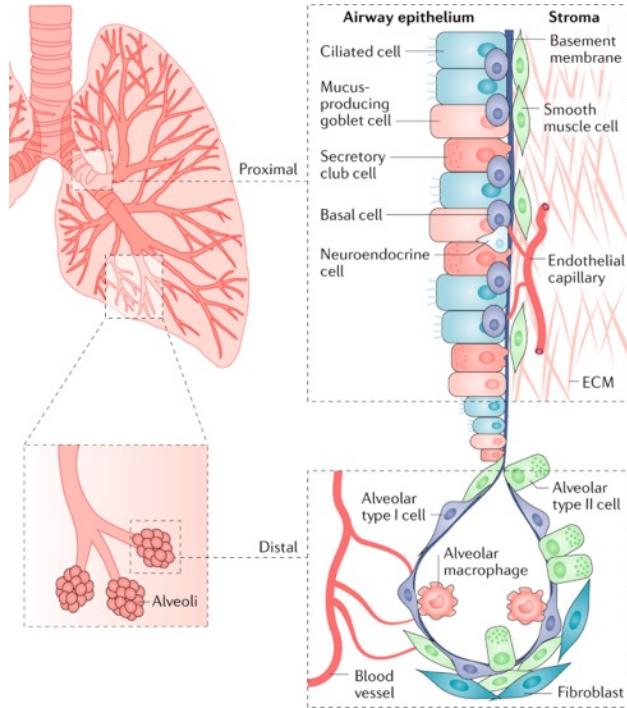


Why single-cell genomics

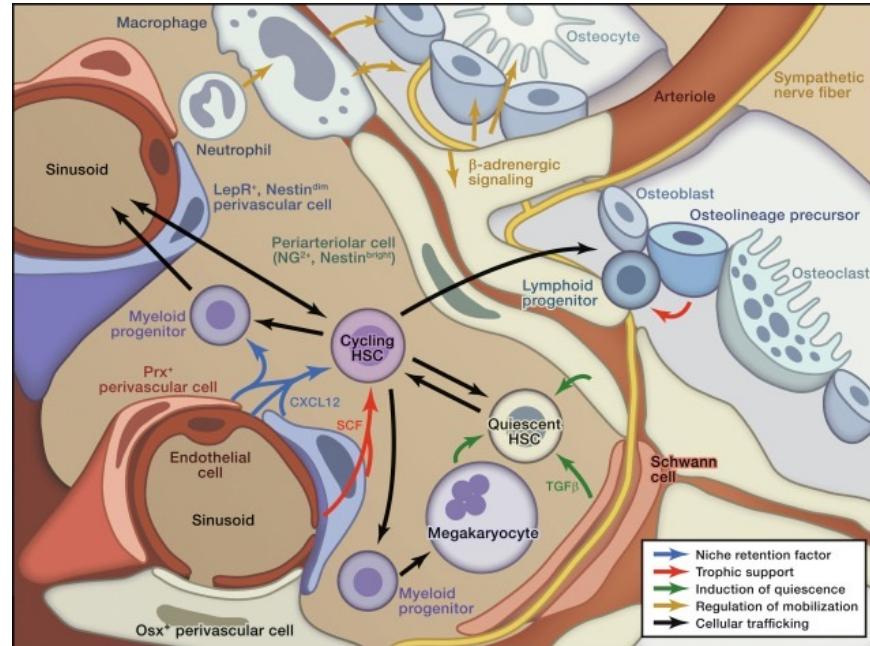
Heterogeneity in tissue homeostasis

Lung



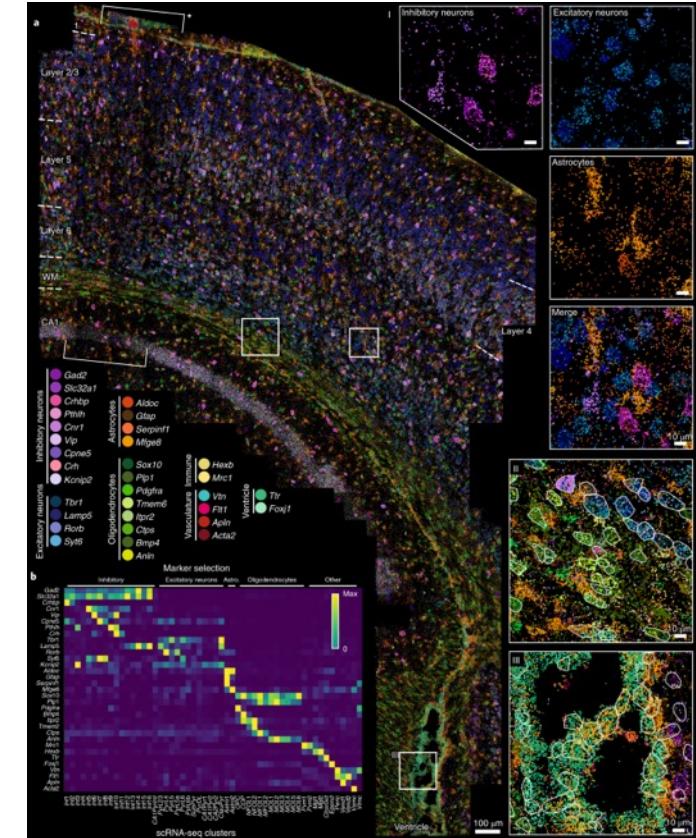
Altorki et. al., 2019

Bone Marrow



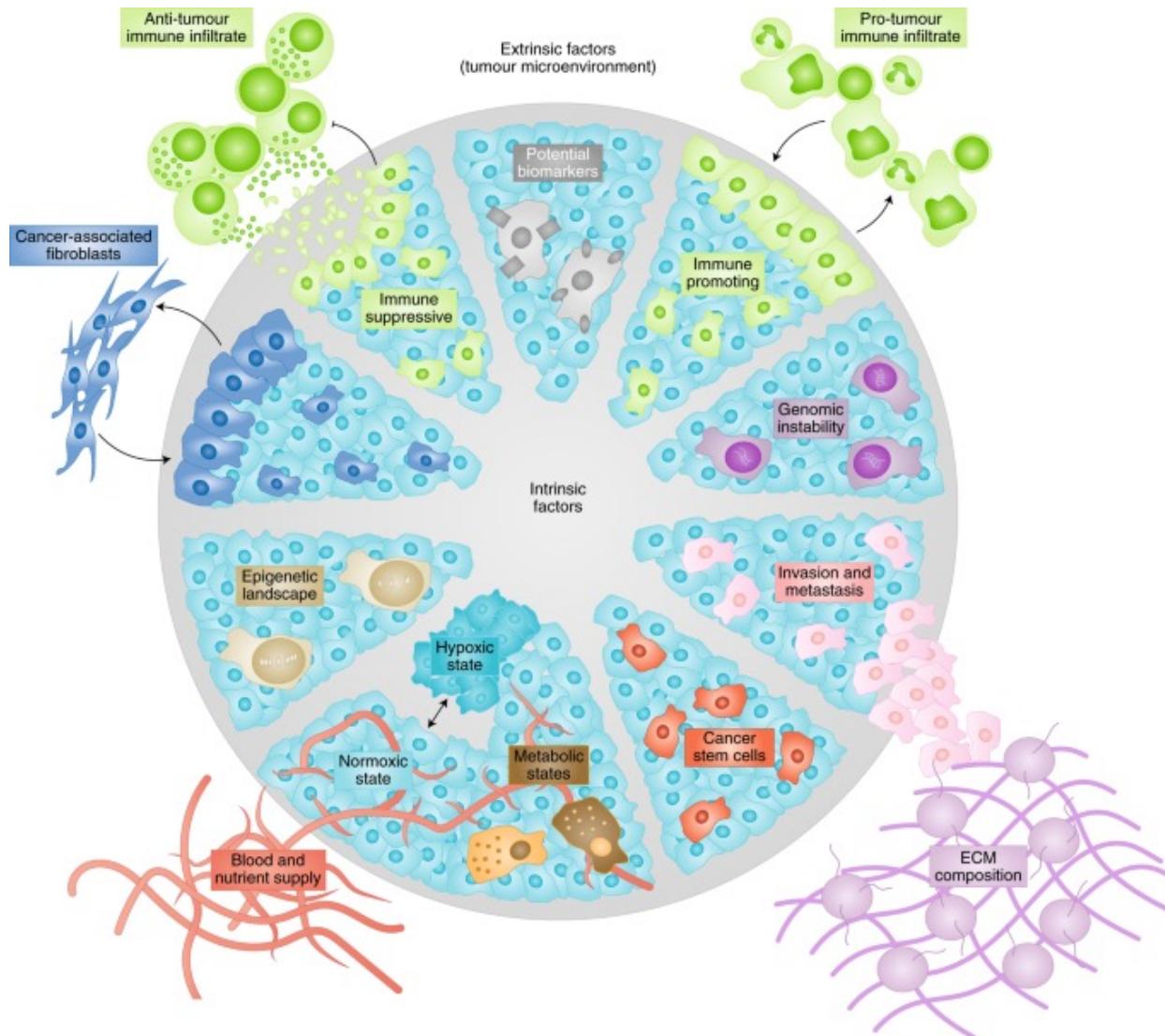
Hoffman et. al., 2020

Somatosensory Cortex



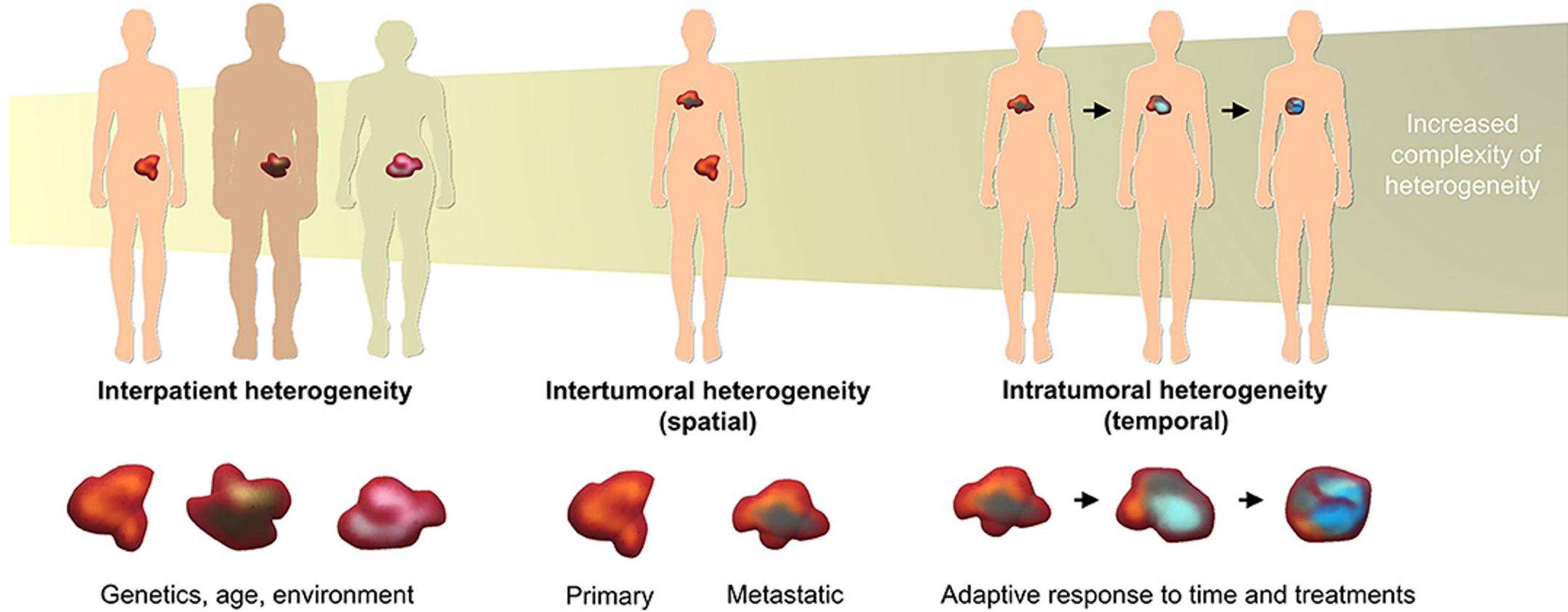
Codeluppi et. al., 2018

Tumor heterogeneity



Lawson et. al., 2018

Tumor heterogeneity can lead to different responses to therapy

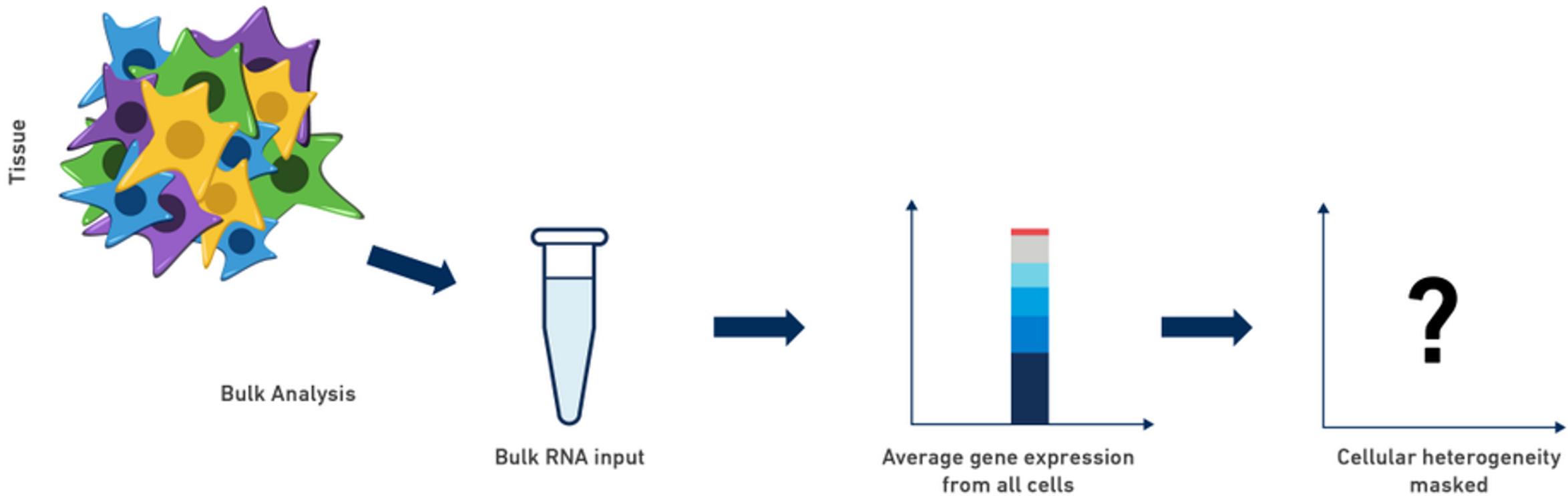


Single-cell technologies profile heterogeneity

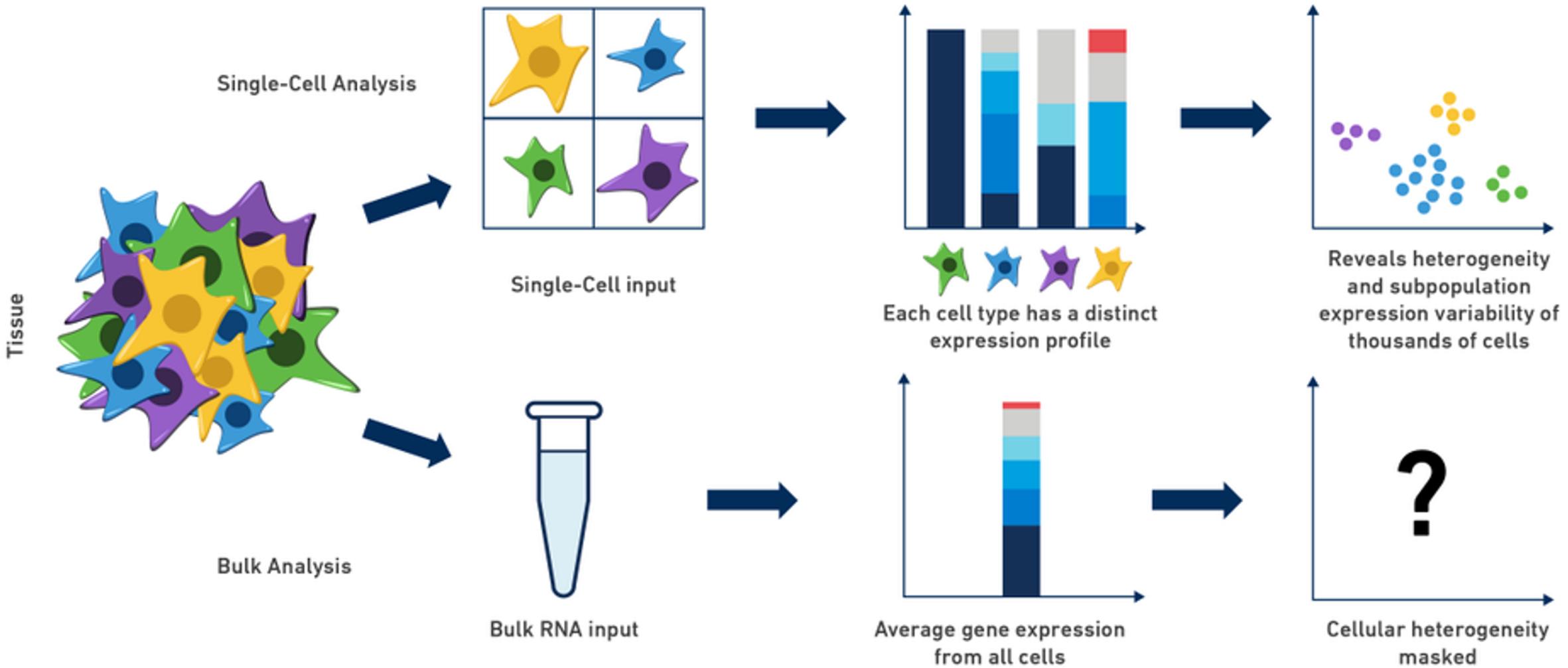
Tissue



Single-cell technologies profile heterogeneity



Single-cell technologies profile heterogeneity



Single-cell technologies profile heterogeneity instead of average

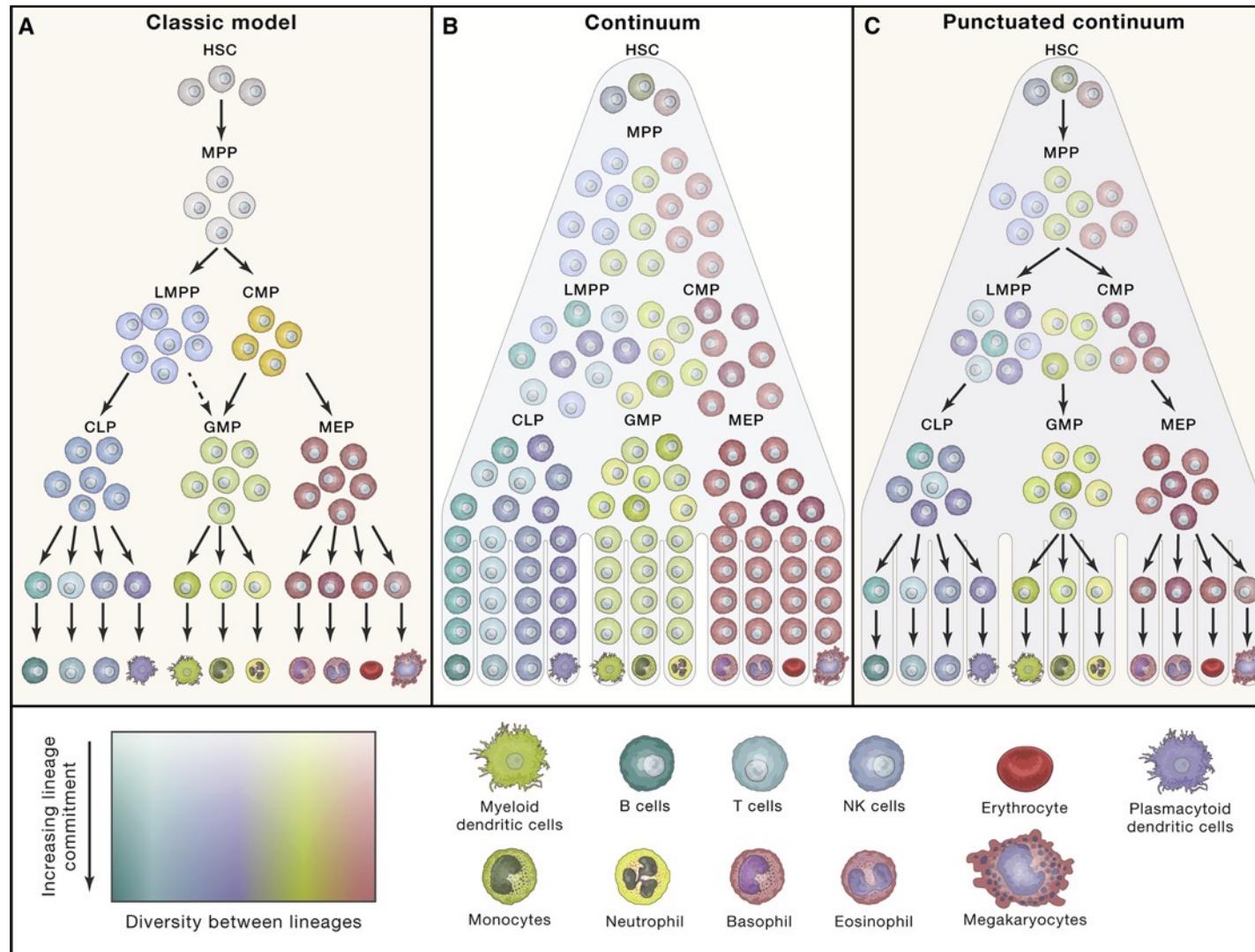


Bulk Genomics



Single-cell Genomics

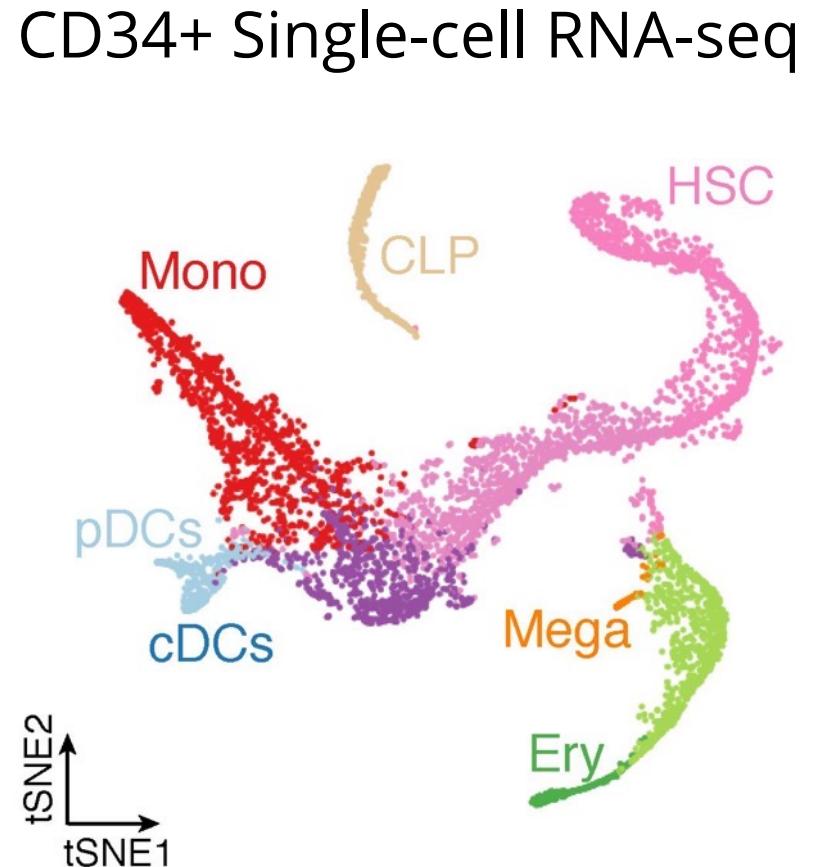
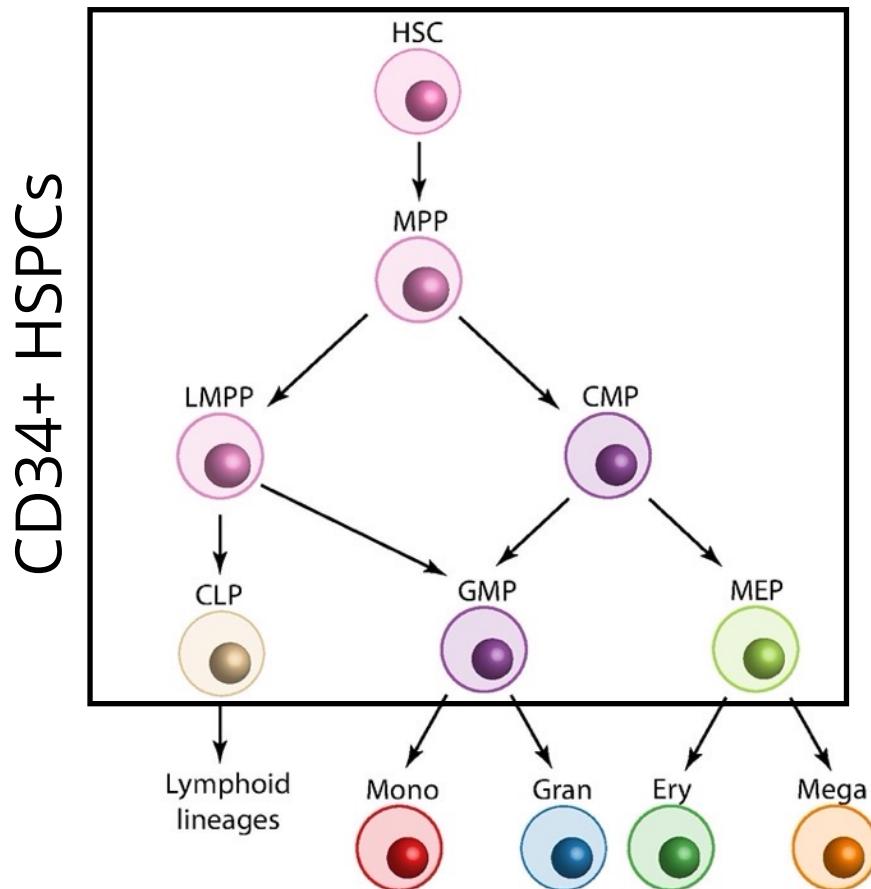
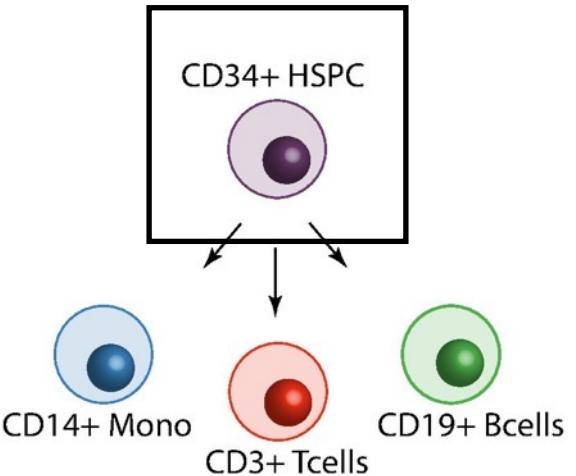
Single-cell genomics: Continuous phenotypic landscape



Ligget et. al., 2020

Single-cell RNA-seq

Continuous cell-state transitions

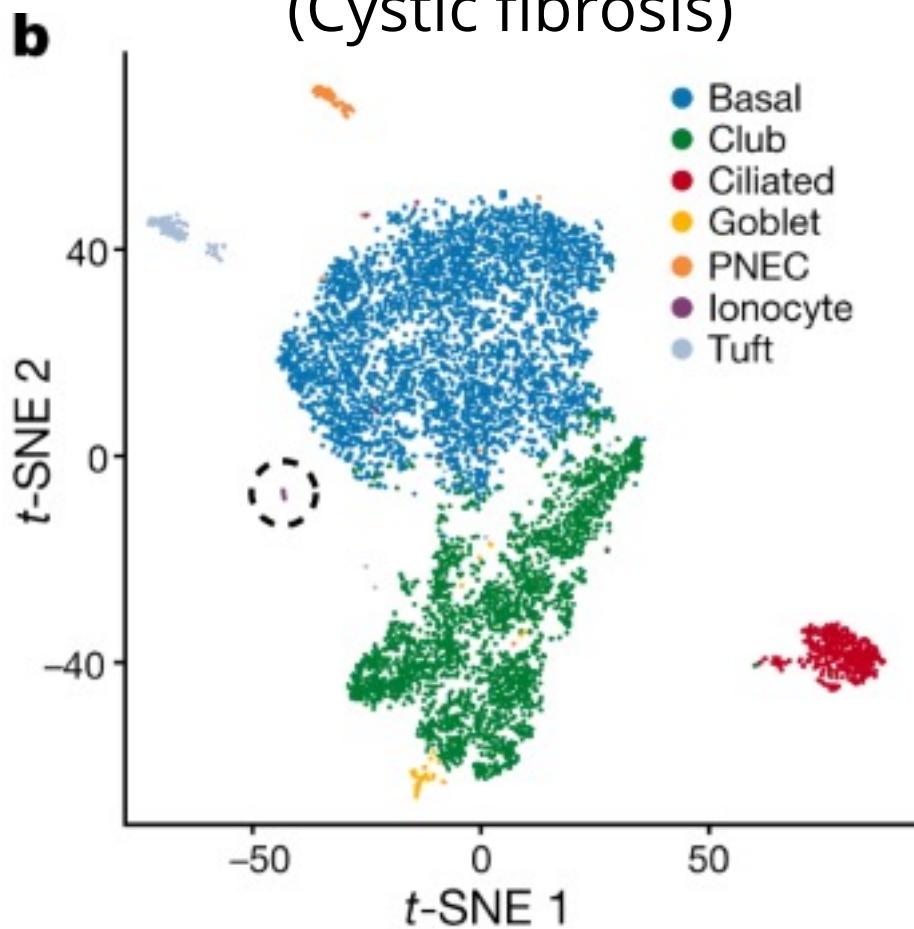


Human hematopoiesis in the bone marrow

Single-cell RNA-seq

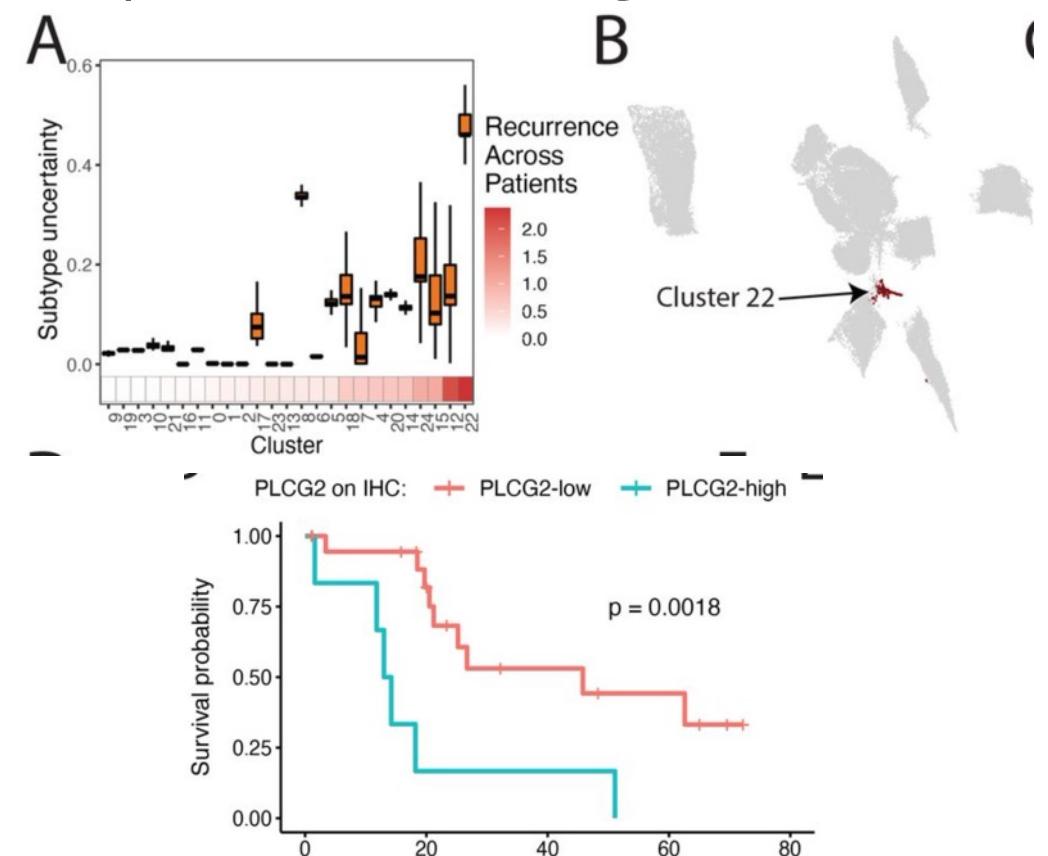
Rare populations

Pulmonary Ionocytes
(Cystic fibrosis)



Monotor et. al., 2018

Rare stem-like cells
(Squamous cell lung carcinoma)



Chan et. al., 2020

Single-cell genomics: How?

