Quality Control for Single Cell RNA-seq Data

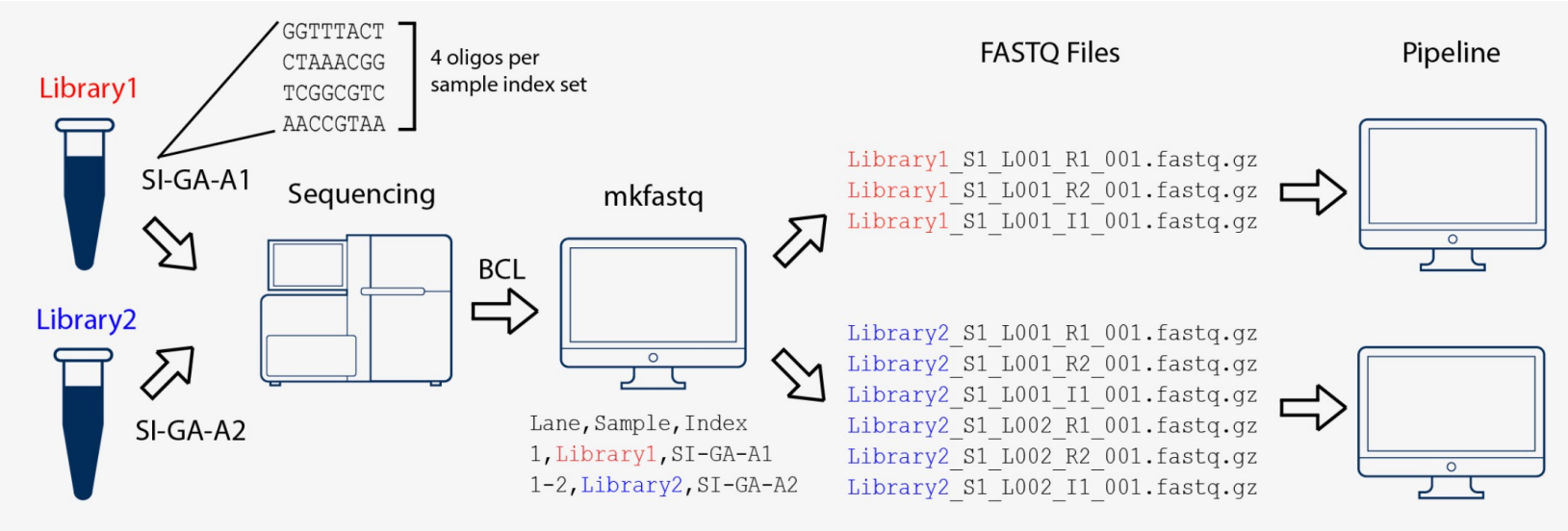
Daniel Chafamo
Seq 28, 2022

Topics

- Overview of steps to go from raw sequencing files to count matrices
- Quality control for single cell RNA-Seq Data
- Normalization of single cell RNA-Seq data

Raw sequencing files to count matrices
Step 1: BCL to FASTQ

- The primary output of Illumina sequencing instruments are per-cycle **base call files** in **BCL** format.
- The first step is to convert these BCL files to fastq files used by most downstream software. This includes separating reads into individual fastq files based on their barcode (demultiplexing), adapter masking/trimming and moving unique molecular identifier (UMI) bases from the read to the fastq header.



Source: 10X Genomics

Raw sequencing files to count matrices
Step 1: BCL to FASTQ

- A FASTQ file is a text file that contains the sequence data which consists of:

1. A sequence identifier with information about the sequencing run and the cluster.
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores.

- Example sequence in a FASTQ

```
@HWI-ST808:130:H0B8YADXX:1:1101:2088:2222:CELL_GGTCCA:UMI_CCCT
AGGAAGATGGAGGAGAGAAGGCGGTGAAAGAGACCTGTAAAAAGCCACCGN
+
@@@DDBD>=AFCF+<CAFHDECII:DGGGHGIGGIIIEHGIIIGIIDHII#
```

Raw sequencing files to count matrices
Step 2: Alignment to reference genome

- The next step is to determine which gene each read originated from. In order to do this the read sequences are mapped to a precompiled genome reference.

- Many tools have been developed for read alignment.

- 10X Cellranger uses the STAR aligner. For each read in the FASTQ, STAR tries to find the longest possible sequence which matches one or more sequences in the reference genome.

- Because of widespread splicing in animal genomes, read alignment against a genome has to be done with a splice-aware aligner.

Raw sequencing files to count matrices
Step 3: Counting reads

- Reads that have been confidently mapped to the transcriptome are then assigned to cells based on their barcode (cell barcode demultiplexing) and the the number of unique RNA molecules corresponding to each gene within each cell are counted (UMI deduplication).

