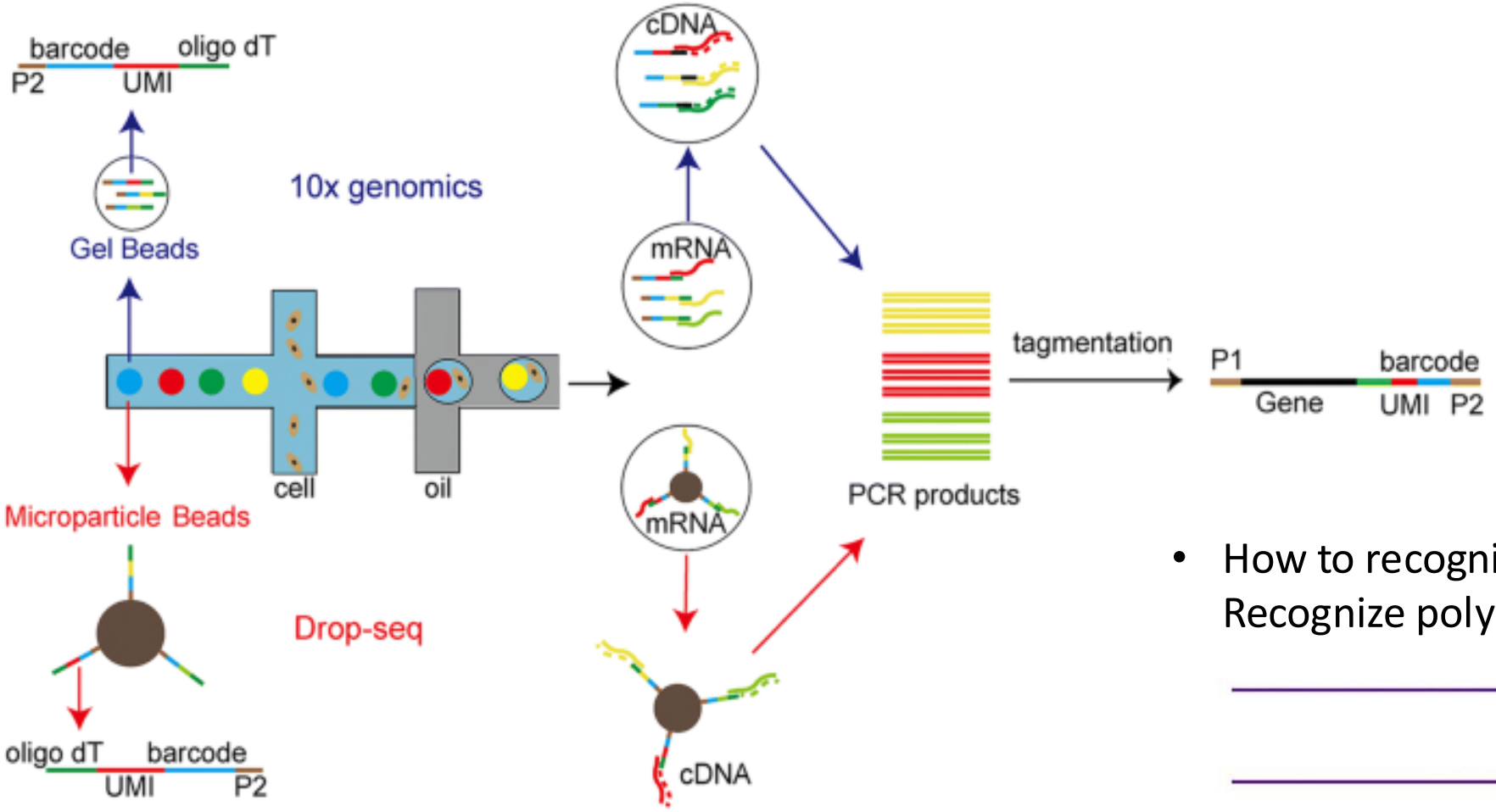
Lecture 2
scRNA-seq technique and count matrix QC

Outline

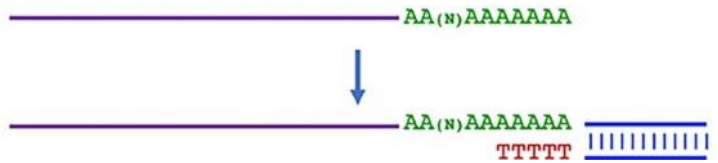
- Measurement error in scRNA-seq experiments

- Quality control of count matrix
  - Doublet removal
  - Ambient RNA correction
  - Remove low-quality cells

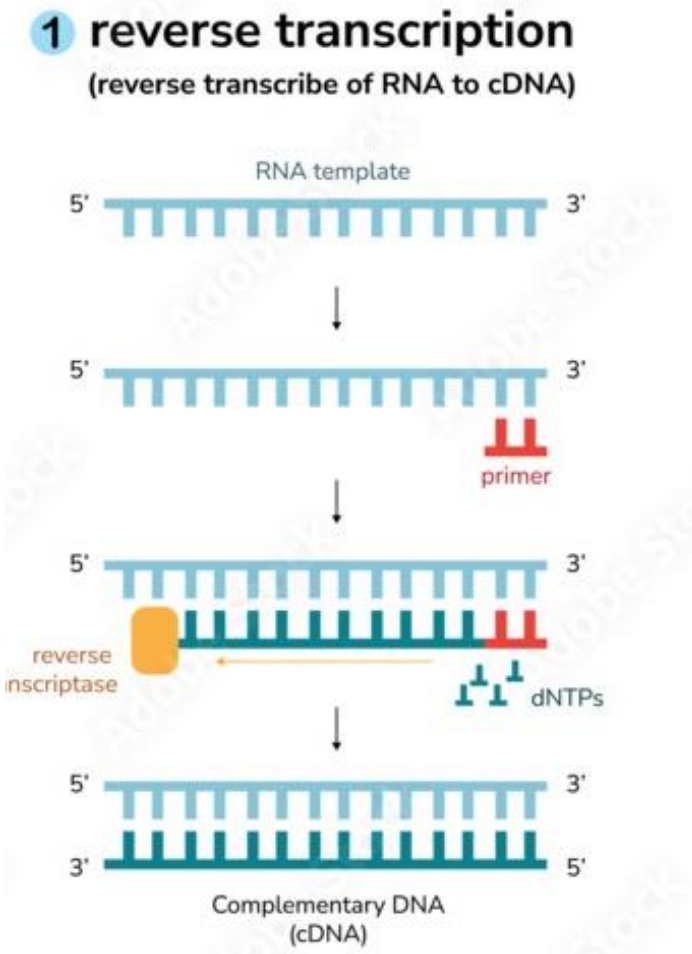
RNA sequencing: reverse transcription, amplification and sequencing

