Tools For Computational Biology – Lecture 18

Introduction to Single-cell genomics

Manu Setty
Basic Sciences Division | Translational Data Science IRC
Fred Hutch

msetty@fredhutch.org

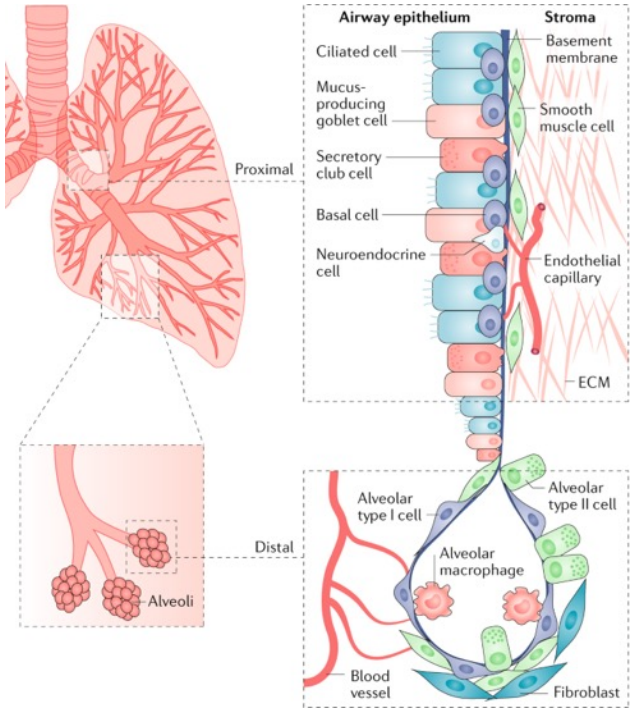
Today’s agenda

- Single-cell genomics: Why?
- Single-cell genomics: How?
- Single-cell RNA-seq preprocessing and analysis

Why single-cell genomics

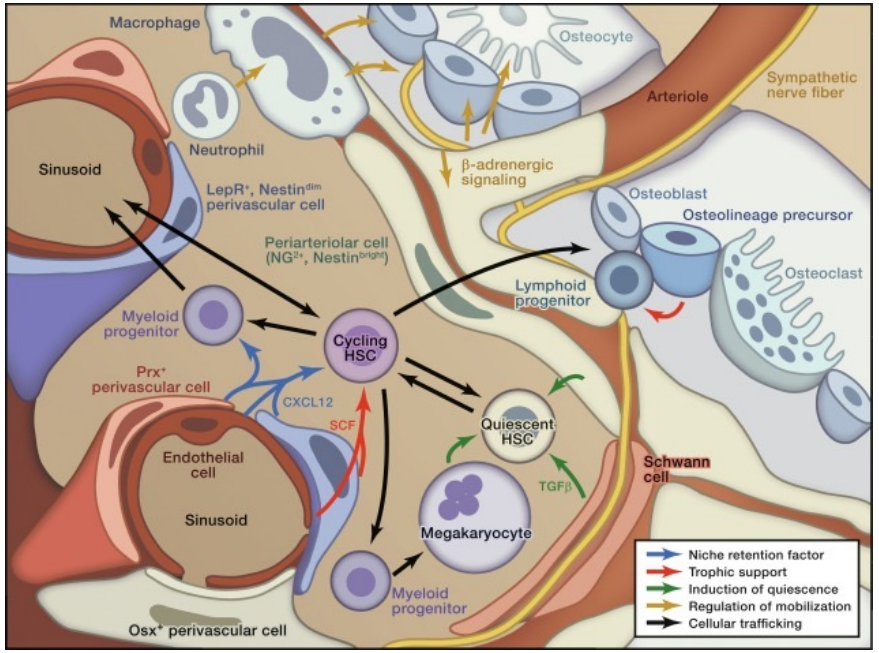
Heterogeneity in tissue homeostasis

Lung



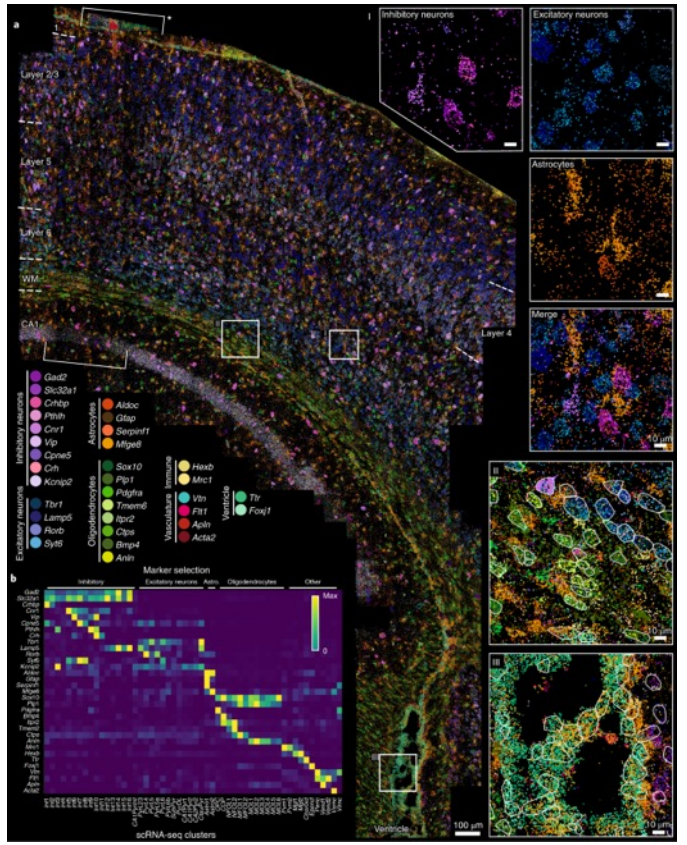
Altorki et. al., 2019

Bone Marrow



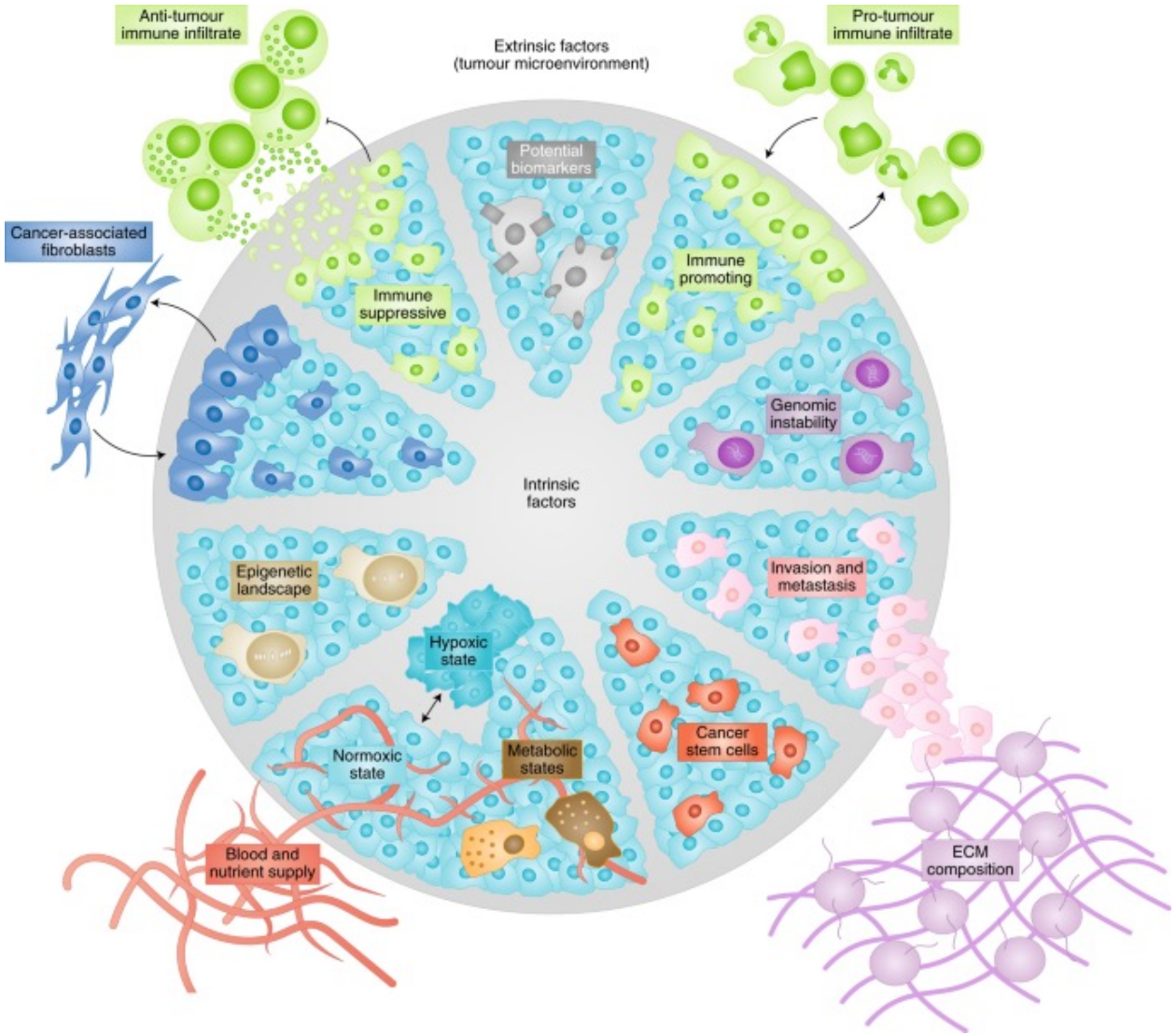
Hoffman et. al., 2020

Somatosenory Cortex



Codeluppi et. al., 2018

Tumor heterogeneity



Lawson et. al., 2018

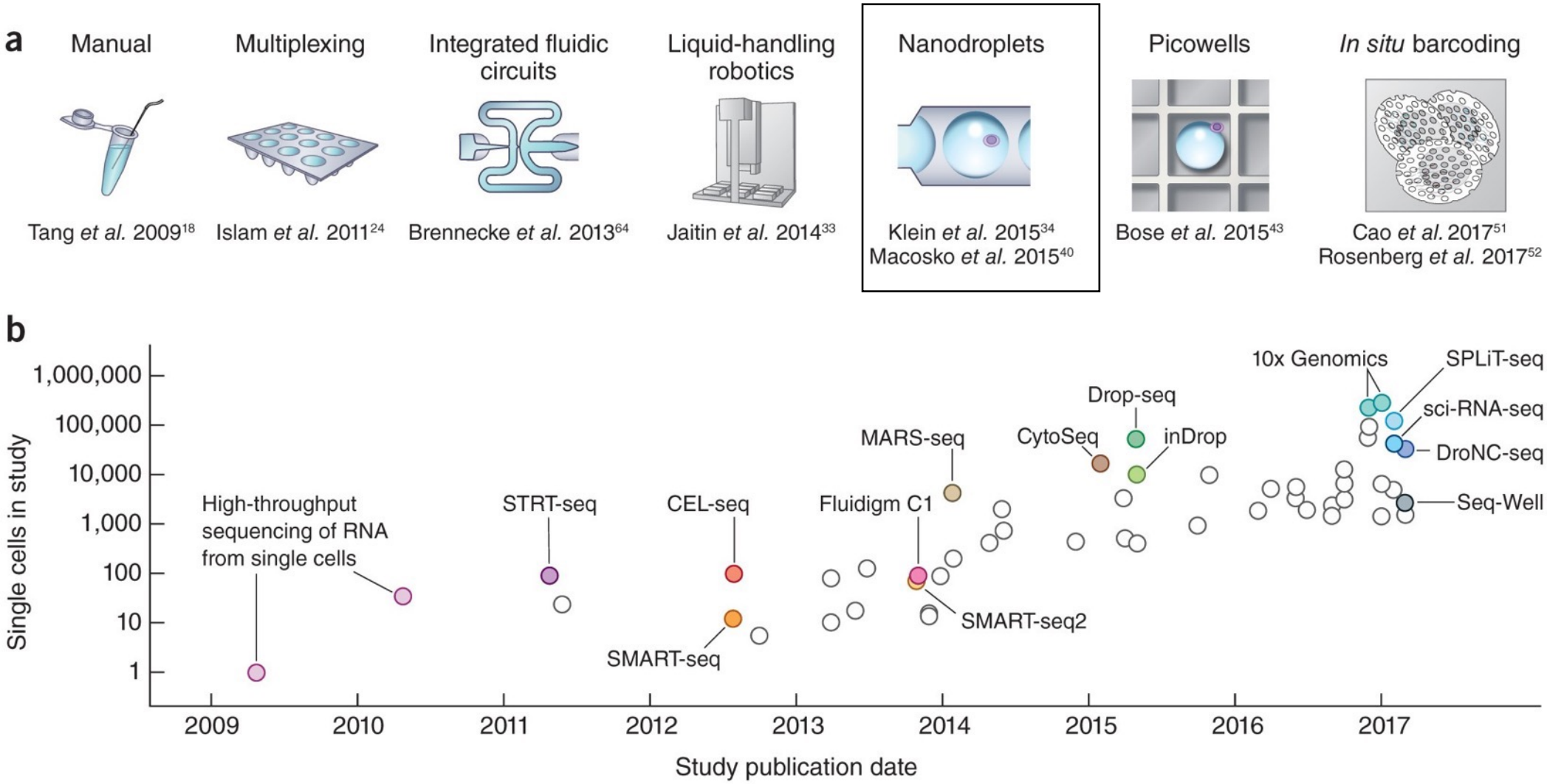
Single-cell technologies profile heterogeneity instead of average

Bulk Genomics

Single-cell Genomics

Single-cell genomics: How?

Evolution of single-cell RNA-seq



Microfluidic single-cell technologies

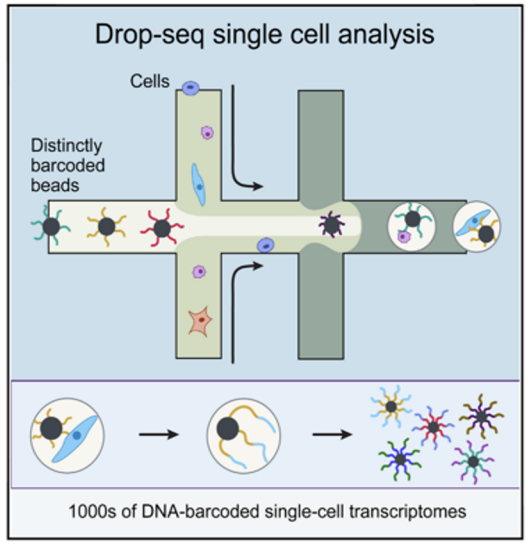
Moved throughput from hundreds to thousands of cells

Resource

Cell

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

Graphical Abstract



Authors

Evan Z. Macosko, Anindita Basu, ..., Aviv Regev, Steven A. McCarroll

Correspondence

emacosko@genetics.med.harvard.edu (E.Z.M.), mccarroll@genetics.med.harvard.edu (S.A.M.)

In Brief

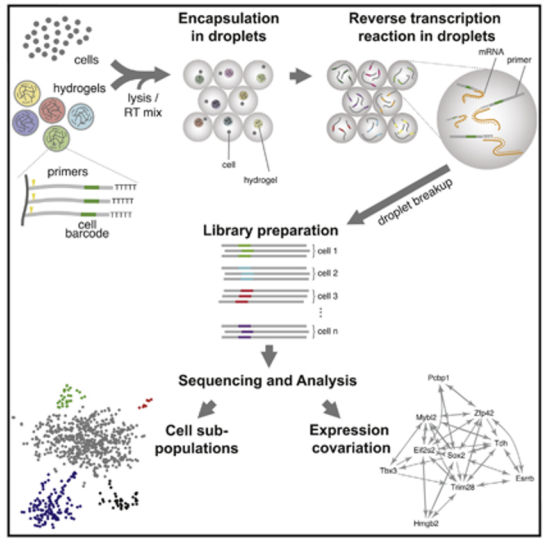
Capturing single cells along with sets of uniquely barcoded primer beads together in tiny droplets enables large-scale, highly parallel single-cell transcriptomics. Applying this analysis to cells in mouse retinal tissue revealed transcriptionally distinct cell populations along with molecular markers of each type.

Resource

Cell

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

Graphical Abstract



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.)

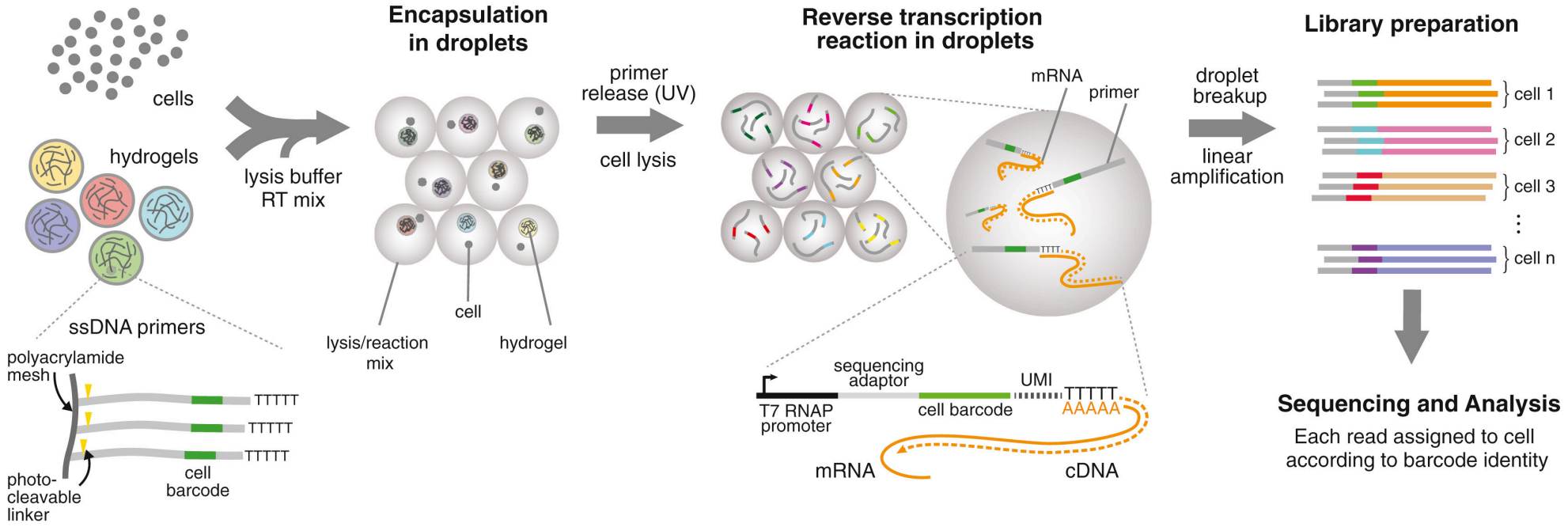
In Brief

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

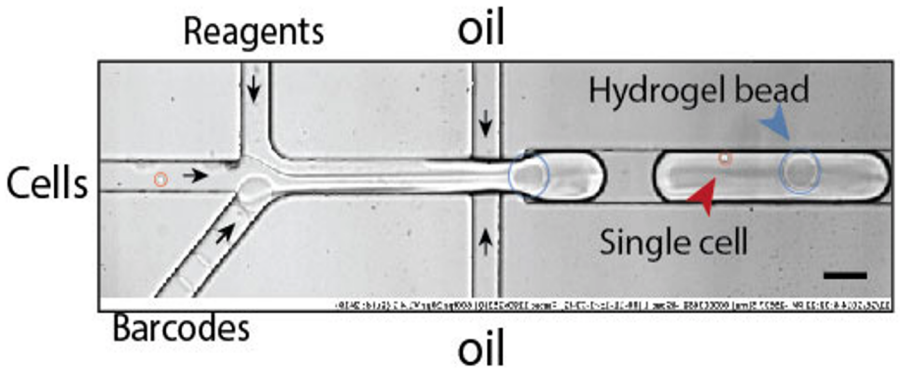


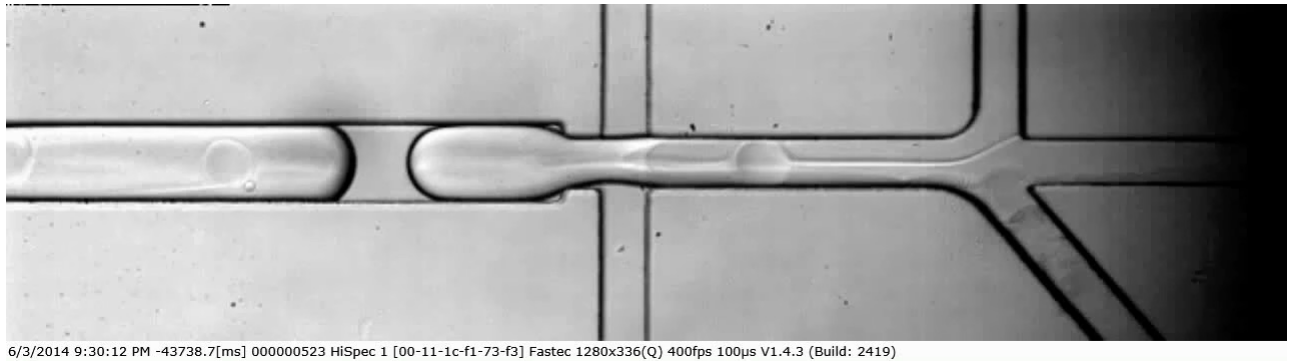
- Droplet-based processing using microfluidics
- Nanoliter scale aqueous drops in oil
- 3' End
- Bead based for cell barcoding
- Uses UMI (Unique Molecular Identifier).

