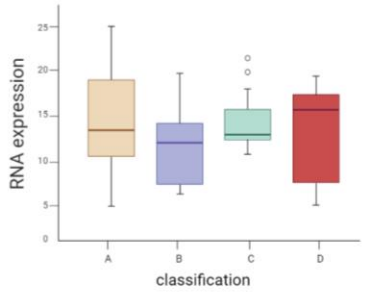


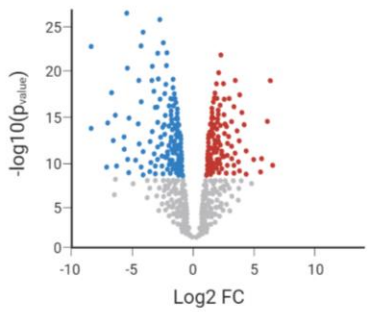


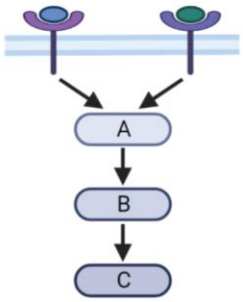
Trajectory analysis (Pseudotime)



Gene expression between classification

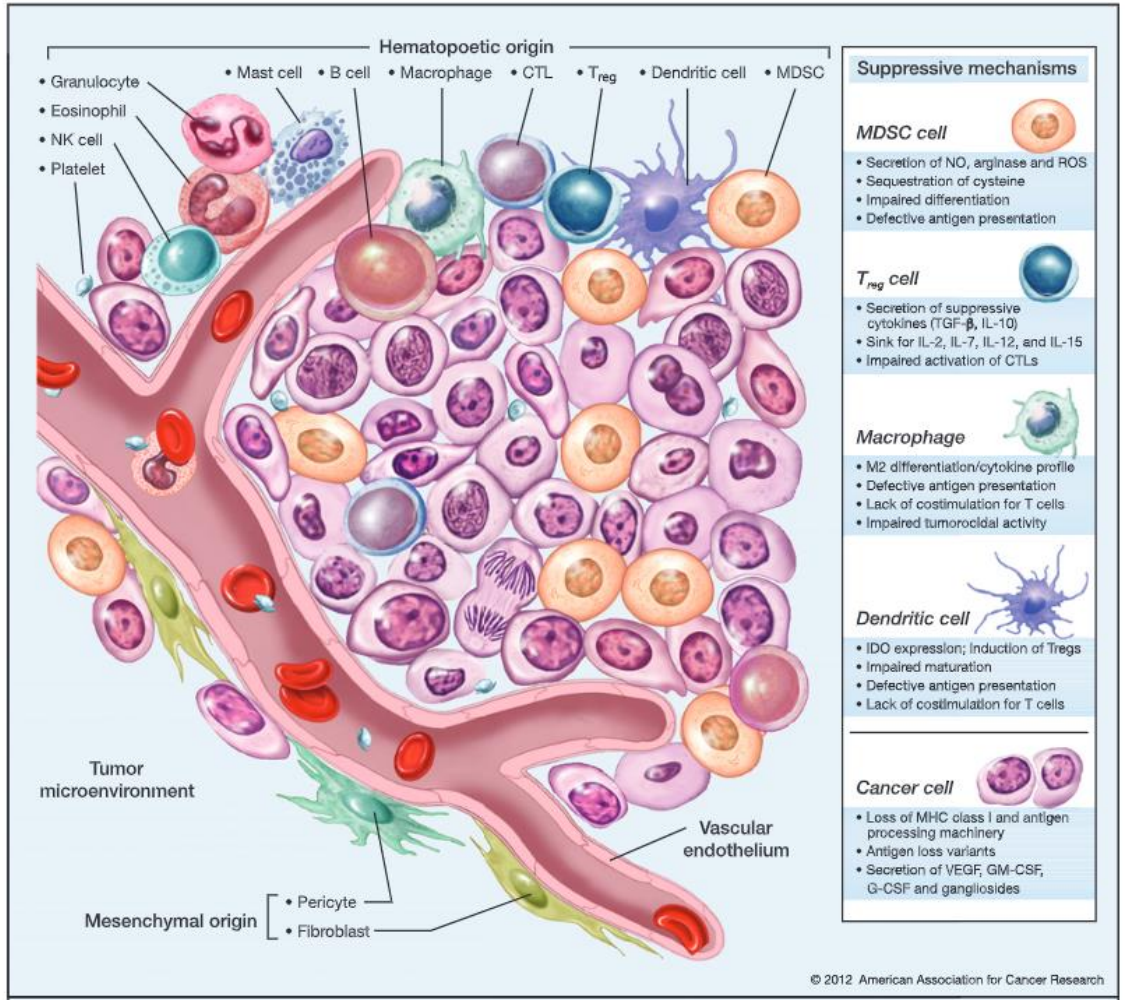






Pathway and functional annotation

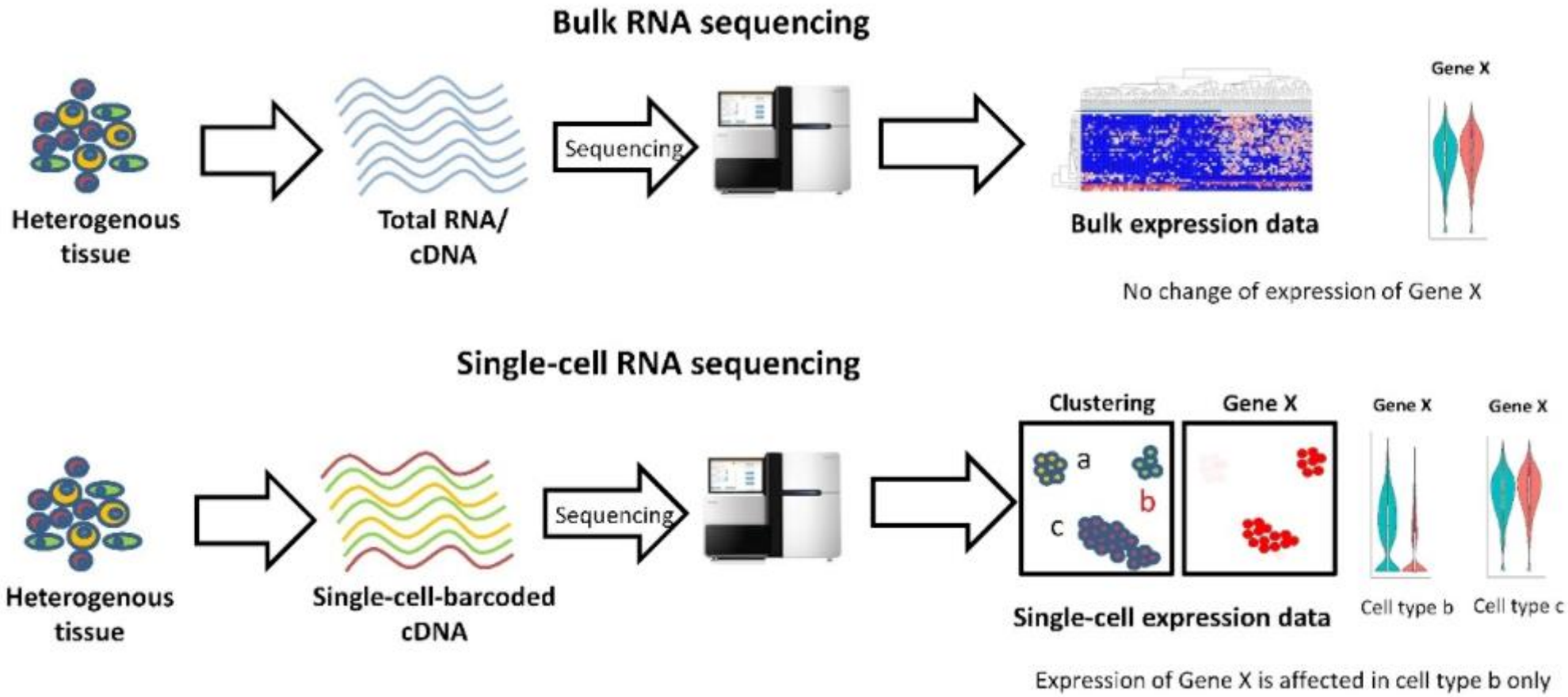
Single-cell RNA sequencing (scRNA-seq) in cancer research



- Immunotherapy
- Immuno-oncology
- Cancer associated fibroblast
- Abnormal vasculature/angiogenesis
- Quantitative and qualitative characterization of heterogenous types of cells in the tumor microenvironment (TME).

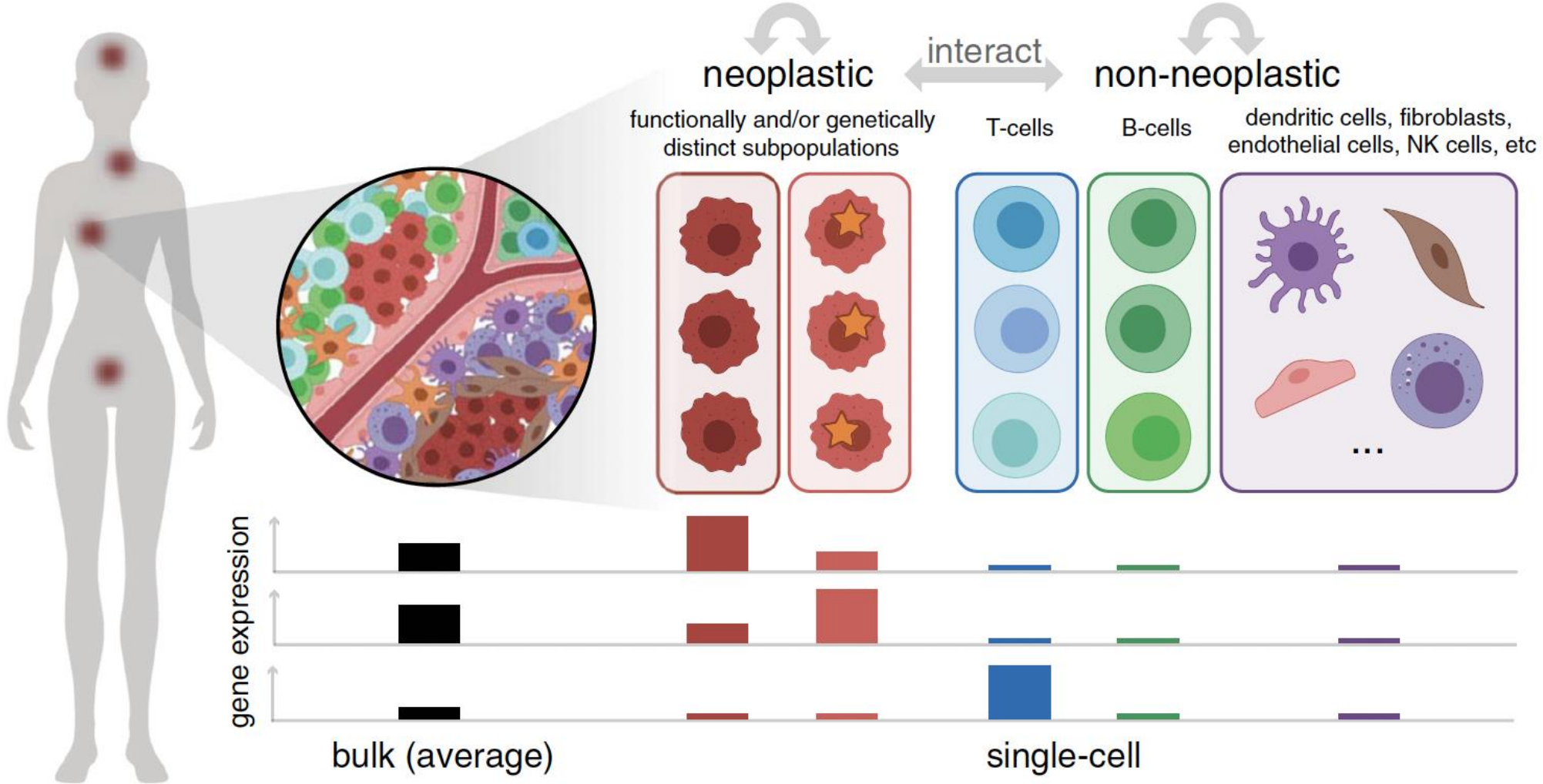
Cancer Res; 72(13); 3125–30.
DOI: 10.1158/0008-5472.CAN-11-4094

Bulk-tissue *versus* single-cell RNA sequencing



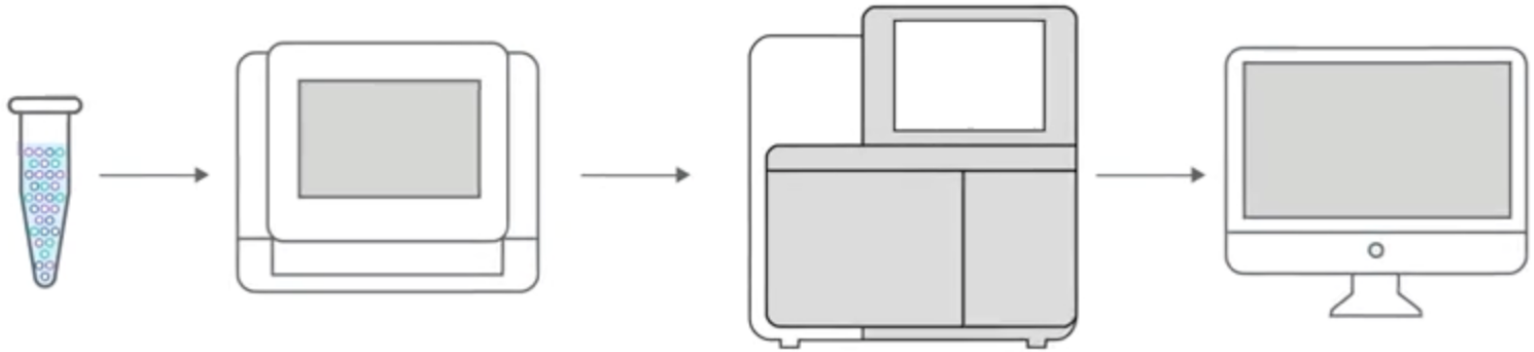
https://www.technologynetworks.com/genomics/articles/recent-advances-in-single-cell-genomics-techniques-324695

