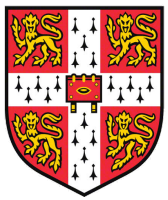


UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Introduction to single-cell RNA-seq analysis - Normalisation

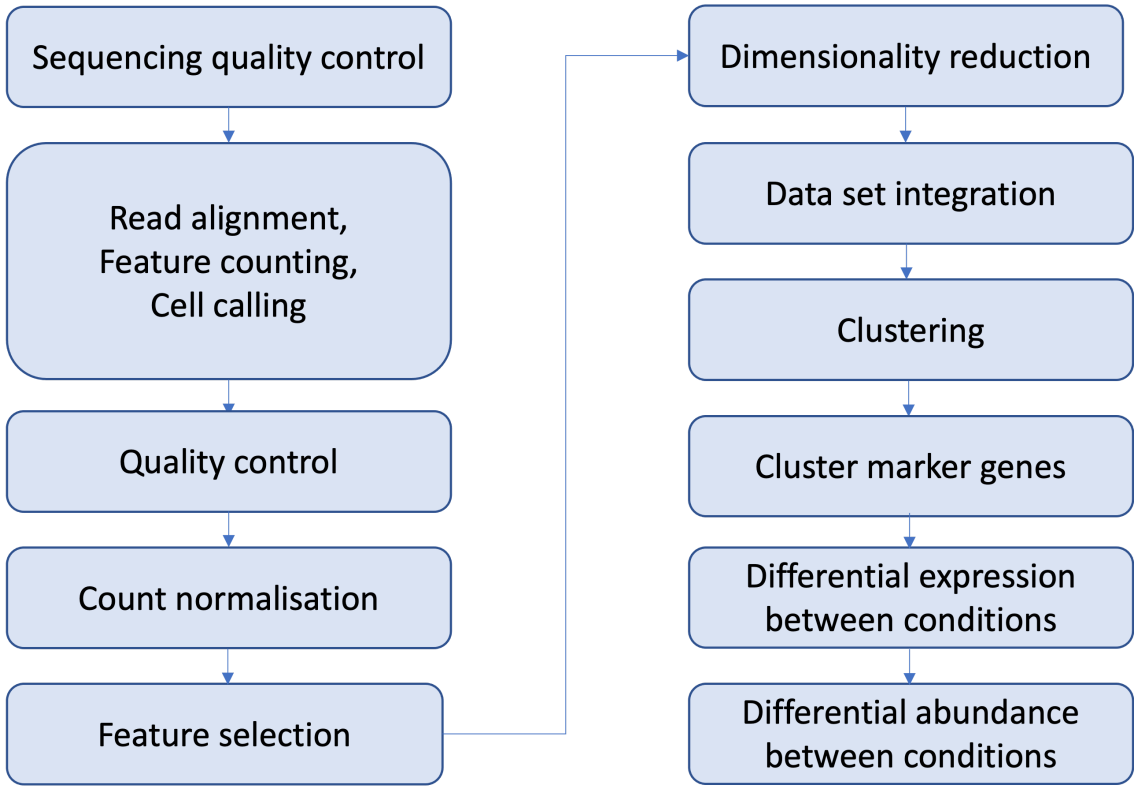
Chandra Chilamakuri and Stephane Ballereau and Adam Reid