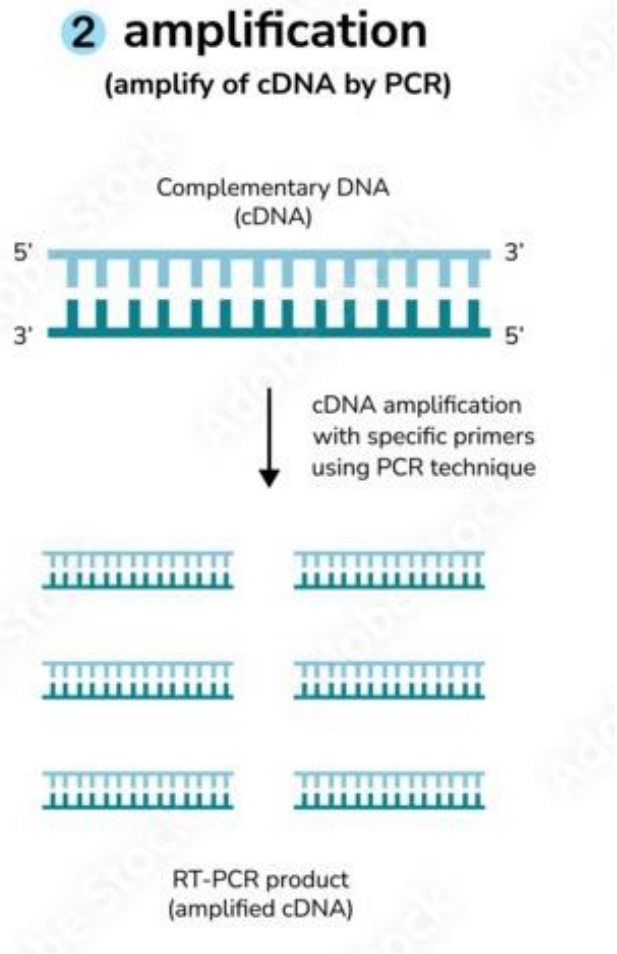


- How to recognize an mRNA fragment? Recognize poly-A tails

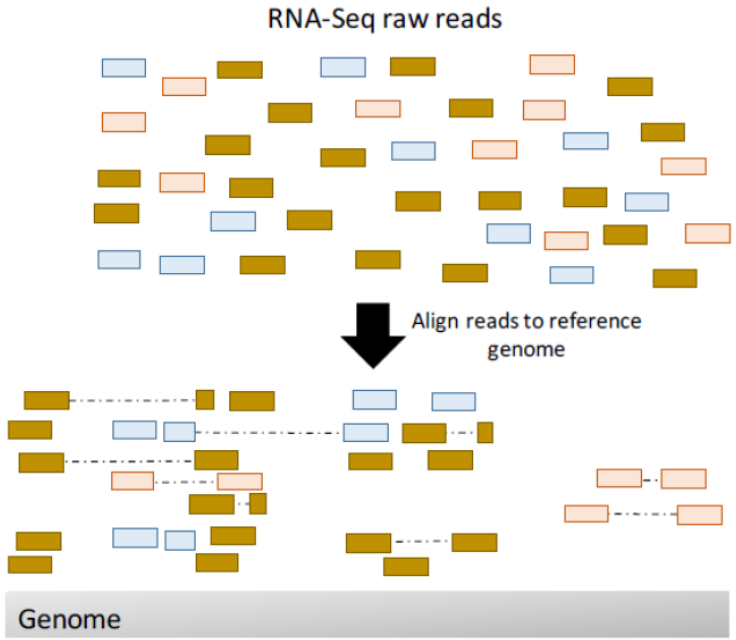


RNA sequencing: reverse transcription, amplification and sequencing





How to know the corresponding gene of the RNA fragment?
Map it back to the genome



Make mRNA fragments more stable

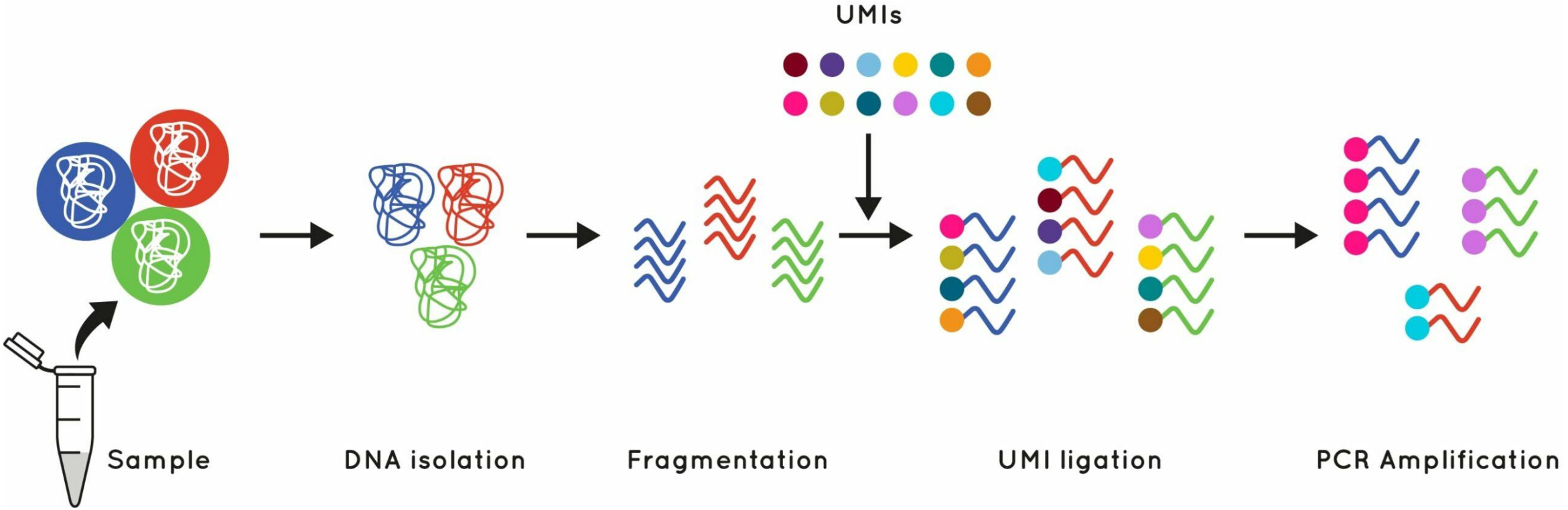
Increase the number of materials to sequence

Cell barcode for demultiplexing

Fig. 2. Schematic of a fragment from a final Chromium™ Single Cell 3' v2 library. *Can be adjusted.

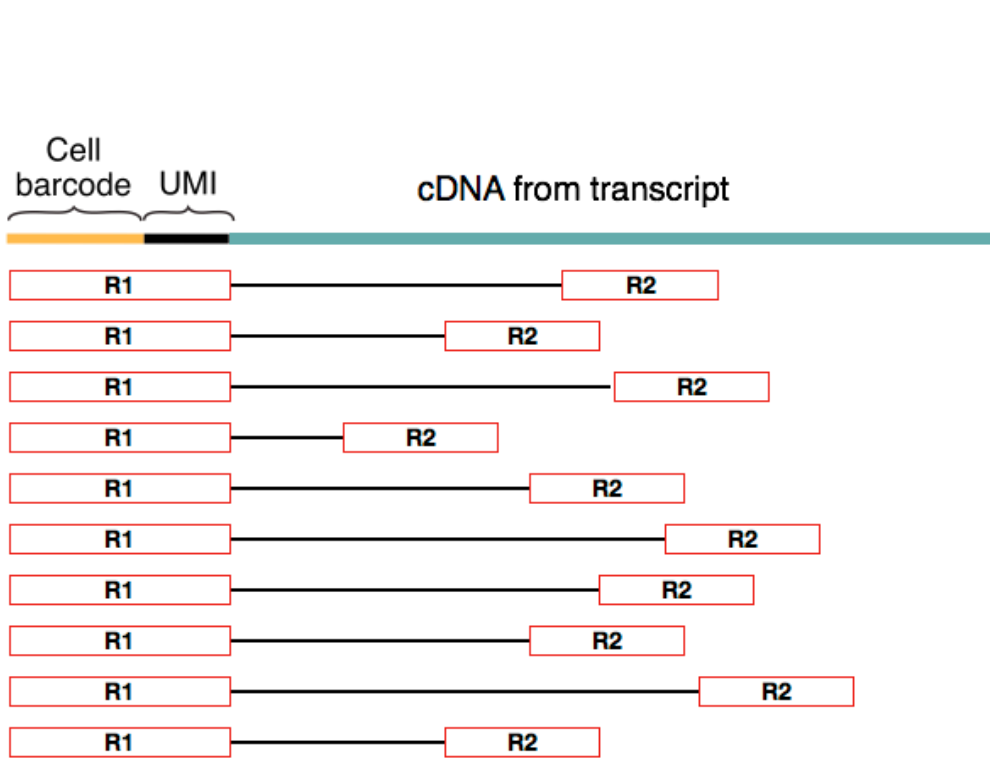
- cell barcode: associate cDNAs to a specific cell
- UMIs: label specific cDNA molecules to avoid amplification bias

