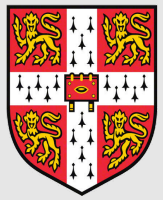


UNIVERSITY OF CAMBRIDGE

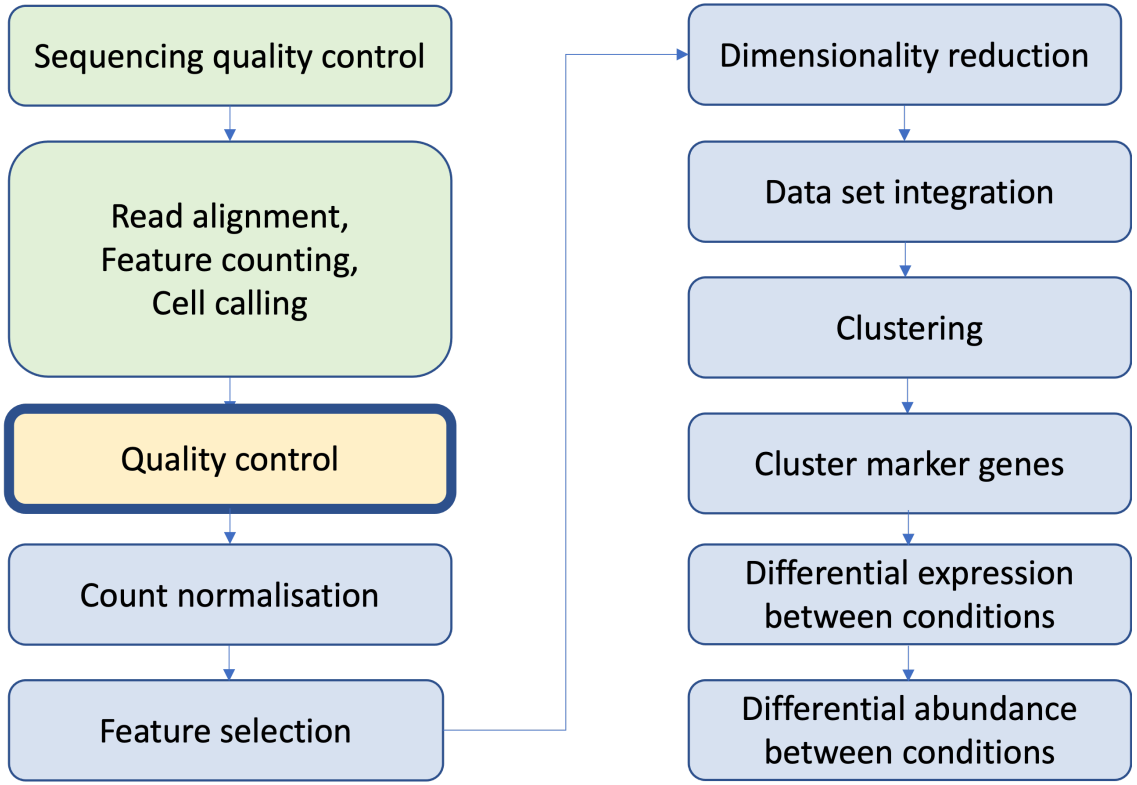
CANCER RESEARCH UK | Cambridge Institute