Single-cell technologies profile heterogeneity instead of average



Bulk Genomics



Single-cell Genomics

How to isolate and capture mRNA from individual cells?

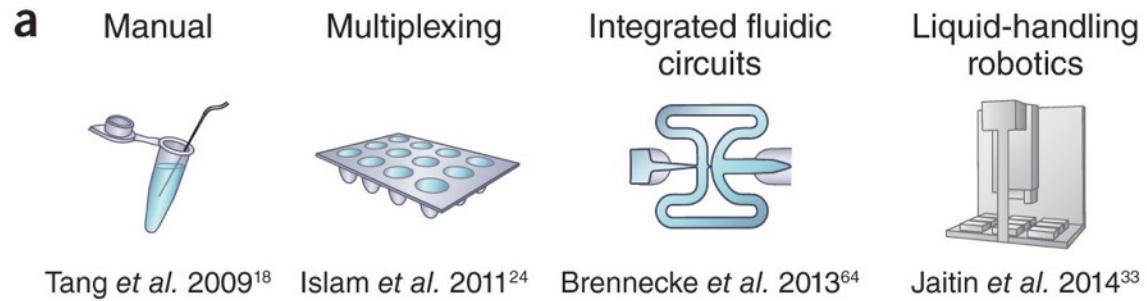
Evolution of single-cell RNA-seq

a Manual

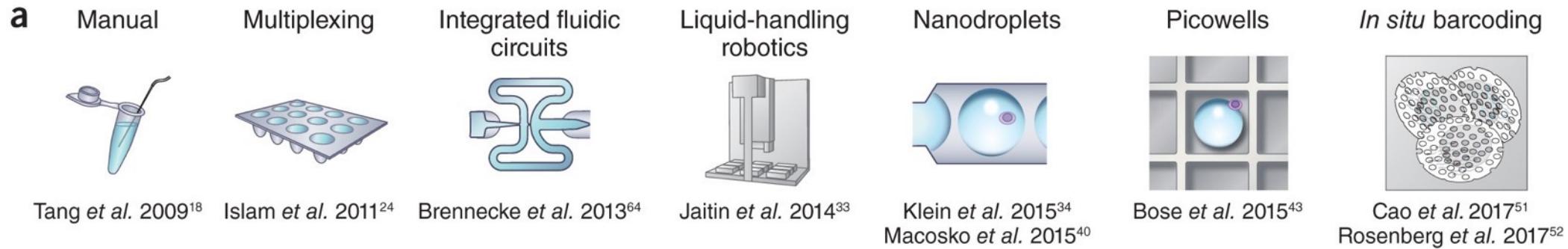


Tang *et al.* 2009¹⁸

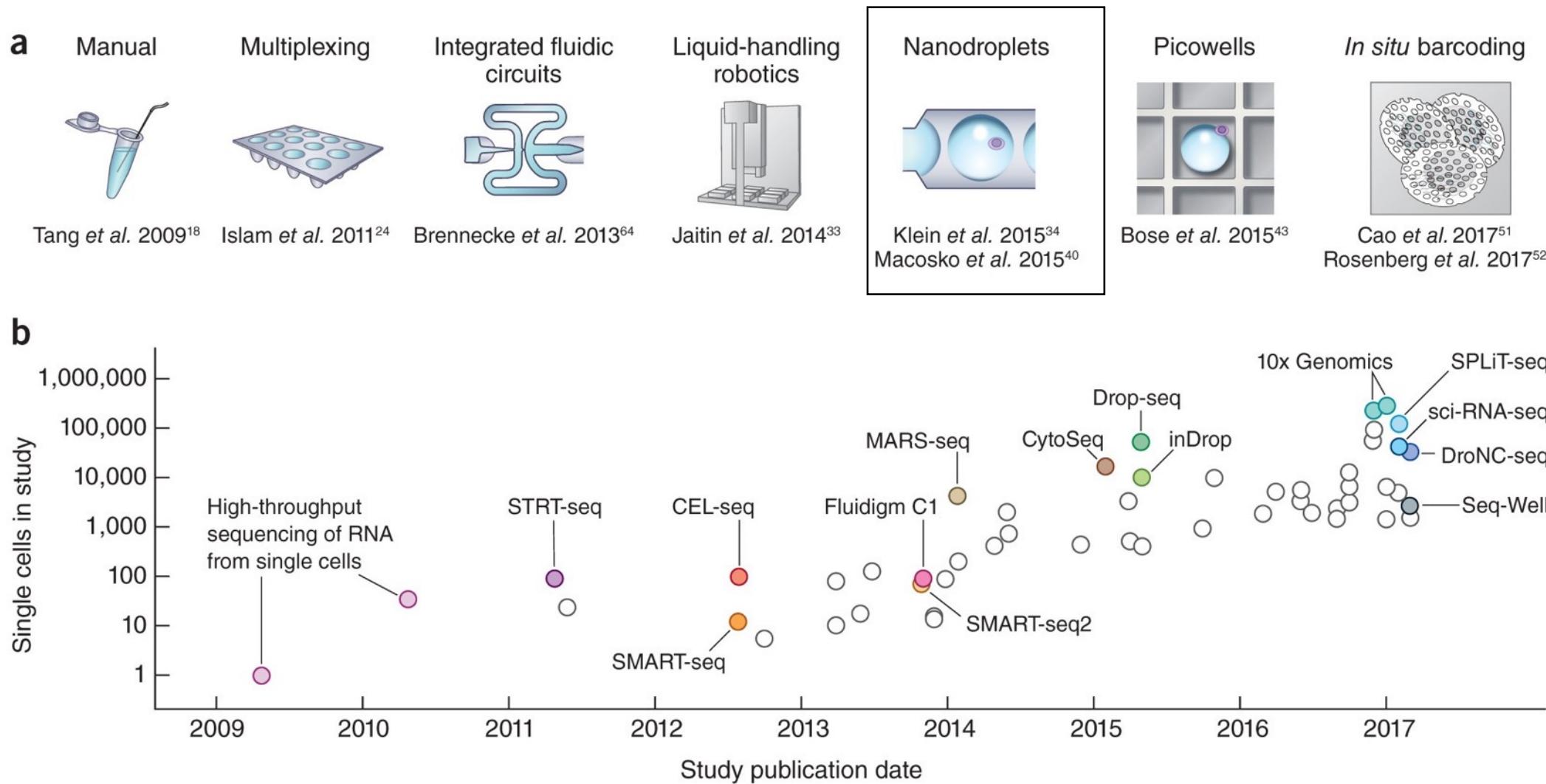
Evolution of single-cell RNA-seq



Evolution of single-cell RNA-seq



Evolution of single-cell RNA-seq



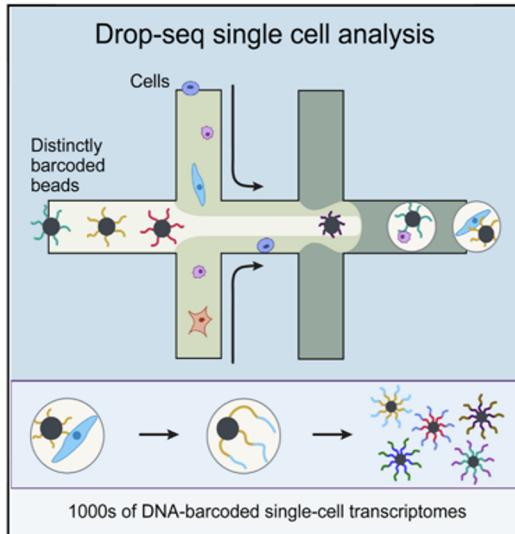
Microfluidic single-cell technologies

Moved throughput from hundreds to thousands of cells

Cell

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

Graphical Abstract

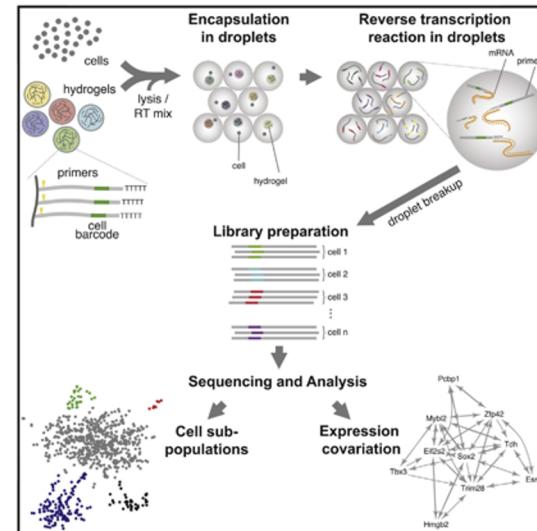


Resource

Cell

Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells

Graphical Abstract



Resource

Authors

Allon M. Klein, Linas Mazutis, ..., David A. Weitz, Marc W. Kirschner

Correspondence

weitz@seas.harvard.edu (D.A.W.), marc@hms.harvard.edu (M.W.K.)

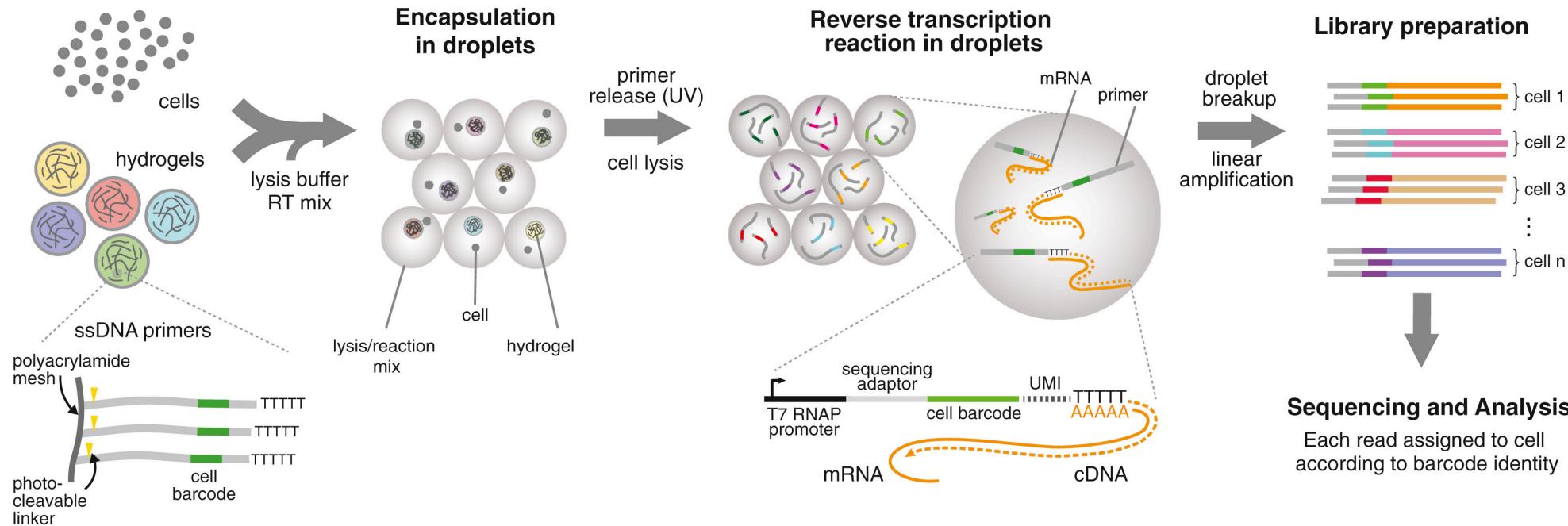
In Brief

Capturing single cells along with a set of uniquely barcoded primers in tiny droplets enables single-cell transcriptomics of a large number of cells in a heterogeneous population. Applying this analysis to mouse embryonic stem cells reveals their population structure, gene expression relationships, and the heterogeneous onset of differentiation.

Klein et al *Cell* 2015
Macosko et al *Cell* 2015

Microfluidic single-cell technologies

Moved throughput from hundreds to thousands of cells

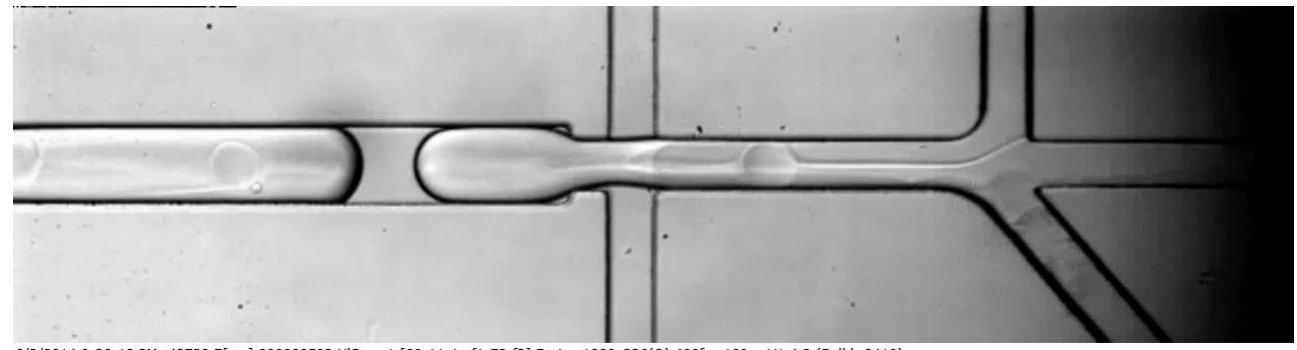
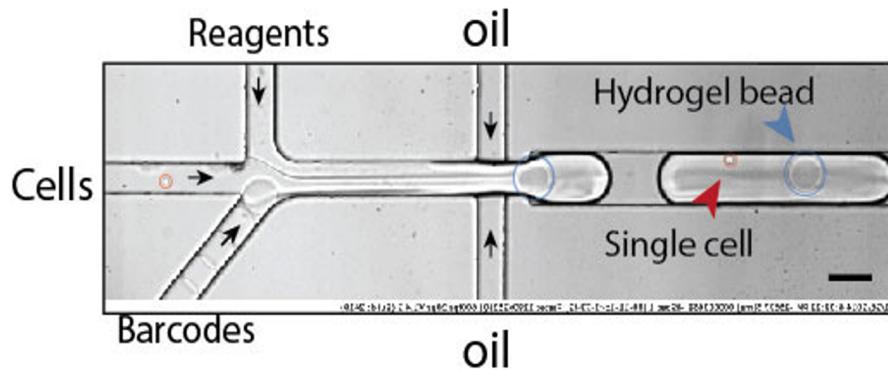


- Droplet-based processing using microfluidics
- Nanoliter scale aqueous drops in oil
- 3' End
- Bead based for cell barcoding
- Uses UMI (Unique Molecular Identifier).

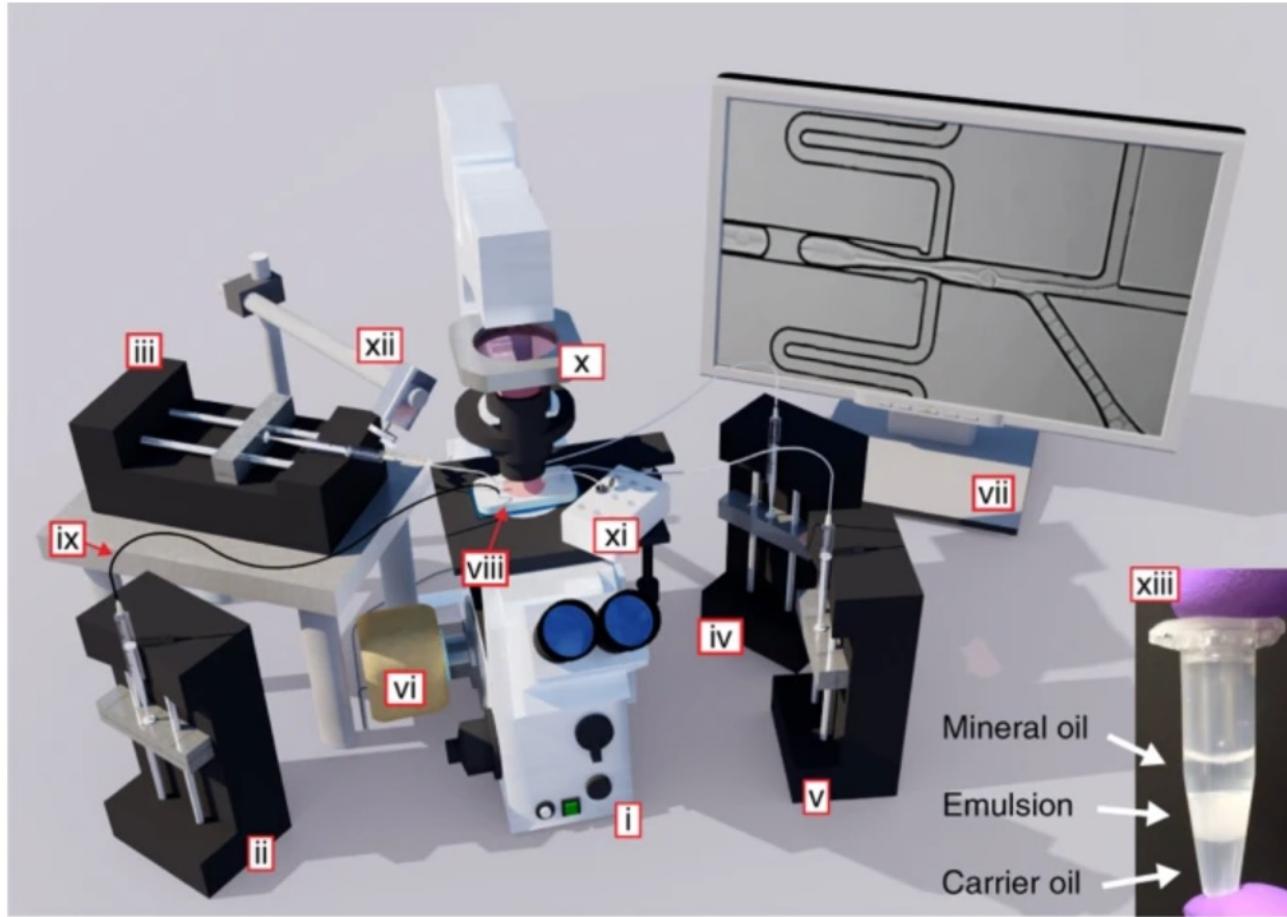
Klein et al *Cell* 2015
Macosko et al *Cell* 2015

Microfluidic single-cell technologies

Moved throughput from hundreds to thousands of cells



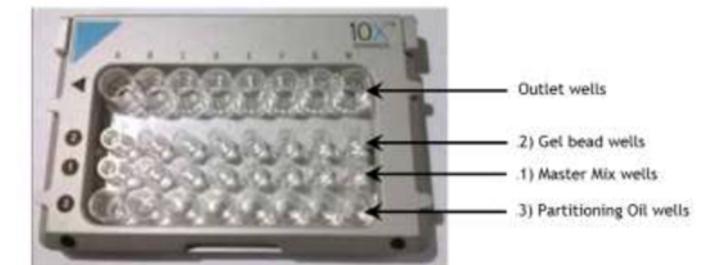
InDrop setup



(i) Inverted bright-field microscope; (ii–v) syringe pumps; (vi) fast speed camera; (vii) computer; (viii) microfluidic chip; (ix) barcoding hydrogel bead containing tubing, protected from ambient light by insertion into a second, opaque tubing, or alternatively, wrapped in aluminum foil; (x) red band-pass filter (≥ 600 nm); (xi) ice-cold rack containing the collection tube; (xii) cell mixer; (xiii) photograph of the collection tube showing three distinct phases.

10X Genomics: Massively parallel sequencing

- Droplet-based, 3' mRNA.
 - GEM (Gel Bead in Emulsion)
- Standardized instrumentation and reagents
- More high-throughput scaling to tens of thousands.
- Less processing time, easier training.
- Cell Ranger software for preprocessing data



10X Genomics

Human Cell Atlas

Mapping the Human Body at the Cellular Level

Community generated, multi-omic,
open data processed by standardized pipelines



4.5M
CELLS



33
ORGANS



289
DONORS



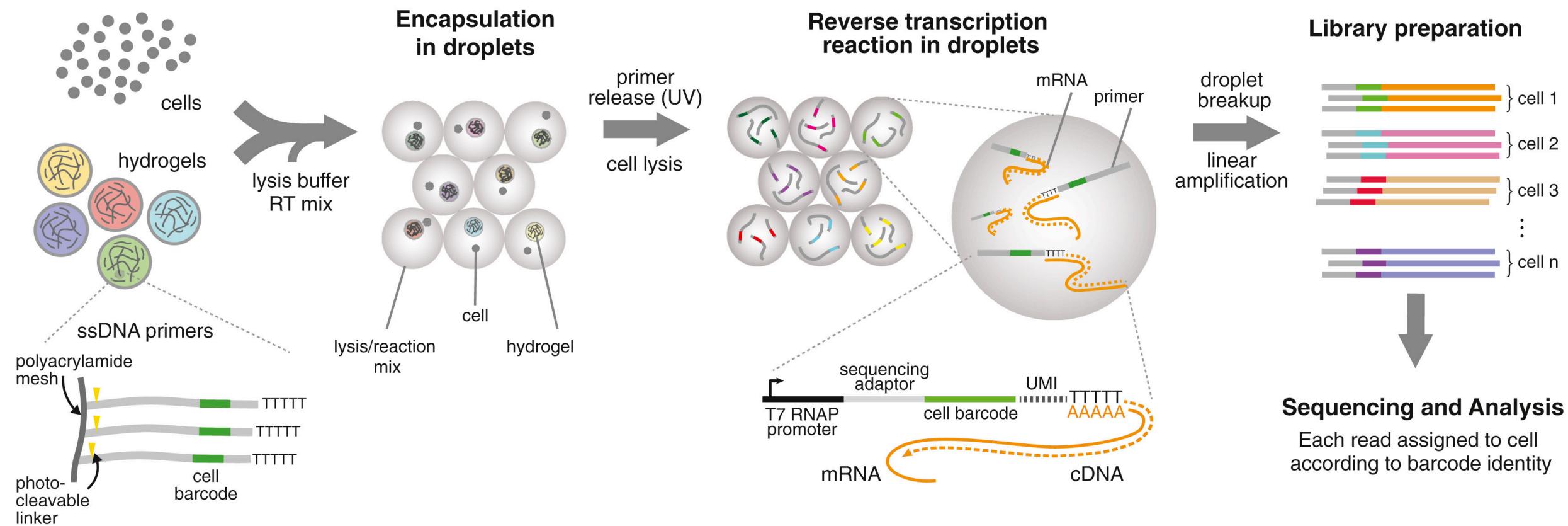
28
PROJECTS



81
LABS

Single-cell RNA-seq: Preprocessing and Analysis

Microfluidic single-cell technologies



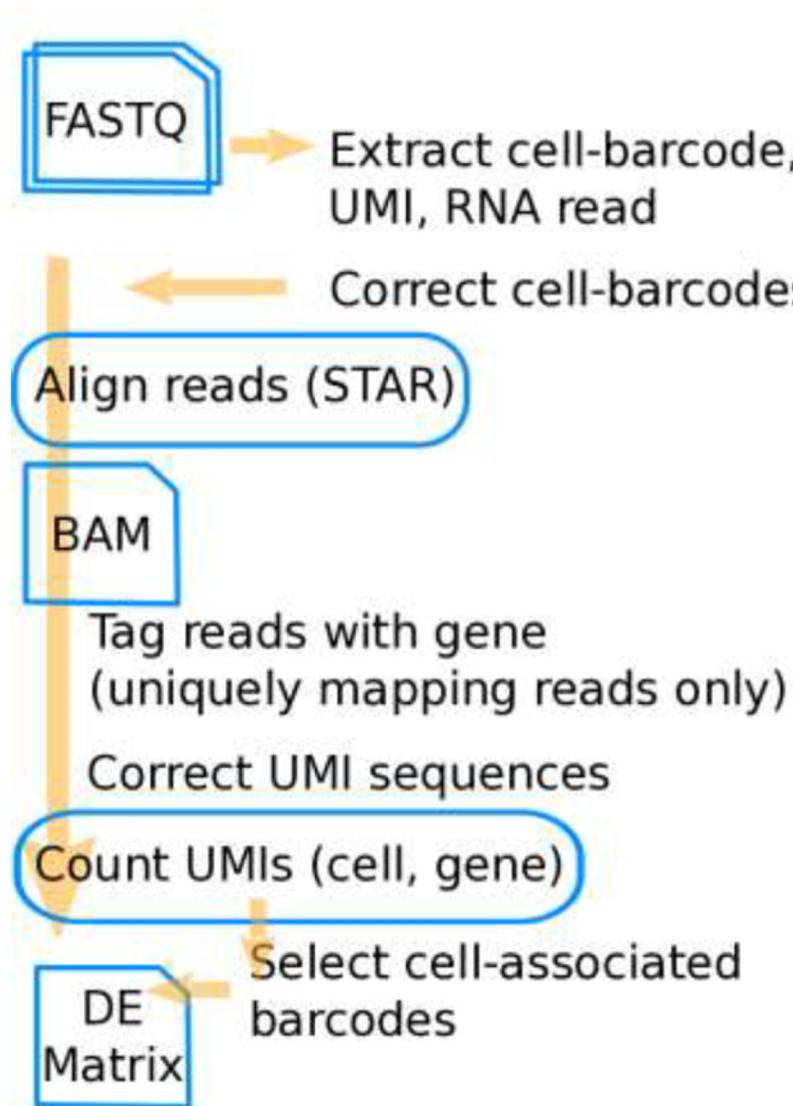
Cell Barcodes and UMIs

- Cell Barcode: Unique cell identifier – Whitelist
- UMI: Unique molecular identifier – Random 8mer

	Cellular barcode	UMI	
Cell 1	TTGCCGTGGTGT	GGCGGGGA.....	CGGTGTTA] <i>DDX51</i>
	TTGCCGTGGTGT	TATGGAGG.....	CCAGCACC] <i>NOP2</i>
	TTGCCGTGGTGT	TCTCAAGT.....	AAAATGGC] <i>ACTB</i>
Cell 2	CGTTAGATGGCA	GGGCCGGG.....	CTCATAGT] <i>LBR</i>
	CGTTAGATGGCA	ACGTTATA.....	ACGGTAC] <i>ODF2</i>
	CGTTAGATGGCA	TCGAGATT.....	AGCCCTTT] <i>HIF1A</i>
Cell 3	AAATTATGACGA	AGTTTGTA.....	GGGAATTAA] <i>ACTB</i> ← 2 reads, 1 molecule
	AAATTATGACGA	AGTTTGTA.....	AGATGGGG]
	AAATTATGACGA	TGTGCTTG.....	GACTGCAC] <i>RPS15</i>
Cell 4	GTAAACGTACC	CTAGCTGT.....	GATTTTCT] <i>GTPBP4</i>
	GTAAACGTACC	GCAGAAAGT.....	GTTGGCGT] <i>GAPDH</i>
	GTAAACGTACC	AAGGCTTG.....	CAAAGTTC]
	GTAAACGTACC	TTCCGGTC.....	TCCAGTCG] <i>ARL1</i> ← 2 reads, 2 molecules

(Thousands of cells)