Bulk-tissue *versus* single-cell RNA sequencing

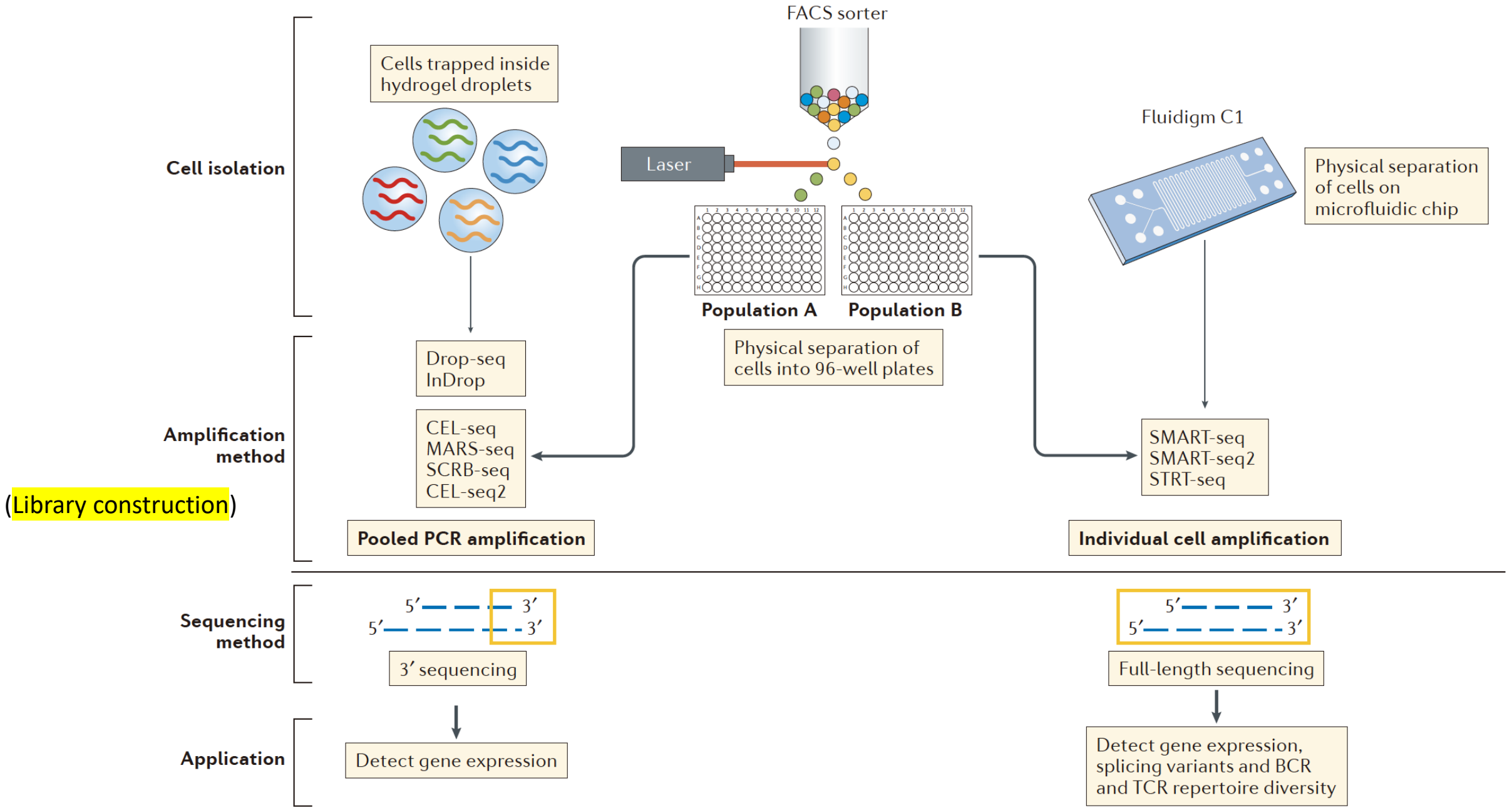


Fan et al. *Experimental & Molecular Medicine* (2020) 52:1452–1465

Workflow of scRNA-seq experiment



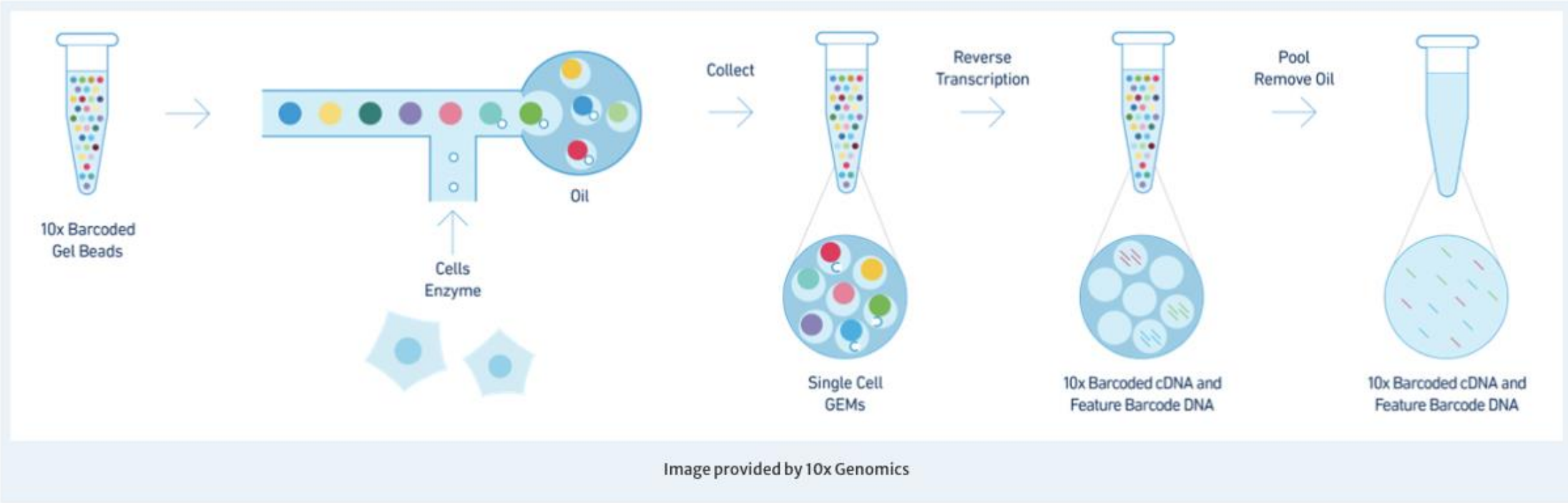
Cell isolation Library construction Sequencing Data analysis



(Library construction)

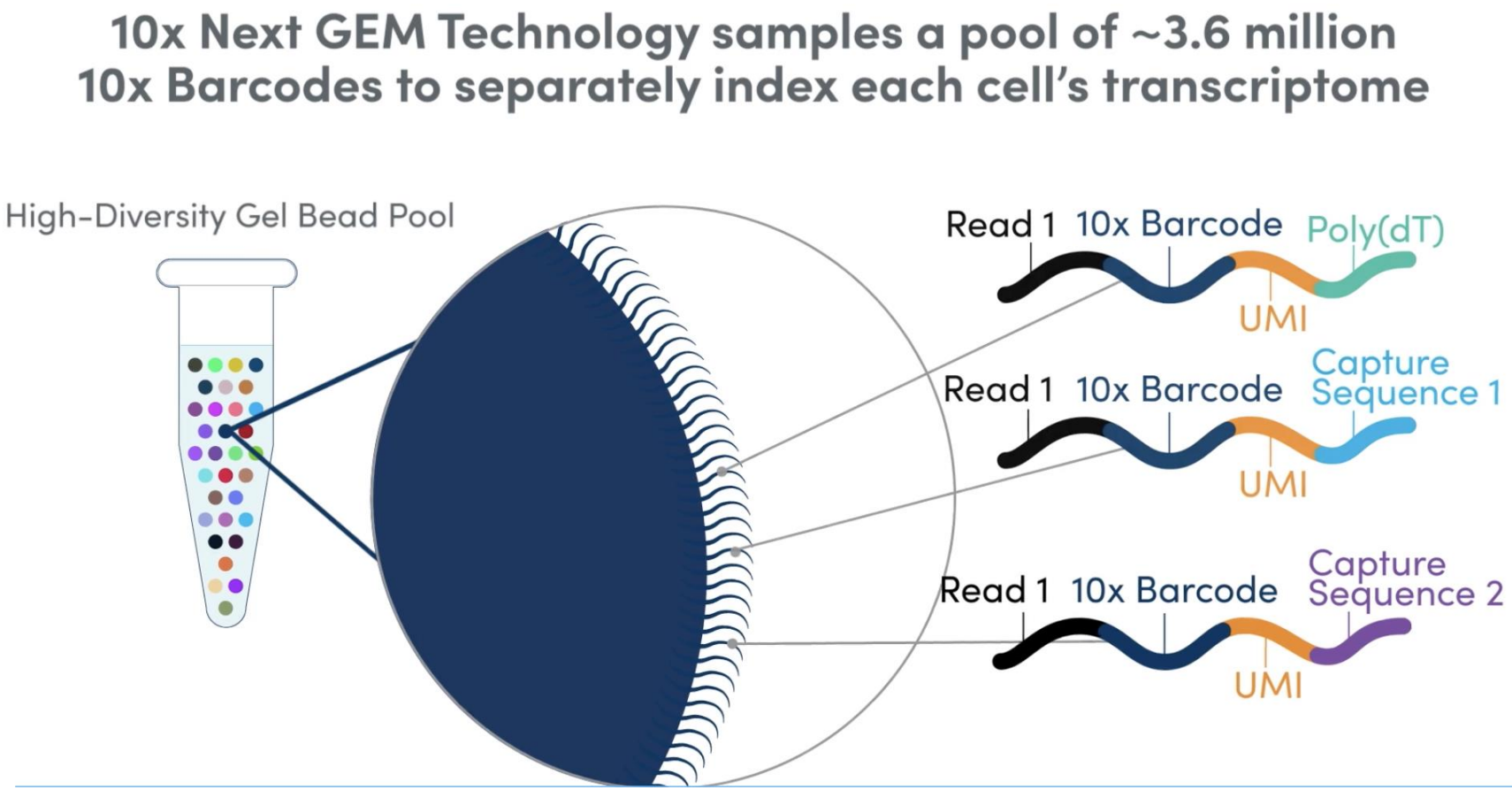
NATURE REVIEWS IMMUNOLOGY VOLUME 18, JANUARY 2018, 35

Workflow (10X Genomics' Chromium system):

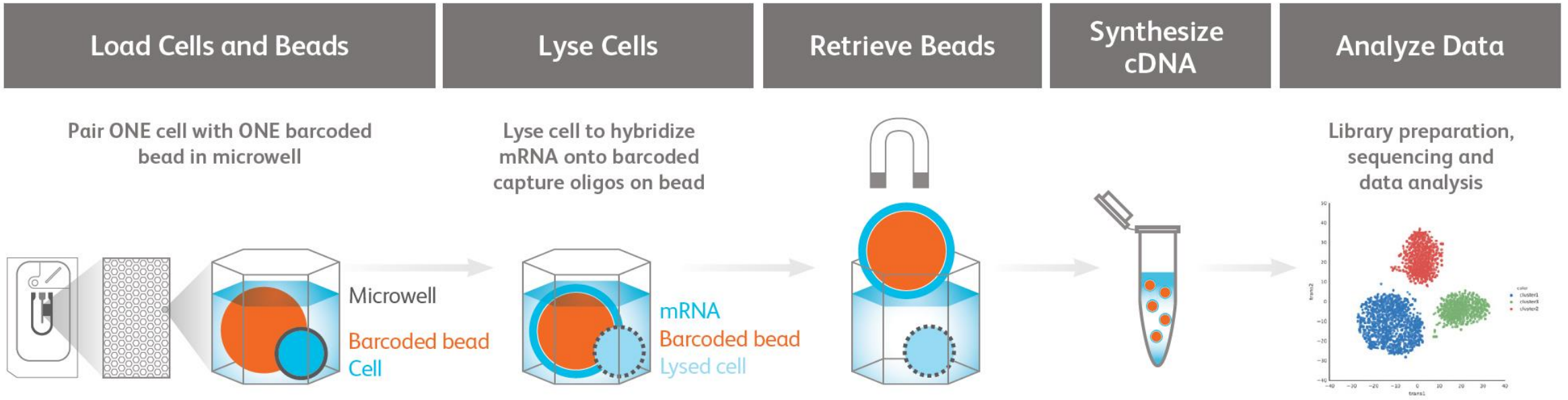


See how it works (https://www.10xgenomics.com/products/single-cell-gene-expression)

Gel beads

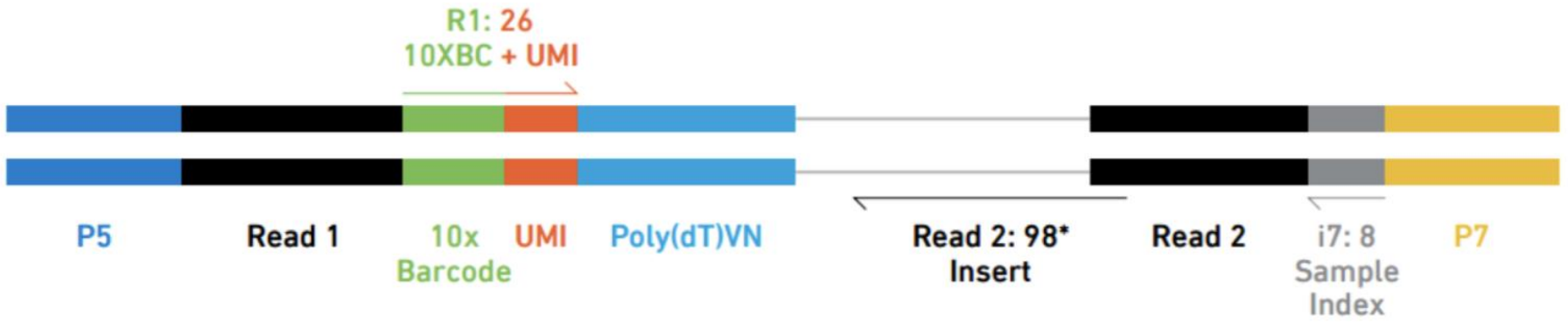

10x Next GEM Technology samples a pool of ~3.6 million
10x Barcodes to separately index each cell's transcriptome
High-Diversity Gel Bead Pool
Read 1
10x Barcode
Poly(dT)
UMI
Read 1
10x Barcode
Capture Sequence 1
UMI
Read 1
10x Barcode
Capture Sequence 2
UMI

Workflow (BD Rhapsody system):