- The result is the gene x cell matrix that is the starting point for downstream analysis:

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

Terra: a scalable platform for biomedical research

Core capabilities designed to support research

Data Library

Access public and access-controlled datasets

Workspaces

Bring together data and tools into secure, shareable units

Workflows

Run workflows at scale; bring your own or explore community favorites

Interactive Analysis

Analyze data with built-in applications like Jupyter Notebooks, RStudio, Galaxy

BROAD INSTITUTE

https://terra.bio

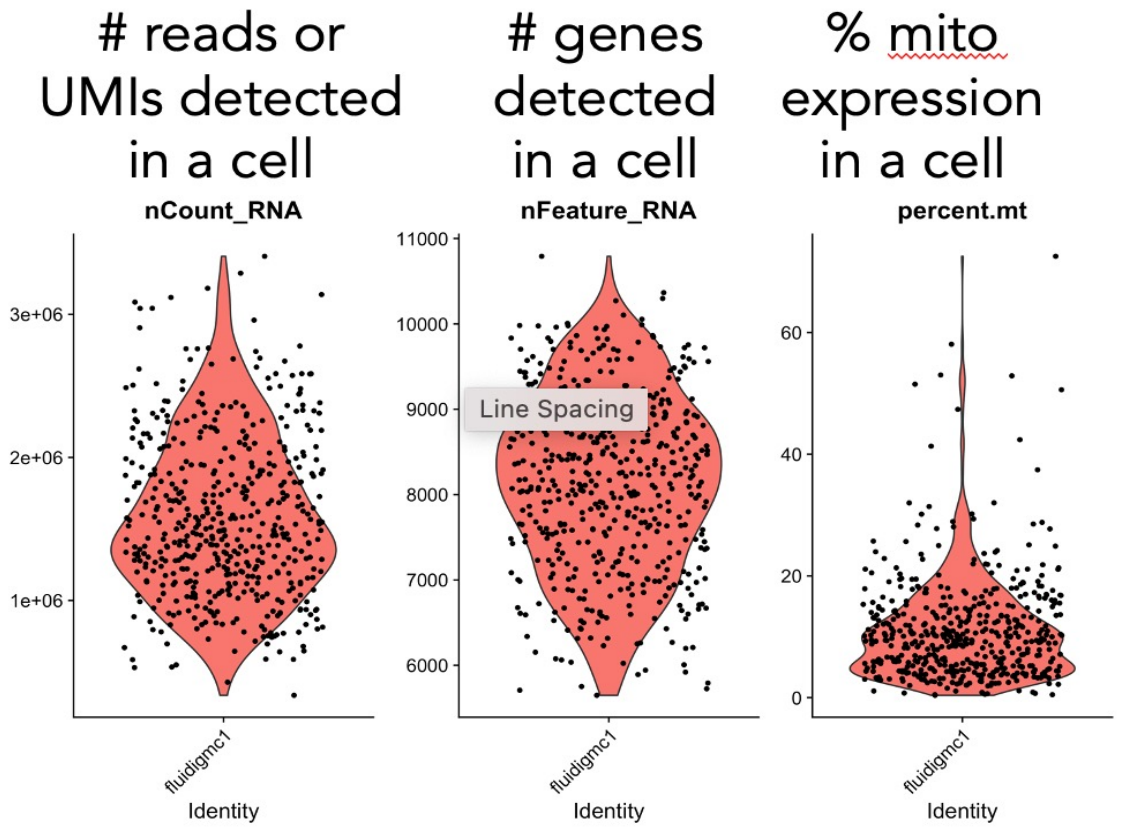
Quality Control

- Goals:
  - Filter the data to only include true cells that are of high quality
    - Remove empty droplets
    - Remove doublets
    - Remove dead or damaged cells
    - Remove ambient RNA
  - Identify any failed samples and either try to salvage the data or remove from analysis

- Challenges:
  - Delineating cells that are poor quality from less complex cells
  - Choosing appropriate thresholds for filtering, to keep high quality cells without removing biologically relevant cell types

Fundamental QC metrics

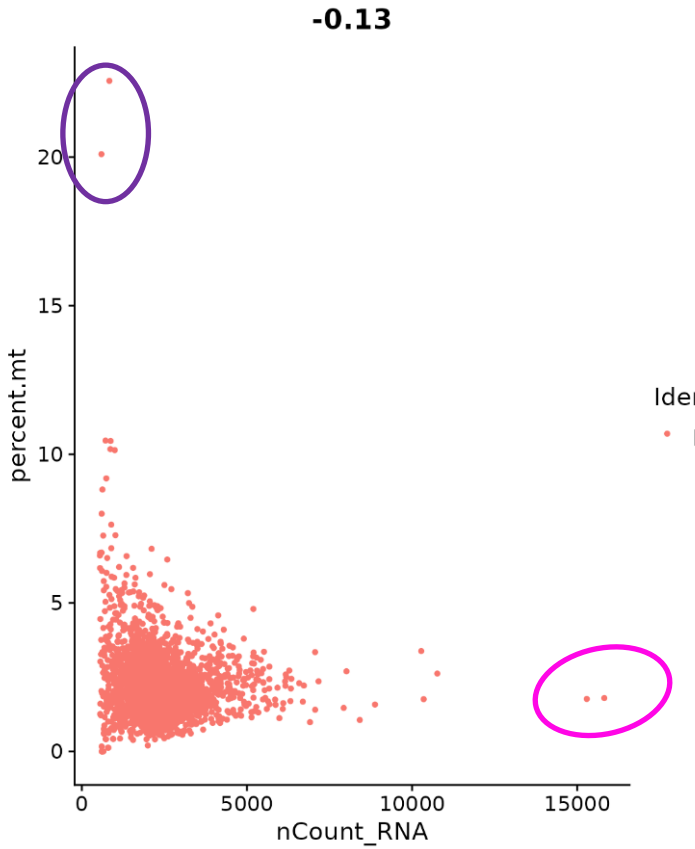
- Cell QC is commonly performed based on three principal QC covariates:
  - the number of counts per barcode (count depth),
  - the number of genes per barcode, and
  - the fraction of counts from mitochondrial genes per barcode