10X Preprocessing Pipeline: CellRanger



Single-cell RNA-seq

*Scale & Resolution X *Noise & Sparsity**

Genes

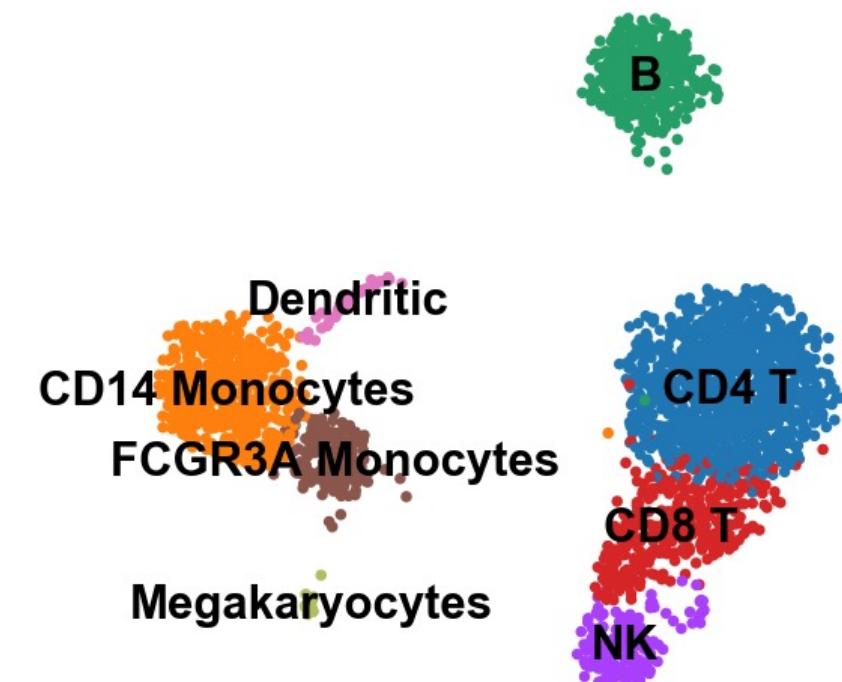
Cells	0	0	0	0	20	0	0	0	0	0	0	0	0	76	0	0	0	62
	0	0	0	0	0	0	0	0	0	0	0	0	0	87	0	0	0	0
	0	38	0	0	0	0	0	0	2	0	0	0	0	0	90	0	0	0
	99	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0
	0	0	0	0	0	24	0	0	0	0	0	0	0	7	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	40	0	0	0	5	0	84	0	0	0	0	0	0
	10	36	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0
	0	0	0	48	0	95	0	3	0	73	0	0	0	8	0	0	0	0
	0	10	0	0	0	0	0	0	0	0	0	0	0	53	0	0	52	0
	0	0	3	0	0	77	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	52	0	0	0	0	94	0	0	0	0	56	0	0	0	0	0	0
	0	0	0	0	0	90	0	0	0	85	45	0	0	0	0	0	0	0
	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75
	0	0	22	0	0	0	0	0	0	65	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	54	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	47	0	0	0	0	0	0	0	0	0	0	0	0	54
	0	0	0	0	0	0	0	0	0	0	0	0	0	84	0	0	0	0
	0	0	0	0	0	0	94	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	26
	0	71	0	46	0	0	0	0	0	52	0	0	27	0	0	76	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	72	69	0	0	77
	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0
	0	67	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	94	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	80	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	93	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	62	0	0	0	5	0	0	0
	0	0	0	30	0	0	0	0	0	48	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	79	0	0	0	0	0	0	0	0	0
	0	0	30	74	0	0	0	22	0	0	52	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	73	0	35	0	0	0	0	0	0	0	0
	0	0	90	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0
	0	90	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0
	0	0	0	21	0	0	0	0	0	0	0	0	0	0	39	0	0	0
	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0
	0	0	12	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0
	0	0	0	8	0	0	0	0	0	0	6	0	0	0	0	0	0	0
	0	0	0	11	0	0	0	23	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	75	0	0	63	0	0	16	0	0	19	0	36
	0	0	0	0	0	0	90	0	0	0	0	0	0	0	0	0	0	0
	0	96	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	57	0	0	92	0	0	0	46	0	0	0	99

Single-cell RNA-seq

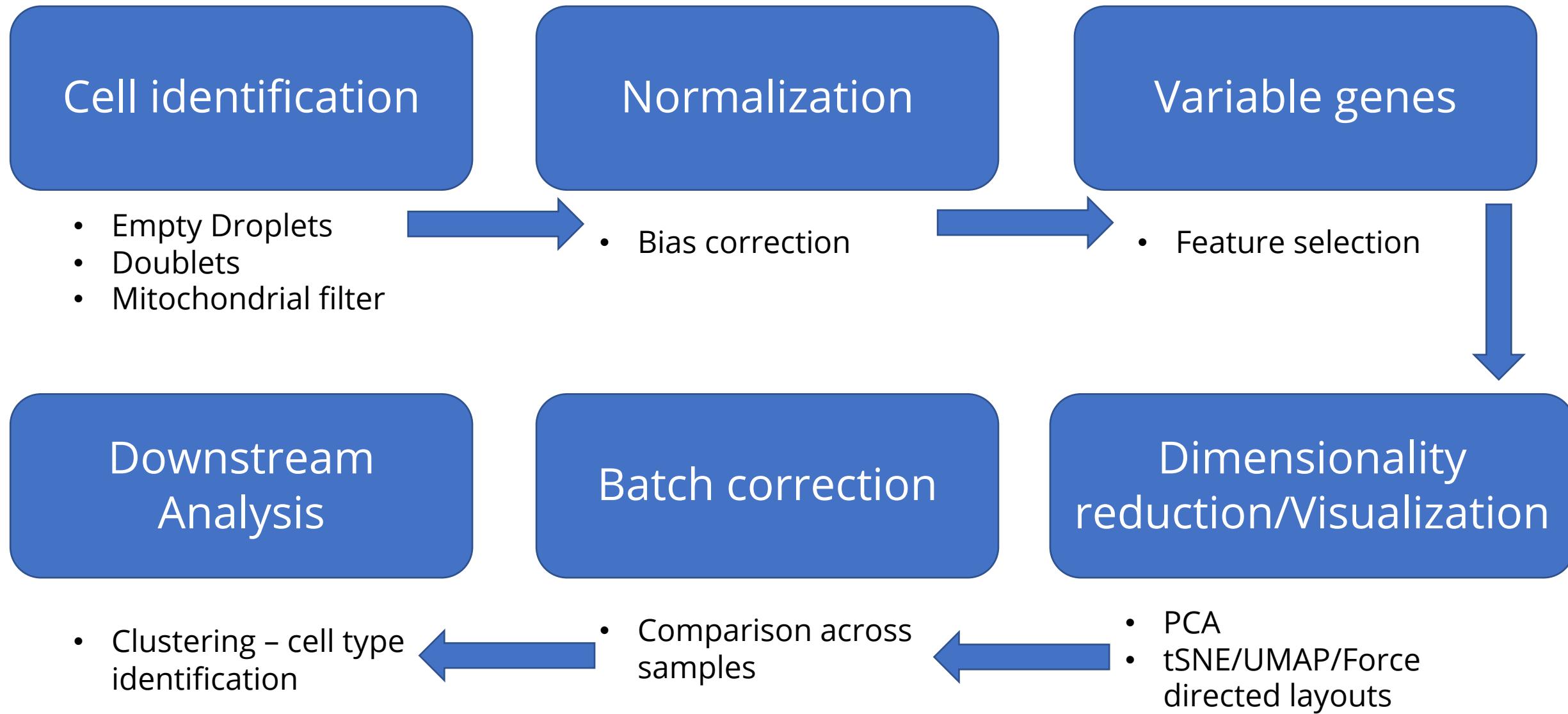
*Scale & Resolution X *Noise & Sparsity**

Genes

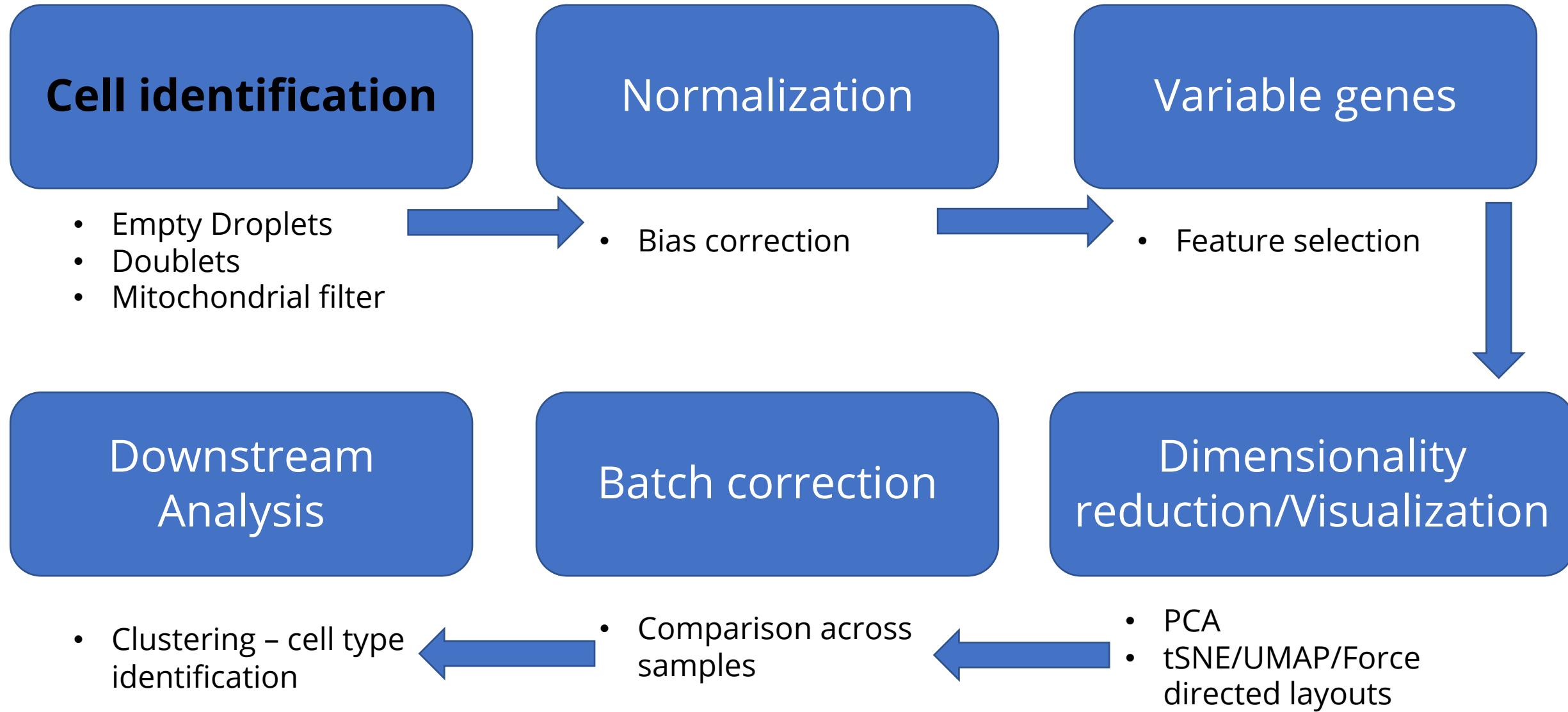
0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	76	0	0	62
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	87	0	0	0
99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	0	0	0
0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	7	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	40	0	0	0	0	5	0	84	0	0	0	0	0	0	0
10	36	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0
0	0	0	48	0	95	0	3	0	73	0	0	0	0	8	0	0	0	0	0
0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	53	0	0	52
0	0	3	0	0	77	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	52	0	0	0	94	0	0	0	0	0	56	0	0	0	0	0	0	0	0
0	0	0	0	90	0	0	0	85	45	0	0	0	0	0	0	0	0	0	0
0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	75
0	0	22	0	0	0	0	0	0	65	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	54	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54
0	0	0	0	0	0	0	0	0	0	0	0	0	0	84	0	0	0	0	0
0	0	0	0	0	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	26
0	71	0	46	0	0	0	0	52	0	0	27	0	0	76	0	0	0	0	87
0	0	0	0	0	0	0	0	0	0	0	0	0	72	69	0	0	0	0	77
0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0
0	67	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	94	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	80	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	93	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	62	0	0	0	5	0	0	0	0	0	0
0	0	0	30	0	0	0	0	0	0	48	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	79	0	0	0	0	0	0	0	0	0	0	0
0	0	30	74	0	0	0	22	0	0	52	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	73	0	35	0	0	0	0	0	0	0	0	0	0	0
0	0	90	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	0
0	90	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0
0	0	0	21	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0
0	0	0	0	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0
0	0	12	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0
0	0	0	8	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0
0	0	0	11	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	75	0	0	63	0	0	16	0	0	19	0	36	0	0
0	0	0	0	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0	0
0	96	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	57	0	0	92	0	0	0	0	46	0	0	0	0	0	99



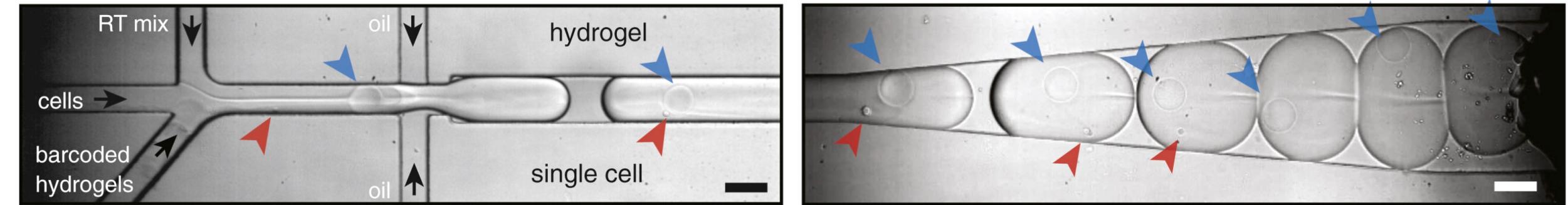
scRNA-seq analysis steps



scRNA-seq analysis steps



scRNA-seq: Empty droplets & Ambient RNA



- Most droplets do not have cells!
 - Ambient RNA

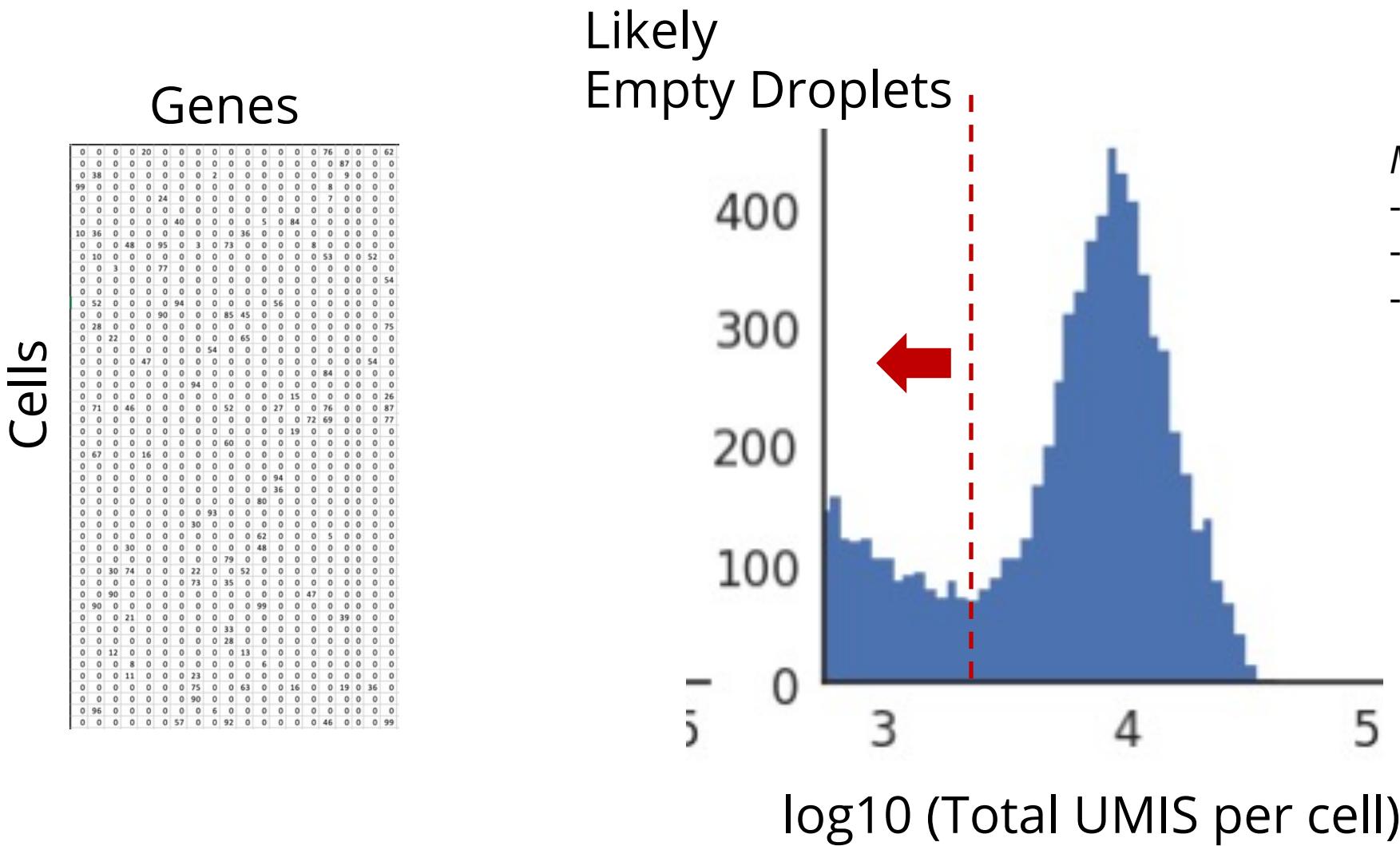
Single-cell RNA-seq

All rows are not real cells

Genes

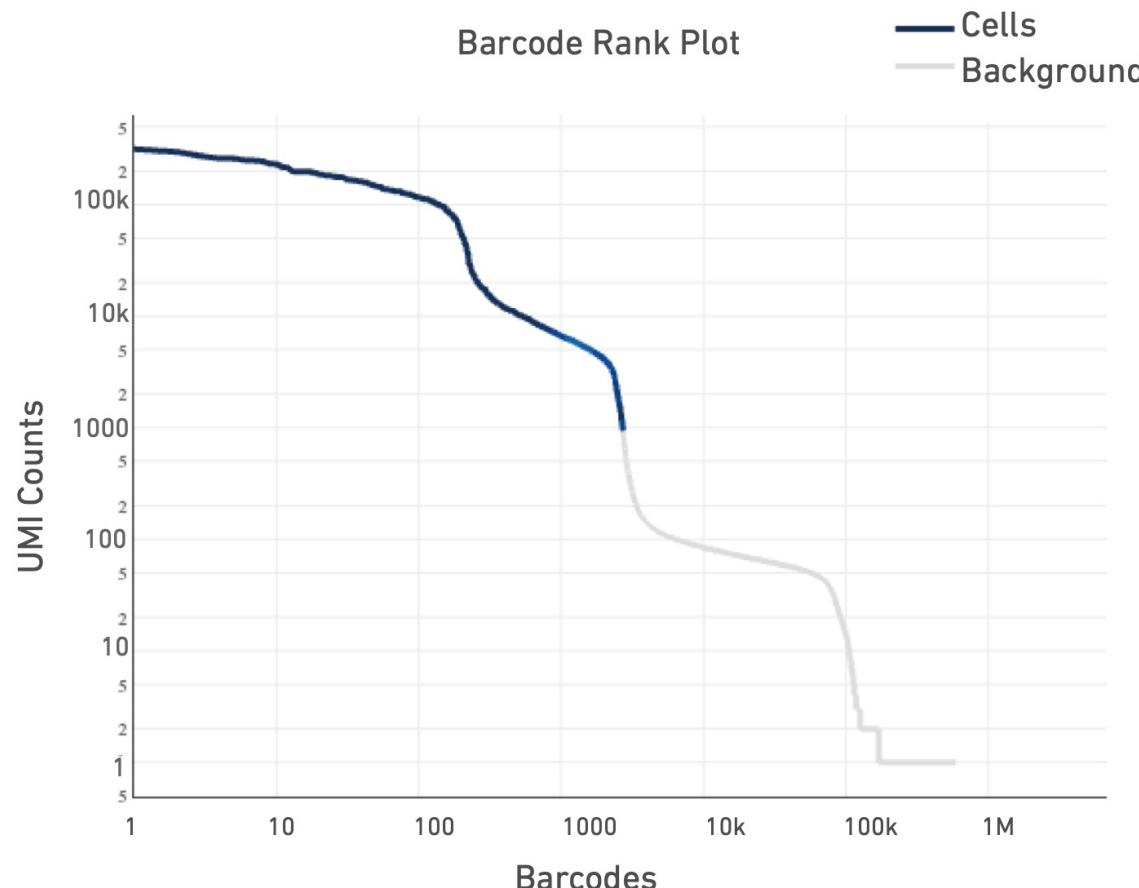
Cells	Genes
0	0 0 0 0 20 0 0 0 0 0 0 0 0 0 0 0 0 0 76 0 0 0 62
0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 87 0 0 0
0	38 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 90 0 0 0
99	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 8 0 0 0 0 0
0	0 0 0 0 0 24 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0 0
0	0 0
0	0 0 0 0 0 40 0 0 0 0 5 0 84 0 0 0 0 0 0 0 0 0
10	36 0 0 0 0 0 0 0 0 36 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 48 0 95 0 3 0 73 0 0 0 0 8 0 0 0 0 0 0 0
0	10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 53 0 0 52 0
0	0 0 3 0 0 77 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 54 0
0	0 0
0	52 0 0 0 0 94 0 0 0 0 0 56 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 90 0 0 0 85 45 0 0 0 0 0 0 0 0 0 0 0
0	28 0 75 0
0	0 0 22 0 0 0 0 0 0 65 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 54 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 47 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 54 0
0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 84 0 0 0 0
0	0 0 0 0 0 0 0 94 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 15 0 0 0 0 0 0 26
0	71 0 46 0 0 0 0 0 52 0 0 27 0 0 76 0 0 0 87 0
0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 72 69 0 0 0 77 0
0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 19 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 0 60 0 0 0 0 0 0 0 0 0 0 0 0 0
0	67 0 0 16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0
0	0 0 0 0 0 0 0 0 0 0 0 0 94 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 0 0 0 36 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 0 0 0 80 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 0 93 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 30 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 0 0 0 62 0 0 0 5 0 0 0 0 0 0 0
0	0 0 0 30 0 0 0 0 0 0 0 48 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 79 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 30 74 0 0 0 22 0 0 52 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 73 0 35 0 0 0 0 0 0 0 0 0 0 0 0
0	0 90 0 0 0 0 0 0 0 0 0 0 0 47 0 0 0 0 0 0 0 0
0	90 0 0 0 0 0 0 0 0 0 0 99 0 0 0 0 0 0 0 0 0 0
0	0 0 21 0 0 0 0 0 0 0 0 0 0 0 0 0 0 39 0 0 0
0	0 0 0 0 0 0 0 0 0 33 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 0 28 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 12 0 0 0 0 0 0 0 13 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 8 0 0 0 0 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 11 0 0 0 23 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 0 75 0 0 63 0 0 16 0 0 19 0 36 0
0	0 0 0 0 0 0 0 90 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	96 0 0 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0	0 0 0 0 0 0 57 0 0 92 0 0 0 0 46 0 0 0 0 0 0

scRNA-seq: Removal of empty droplets



EmptyDrops

- Estimate the likelihood of a gene expression profile derived from an empty droplet using the "obvious" empty droplets as reference



EmptyDrops

- Null hypothesis: Free floating transcripts/ambient RNA are randomly encapsulated into droplets
- Model counts as a Dirichlet – Multinomial distribution
 - Model each genes as a gamma distribution and assume independence

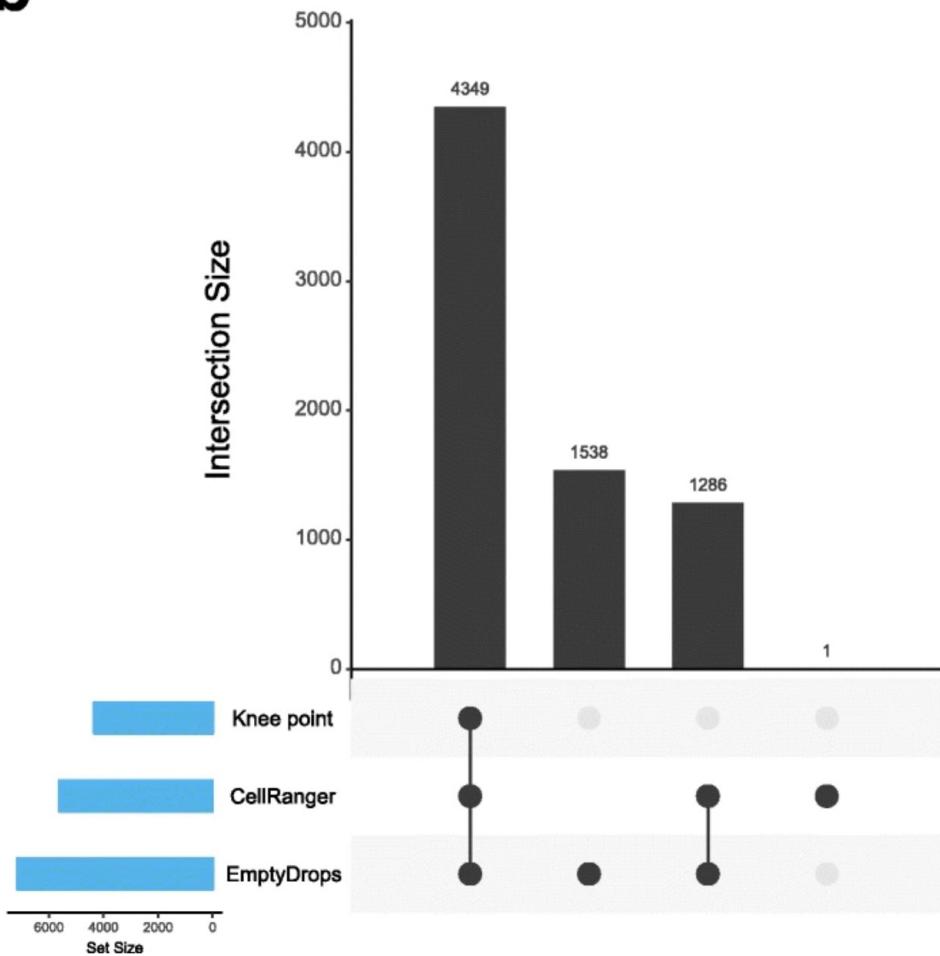
$$L_b = \frac{t_b! \Gamma(\alpha)}{\Gamma(t_b + \alpha)} \prod_{g=1}^N \frac{\Gamma(y_{gb} + \alpha_g)}{y_{gb}! \Gamma(\alpha_g)},$$

Gene g
Barcode b
Total counts t_b
Gene counts y

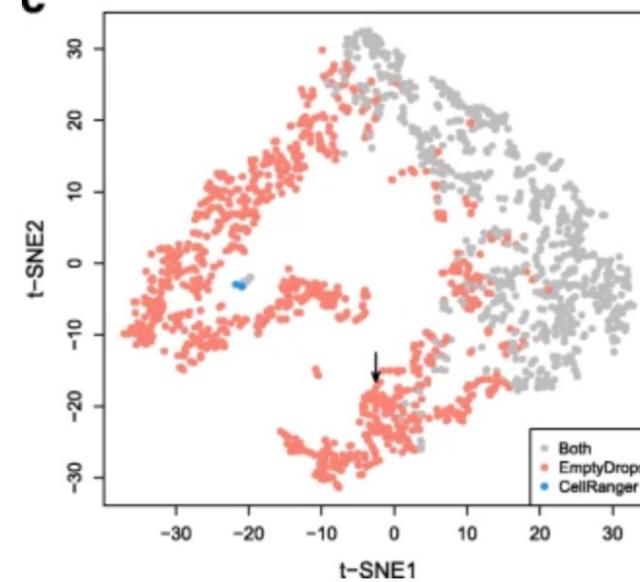
- Estimate parameters by MCMC
- Identify cell containing droplets using FDR

