Single cell transcriptomics (1):
technologies, preprocessing and cell annotation

BBMS 3009: Genome Science (First Semester, 2021)

Dr. Yuanhua Huang / 黃淵華
School of Biomedical Sciences &
Department of Statistics and Actuarial Science

香 港 大 學
THE UNIVERSITY OF HONG KONG

Today's learning objectives

1. Understand the major single-cell technology
2. Pros and Cons of each technology
3. Main steps of computational preprocessing
4. Clustering and cell type annotation

Resources:

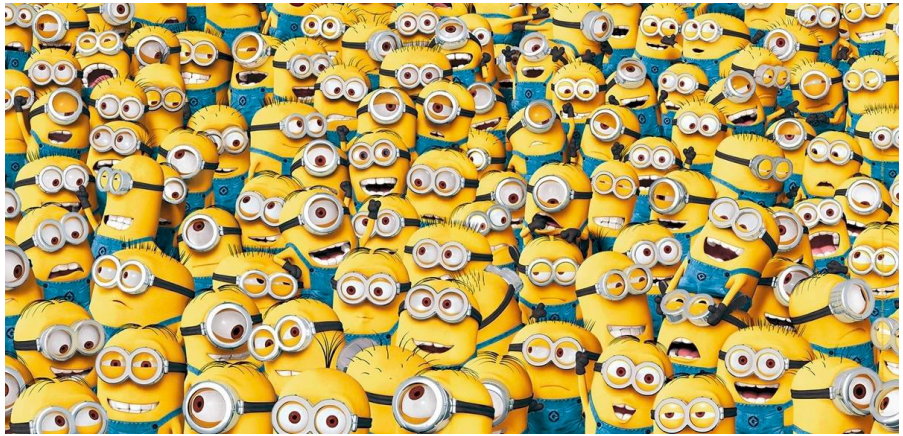
- Analysis of single cell RNA-seq data (Sanger course) https://www.singlecellcourse.org/
- HKUMed Single-cell analysis tutorial workshop (HKU Med) https://statbiomed.github.io/HKU-single-cell-workshop/

What is a minion?

Key features

- Yellow
- Round head
- Big eyes
- Earless
- Wear goggles
- Neckless
- Short legs
- Long arms
- …

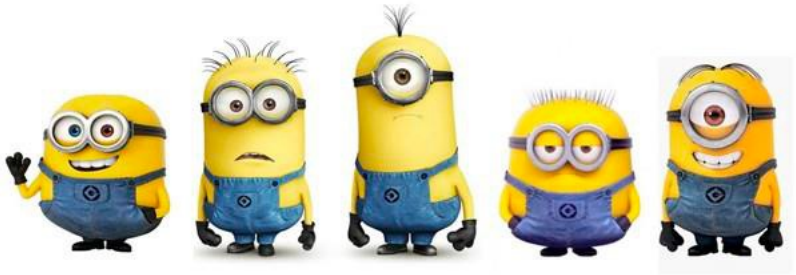
What is a minion?



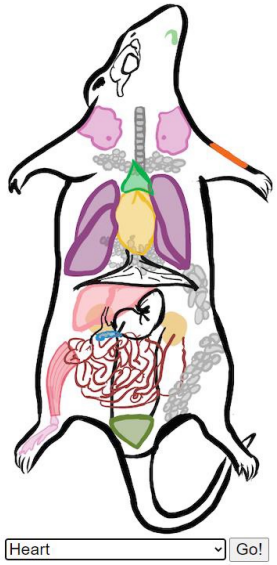
Actually, there are many minions, and they can be classified into several basic character forms based on some key distinguishing features.

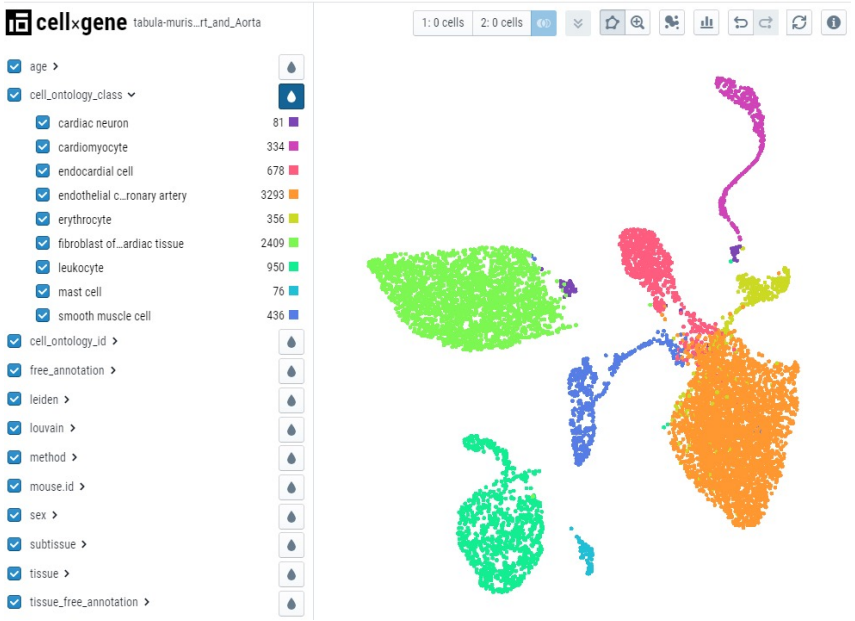
Key distinguishing features

- One vs two eyes
- Height
- Hair style



What are the cells in mouse heart





Similarly, most organs consist of many cells that can be classified into a handful of cell types based on some differentially expressed genes.