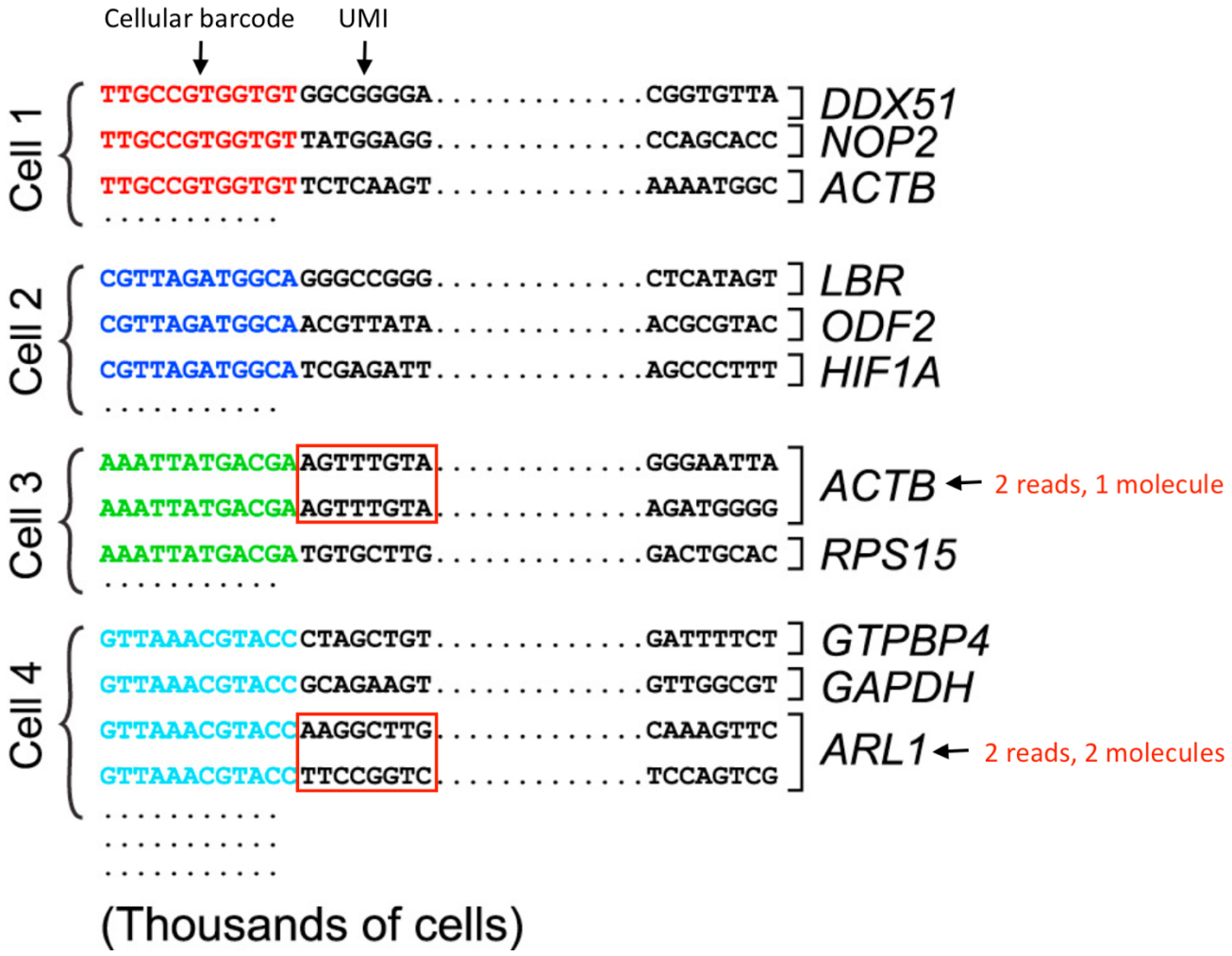
Unique molecular identifier (UMI)



- each initial input cDNA fragment has its own unique tag

Obtain count matrix from reads





https://data-science-sequencing.github.io/Win2018/lectures/lecture16/

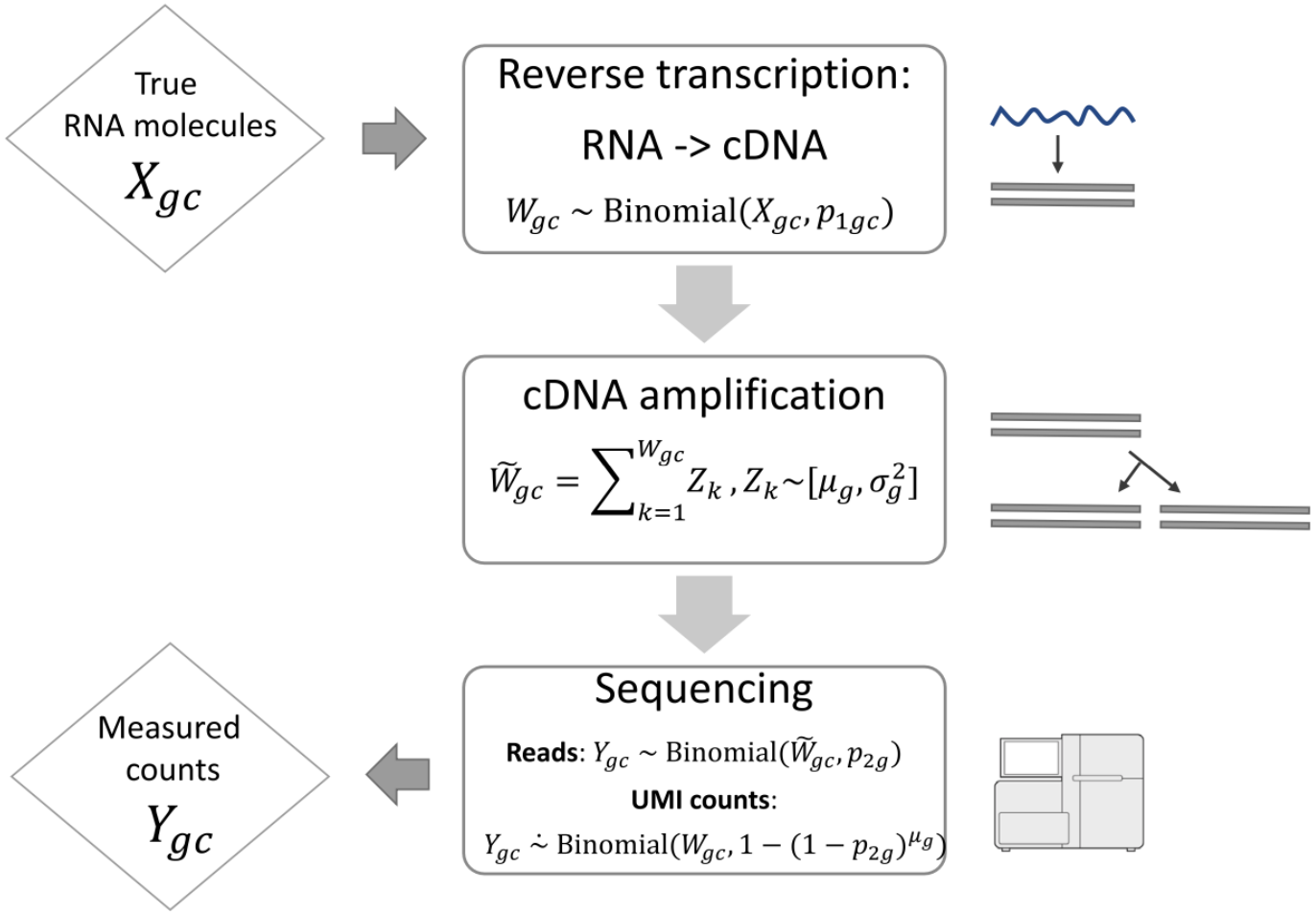
Gene expression count matrix

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

- Understand the cell population
- Characterize each cell
- Understand how gene expressions change across cells and gene-gene relationships

