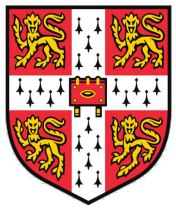


UNIVERSITY OF CAMBRIDGE

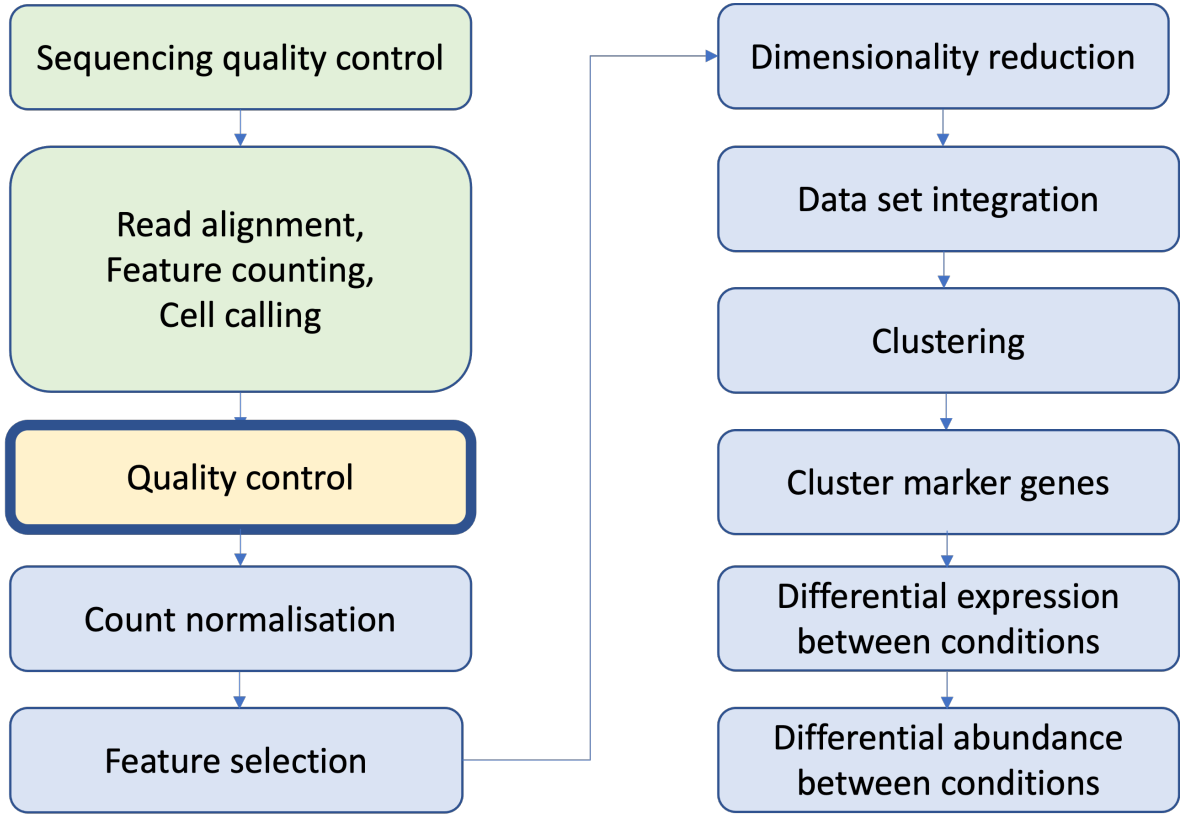
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Introduction to single-cell RNA-seq analysis

Quality Control

12th September 2022

Single Cell RNAseq Analysis Workflow



10x overview

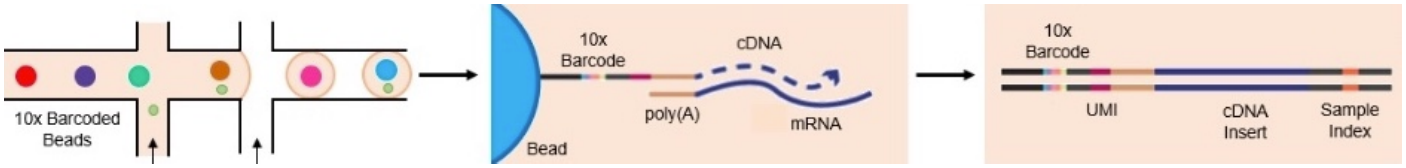


Image source: https://web.genewiz.com/single-cell-faq

Not every droplet is useble

A single happy cell in a droplet is ideal

- Complex transcriptome
- Average number of genes detected

Empty droplet: No cell in a droplet

- No genes detected

Droplet with ambient RNA

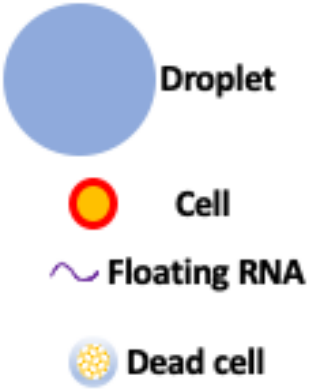
- Low complex transcriptome
- Genes detected much lower than average genes per cell