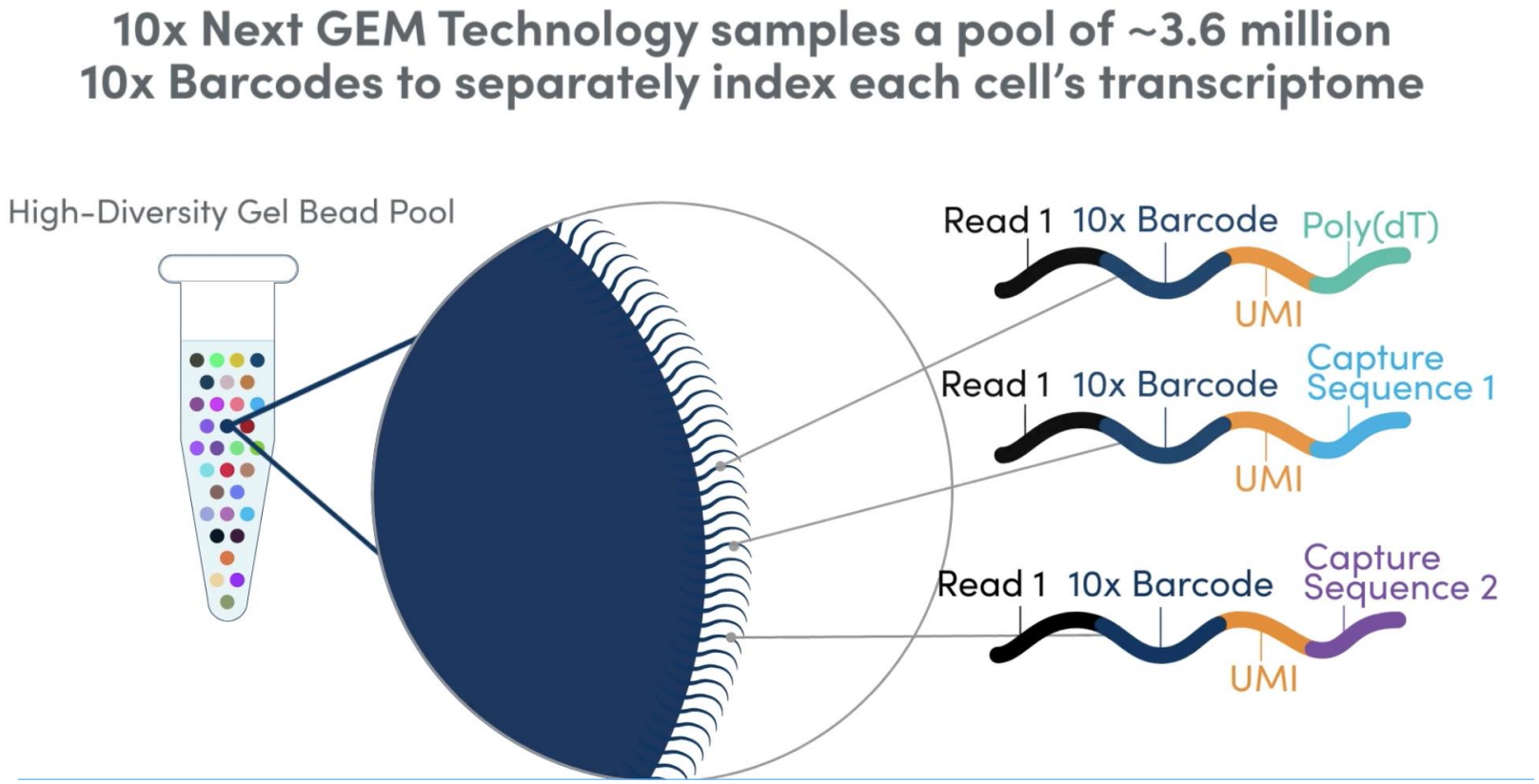
Barcoding (I)

- Cell barcode
- Unique molecular index (UMI) barcode



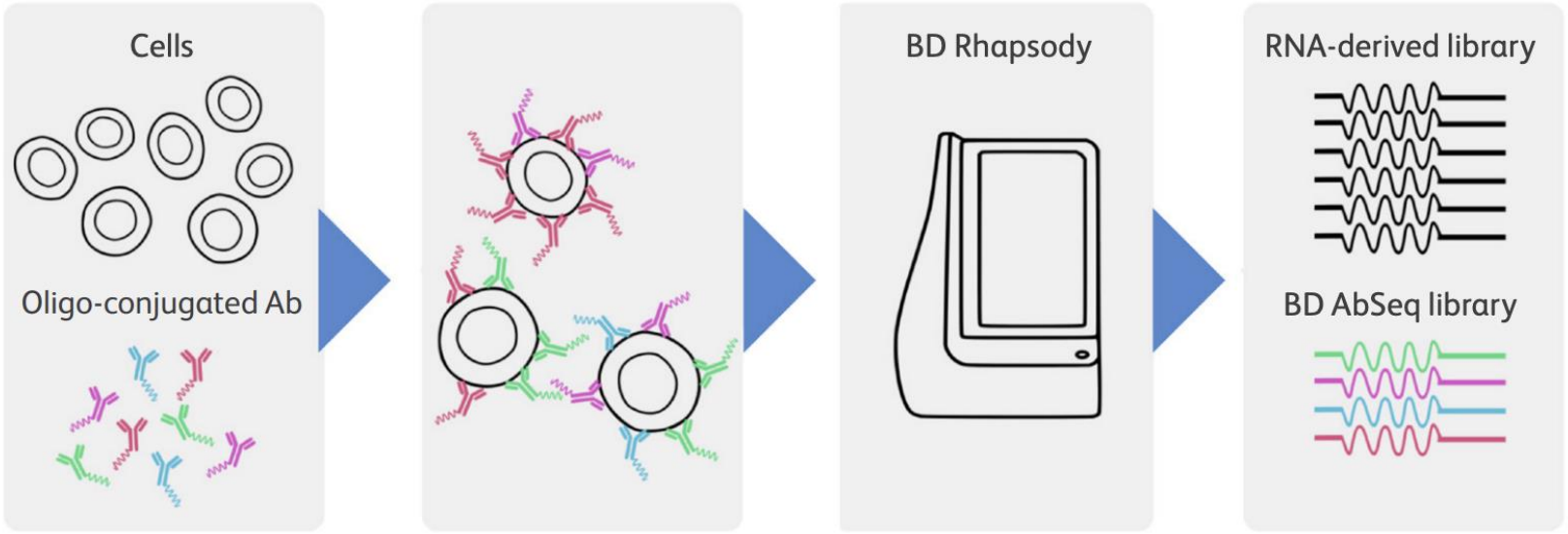
Barcoding (II)

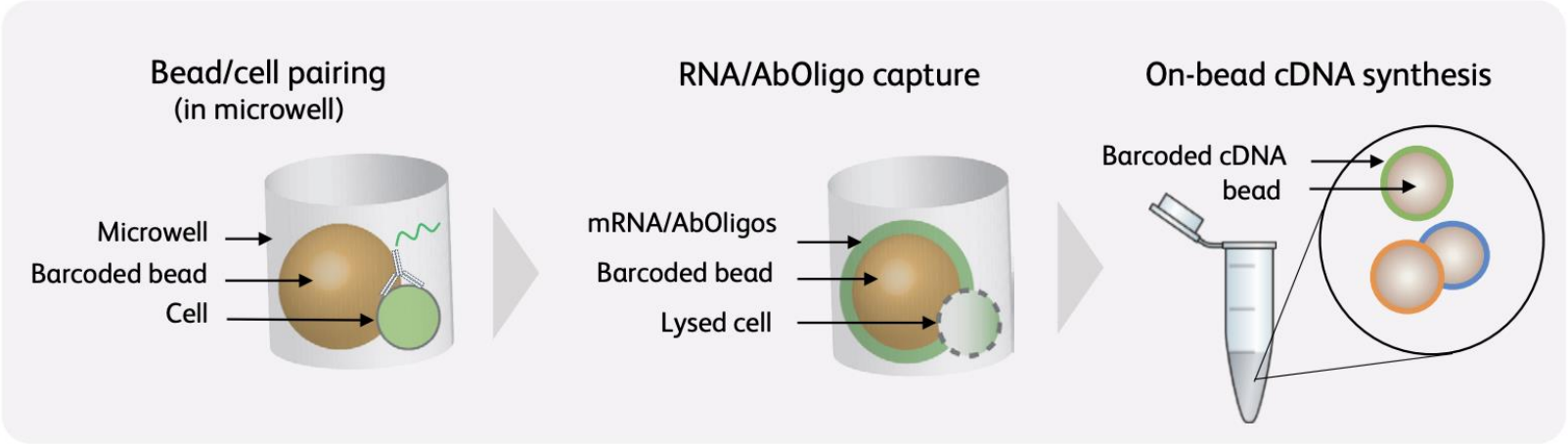
- Feature barcoding (10x)



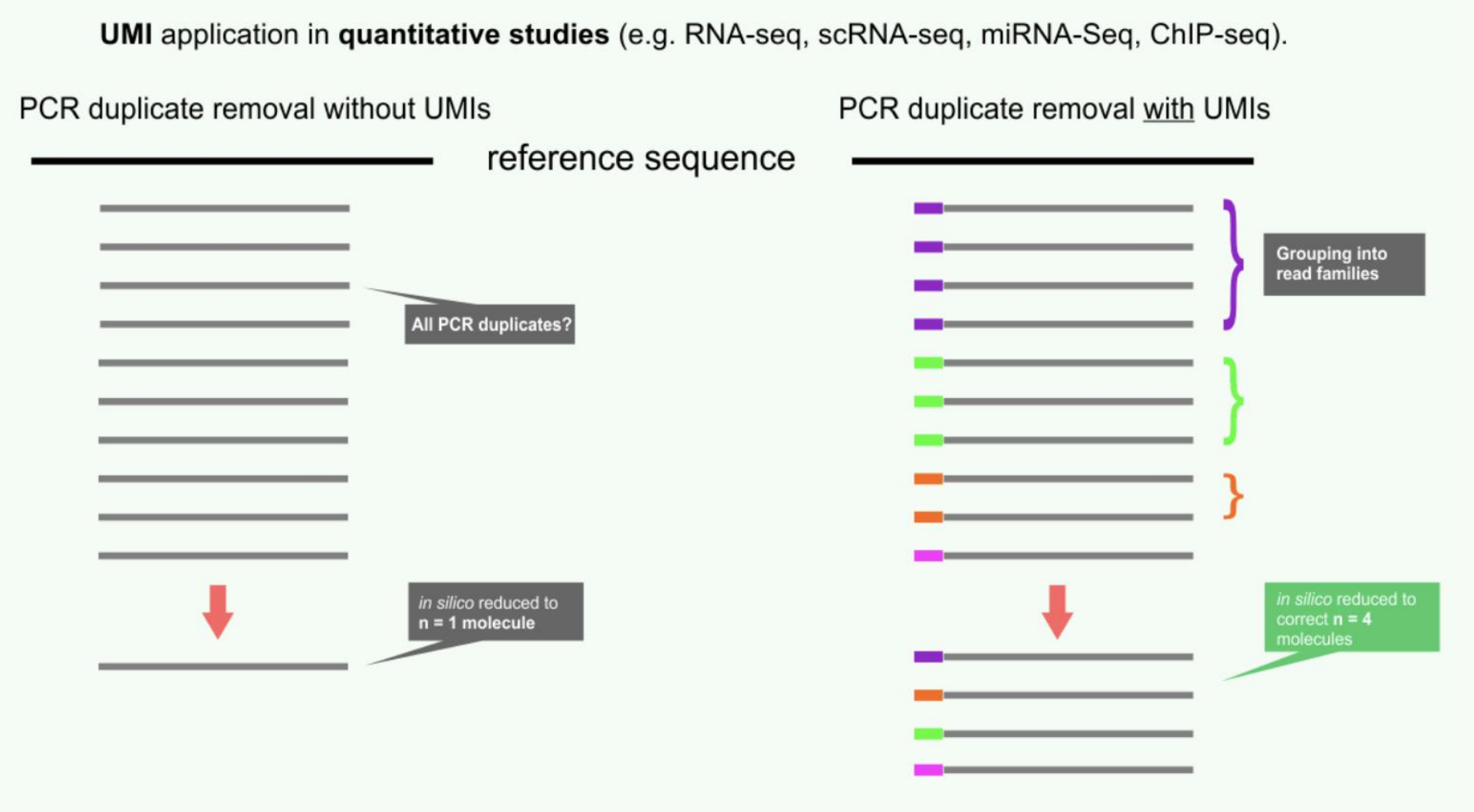
Barcoding (II)

- Ab-seq (BD)



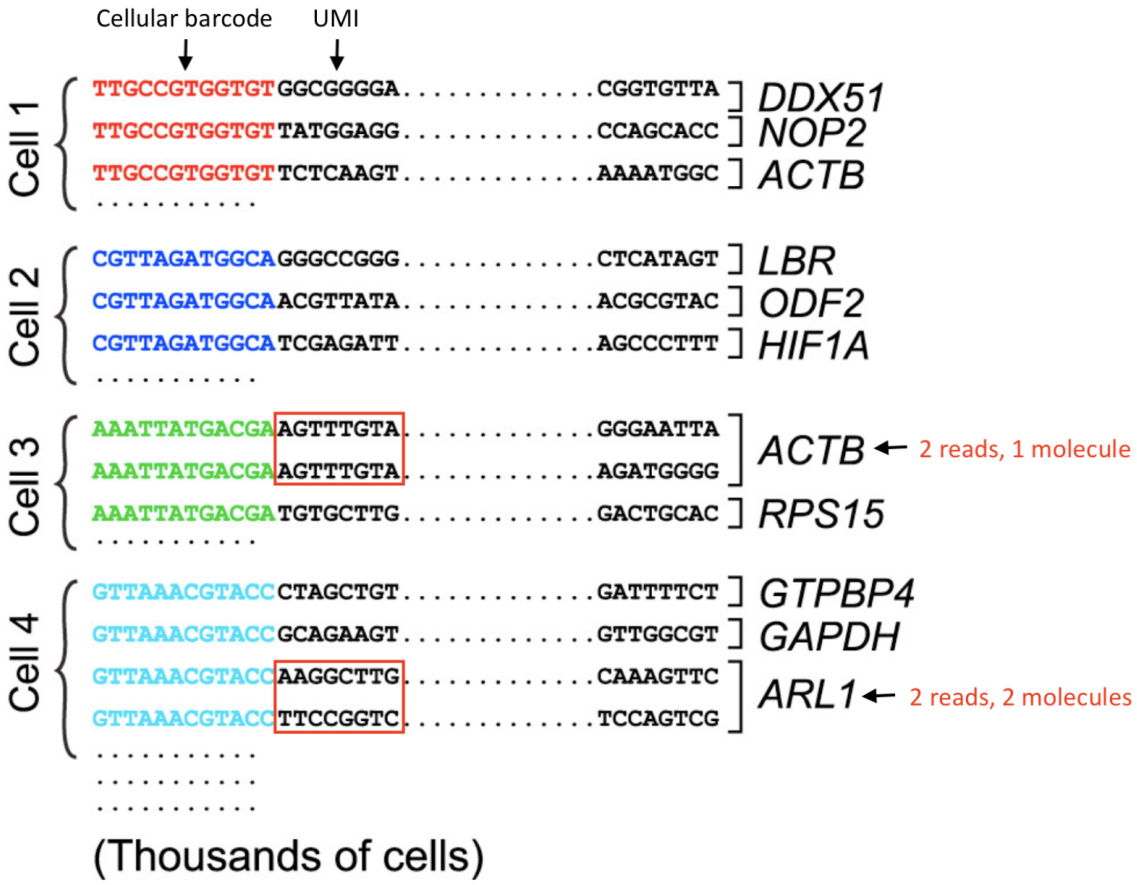


UMI (unique molecular identifier) barcode



https://dnatech.genomecenter.ucdavis.edu/faqs/what-are-umis-and-why-are-they-used-in-high-throughput-sequencing/