- Next class: QC to improve the quality of this matrix, understand noise and signals in the matrix

Propagation of measurement error



- A cell c, a gene g
- For UMI counts, roughly $Y_{gc} \sim \text{Binomial}(X_{gc}, \alpha_{gc})$
- For non-UMI reads:
  - $Y_{gc} = 0$ if $W_{gc} = 0$
  - Y_{gc} can be large if W_{gc} due to amplification
- Most of scRNA-seq data nowadays use UMI

library size

- For UMI counts, roughly

$$Y_{gc} \sim \text{Binomial}(X_{gc}, \alpha_{gc})$$

  where α_{gc} is the cell-gene-specific efficiency
- Assume that $\alpha_{gc} \approx \alpha_c \gamma_g$ where α_c is cell-specific efficiency and γ_g is a gene-specific bias

- Researchers have observed that α_c can vary greatly across cells, but it is typically unidentifiable (will talk more in later slides)

- Library size of a cell: total total sum of UMI counts across all measured genes in a cell

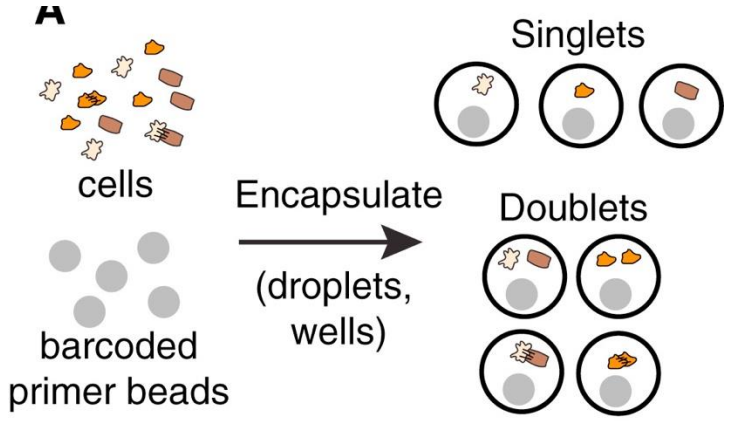
$$l_c = \sum_g Y_{gc}$$

  - Cells with large library size
    - Large cells containing many mRNAs (like neurons), high-quality cells where mRNAs are efficiently captured, doublets
  - Cells with small library size
    - Small cells containing few mRNAs, low-quality cells, empty droplets

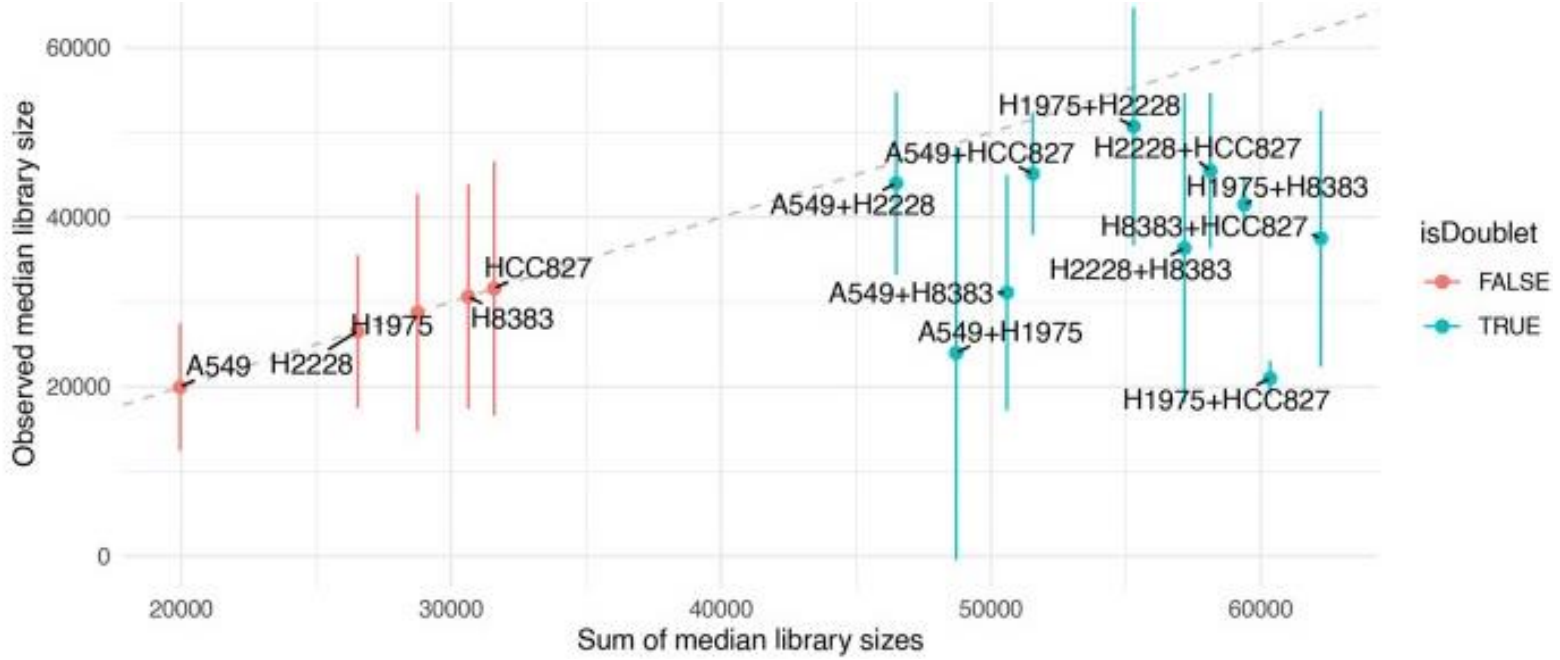
- Library size normalization: Y_{gc} is not comparable across cells, compare relative proportion Y_{gc}/l_c across cells

Doublets

- It is always possible that two (or more) cells share the same barcode
  - Common to have 10% - 20% doublets in scRNA-seq experiments
  - More cells → higher proportion of doublets



- Doublets or multiplets may have relatively large library size, but removing them simply based on library size is not efficient



Germain, Pierre-Luc, et al. "Doublet identification in single-cell sequencing data using scDblFinder." *F1000Research* 10 (2021).

Doublets

- Two major types of doublets
  - Homotypic doublets: formed by cells of the same ”type”
    - Transcriptomic profile looks similar to a singlet
    - Hard to identify but also not that harmful for most data analysis purposes
  - Heterotypic doublets: formed by cells of distinct transcriptional states
    - Possible to identify due to their distinct gene expression profile
- Experimental approaches to identify doublets
  - Very few false positives, but requires special experimental design (not available for most experiments)
  - Example techniques:
    - species mixture: only works for experiments with multiple species
    - demuxlet (Kang et. al. Nature Biotech 2018): use SNP, works for experiments involving multiple individuals
- Computational approaches: identify doublets solely based on count matrix

Scublet (Wolock et. al. Cell Systems, 2019)

- Core idea:
  - Simulate doublet by combing random pairs of cells
  - Remove cells if they are similar to the simulated doublets
  - Do not rely on library size at all

- Simulate pseudo-doublets:
  - the counts for gene g in doublet i with parent cells a and b is $Y_{gi} = Y_{ga} + Y_{gb}$