Droplet with dead cell

- Enriched for mitochondrial genes

Droplet with multiple cell

- Very complex transcriptome
- Genes detected much higher than average genes per cell



Quality Control overview

- Aim of QC is …
  - To remove undetected genes
  - To remove empty droplets
  - To remove droplets with dead cells
  - To remove Doublet/multiplet
  - Ultimately To filter the data to only include true cells that are of high quality
- Above is achieved by …
  - Applying hard cut-off or adaptive cut-off on …
    - Number of genes detected per cell
    - Percent of mitochondrial genes per cell
    - Number of UMIs/transcripts detected per cell

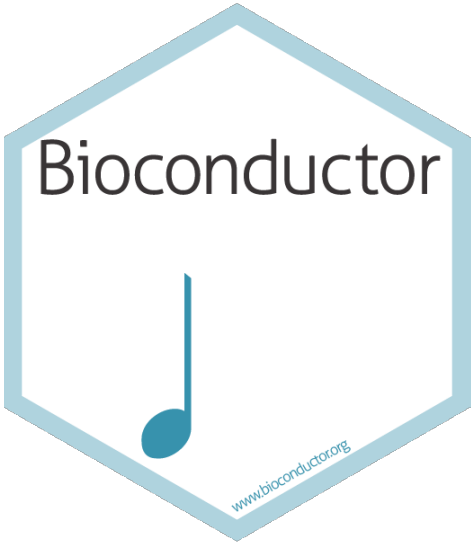
Quality Control

Bioconductor R packages:

- *scran*: Collection functions for interpretation of single-cell RNA-seq data
- *scater*: For focus on quality control and visualization.
- *DropletUtils*: Handling single-cell (RNA-seq) data from droplet technologies such as 10X Genomics

Orchestrating Single-Cell Analysis with Bioconductor
Robert Amezquita, Aaron Lun, Stephanie Hicks, Raphael Gottardo

http://bioconductor.org/books/release/OSCA/



Read CellRanger outputs into R

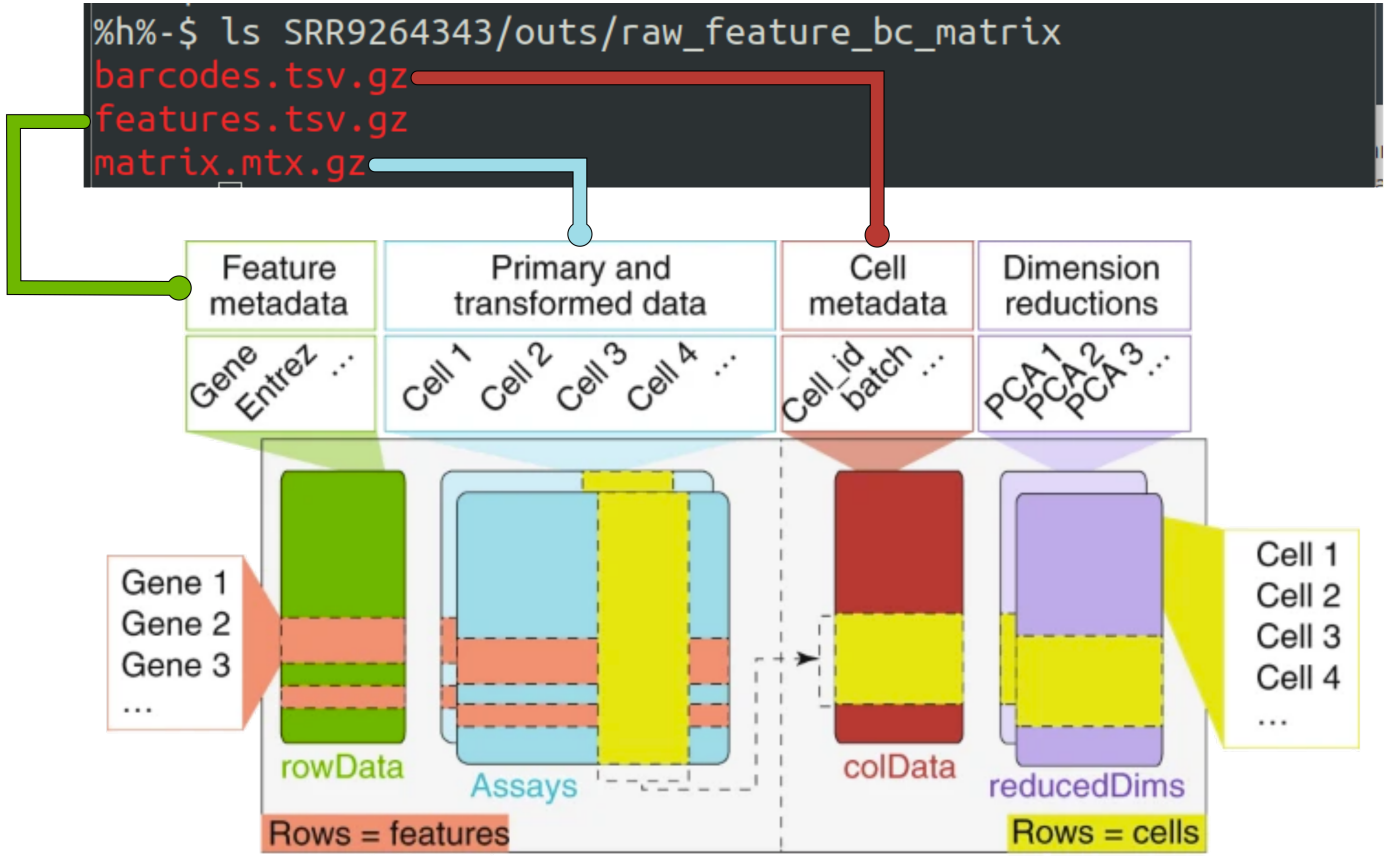
- CellRanger outputs: gives two output folders raw and filtered
- Each folder has three zipped files
  - features.tsv.gz, barcodes.tsv.gz and matrix.mtx.gz
  - raw_feature_bc_matrix
    - All valid barcodes from GEMs captured in the data
    - Contains about half a million to a million barcodes
    - Most barcodes do not actually contain cells
  - filtered_feature_bc_matrix
    - Excludes barcodes that correspond to this background
    - Contains valid cells according to 10x cell calling algorithm
    - Contains 100s to 1000s of barcodes

```
%h%-$ ls SRR9264343/outs/raw_feature_bc_matrix
barcodes.tsv.gz
features.tsv.gz
matrix.mtx.gz
```