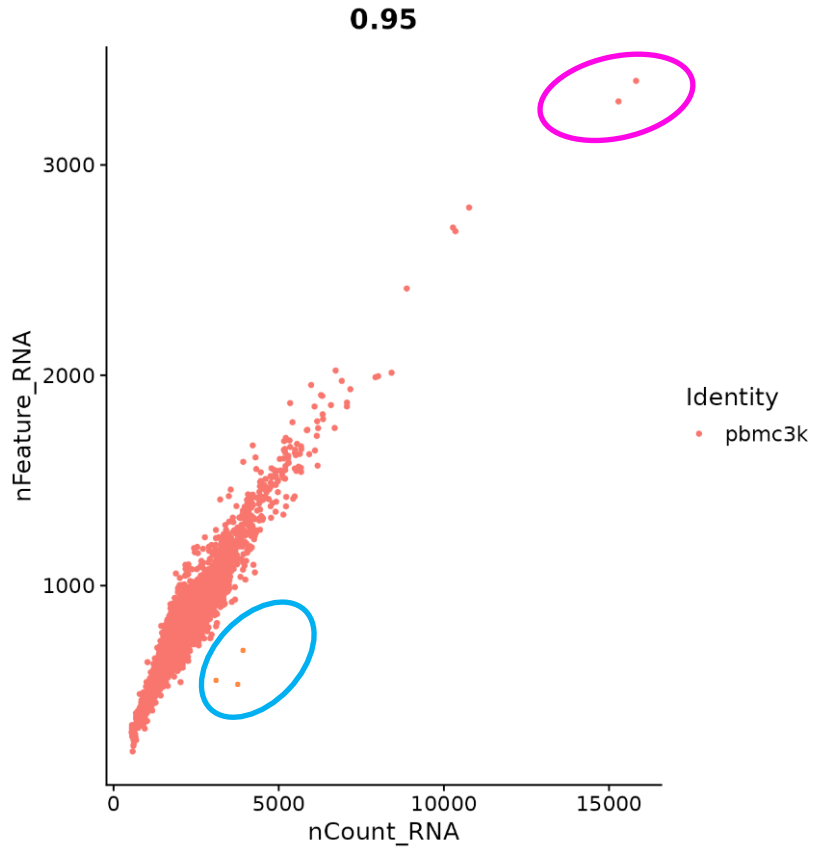


Fundamental QC metrics

It is sometimes helpful to use a combination of these metrics to filter out cells.

- **Low nUMI and high % mitochondrial -** Cells captured but lost a lot of the mRNA, and the mitochondrial genes were protected and retained.
- **High nUMI & low nGene ratio –** low quality library or capture rate
- **High nUMI & high nGene –** doublets

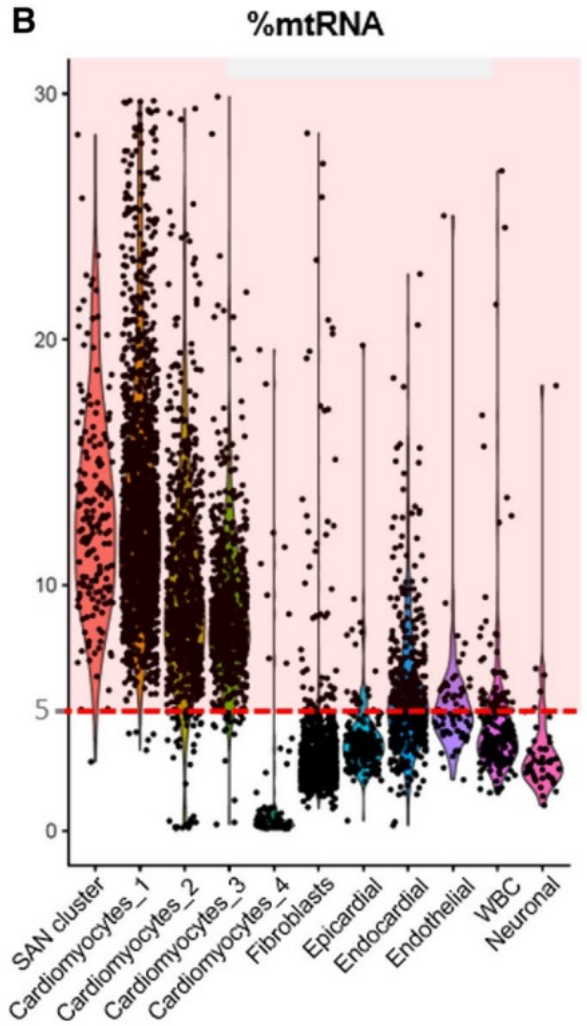




Appropriate quality control filters vary with platform and cell types

- Different platforms set different expectations
  - Example:. Smart-Seq2 often yields more genes detected per cell than 10x Chromium.

- Different cell types set different expectations
  - Immune cells normally have fewer genes detected per cell than non-immune cells
  - Malignant cells normally have more genes detected per cell than non-malignant cells
  - Cardiac cells normally have higher percent of mito genes per cells then other cells



Source:

Tips

- It may be necessary to revisit quality control decisions multiple times when analyzing data. Often it is beneficial to start with permissive QC thresholds and investigate the effects of these thresholds before going back to perform more stringent QC.