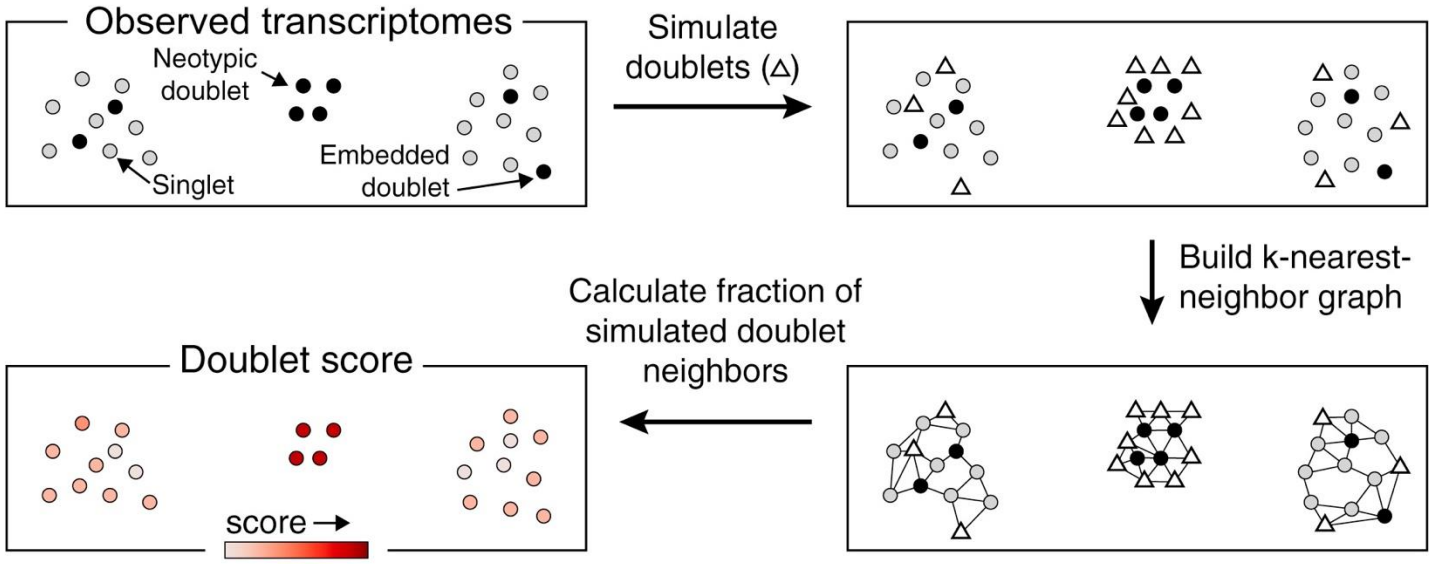
- KNN classifier to identify cells similar to the pseudo-doublets
  - Merge observe cells and pseudo-doublets and preprocess the merged data: Normalization, identify highly variable genes, scaling, PCA (more details in Lecture 3)
  - Find k nearest neighbors of each cell using Euclidean distance (by default)
  - q_i: (slightly adjusted) proportion of pseudo-doublets in k nearest neighbors of cell i

$$q_i = \frac{k_d(i)+1}{k_{adj}+2}$$

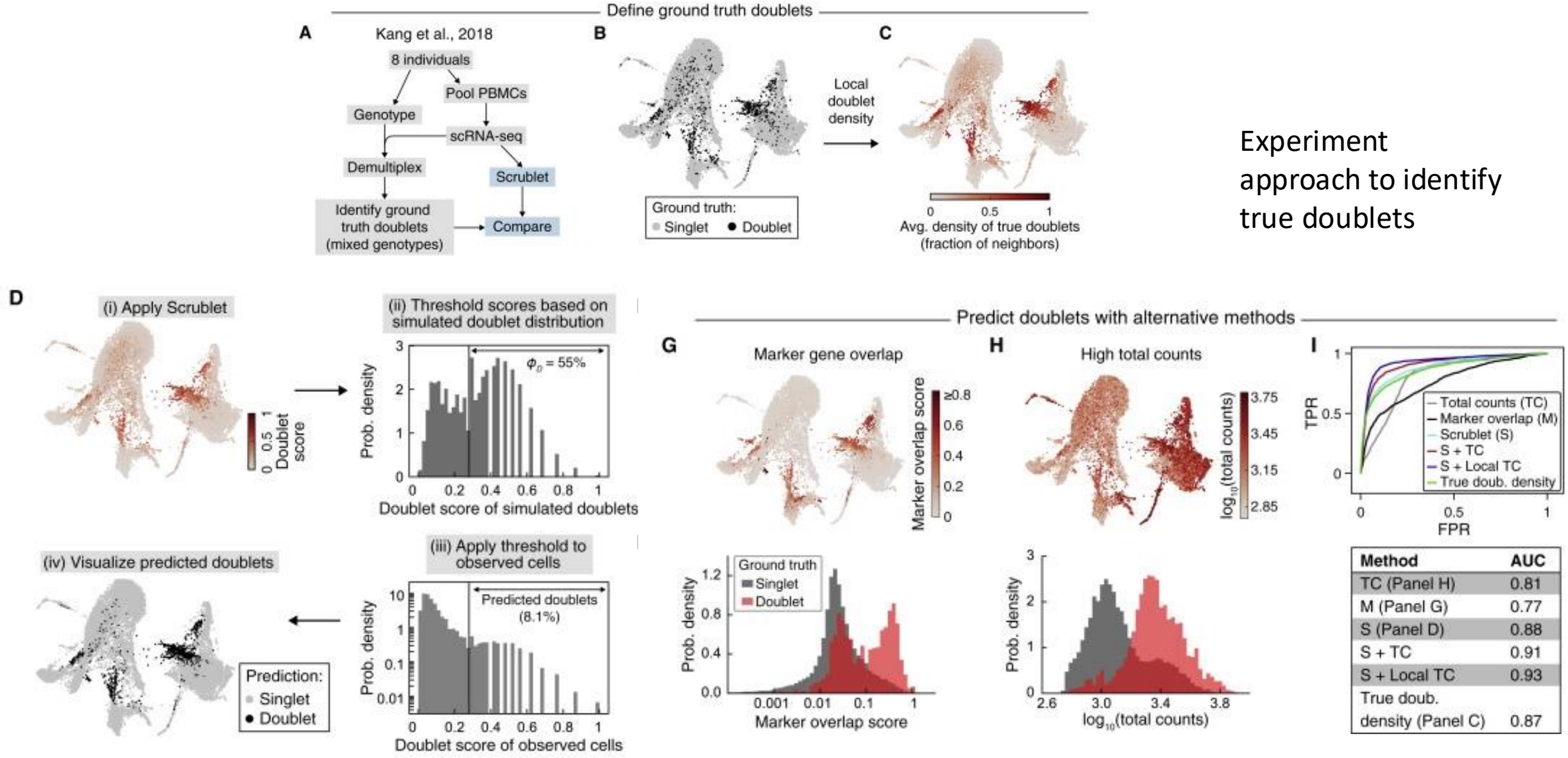
  - Remove a cell if $q_i > c_0$ where c_0 is some threshold
    - In the paper, they defined some Bayesian likelihood L_i which is monotone increasing in q_i

Scublet (Wolock et. al. Cell Systems, 2019)

- Two key tuning parameters: k and c_0
  - k: they used an adjusted k: k_{adj} = round(k·(1+r)) where $k = \text{round}(0.5\sqrt{\text{number of cells}})$ and r≥2 (they found this formula empirically)
  - c_0 The distribution of q_i is empirically bimodal and they define c_0 as valley between two modes