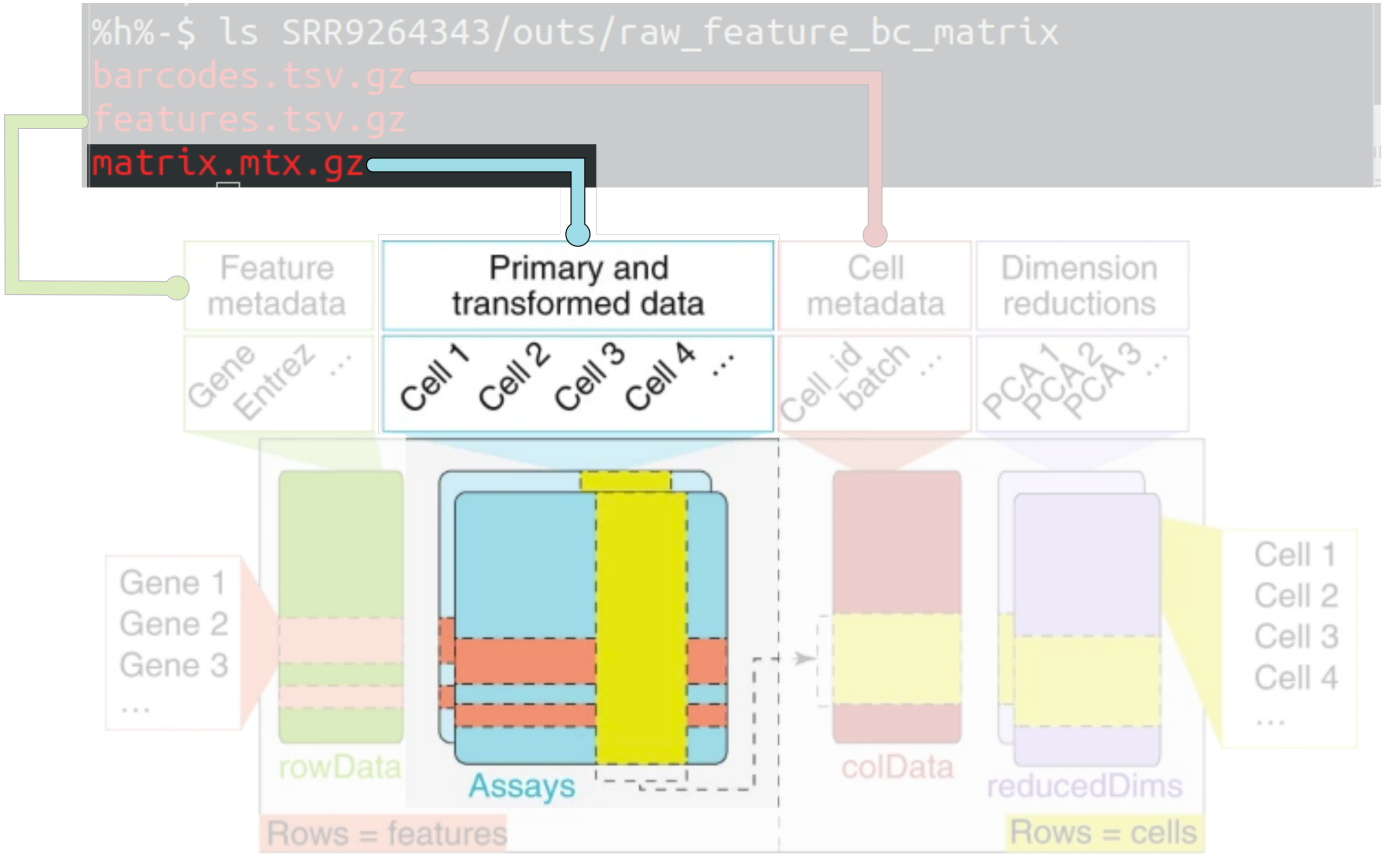
Single Cell Experiment Vocabulary alert

- cell = Barcode = droplet