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

Microfluidic single-cell technologies

Moved throughput from hundreds to thousands of cells





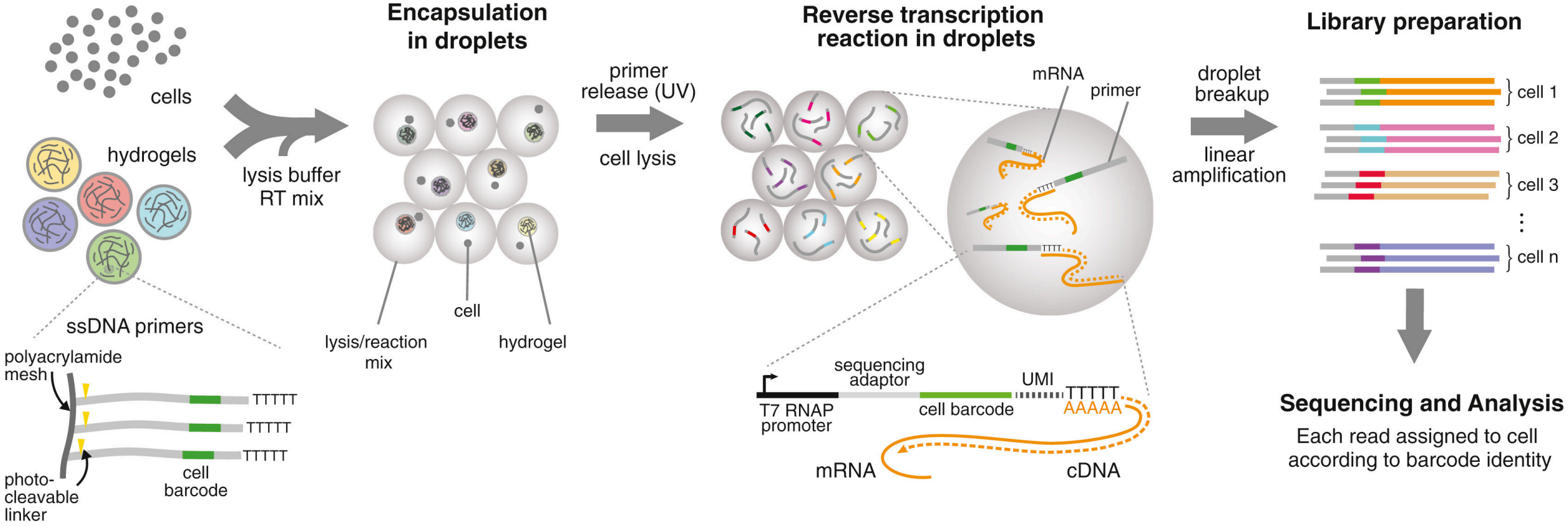
Klein et al *Cell* 2015
Macosko et al *Cell* 2015

CellXGene interactive tool

```
cellxgene launch 10x_pbmc.h5ad

http://localhost:5005
```

Single-cell RNA-seq: Preprocessing and Analysis

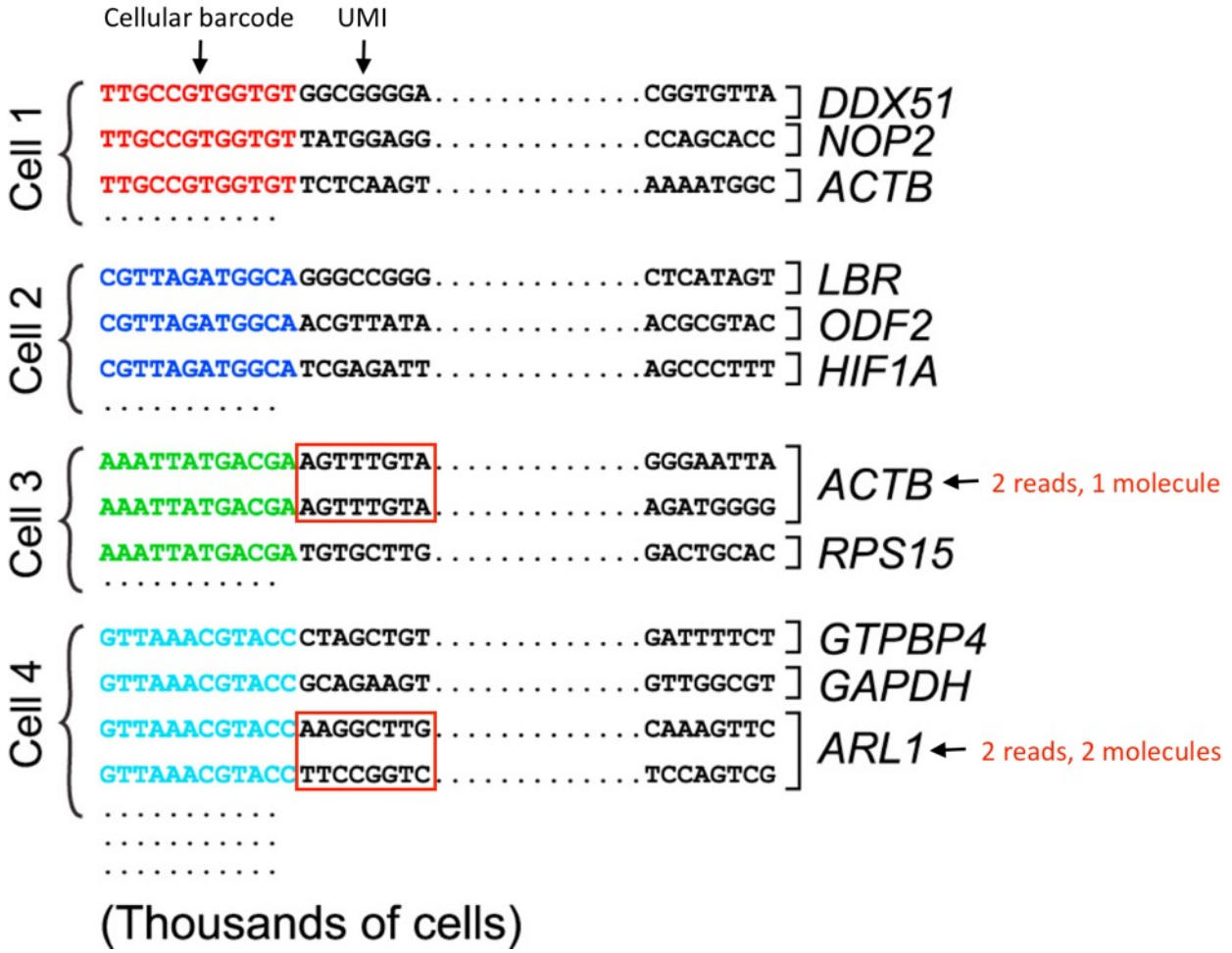
Microfluidic single-cell technologies



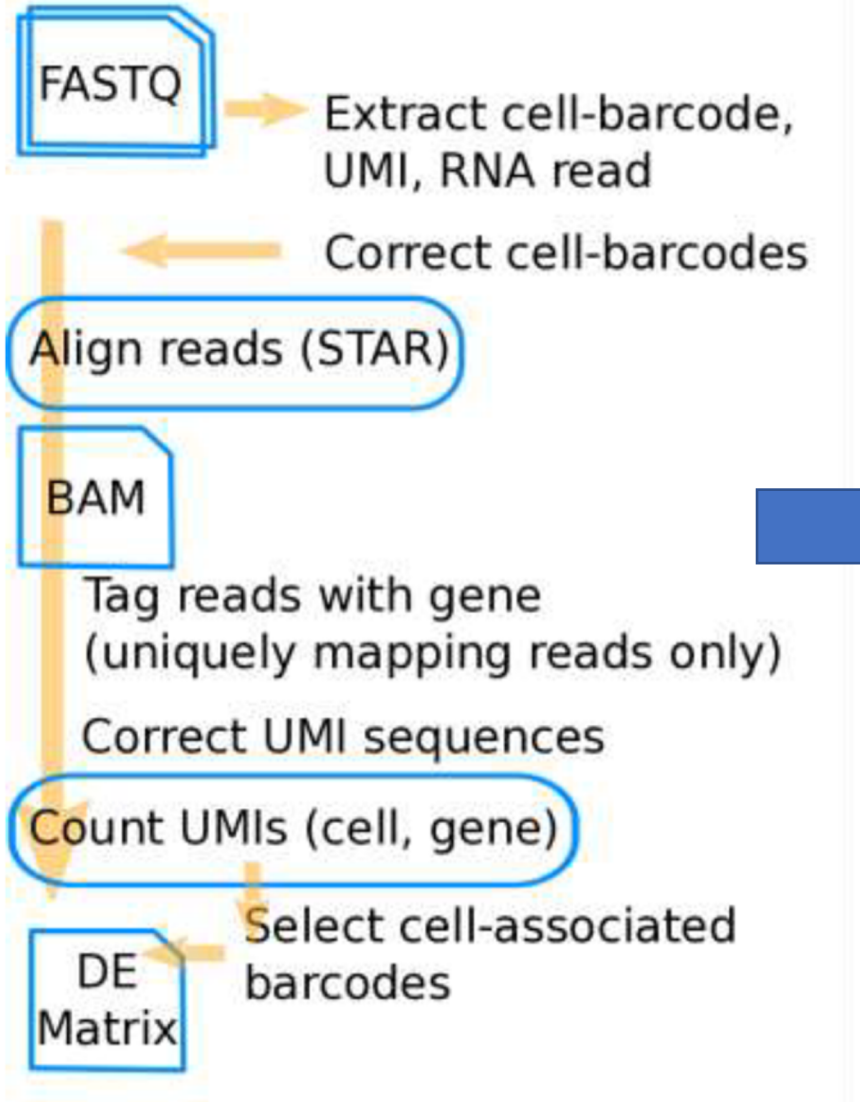
Klein et al *Cell* 2015

Cell Barcodes and UMIs

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



10X Preprocessing Pipeline: CellRanger



Genes

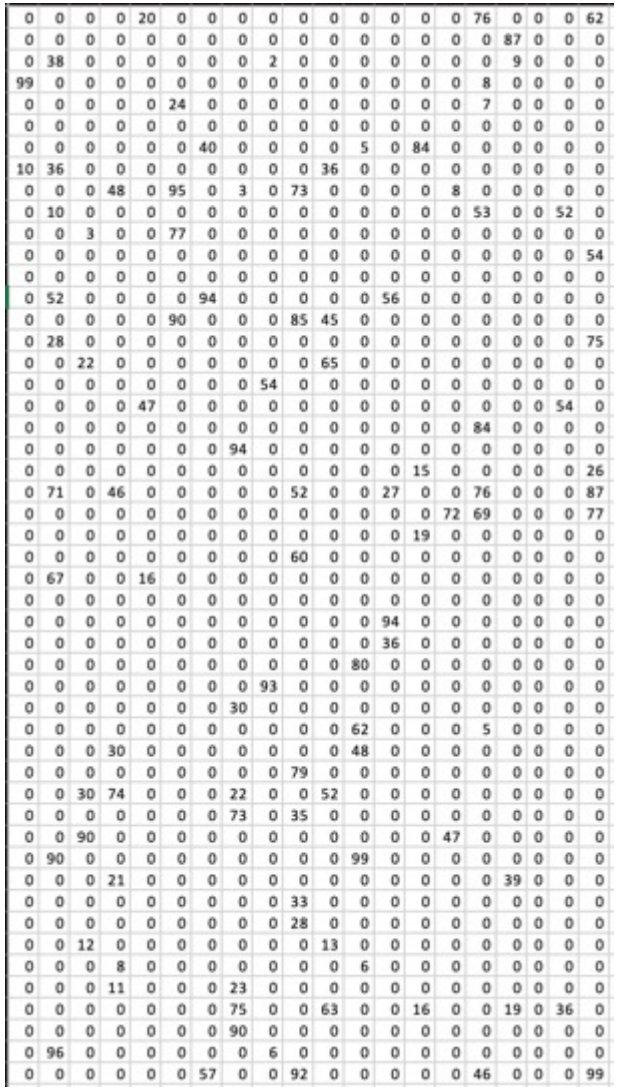
Cells

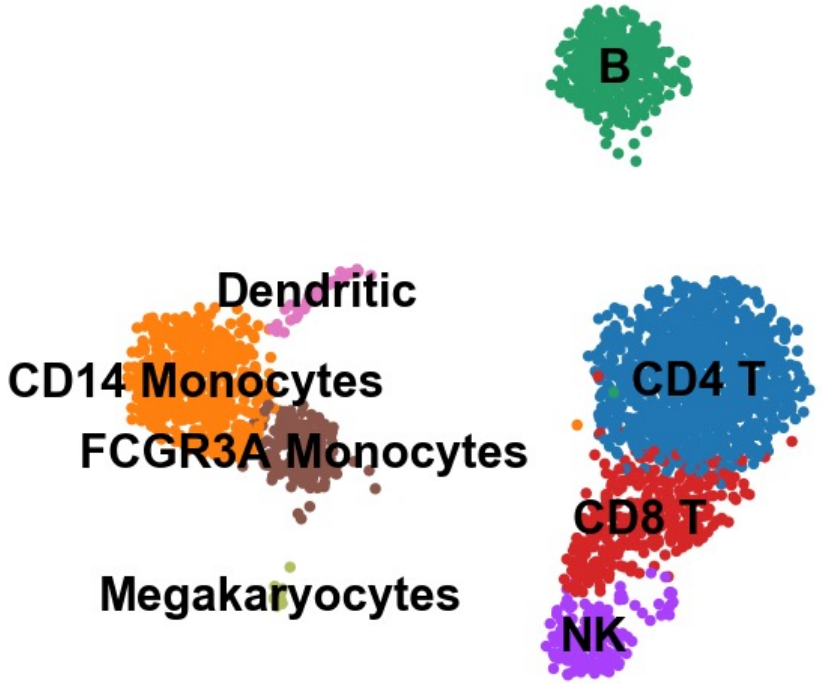
Single-cell RNA-seq

*Scale & Resolution X *Noise & Sparsity**

Genes

Cells

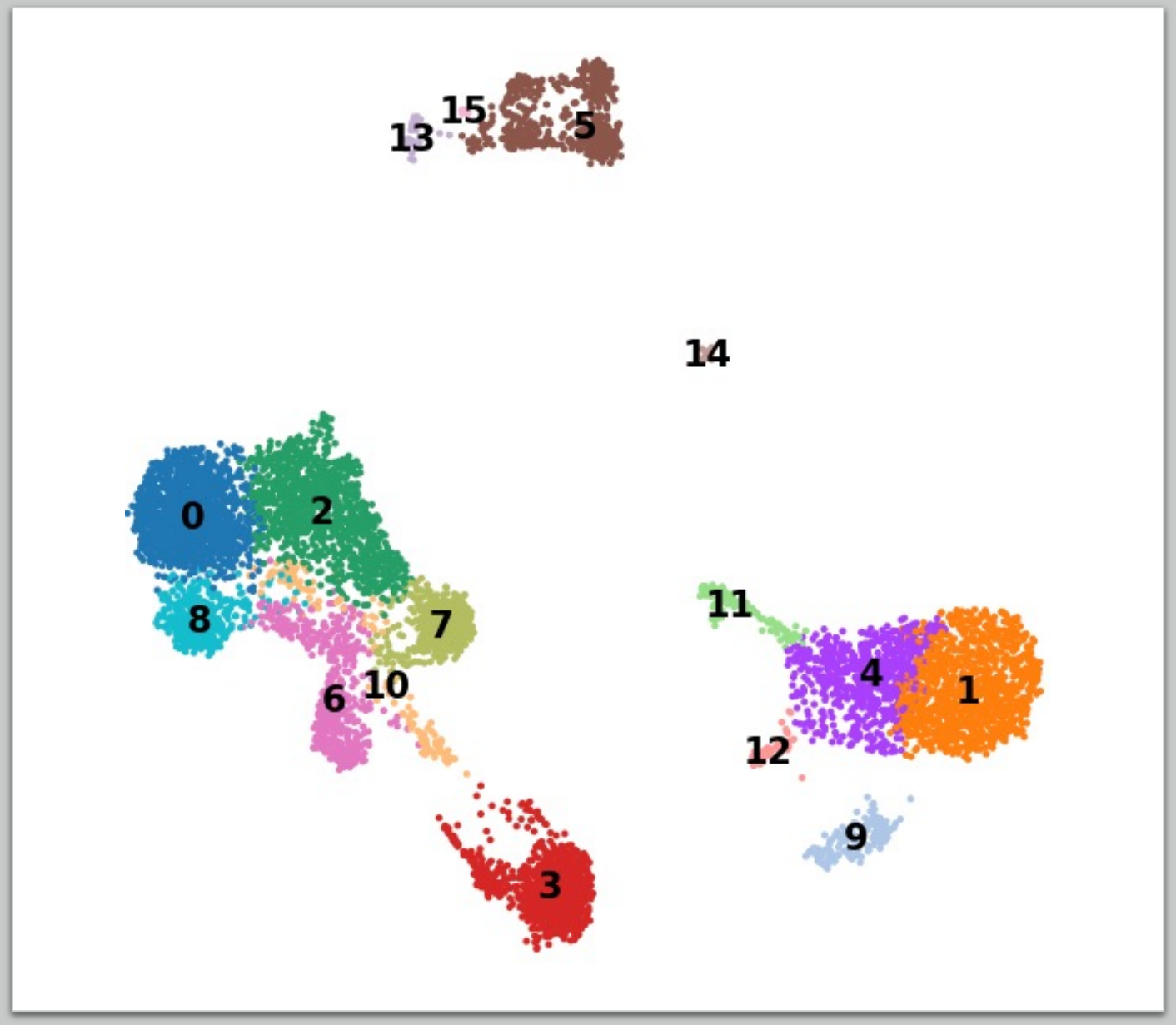




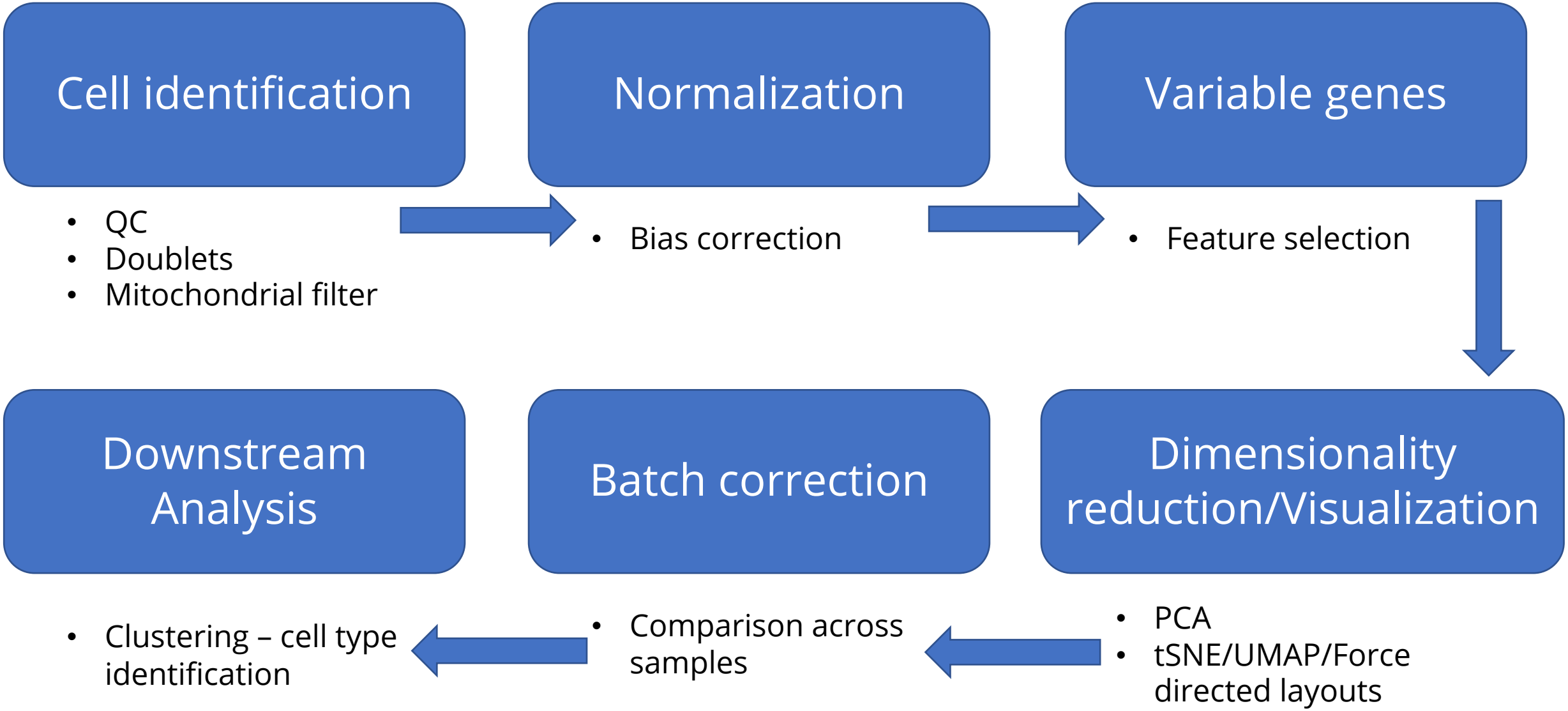
A black-box view

```
import scanpy as sc

ad = sc.read( <counts file> )
sc.pp.normalize_total(ad)
sc.pp.log1p(ad)

sc.pp.highly_variable_genes(ad)
sc.pp.pca(ad)

sc.pp.neighbors(ad)
sc.tl.leiden(ad)
sc.tl.umap(ad)
```