An example

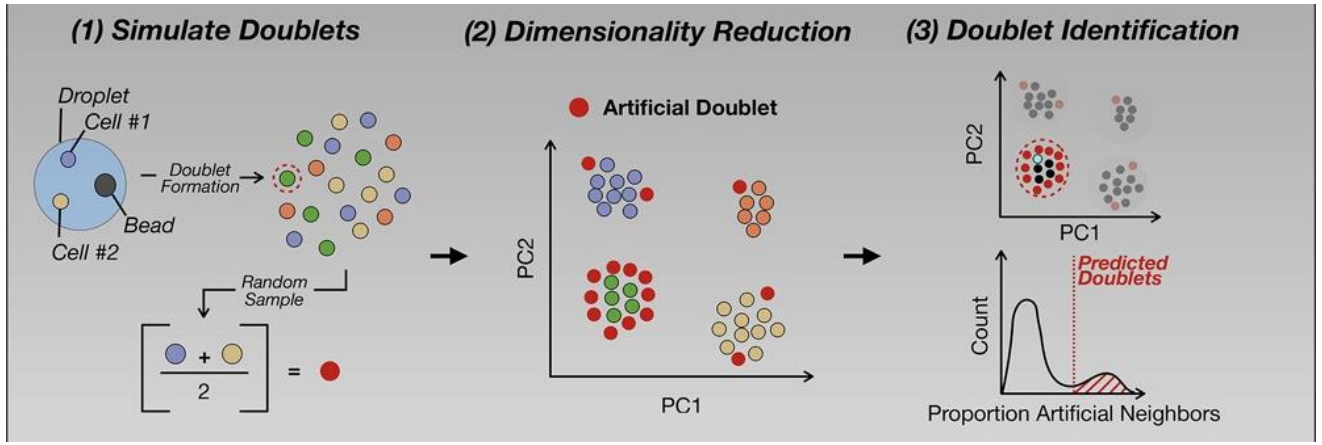


Method	AUC
TC (Panel H)	0.81
M (Panel G)	0.77
S (Panel D)	0.88
S + TC	0.91
S + Local TC	0.93
True doub. density (Panel C)	0.87

Experiment approach to identify true doublets

DoubletFinder (McGinnis et. al. Cell Systems, 2019)

- Same idea as Scublet
  - 25% pseudo-doublets in the merged data



- Different ways to choose tuning parameters: k and c_0
  - k: choose k to maximize the bimodality coefficient of the distribution of q_i
    - Bimodality coefficient (formula from SAS)

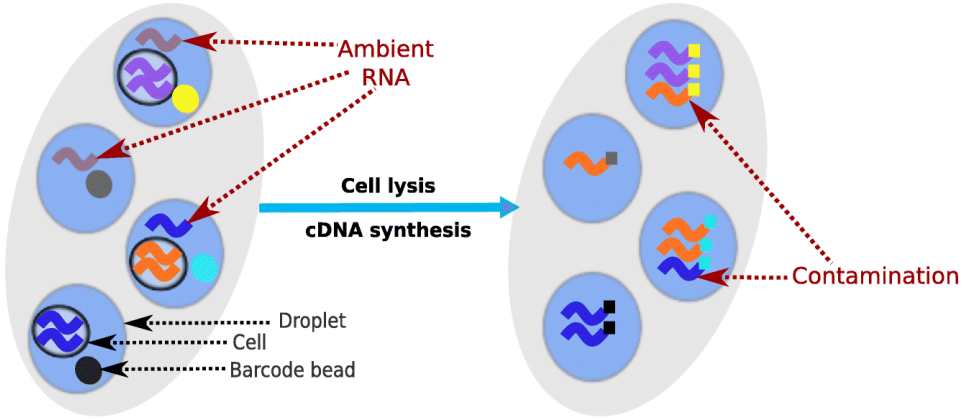
$$BC = \frac{\gamma^2+1}{\kappa+\frac{3(n-1)^2}{(n-2)(n-3)}}$$

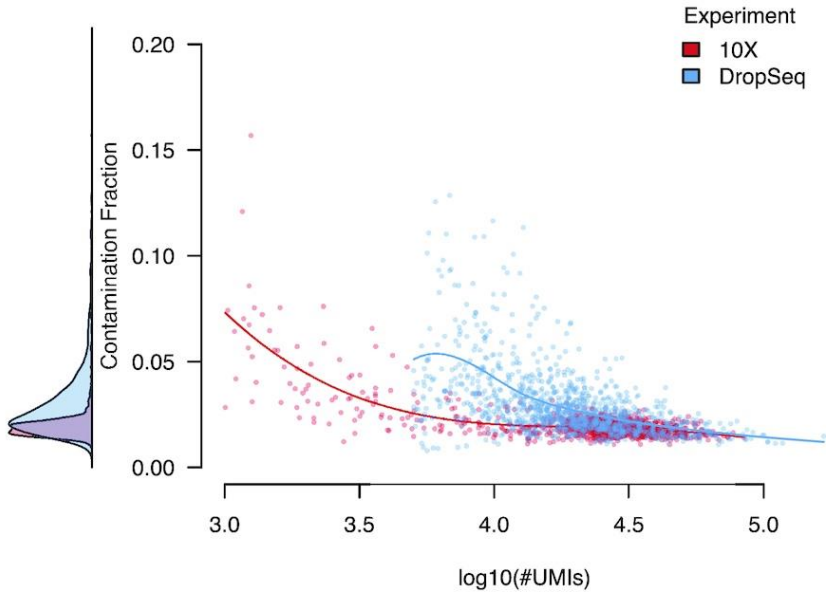
γ skewness,
κ kurtosis

    - Not very ideal, so they used a modified version
  - c_0: a pre-given proportion of doublets need to be detected
- DoubletFinder performs slightly better than Scublet in a benchmarking study (Xi and Li, Cell Systems 2021)

Ambient RNA

- In Droplet-based scRNA-seq platforms, a droplet can contain isolated RNAs even if it does not contain a cell
- Ambient RNA: pool of mRNA molecules that have been released in the cell suspension
- Ambient RNA also brings contamination to droplets that contain cells
- Ratio of contaminated RNA on average can be low ($\sim 2\%$, less than 10%), but the contamination rate can vary greatly across cells
- Why may we separate ambient RNA from mRNAs in the cell? → empty droplets serve as negative controls





EmptyDrops (Lun et. al. Genome Biology, 2019)

- Typically, we can identify droplets with no cells by the library size (library size too small)

- This paper argued that such method discards small cells with low RNA content
- Goal: rescue true cells with small library size

- This paper only detect empty droplets, it does not correct for ambient RNA in droplets with cells

- Core idea: find empty droplets use both the library size and gene expression profile
  - Learn an initial ambient profile
    - Estimate empty droplet gene expression distribution

  - Compute a p-value for each barcode to test whether the barcode is not an empty droplet

  - Keep barcodes as “cells” if they have small p-values or large enough library size

EmptyDrops (Lun et. al. Genome Biology, 2019)

- Estimate empty droplet gene expression distribution
  - Select barcodes whose library sizes are less than T as an initial pool of empty droplets
  - Assume that gene expressions in an empty droplet i follows

    $$(Y_{1i}, \cdots, Y_{Gi}) \sim \text{Dirichlet\_multinomial}(l_i, (\alpha_0 \tilde{p}_1, \cdots, \alpha_0 \tilde{p}_G))$$

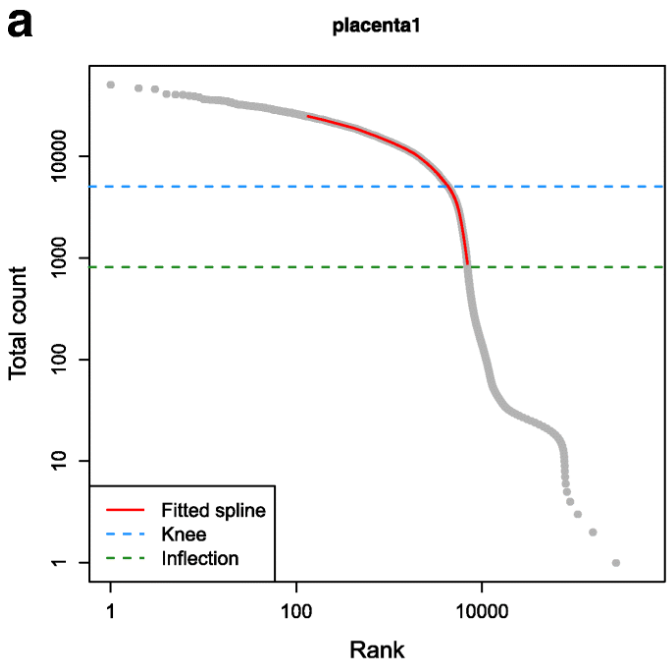
    [check Wikipedia for the definition]
  - $\tilde{p}_g$ is obtained by some empirical Bayes estimate to avoid reaching 0
  - α_0 estimated by maximum likelihood estimation given an estimated $\tilde{p}_g$