From sequence reads to cell-gene matrix (table)



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

Data preprocessing: quality control

- Cells
  - Library size filtering
  - Low-complexity cells filtering
  - Filtering of dead or dying cells
  - Doublet (multiplet)
- Genes
  - Filtering of low-abundance genes

Data preprocessing: quality control

- Cell ranger (10x)

10x GENOMICS™ Cell Ranger · FVB ·

SUMMARY ANALYSIS

Estimated Number of Cells

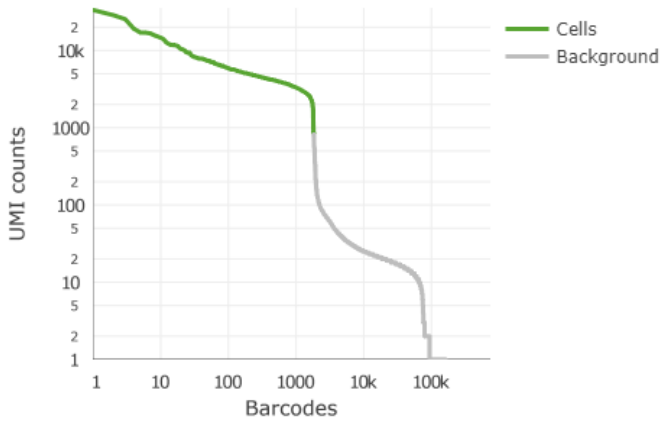
1,820

Mean Reads per Cell	Median Genes per Cell
101,825	1,151

Sequencing

Number of Reads	185,323,111
Valid Barcodes	98.3%
Sequencing Saturation	92.5%
Q30 Bases in Barcode	98.0%
Q30 Bases in RNA Read	80.0%
Q30 Bases in Sample Index	94.7%
Q30 Bases in UMI	98.2%

Cells



Estimated Number of Cells	1,820
Fraction Reads in Cells	92.4%
Mean Reads per Cell	101,825
Median Genes per Cell	1,151
Total Genes Detected	14,589
Median UMI Counts per Cell	3,429

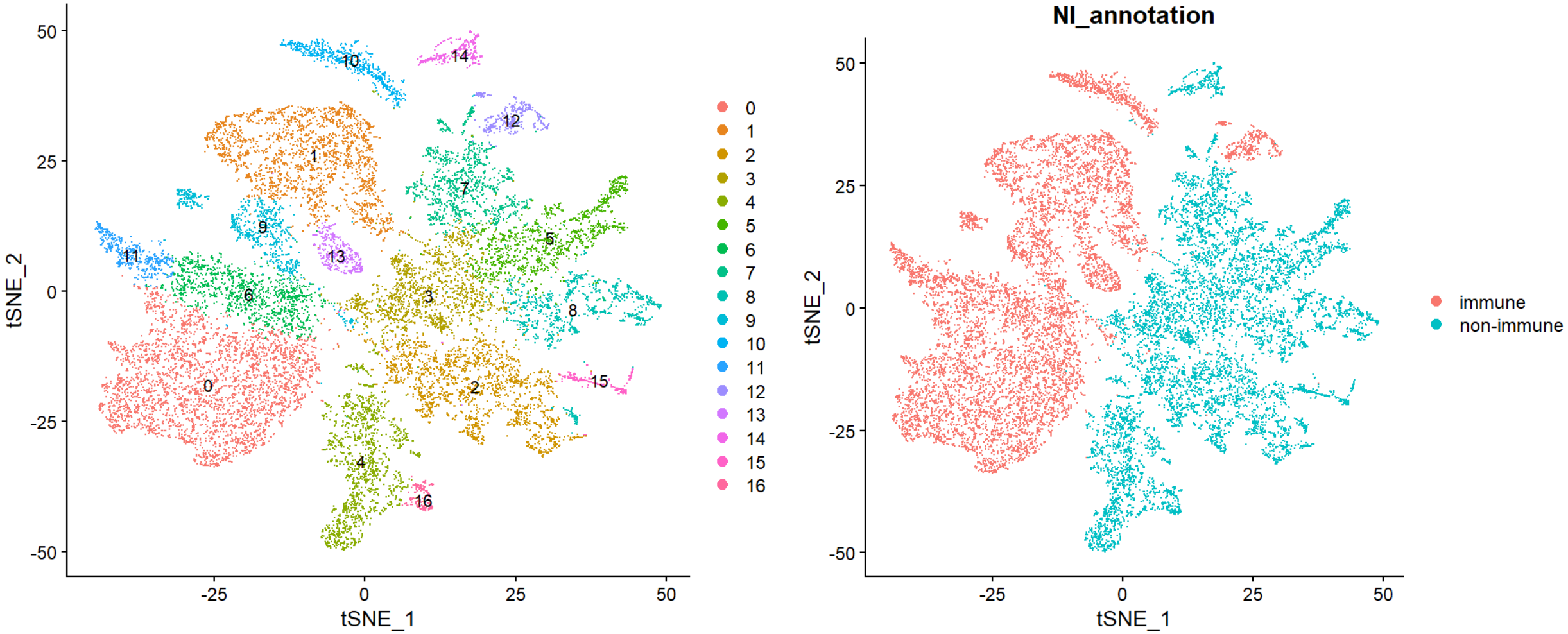
Normalization

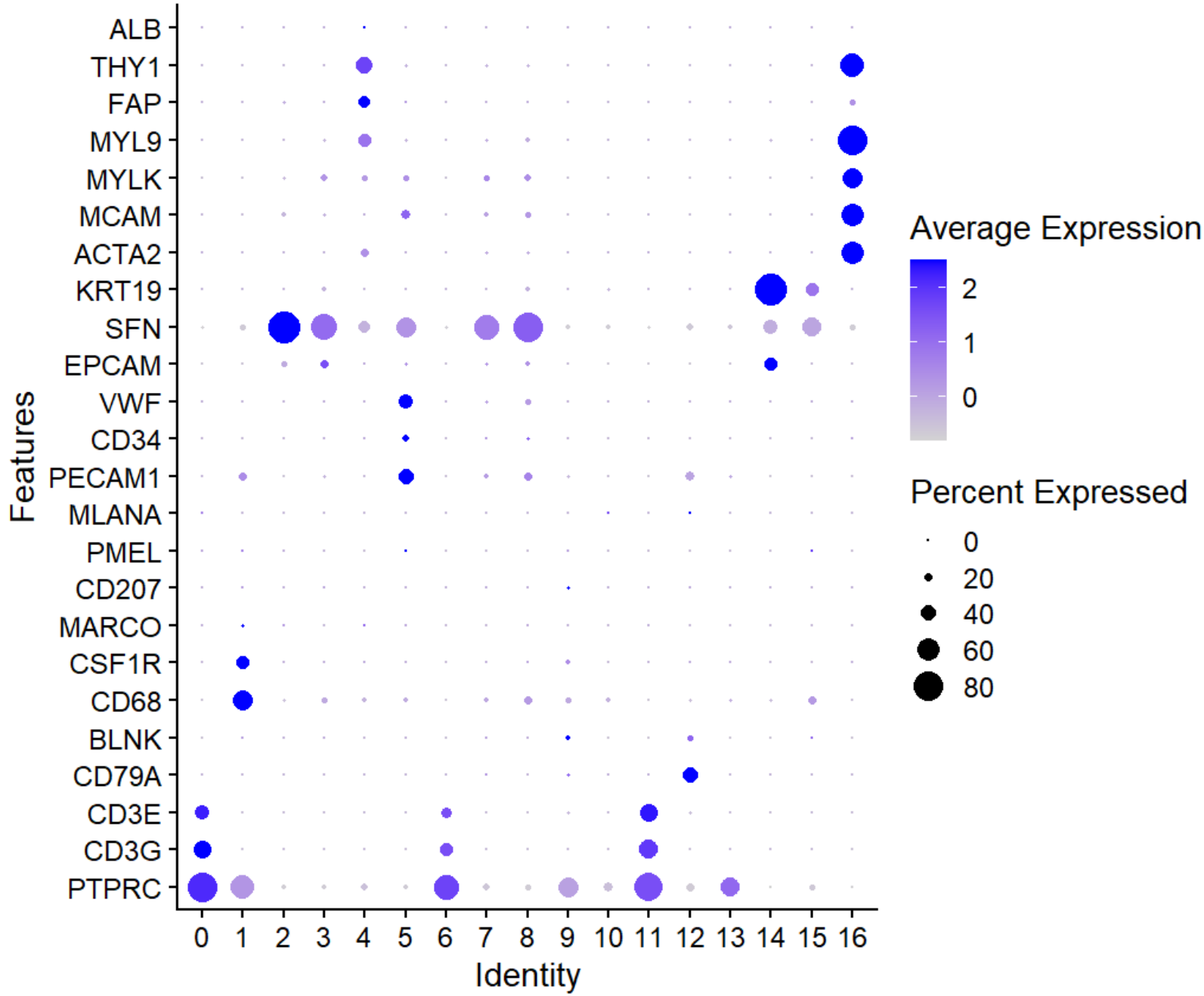
- Addressing limitations presented by low input material and various forms of bias or noise present in the sequencing process.
- Some rely on spike-in genes, molecules added in known quantities to serve as a basis for a normalization model.
- By default, Seurat implements a global-scaling normalization method “LogNormalize” that normalizes the gene expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result.

Dimension reduction, clustering and cell typing

- Highly variable genes (HVG) selection.
- Principle component analysis (PCA)
- tSNE/UMAP.
- Differential expressed genes-based cell type markers matching

HVG.method=mvp (523)
n.pcs=15
Resoulation=0.5




ALB
THY1
FAP
MYL9
MYLK
MCAM
ACTA2
KRT19
SFN
EPCAM
VWF
CD34
PECAM1
MLANA
PMEL
CD207
MARCO
CSF1R
CD68
BLNK
CD79A
CD3E
CD3G
PTPRC
Features
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Identity
Average Expression
2
1
0
Percent Expressed
0
20
40
60
80

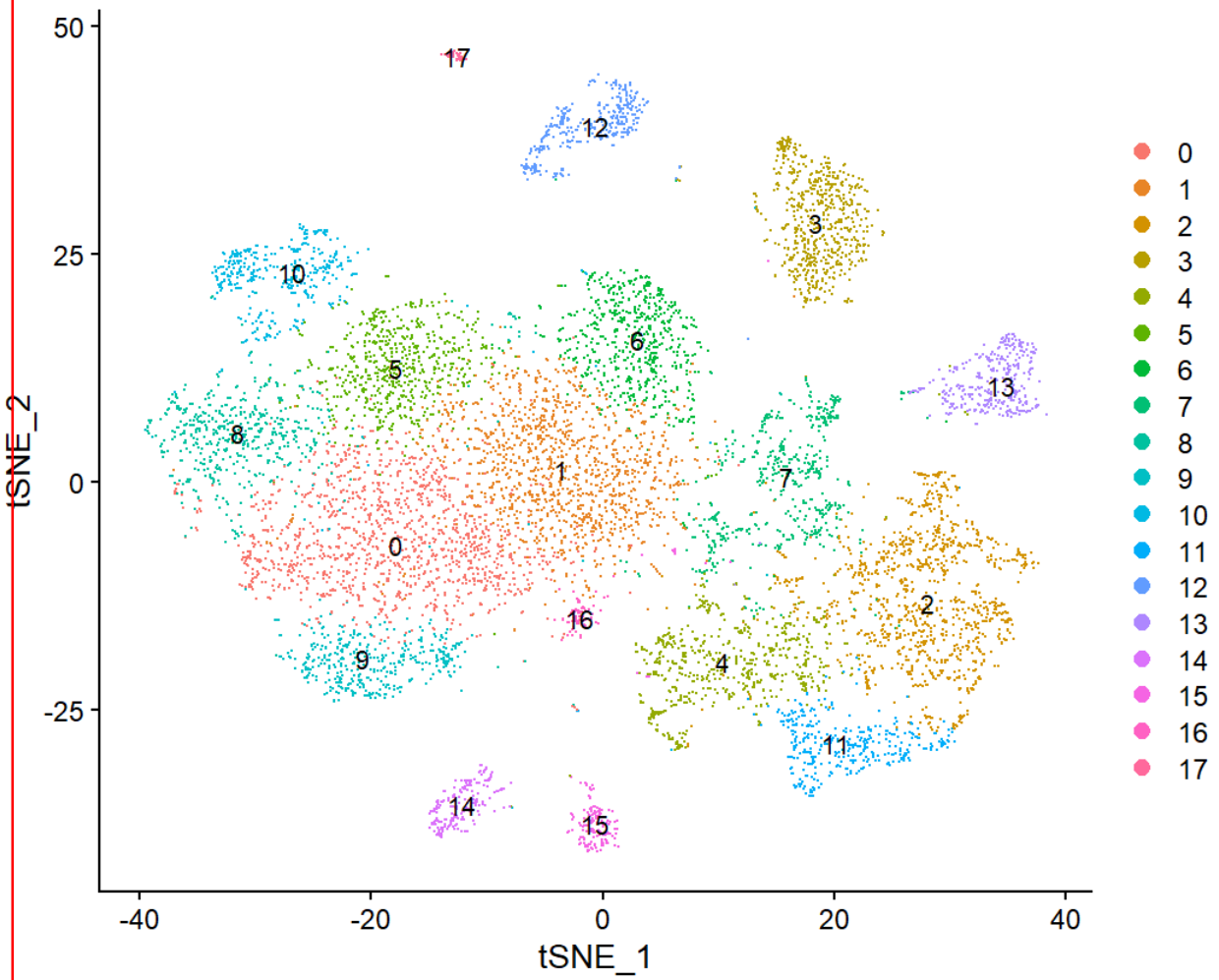
ALL cells. p_val_adj<0.05 & avg_log2FC>0.5