- Transcript = UMI

The *SingleCellExperiment* object

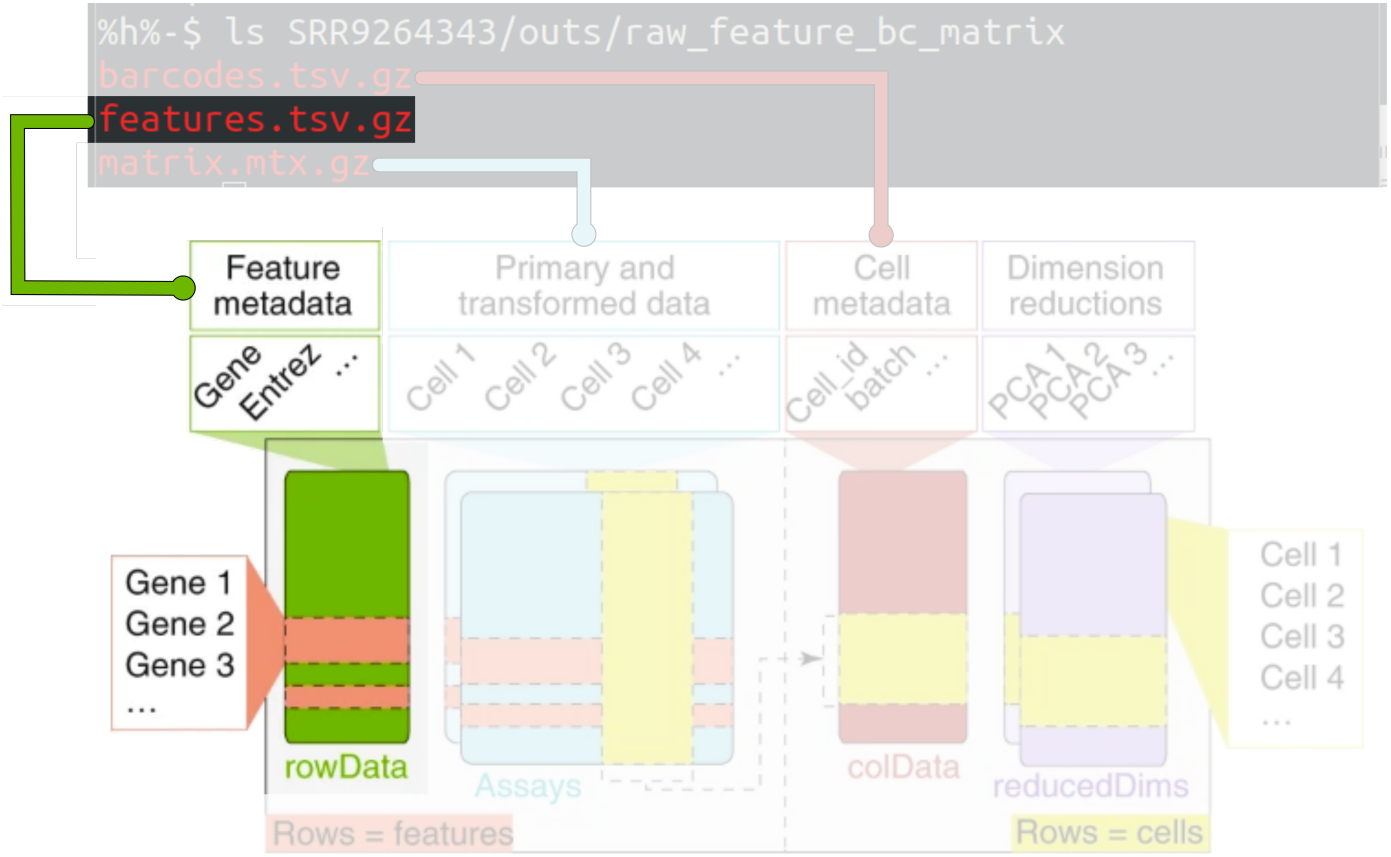