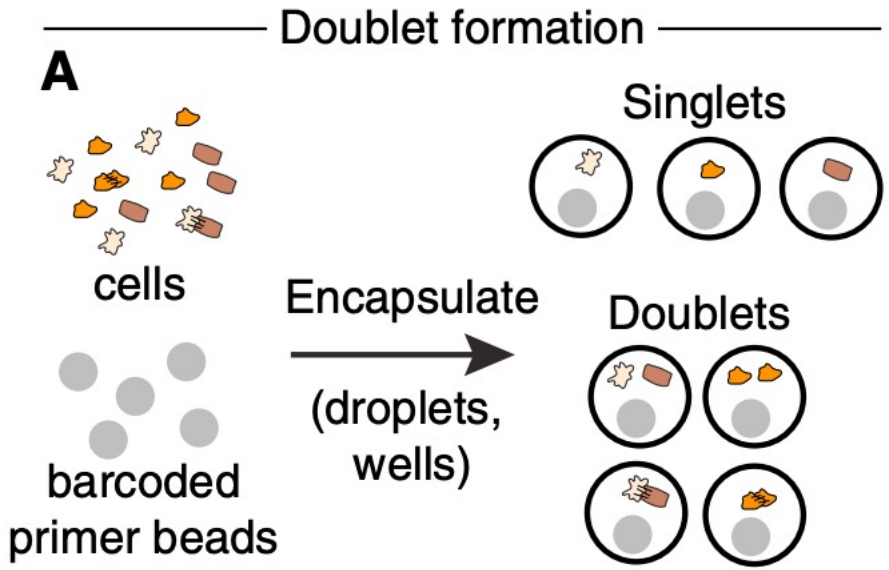
- If the distribution of QC covariates differ between samples, QC thresholds should be determined separately for each sample to account for sample quality differences.

- Always visualize QC metrics per cluster in order to flag any biases in QC filters that have been applied.

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial.

Identifying Doublets in Single-Cell RNA-Seq Data

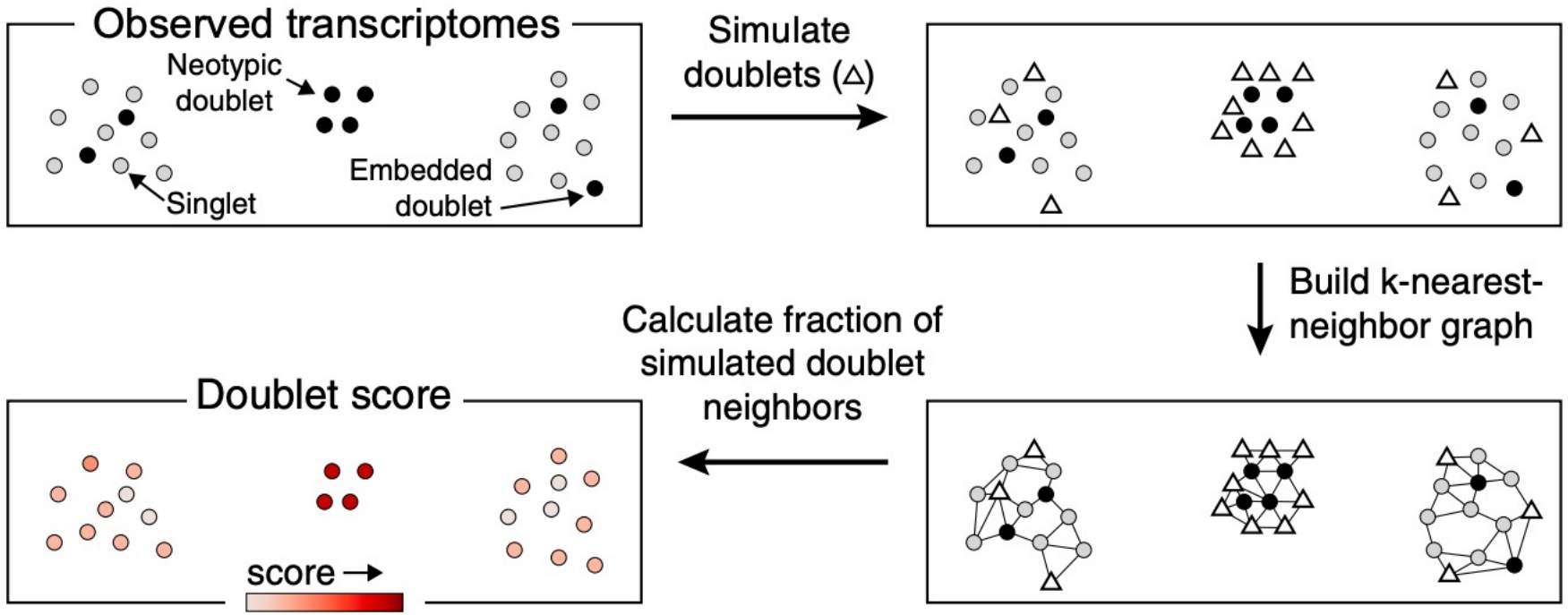
- Doublets = two or more cells captured together
- Doublets will contain gene counts from the two combined cells.
- Some doublets will be filtered by the basic QC steps since doublets will have a high number of genes and number of UMIs