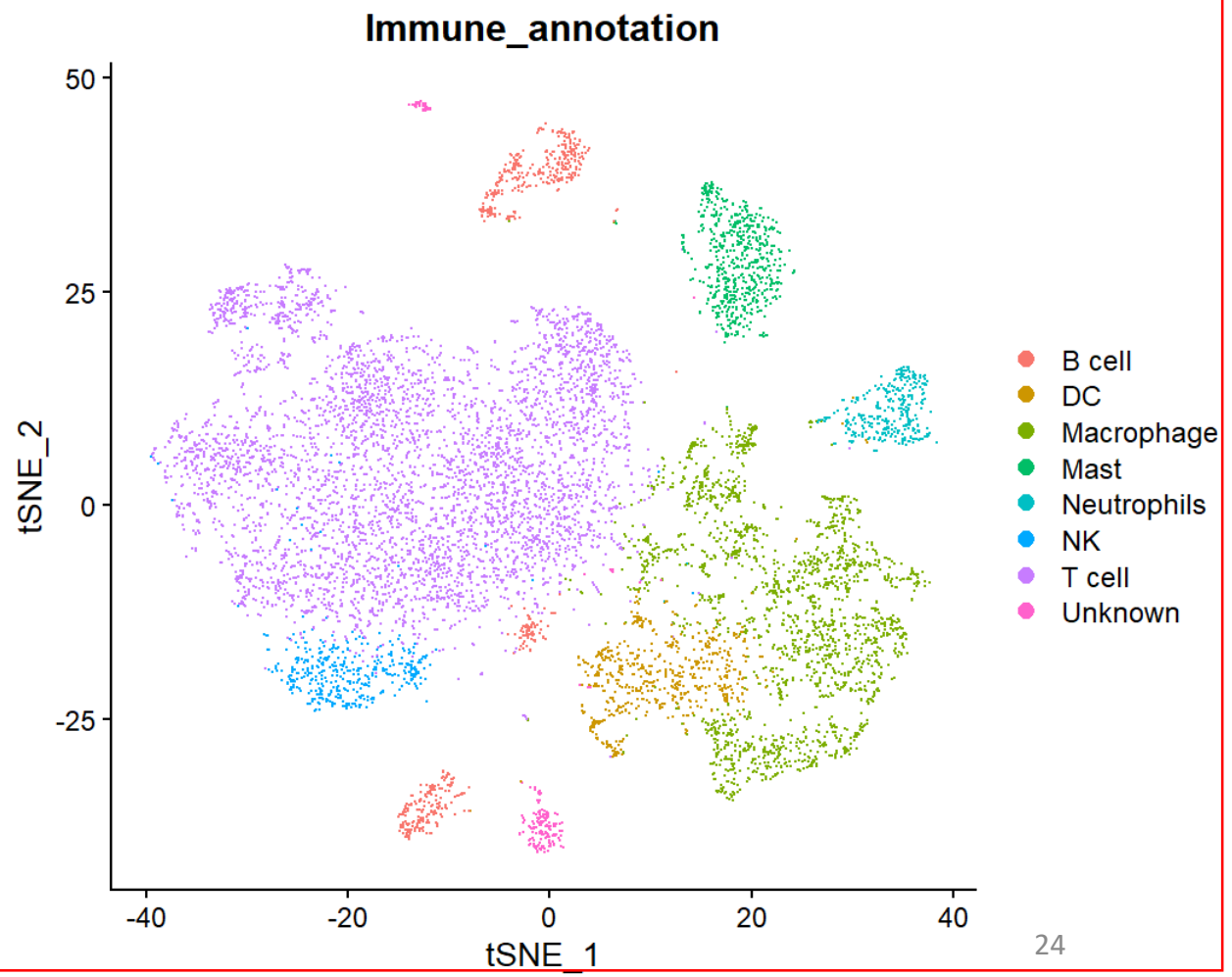
p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene	CellType
0.00E+00	2.314801	0.459	0.06	0.00E+00	0	CD3G	Immune Cells
0.00E+00	1.946833	0.778	0.284	0.00E+00	0	PTPRC	Immune Cells
0.00E+00	1.942074	0.357	0.049	0.00E+00	0	CD3E	Immune Cells
0.00E+00	2.311227	0.531	0.068	0.00E+00	1	CD68	Immune Cells
0.00E+00	1.731657	0.308	0.012	0.00E+00	1	CSF1R	Immune Cells
0.00E+00	2.322972	0.861	0.281	0.00E+00	2	SFN	Epithelial
1.80E-203	1.005902	0.701	0.309	6.04E-199	3	SFN	Epithelial
2.45E-163	0.777766	0.187	0.035	8.23E-159	3	EPCAM	Epithelial
0.00E+00	2.346925	0.41	0.018	0.00E+00	4	THY1	Fibroblasts
0.00E+00	1.792353	0.312	0.033	0.00E+00	4	MYL9	Fibroblasts
0.00E+00	1.658452	0.282	0.014	0.00E+00	4	FAP	Fibroblasts
7.60E-220	1.265682	0.186	0.025	2.55E-215	4	ACTA2	Fibroblasts
0.00E+00	2.651813	0.37	0.017	0.00E+00	5	VWF	Endothelial
0.00E+00	2.089985	0.382	0.065	0.00E+00	5	PECAM1	Endothelial
0.00E+00	1.127696	0.156	0.007	0.00E+00	5	CD34	Endothelial
2.94E-185	1.378154	0.222	0.039	9.88E-181	5	MCAM	Fibroblasts
4.43E-28	0.543788	0.126	0.053	1.49E-23	5	MYLK	Fibroblasts
2.63E-151	1.271469	0.665	0.356	8.83E-147	6	PTPRC	Immune Cells
8.57E-79	1.113987	0.307	0.122	2.87E-74	6	CD3G	Immune Cells
1.38E-64	1.090453	0.246	0.096	4.63E-60	6	CD3E	Immune Cells
3.50E-112	0.697486	0.67	0.319	1.18E-107	7	SFN	Epithelial
2.18E-28	0.68877	0.128	0.053	7.30E-24	7	MYLK	Fibroblasts
3.59E-205	1.132724	0.791	0.316	1.20E-200	8	SFN	Epithelial
3.28E-34	0.509073	0.194	0.08	1.10E-29	8	PECAM1	Endothelial
3.90E-33	0.558268	0.133	0.047	1.31E-28	8	MCAM	Fibroblasts
4.27E-31	0.574228	0.141	0.053	1.43E-26	8	MYLK	Fibroblasts
4.94E-105	1.218137	0.057	0.002	1.66E-100	9	CD207	Immune Cells
9.29E-59	0.567891	0.071	0.009	3.12E-54	9	BLNK	Immune Cells

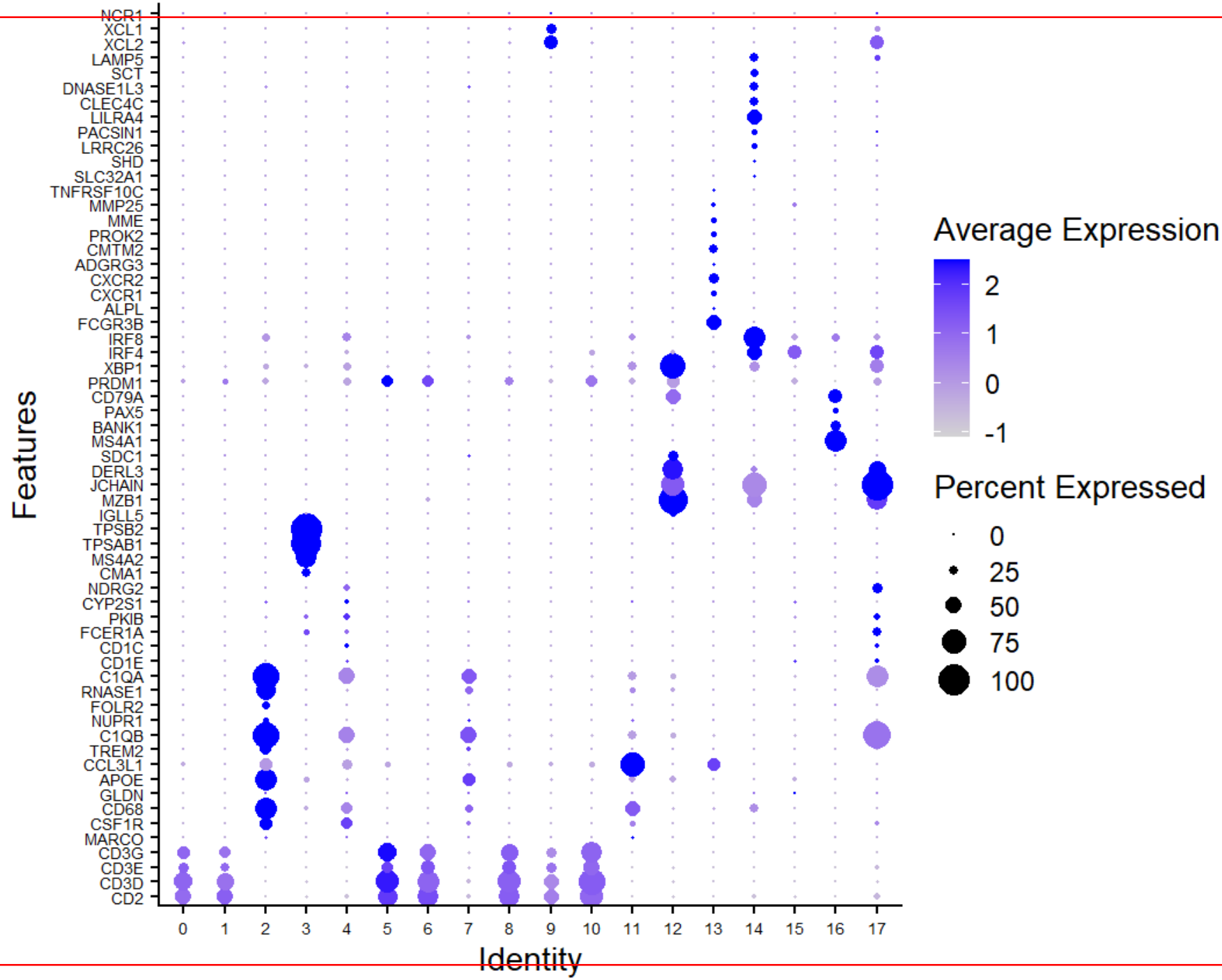
p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene	CellTypeA
1.68E-135	1.42925	0.459	0.097	5.63E-131	11	CD3E	Immune Cells
4.10E-119	1.244529	0.506	0.124	1.37E-114	11	CD3G	Immune Cells
3.24E-81	1.082032	0.776	0.365	1.09E-76	11	PTPRC	Immune Cells
0.00E+00	1.615516	0.4	0.006	0.00E+00	12	CD79A	Immune Cells
2.03E-10	0.774802	0.478	0.374	6.82E-06	13	PTPRC	Immune Cells
0.00E+00	4.45157	0.846	0.024	0.00E+00	14	KRT19	Epithelial
3.18E-109	1.174967	0.308	0.043	1.07E-104	14	EPCAM	Epithelial
1.46E-110	2.506583	0.312	0.034	4.91E-106	15	KRT19	Epithelial
0.00E+00	3.012302	0.792	0.05	0.00E+00	16	MYL9	Fibroblasts
1.08E-288	3.239483	0.585	0.034	3.63E-284	16	ACTA2	Fibroblasts
3.45E-247	2.67985	0.629	0.046	1.16E-242	16	THY1	Fibroblasts
2.04E-218	2.266783	0.597	0.047	6.83E-214	16	MCAM	Fibroblasts
9.55E-147	1.94026	0.528	0.054	3.20E-142	16	MYLK	Fibroblasts

cluster	判定	cluster	判定
0	Immune Cells	9	Immune Cells
1	Immune Cells	10	Immune Cells
2	Non-immune Cells	11	Immune Cells
3	Non-immune Cells	12	Immune Cells
4	Non-immune Cells	13	Immune Cells
5	Non-immune Cells	14	Non-immune Cells
6	Immune Cells	15	Non-immune Cells
7	Non-immune Cells	16	Non-immune Cells
8	Non-immune Cells		

HVG.method=mvp
n.pcs=30
Resoulation=0.8






NCR1
XCL1
XCL2
LAMP5
SCT
DNASE1L3
CLEC4C
LILRA4
PACSIN1
LRRC26
SHD
SLC32A1
TNFRSF10C
MMP25
MME
PROK2
CMTM2
ADGRG3
CXCR2
CXCR1
ALPL
FCGR3B
IRF8
IRF4
XBP1
PRDM1
CD79A
PAX5
BANK1
MS4A1
SDC1
DERL3
JCHAIN
MZB1
IGLL5
TPSB2
TPSAB1
MS4A2
CMA1
NDRG2
CYP2S1
PKIB
FCER1A
CD1C
CD1E
C1QA
RNASE1
FOLR2
NUPR1
C1QB
TREM2
CCL3L1
APOE
GLDN
CD68
CSF1R
MARCO
CD3G
CD3E
CD3D
CD2
Features
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Identity
Average Expression
2
1
0
-1
Percent Expressed
0
25
50
75
100