EmptyDrops

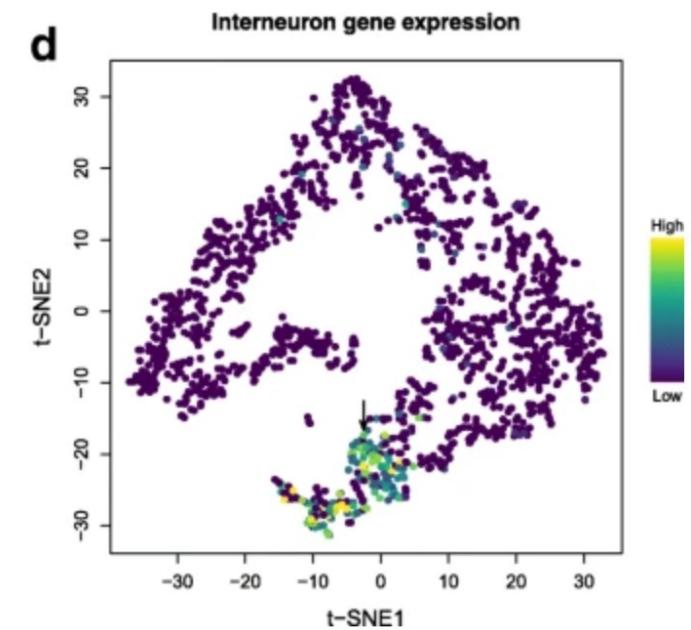
b



c

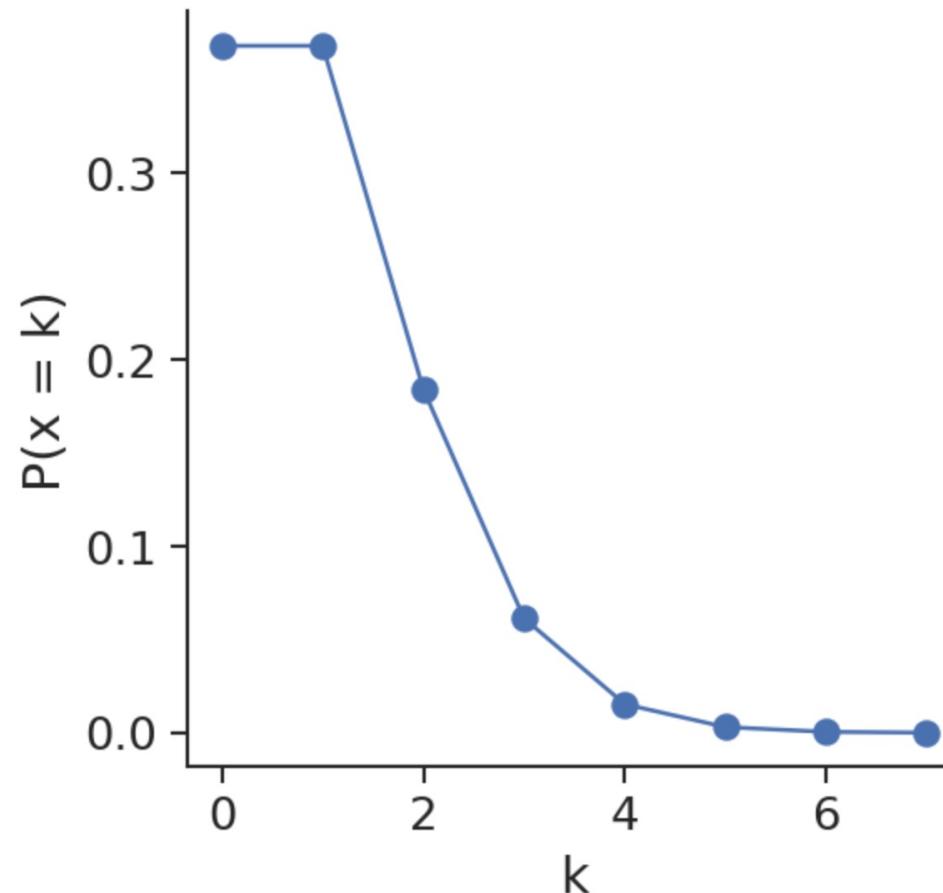


d

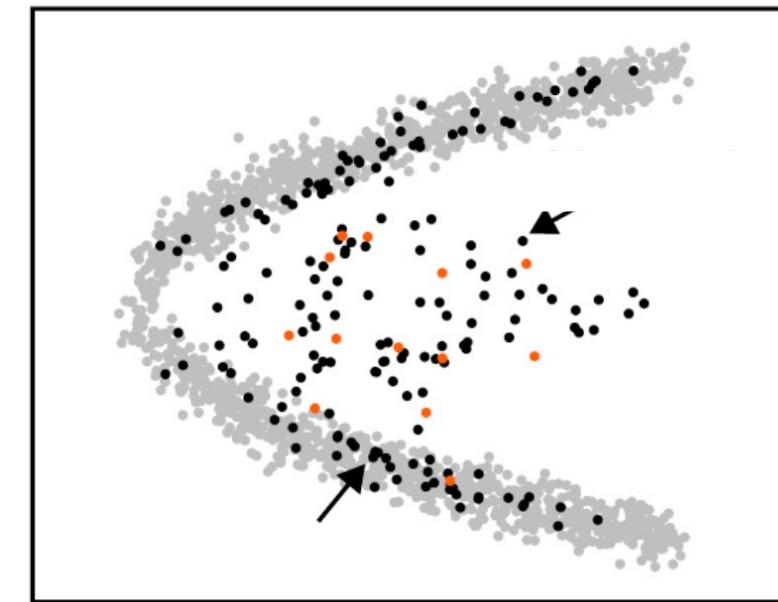
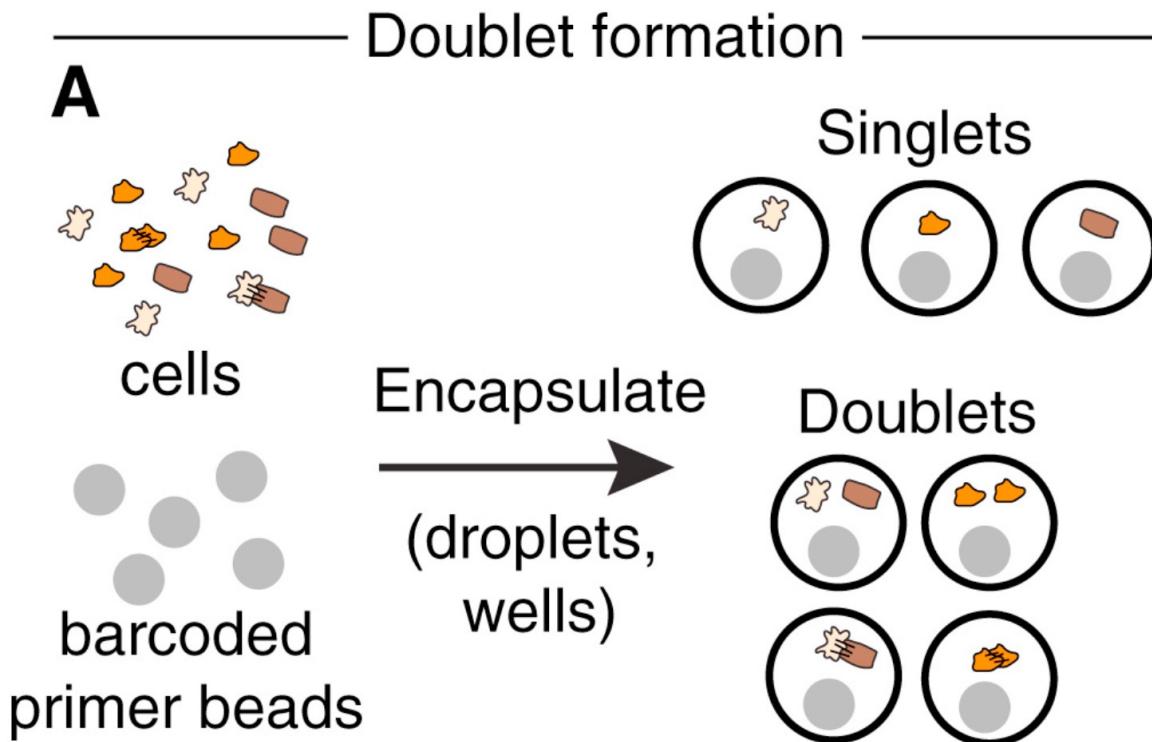


Cell containing droplets

- Cell encapsulation follows a Poisson distribution – are there reasons beyond ambient RNA, that can lead to misleading biology

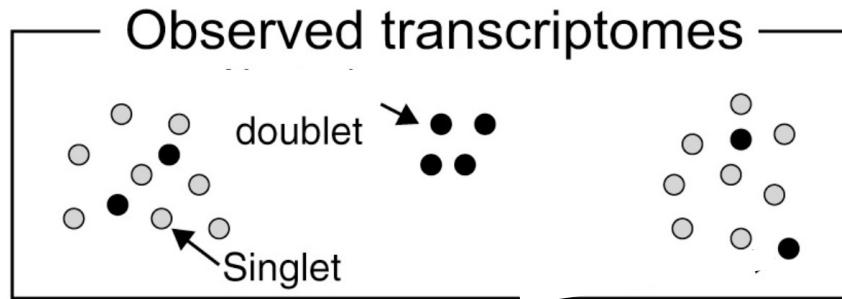


Doublets



How to detect doublets?

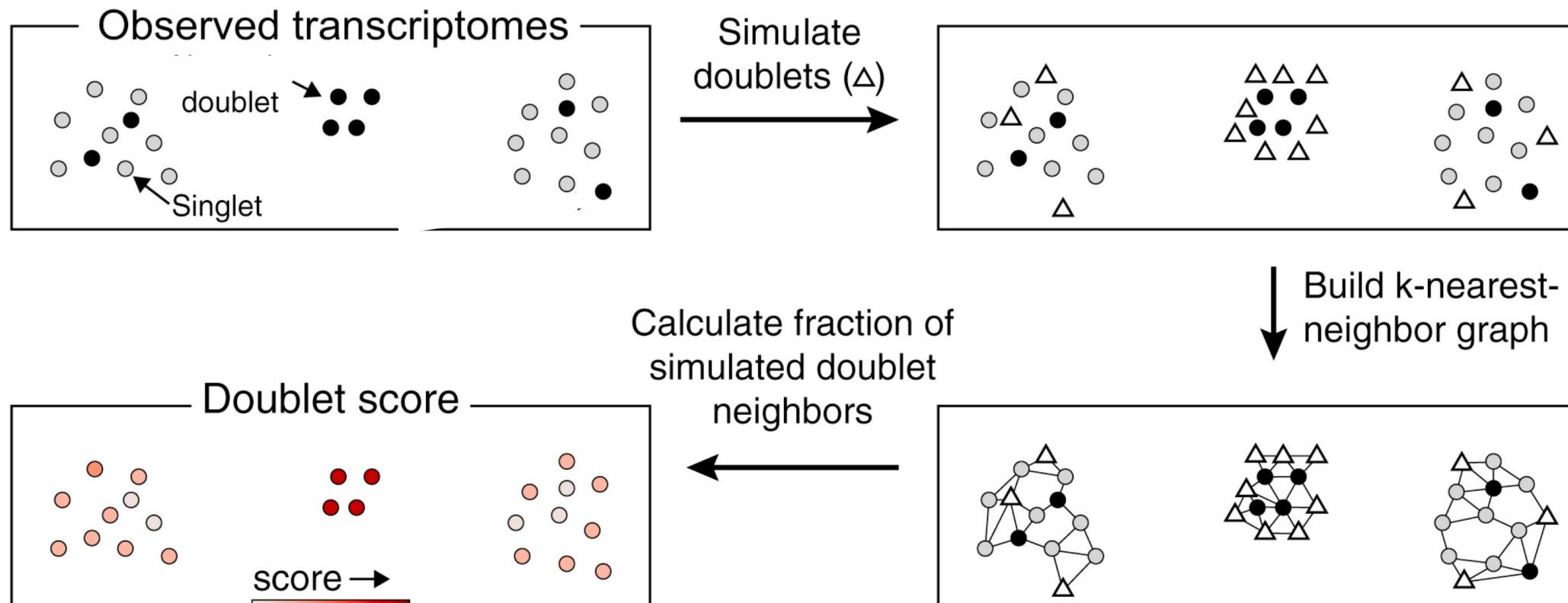
- Assume:
 - Multiplets / Doublets are rare
 - Constituent singlets are present



How to detect doublets?

- Assume:

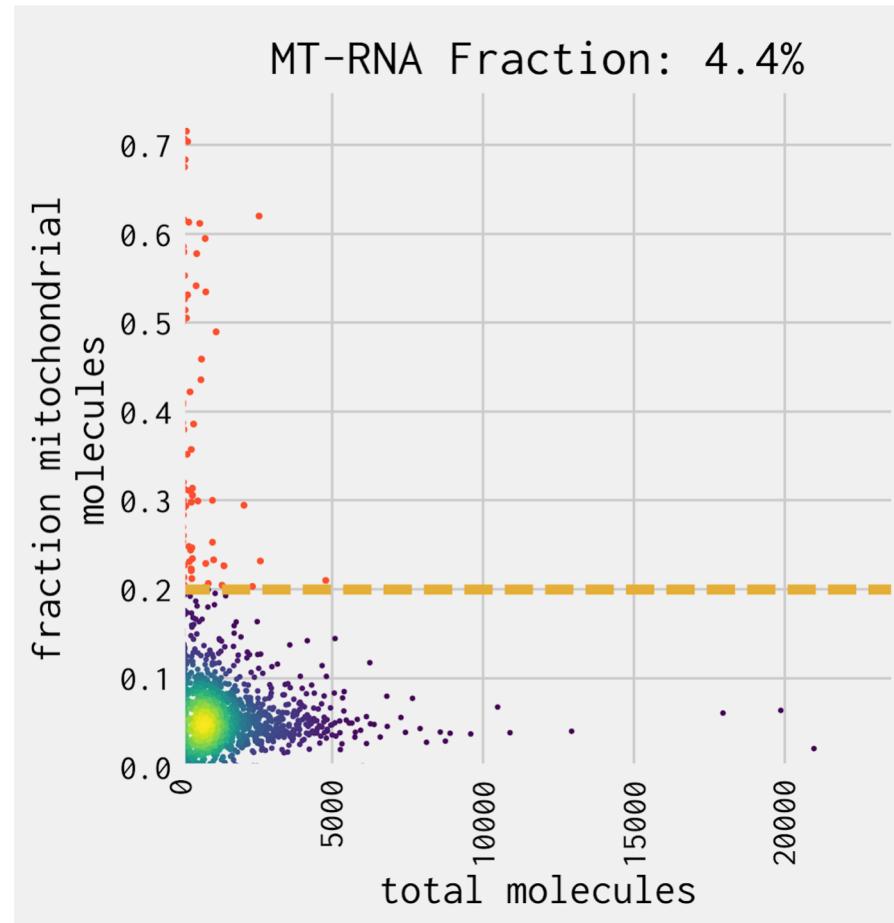
- Multiplets / Doublets are rare
- Constituent singlets are present



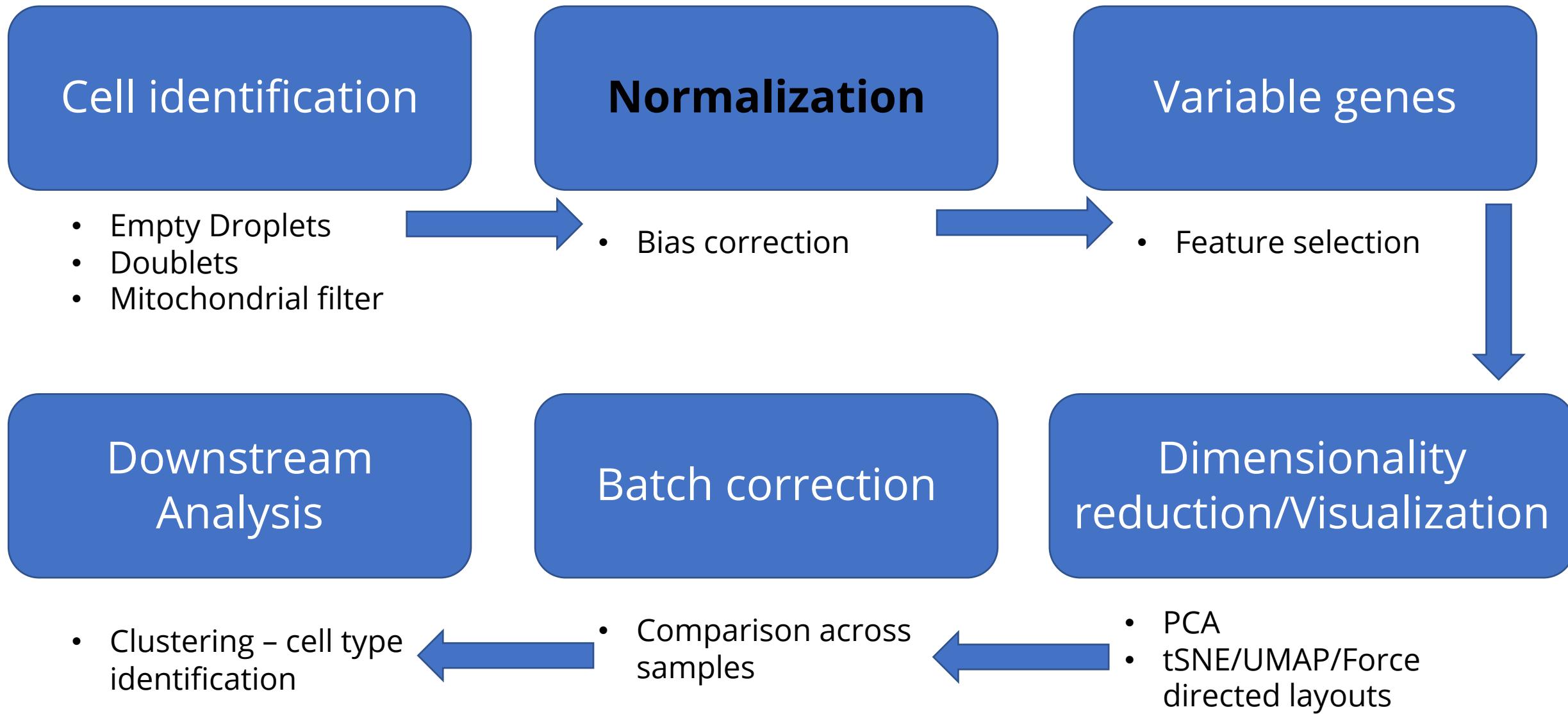
- Can this detect all possible multiplets?

Mitochondrial filter

- Calculate fraction of molecules from MT genes
- Exclude cells with > 20% (optional)

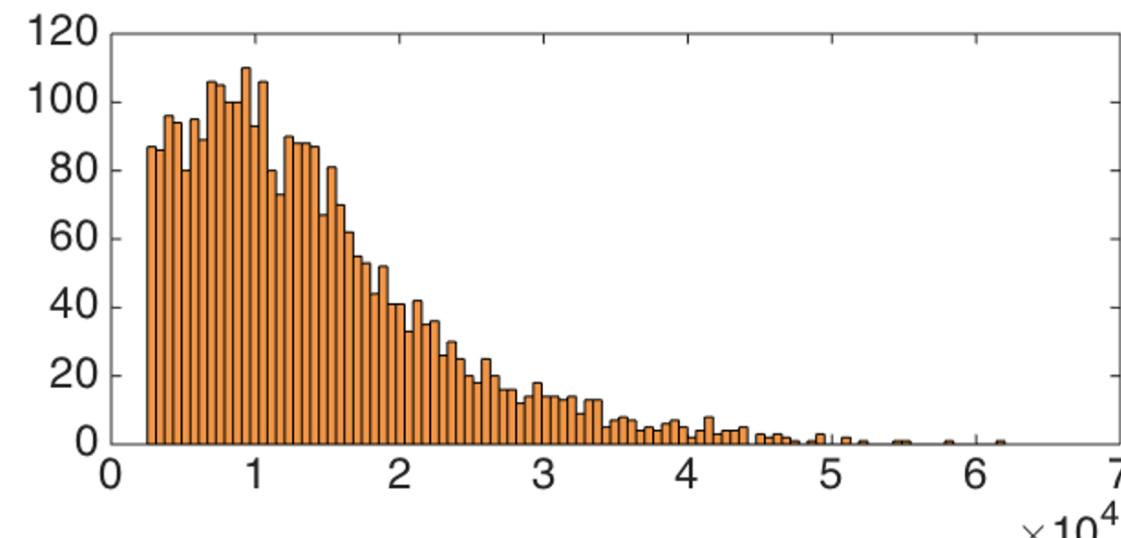


scRNA-seq analysis steps



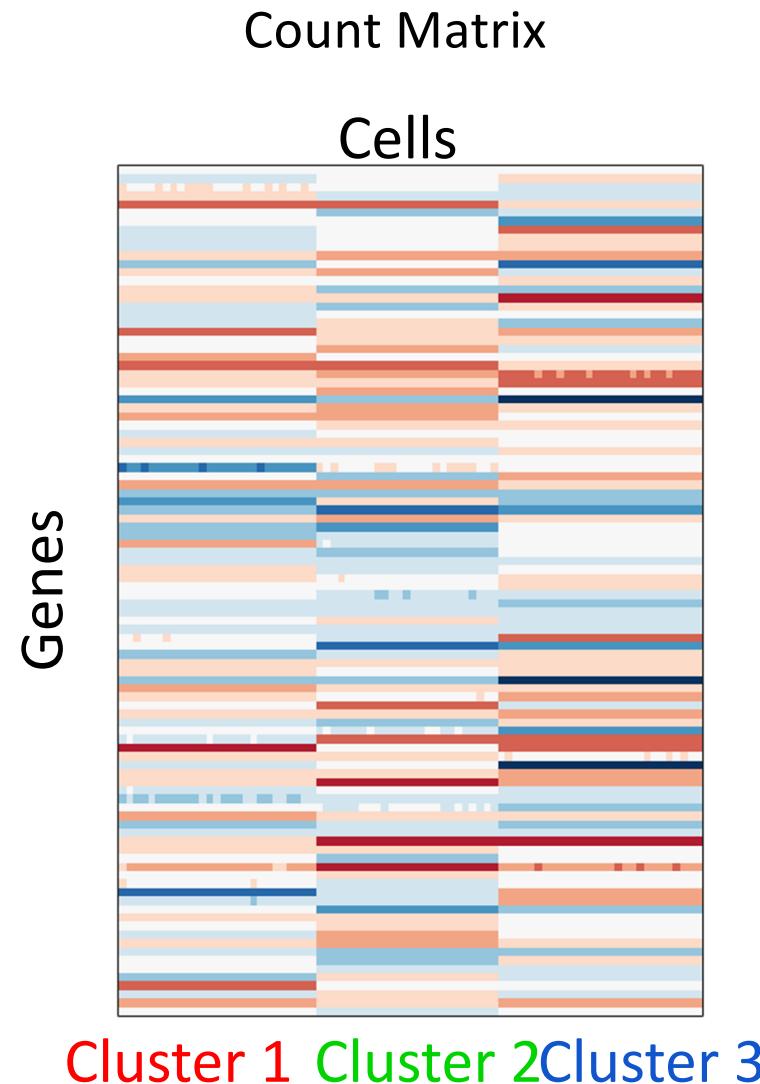
Why Normalize?

- Larger skew in distribution of total molecules (UMIs) per cell, i.e. library size
- Expression values not comparable across cells
 - *Measuring distance between cells*



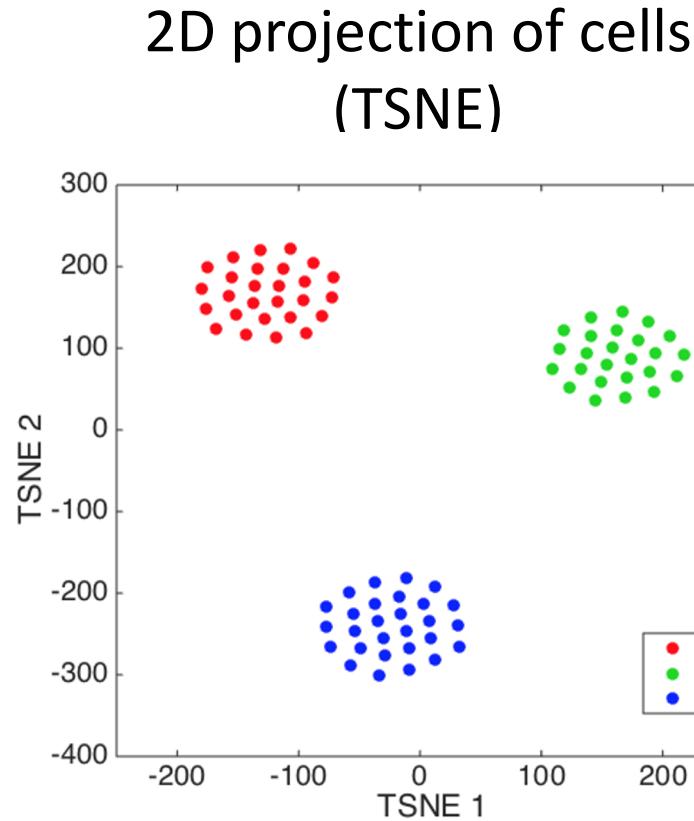
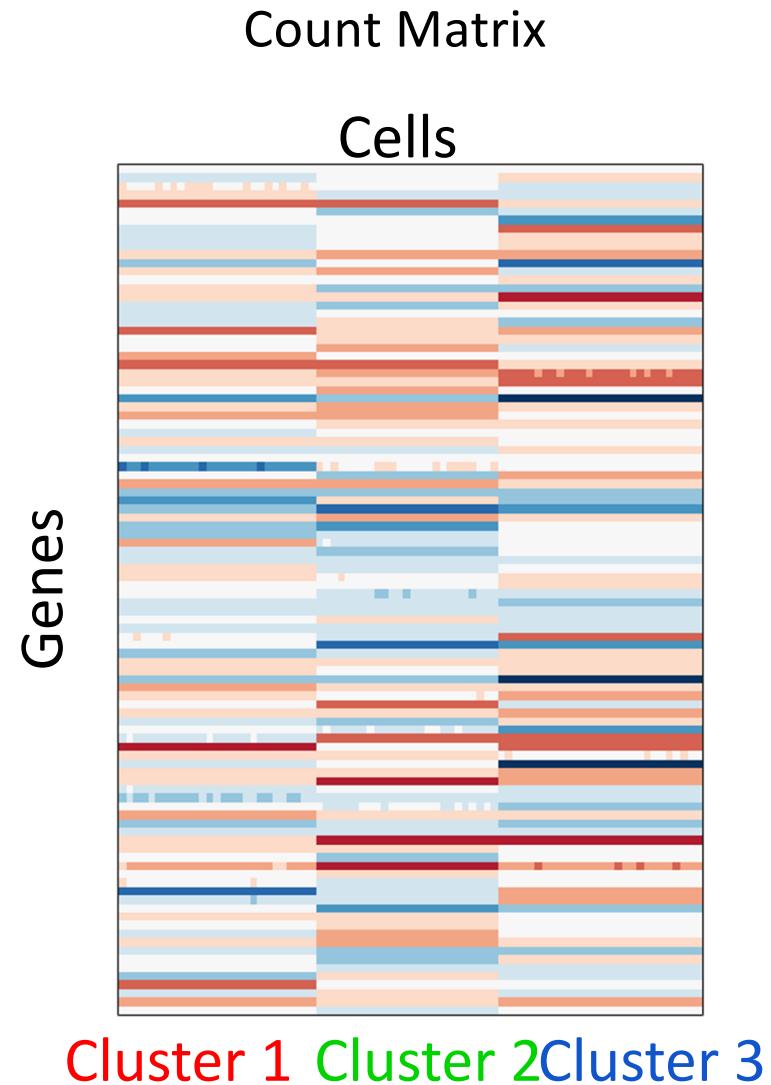
Histogram of UMI count in
example SC dataset
From Zeisel, Science 2014

Why is this problematic?

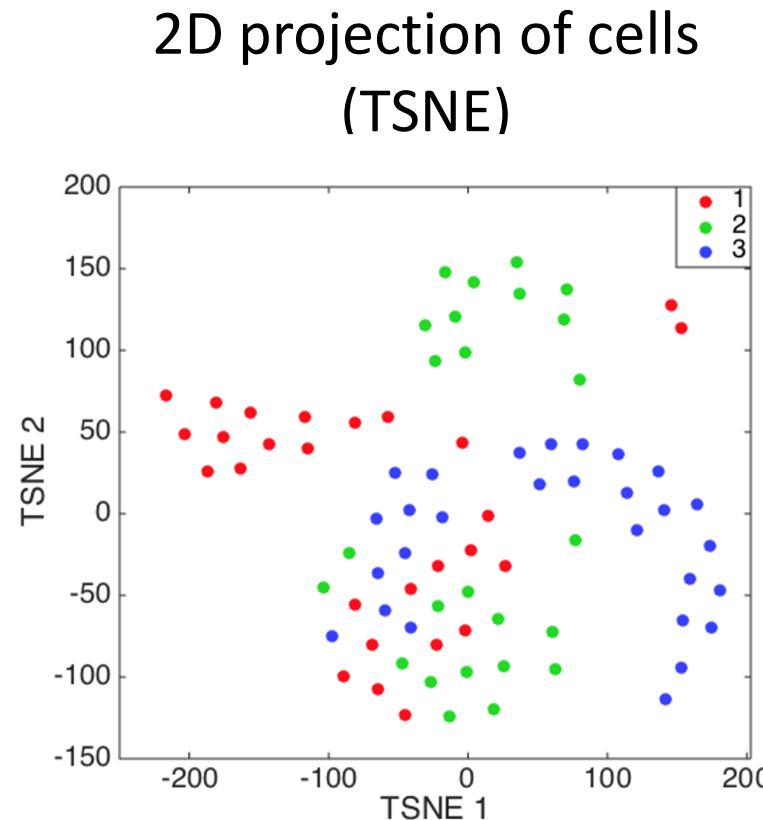
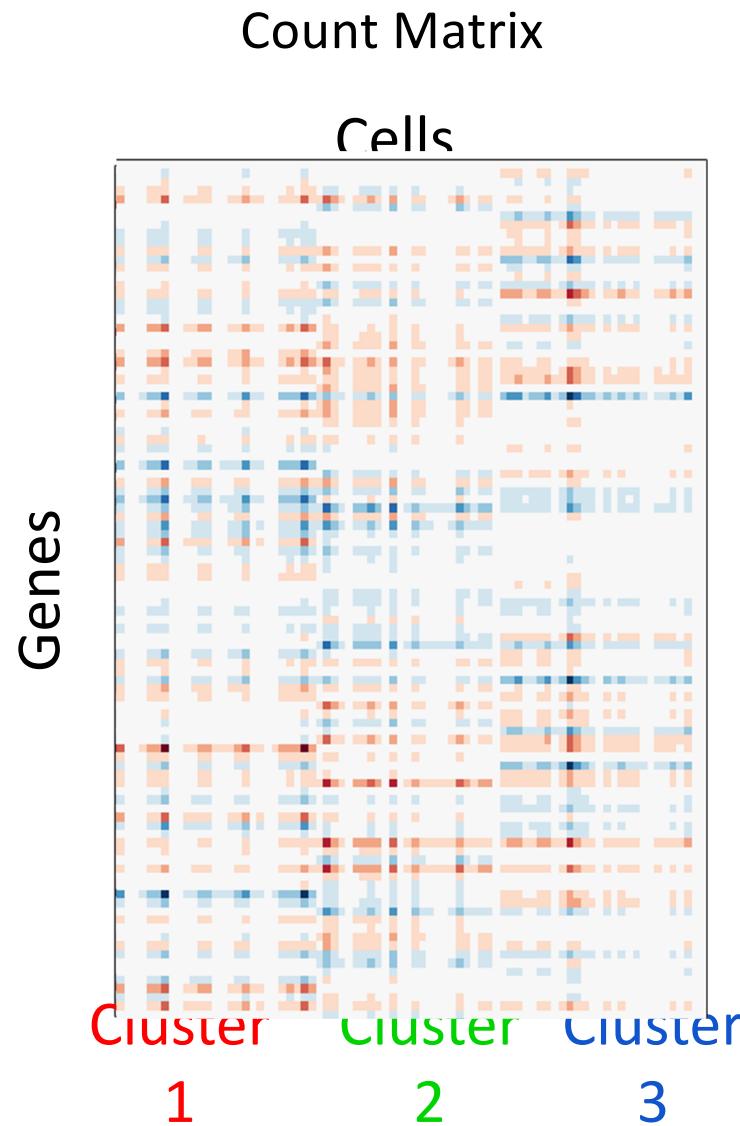


Slides courtesy of Elham Azizi

Why is this problematic?

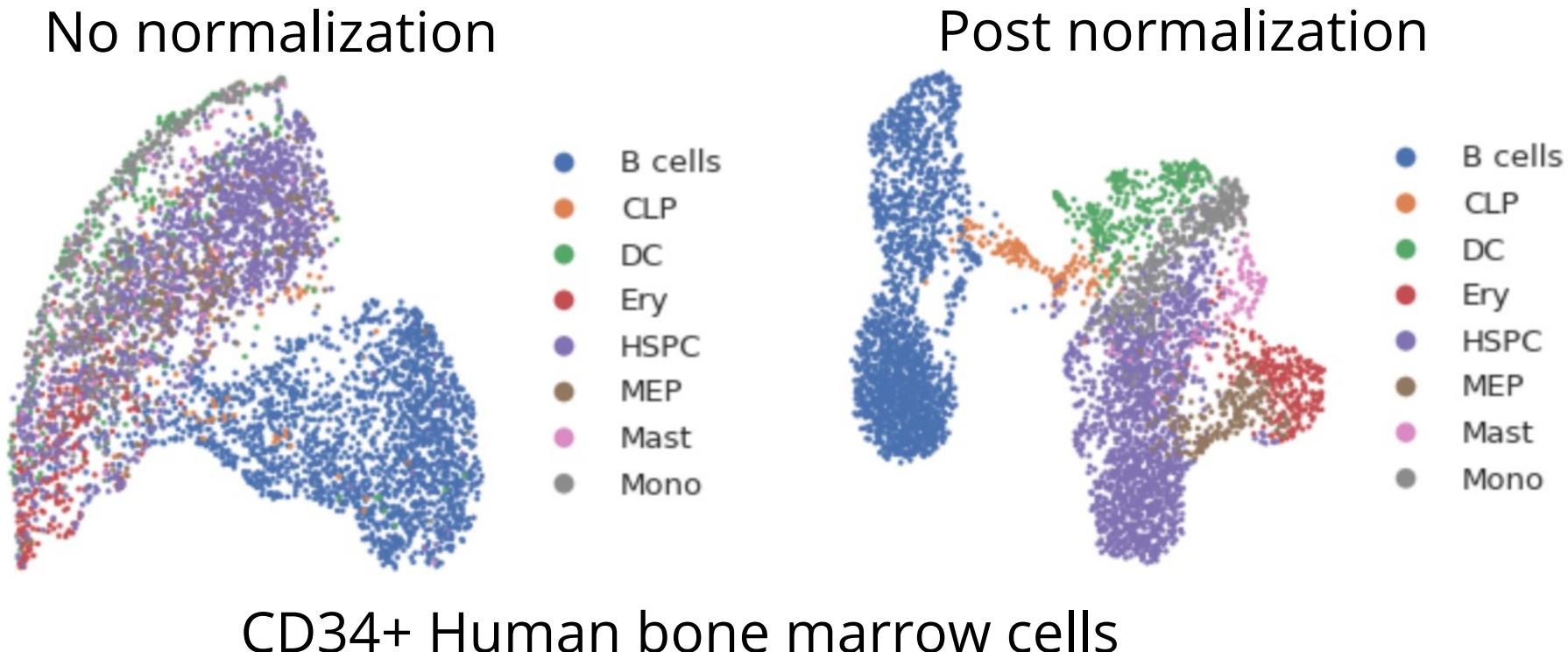


Why is this problematic?