Tabula Muris Senis: https://tabula-muris-senis.ds.czbiohub.org/

Single cell vs bulk RNA-seq

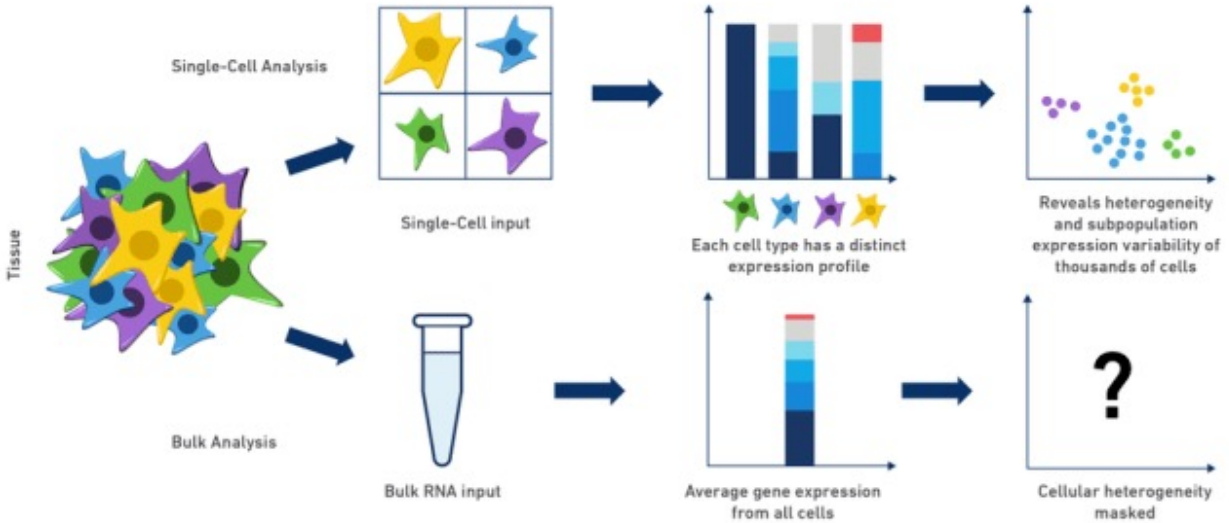
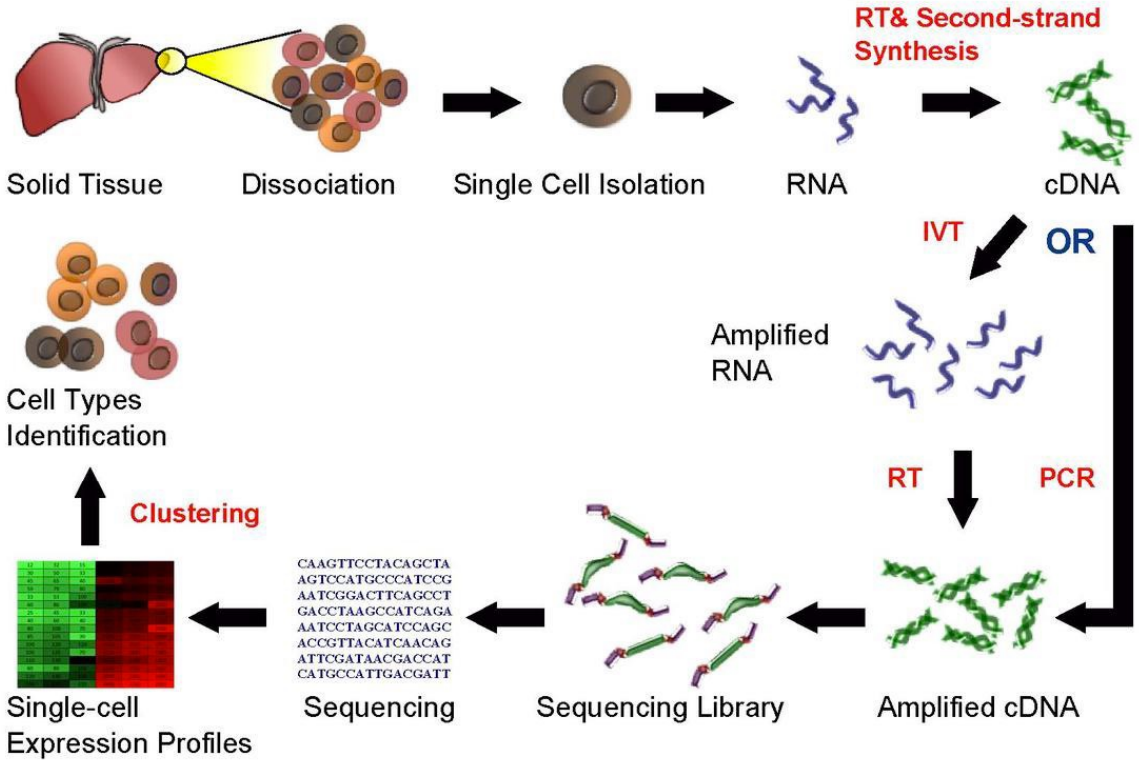