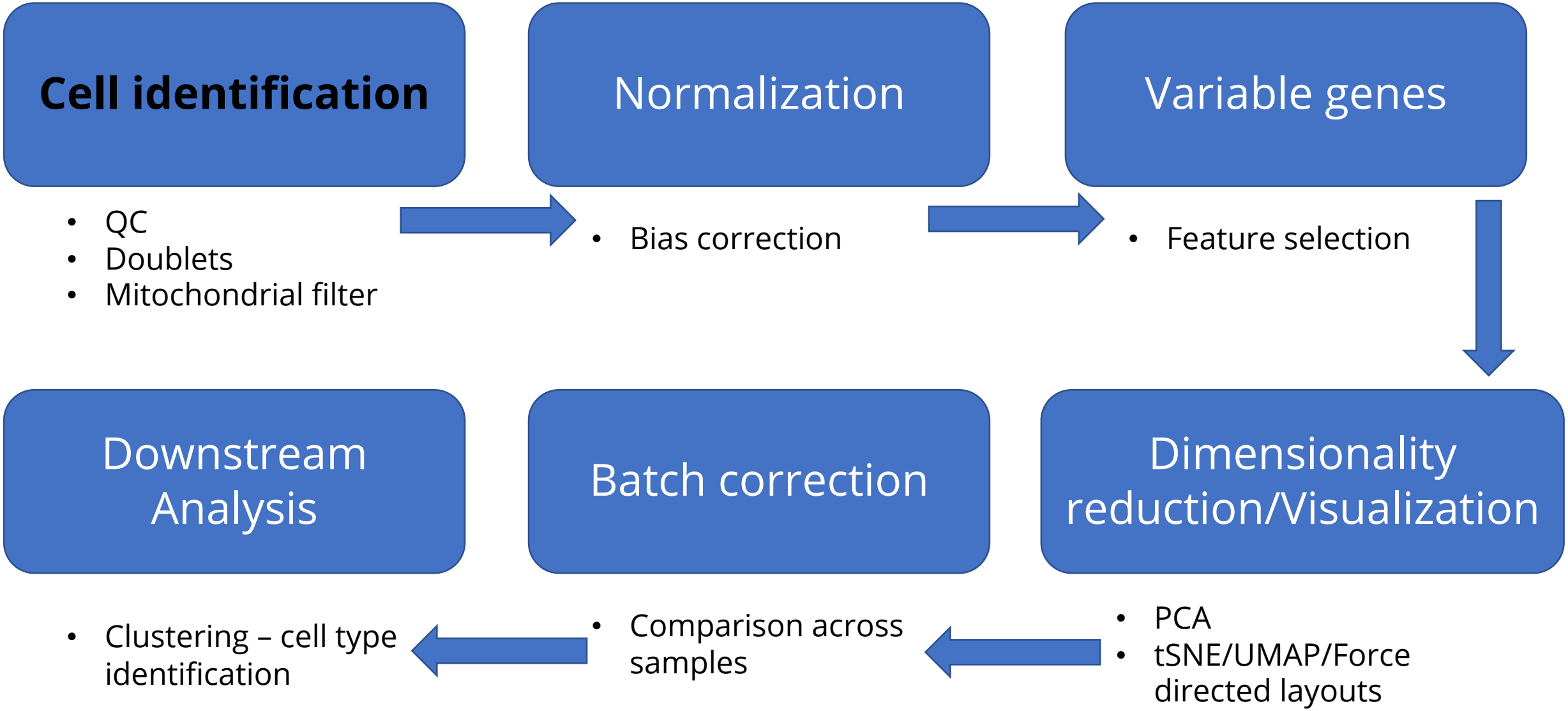
Click here for an example



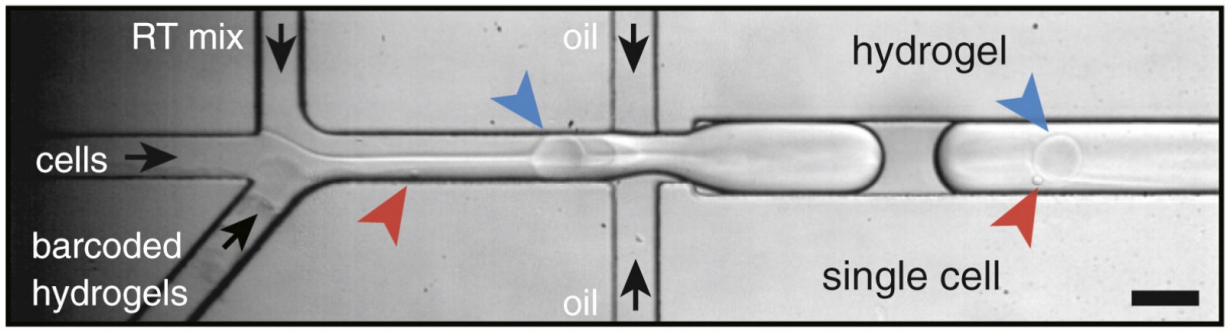
scRNA-seq analysis steps

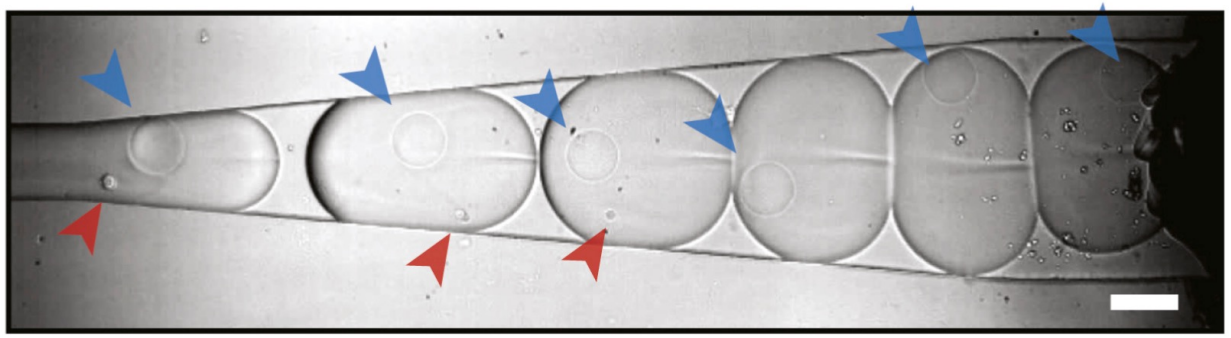


scRNA-seq analysis steps



scRNA-seq: Empty droplets & Ambient RNA





- Most droplets do not have cells!
  - Ambient RNA

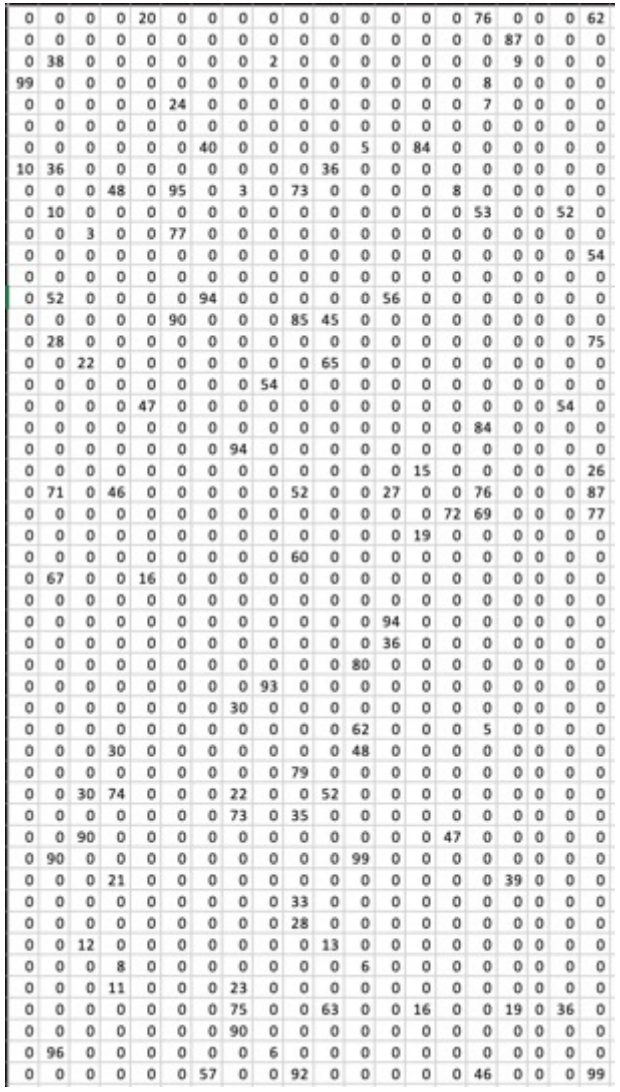
Klein et al *Cell* 2015

Single-cell RNA-seq

All rows are not real cells

Genes

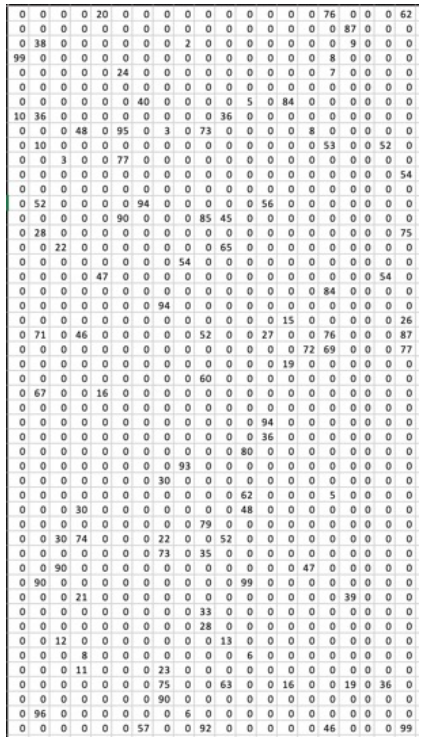
Cells



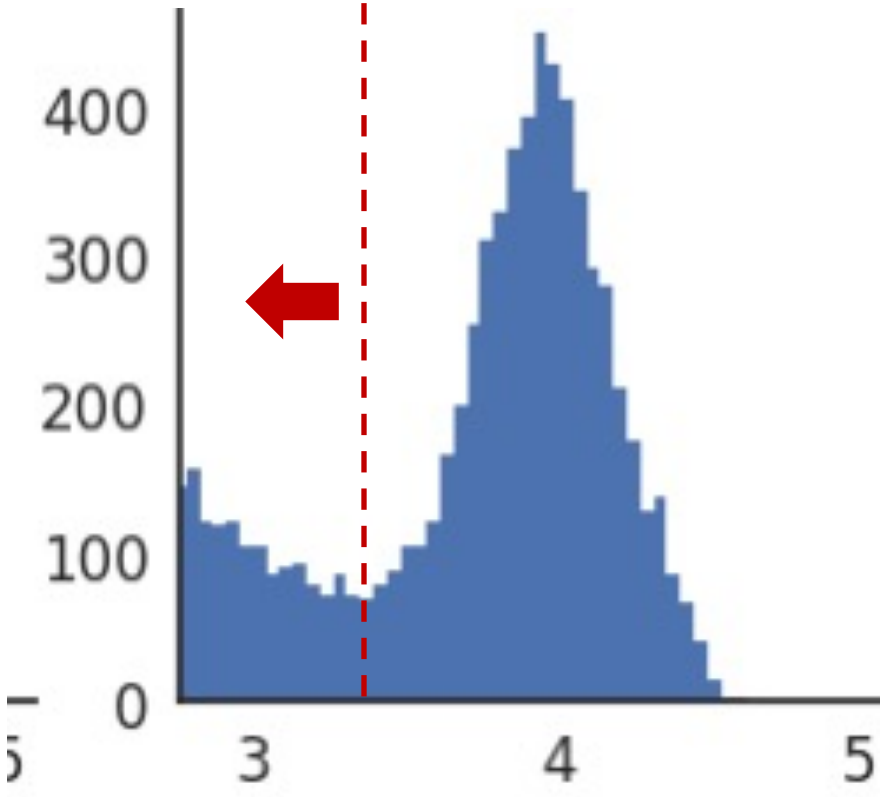
scRNA-seq: Removal of empty droplets

Genes

Cells



Likely
Empty Droplets



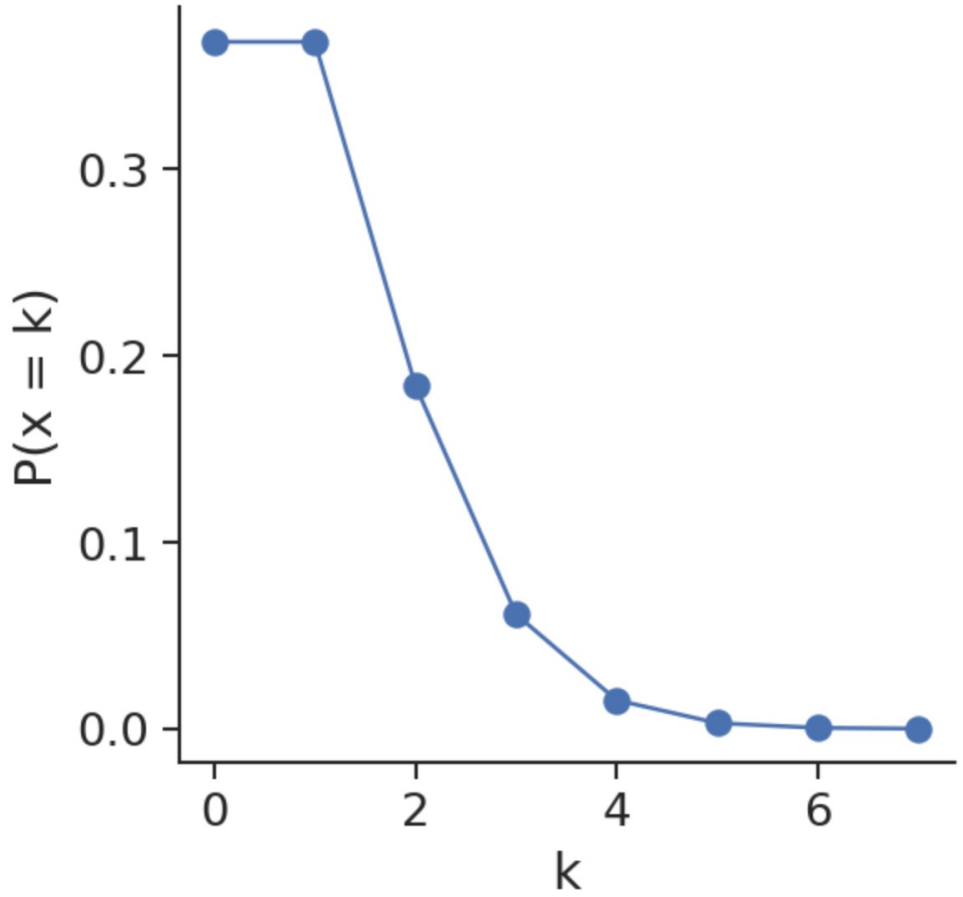
log10 (Total UMIS per cell)

Modeling approaches:

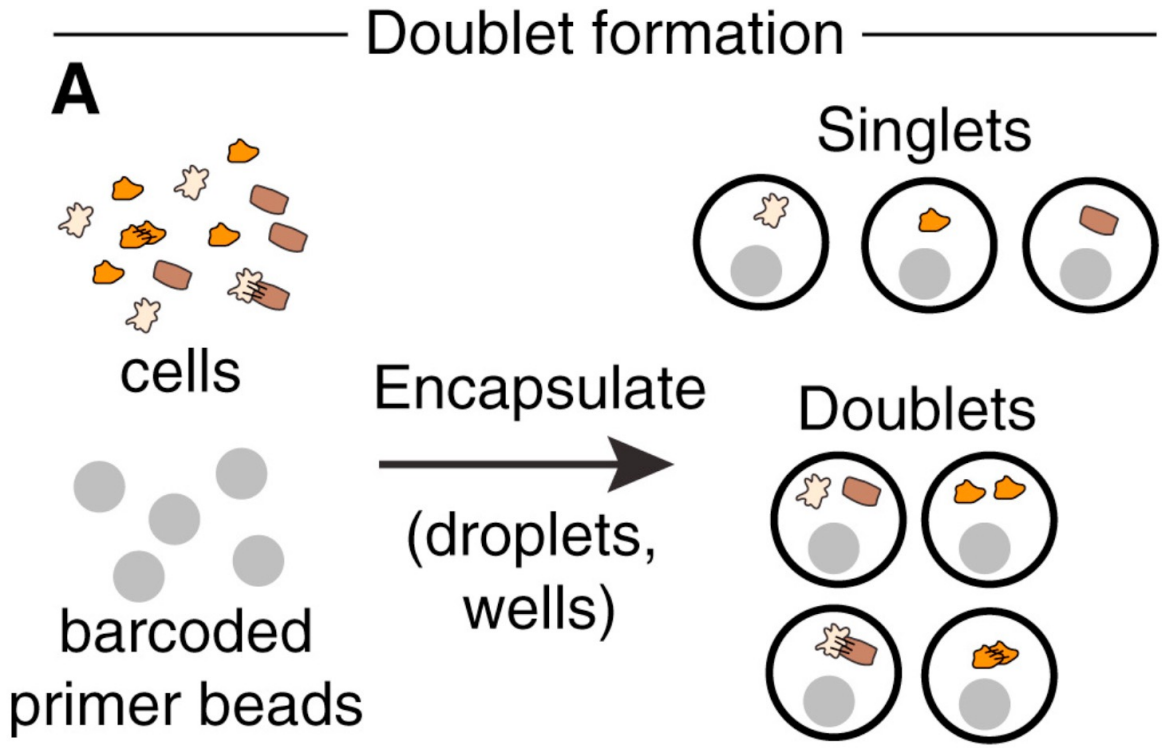
- *EmptyDrops (Lun et al 2019)*
- *SoupX (Young et al 2020)*
- *DecontX (Yang et al 2020)*

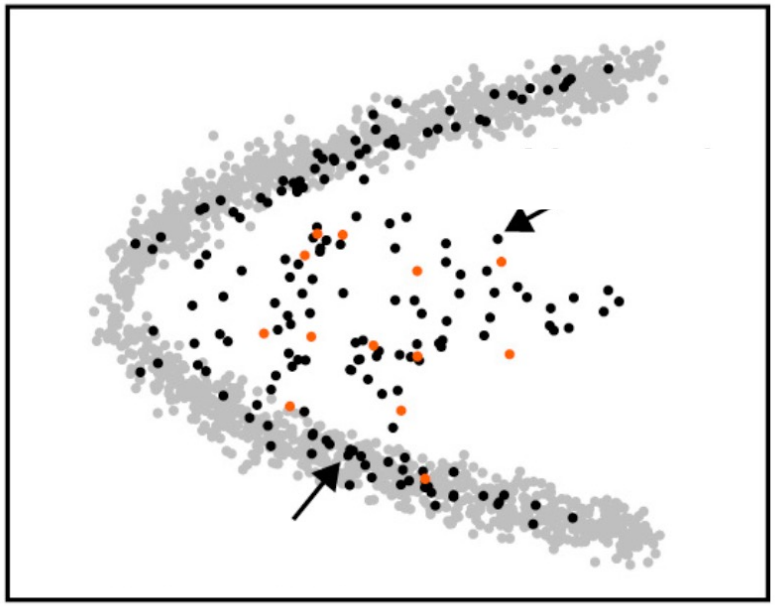
Cell containing droplets

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



Doublets

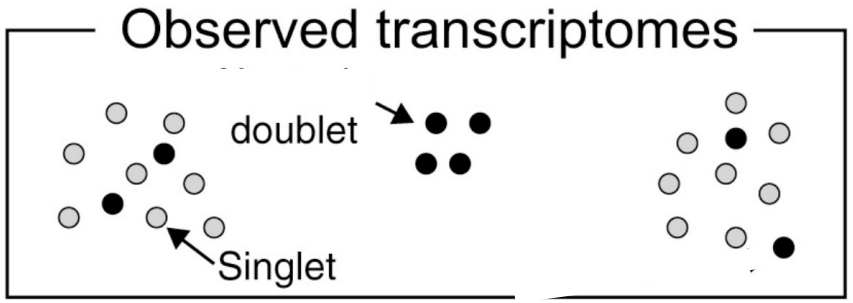




Wolock et al *2019*

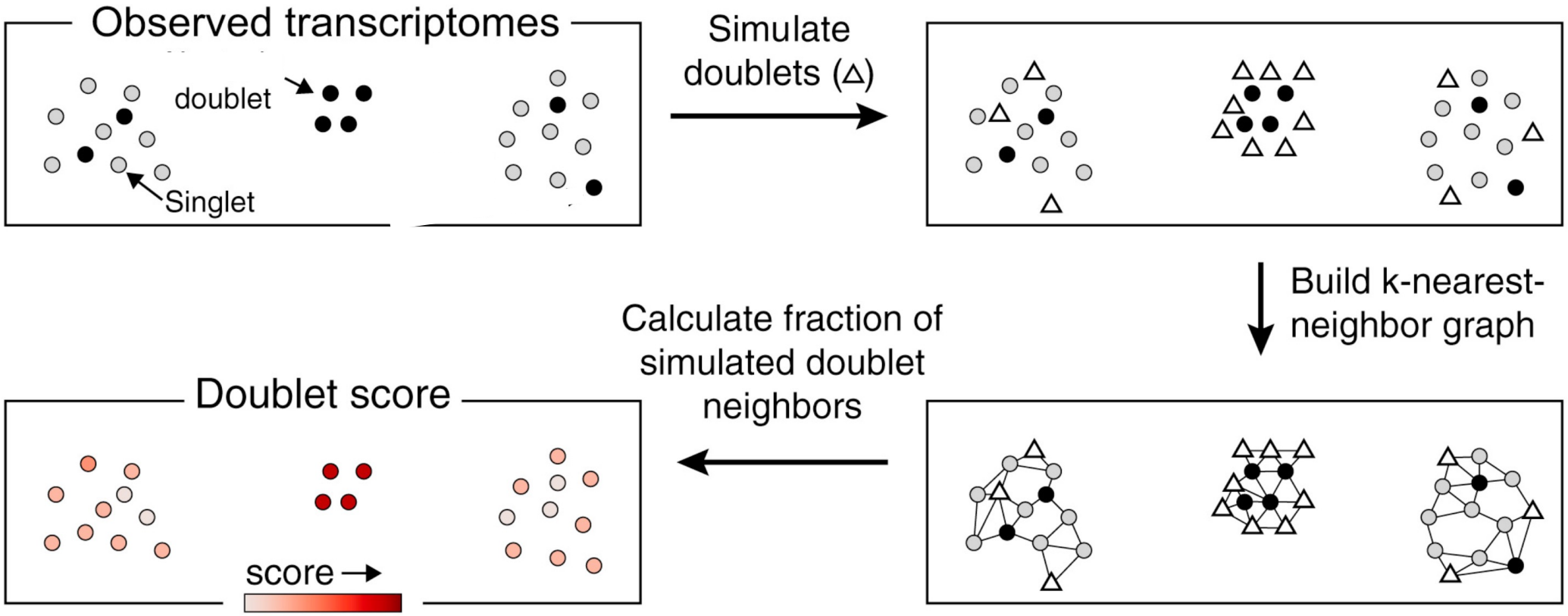
How to detect doublets?

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



How to detect doublets?

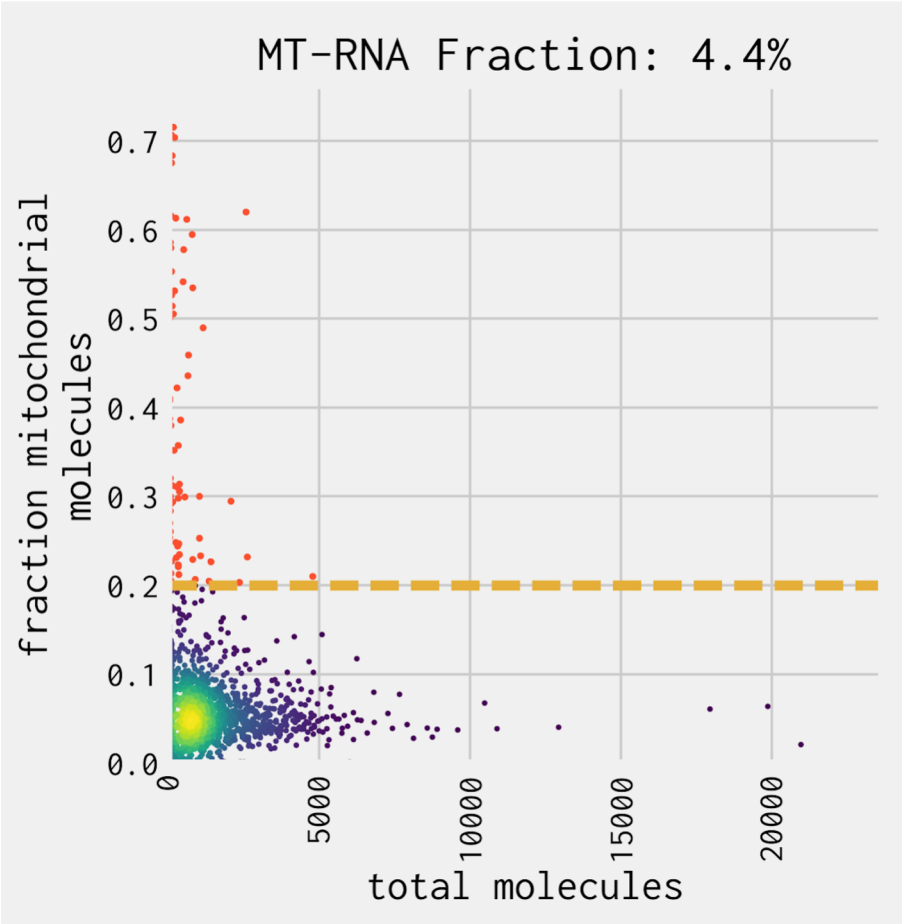
- Assume:
  - Multiplets / Doublets are rare
  - Constituent singlets are present



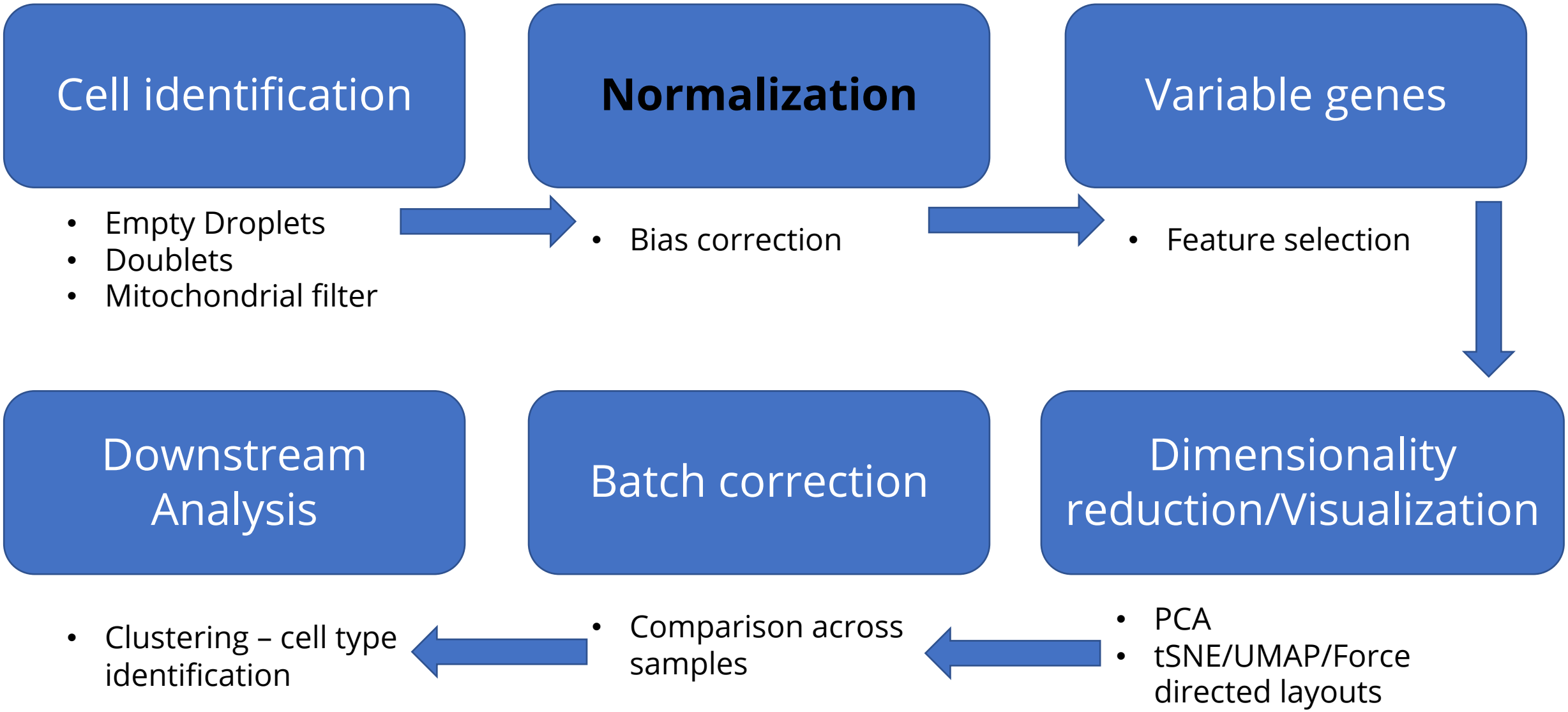
  - Can this detect all possible multiplets?

Mitochondrial filter

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

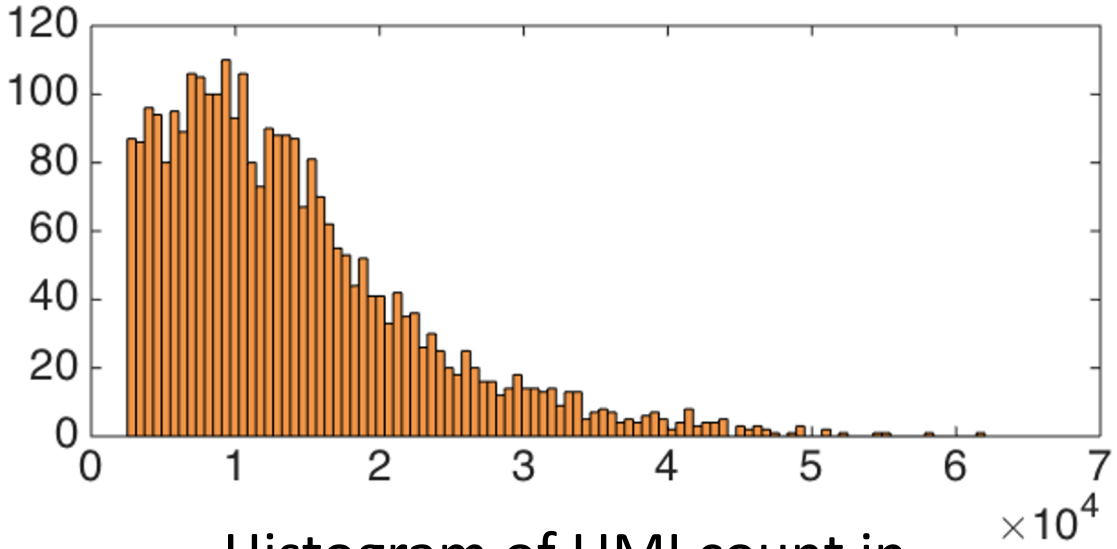


scRNA-seq analysis steps



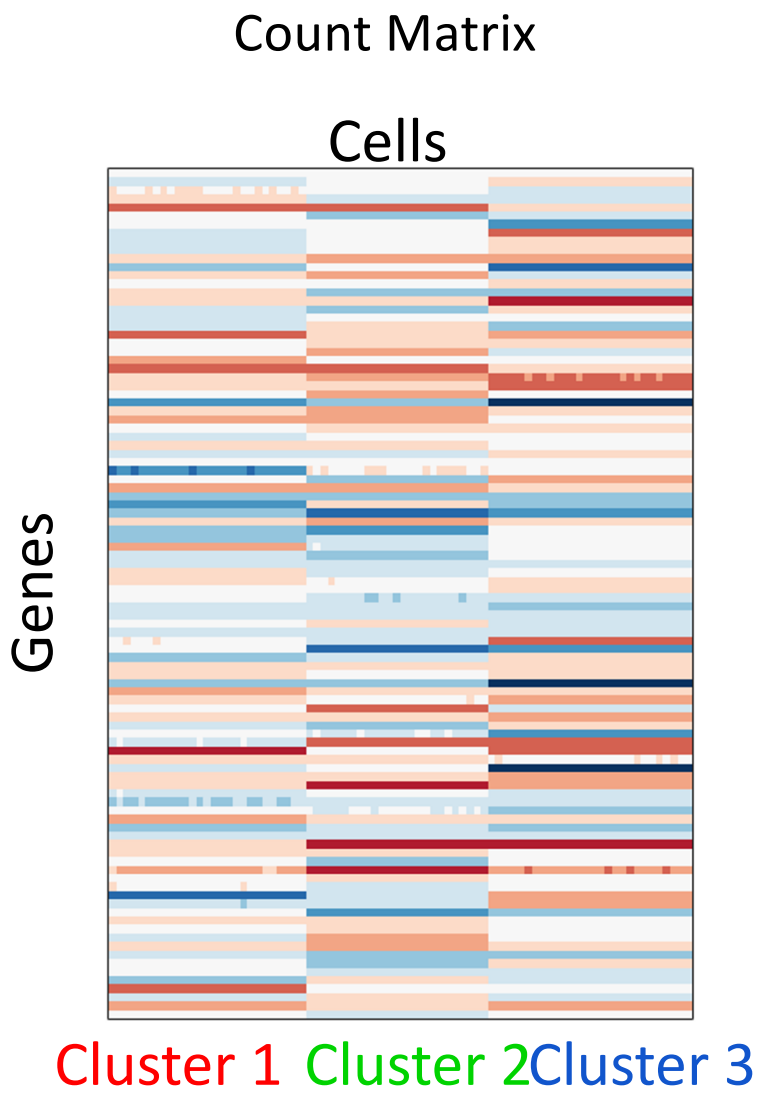
Why Normalize?