25/01/2023

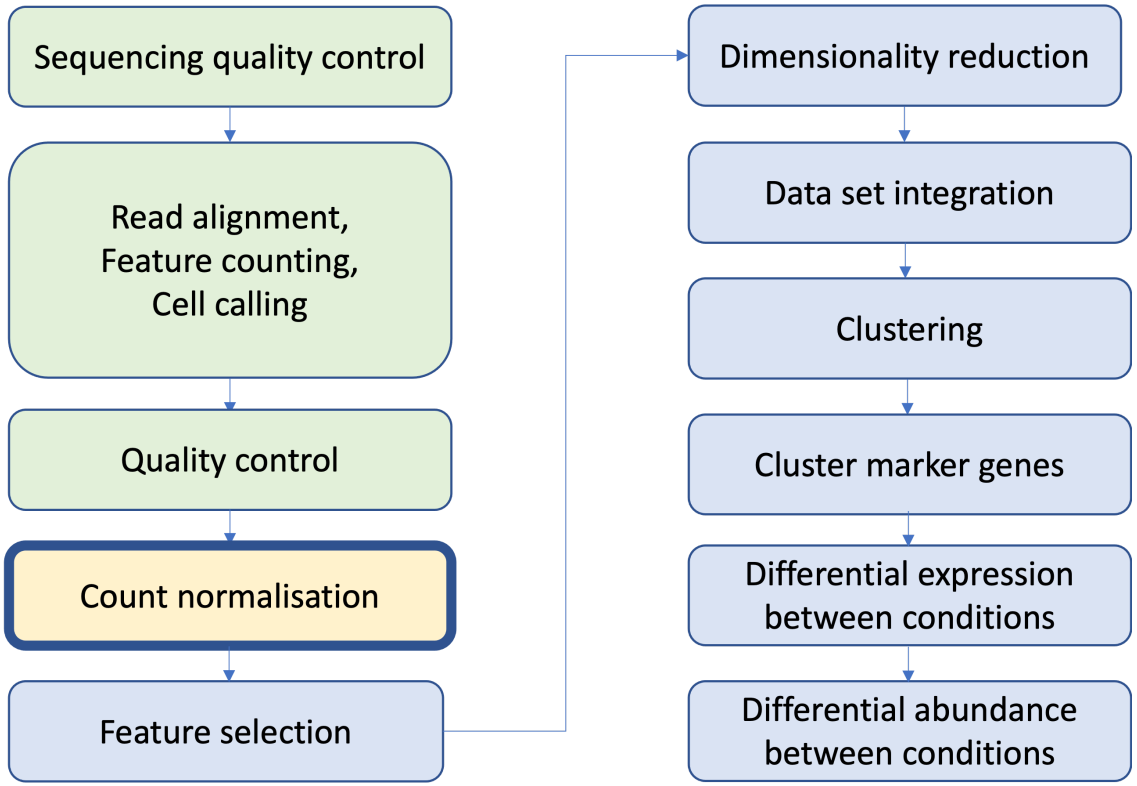
Outline

- Motivation
- Biases
  - Depth bias
  - Composition bias
  - Mean-variance correlation
- Normalisation strategies
- Deconvolution