Introduction to single-cell RNA-seq analysis

Quality Control

6th January 2026

Single Cell RNAseq Analysis Workflow



10x overview

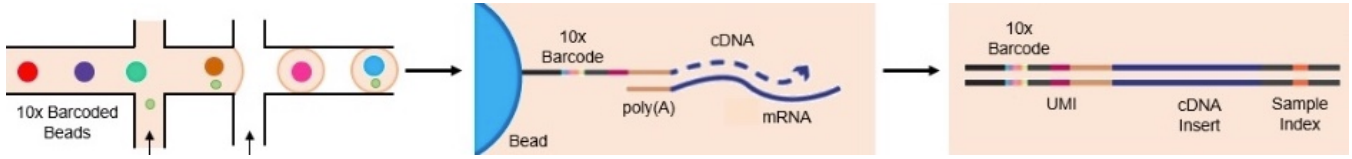


Image source: https://web.genewiz.com/single-cell-faq

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

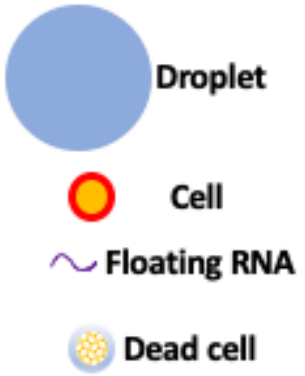
- Low complex transcriptome
- Genes detected much lower than average genes per cell

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

Seurat (v5) *Hao, Y., Stuart, T., Kowalski, M.H. et al.*

Take care of version differences!

https://satijalab.org/seurat/

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

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

Single-Cell Best Practices *Heumos, L., Schaar, A.C., Lance, C. et al.*

https://www.sc-best-practices.org/preamble.html

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

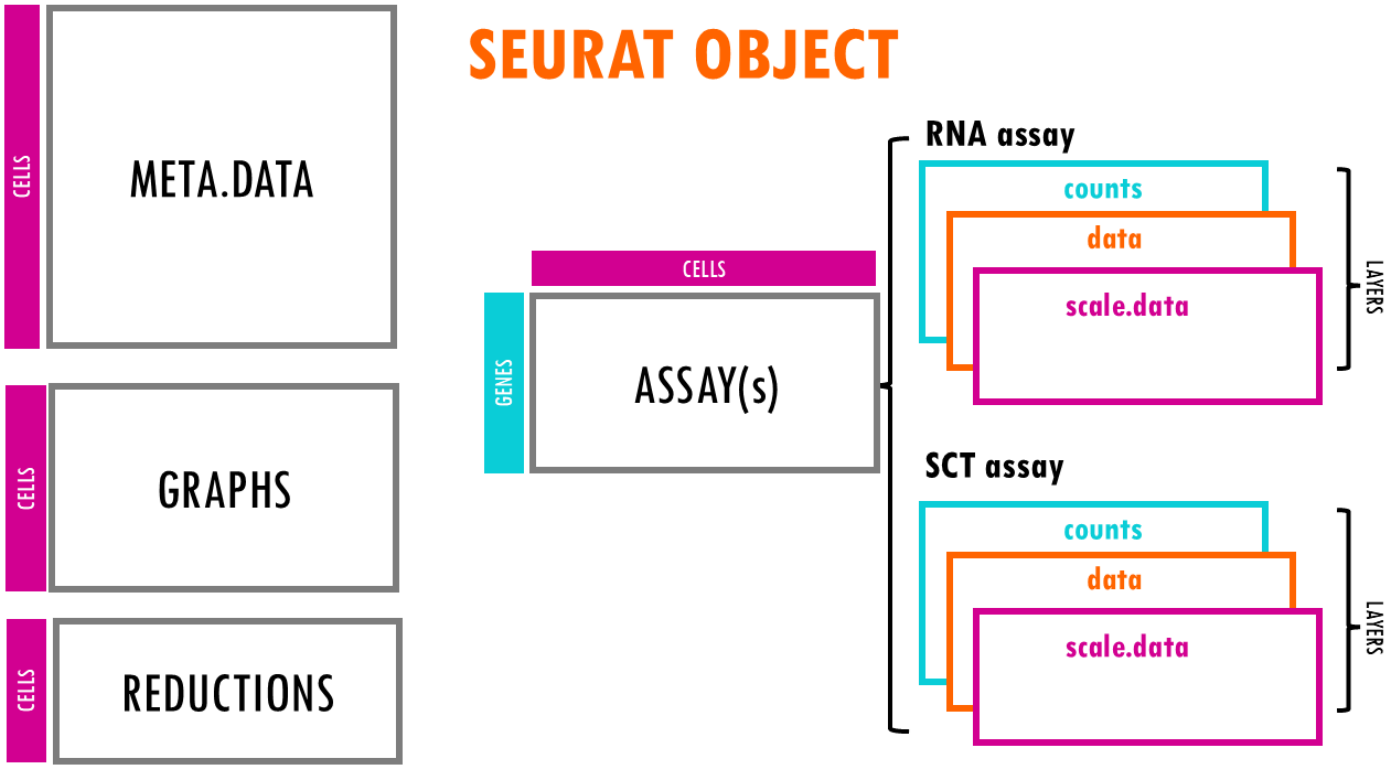
```
%h%-$ ls SRR9264343/outs/raw_feature_bc_matrix
barcodes.tsv.gz
features.tsv.gz
matrix.mtx.gz
```