- 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?



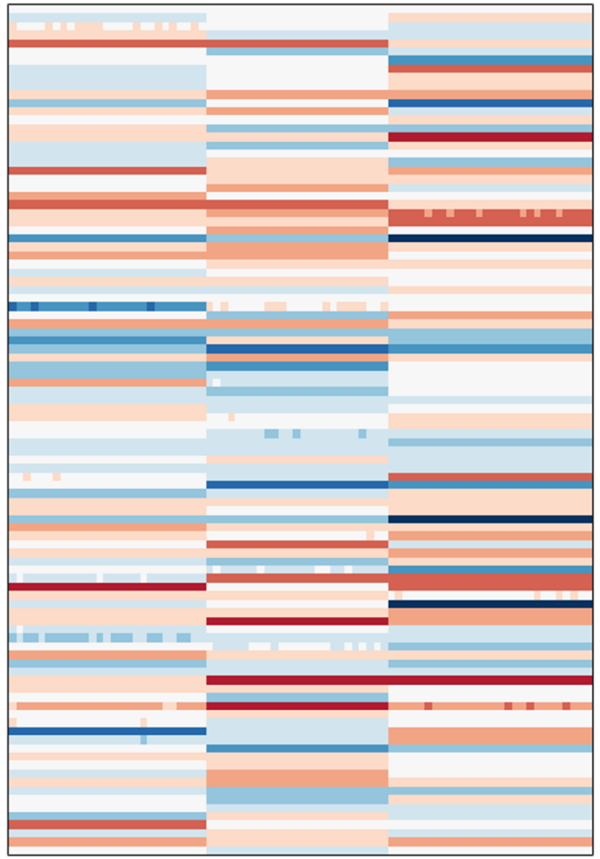
Slides courtesy of Elham Azizi

Why is this problematic?

Count Matrix

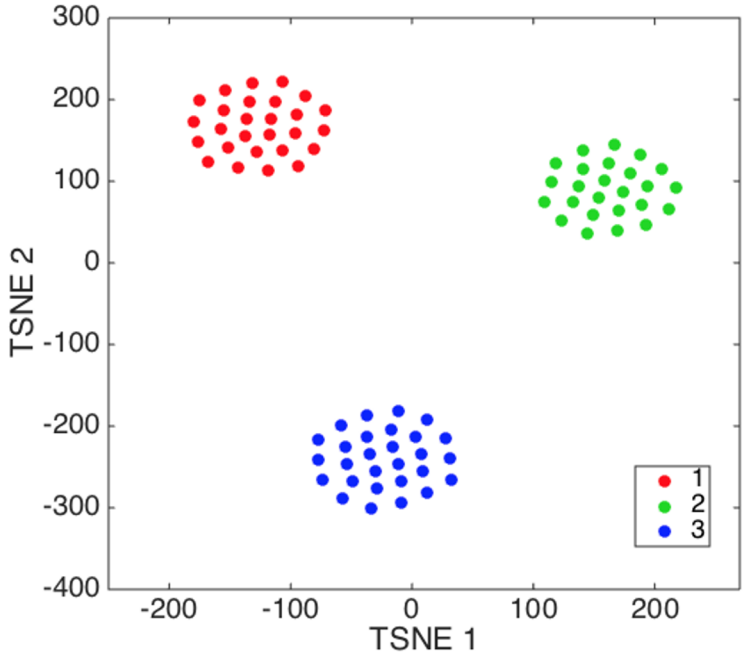
Cells

Genes



Cluster 1 Cluster 2Cluster 3

2D projection of cells (TSNE)

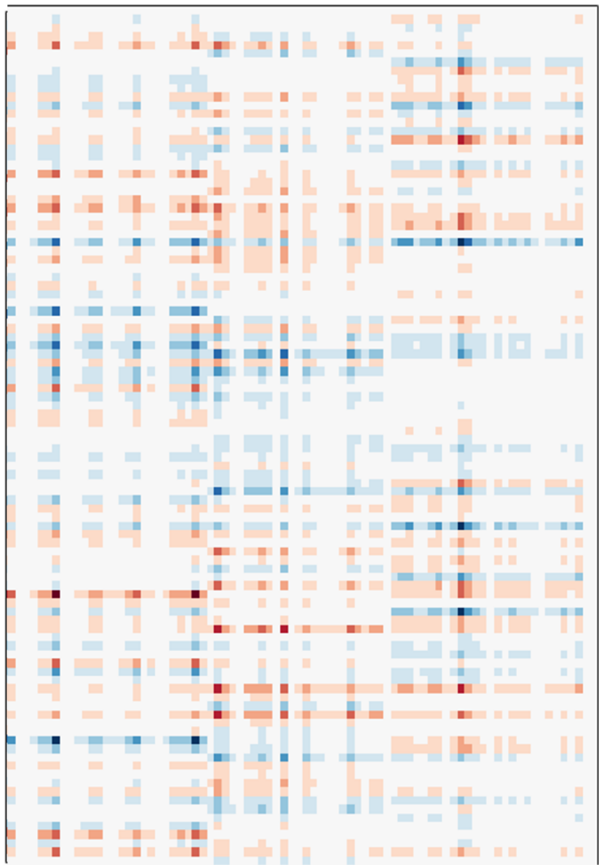


Why is this problematic?

Count Matrix

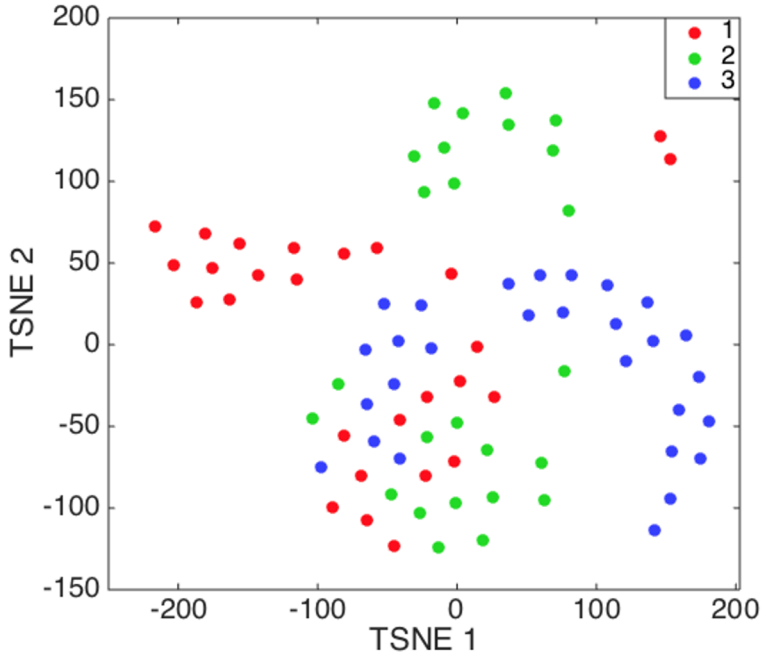
Cells

Genes



Cluster 1 Cluster 2Cluster 3

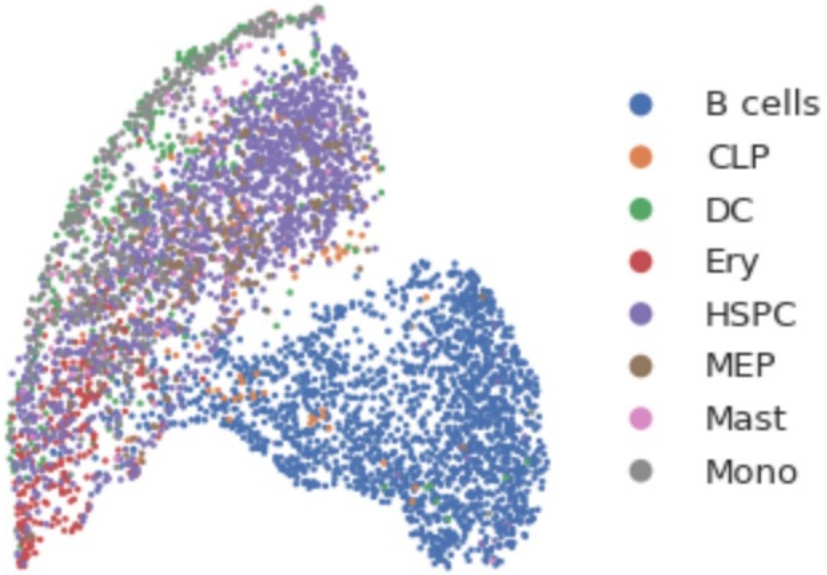
2D projection of cells (TSNE)



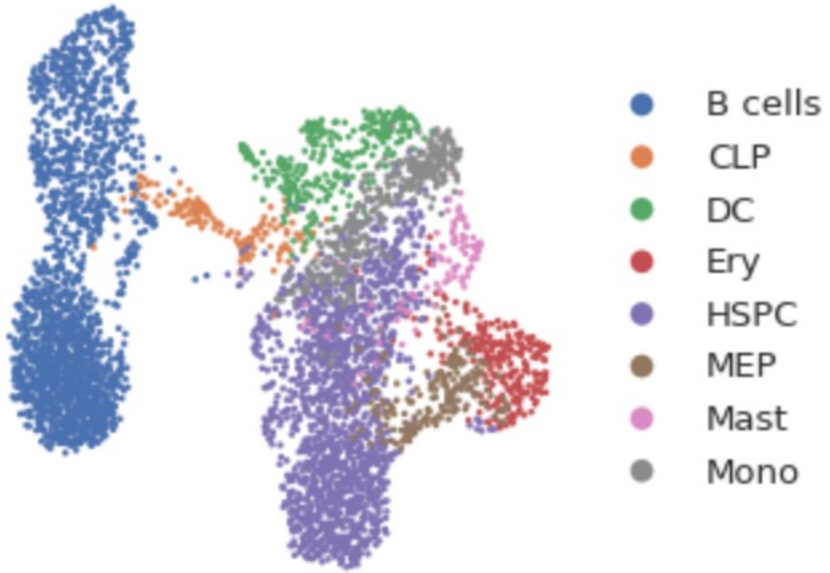
Normalization

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

No normalization

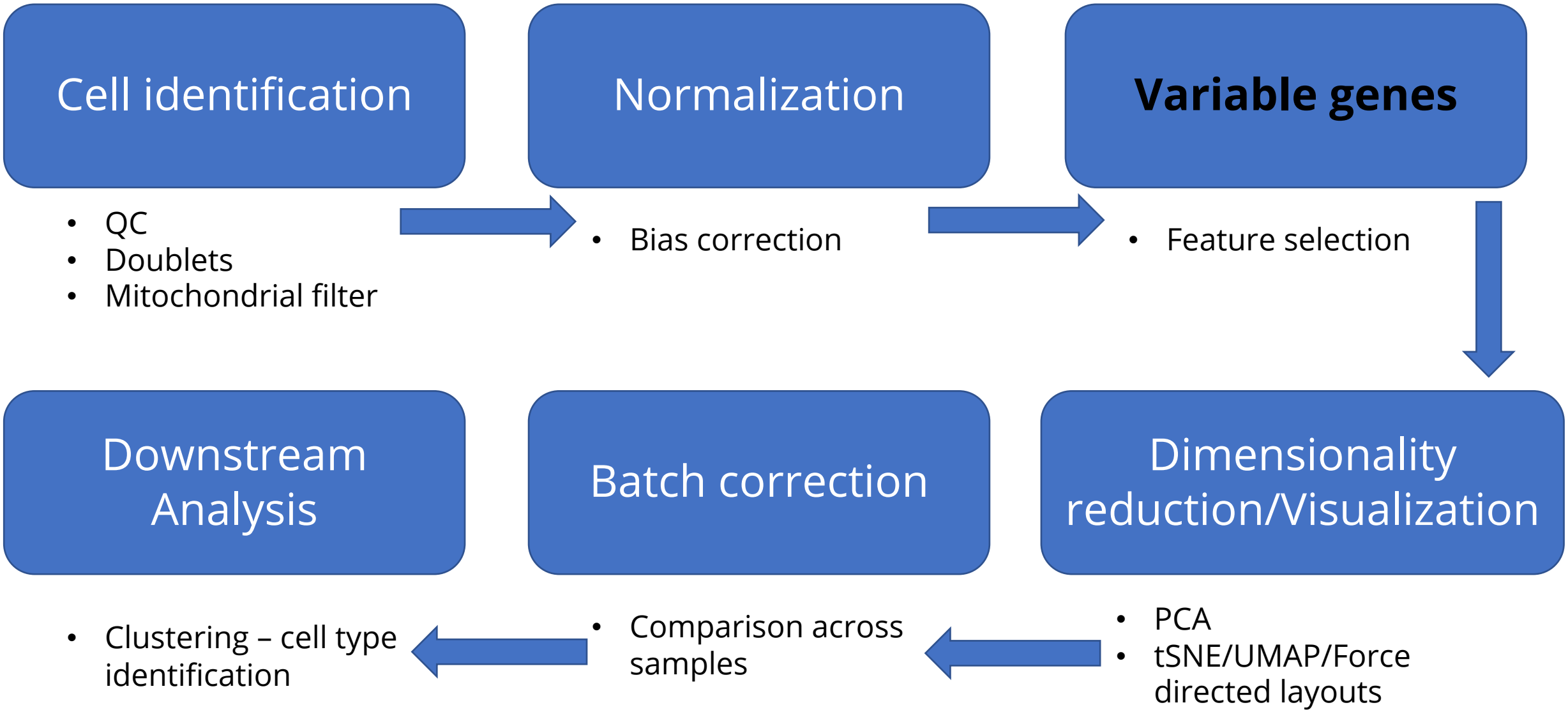


Post normalization



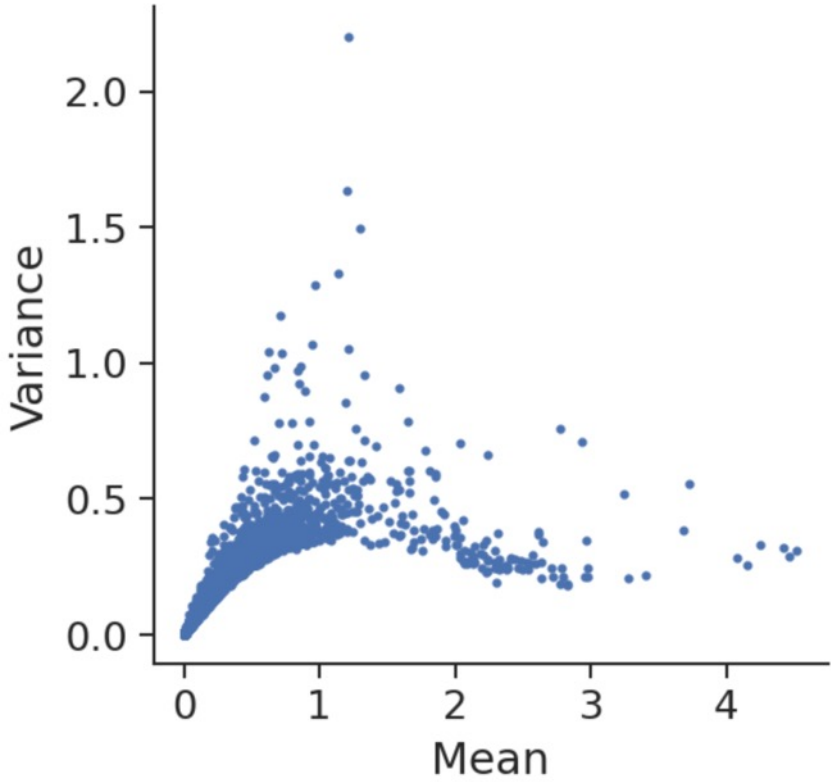
CD34+ Human bone marrow cells

scRNA-seq analysis steps



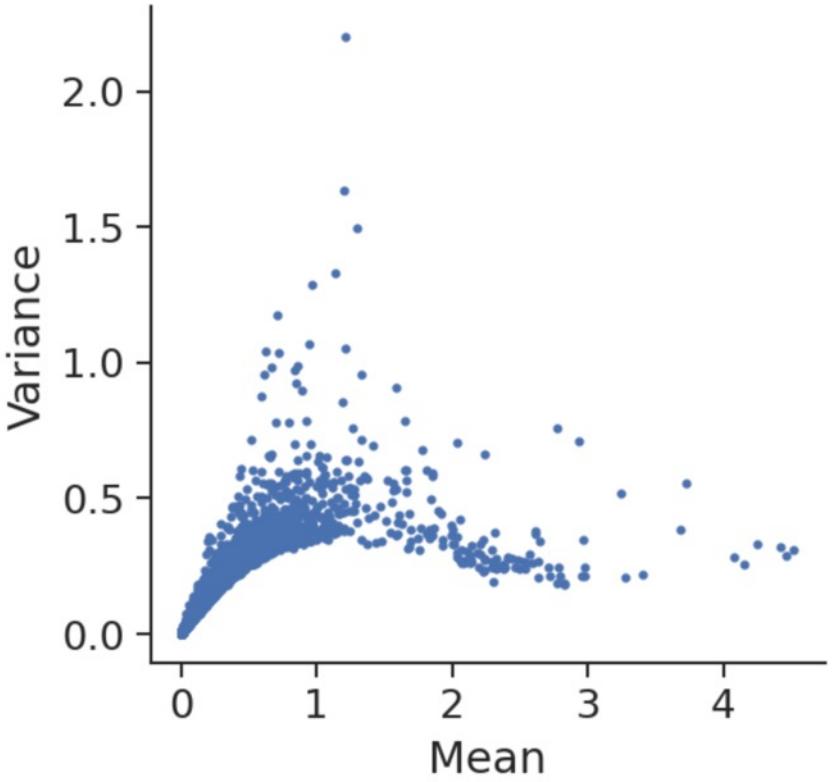
Feature selection

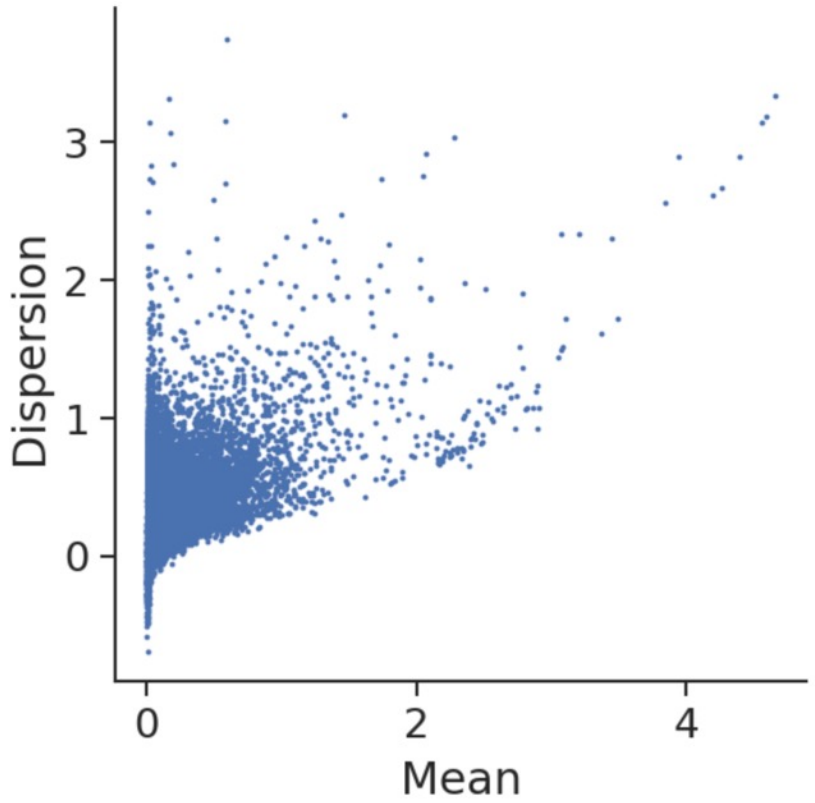
- Cell states are defined by expression of subsets of genes
- Goal of feature selection: Select genes that inform the biology rather than genes that represent random noise
- Possible Solution: How variable is the gene across cells?