Figure adapted from 10X Genomics

Single-cell RNA sequencing workflow

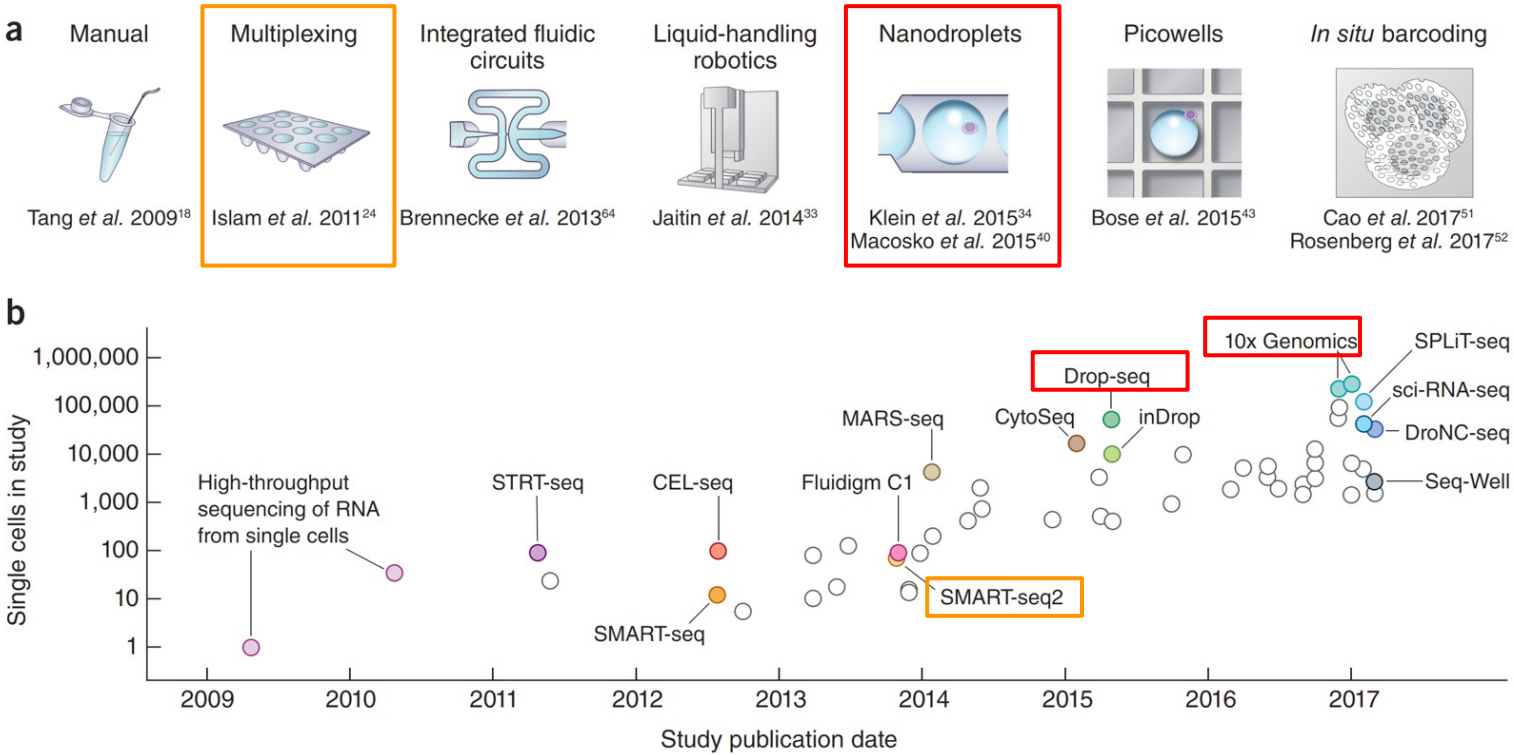


Major unique steps (challenges) in scRNA-seq comparing to bulk RNA-seq

- Cell isolation
- Molecule capture efficiency
- cDNA amplification bias

https://en.wikipedia.org/wiki/Single_cell_sequencing

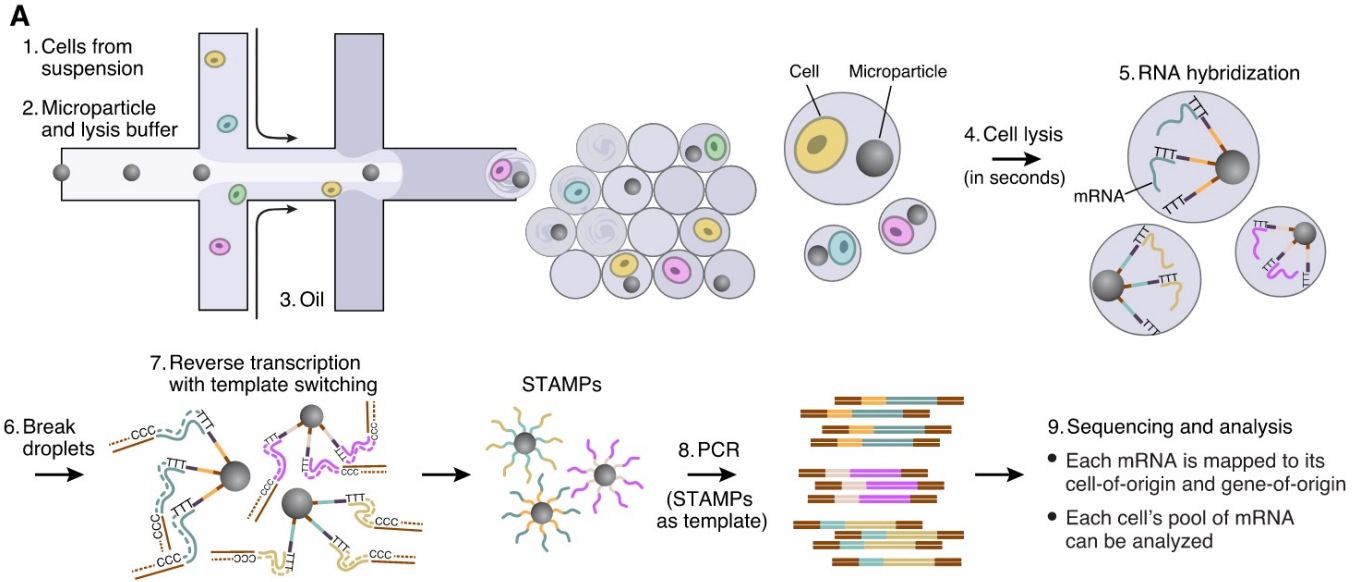
Revolution of scRNA-seq technology



Svensson et al. *Nature Protocols* (2018)