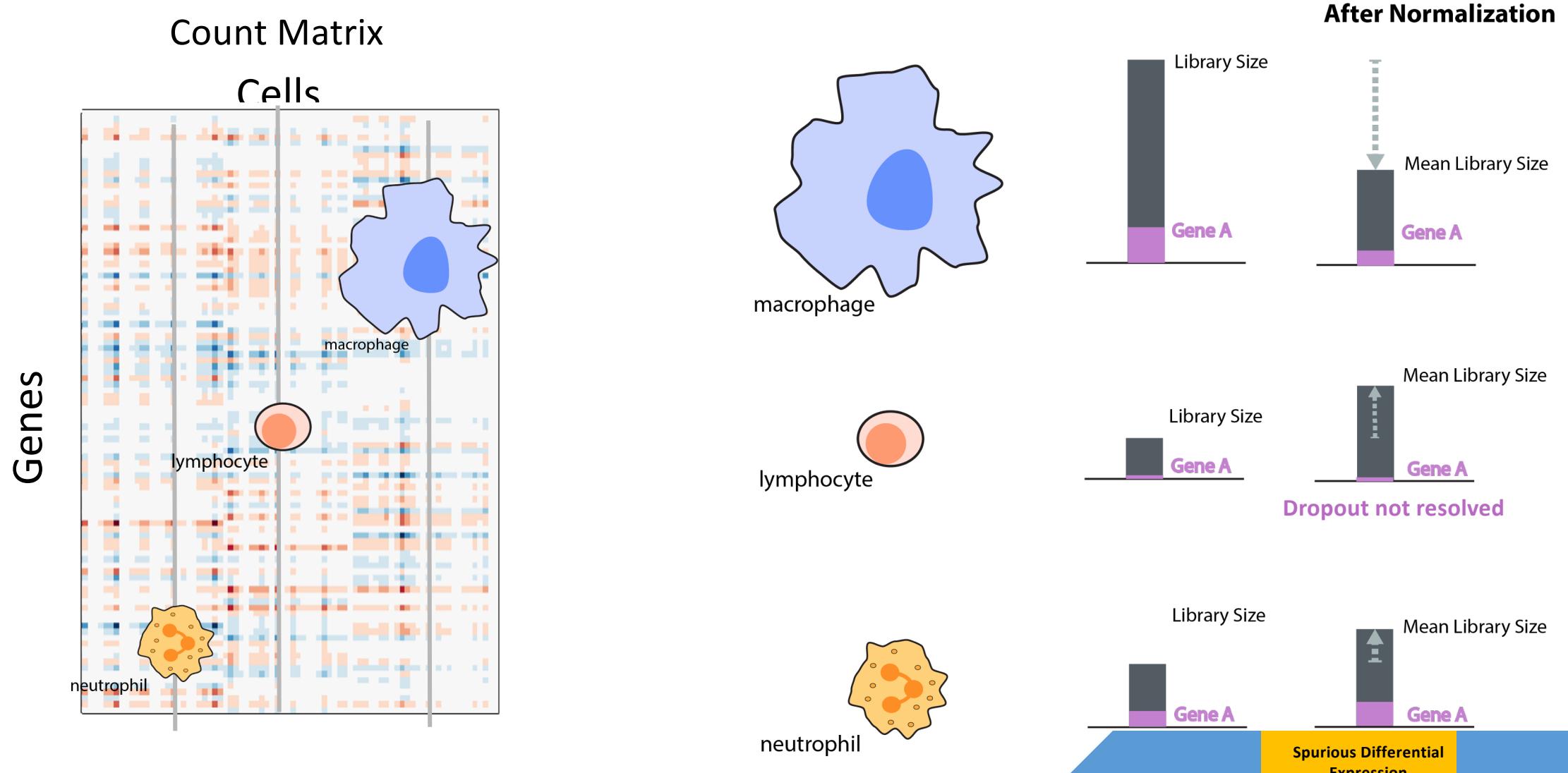


Normalization

- Global normalization:
 - Divide counts by total molecules in each cell
 - Multiply by median [To avoid numerical issues]
- Log transform of the data

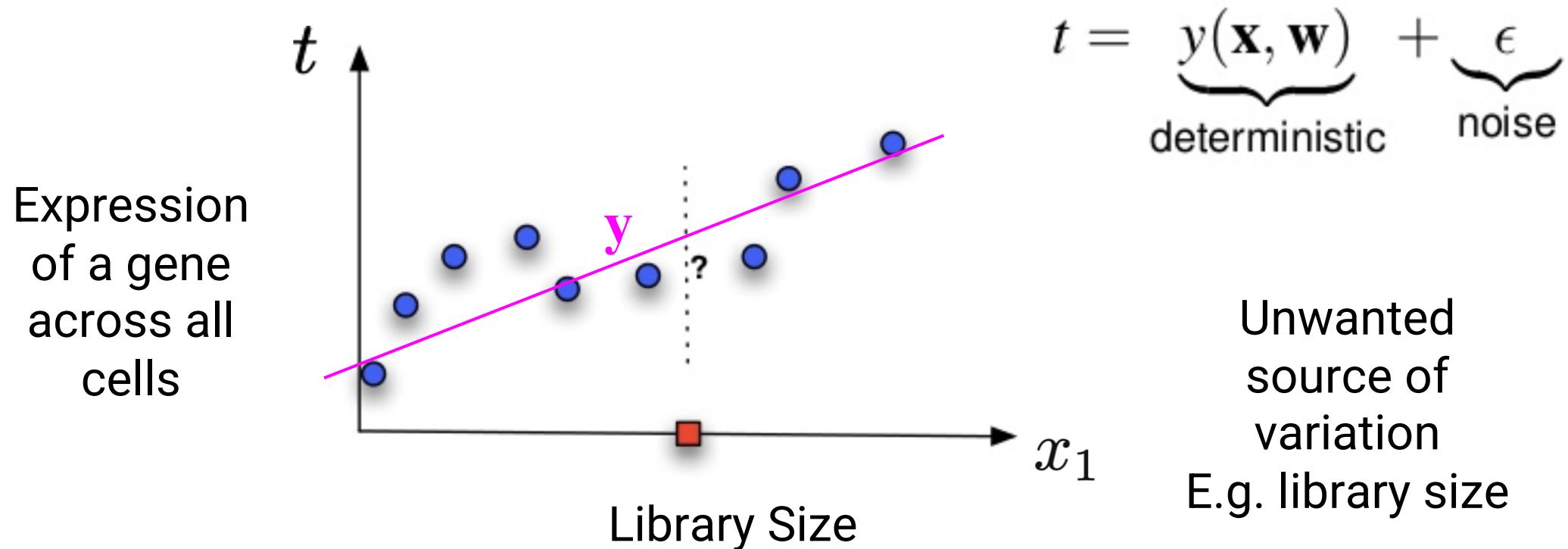


Normalizing by global scaling



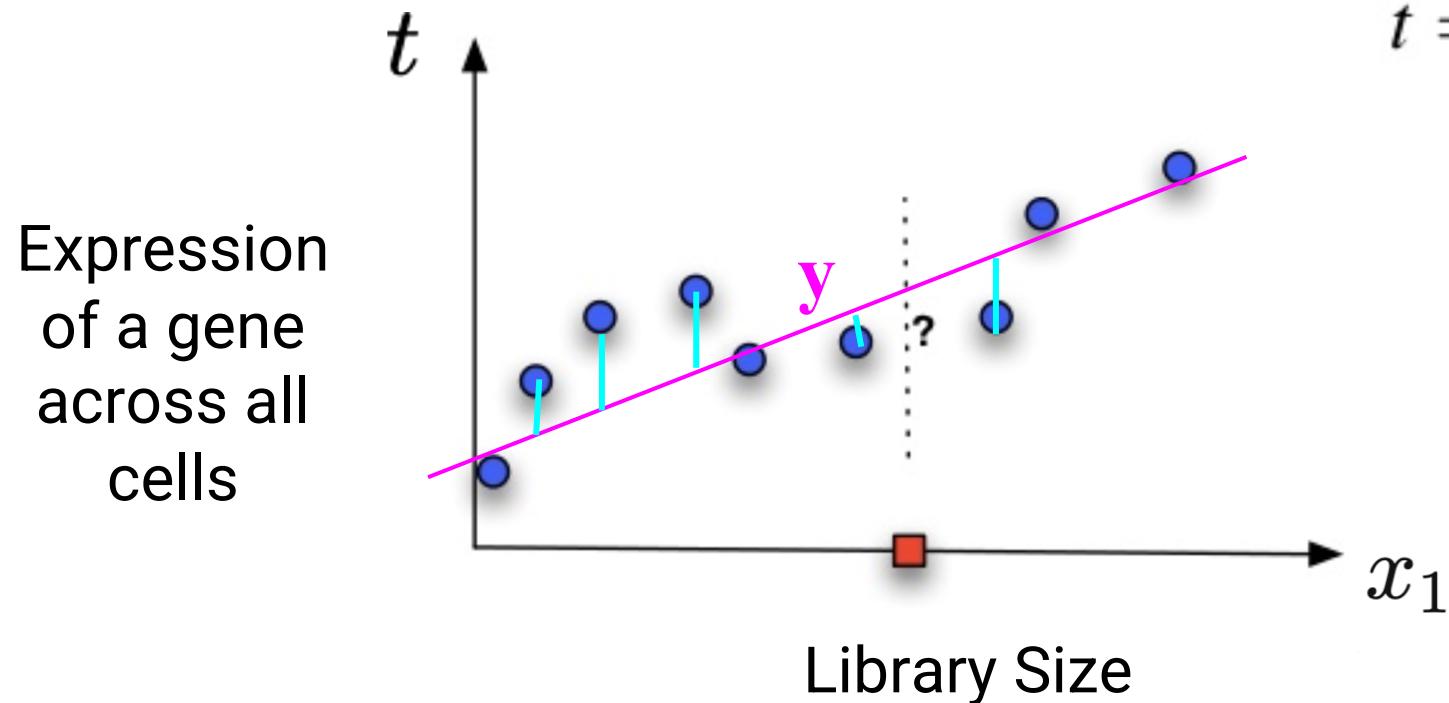
Simplistic Approach: Regressing out library size

- A simplistic approach to normalization $y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1$



Regressing out library size

- If noise is not small
 - Bad fit, x not predictive of t
 - **Where is the real biological variance?**



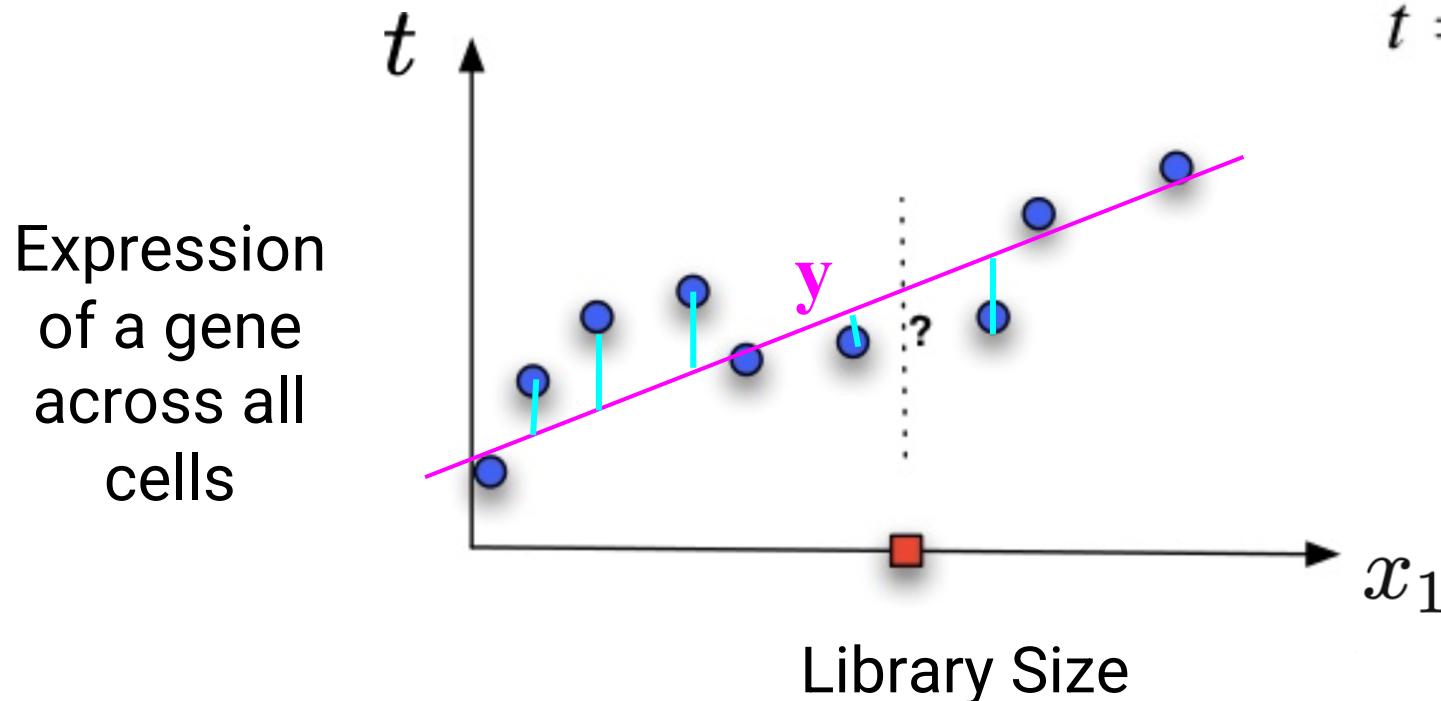
$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1$$

$$t = \underbrace{y(\mathbf{x}, \mathbf{w})}_{\text{deterministic}} + \underbrace{\epsilon}_{\text{noise}}$$

Unwanted source of variation
E.g. library size

Regressing out library size

- If noise is not small
 - Bad fit, x not predictive of t
 - Where is the real biological variance?



$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1$$

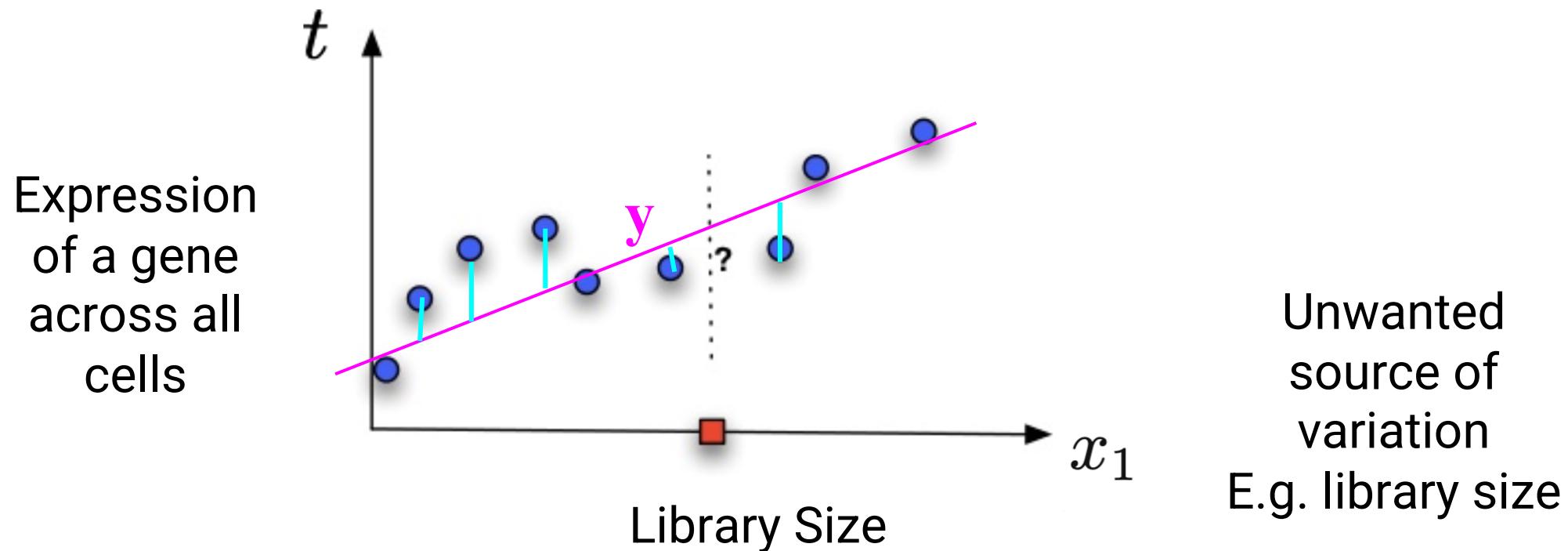
$$t = \underbrace{y(\mathbf{x}, \mathbf{w})}_{\text{deterministic}} + \underbrace{\epsilon}_{\text{noise}}$$

Unwanted source of variation
E.g. library size

Regressing out library size

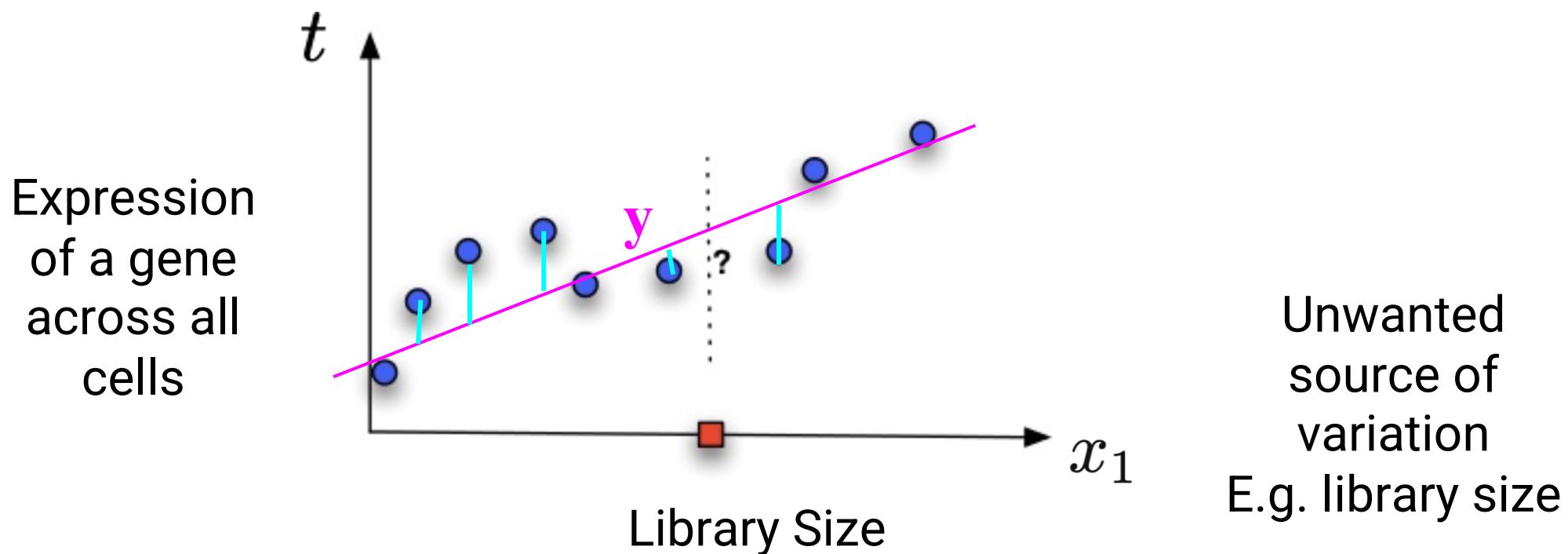
- Model error:

$$\epsilon = t - \hat{y} = t - w_0 - w_1 x_1$$



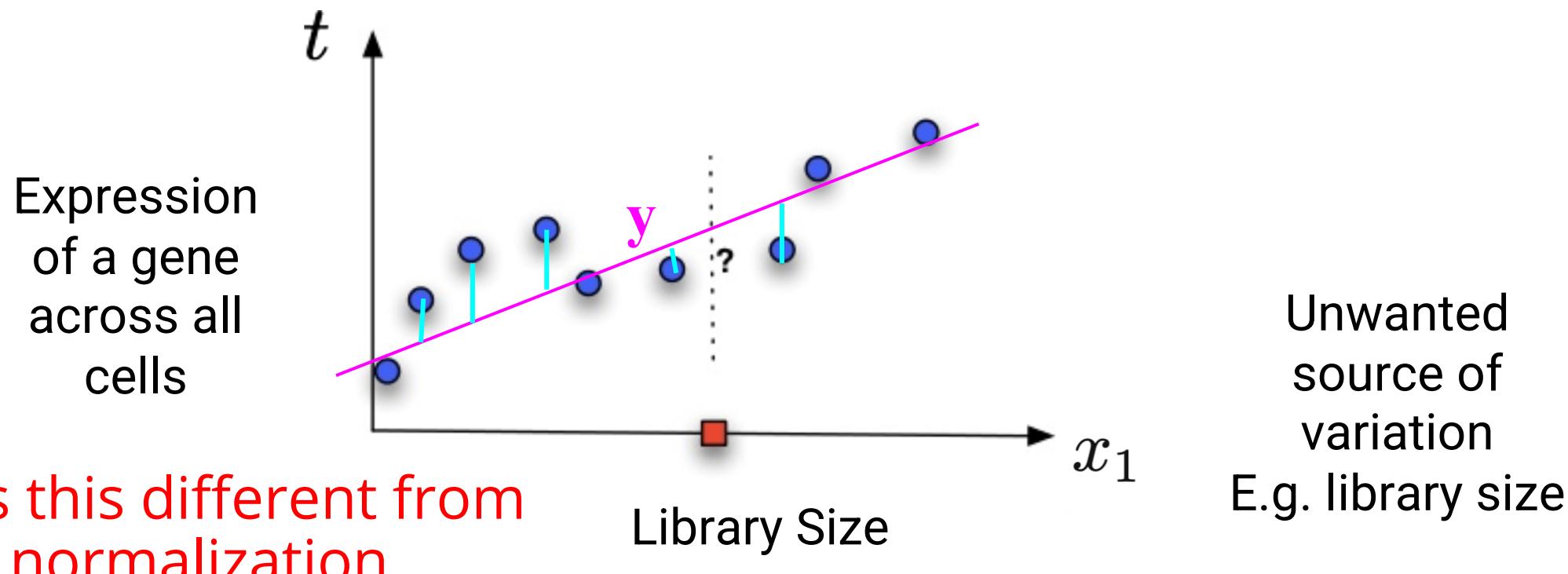
Regressing out library size

- Model error: $\epsilon = t - \hat{y} = t - w_0 - w_1 x_1$
- Corrected expression: $t_{adj} = t - w_1 x_1 = \epsilon + w_0$



Regressing out library size

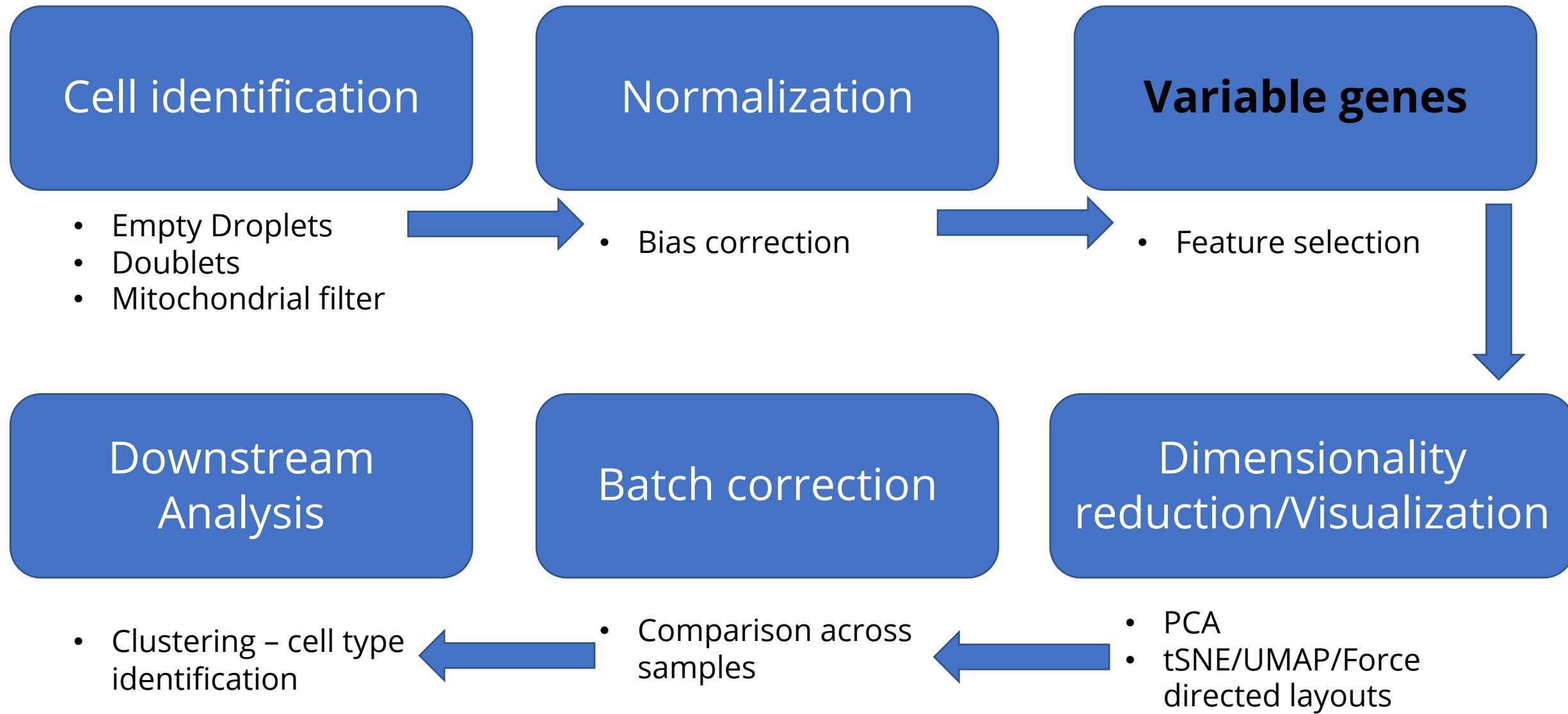
- Model error: $\epsilon = t - \hat{y} = t - w_0 - w_1 x_1$
- Corrected expression: $t_{adj} = \epsilon - w_1 x_1 = \epsilon + w_0$



Normalization

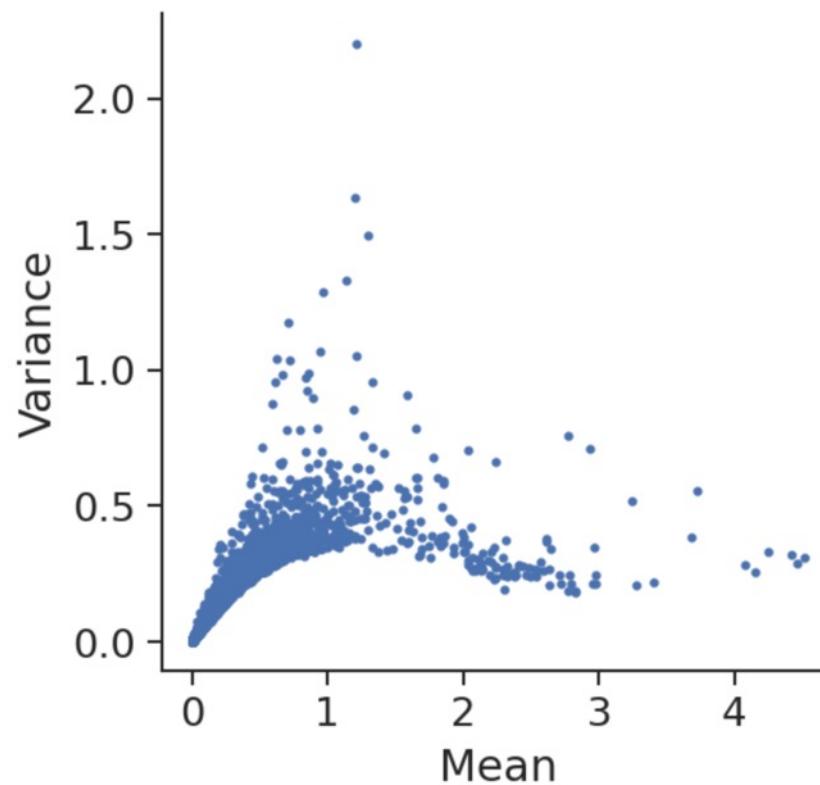
- SCNorm
- Sctransform
- Gamma Regression Model (GRM)
- scRAN
- BASiCS
- Linnorm
 -

scRNA-seq analysis steps



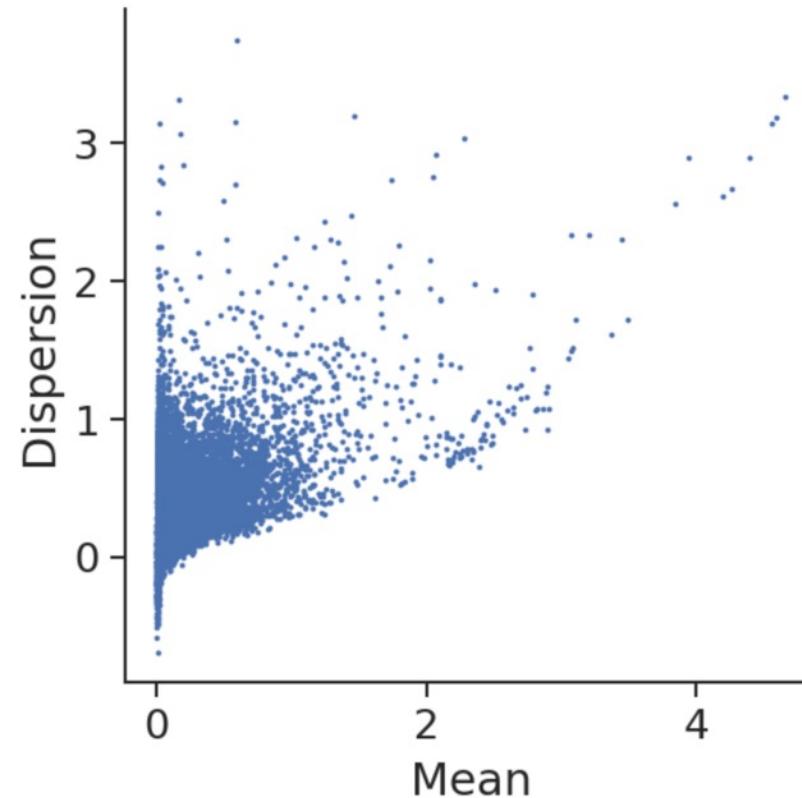
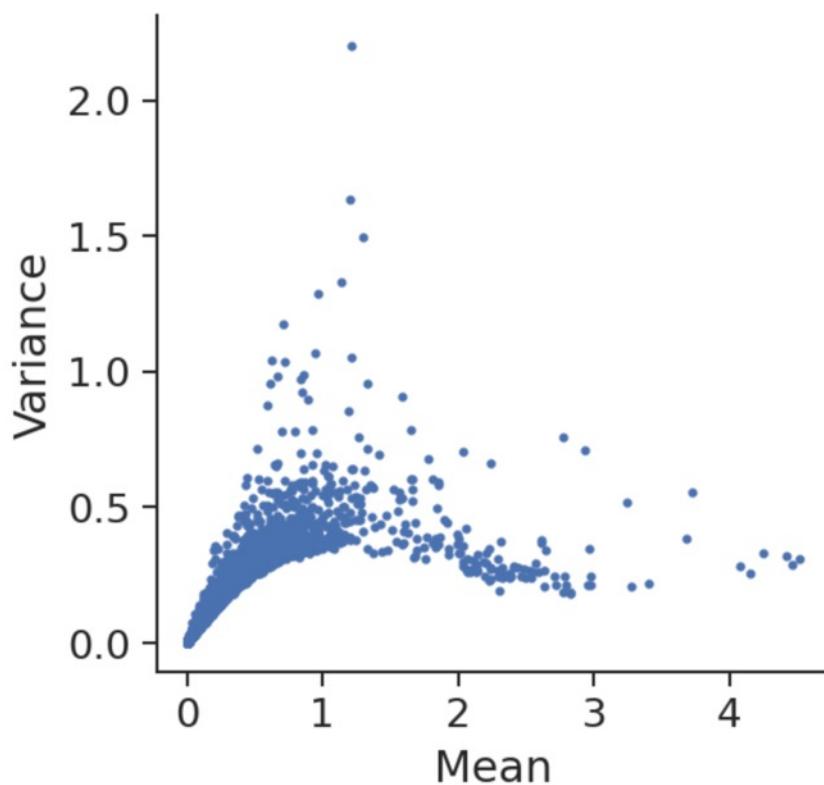
Feature selection

- Cell states are defined by expression of subsets of genes
- Goal of feature selection: Select genes that inform the biology rather than genes that represent random noise
- Possible Solution: How variable is the gene across cells?