Feature selection

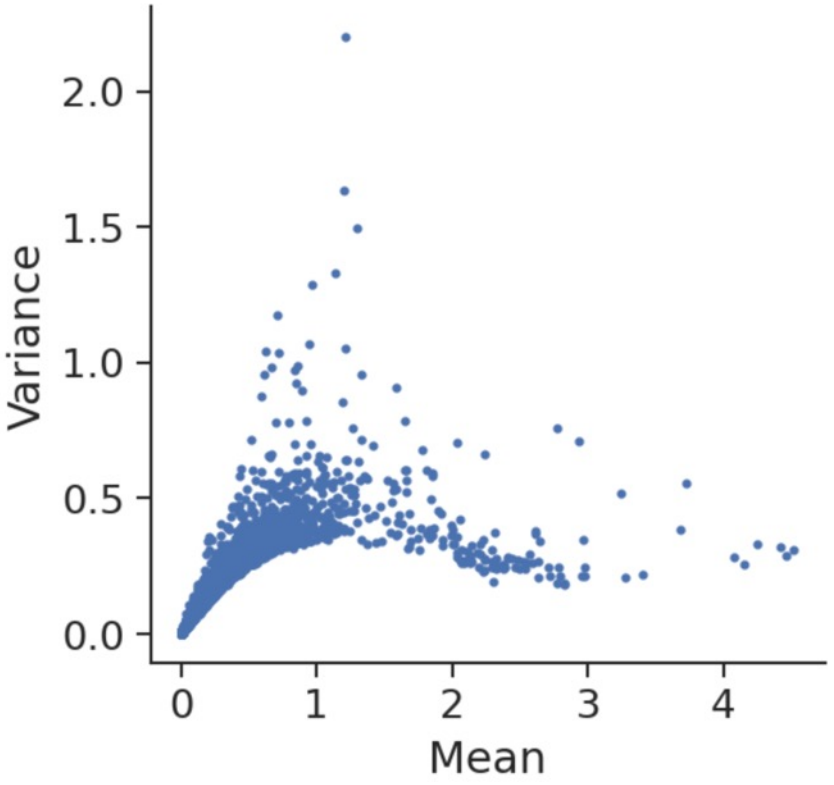
$$Dispersion = \frac{Variance}{Mean}$$

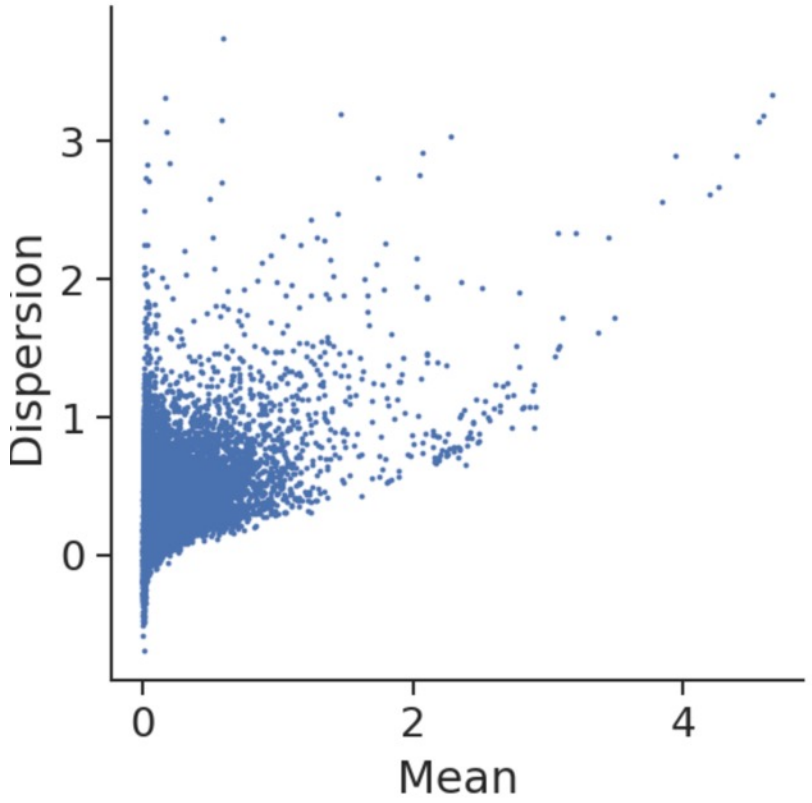


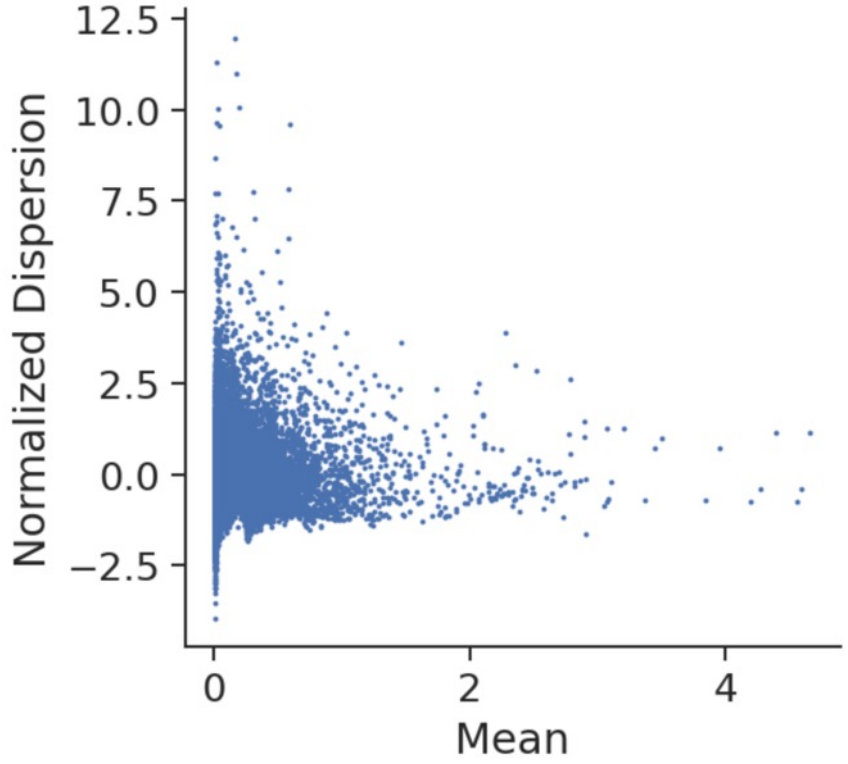


Feature selection: Share information across genes

$$Normalized\ Dispersion = \frac{(Dispersion\ - Mean(Dispersion\ in\ expression\ mean\ bin)}{Std(Dispersion\ in\ expresion\ mean\ bin)}$$

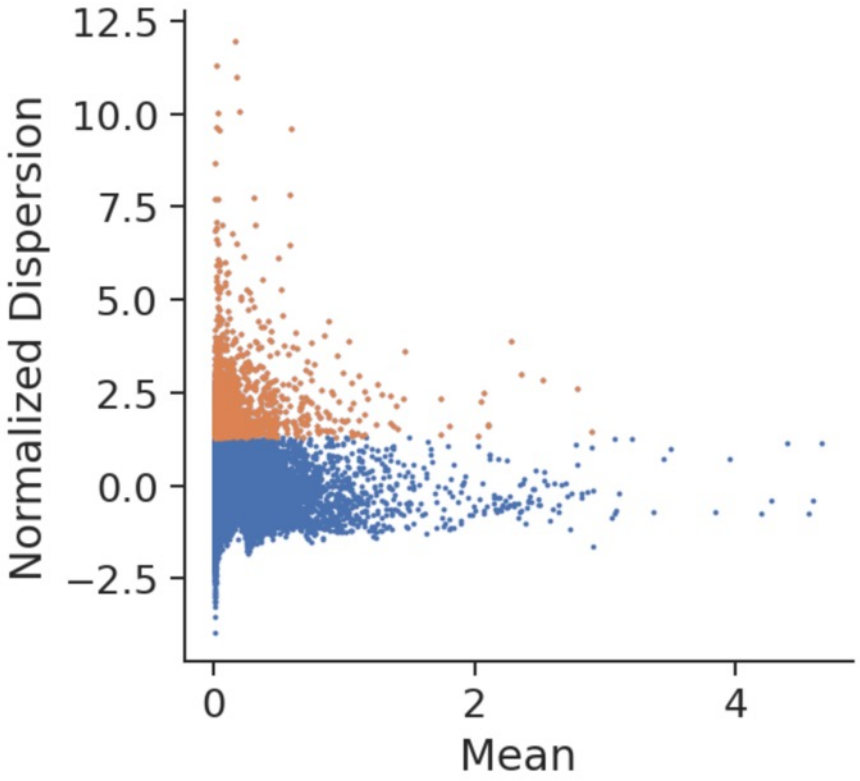






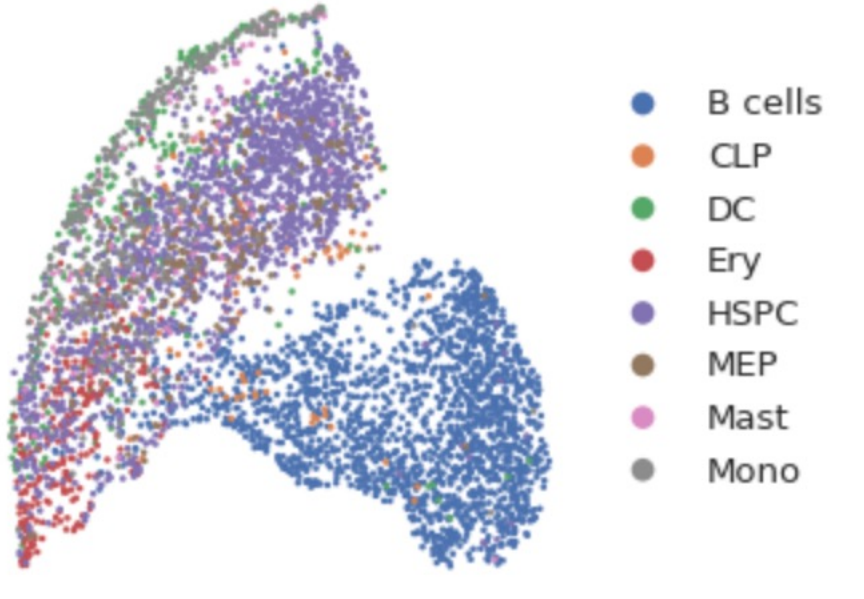
Feature selection: Share information across genes

$$Normalized\ Dispersion = \frac{(Dispersion\ - Mean(Dispersion\ in\ expression\ mean\ bin)}{Std(Dispersion\ in\ expresion\ mean\ bin)}$$

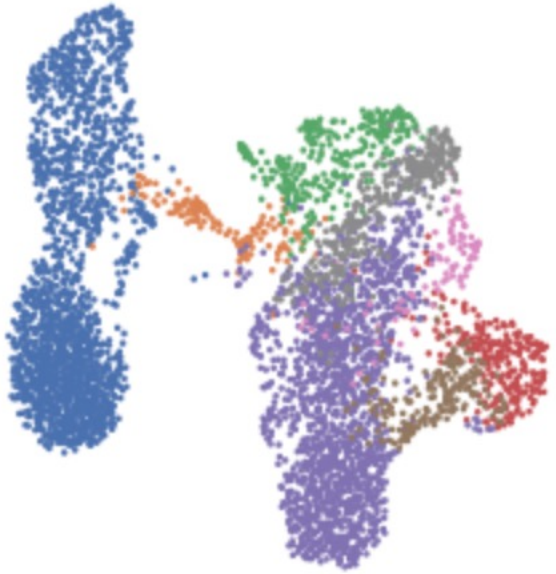


Feature selection: highly variable genes

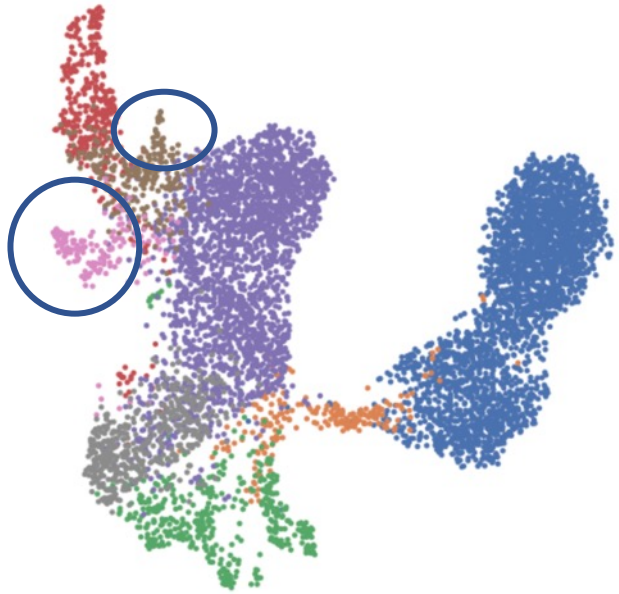
No normalization



Post normalization

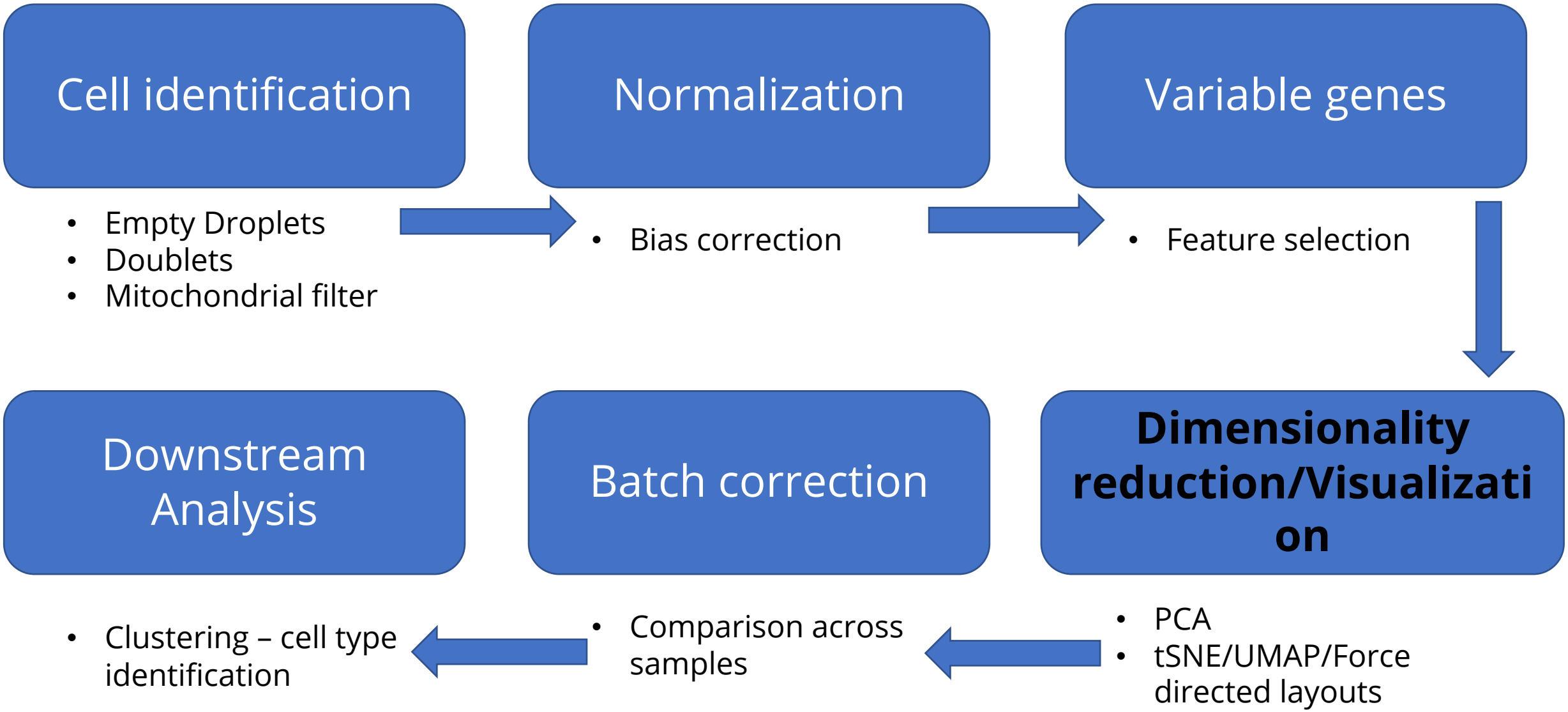


Feature selection



CD34+ Human bone marrow cells

scRNA-seq analysis steps

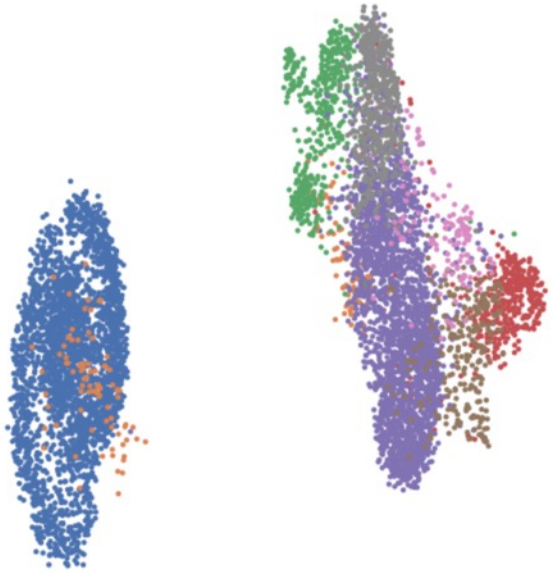


Measuring distance between cells

- Possible Solution: Euclidean distance between normalized, selected genes

Feature selection

Euclidean distance
In gene expression space

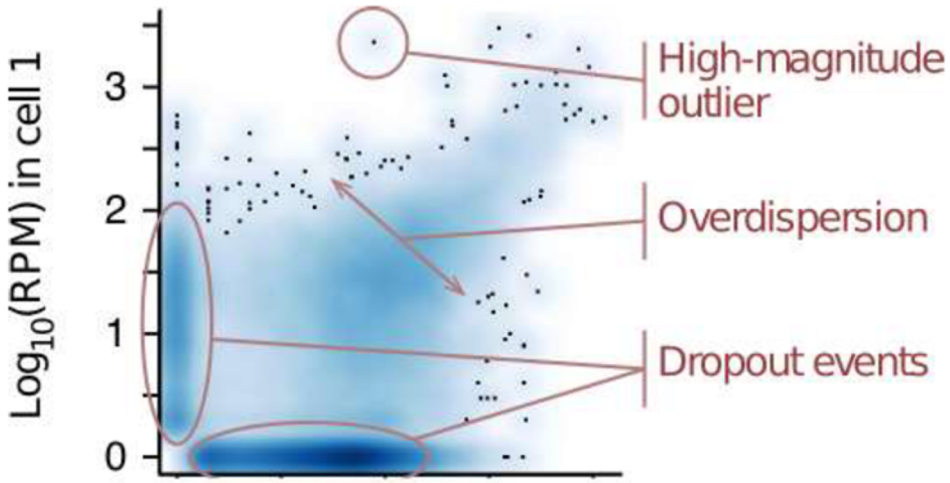


CD34+ Human bone marrow cells

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 under-estimates of true counts



BRIEF COMMUNICATIONS

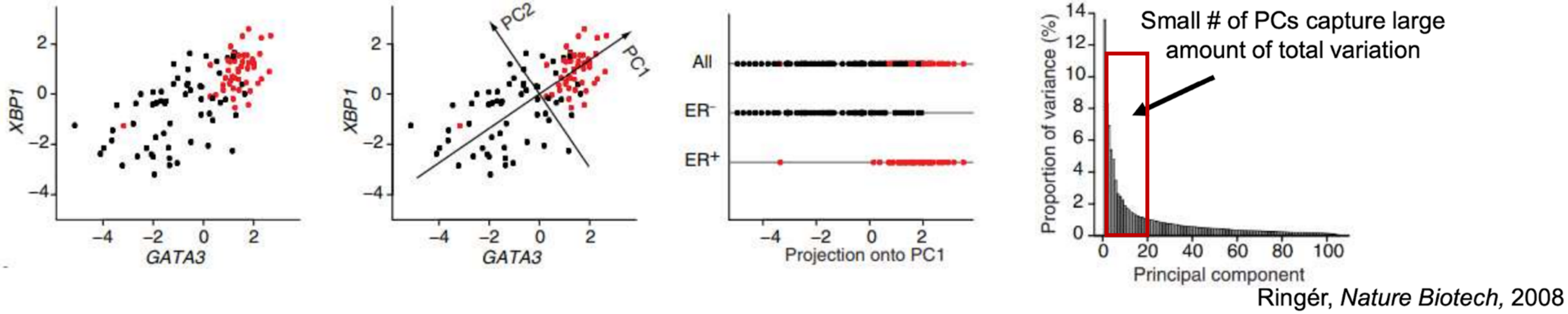
Bayesian approach to single-cell differential expression analysis

Peter V Kharchenko[1–3], Lev Silberstein[3–5] & David T Scadden[3–5]

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

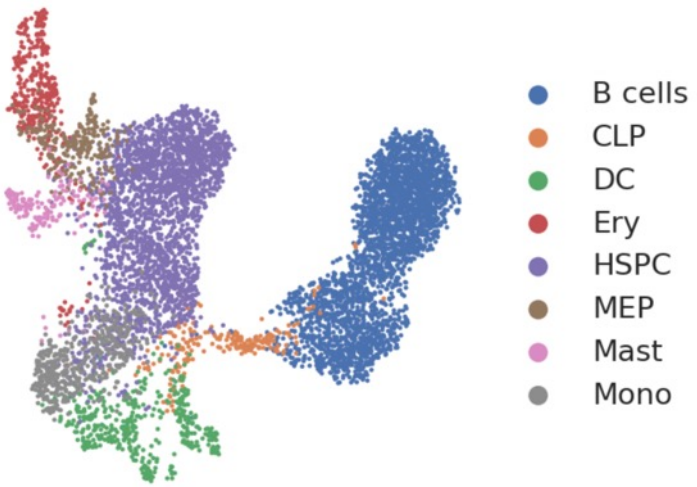
- Selection of subset of PCs: Dimensionality reduction

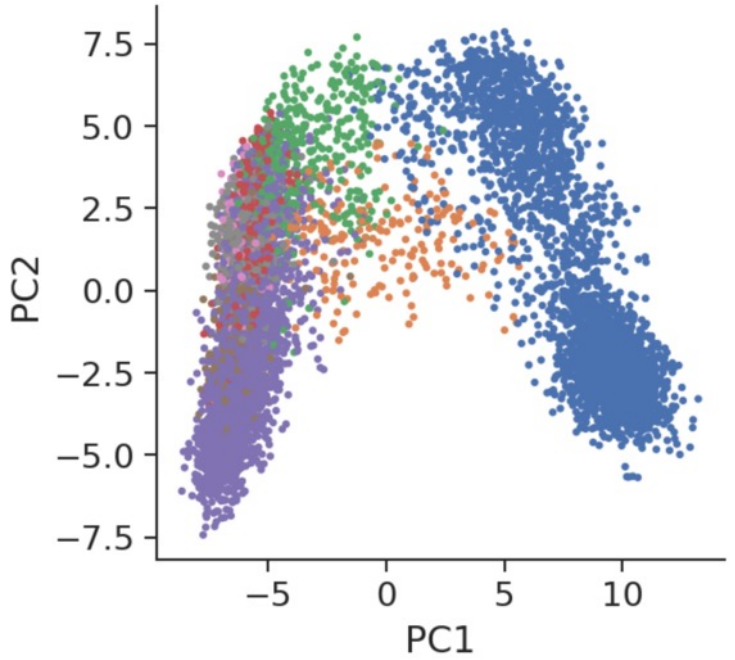
Dimensionality Reduction

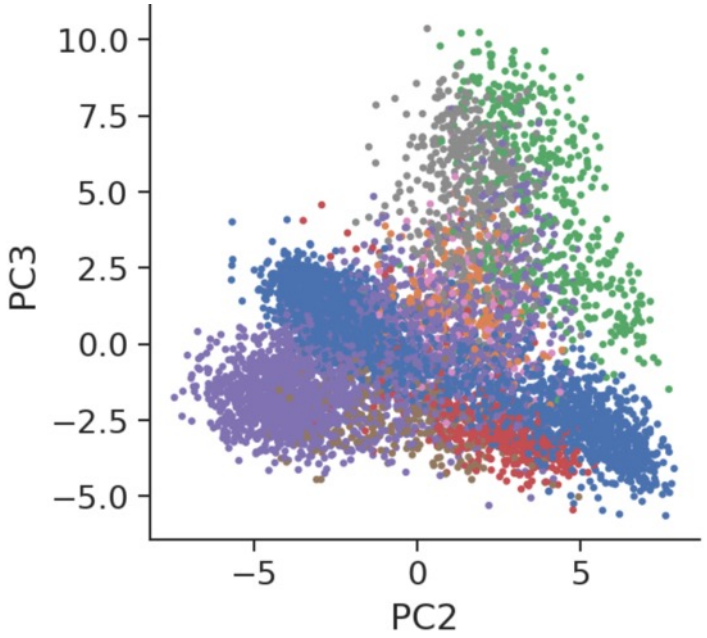
- Reduce the number of dimensions of data while preserving high dimensional information