With fluorescence-activated cell sorting (FACS)

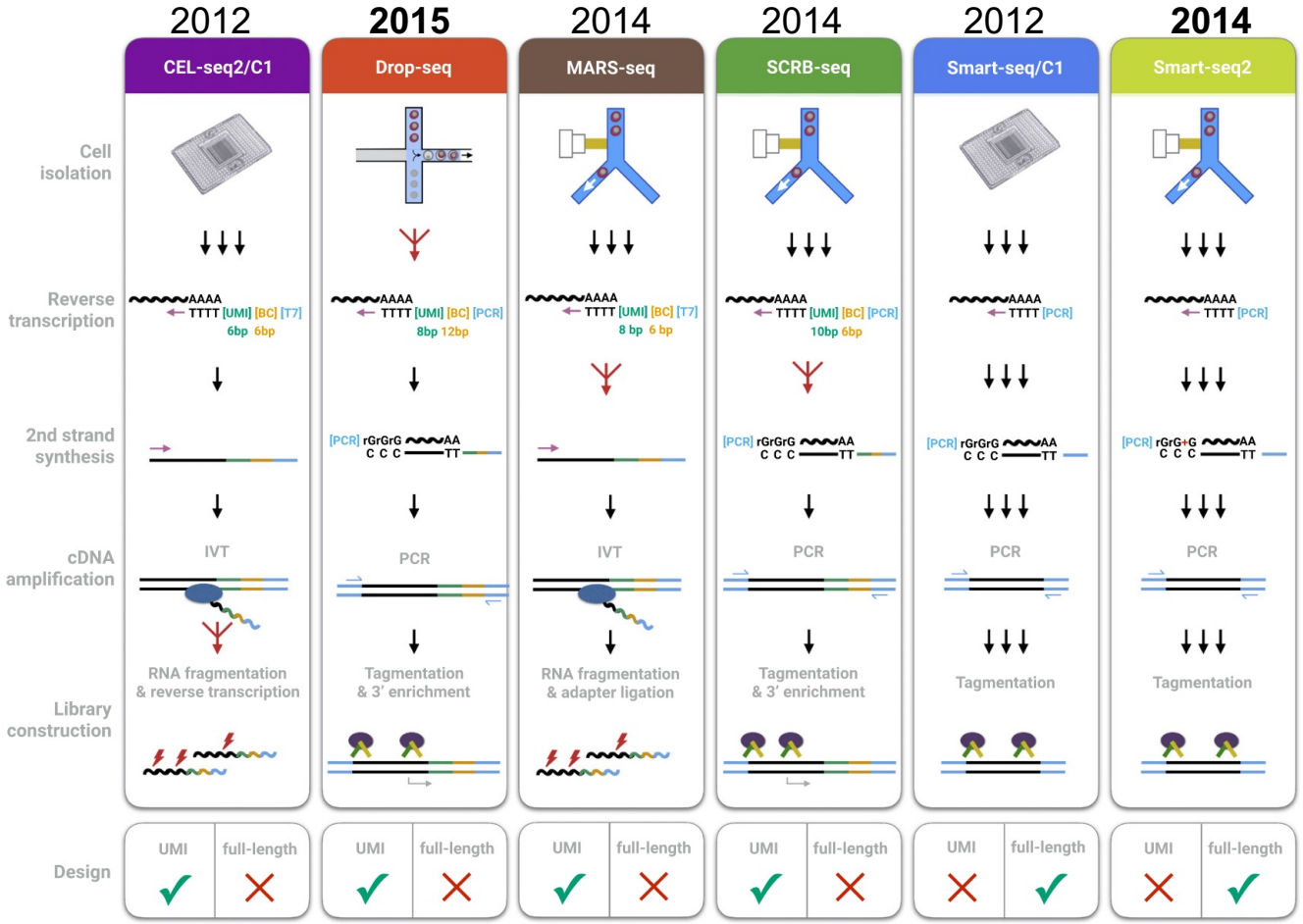
Droplet-based technology; now dominating



10X Genomics is a commercial provider of droplet based scRNA-seq platform

Drop-seq; Macosko et al. *Cell* 2015

Comparison of scRNA-seq technologies



Key difference

- Unique molecular identifier (UMI); always contradict with full length
- Cell isolation: affecting the throughput and requirement of input cell numbers

Ziegenhain et al., Mol Cell, 2017
Kivioja et al. Nat Methods, 2011

Comparison of scRNA-seq technologies

- With UMI (vs non-UMI, i.e., full length)
  - Remove the bias from PCR amplifications;
  - Only one end of the molecule (3’ or 5’);
  - Mainly gene level: missing most alternative splicing and genotype info;
  - Much lower number of sequencing reads needed.

- Well-based vs droplet-based
  - Well-based: small number of cells, e.g., embryos
  - Droplet-based: large number of cells (both out and input requirement)

Open challenges in single-cell sequencing

- How to isolate cells, e.g., from a solid tissue?
  - Factor to consider: throughput, sensitivity, doublet rate (10% for 10K cells droplet-based platform; the more cells the higher doublet rates)