Single Cell Experiment Vocabulary alert

- cell = Barcode = droplet
- Transcript = UMI

The *Seurat* object

https://biostatsquid.com/seurat-objects-explained/



The Assays Slot

assays:

This slot stores the raw and processed data in different forms. It is a list of Assay objects, each representing a specific type of data.

Examples: RNA: The most commonly used assay, containing the raw and processed RNA counts. SCT: Stores data processed using SCTransform (a normalization method). integrated: Contains integrated data when datasets have been merged. Each assay can contain matrices like counts, data, and scale.data.

Unlike the bioconductor *SingleCellExperiment* Object there is not a specific place to store feature data (gene annotations etc) in the Seurat object. This information is typically stored within the assay itself or in the metadata associated with the object as a whole.

The Metadata slot

meta.data:

A data.frame containing metadata associated with each cell. This can include cell type annotations, experimental conditions, or other variables related to the cells.

Example columns: cell_type, batch, condition, cluster.

The Reductions Slot

reductions:

A list of dimensionality reductions applied to the data. These are used for visualizations like PCA, t-SNE, or UMAP.

Examples: pca, tsne, umap. Each reduction stores a dimensionality reduction object, which contains information about the reduced coordinates (e.g., UMAP coordinates) and additional metadata like the variance explained by principal components.

The Other Slots

graphs: A list of graphs (usually a nearest-neighbor graph) that are used for clustering and other analyses. The graph typically represents relationships between cells based on gene expression similarity.

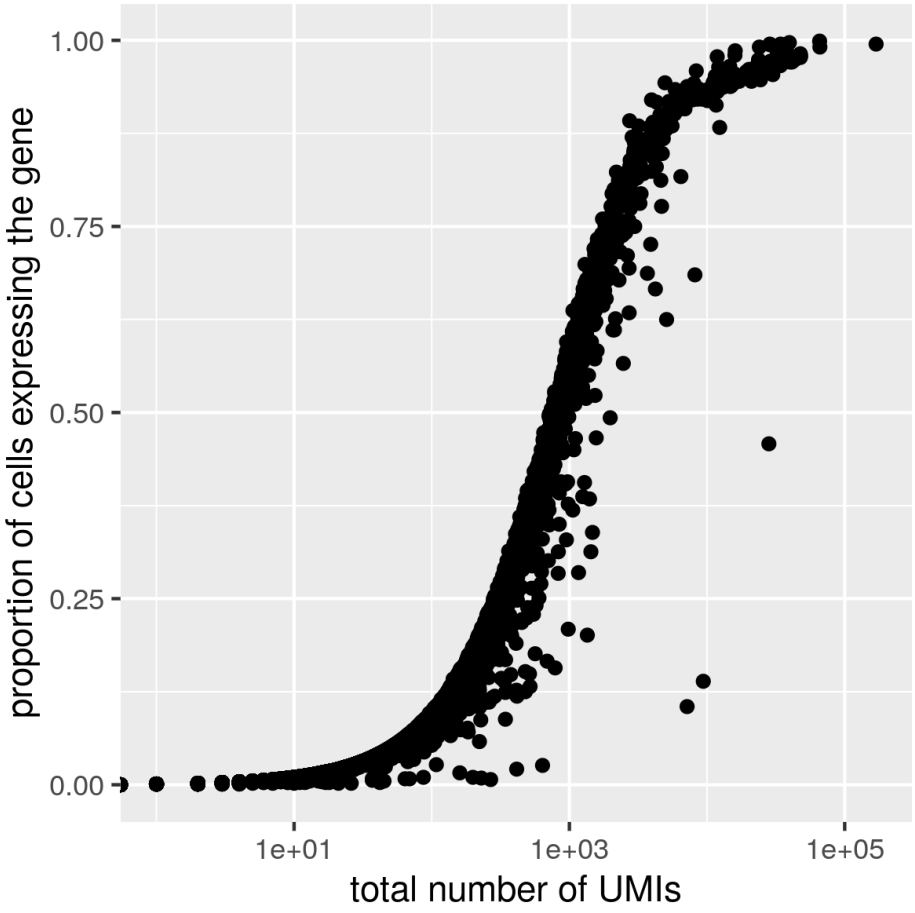
Examples: RNA_snn (a shared nearest-neighbor graph for RNA-seq data), pca_snn.