- Overcome noise in high dimensions
- Computational efficiency
- Visualization

PCA for visualization





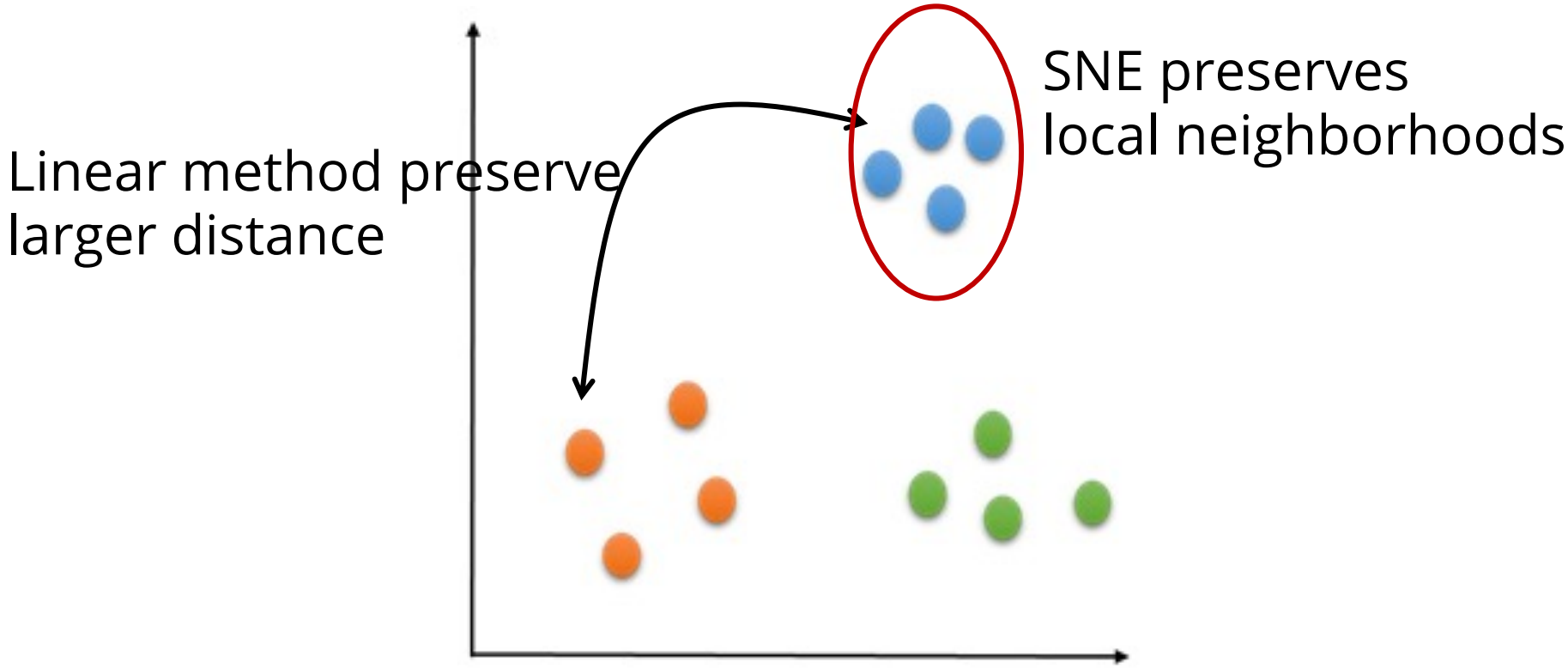


CD34+ Human bone marrow cells

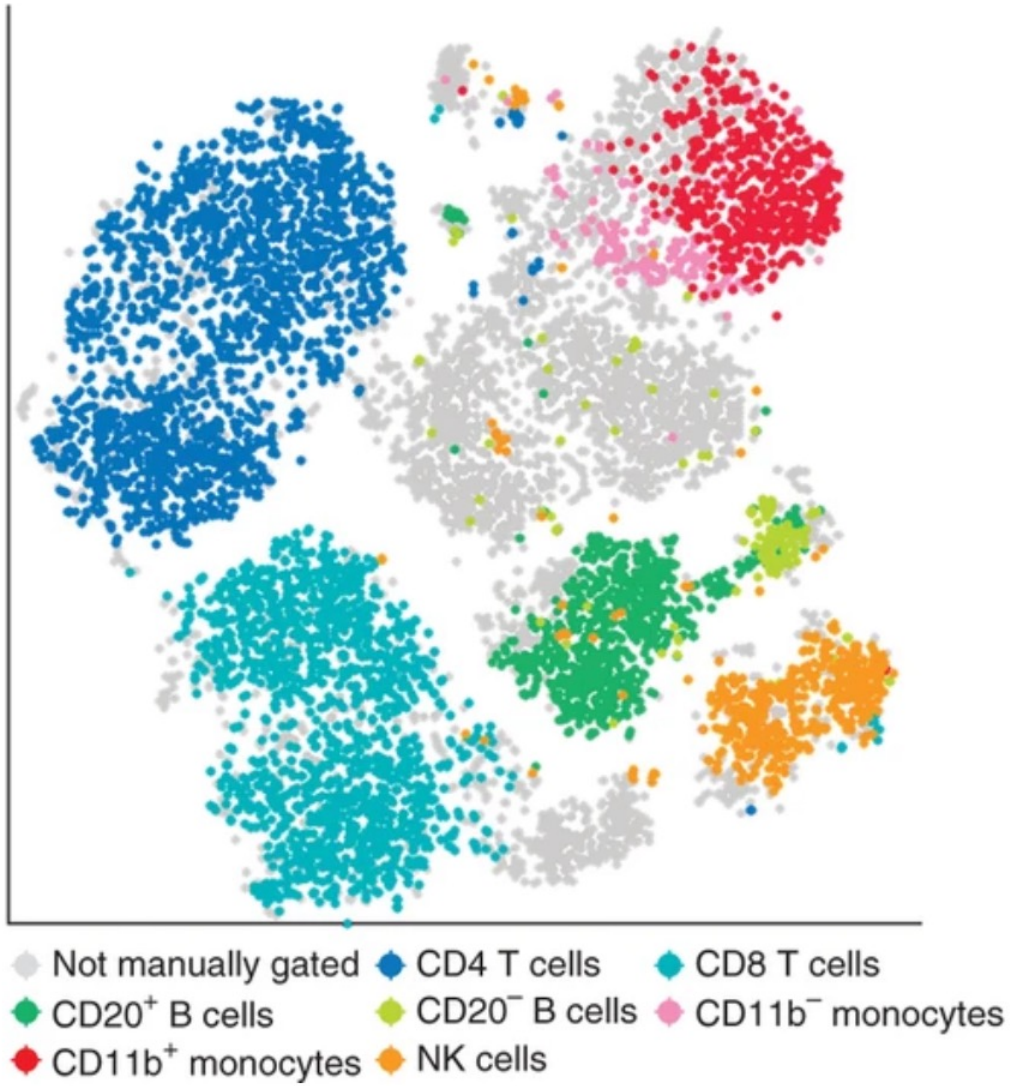
- Two axes can only capture so much information
  - Not explicitly modeled to capture as much of higher order information in 2D

Dimensionality Reduction for Visualization: tSNE

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



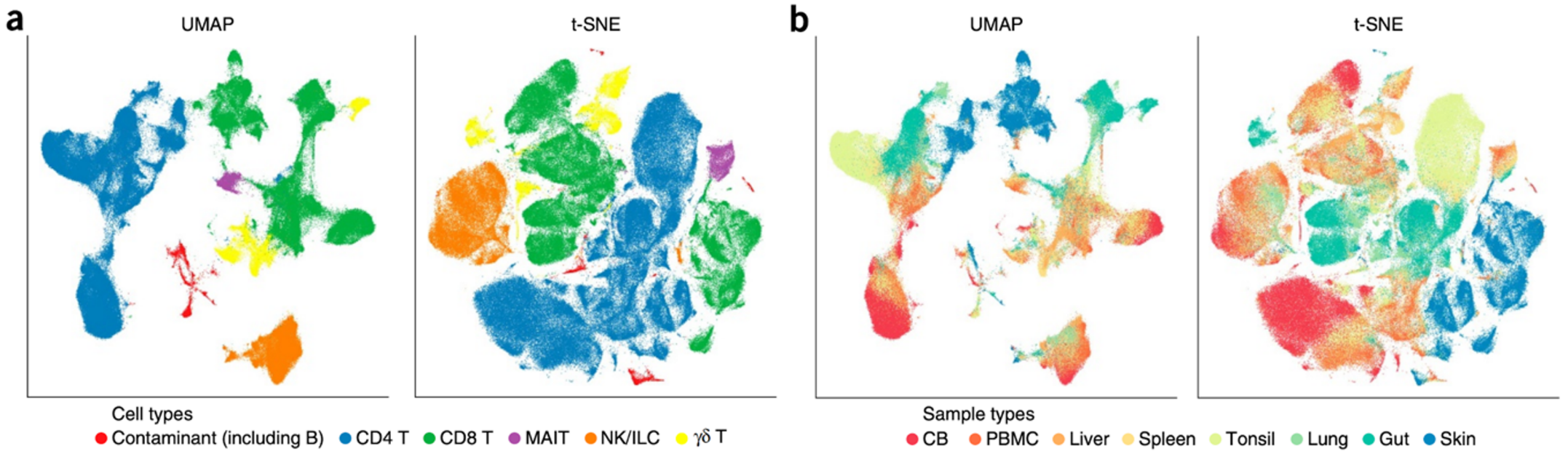
tSNE for single-cells



David Amir et. al, 2013

Visualization: UMAP

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



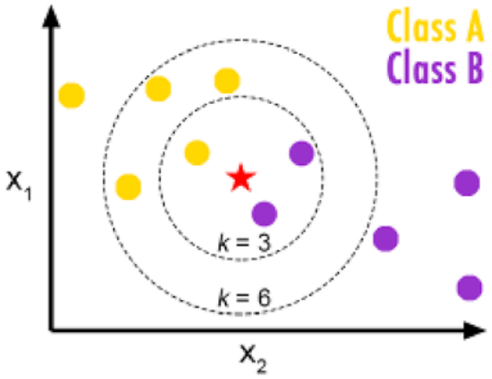
Becht et. al., 2019

Visualization: Force directed layouts

Goal: To visualize the structure of our data

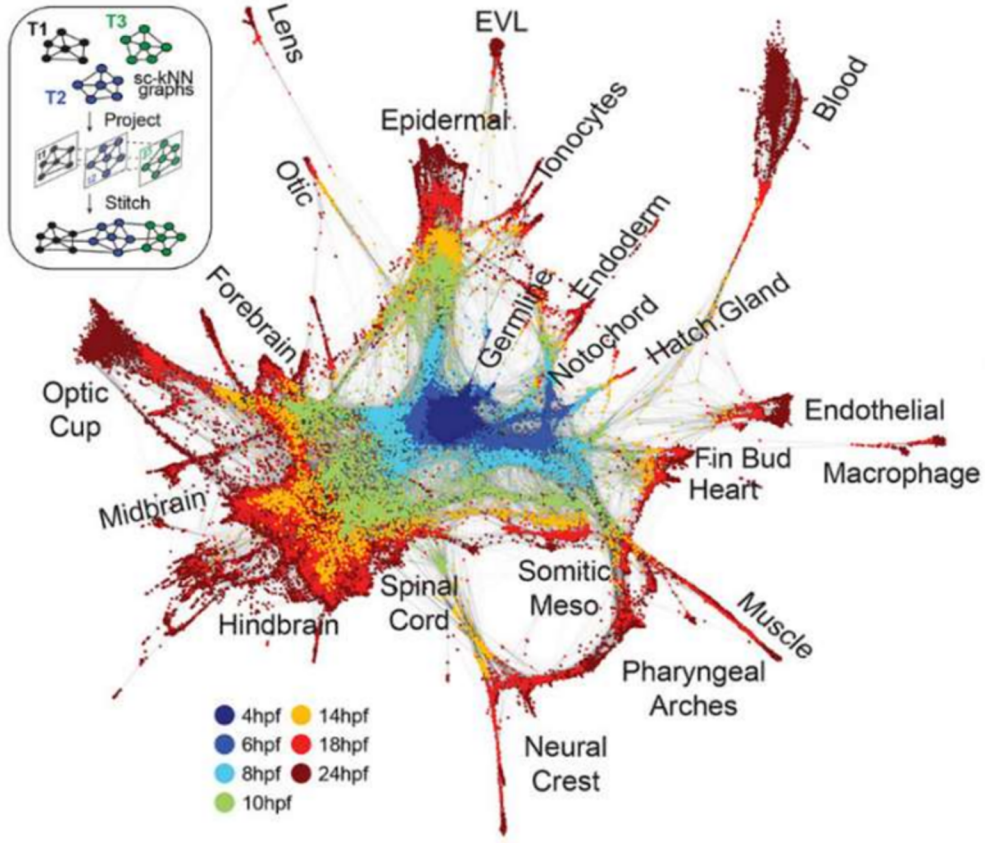
Force-directed graphs

Visualize cells based on nearest neighbor structures



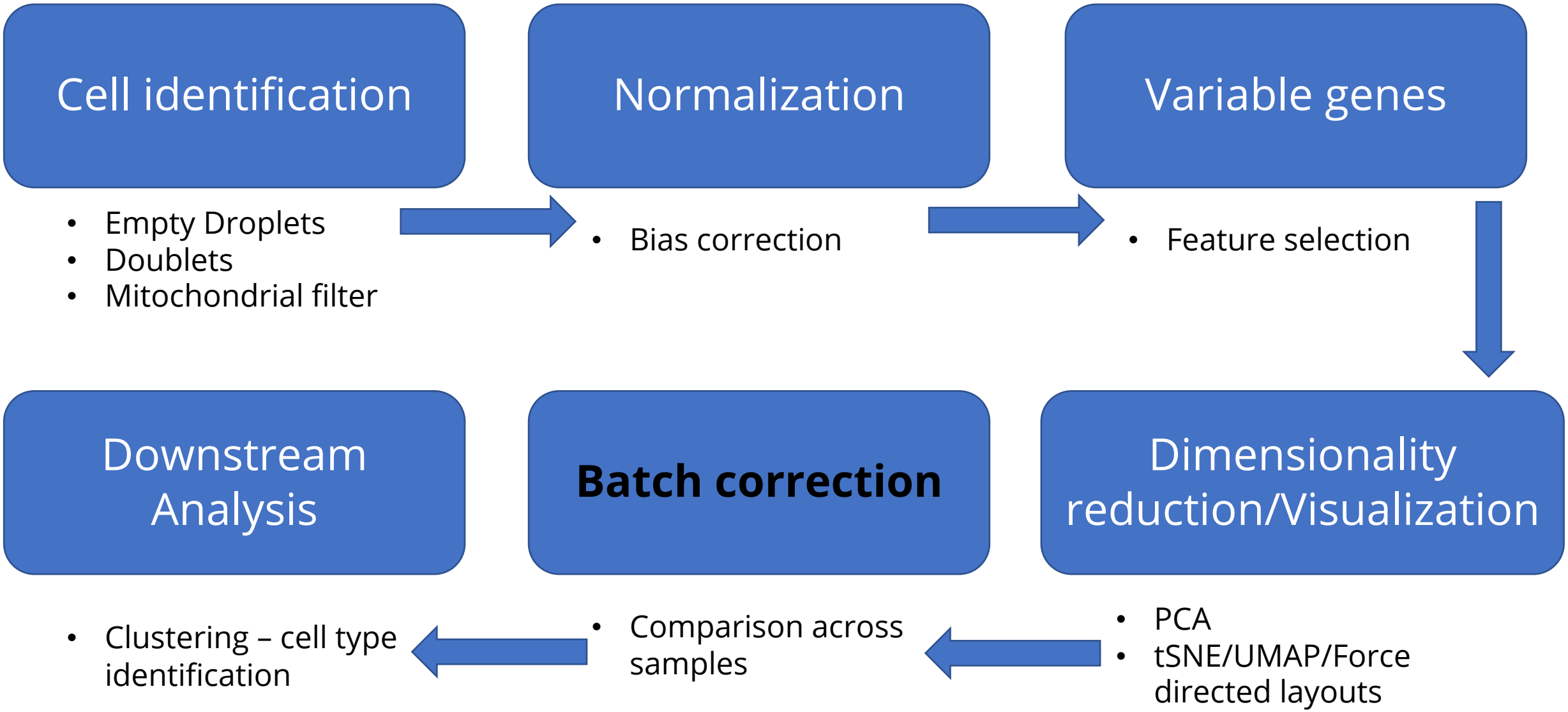
Repulsion between nodes

Attractive forces added to edges connecting nodes (spring functions)