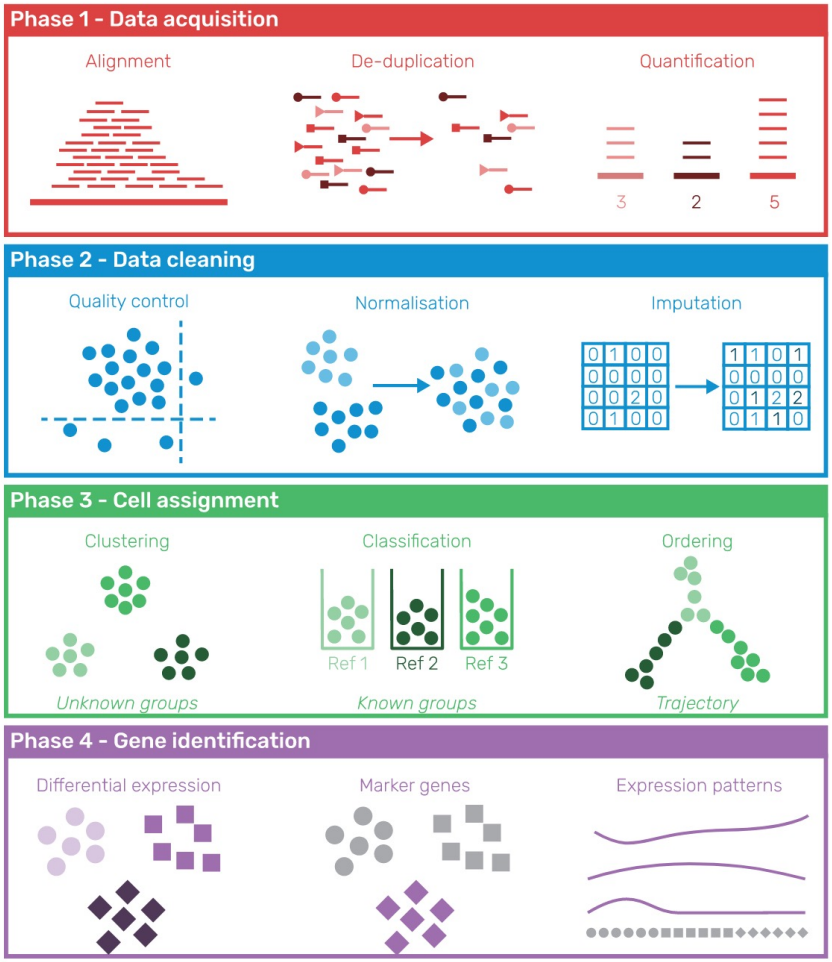
- Unsatisfied capture efficiency: the RNA contents are very low in an individual cell (some gene may only have one or two RNA molecules)

- How to balance the number of cells and coverage per cell?
  - Budget limit: 300 million reads → 2,500 USD
  - In bulk RNA-seq: ~60 Million reads / sample; 5 samples
  - In scRNA-seq (full length based): 1 million reads / cell * 300 cells
  - In scRNA-seq (UMI-based): 50,000 reads / cell * 6000 cells

Today’s learning objectives

1. Understand the major single-cell technology
2. Pros and Cons of each technology
3. Main steps of computational preprocessing
4. Clustering and cell type annotation

Major steps of computational analysis



Data acquisition

- STAR-solo;
- Kallisto bustools;
- CellRanger for 10X Genomics data

Core algorithm is the same as bulk RNA-seq

Downstream analysis platforms

- Seurat (R package);
- Scanpy (Python package);
- Other tools for specific task(s): follow their online tutorial

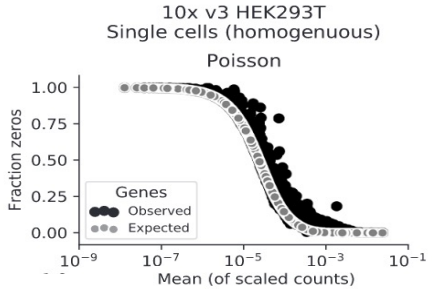
Zappia et al. Plos Comp Bio, 2018

Preprocessing of count matrix (1)

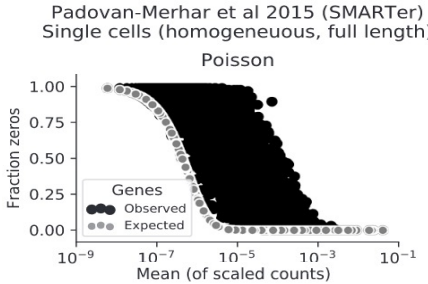
- Reads count matrix (SMART-seq2) or UMI count matrix
- Extremely high proportion of zero values

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
ISG15	0	0	1	9	0	1	0	0	0	3	0	0	1	5	0	2
AURKAIP1	0	1	0	1	0	0	0	1	0	0	2	0	2	1	0	0
MRPL20	1	0	1	0	0	0	0	1	0	0	2	0	0	1	0	0
SSU72	0	1	0	3	0	0	0	0	0	0	0	0	0	1	0	0
C1orf86	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
RER1	0	1	1	1	0	0	0	0	0	0	0	1	0	1	0	0
TNFRSF14	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
RPL22	1	3	1	2	0	0	0	1	0	0	3	2	2	1	1	2
PARK7	0	0	2	0	0	0	0	0	0	1	1	0	2	0	0	0
ENO1	0	2	2	2	0	0	1	0	0	0	1	0	0	1	0	0
AGTRAP	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
EFHD2	1	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0
NECAP2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SDHB	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CAPZB	1	1	3	1	2	0	1	2	0	1	1	0	0	1	0	0
MINOS1	0	0	1	0	0	0	0	0	0	0	1	1	2	0	0	0
CDA	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0	0
DDOST	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
CDC42	0	0	1	2	1	1	1	0	0	0	0	0	0	1	1	0
HNRNPR	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
RPL11	41	39	24	19	3	14	17	20	4	6	43	33	14	16	7	10
SRRM1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
SYF2	0	0	0	0	1	1	0	0	0	0	3	1	0	1	0	1

10x Genomics data (PBMC)



UMI-based:
Lower expression → more zeros

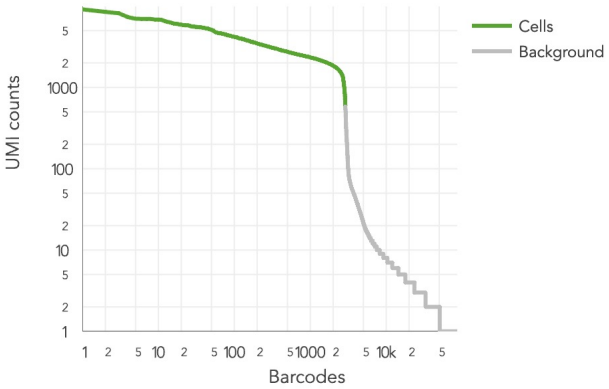


Non-UMI-based:
Much higher proportion of zeros (zero inflation); so-called "dropout", a bit misleading.