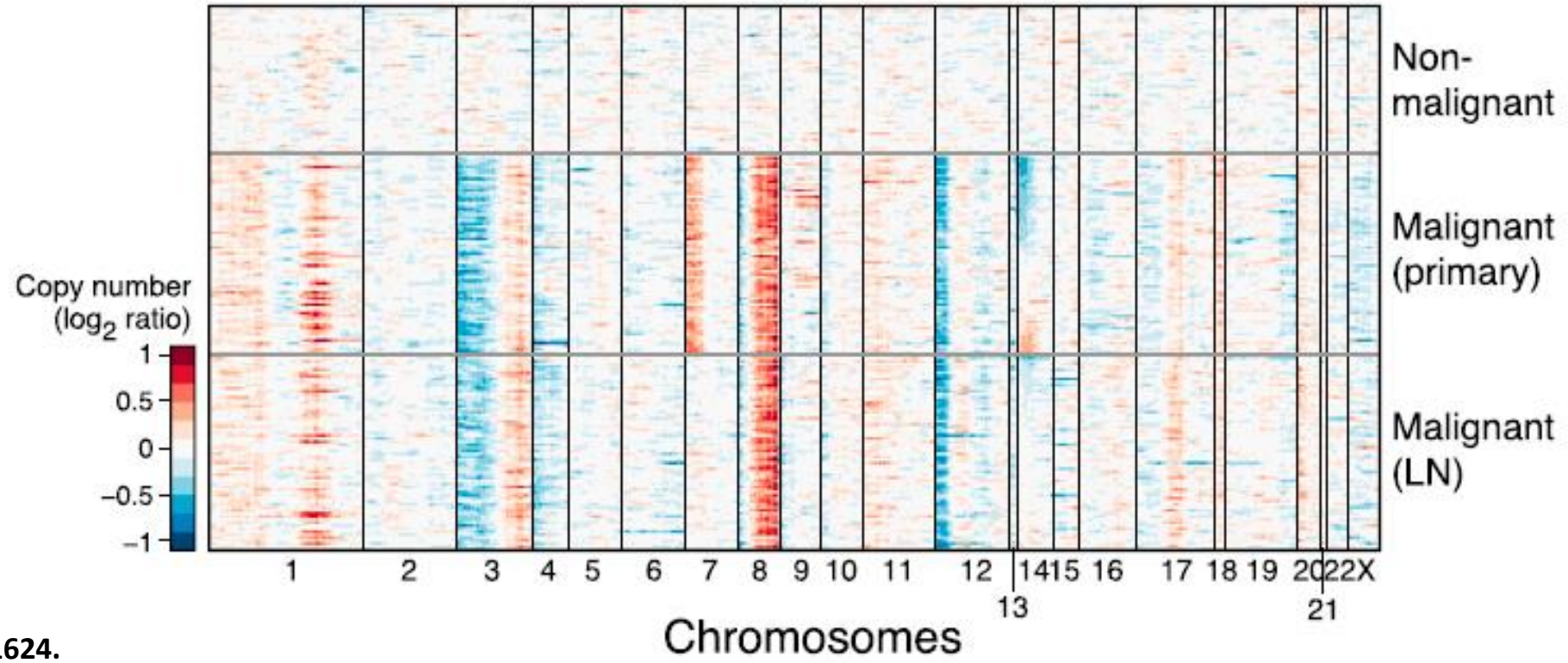
Immune cells. p_val_adj<0.05 & avg_log2FC>0.5

p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene	CellType
5.44E-84	0.900903	0.579	0.317	1.82E-79	0	CD3D	Tcell
1.78E-61	0.686999	0.75	0.614	5.97E-57	0	PTPRC	Immune Cells
4.66E-45	0.715665	0.48	0.293	1.56E-40	0	CD2	Tcell
7.85E-42	0.830841	0.377	0.219	2.63E-37	0	CD3G	Tcell
1.54E-28	0.771052	0.295	0.176	5.15E-24	0	CD3E	Tcell
8.87E-53	0.907278	0.484	0.294	2.98E-48	1	CD2	Tcell
1.54E-46	0.653406	0.532	0.326	5.15E-42	1	CD3D	Tcell
1.17E-28	0.554761	0.676	0.626	3.94E-24	1	PTPRC	Immune Cells
1.04E-20	0.582989	0.338	0.226	3.50E-16	1	CD3G	Tcell
3.39E-16	0.644849	0.268	0.181	1.14E-11	1	CD3E	Tcell
0.00E+00	3.740999	0.852	0.11	0.00E+00	2	C1QA	Macrophage
0.00E+00	3.689785	0.682	0.079	0.00E+00	2	APOE	Macrophage
0.00E+00	3.566536	0.849	0.118	0.00E+00	2	C1QB	Macrophage
0.00E+00	3.477929	0.624	0.037	0.00E+00	2	RNASE1	Macrophage
0.00E+00	2.455359	0.682	0.081	0.00E+00	2	CD68	Macrophage
0.00E+00	1.776184	0.393	0.04	0.00E+00	2	CSF1R	Macrophage
0.00E+00	1.707771	0.322	0.013	0.00E+00	2	TREM2	Macrophage
0.00E+00	1.654535	0.233	0.004	0.00E+00	2	FOLR2	Macrophage
2.63E-122	0.883293	0.132	0.012	8.82E-118	2	NUPR1	Macrophage
6.40E-58	0.531884	0.206	0.063	2.15E-53	2	IRF8	Bcell_Mem
0.00E+00	7.135617	0.983	0.008	0.00E+00	3	TPSB2	Mast
0.00E+00	6.141054	0.944	0.004	0.00E+00	3	TPSAB1	Mast
0.00E+00	3.946708	0.636	0.002	0.00E+00	3	MS4A2	Mast
0.00E+00	2.66538	0.274	0	0.00E+00	3	CMA1	Mast
2.38E-126	1.158152	0.136	0.009	7.98E-122	3	FCER1A	Dendritic
2.95E-185	1.582279	0.105	0.001	9.91E-181	4	CD207	Dendritic
1.43E-180	0.818816	0.12	0.002	4.79E-176	4	CD1C	Dendritic
1.14E-150	1.001361	0.188	0.015	3.81E-146	4	PKIB	Dendritic
9.38E-140	0.851203	0.169	0.013	3.15E-135	4	NDRG2	Dendritic
1.74E-119	1.017558	0.329	0.063	5.83E-115	4	CSF1R	Macrophage
1.81E-71	0.563279	0.508	0.18	6.08E-67	4	C1QB	Macrophage
2.78E-70	0.989063	0.267	0.066	9.31E-66	4	IRF8	Bcell_Mem

p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene	CellType
1.55E-129	1.637698	0.73	0.327	5.20E-125	5	CD3D	Tcell
1.46E-101	1.617986	0.555	0.218	4.90E-97	5	CD3G	Tcell
8.80E-74	1.27753	0.607	0.299	2.95E-69	5	CD2	Tcell
7.36E-34	1.367804	0.332	0.167	2.47E-29	5	PRDM1	Bcell_PB
2.26E-30	1.048553	0.352	0.182	7.59E-26	5	CD3E	Tcell
1.58E-27	0.656079	0.73	0.626	5.29E-23	5	PTPRC	Immune Cells
1.12E-73	1.074226	0.654	0.296	3.76E-69	6	CD2	Tcell
2.24E-57	0.822928	0.684	0.33	7.51E-53	6	CD3D	Tcell
9.34E-45	0.839794	0.422	0.177	3.13E-40	6	CD3E	Tcell
2.11E-41	0.718043	0.484	0.224	7.07E-37	6	CD3G	Tcell
6.20E-27	0.864994	0.334	0.168	2.08E-22	6	PRDM1	Bcell_PB
9.81E-68	1.577921	0.476	0.184	3.29E-63	7	C1QB	Macrophage
7.02E-60	1.444837	0.445	0.179	2.35E-55	7	C1QA	Macrophage
1.28E-56	1.862498	0.369	0.134	4.30E-52	7	APOE	Macrophage
7.91E-16	1.042763	0.2	0.098	2.65E-11	7	RNASE1	Macrophage
9.25E-06	0.917749	0.2	0.146	3.10E-01	7	CD68	Macrophage
1.07E-69	0.880441	0.737	0.329	3.60E-65	8	CD3D	Tcell
2.55E-56	0.84848	0.647	0.299	8.54E-52	8	CD2	Tcell
1.29E-49	0.864453	0.522	0.223	4.31E-45	8	CD3G	Tcell
2.35E-46	0.710685	0.866	0.618	7.88E-42	8	PTPRC	Immune Cells
2.57E-42	0.837578	0.428	0.178	8.61E-38	8	CD3E	Tcell
0.00E+00	3.148858	0.404	0.024	0.00E+00	9	XCL2	NK
0.00E+00	2.958628	0.302	0.01	0.00E+00	9	XCL1	NK
2.46E-68	0.907235	0.825	0.332	8.24E-64	10	CD3D	Tcell
3.47E-58	0.749073	0.64	0.223	1.16E-53	10	CD3G	Tcell
3.40E-52	0.6937	0.741	0.301	1.14E-47	10	CD2	Tcell
2.87E-40	0.585098	0.492	0.18	9.63E-36	10	CD3E	Tcell
7.50E-279	3.716229	0.763	0.13	2.52E-274	11	CCL3L1	Macrophage
1.69E-60	1.020938	0.451	0.136	5.65E-56	11	CD68	Macrophage
1.71E-17	0.515768	0.266	0.116	5.73E-13	11	XBP1	Bcell_PB
3.58E-15	0.747361	0.185	0.075	1.20E-10	11	IRF8	Bcell_Mem

Identification of cancer cells

- InferCNV (Infer Copy Number Variation)
  - identify malignant cells



Cell 2017 (171), 1611-1624.

Quantitative analysis

(1) Cell proportion analysis

(2) Dot plot

(3) Violin plot

(4) Volcano plot (differential expression gene analysis)

Pathway analysis and functional annotation

- GSVA (gene set variation analysis)
- GSEA/ssGSEA

GSEA to GSVA

Hänzelmann *et al. BMC Bioinformatics* 2013, **14**:7
http://www.biomedcentral.com/1471-2105/14/7

BMC Bioinformatics

SOFTWARE Open Access

GSVA: gene set variation analysis for microarray and RNA-Seq data

Sonja Hänzelmann[1,2], Robert Castelo[1,2*] and Justin Guinney[3*]

Abstract

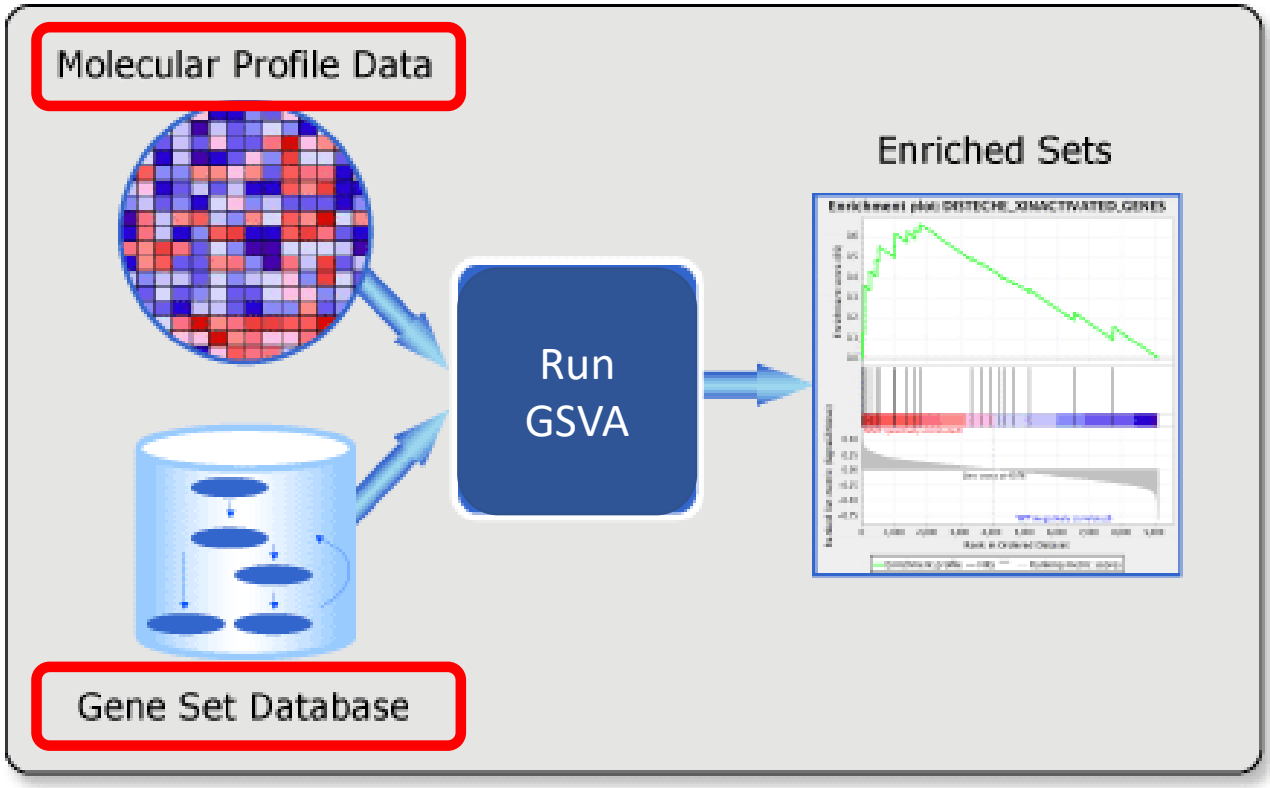
Background: Gene set enrichment (GSE) analysis is a popular framework for condensing information from gene expression profiles into a pathway or signature summary. The strengths of this approach over single gene analysis include noise and dimension reduction, as well as greater biological interpretability. As molecular profiling experiments move beyond simple case-control studies, robust and flexible GSE methodologies are needed that can model pathway activity within highly heterogeneous data sets.

Results: To address this challenge, we introduce Gene Set Variation Analysis (GSVA), a GSE method that estimates variation of pathway activity over a sample population in an unsupervised manner. We demonstrate the robustness of GSVA in a comparison with current state of the art sample-wise enrichment methods. Further, we provide examples of its utility in differential pathway activity and survival analysis. Lastly, we show how GSVA works analogously with data from both microarray and RNA-seq experiments.

Conclusions: GSVA provides increased power to detect subtle pathway activity changes over a sample population in comparison to corresponding methods. While GSE methods are generally regarded as end points of a bioinformatic analysis, GSVA constitutes a starting point to build pathway-centric models of biology. Moreover, GSVA contributes to the current need of GSE methods for RNA-seq data. GSVA is an open source software package for R which forms part of the Bioconductor project and can be downloaded at http://www.bioconductor.org.

GSVA input

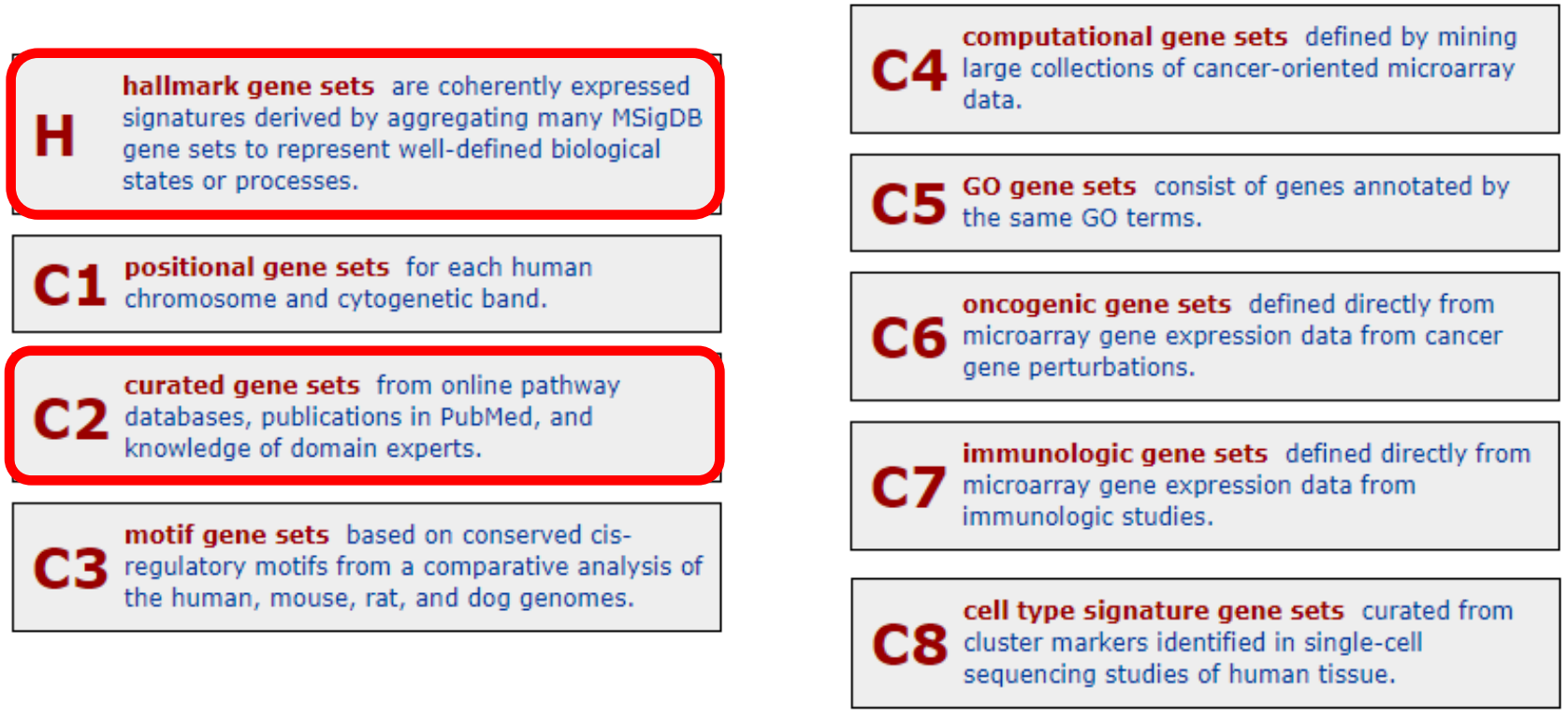
- Molecular profile data (user): Microarray, RNA-seq, …
- Gene set database (reference database): Molecular signatures database



From GSEA website

Reference database

- The Molecular signatures database (MSigDB) gene sets are divided into 9 major collections.