The Counts Matrix



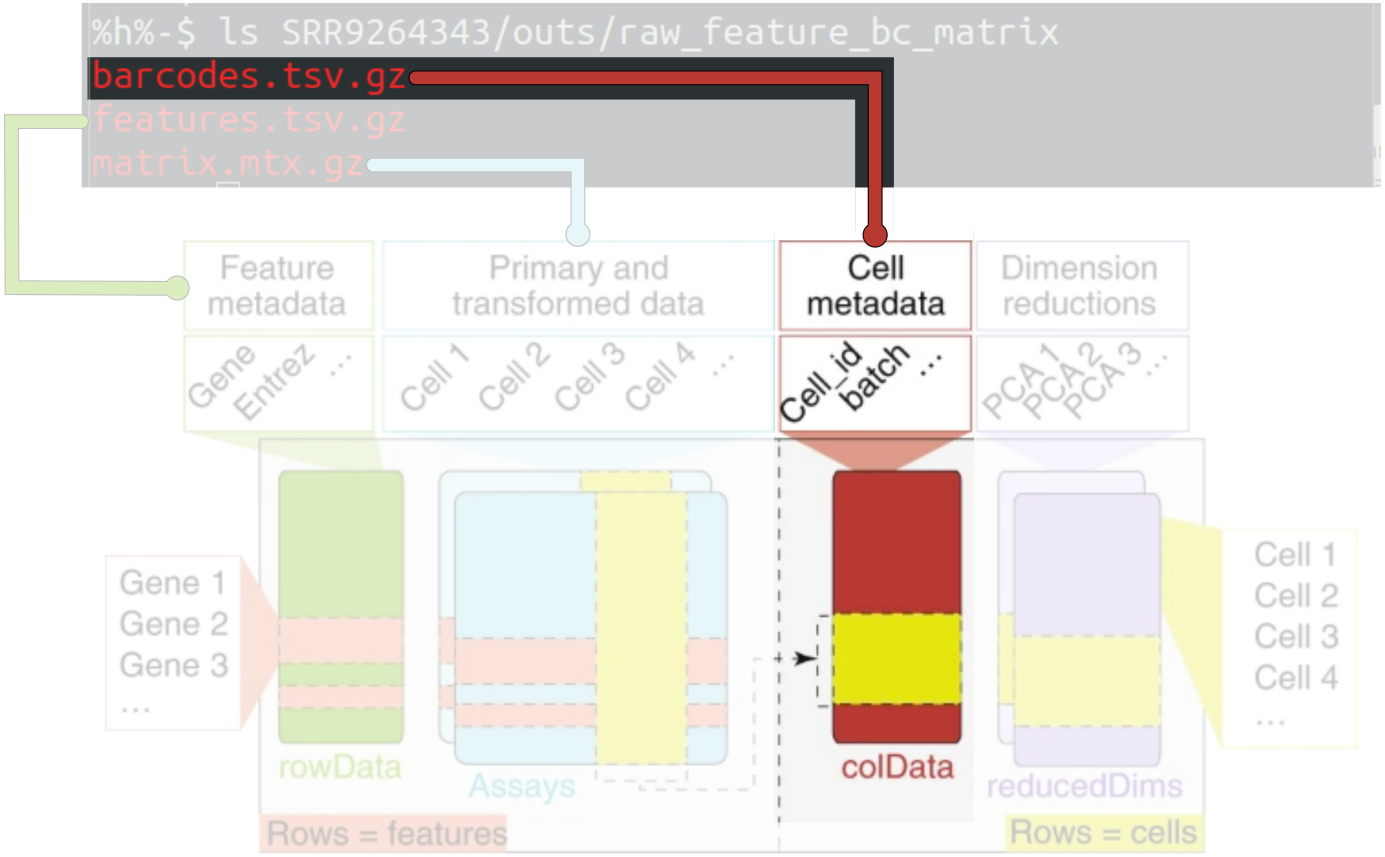
To access counts from sce object: **`counts(sce)`**