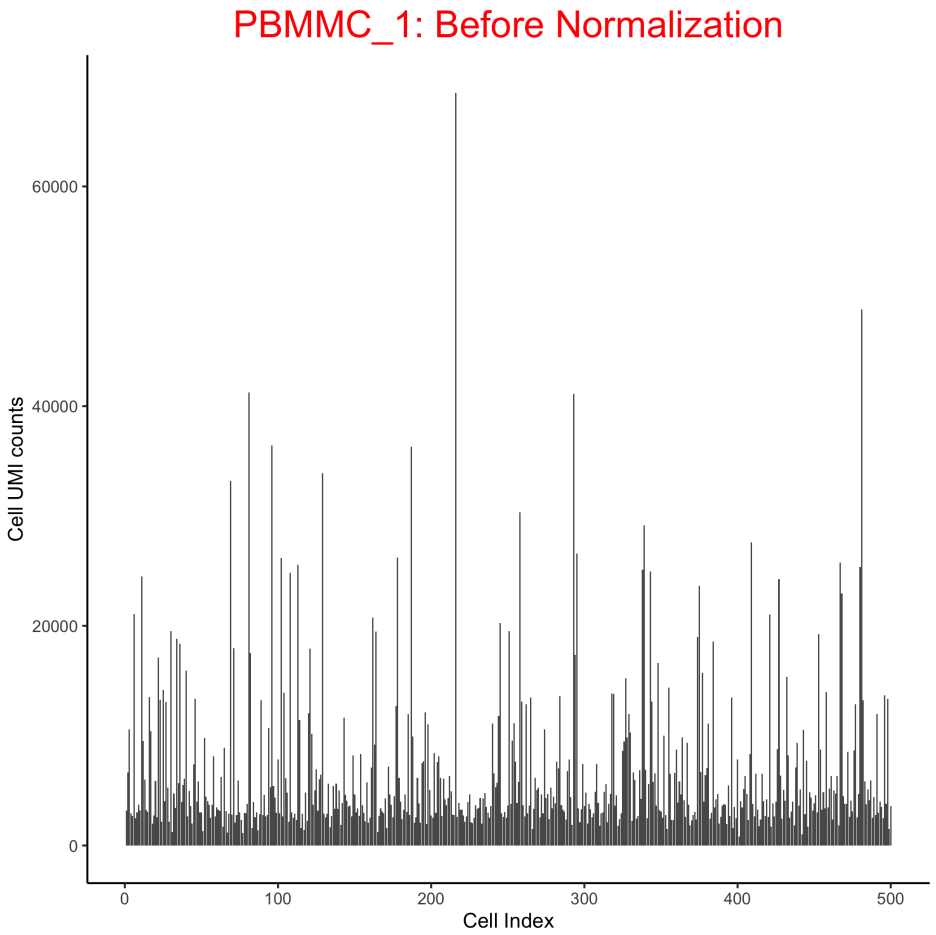
Workflow


Sequencing quality control
Read alignment,
Feature counting,
Cell calling
Quality control
Count normalisation
Feature selection
Dimensionality reduction
Data set integration
Clustering
Cluster marker genes
Differential expression
between conditions
Differential abundance
between conditions

Workflow


Sequencing quality control
Read alignment,
Feature counting,
Cell calling
Quality control
Count normalisation
Feature selection
Dimensionality reduction
Data set integration
Clustering
Cluster marker genes
Differential expression
between conditions
Differential abundance
between conditions

Raw UMI counts distribution



Why do UMI counts differ among the cells?

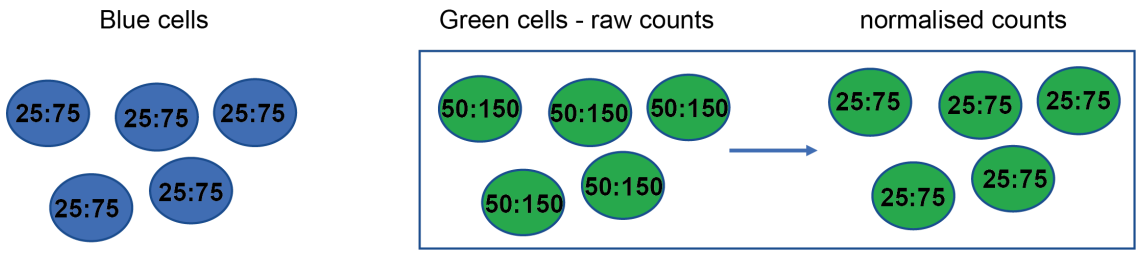
- We derive biological insights downstream by comparing cells against each other.
- But the UMI count differences makes it harder to compare cells.
- Why do total transcript molecules (UMI counts) detected between cells differ?
  - Biological:
    - Cell subtype differences - size and transcriptional activity, variation in gene expression
  - Technical: scRNA data is inherently noisy
    - Low mRNA content per cell
    - cell-to-cell differences in mRNA capture efficiency
    - Variable sequencing depth
    - PCR amplification efficiency

Normalization reduces technical differences so that differences between cells are not technical but biological, allowing meaningful comparison of expression profiles between cells.

Depth bias

Consider two genes A:B, in two cells types, blue and green.

We normalize here by dividing UMI counts for each gene by the total UMI counts in a cell and multiplying by 100.

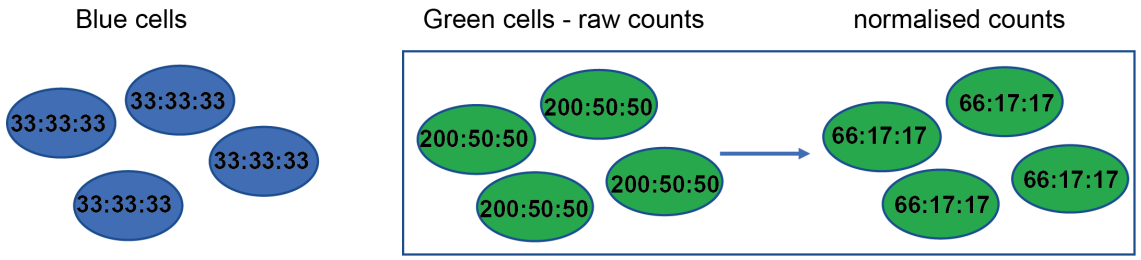


There is no differential expression, we have just sequenced twice as much in the second cell type.