clusters: This stores the cluster assignments for each cell after a clustering analysis (e.g., Louvain or Leiden clustering). It is typically stored in the meta.data slot but can also be stored in a separate slot.

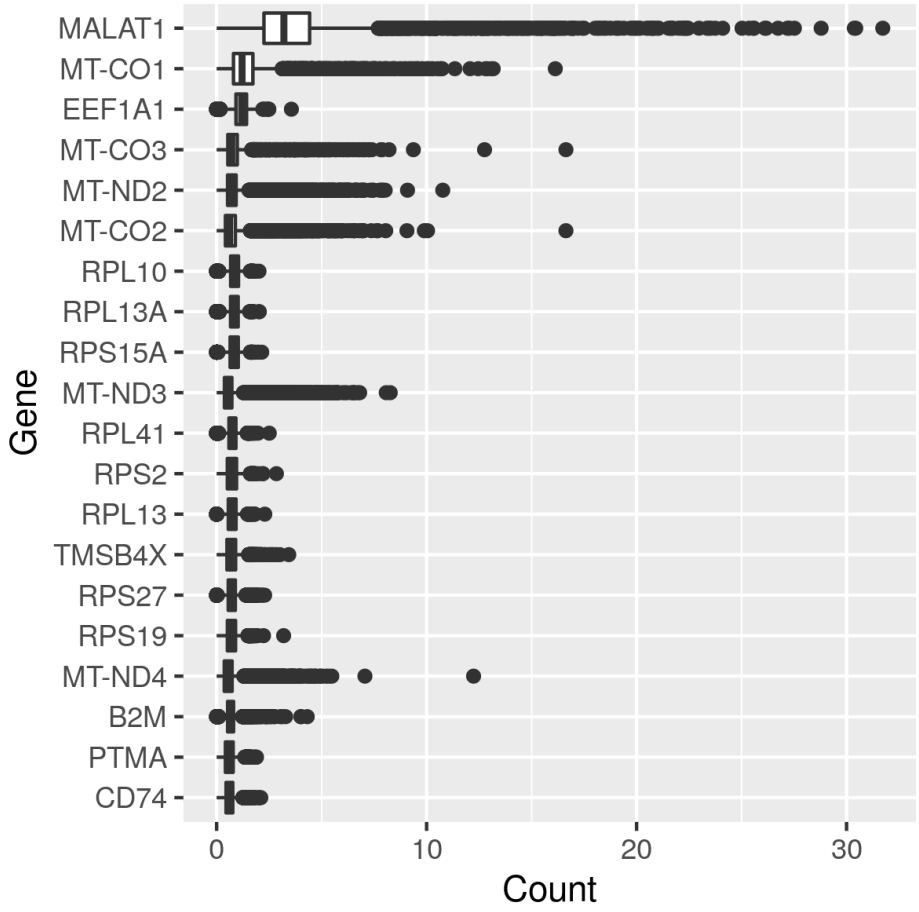
commands: A record of the commands used to generate the object. This can help in reproducibility by storing a log of the methods and operations that were applied to create or manipulate the Seurat object.

misc: This slot is used to store arbitrary information that doesn't fit into the other slots. It can be used to store additional analysis results or custom annotations.