Wollock et al., Cell Syst., 2019

Identifying Doublets in Single-Cell RNA-Seq Data

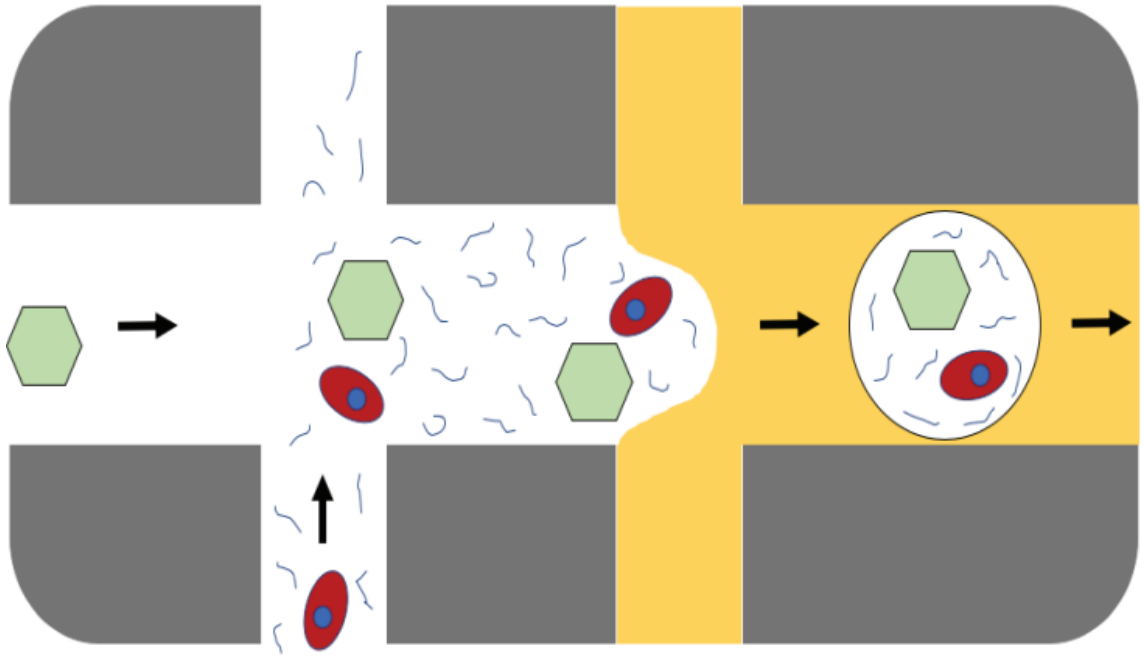
- A simple heuristic to check for doublet clusters is to see if they express gene markers of two or more disparate celltypes.
- Advanced methods such as Scrublet and DoubletFinder use simulations to determine doublet scores.



Wollock et al., Cell Syst., 2019

Detecting empty drops and correcting for ambient RNA

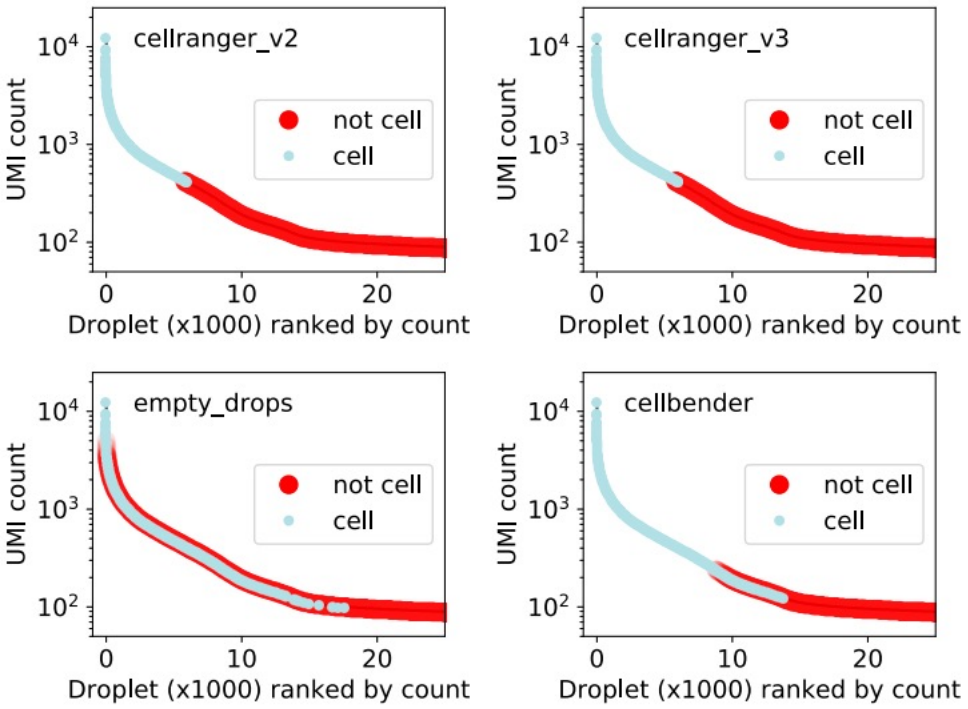
- Ambient gene expression refers to counts that do not originate from a barcoded cell, but from other lysed cells whose mRNA contaminated the cell suspension prior to library construction.
- Sequencing errors in barcodes and Barcode swapping also contribute to ‘ambient’ RNA counts.
- The presence of background RNA can lead to systematic biases and batch effects in various downstream analyses such as differential expression and marker gene discovery.