Feature metadata



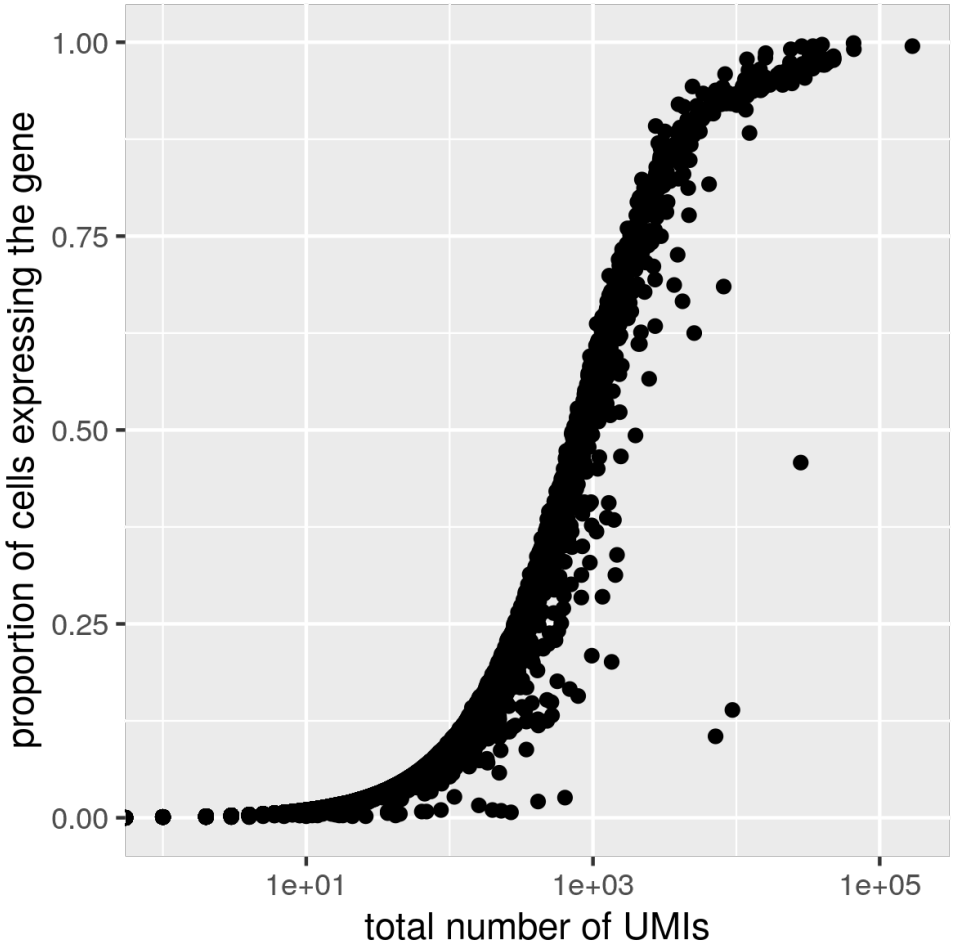
To access gene metadata from sce object: **`rowData(sce)`**

Droplet annotation (Cell metadata)