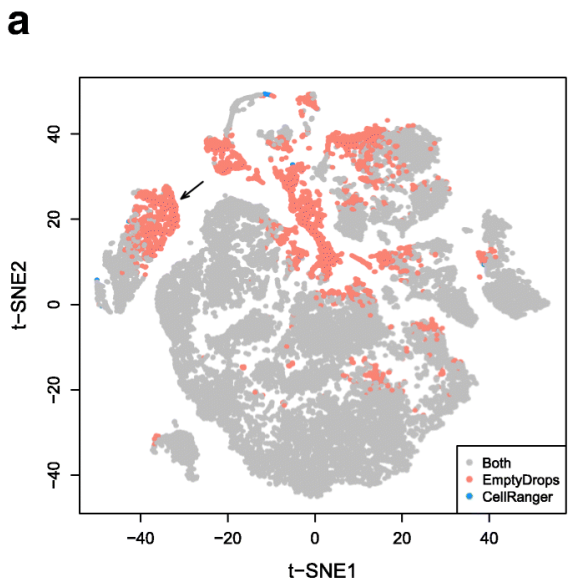
- Compute p-value to test whether a barcode is not an empty droplet
  - Essentially test whether an observation comes from a known distribution
  - Basically, you check if the observation b is at the tail of the density (likelihood in the paper)
  - Monte Carlo calculation of tail probability
    - Sample N new observations from the above estimated empty droplet distribution, get the density $L_{1b}, \cdots L_{Nb}$
    - Calculate p-value as proportion of $L_{1b}, \cdots L_{Nb}$ that are smaller than L_b (density of b)

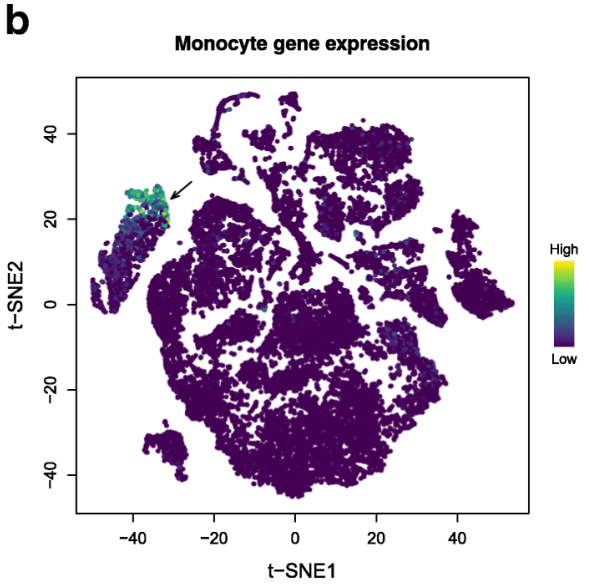
- Barcode selection
  - BH correction of p-values and select a barcode if library size $l_i > U$ where U is a knee point

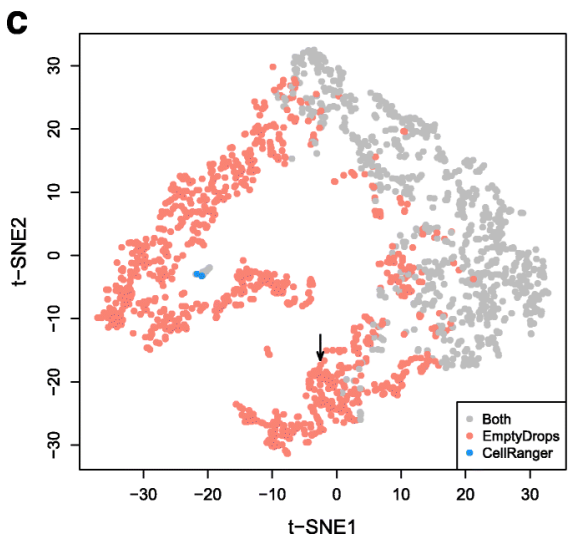
Conventional method

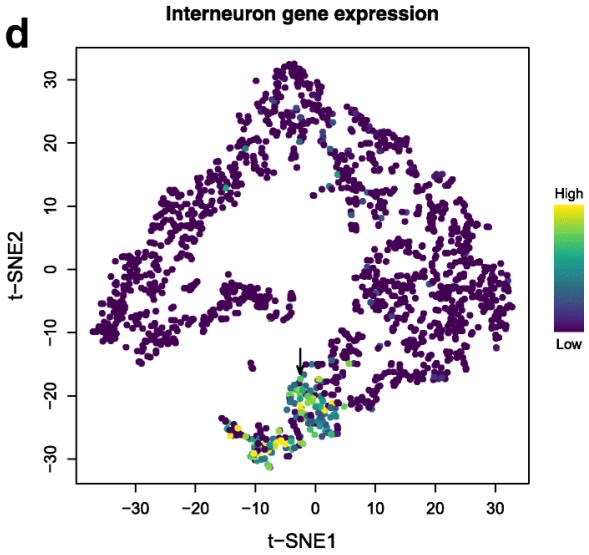
Some results







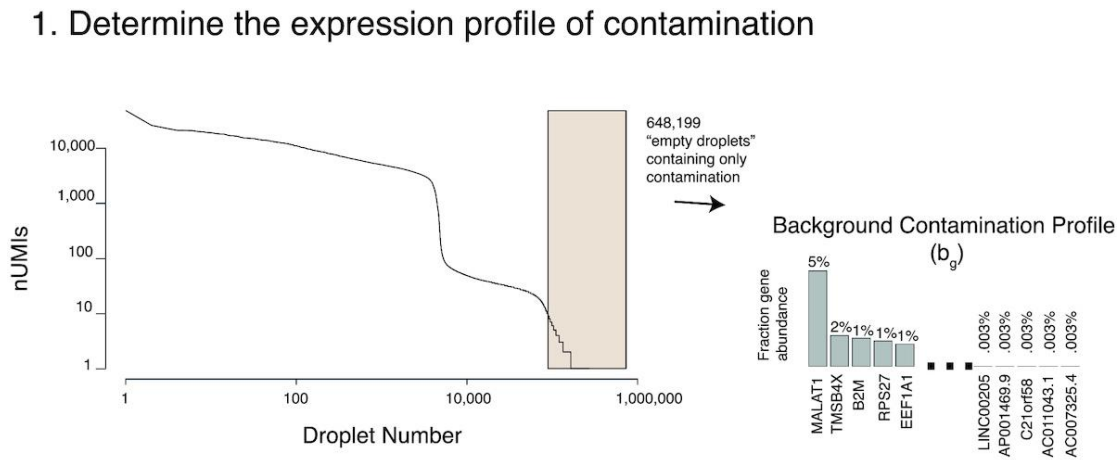




SoupX (Young et. al. GigaScience, 2020)

- Correct for ambient RNA confounding in cells

- Core idea:
  - Estimate ambient RNA gene expression profile from empty droplet (similar to EmptyDrops)