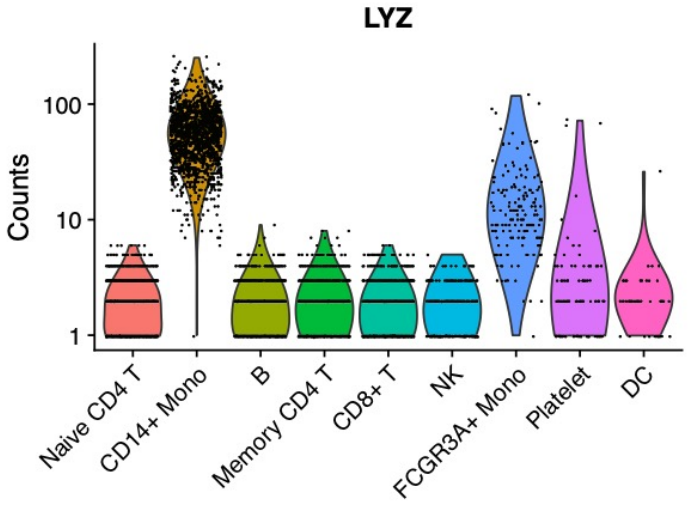


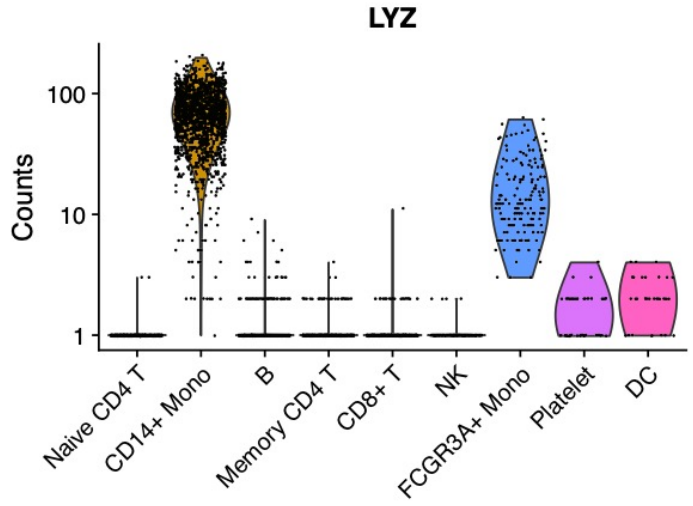
CellBender: Tool to detect empty drops and correct ambient RNA

CellBender removes ambient background and barcode swapping via deep learning

- successfully learns and subtracts background noise and artifactual counts from non-empty droplets and leads to significantly increased amplitude and specificity of differential gene expression







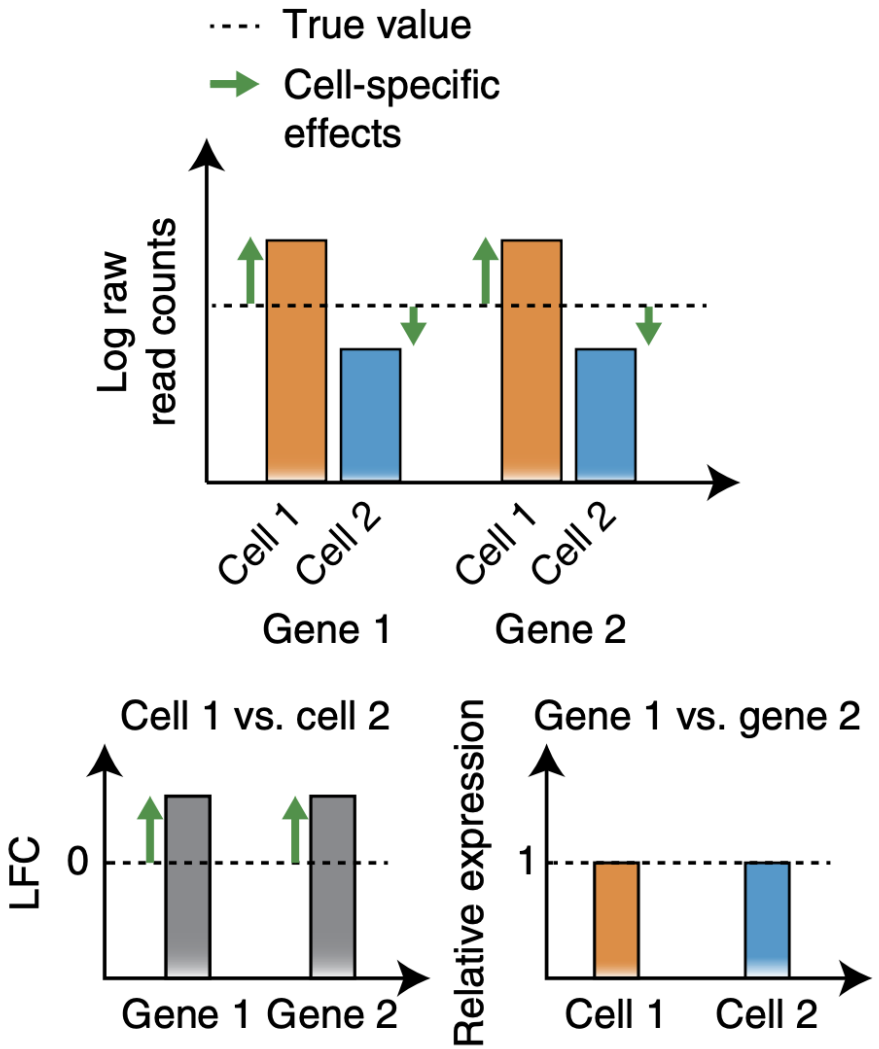
Fleming et al. *bioRxiv* (2019)

Normalization of gene expression data

- Normalization is the process of adjusting raw data to account for unwanted factors that prevent direct comparison of expression measures.