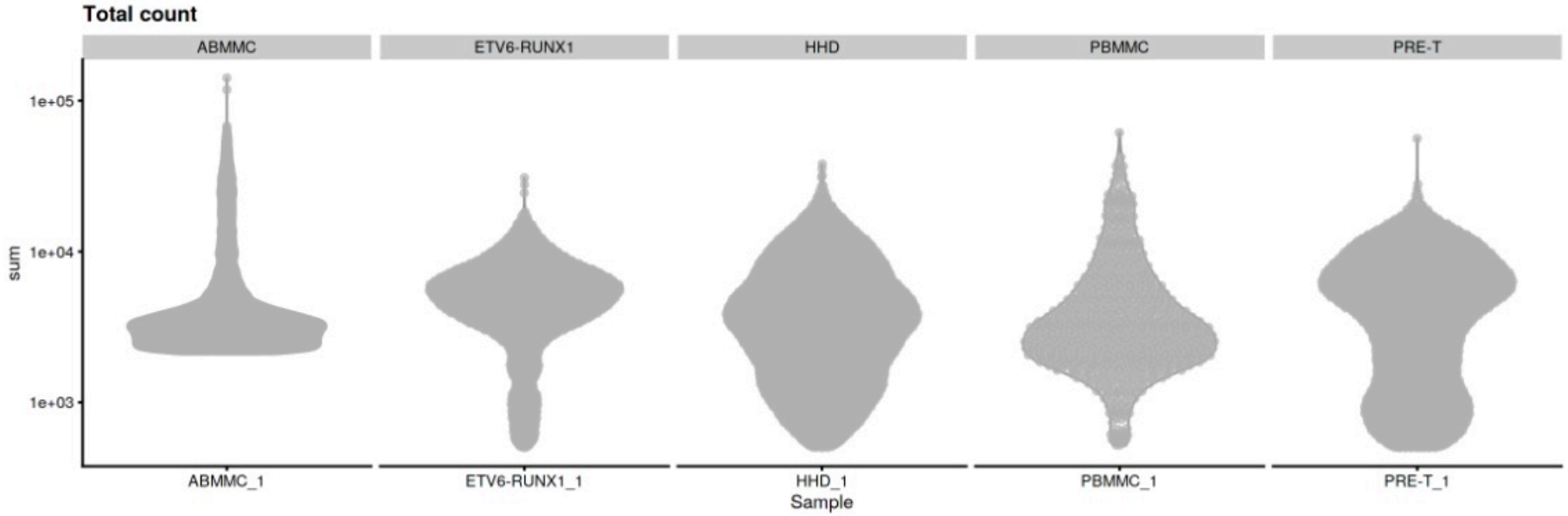
Properties of RNAseq data - Total UMIs



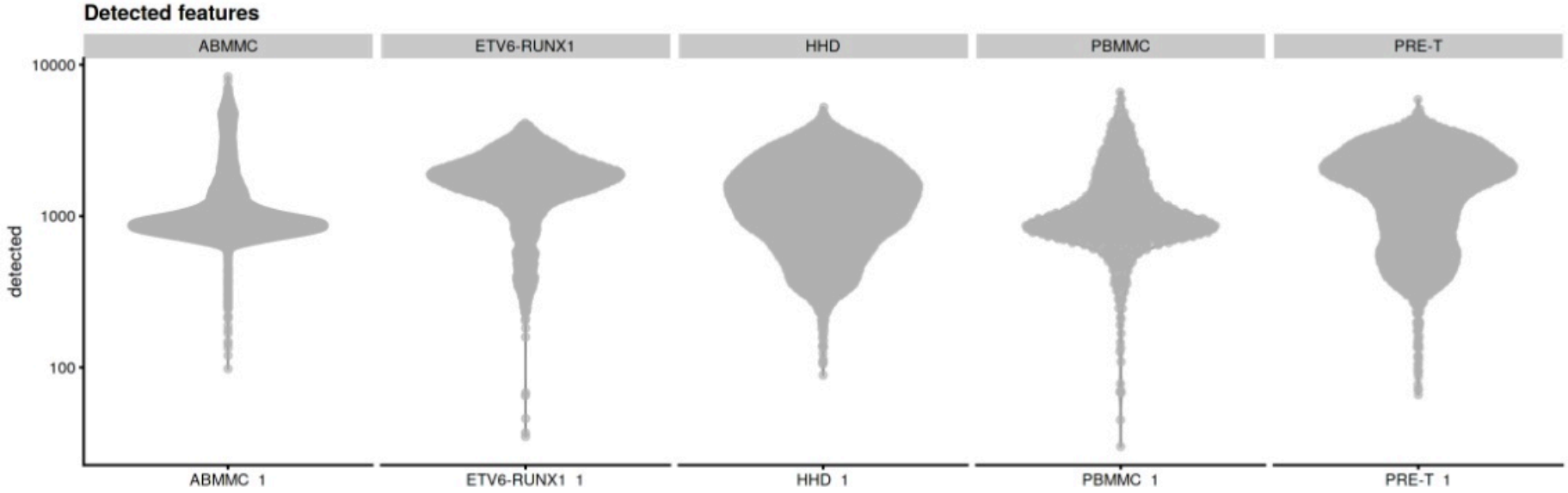
Properties of RNAseq data - Distribution of counts for a gene across cells