From GSEA website

GSVA algorithm

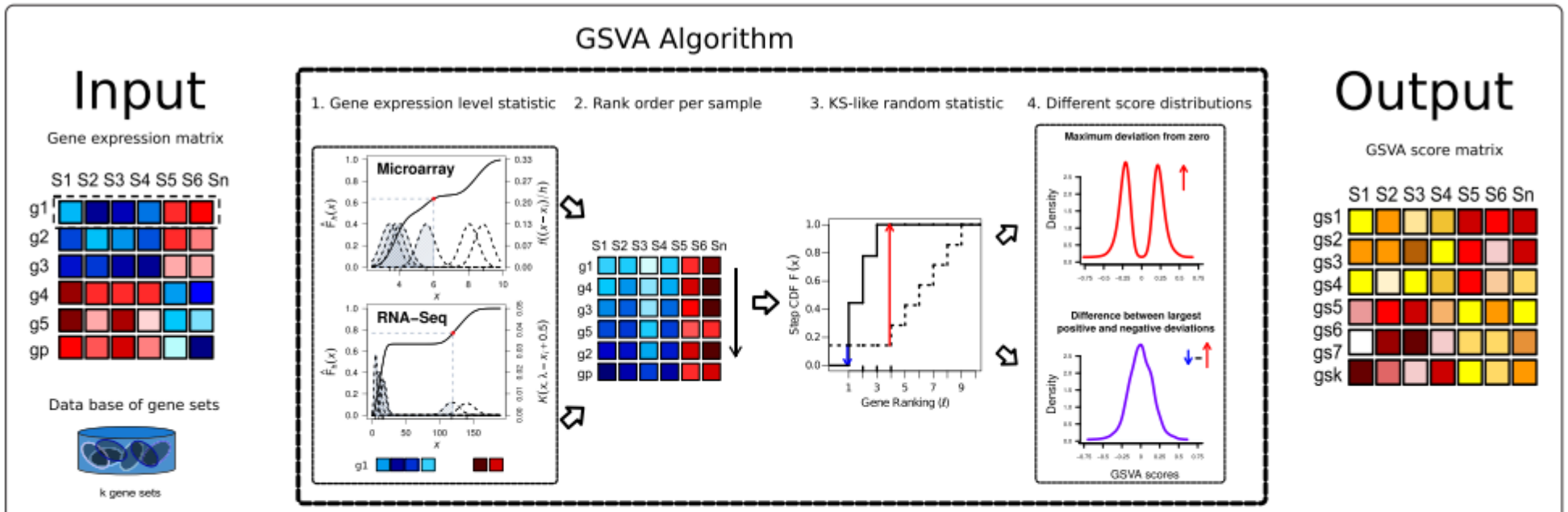


Figure 1 GSVA methods outline. The input for the GSVA algorithm are a gene expression matrix in the form of log2 microarray expression values or RNA-seq counts and a database of gene sets. 1. Kernel estimation of the cumulative density function (kcdf). The two plots show two simulated expression profiles mimicking 6 samples from microarray and RNA-seq data. The *x*-axis corresponds to expression values where each gene is lowly expressed in the four samples with lower values and highly expressed in the other two. The scale of the kcdf is on the left *y*-axis and the scale of the Gaussian and Poisson kernels is on the right *y*-axis. 2. The expression-level statistic is rank ordered for each sample. 3. For every gene set, the Kolmogorov-Smirnov-like rank statistic is calculated. The plot illustrates a gene set consisting of 3 genes out of a total number of 10 with the sample-wise calculation of genes inside and outside of the gene set. 4. The GSVA enrichment score is either the maximum deviation from zero (top) or the difference between the two sums (bottom). The two plots show two simulations of the resulting scores under the null hypothesis of no gene expression change (see main text). The output of the algorithm is a matrix containing pathway enrichment scores for each gene set and sample.

Trajectory (pseudotime) analysis

- Developmental biology: cell differentiation
- Immunology: cell differentiation
- Cancer research: origin of cancer cells

- Example

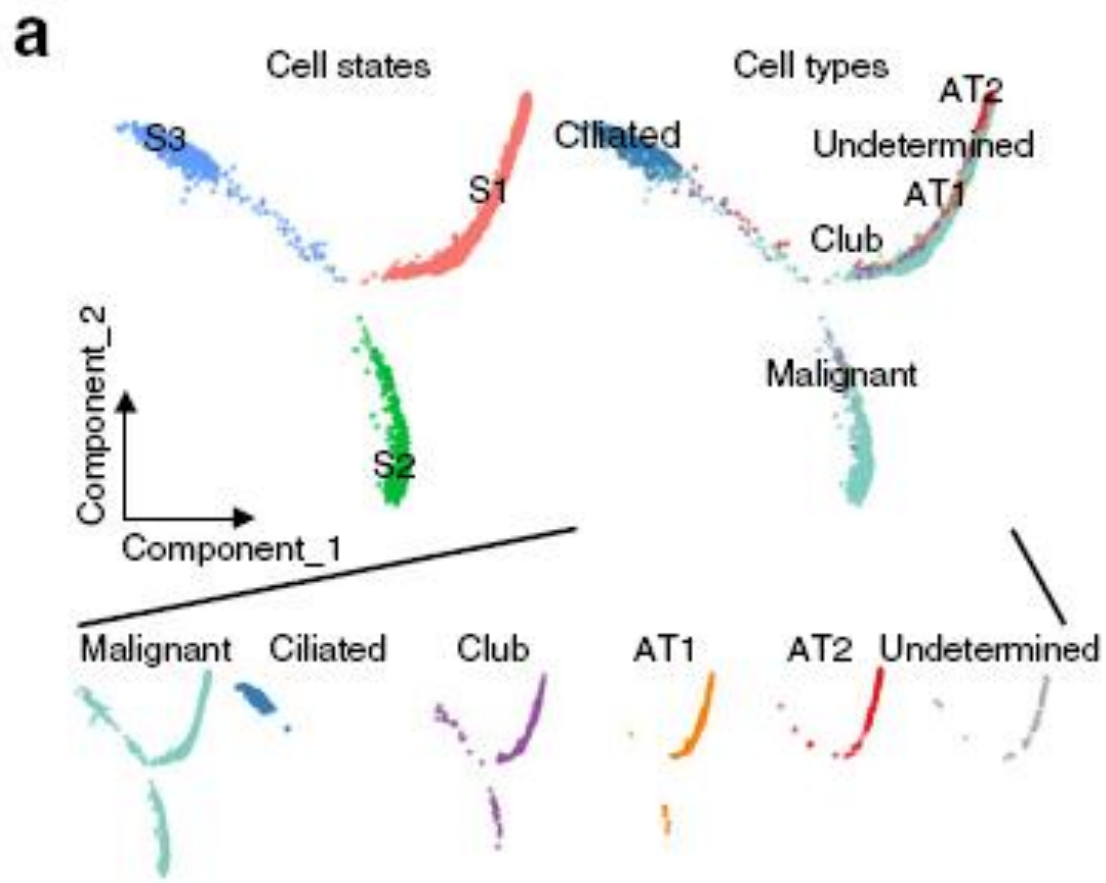


Fig. 2 Identification of novel cancer cell signature tS2 in LUAD and the connection to poor survival. a Unsupervised transcriptional trajectory of malignant and normal epithelial cells from Monocle (version 2), colored by cell states and subsets.
(Nature Comimunications, 2020)

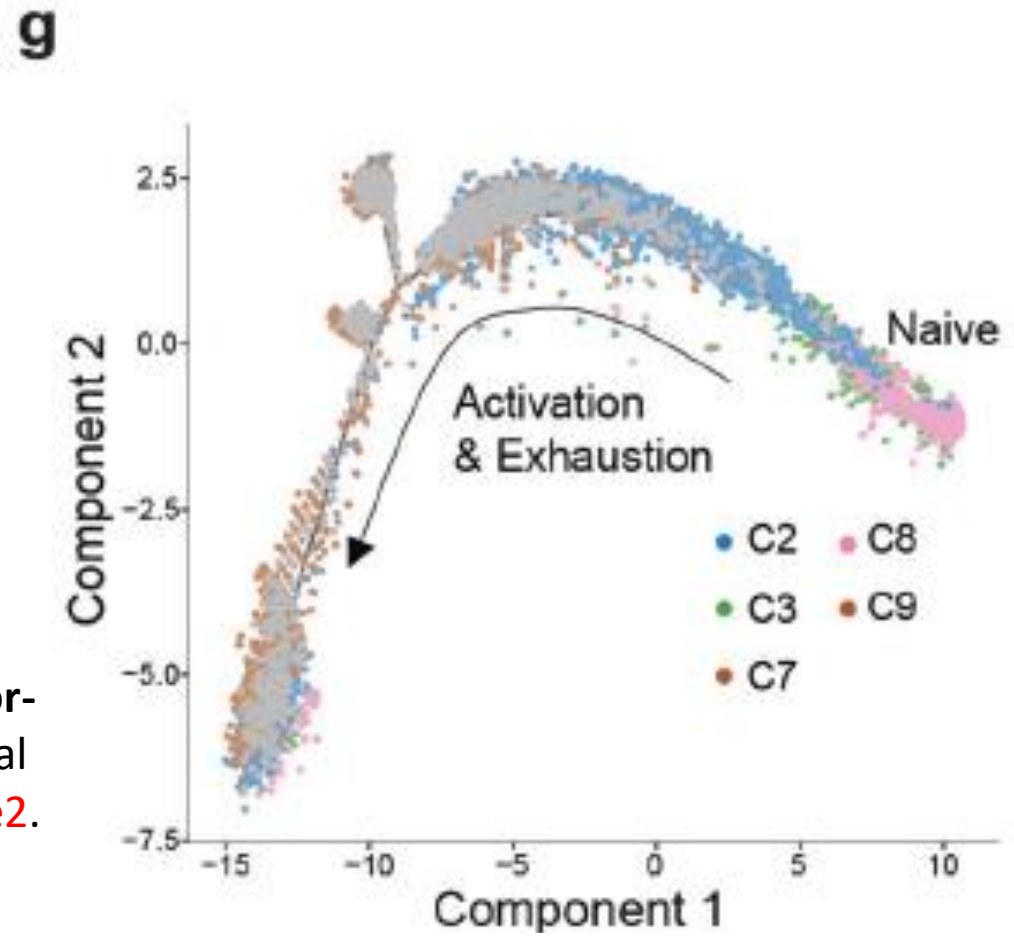


Fig. 3 Assessing the functional states of tumor-infiltrating T cells in NPC. g The developmental trajectory of CD8+ T cells inferred by Monocle2.
(Cell Research,2020)

Step 1: choosing genes that define progress
completely "unsupervised"
"semi-supervised"

Step 2: reducing the dimensionality of the data
dimensionality reduction
Reversed Graph Embedding

Step 3: ordering the cells in pseudotime
Monocle assumes that the trajectory has a tree structure, with one end of it the "root", and the others the "leaves".

A cell's pseudotime value is the distance it would have to travel to get back to the root.

Monocle doesn't know a priori which of the trajectory of the tree to call the "beginning.