- The goal of normalization is for differences in normalized read counts to represent differences in true expression.

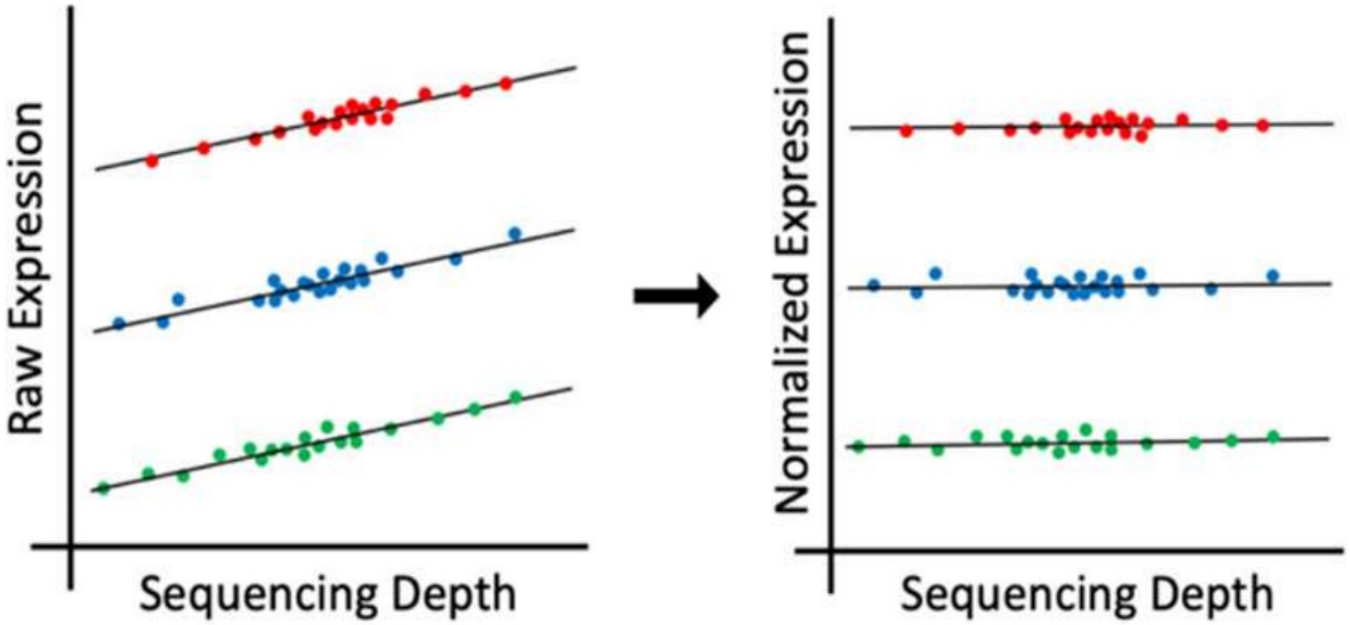
- In practice, single cell RNA-Seq normalization involves two steps:
  - Scaling
  - Transformation



Why normalize gene expression within a cell?

1. Many technical factors such as sequencing depth can introduce bias into the raw read counts obfuscating true signal.

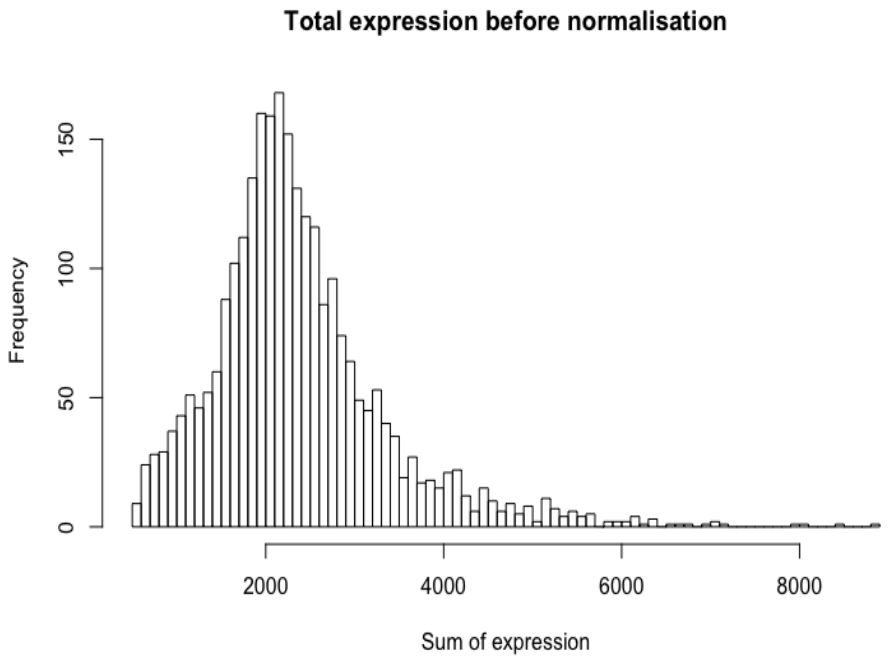
- sequencing depth = total number of molecules sequenced
- When one cell has a higher sequencing depth than another, even non-differentially expressed genes will tend to have higher read counts in that cell