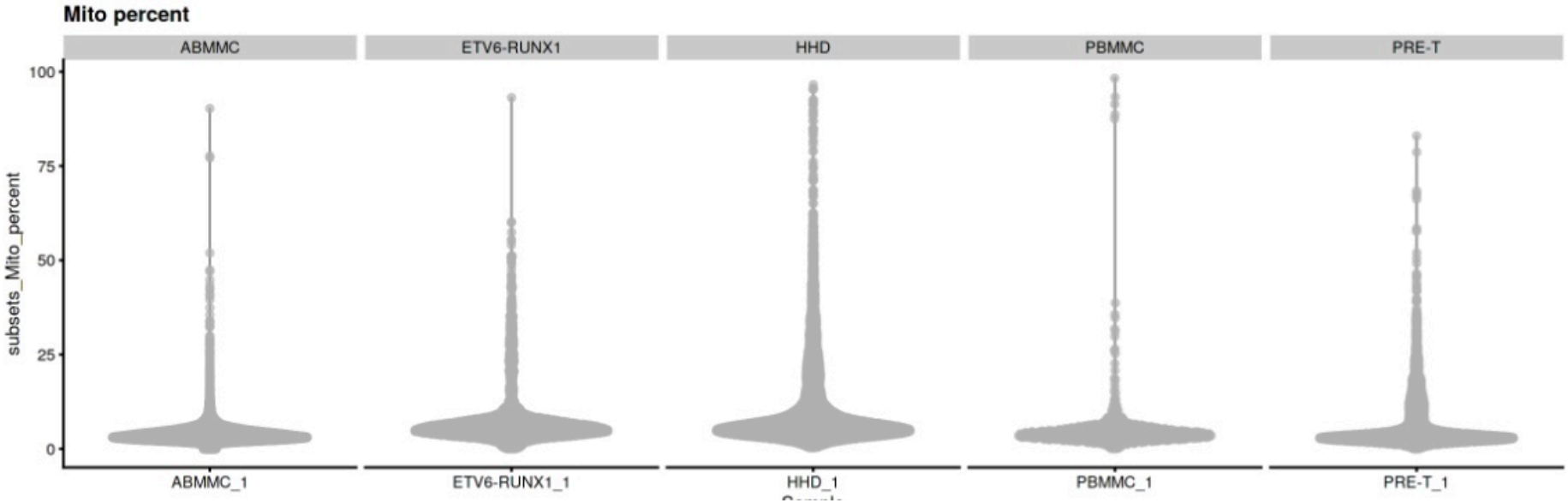
Properties of RNAseq data - Distribution of UMI counts



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**