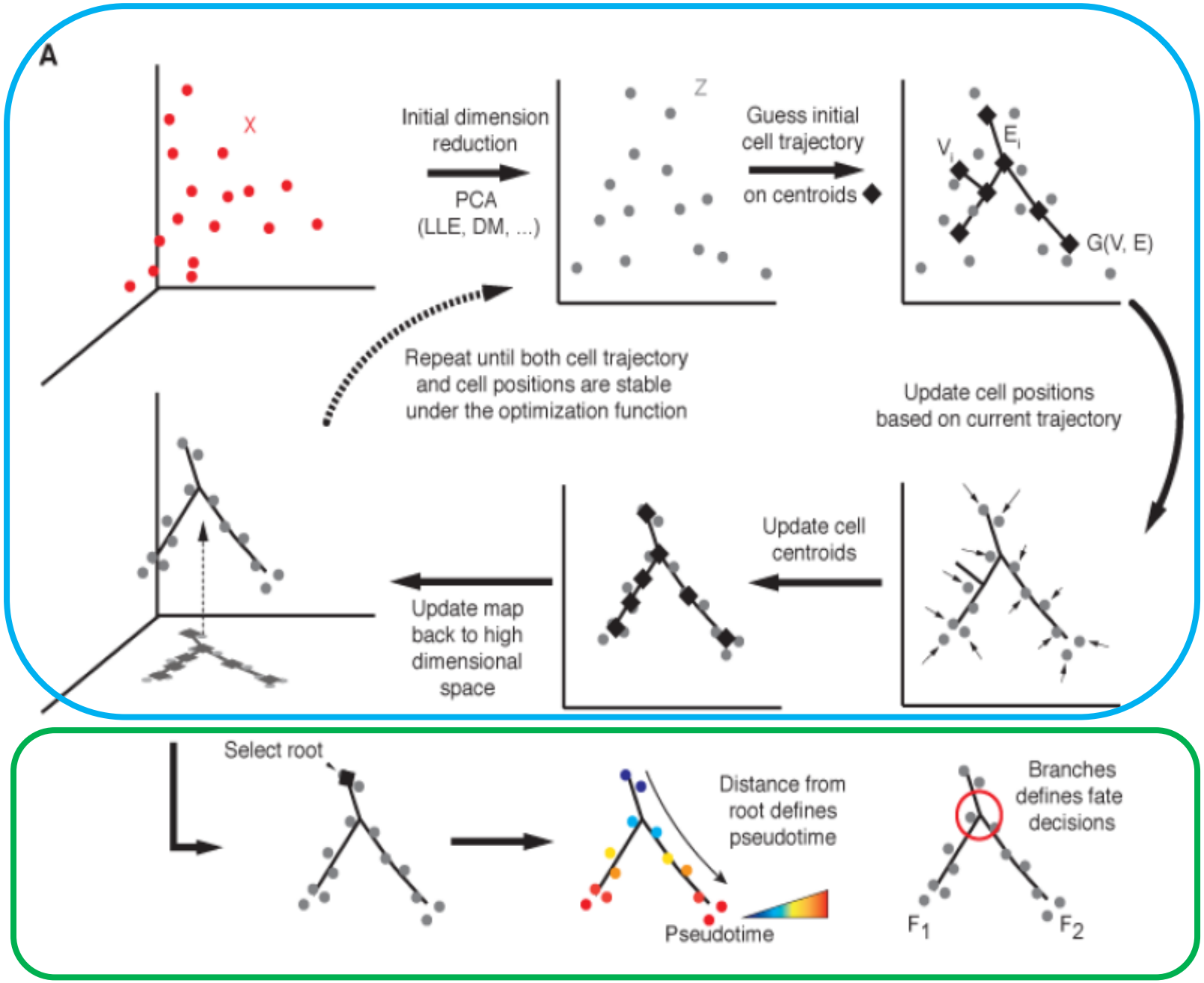
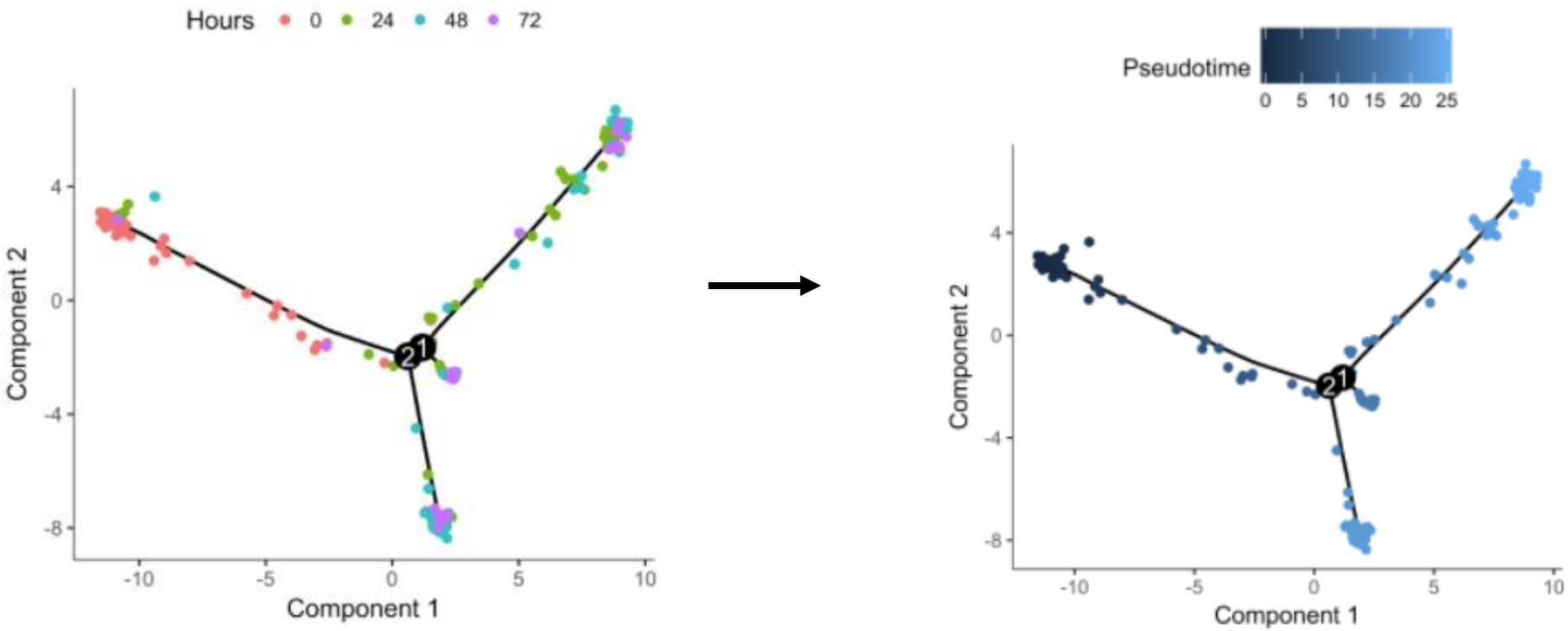


Figure 1 (Nat Methods. 2017)

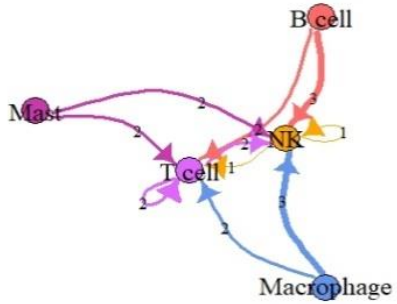
Step 3: ordering the cells in pseudotime

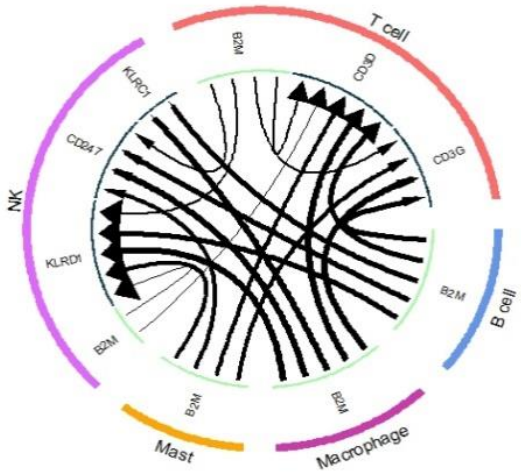


- http://cole-trapnell-lab.github.io/monocle-release/

Advanced analysis: cell-cell interaction

		CellChat	SingleCellSignalR	iTALK	CellPhoneDB
Database	Multisubunit structure	Y	N	N	Y
	Cofactors	Y	N	N	N
	Structured pathways	Y	N	N	N
	Number of interactions	2021	3251	2648	1396
	Species	mouse/human	mouse/human	human	human
Input data	Preprocessed data matrix	Y	Y	Y	Y
	Low-dimensional space	Y	N	N	N
	Multiple datasets	Y	N	Y	N
Model	Methodology	Mass Action Law + Statistical test	Regularized product + Thresholding	DEG analysis	Mean expression + Statistical test
	Cell proportion	Y	N	N	N
Systems analysis	Infer signaling roles of cells	Y	N	N	N
	Predict key incoming and outgoing signals	Y	N	N	N
	Predict signaling patterns	Y	N	N	N
	Classify signaling networks	Y	N	N	N
	Identify conserved vs. context-specific signaling	Y	N	Y	N
Visualization	Hierarchical plot	Y	N	N	N
	Circle plot	Y	N	Y	N
	Bubble plot	Y	N	N	Y
	Alluvial plot	Y	N	N	N
Language		R	R	R	Python





Cell and Gene type (a-e)

Gene type
- Cytokine/chemokine
- Immune checkpoint
- Growth factor
- Other
- Receptor

Relative expression level (b-c)

Ligand — High / Low

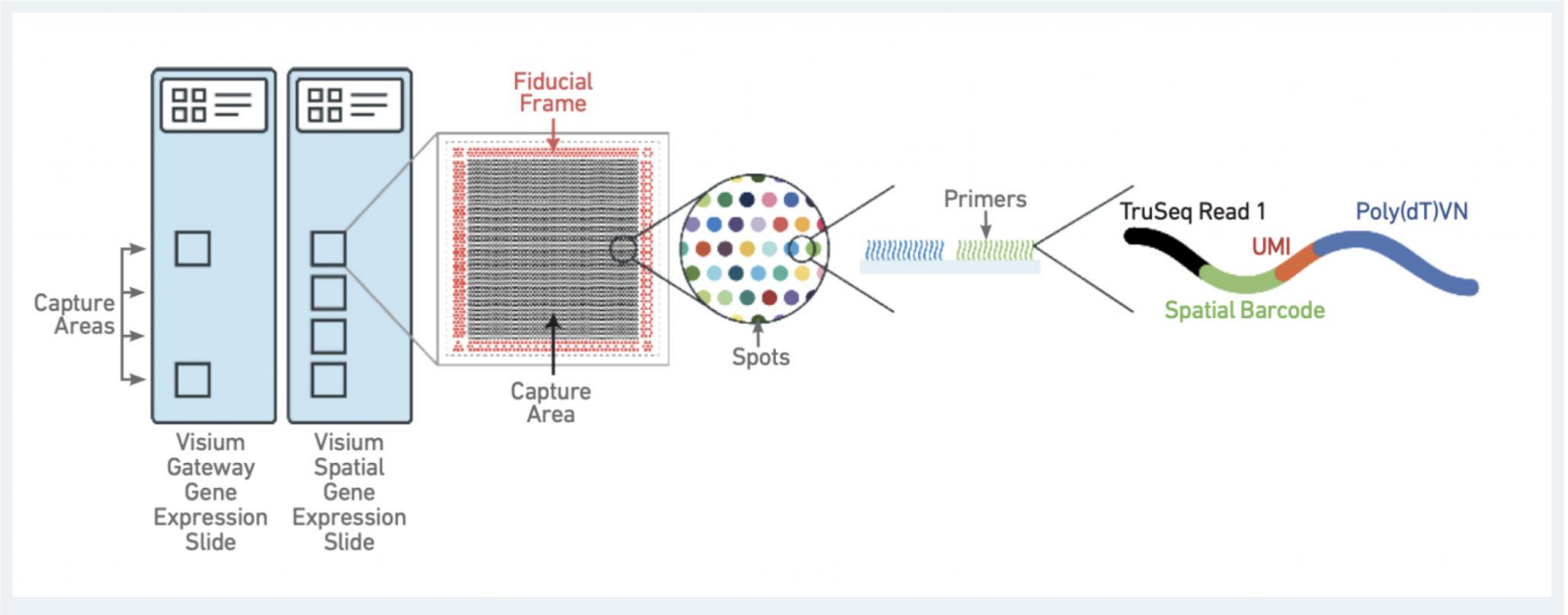
Receptor — High / Low

Nat Commun.2021; 12: 1088.

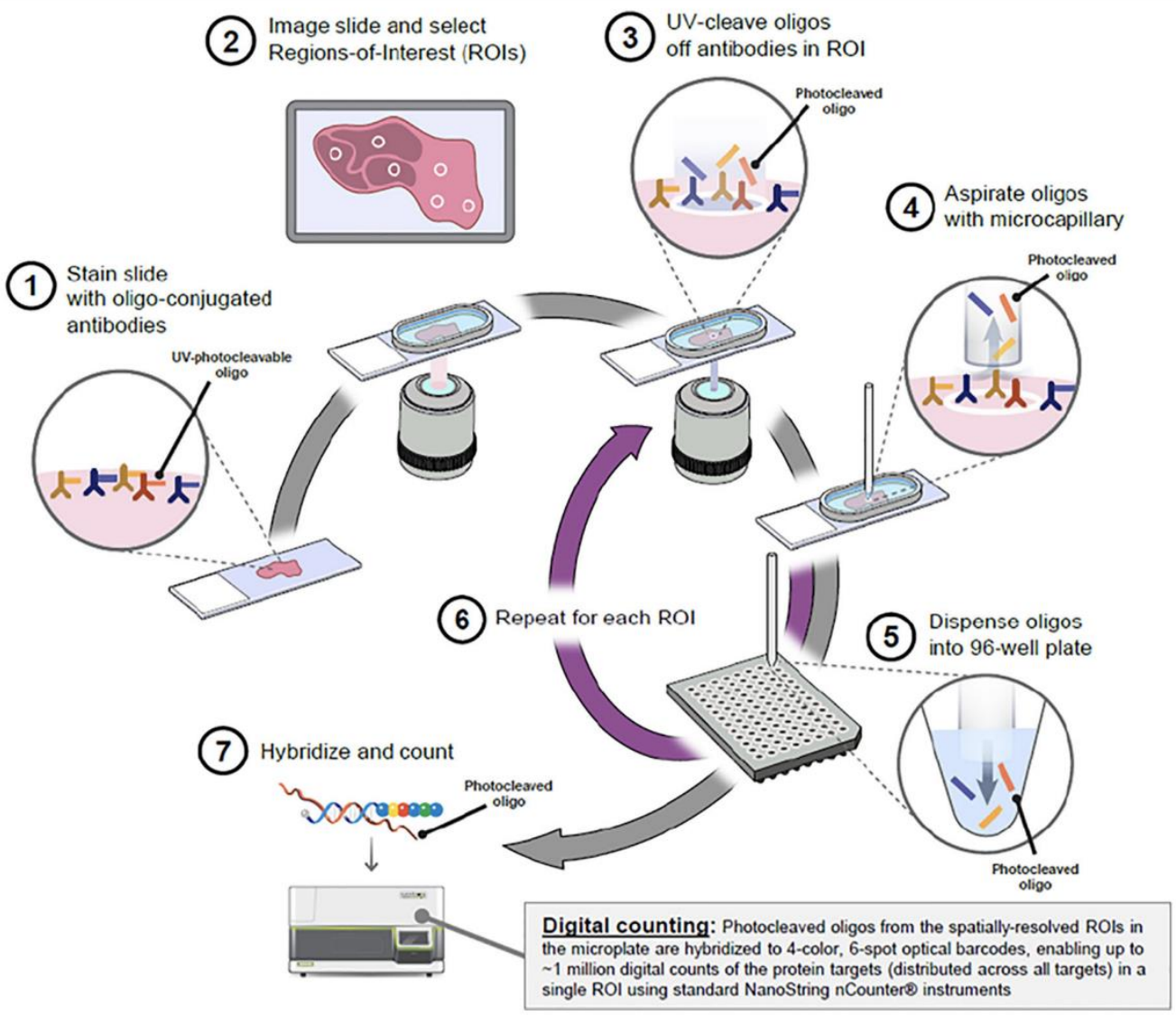
Challenges and limitations and Perspectives

- Preparation of single cell suspension
- Cell typing
- Spatial information
- Spatial genomics/proteomics

Visium: spatial transcriptome from 10x genomics



DSP (Digital Spatial Profiler) of NanoString



Thank you!