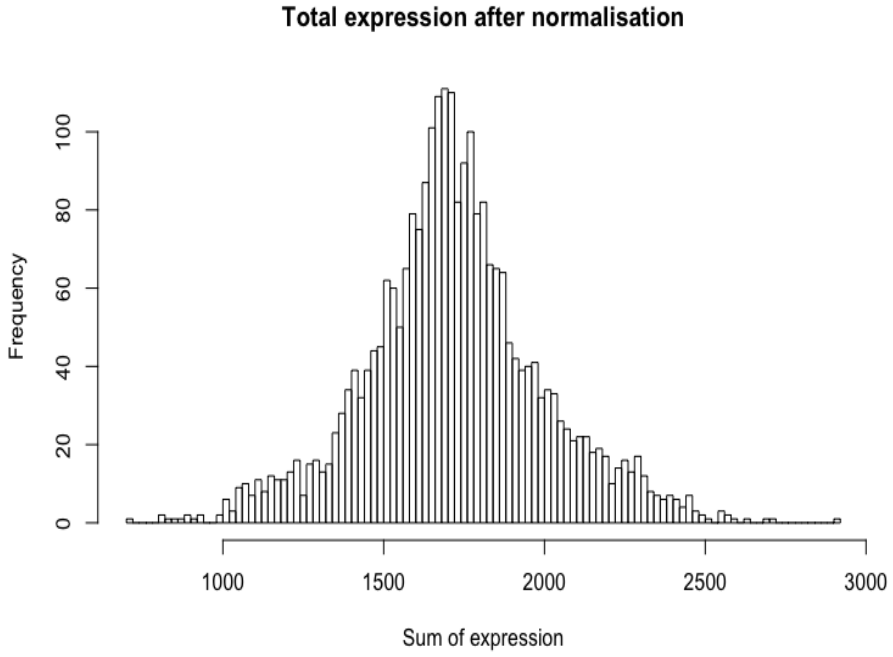


Why normalize gene expression within a cell?

2. There are typically extreme values in distribution of gene expression

- Normalization, especially the transformation step, reduces the skew-ness of the data to approximate the assumption of many downstream analysis tools that the data are normally distributed.





Why normalize gene expression within a cell?

2. There are typically extreme values in distribution of gene expression

- Normalization, especially the transformation step, reduces the skew-ness of the data to approximate the assumption of many downstream analysis tools that the data are normally distributed.
- Also prevents downstream analysis from being completely dominated by differences among the most highly expressed genes. Log transformation for instance

Raw data

	Cell Type A	Cell Type B	Δ
Gene 1	1	2	1
Gene 2	100	200	100

Log_2 transform

Cell Type A	Cell Type B	Δ
0	1	1
6.64	7.64	1

Normalization of gene expression data

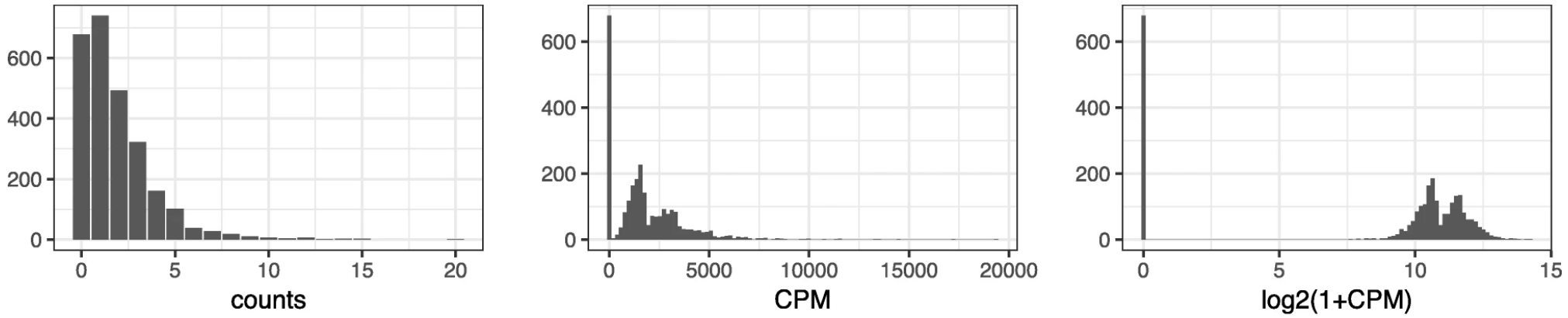
- The most commonly used normalization protocol is count depth scaling, also referred to as “counts per million” or CPM normalization.

- To perform CPM:

  - Gene expression measurements for each cell are normalized by the total gene expression or median gene expression
  - Gene expression values then scaled to sum to 10,000 (typically),
  - Finally, these values are log-transformed: log(CPM+1).

Is standard normalization appropriate?

- Normalization methods perform poorly when their assumptions are violated
- Current approaches to normalization and transformation artificially distort differences between zero and nonzero counts.
- Advanced methods like SCTransform can address some of the limitations
  "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression." Hafemeister et al. Genome Biology (2019)



(a) UMI counts (b) counts per million (CPM) (c) log of CPM