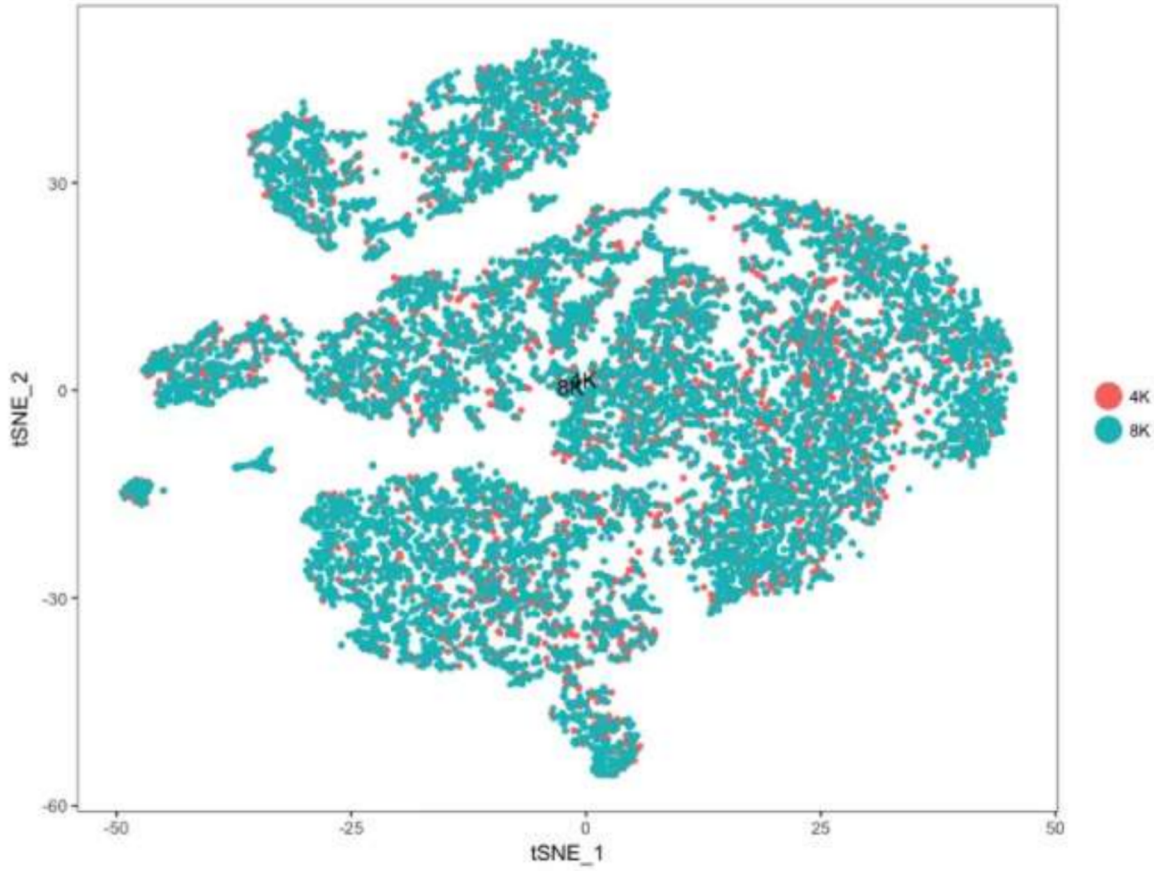
Wagner *et al., Science,* 2018

scRNA-seq analysis steps

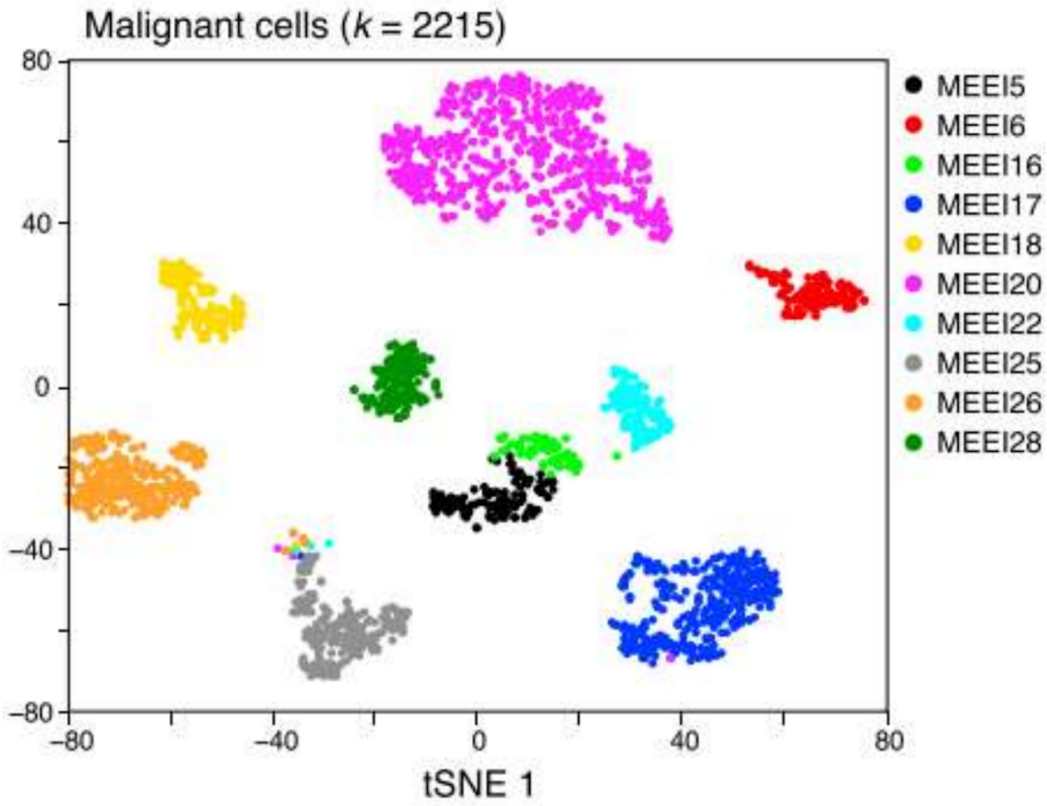


Real heterogeneity –vs- technical noise

Technical replicate of PBMCs has near-perfect overlap

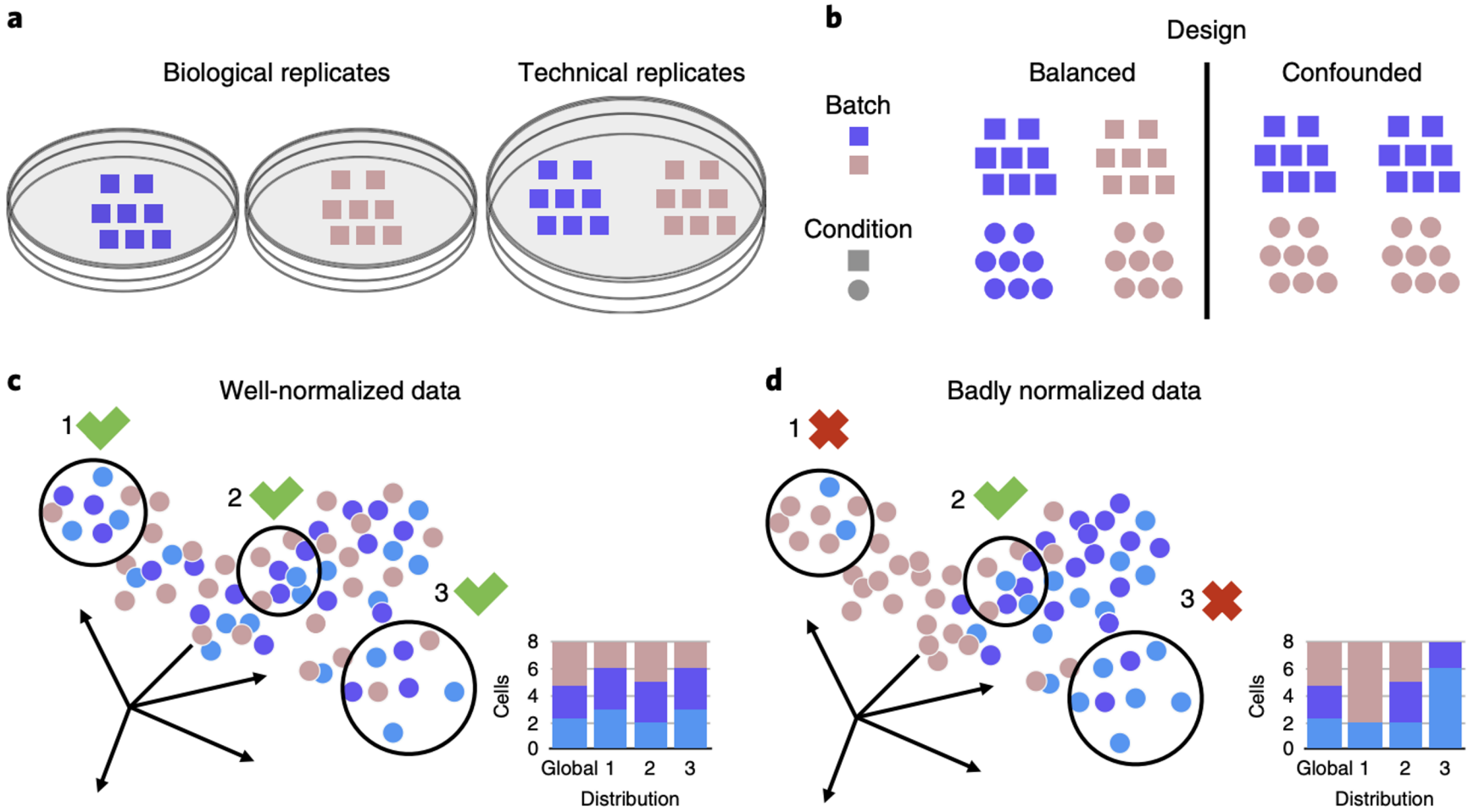


Cancer cells dramatically different between patients



Assessing normalization and batch effects

Evaluate mixing of samples



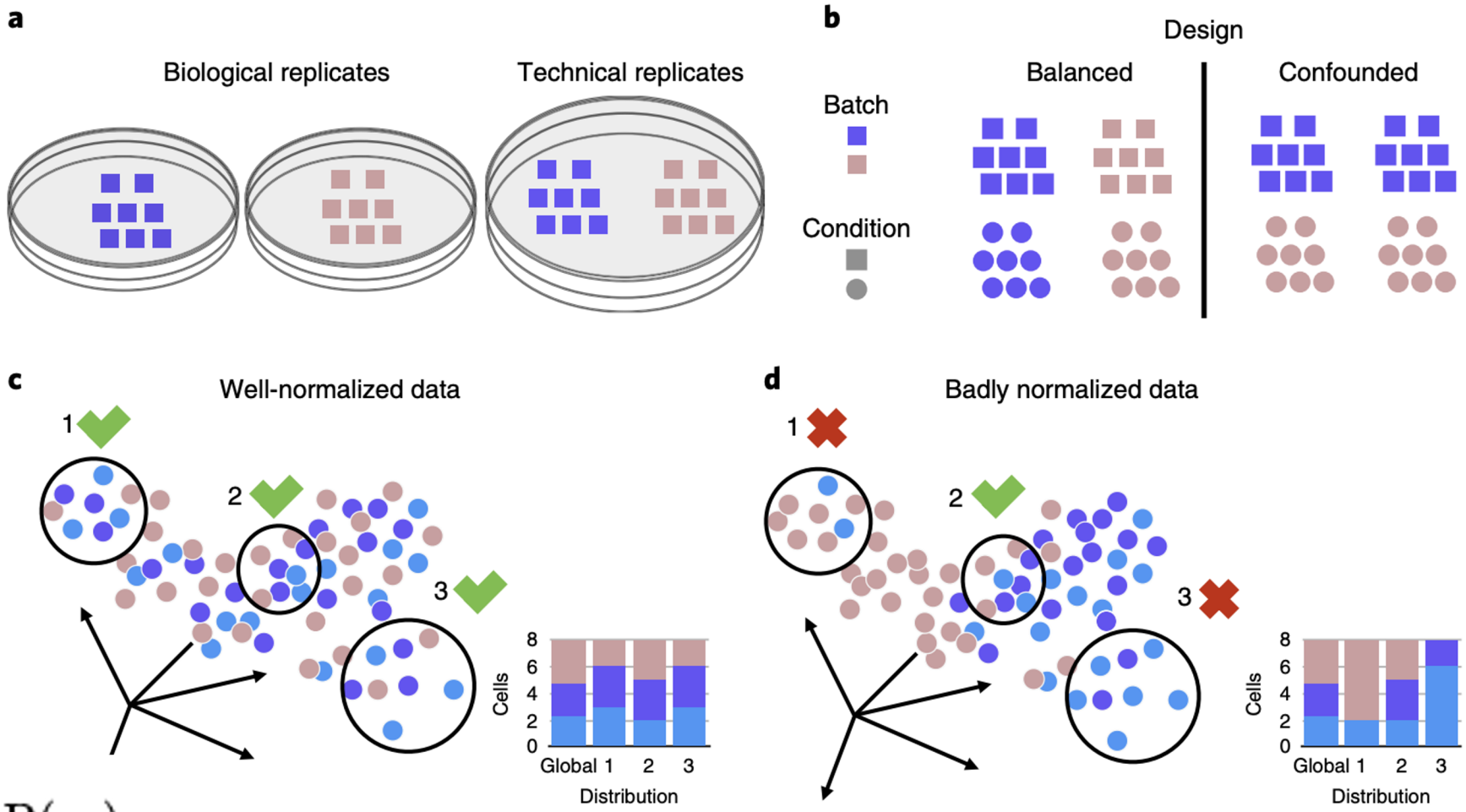
Buttner et. al, 2018

Assessing normalization and batch effects

Select random neighborhoods of fixed size

Compute Shannon Entropy of distribution across samples

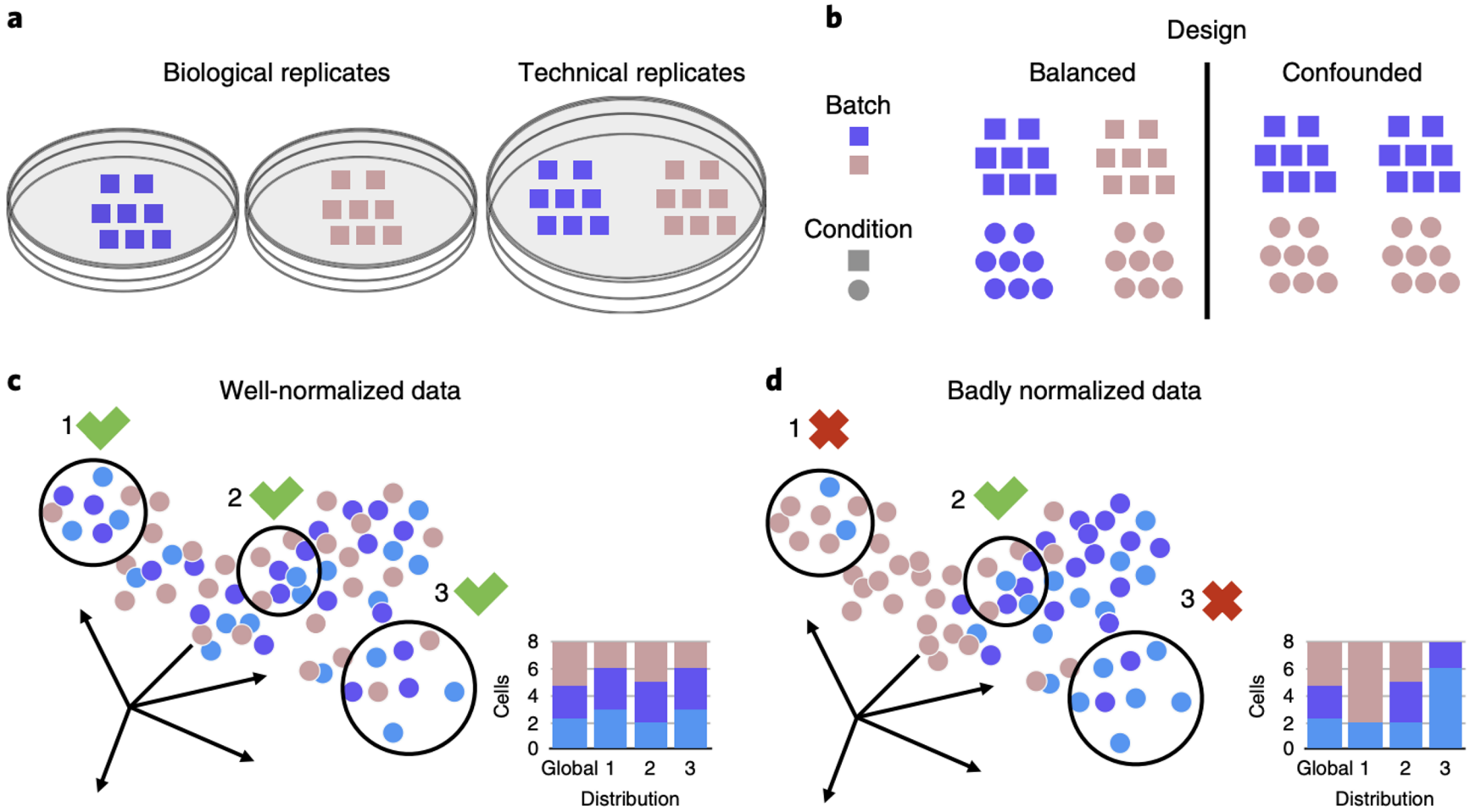
$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$$



Buttner et. al, 2018

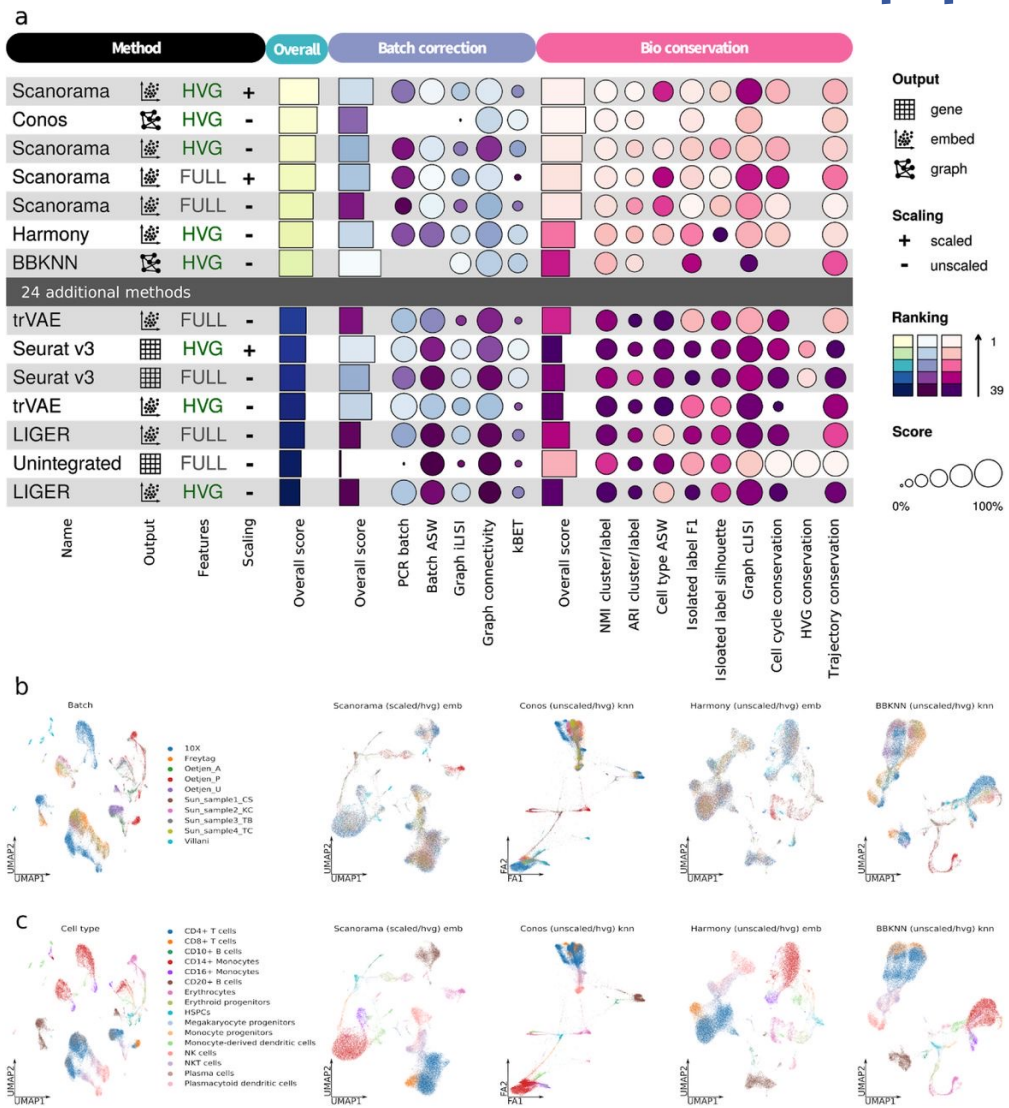
Assessing normalization and batch effects

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



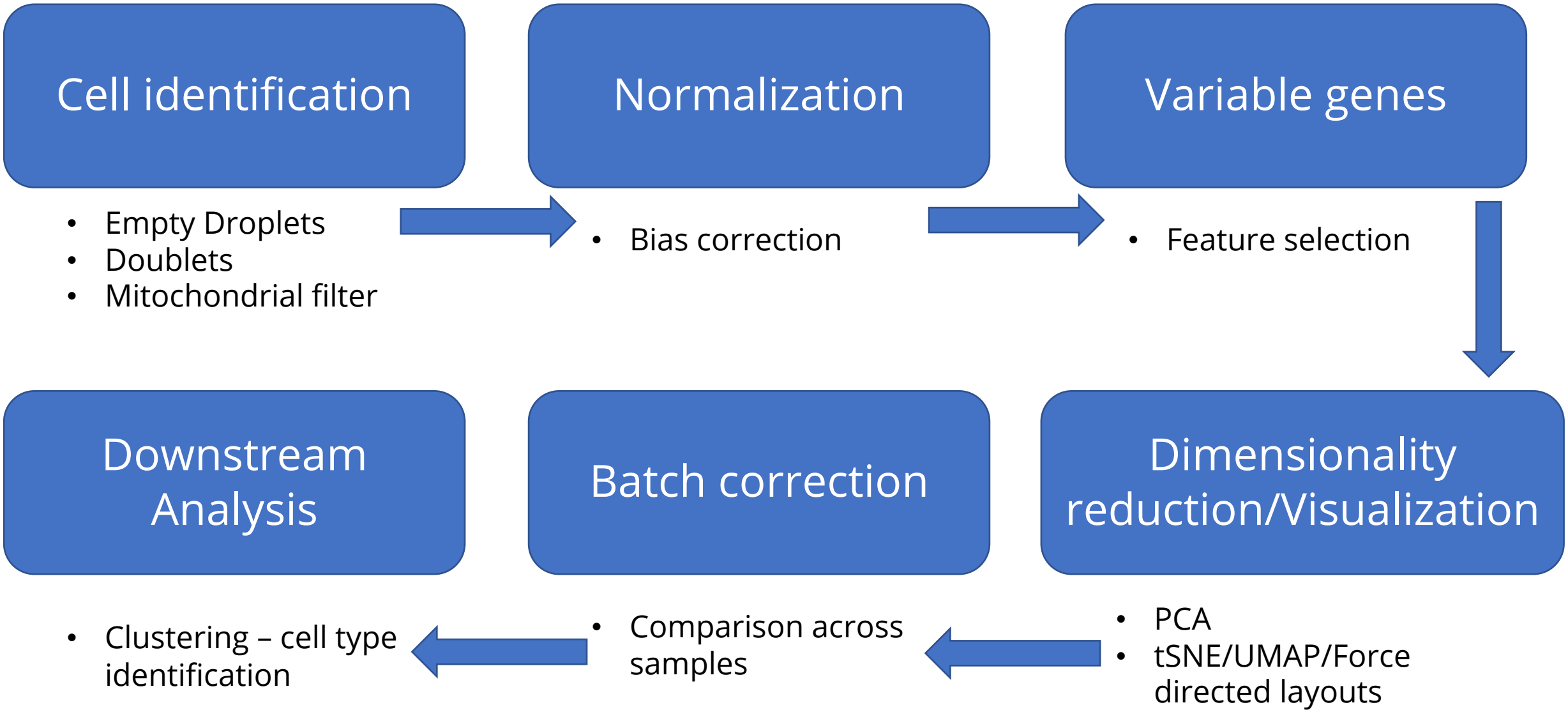
Buttner et. al, 2018

Batch effect correction approaches



Luecken et. al, 2020

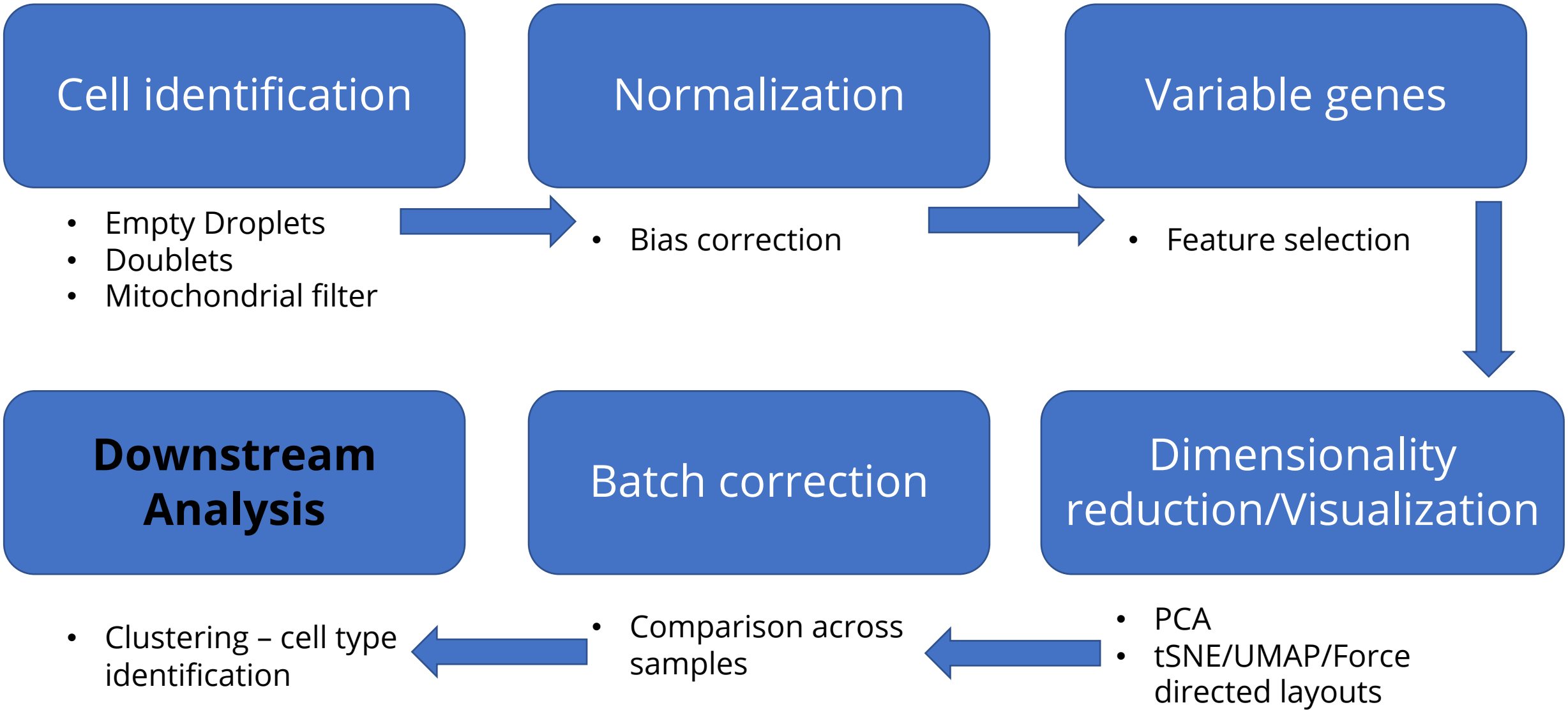
scRNA-seq analysis steps



Other corrections

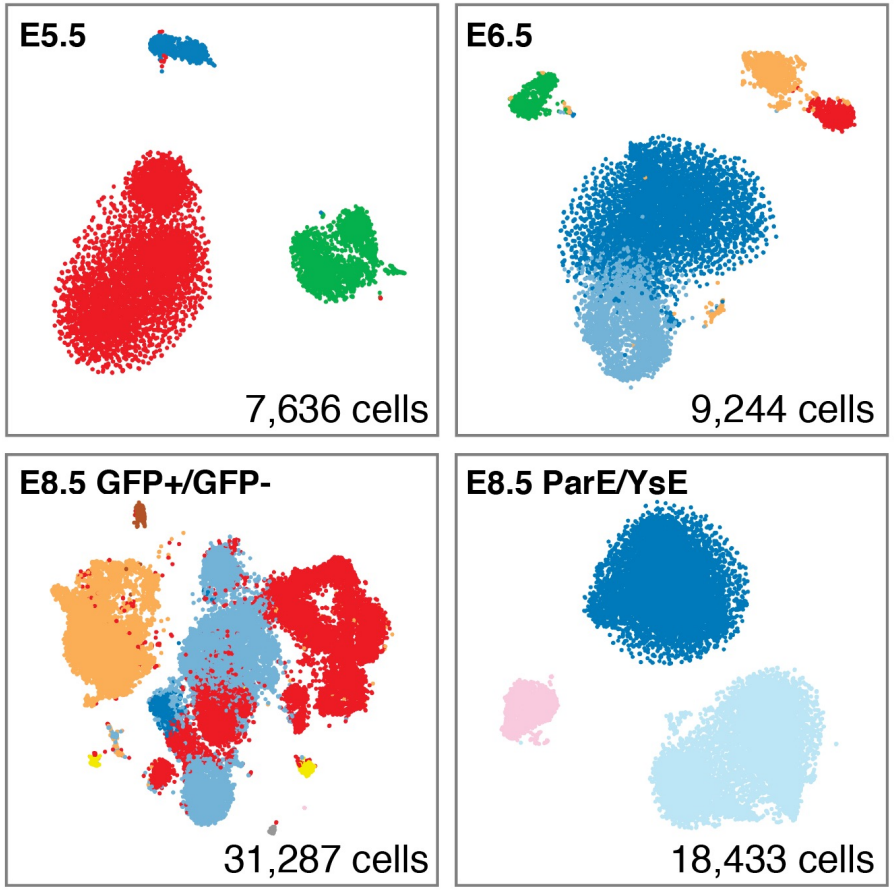
- Cell cycle correction
- Gene expression imputation

scRNA-seq analysis steps

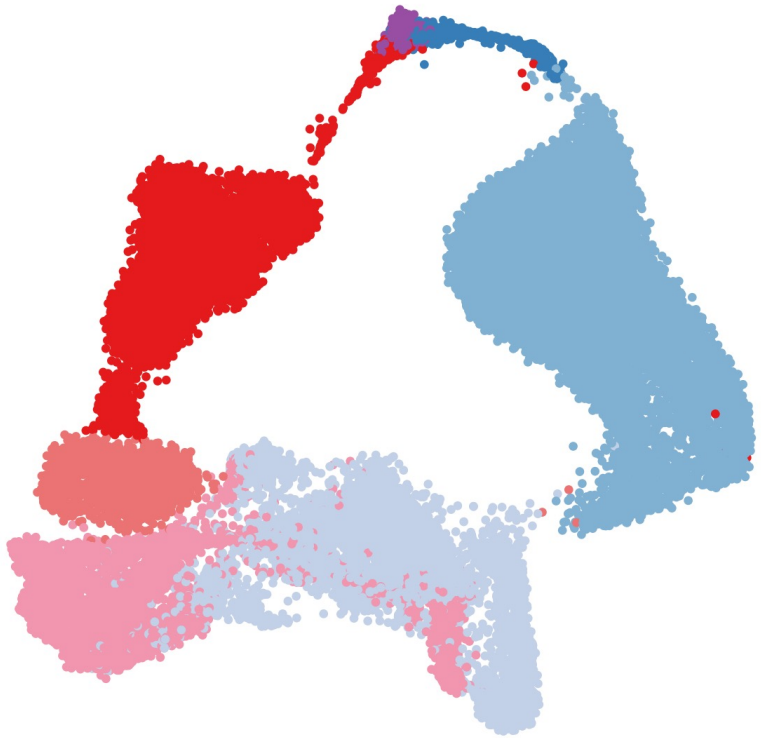


Single-cell data

Discrete clusters



Continuous trajectories

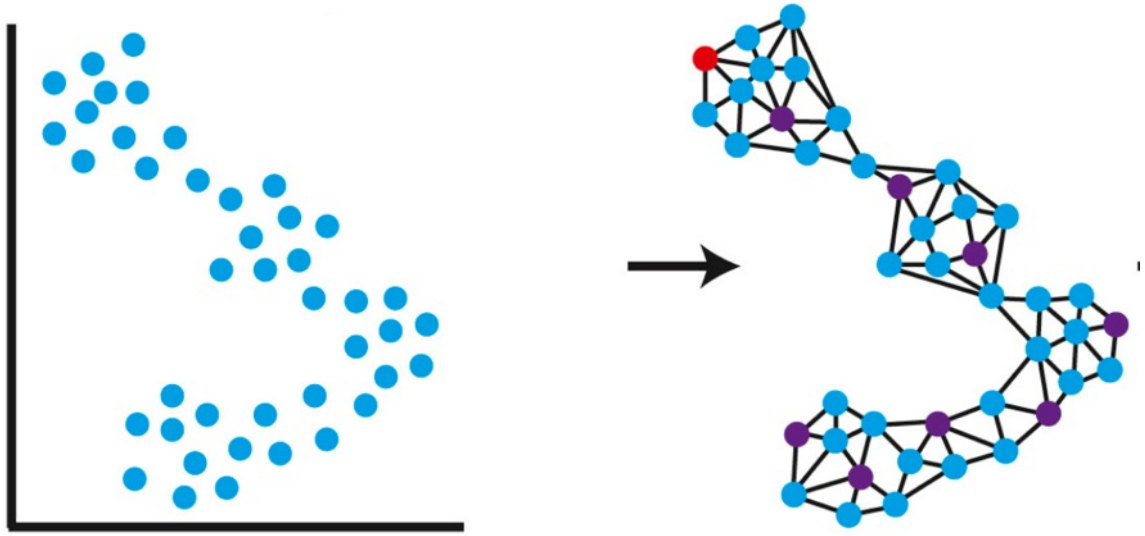


ICM
PrE
VE
Gut tube: VE
EPI
DE
Gut tube: DE

Nearest neighbor graphs

- For each point, find k nearest points using Euclidean distance

$$Graph\ G = (V, E)$$
$$V: Set\ of\ vertices$$
$$E: Set\ of\ edges$$



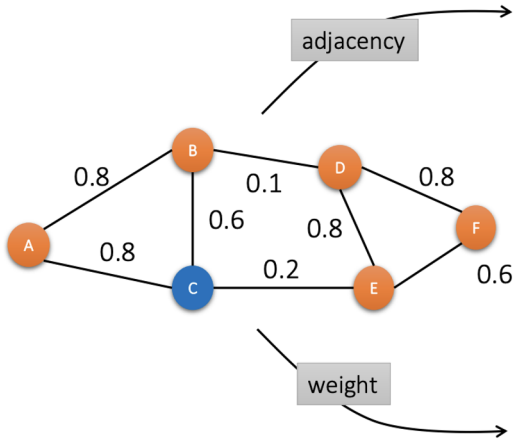
Graph adjacency matrices

- Graph can also be a represented as an adjacency matrix o

$$Graph\ G = (V, E)$$

$$Adjacency\ matrix\ A \in \mathbb{R}^{|V| \times |V|}$$

$$A_{ij} = E(i, j)$$

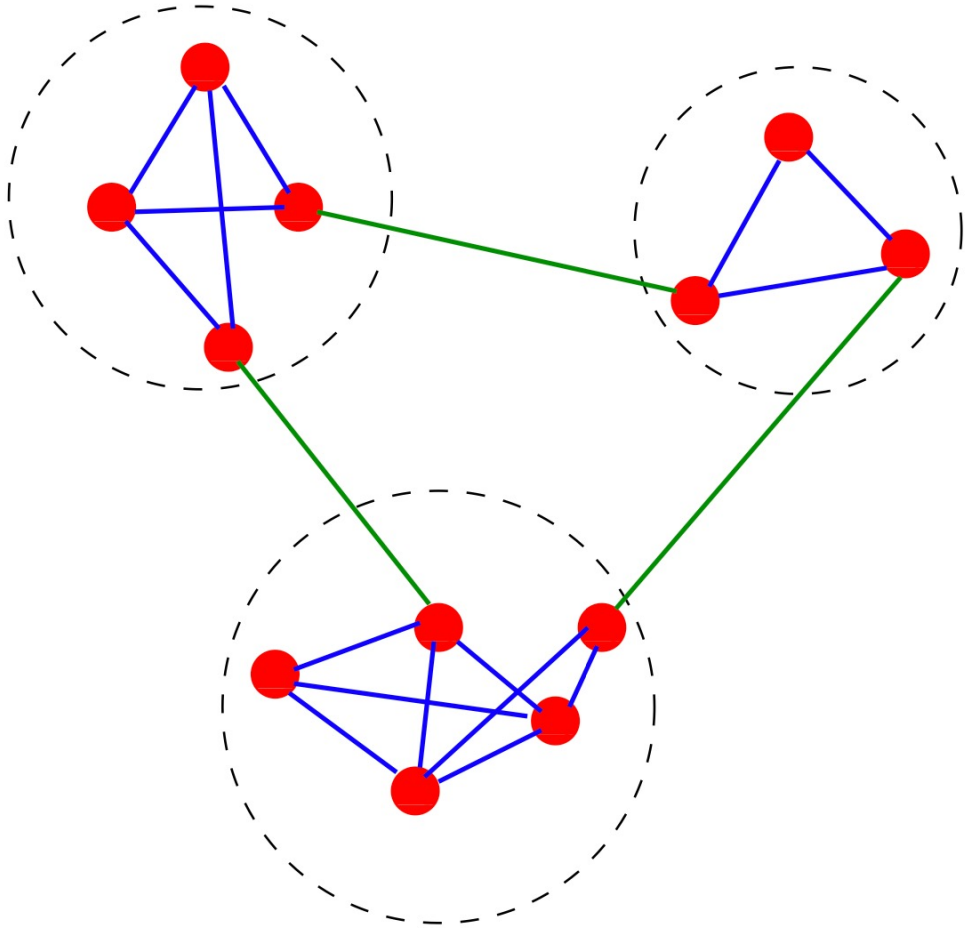


$$A = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

$$W = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0 & 0.6 & 0.1 & 0 & 0 \\ 0.8 & 0.6 & 0 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0 & 0.8 & 0.8 \\ 0 & 0 & 0.2 & 0.8 & 0 & 0.6 \\ 0 & 0 & 0 & 0.8 & 0.6 & 0 \end{pmatrix} \end{array}$$