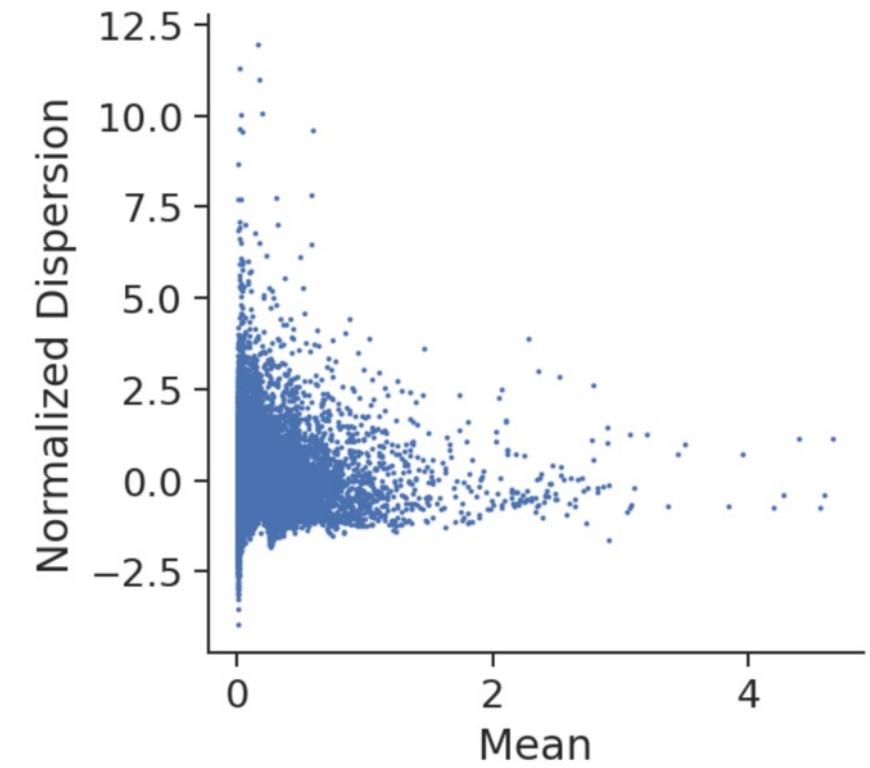
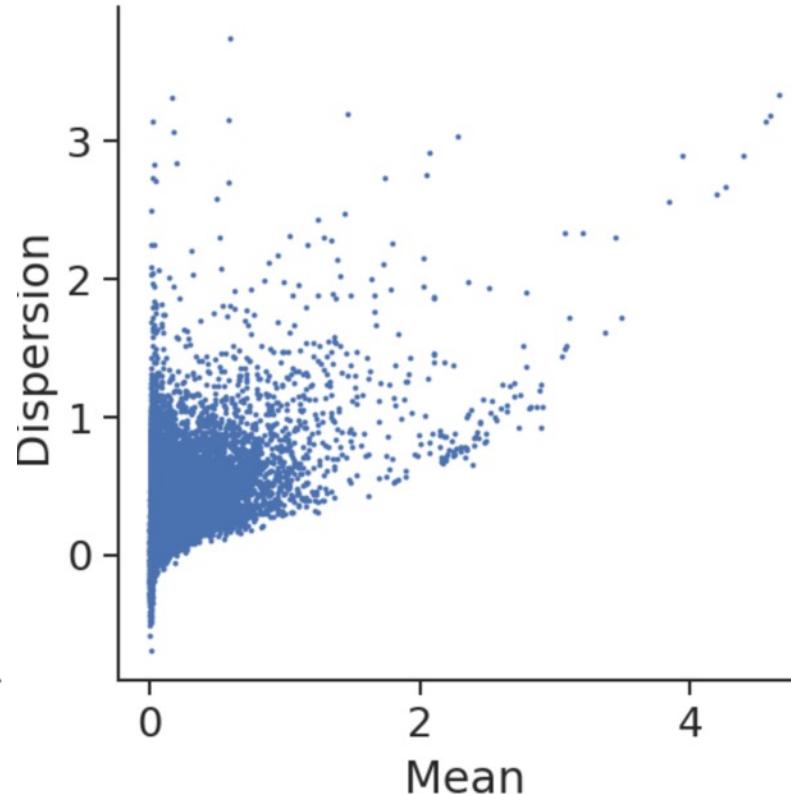
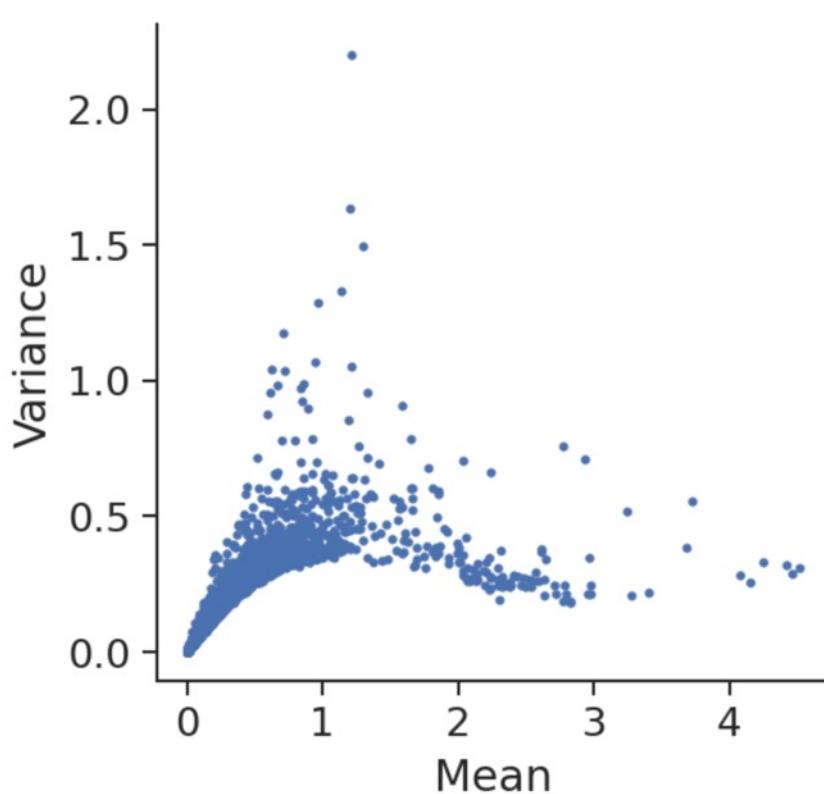
Feature selection

$$Dispersion = \frac{Variance}{Mean}$$



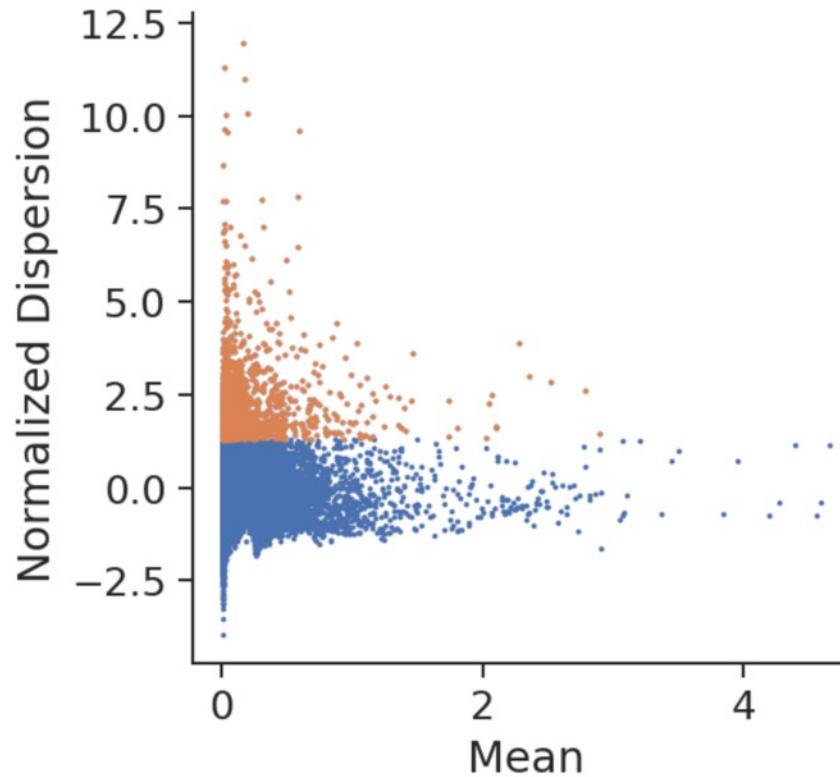
Feature selection: Share information across genes

$$\begin{aligned} \text{Normalized Dispersion} \\ = \frac{\text{Dispersion} - \text{Mean}(\text{Dispersion in expression mean bin})}{\text{Std}(\text{Dispersion in expression mean bin})} \end{aligned}$$



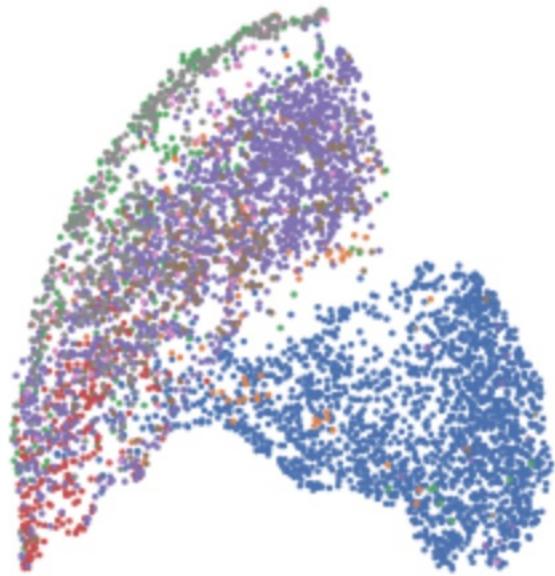
Feature selection: Share information across genes

$$\begin{aligned} \text{Normalized Dispersion} \\ = \frac{\text{Dispersion} - \text{Mean}(\text{Dispersion in expression mean bin})}{\text{Std}(\text{Dispersion in expression mean bin})} \end{aligned}$$

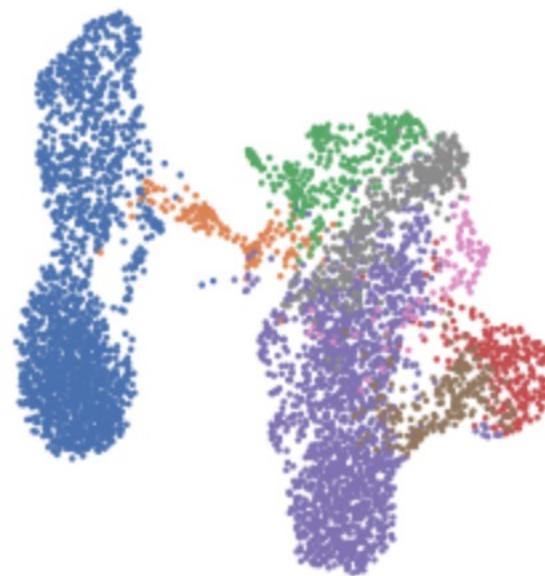


Feature selection: highly variable genes

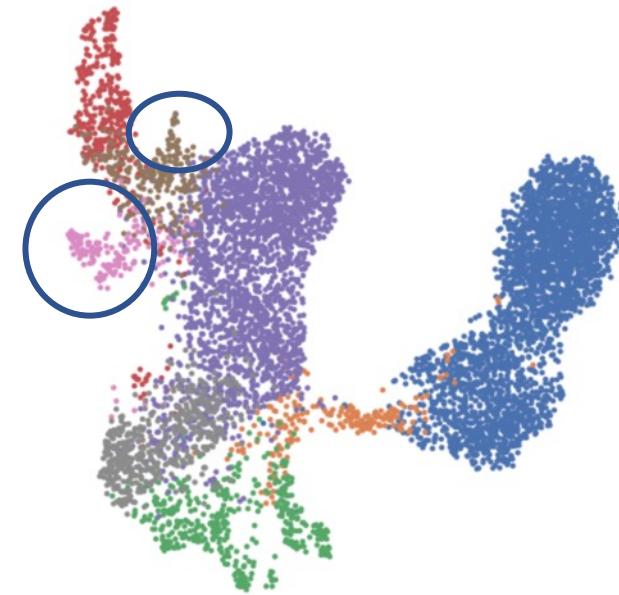
No normalization



Post normalization

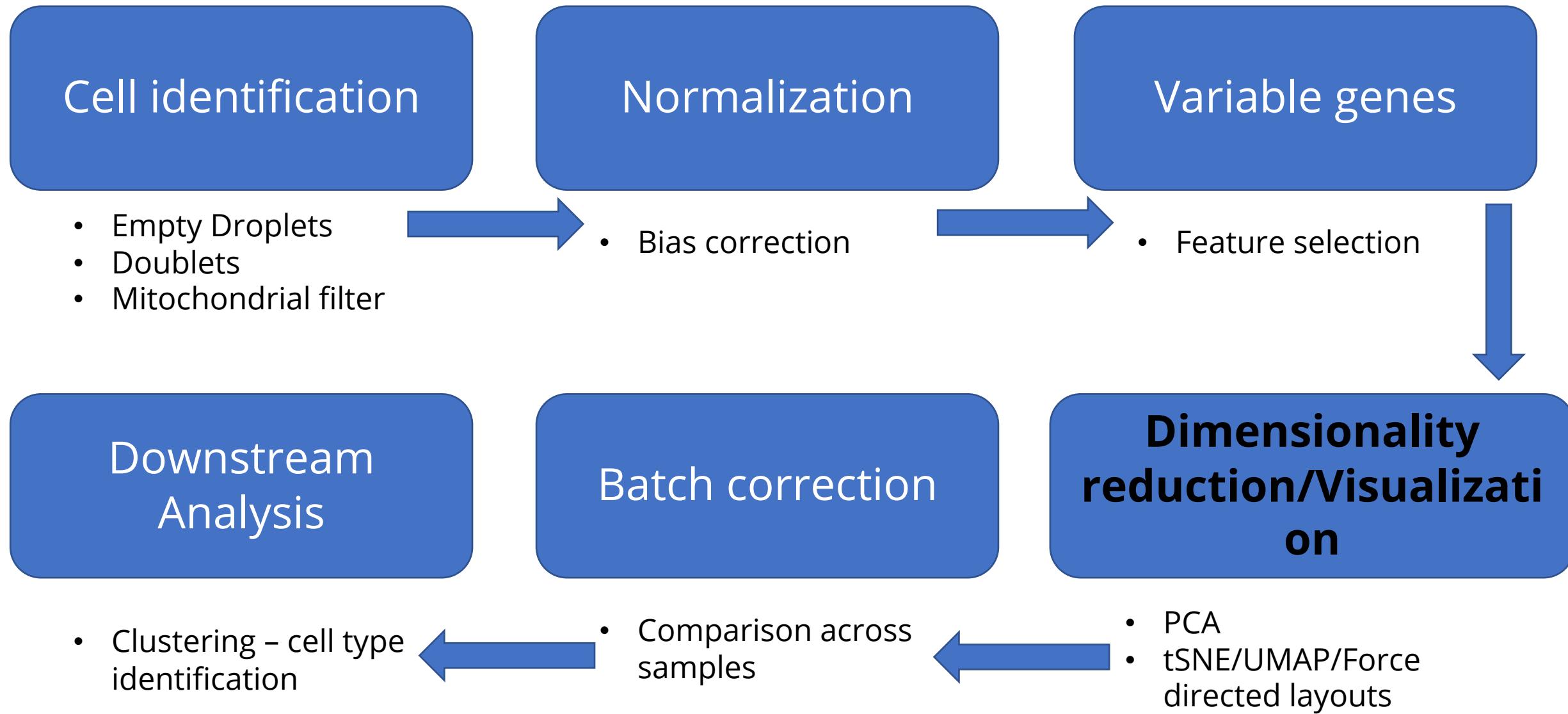


Feature selection



CD34+ Human bone marrow cells

scRNA-seq analysis steps

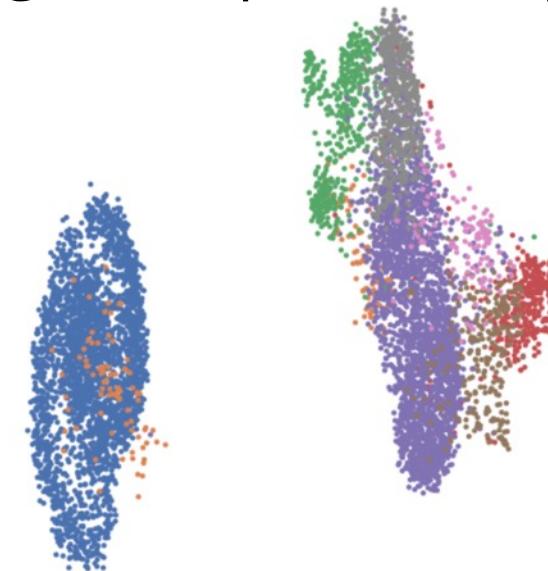


Measuring distance between cells

- Possible Solution: Euclidean distance between normalized, selected genes



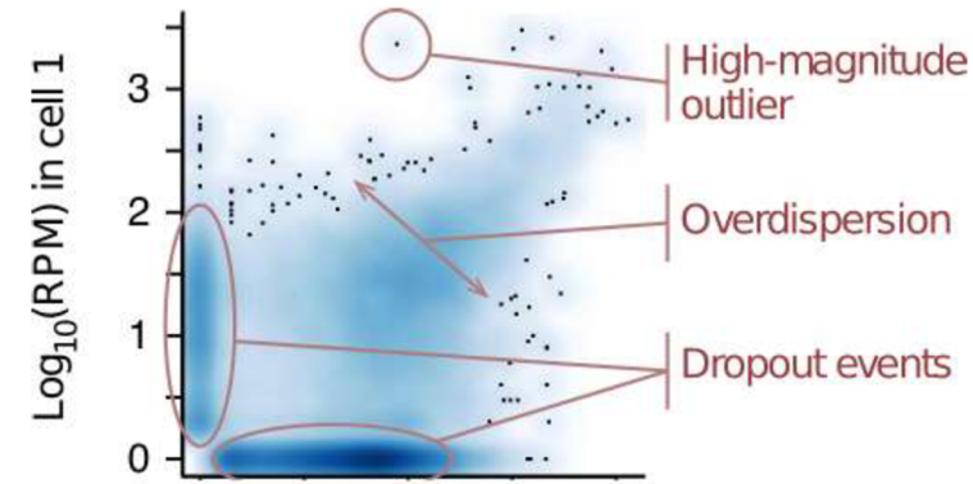
Euclidean distance
In gene expression space



CD34+ Human bone marrow cells

Single-cell RNA Noise: Dropouts

- ~5-10% of transcripts in a cell are captured
- Further loss during reverse transcription
- Genes with higher expression have fewer zeros
- Non-zero values are also underestimates of true counts



BRIEF COMMUNICATIONS

Bayesian approach to
single-cell differential
expression analysis



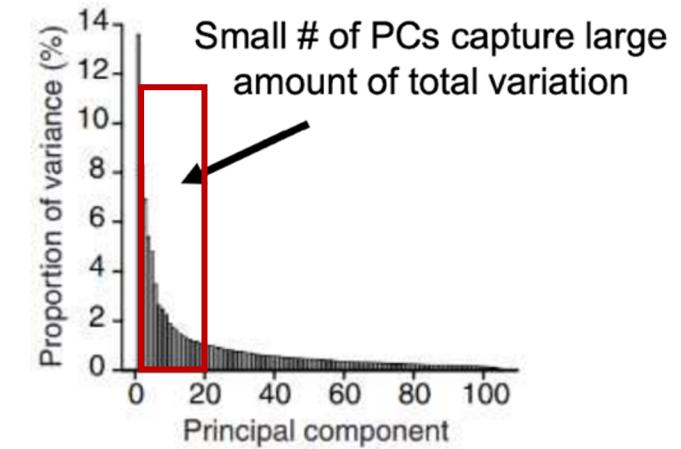
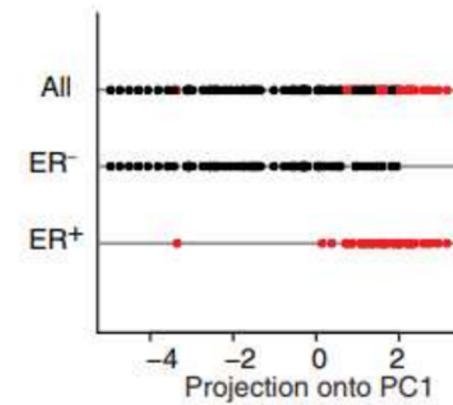
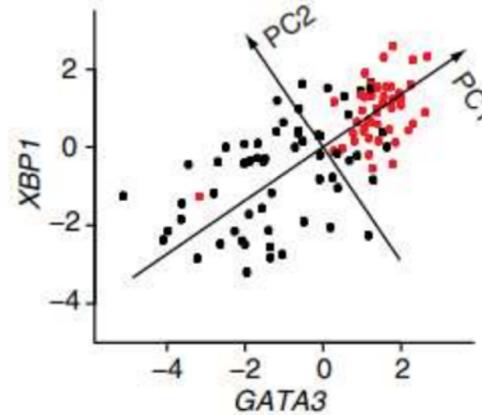
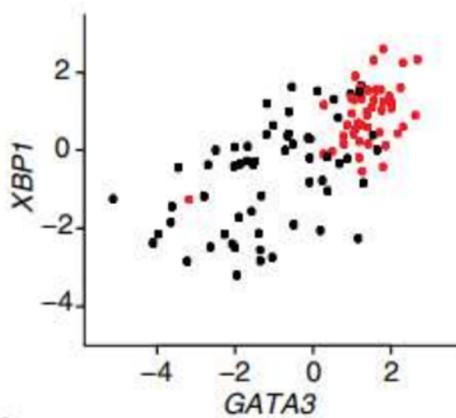
Peter V Kharchenko¹⁻³, Lev Silberstein³⁻⁵ &
David T Scadden³⁻⁵

© 2014 Nature America, Inc.

Alternative solution: PCA

- Cell states are defined by co-regulated gene modules
- PCA as a proxy to identify these genes modules

PCA effectively defines new axes through the data that capture the highest amount of variation possible



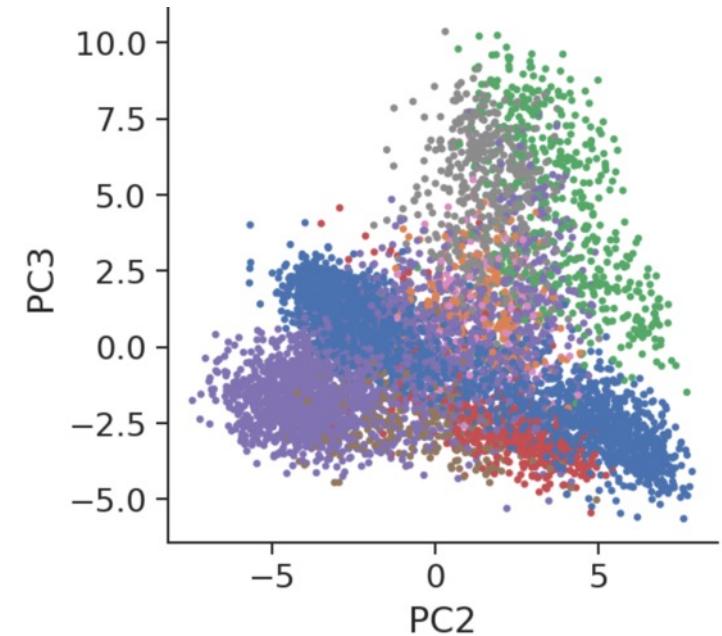
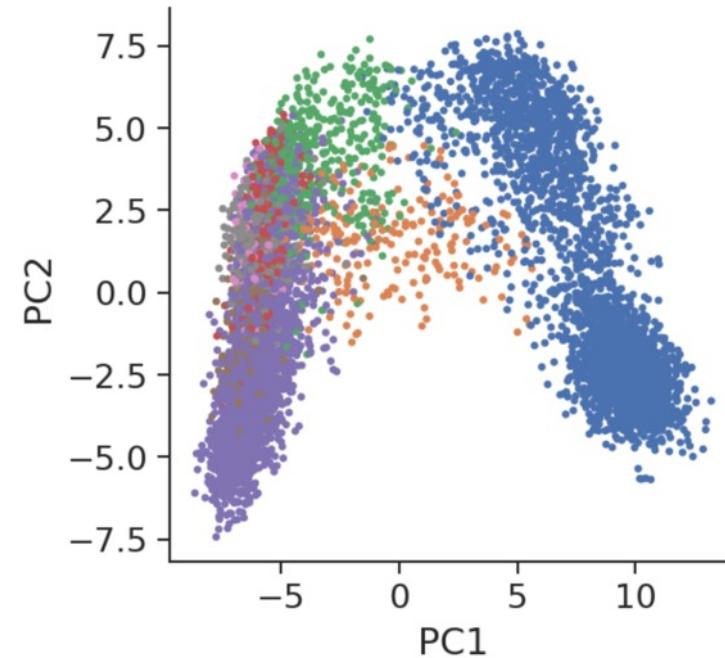
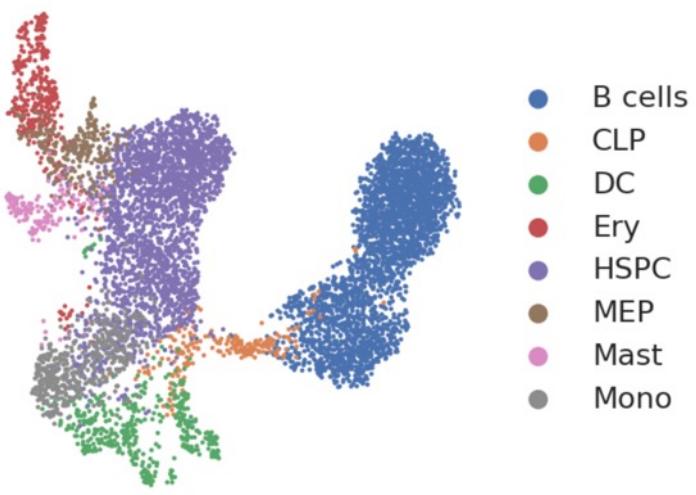
Ringér, *Nature Biotech*, 2008

- Selection of subset of PCs: Dimensionality reduction

Dimensionality Reduction

- Reduce the number of dimensions of data while preserving high dimensional information
- Overcome noise in high dimensions
- Computational efficiency
- Visualization