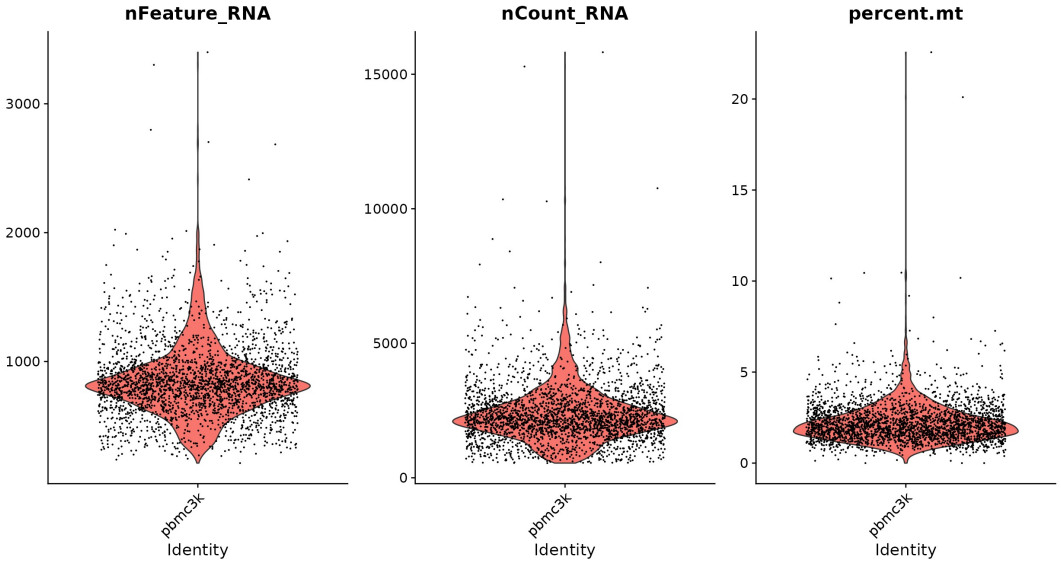
Svensson. *Nature BioTech, 2020*

Preprocessing of count matrix (2)

- Quality control: filter low quality cells
- Cellranger has a build-in function to select valid cell barcodes vs empty drops (there are also other methods proposed)
- A further filtering of cells based on n_genes, n_UMIs, or percent of mitochondrial reads

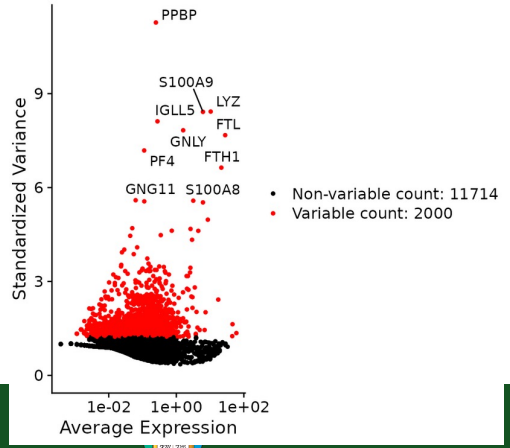


https://satijalab.org/seurat/articles/pbmc3k_tutorial.html



Preprocessing of count matrix (3)

- Normalization (related to bulk RNA-seq in *Transcriptomics II*):
  - UMI-based: Count per million; log(CPM + 1)
  - Non-UMI based: RPKM or TPM; log(RPKM + 1) or log(TPM + 1)
- Select informative genes:
  - From the Gencode annotation, there are >30K genes; a substantial proportion of them have no coverage in almost all cells
  - We only want to keep the highly variable genes in the cell population
- Potential further preprocessing:
  - Standardize all genes (rescale to same variance)
  - Regress meta covariate for each gene