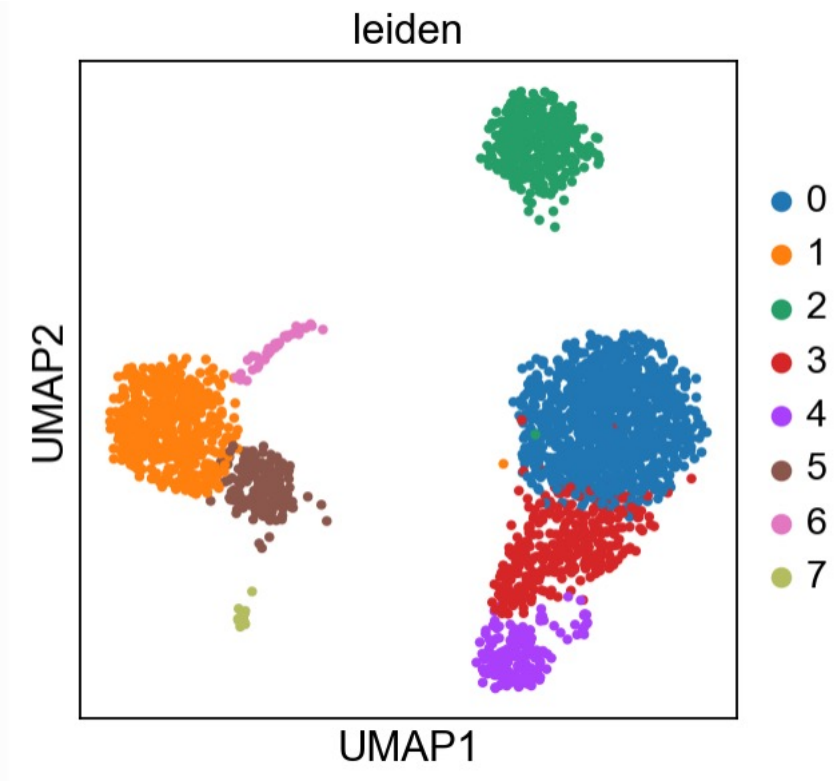
Graph based clustering

- Community detection
- Identify clusters of nodes or “communities” with high density of edges within and low density of edges across communities



Fortunato 2009

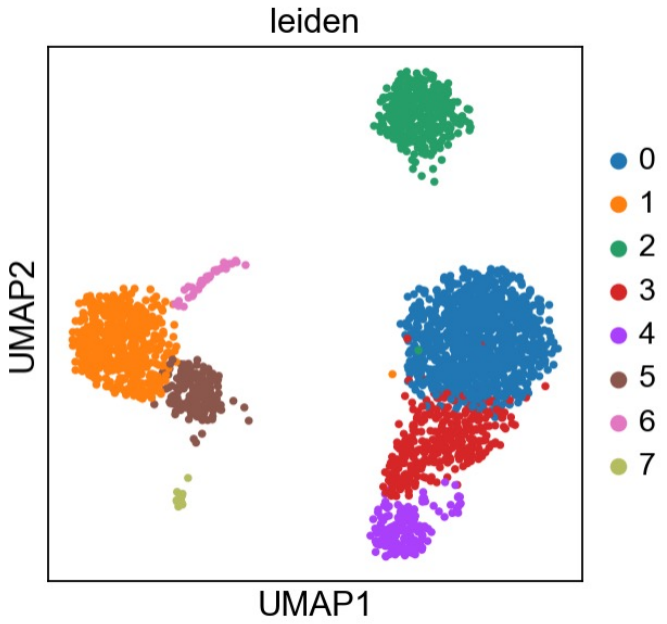
Leiden clustering in single-cell data

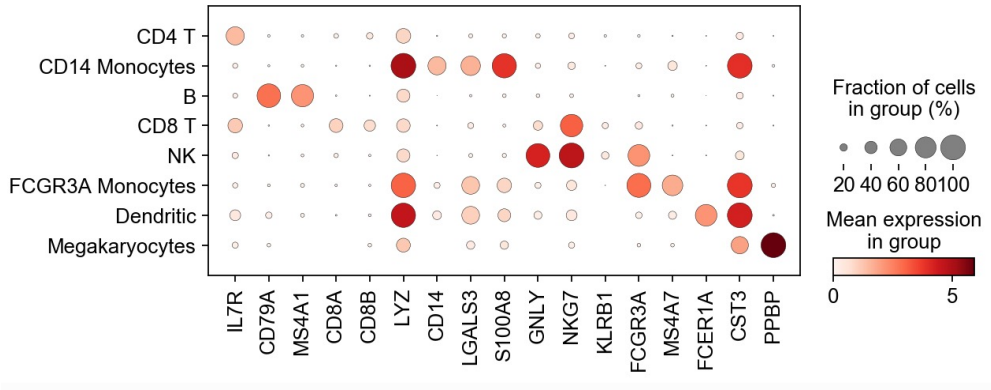


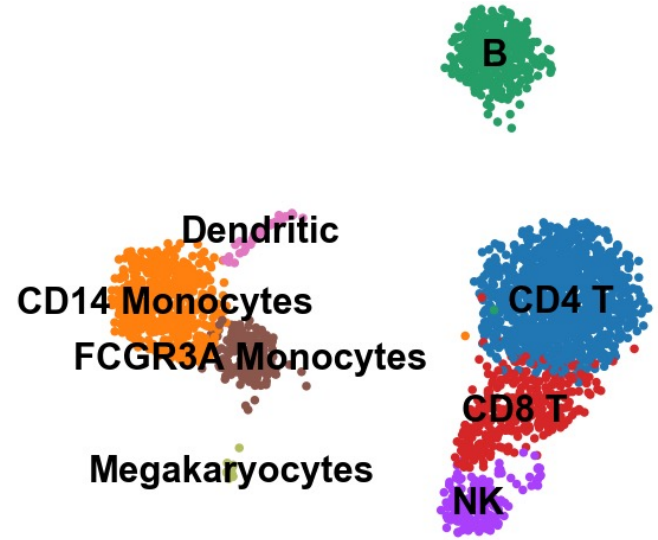
Butler et. al. 2018

Cell type identification

- Marker based identification

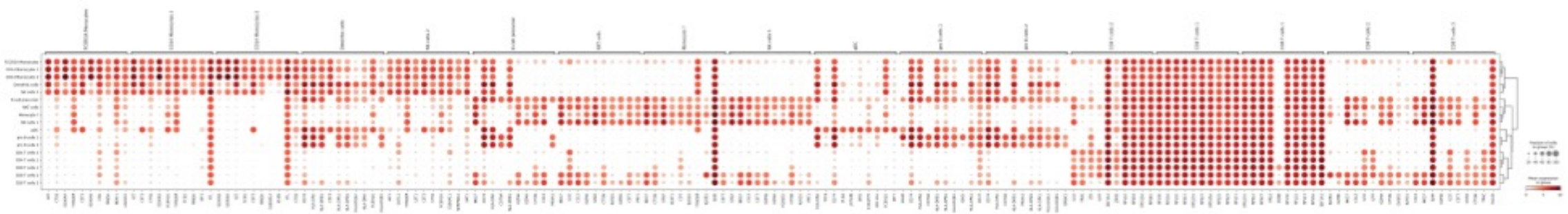






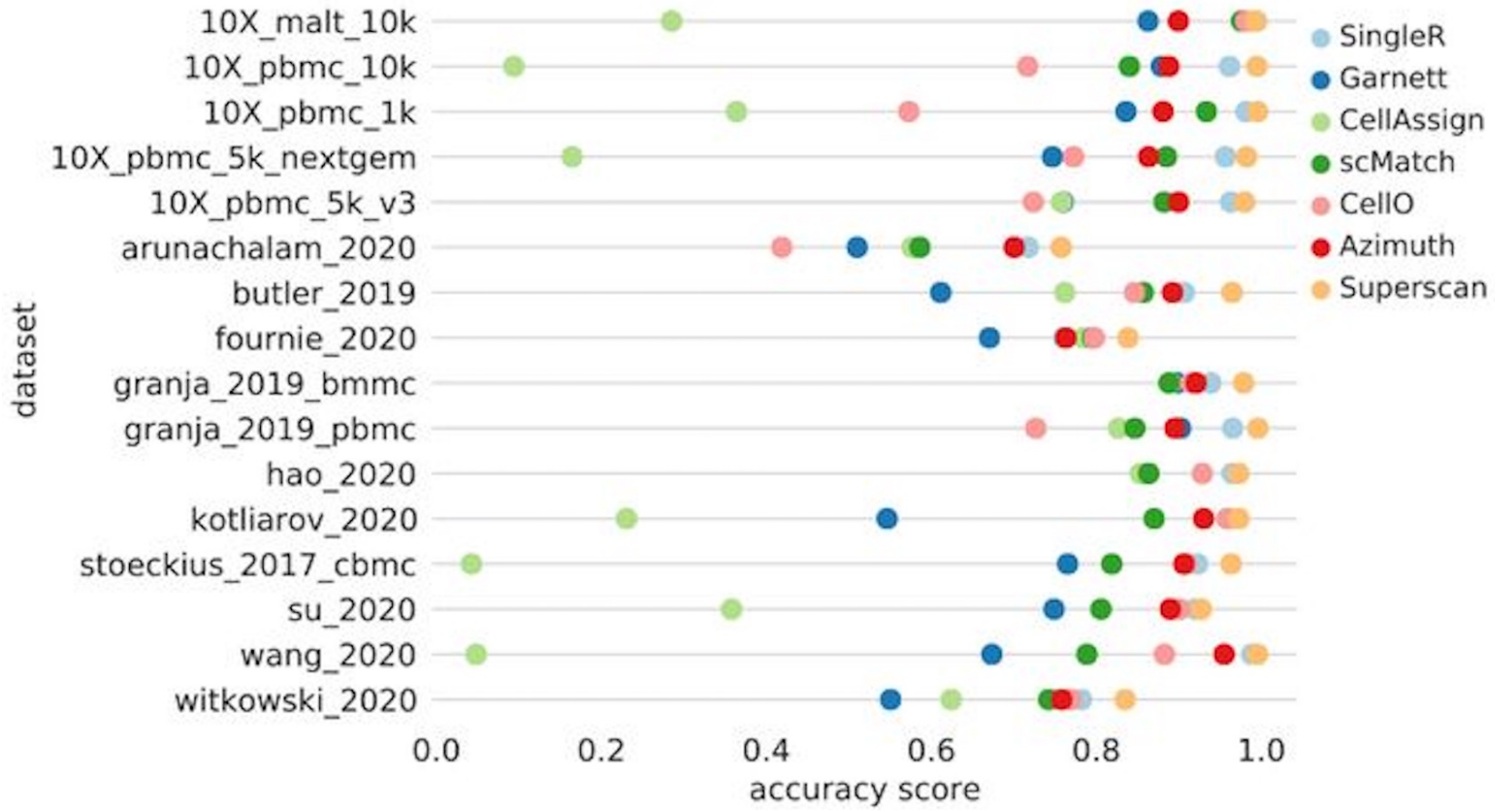
Cell type identification

- Differential expression



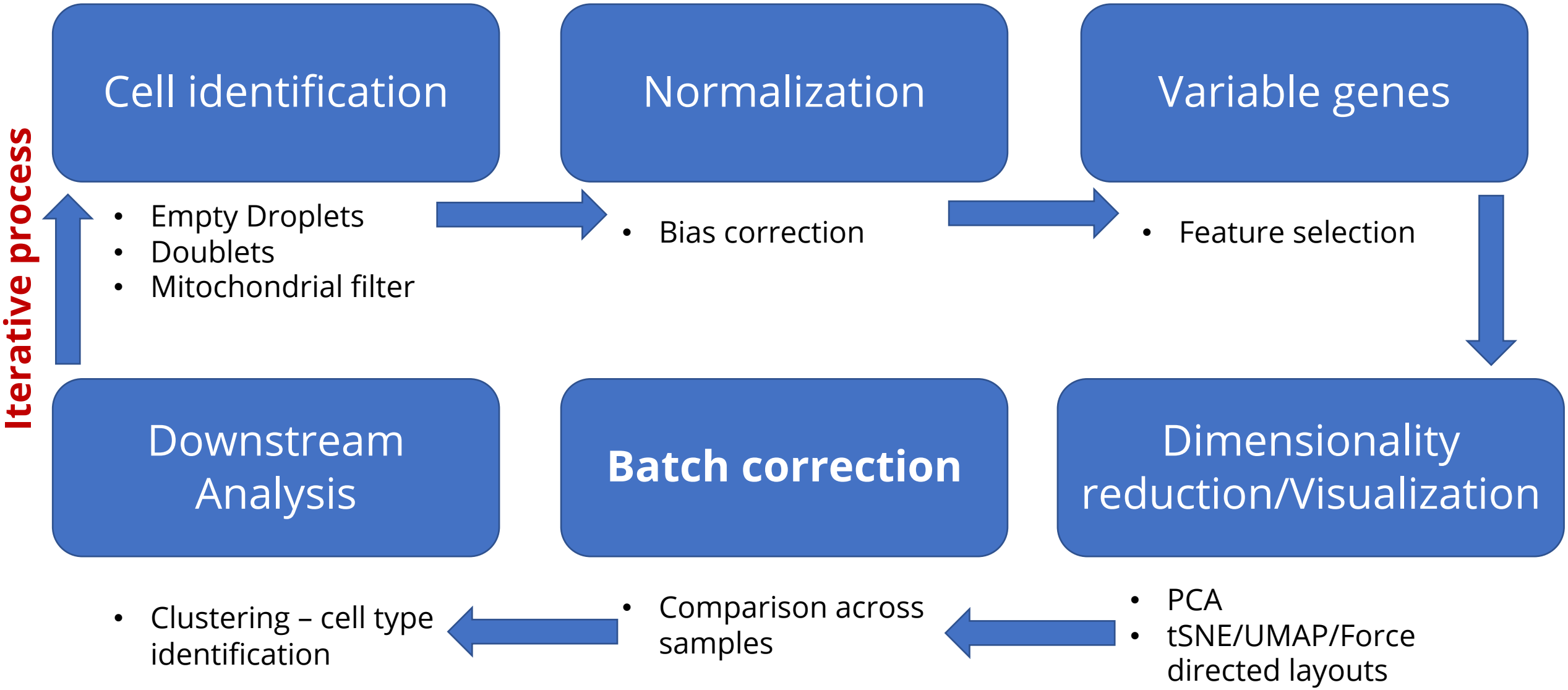
Cell type identification

- Supervised approaches: Train on manually labeled cells
- Superscan



Shasha et. al. 2021

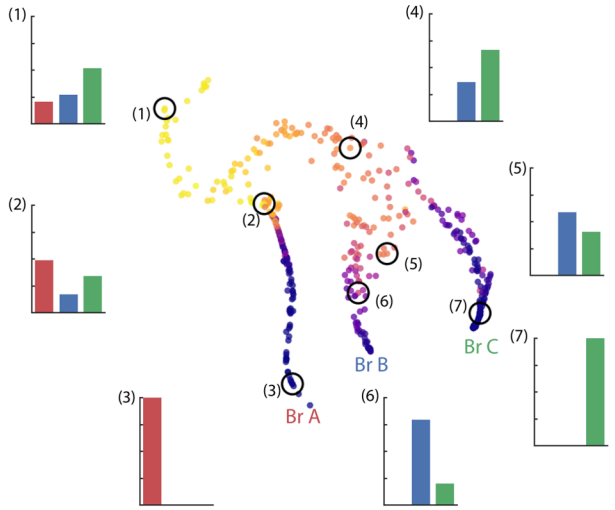
scRNA-seq analysis steps



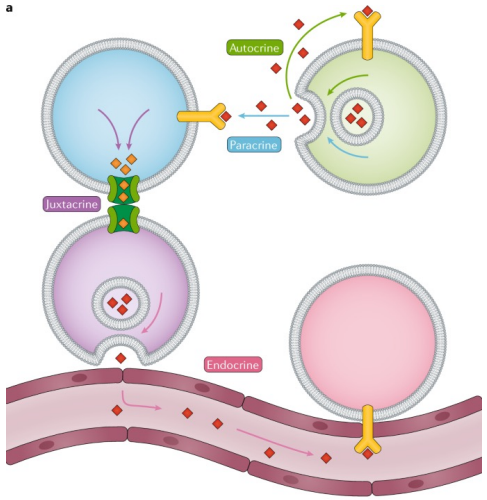
Interpreting single-cell data

Interpreting single-cell data

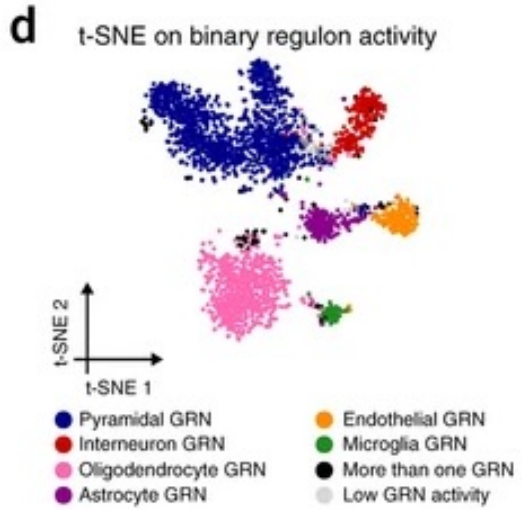
Trajectory analysis



Cell communication



Regulatory Networks

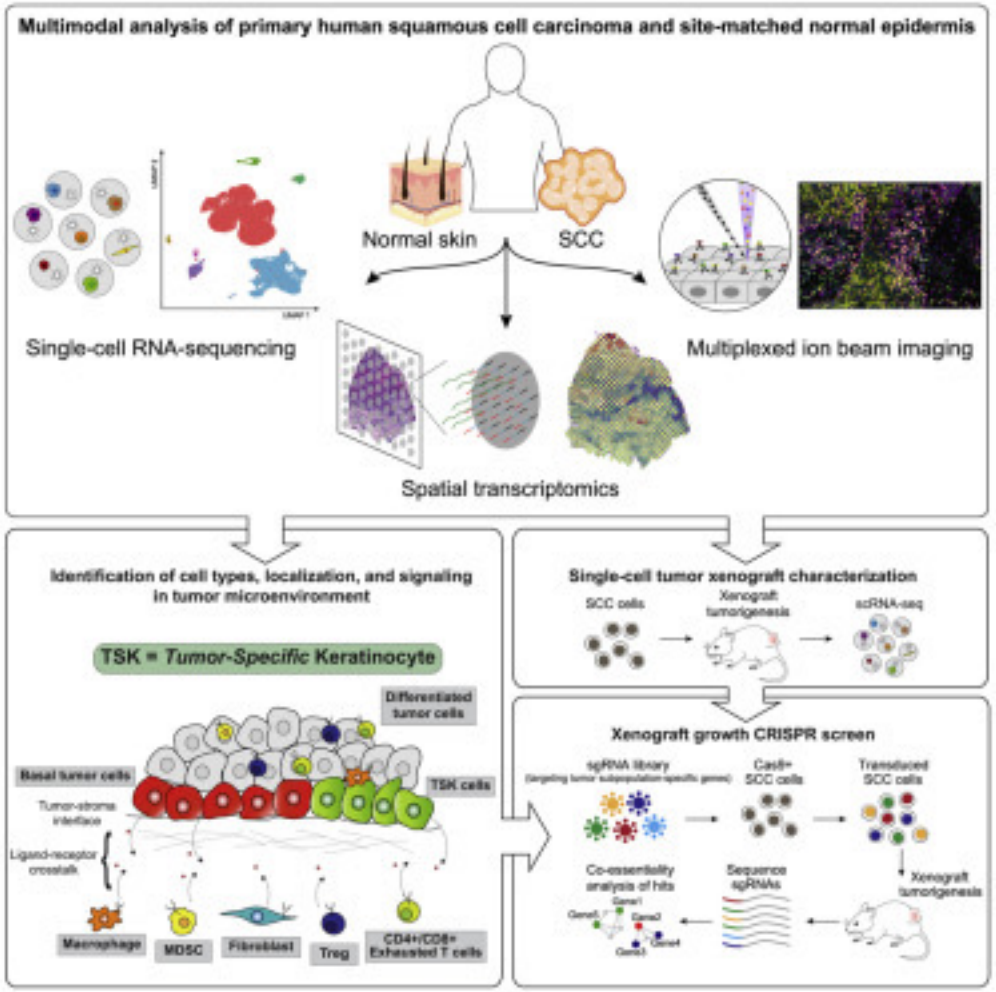


Where to find data?

Major data resources – Healthy/Normal

- Human Cell Atlas
  - Fetal Cell Atlas
- Tabular Muris

Major Data resources - GEO



Supplementary file	Size	Download	File type/resource
GSE144236_CAL27_counts.txt.gz	20.9 Mb	(ftp)(http)	TXT
GSE144236_CAL27_vitro_counts.txt.gz	20.0 Mb	(ftp)(http)	TXT
GSE144236_SCC13_counts.txt.gz	11.2 Mb	(ftp)(http)	TXT
GSE144236_XG_TME_counts.txt.gz	3.4 Mb	(ftp)(http)	TXT
GSE144236_cSCC_counts.txt.gz	127.2 Mb	(ftp)(http)	TXT
GSE144236_patient_metadata_new.txt.gz	648.8 Kb	(ftp)(http)	TXT

SRA Run Selector

Processed data are available on Series record

Raw data are available in SRA

Count matrices (post QC) and metadata are typically made available

Interactive browsers

Mouse endoderm atlas

endoderm-explorer.com

A large number of studies set up webapps for interacting with the data

Lecture 19

- Hands-on fun with single-cell RNA-seq data!