PCA for visualization

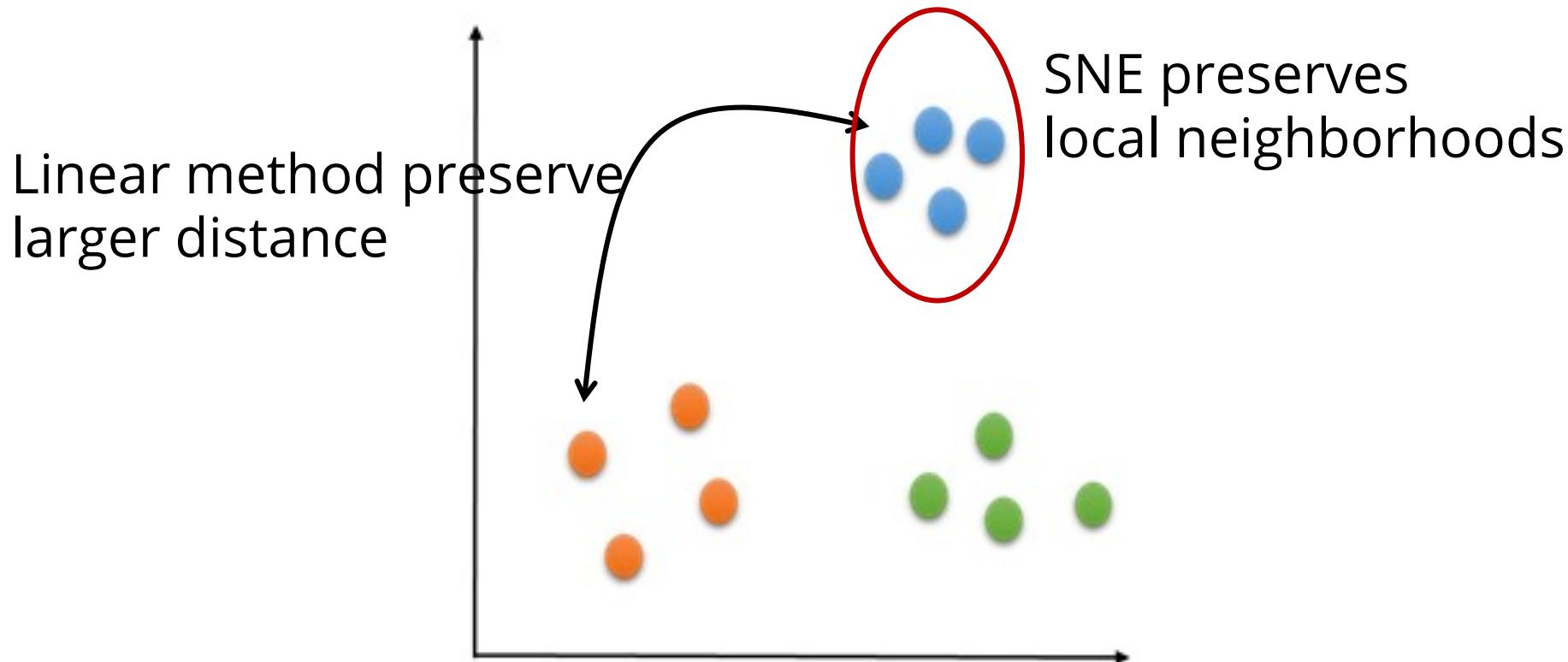


CD34+ Human bone marrow cells

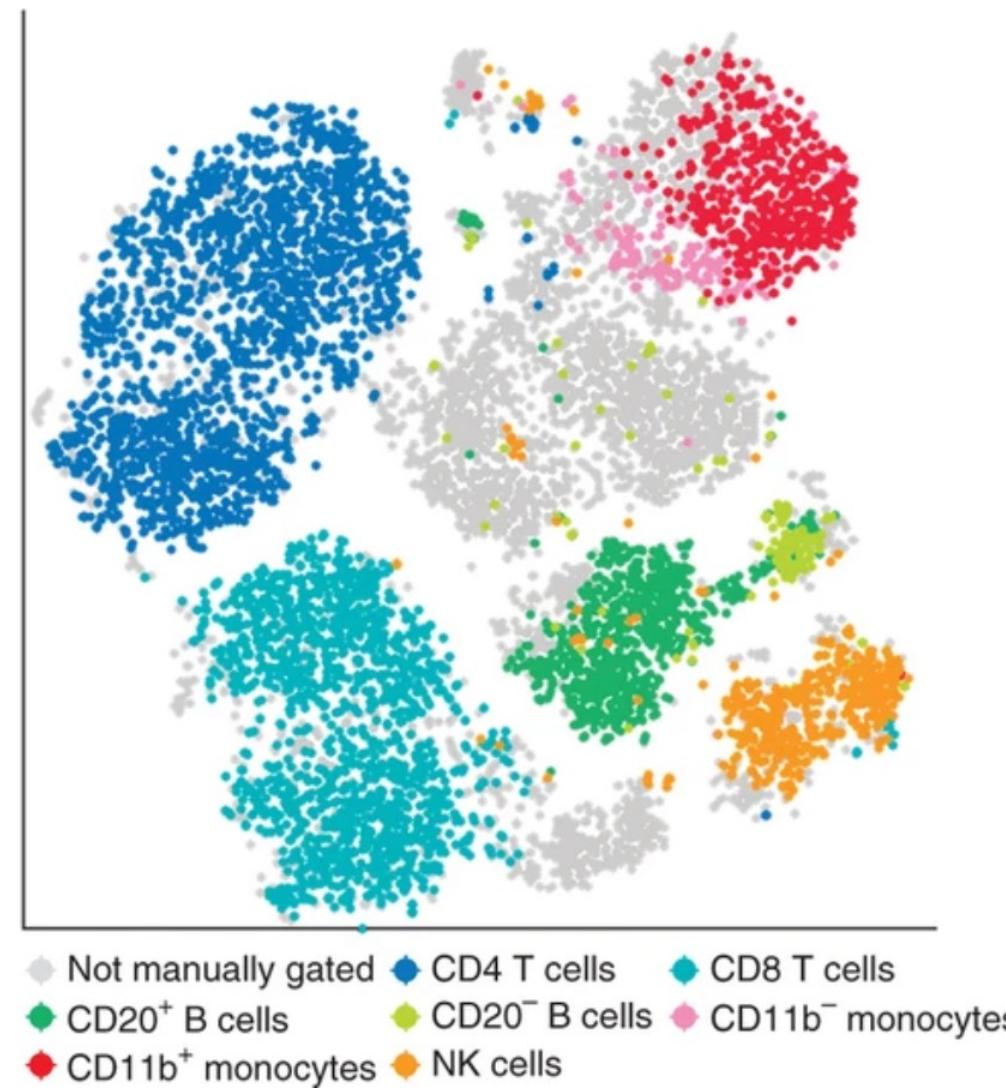
- Two axes can only capture so much information
 - Not explicitly modeled to capture as much of higher order information in 2D

Dimensionality Reduction for Visualization: tSNE

- SNE: Stochastic Neighborhood Embedding
- Goal: Compute a low dimensional representation that best preserves the local neighborhoods of cells

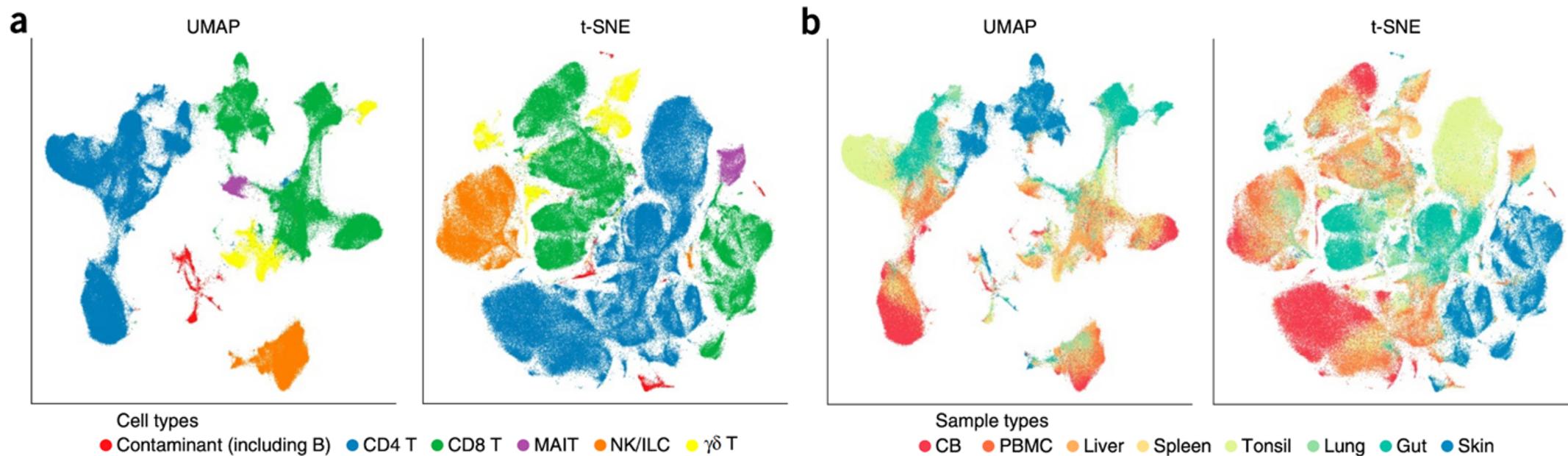


tSNE for single-cells



Visualization: UMAP

- Claim: Better preservation of global structure compared to tSNE while also preserving local structure

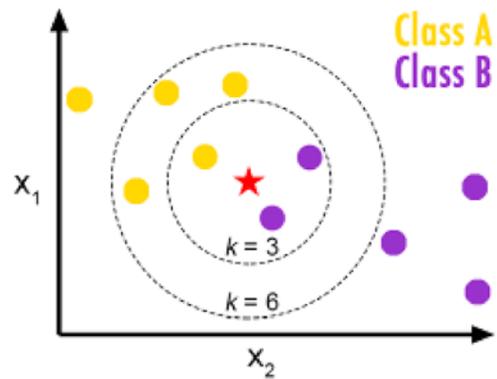


Visualization: Force directed layouts

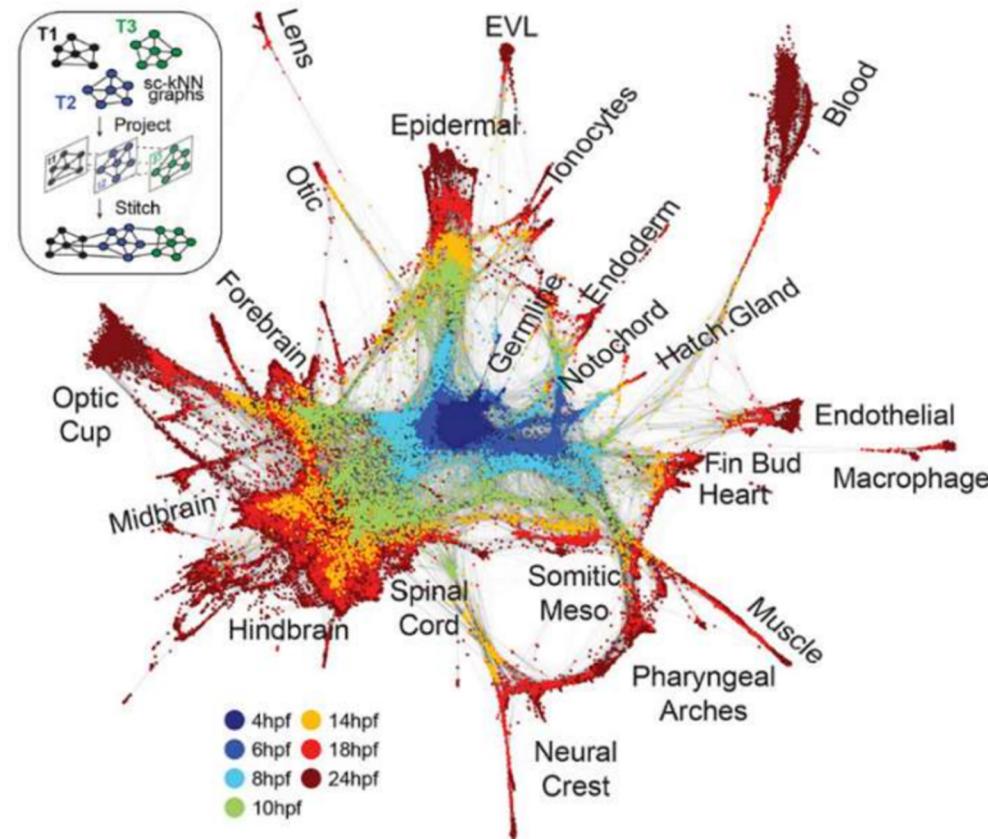
Goal: To visualize the structure of our data

Force-directed graphs

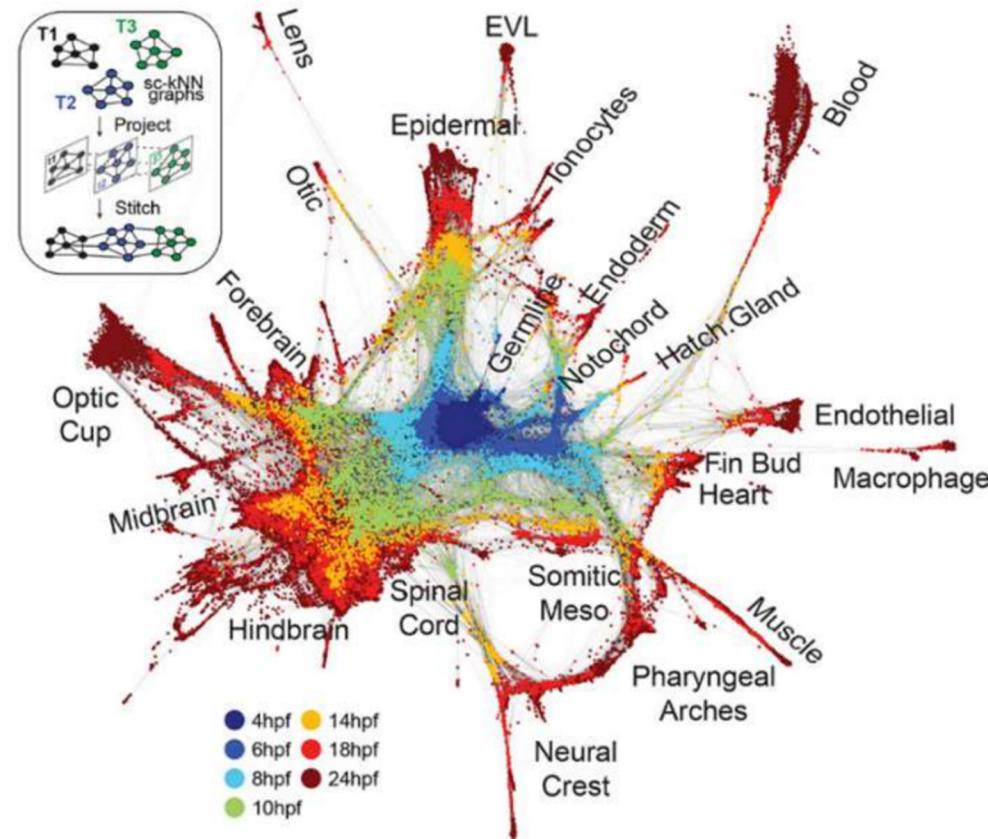
Visualize cells based on nearest neighbor structures



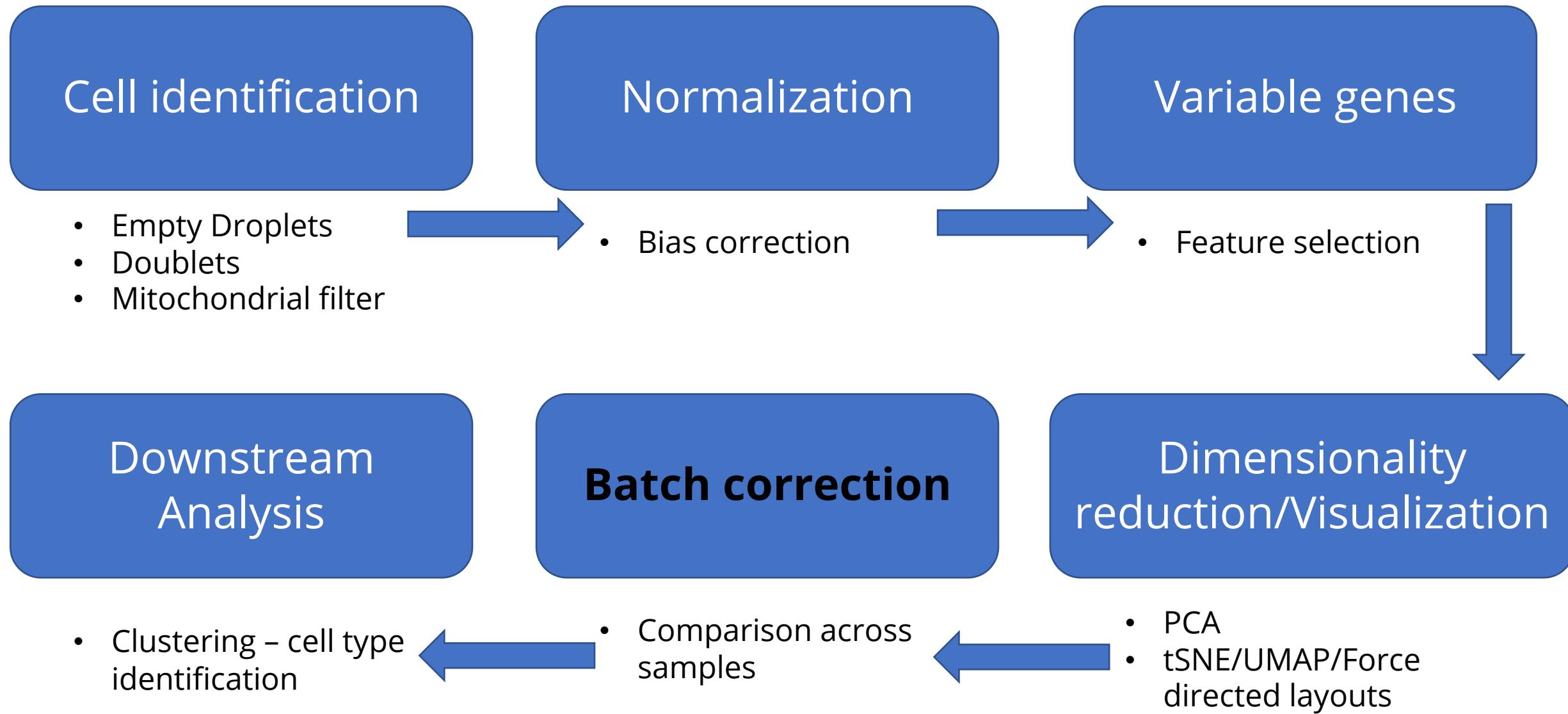
Repulsion between nodes
Attractive forces added
to edges connecting
nodes (spring functions)



Wagner et al., Science, 2018

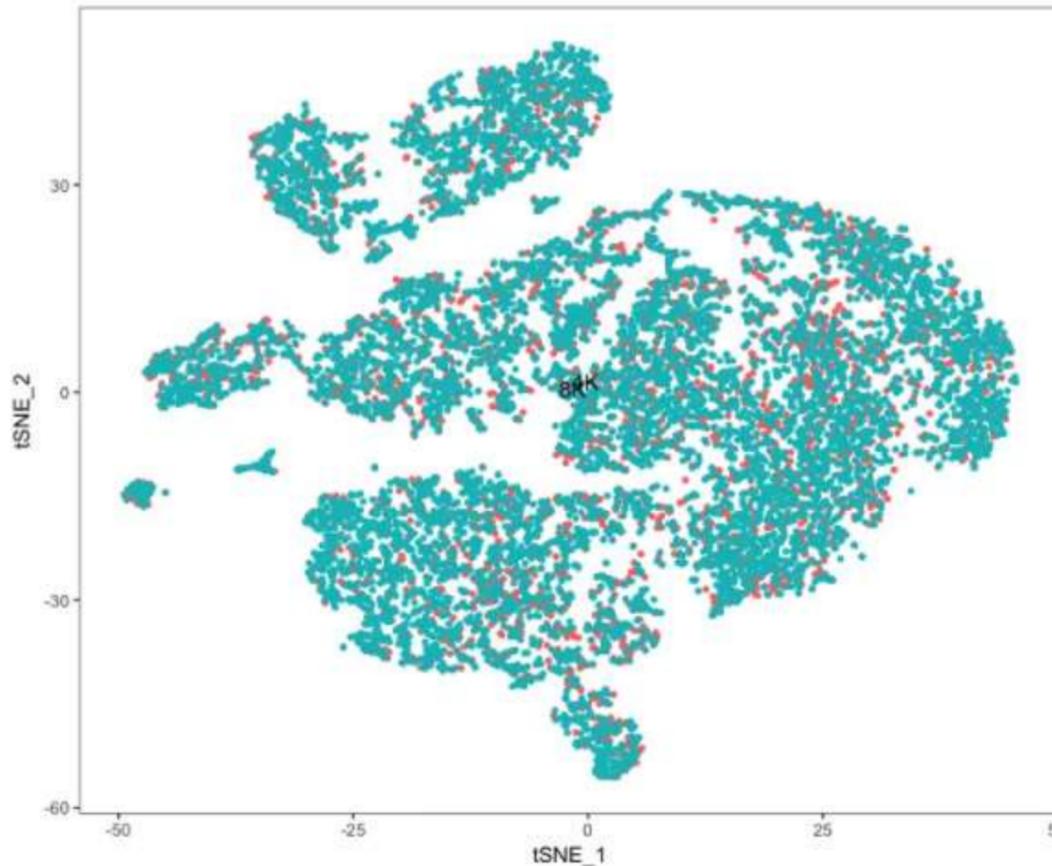


scRNA-seq analysis steps

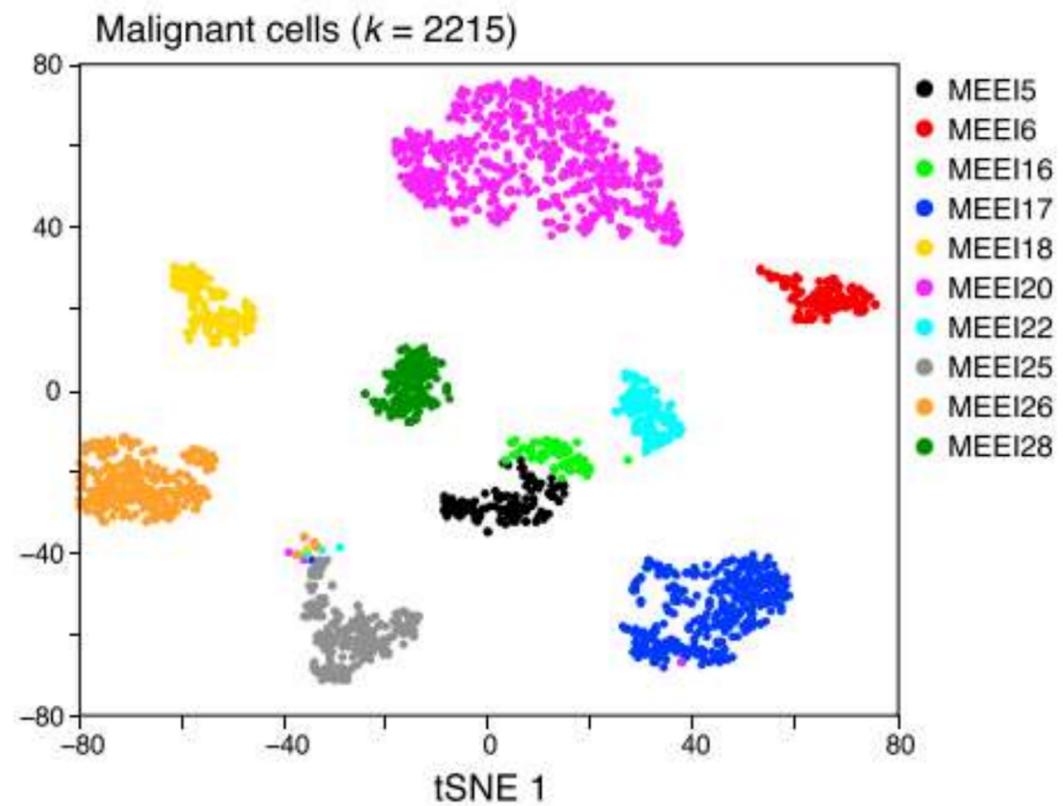


Real heterogeneity –vs- technical noise

**Technical replicate of PBMCs
has near-perfect overlap**

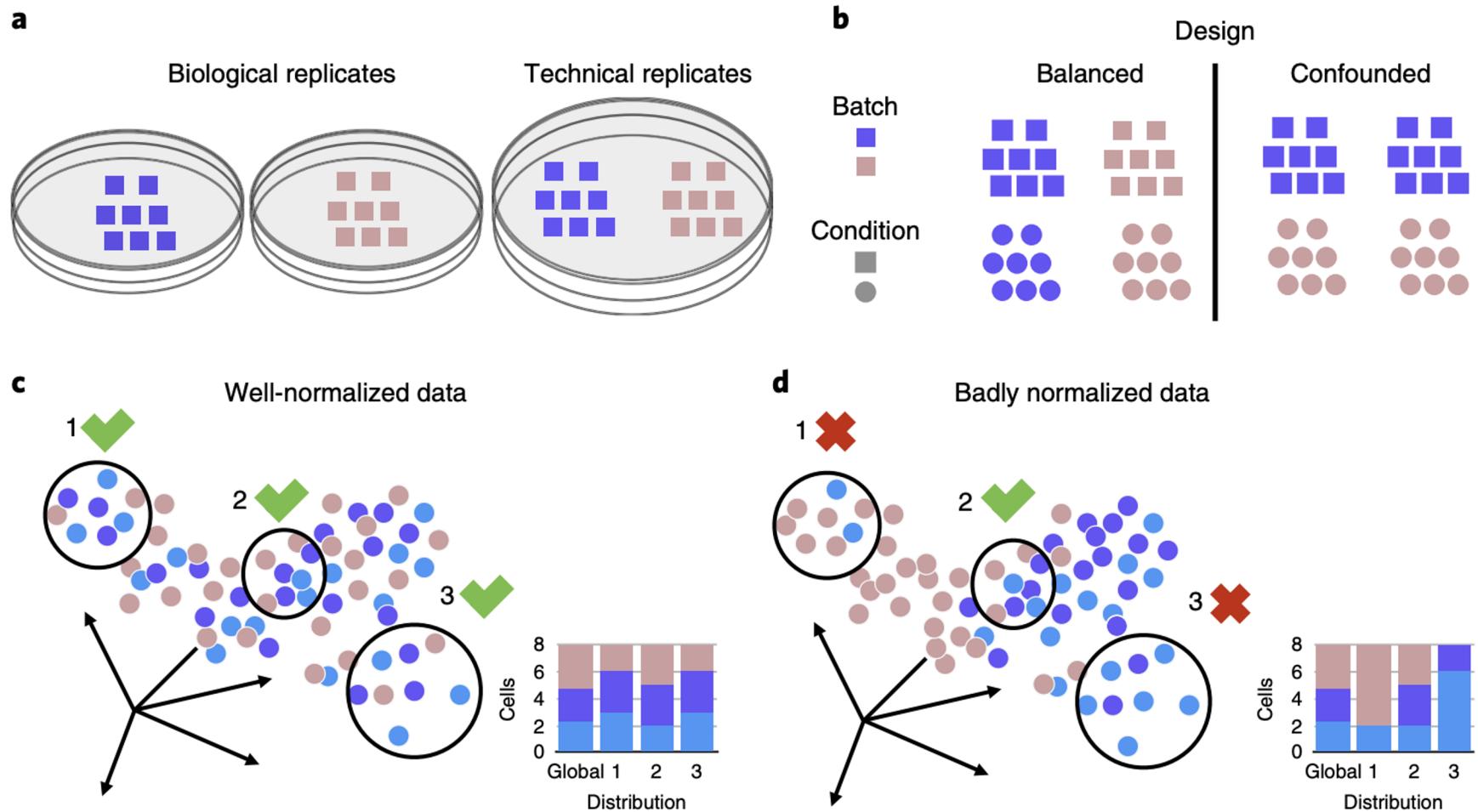


**Cancer cells dramatically
different between patients**



Assessing normalization and batch effects

Evaluate mixing of samples

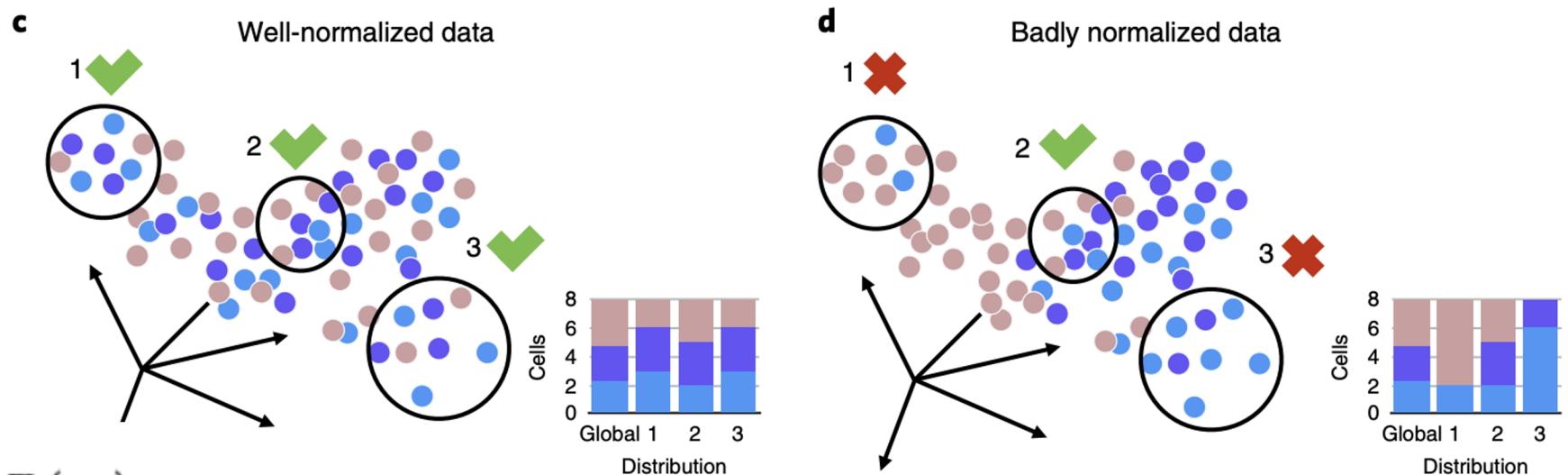
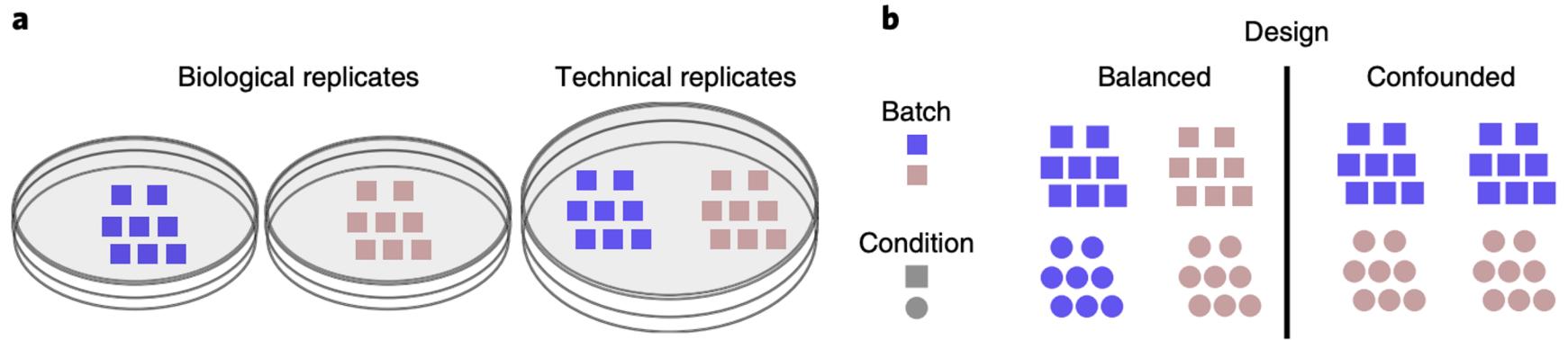