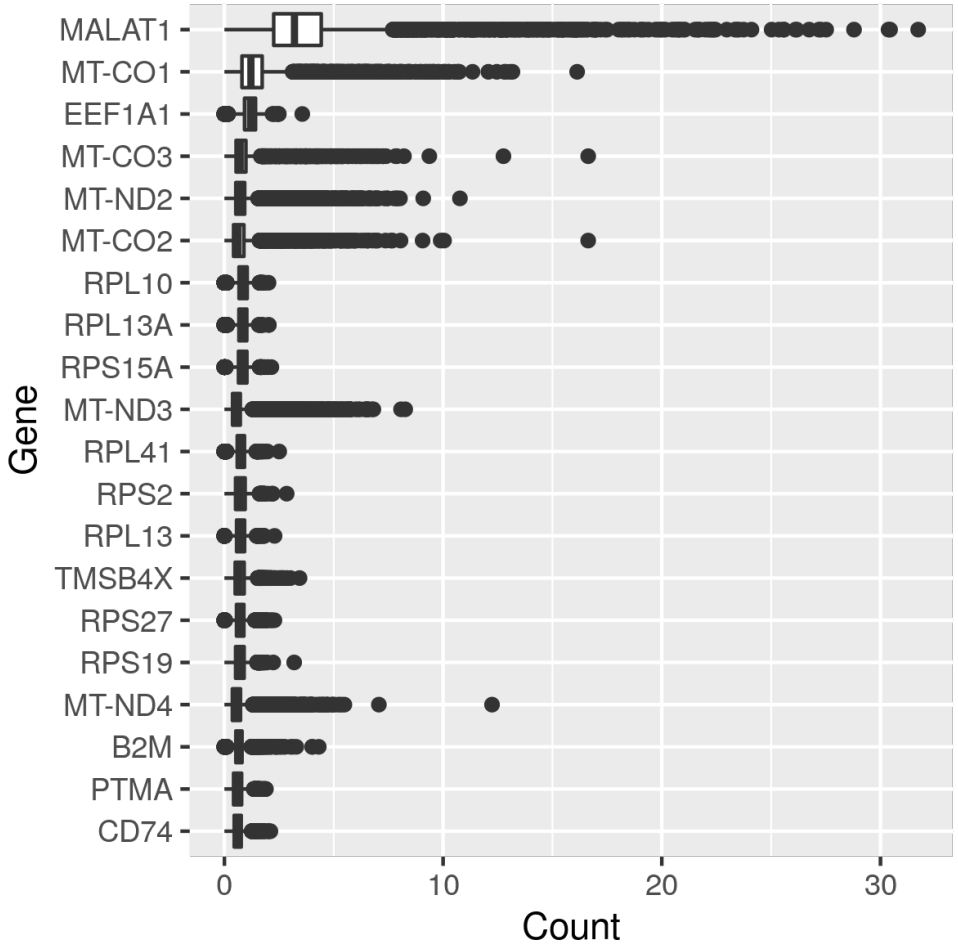


To access cell metadata from sce object: **`colData(sce)`**

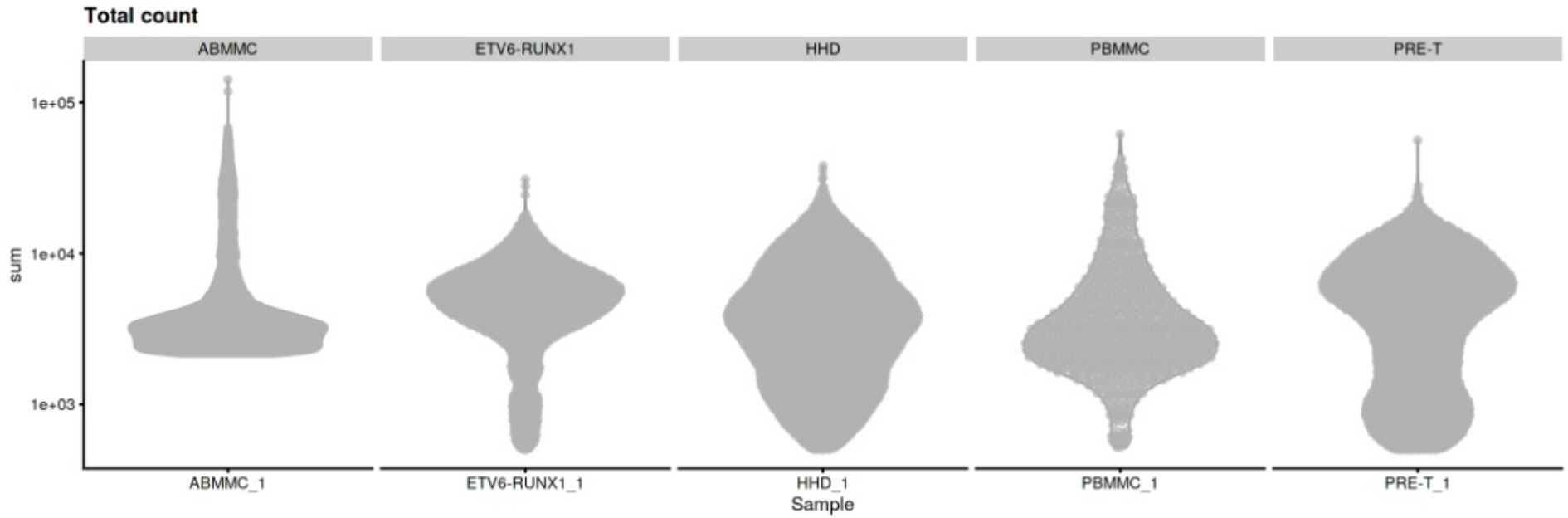
Properties of RNAseq data - Total UMIs



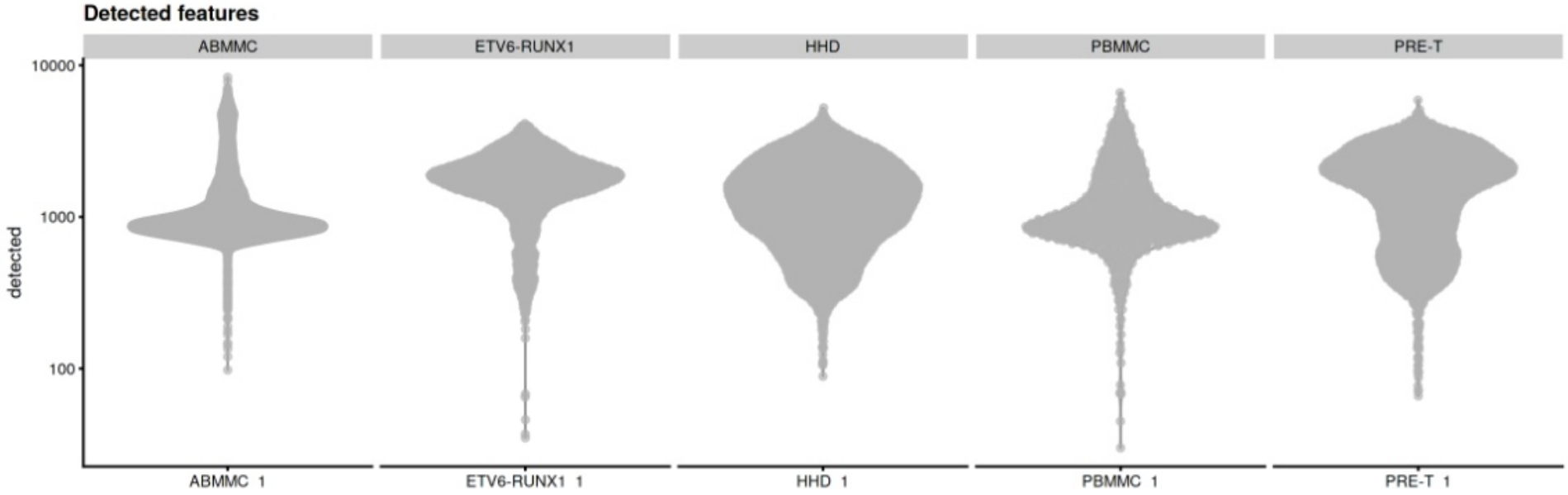
Properties of RNAseq data - Distribution of counts for a gene across cells



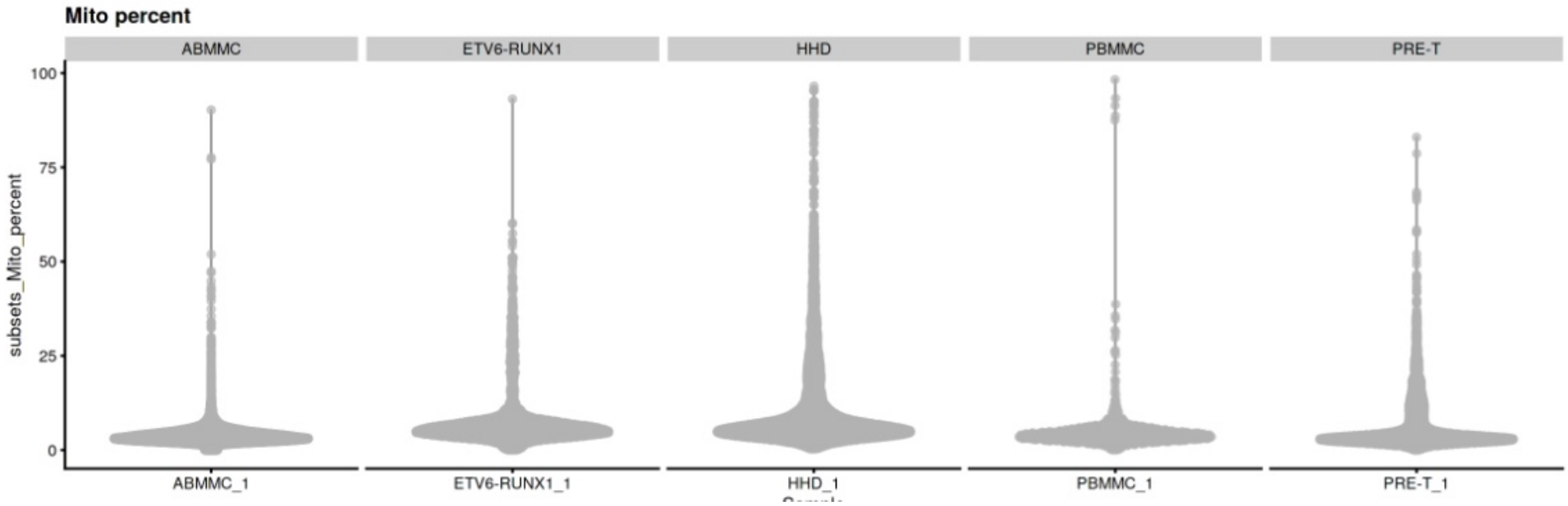
Properties of RNAseq data - Distribution of UMI counts



Properties of RNAseq data - Distribution of genes per cell



Properties of RNAseq data - Distribution of mitochondrial genes



Challenges

- Selecting appropriate thresholds for filtering, so that high quality cells are kept without removing biologically relevant cell types
  - Differentiating poor quality cells from less complex ones
  - Differentiating transcriptionally active cell types from multiplets/doublets
  - Distinguishing dead cells from those cells that express a high proportion of mitochorial genome

Recommendations

- Ensure that you know what types of cells you expect to be present before performing the QC.
- Are you expecting to find low complexity cells in your sample or cells with higher levels of mitochondrial expression?
- **When assessing the quality of our data, we must take this biology into consideration**