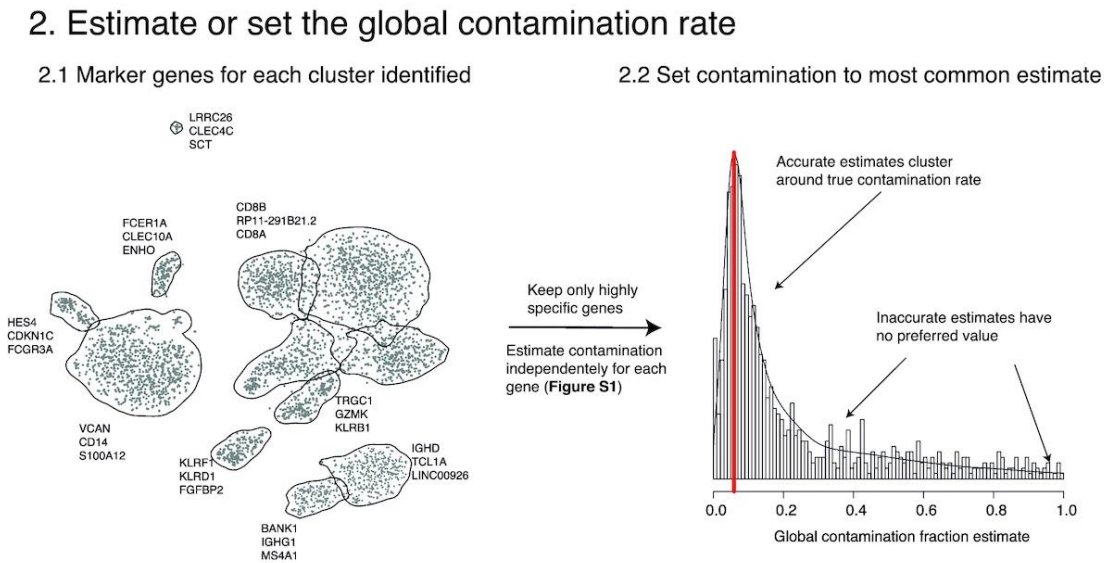
  - Use marker genes to determine proportion of contamination in each cell
  - Remove the estimated ambient RNA count for each gene from the observed counts

SoupX (Young et. al. GigaScience, 2020)

- Use marker genes to determine proportion of contamination in each cell

$$Y_{gc} = m_{gc} + o_{gc}$$

  - $o_{gc} = l_c \rho_c b_g$: ρ_c contamination rate in each cell
  - "Negative control" genes
    Assume that the marker genes for one cell cluster has zero expression in other cells
  - If gene g is a negative control for the cell, then $m_{gc} = 0$ and $Y_{gc}/(l_c b_g) \approx \rho_c$
  - Estimate ρ_c as the mode of the gene-specific estimated rates

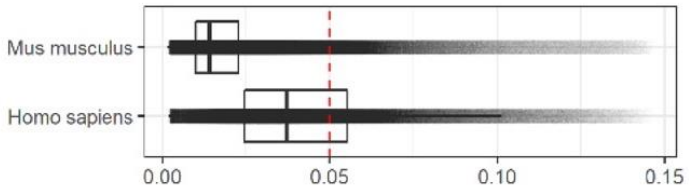


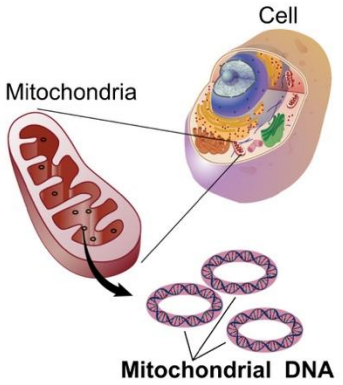
- Some adjustments to provide a good estimate of o_{gc} (need to be an integer, no greater than Y_{gc})

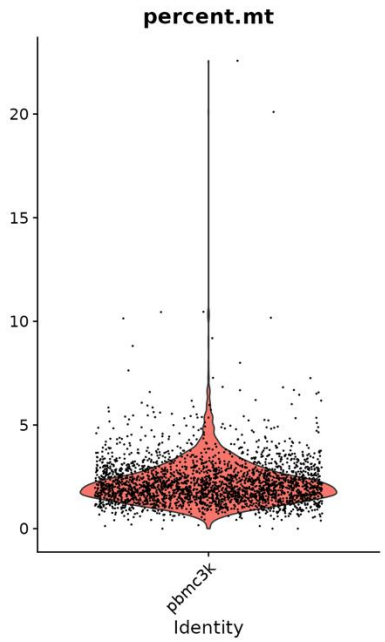
Low-quality cell filtering

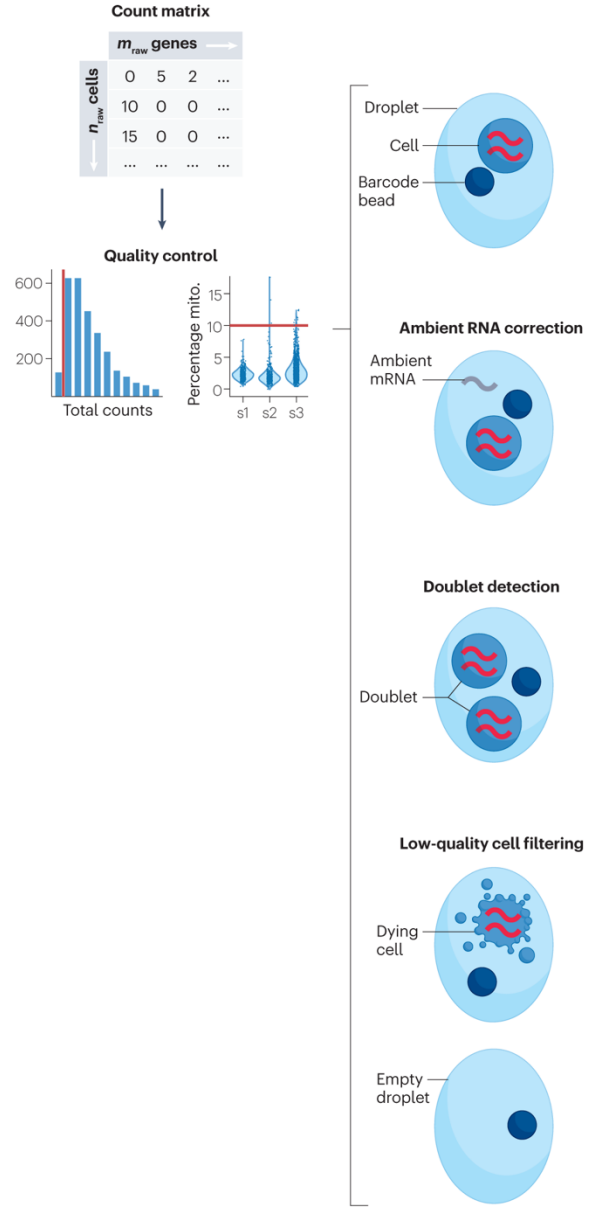
- Remove low-quality cells
  - Mitochondria also have DNA and can transcribe into RNA
  - Mitochondrial mRNA also have poly-A tail that are captured in scRNA-seq
  - High expression levels of mitochondrial genes can be an indicator of lysing cells

  - Remove cells that have a high proportion of reads from mitochondrial genes (default 5%)
    - Maybe better to use 10% for human cells (Osorio and Cai, Bioinformatic 2021)









[Heumos et. al., Nature reviews genetics 2023]

Related papers

- Wolock, S. L., Lopez, R., & Klein, A. M. (2019). Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell systems, 8(4), 281-291.
- McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell systems, 8(4), 329-337.
- Lun, A. T., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., Participants in the 1st Human Cell Atlas Jamboree, & Marioni, J. C. (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome biology, 20, 1-9.
- Young, M. D., & Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience, 9(12), giaa151.
- Osorio, D., & Cai, J. J. (2021). Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics, 37(7), 963-967.
- Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., ... & Theis, F. J. (2023). Best practices for single-cell analysis across modalities. Nature Reviews Genetics, 24(8), 550-572.