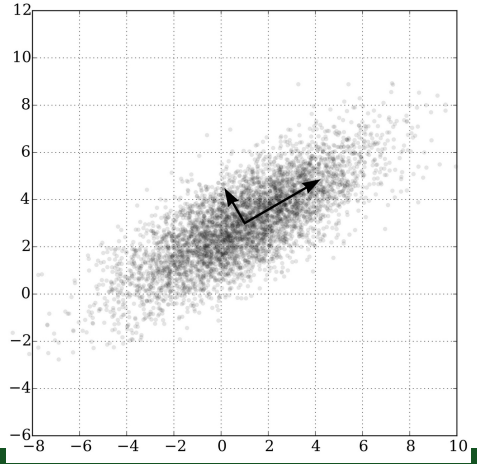


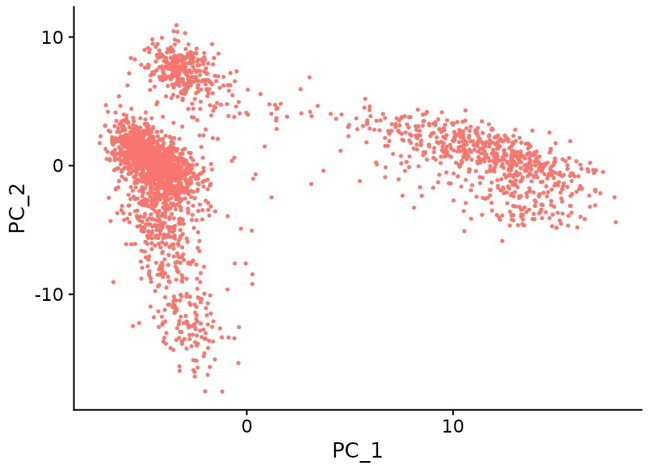
Today’s learning objectives

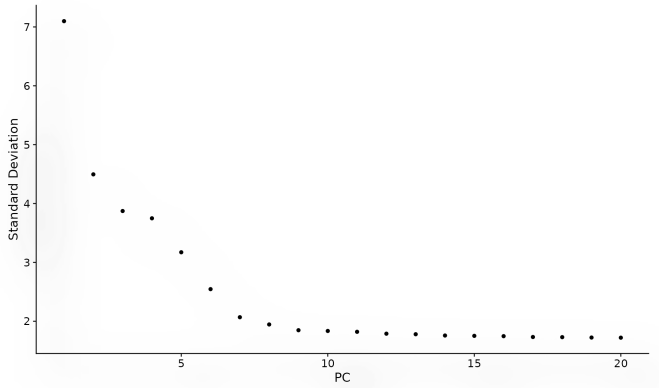
1. Understand the major single-cell technology
2. Pros and Cons of each technology
3. Main steps of computational preprocessing
4. Clustering and cell type annotation

Dimension reduction

- The high dimensional data (e.g., 3K cells x 1K HV_genes) is sparse, and difficult to discover the underlying patterns
- Dimension reduction provide dense representation of each cell, often less noisy and more computational efficient (e.g., 3K cells x 50 PCs)
- Principal component analysis (PCA); invented in 1901 by Karl Pearson
  - A linear transformation to orthogonal and most variable new dimensions

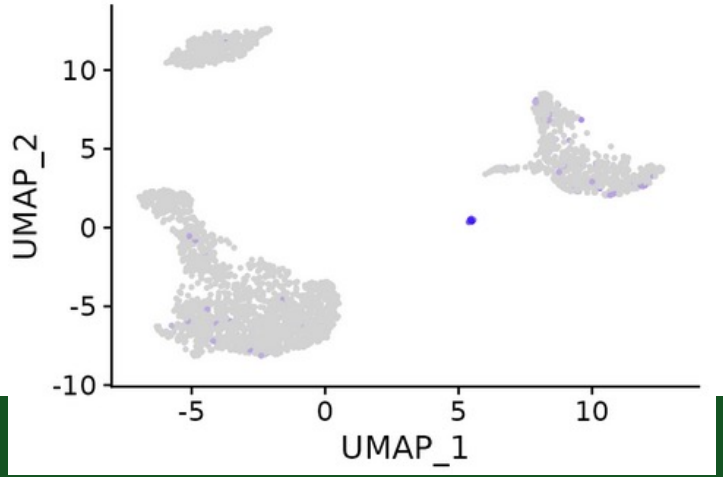






Visualization

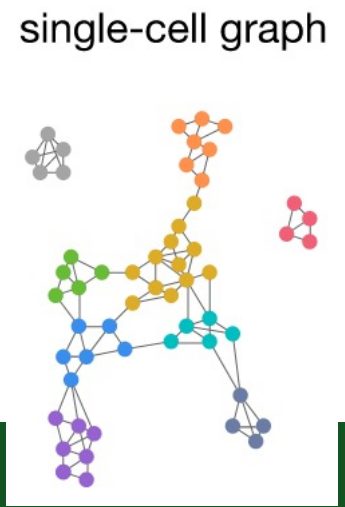
- An extreme dimension reduction to 2D or 3D
- PCA: Global structure and linear transformation
- t-SNE or UMAP: non-linear transformation but better retaining local structure, e.g., separating cell types
  - This type of methods are very popular recently
  - Can be easily mis-interpret, especially distance between dots or groups; they don't have physical meaning

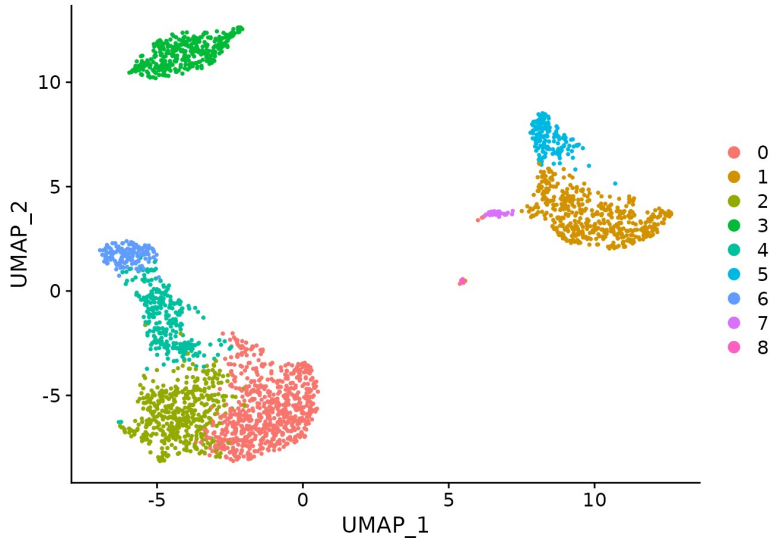


Clustering of cells

- One of the most interesting patterns of the scRNA-seq is the sub-population structure. In machine learning, it calls clustering.
- Common methods: K-means, and Graph-based method (K-nearest neighbour graph), hierarchical clustering
- Generally, it is not ideal to perform clustering with only 2D data (e.g., t-SNE), better to use more dimensions (e.g., 50 PCs)