Simple library size normalization accounts for the depth bias

Composition bias

Consider three genes A:B:C, in two cell types.



Just one gene is DE but library size normalization makes all look differentially expressed after normalisation

The deconvolution approach will we use takes account of both depth and compositions biases

Mean-variance correlation

Mean and variance of raw counts for genes are correlated

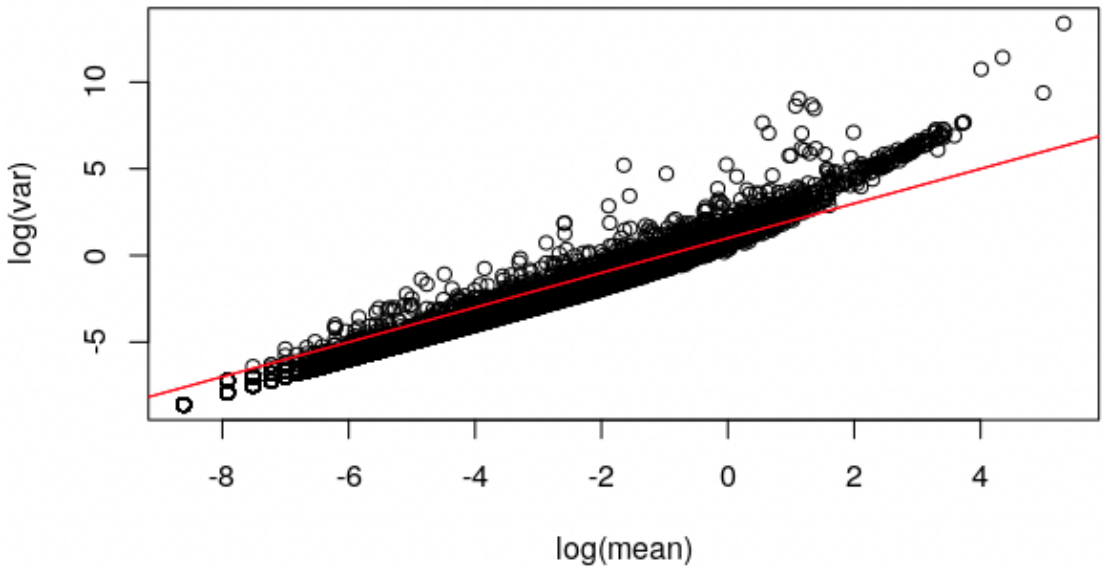
More highly expressed genes tend to look more variable because larger numbers result in higher variance

A gene expressed at a low level tends to have a low variance across cells:

var(c(2,4,2,4,2,4,2,4)) = 1.14

A gene with the same proportional differences between cells, but expressed at a higher level will have higher variance:

var(c(20,40,20,40,20,40,20,40)) = 114.29



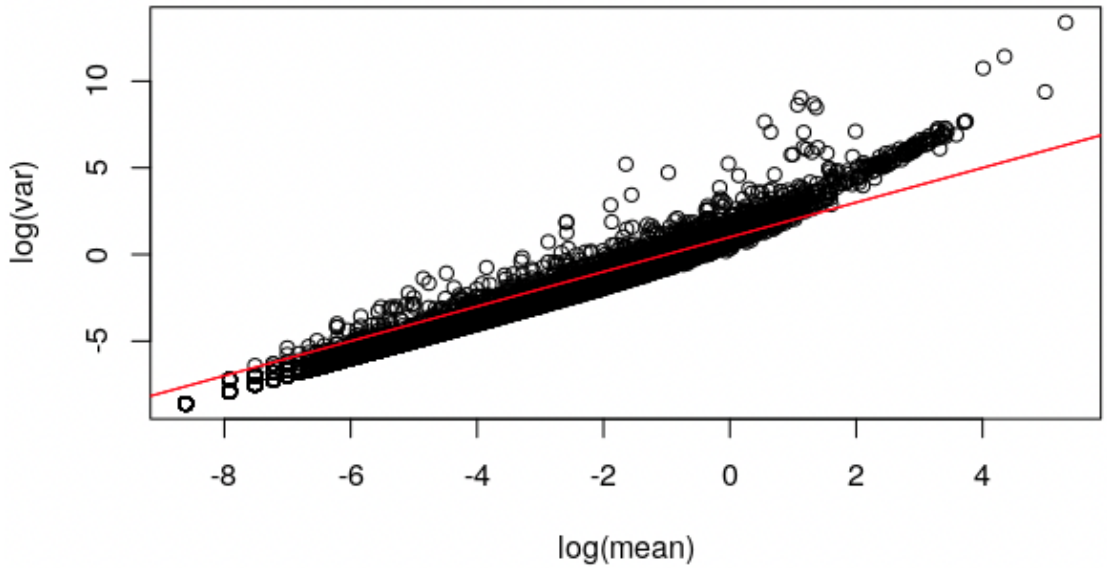
Mean-variance correlation

If we take the logs of the expression values, the variances are the same for both genes:

var(log(c(2,4,2,4,2,4,2,4))) = 0.14

var(log(c(20,40,20,40,20,40,20,40))) = 0.14

This “variable stabilising transformation” helps to remove the correlation between mean and variance



General principle behind normalisation

Normalization has two steps

1. Scaling
   - Calculate size factors or normalization factors that represents the relative depth bias in each cell
   - Scale the counts for each gene in each cell by dividing the raw counts with cell specific size factor
2. Transformation: Transform the data after scaling
   - Per million (e.g. CPM)
   - log2 (e.g. Deconvolution)
   - Pearson residuals (eg. sctransform)

Bulk RNAseq methods are not suitable for scRNAseq data

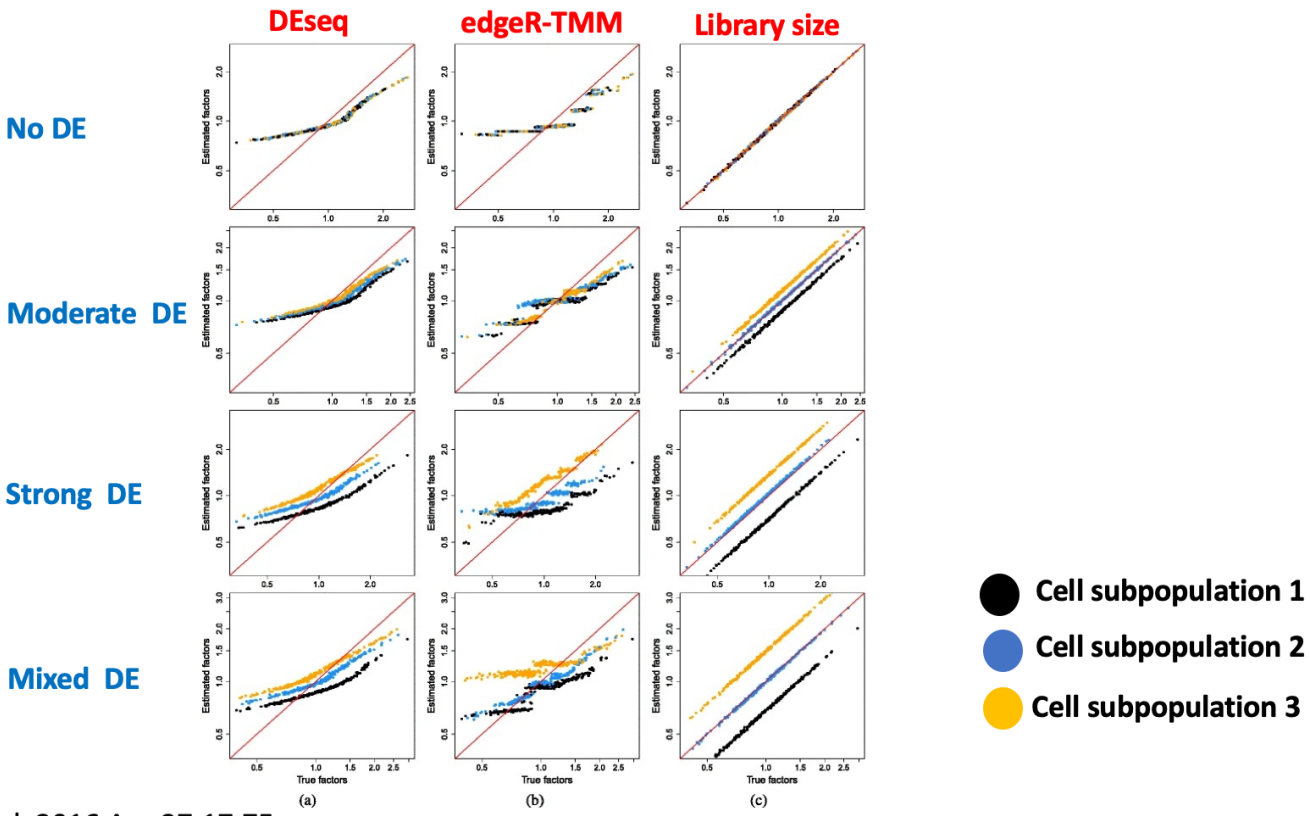
CPM: convert raw counts to counts-per-million (CPM)