Assessing normalization and batch effects

Select random neighborhoods of fixed size

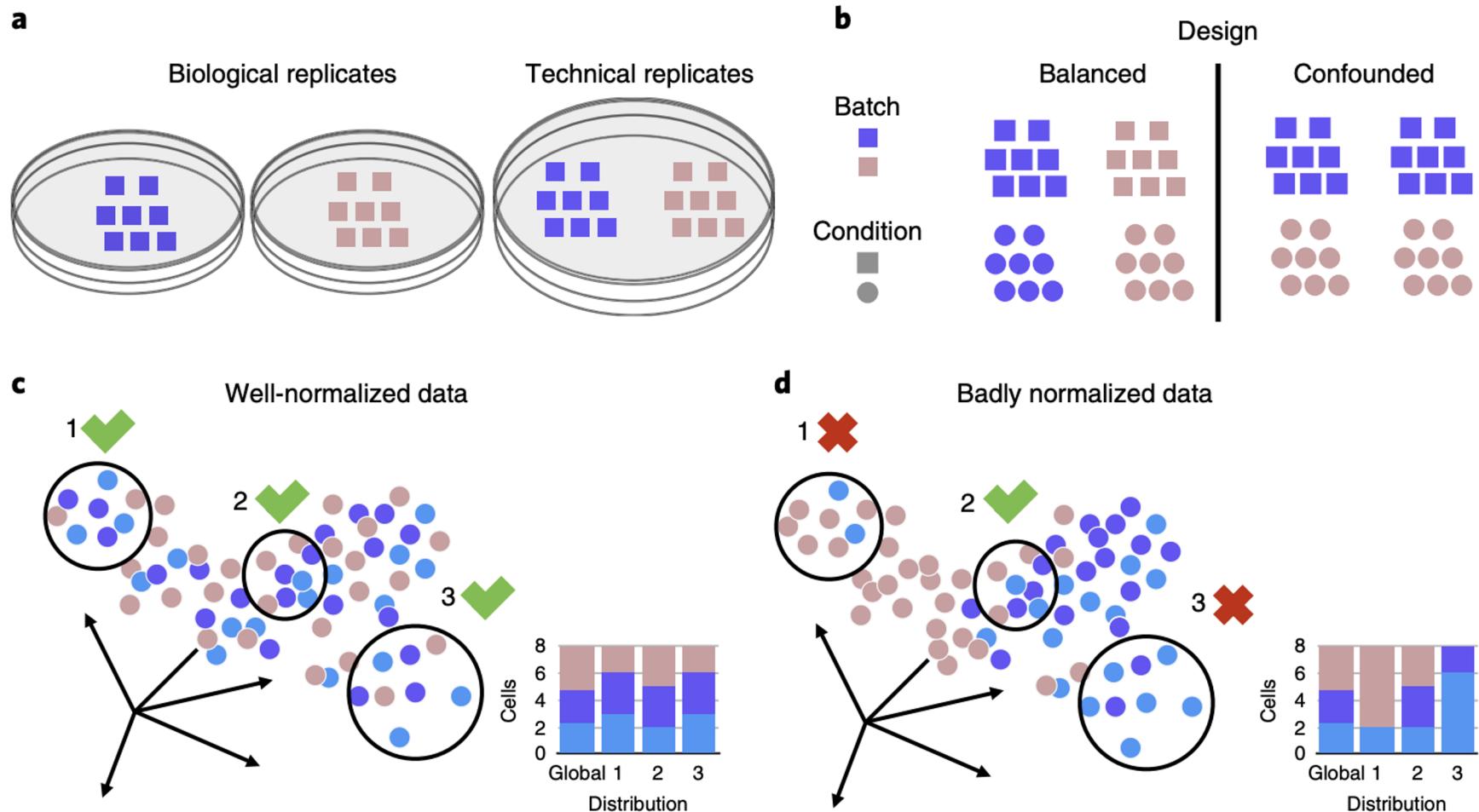
Compute Shannon Entropy of distribution across samples

$$H(X) = - \sum_{i=1}^n P(x_i) \log_b P(x_i)$$

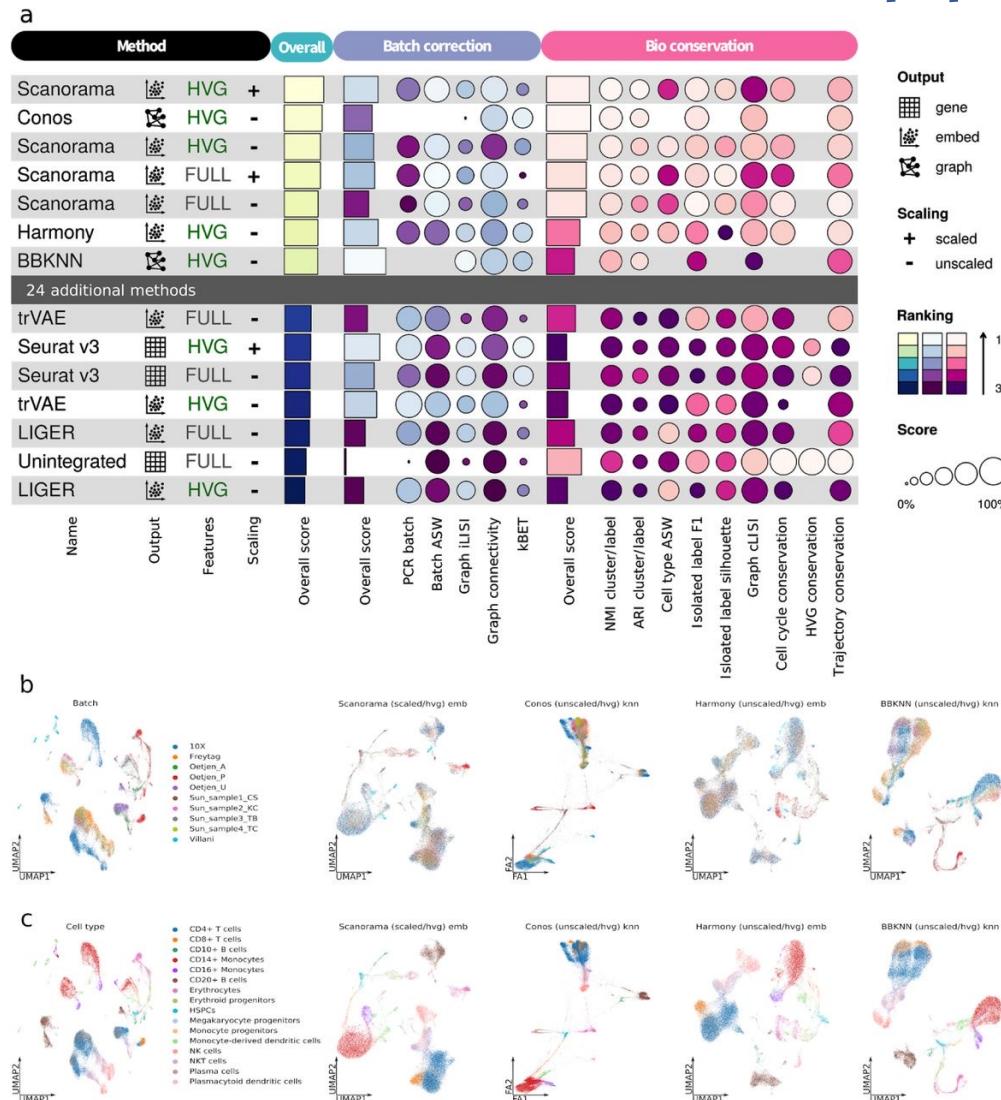


Assessing normalization and batch effects

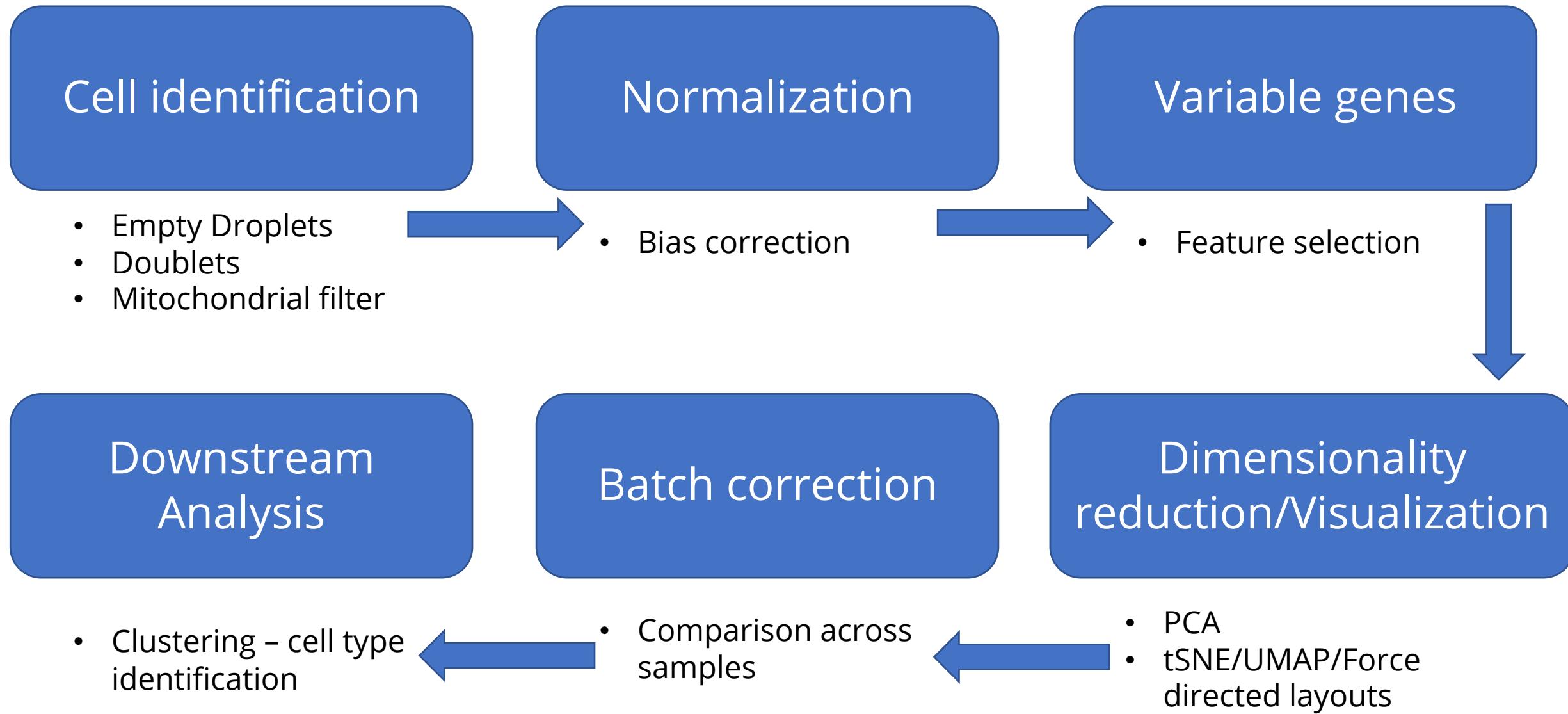
kBET:
Chi-squared test in
random
neighborhoods,
followed by
averaging of binary
test results



Batch effect correction approaches



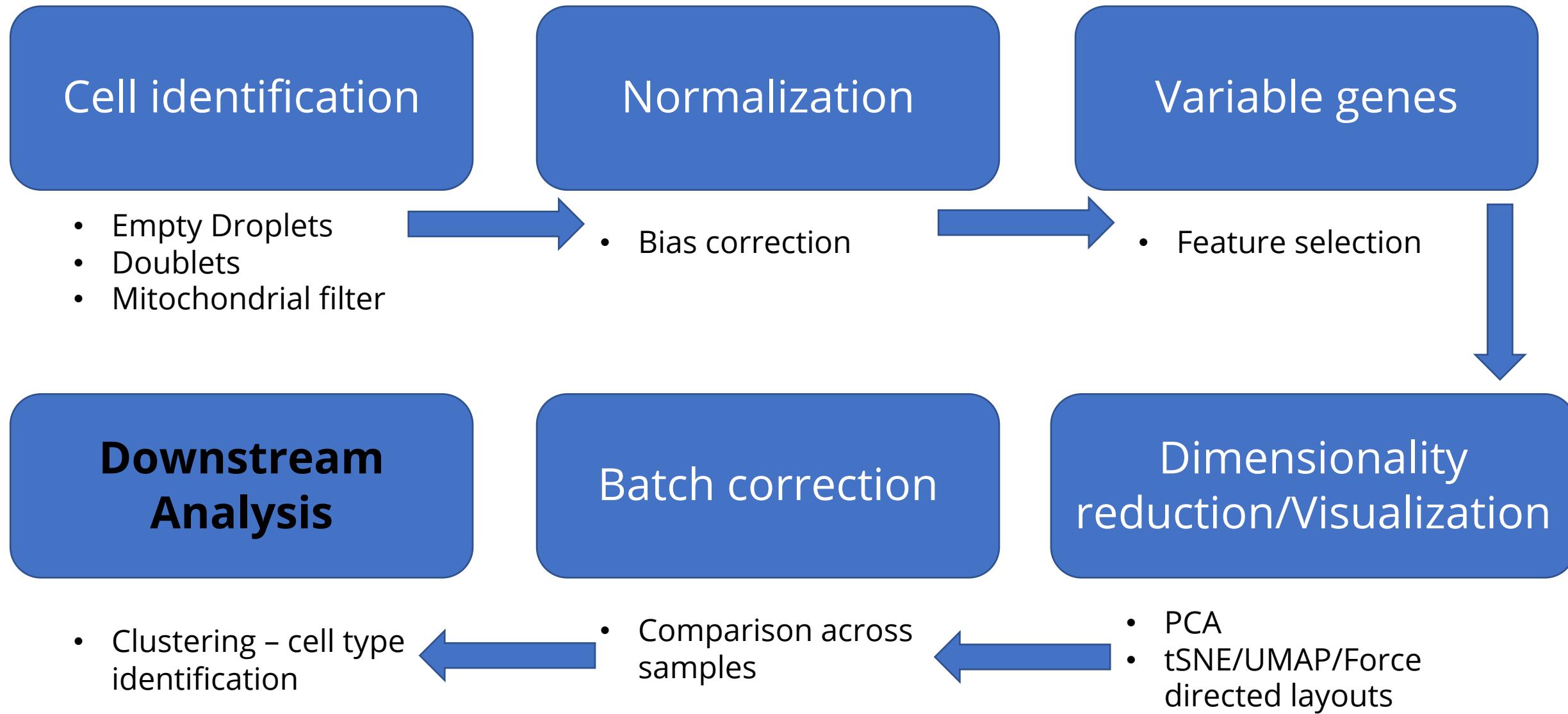
scRNA-seq analysis steps



Other corrections

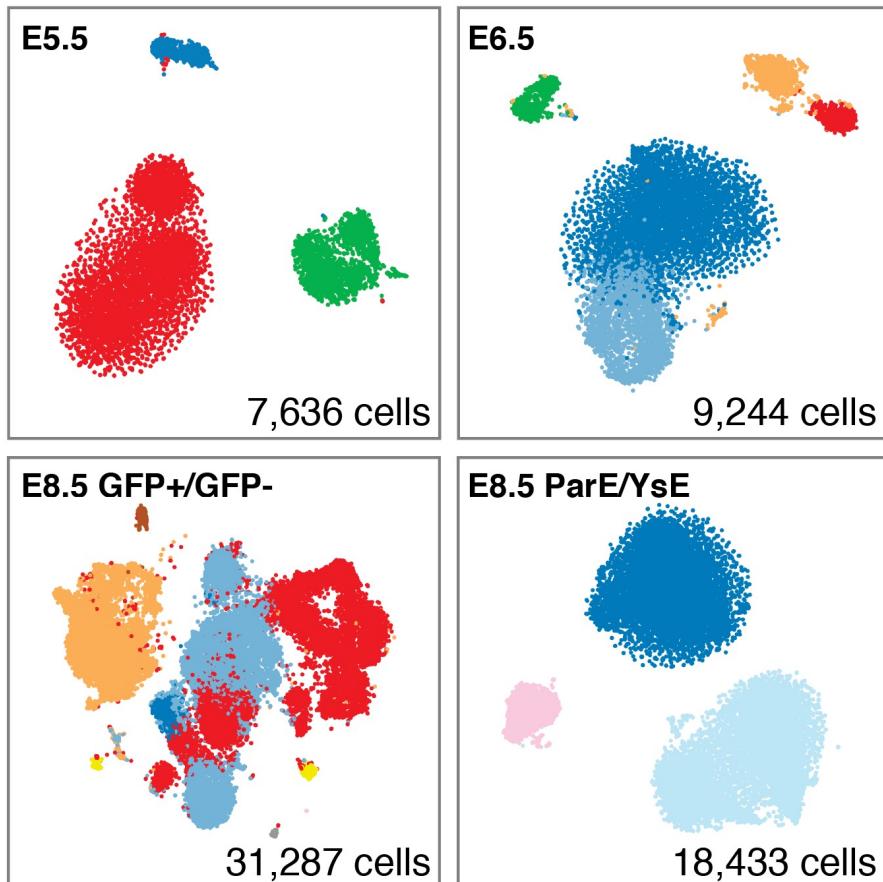
- Cell cycle correction
- Gene expression imputation

scRNA-seq analysis steps

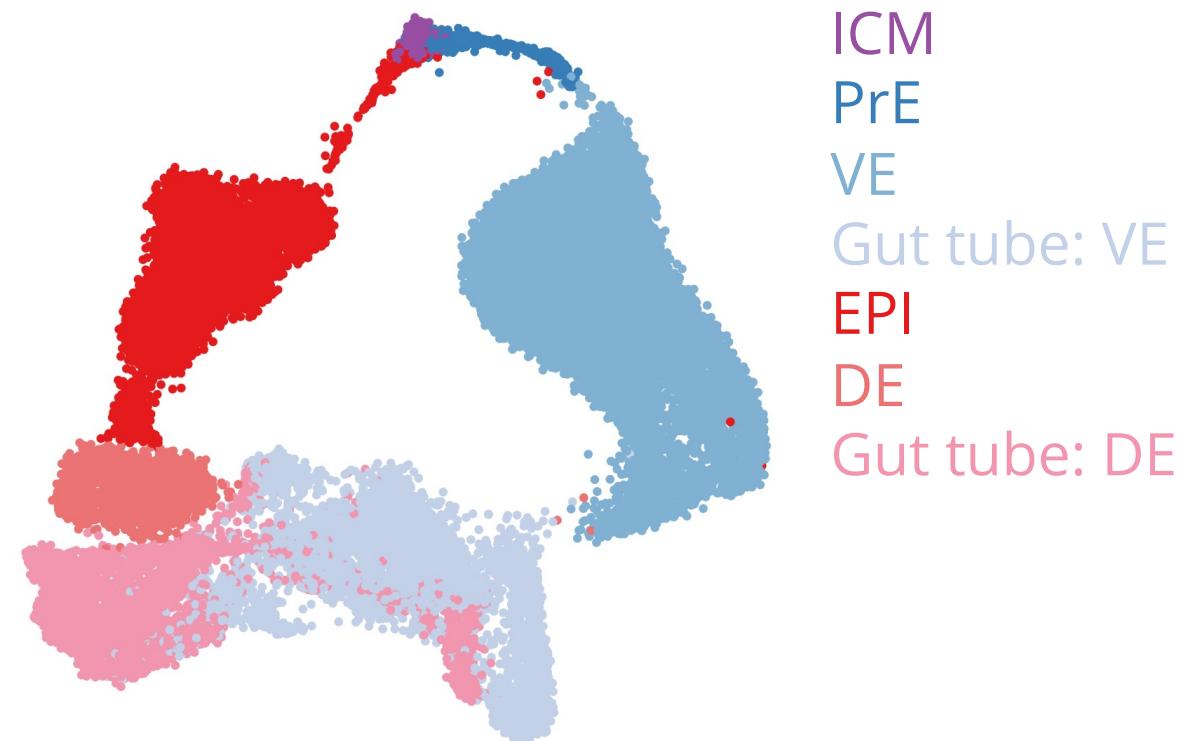


Single-cell data

Discrete clusters

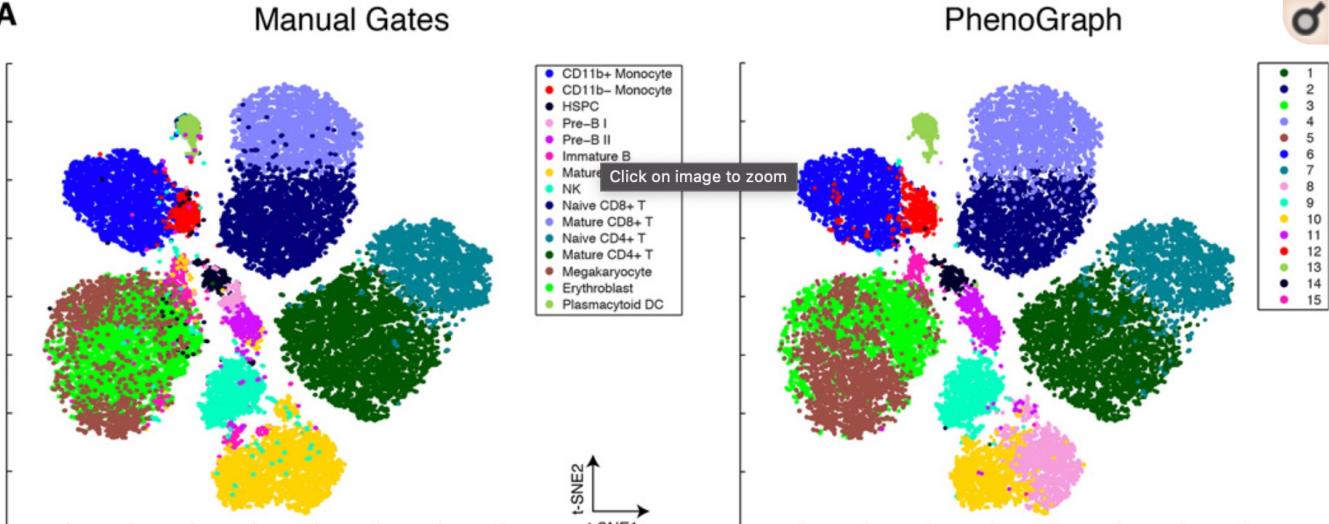


Continuous trajectories

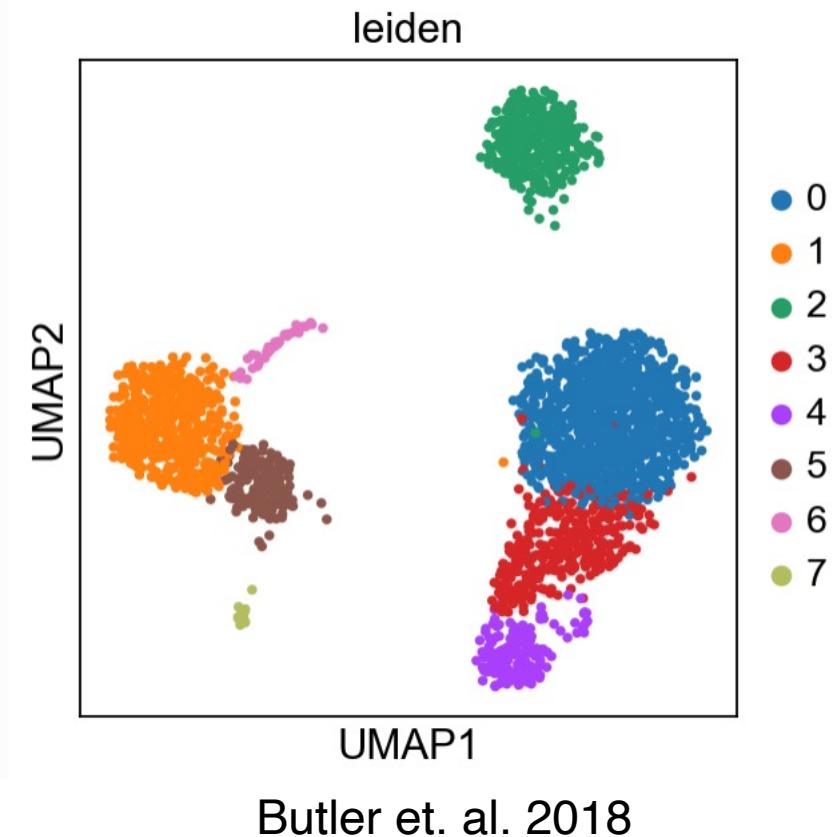


Louvain clustering in single-cell data

A

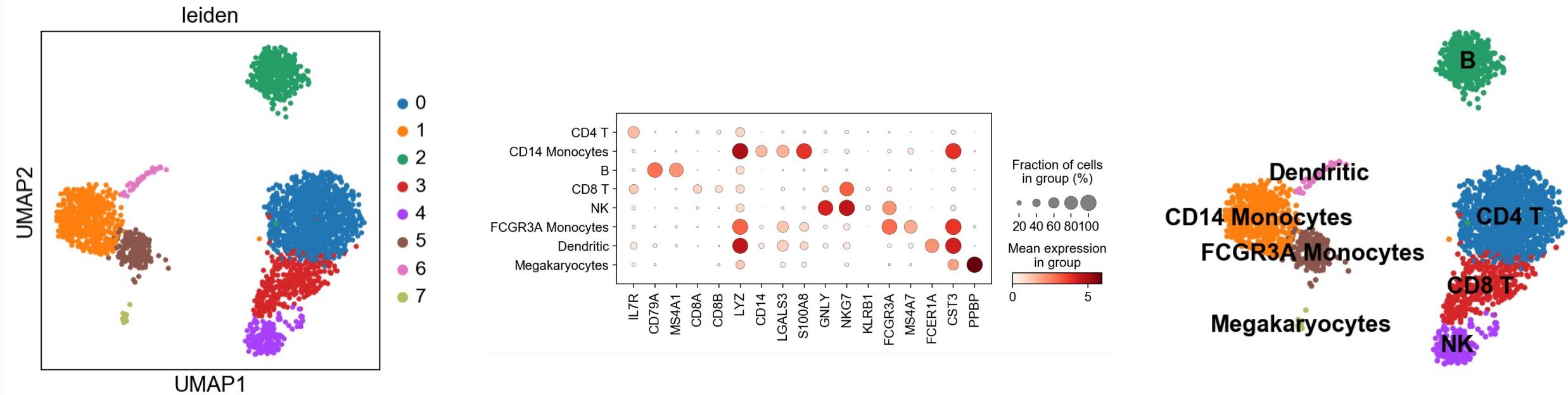


Levine et. al. 2015



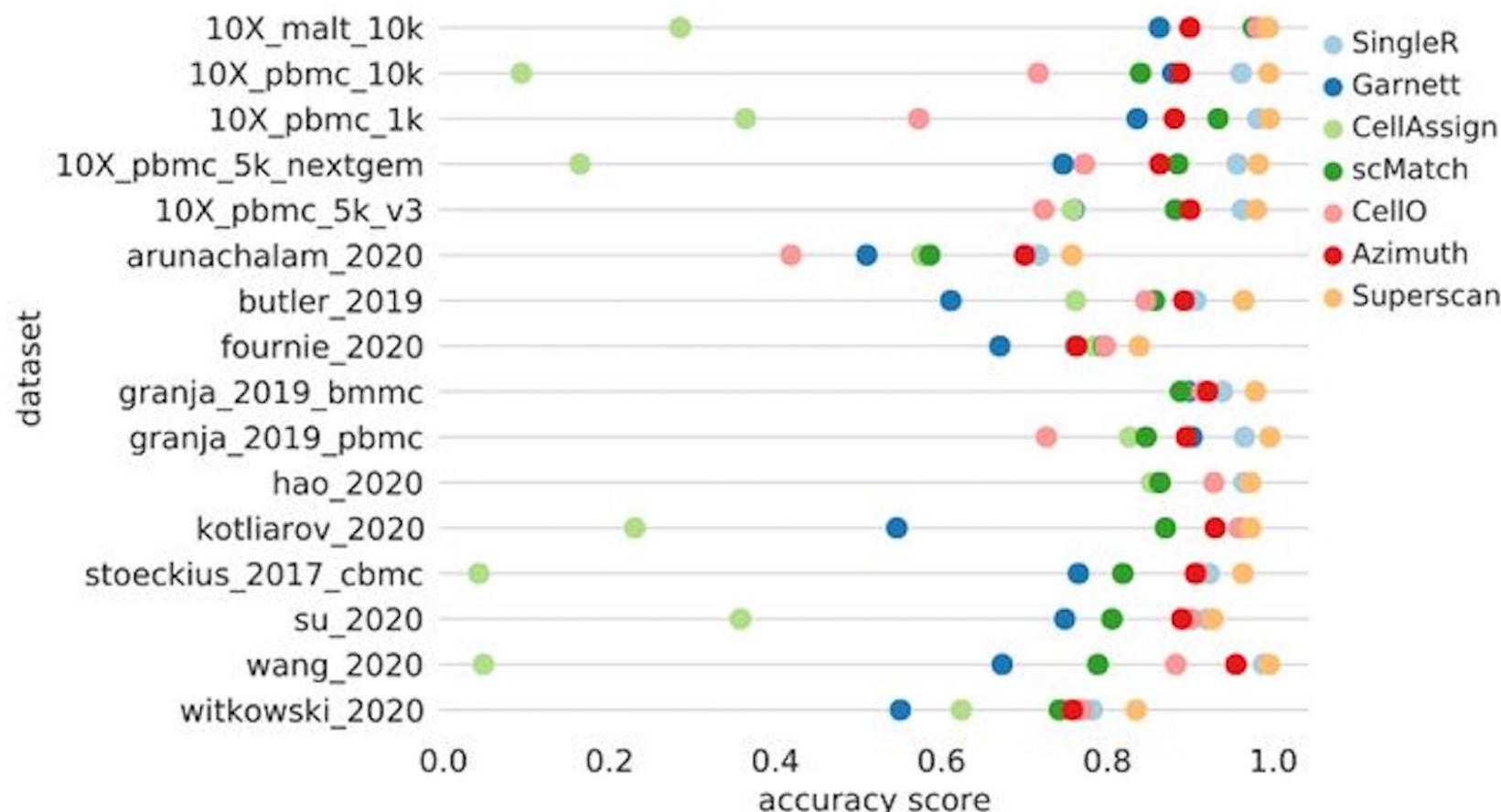
Cell type identification

- Marker based identification



Cell type identification

- Supervised approaches: Train on manually labeled cells
- Superscan: XGBoost classifier



scRNA-seq analysis steps