KNN graph; Partition at a certain resolution, e.g., with Louvain algorithm





Wolf et al, Gen Biol, 2019

Annotation of clusters

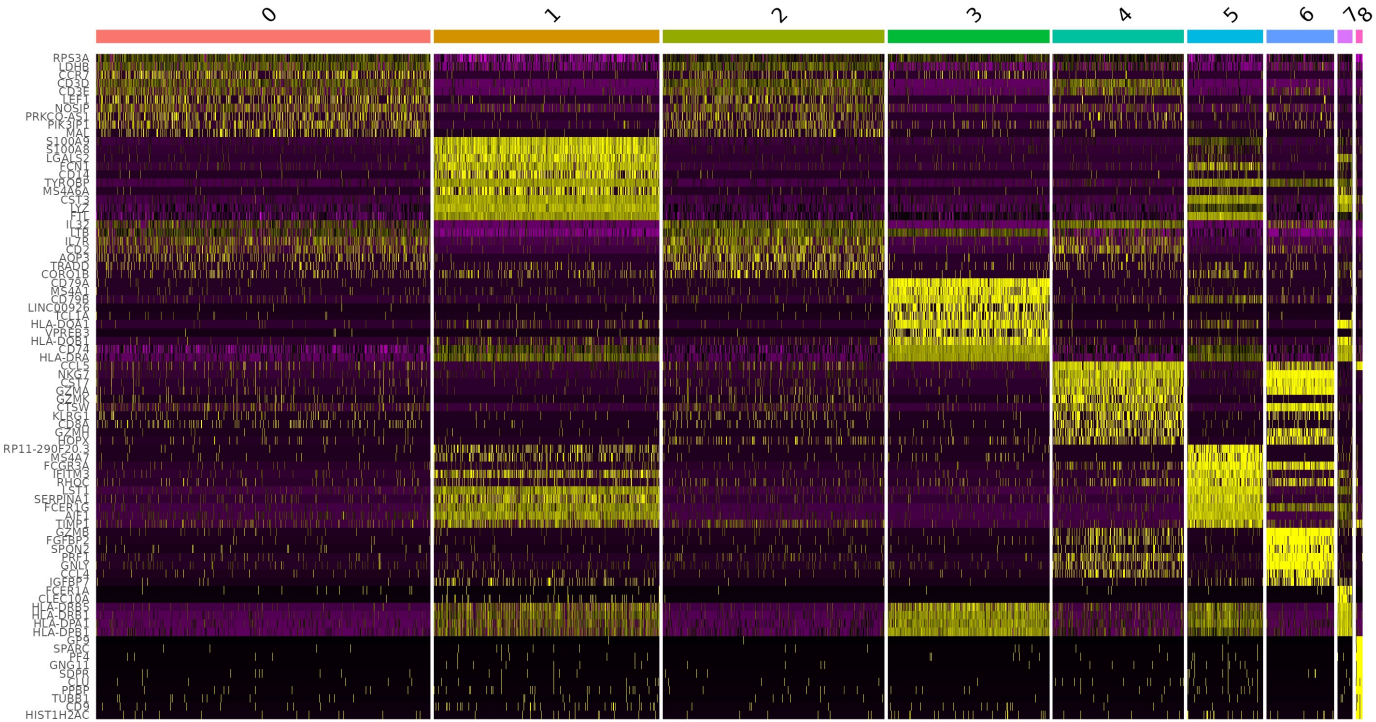
- Finding differentially expressed features (cluster biomarkers)
  - GLM regression-based hypothesis testing (we learned in bulk RNA-seq)
  - More computational efficient methods (we have many more replicates now)
  - Wilcoxon rank-sum (Mann-Whitney-U), MAST, or even t-test
- DE genes: one cell type vs the rest

```
# find all markers of cluster 2
cluster2.markers <- FindMarkers(pbmc, ident.1 = 2, min.pct = 0.25)
head(cluster2.markers, n = 5)
```

```
##            p_val avg_log2FC pct.1 pct.2    p_val_adj
## IL32 2.593535e-91  1.2154360 0.949 0.466 3.556774e-87
## LTB  7.994465e-87  1.2828597 0.981 0.644 1.096361e-82
## CD3D 3.922451e-70  0.9359210 0.922 0.433 5.379250e-66
## IL7R 1.130870e-66  1.1776027 0.748 0.327 1.550876e-62
## LDHB 4.082189e-65  0.8837324 0.953 0.614 5.598314e-61
```

Annotation of clusters

- Using the identified DE genes to interpret the clusters



Cluster ID	Markers	Cell Type
0	IL7R, CCR7	Naive CD4+ T
1	CD14, LYZ	CD14+ Mono
2	IL7R, S100A4	Memory CD4+
3	MS4A1	B
4	CD8A	CD8+ T
5	FCGR3A, MS4A7	FCGR3A+ Mono
6	GNLY, NKG7	NK
7	FCER1A, CST3	DC
8	PPBP	Platelet

Annotation of clusters: (semi-)automated

- The above annotation procedure requires a strong domain knowledge
- Computational methods: further using maker gene database to automatically annotate the clusters into known cell types
- Given large human and mouse cell atlas available, and good annotation have been generated in many good datasets, we can use them as training data, then directly predict our query data with machine learning algorithms, like predict dogs and cats from images.
  - Azimuth is one of them: https://azimuth.hubmapconsortium.org
  - More tools are under fast development

Any questions?

1. Understand the major single-cell technology
2. Pros and Cons of each technology
3. Main steps of computational preprocessing
4. Clustering and cell type annotation

Resources:

- Analysis of single cell RNA-seq data (Sanger course) https://www.singlecellcourse.org/
- HKUMed Single-cell analysis tutorial workshop (HKU Med) https://statbiomed.github.io/HKU-single-cell-workshop/