- for each cell
- by dividing counts by the library size then multiplying by 1.000.000.
- does not address compositional bias caused by highly expressed genes that are also differentially expressed between cells.

DESeq's size factor

- For each gene, compute geometric mean across cells
- For each cell
  - compute for each gene the ratio of its expression to its geometric mean,
  - derive the cell's size factor as the median ratio across genes.
- Not suitable for sparse scRNA-seq data as the geometric mean is computed on non-zero values only.

Bulk RNA-seq normalization methods fail for scRNA-seq data



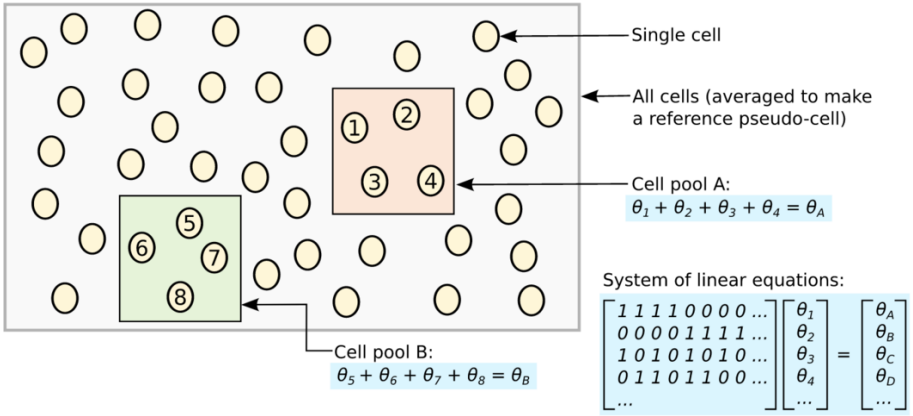
Lun AT *et al*. Genome Biol. 2016 Apr 27;17:75.

Deconvolution

Deconvolution strategy Lun et al 2016:

The deconvolution method consists of several key steps:

- Defining a pool of cells
- Summing expression values across all cells in the pool
- Normalizing the cell pool against an average reference, using the summed expression values
- Repeating this for many different pools of cells to construct a linear system
- Deconvolving the pool-based size factors to their cell-based counterparts (Fig. 3)



Steps:

- compute scaling factors by pooling cells
- apply scaling factors to get scaled data
- log2 transform the data

Recap

- We get different total counts for each cell due to technical factors (depth bias)
- A simplistic library size normalisation (e.g. CPM) removes a large part of this bias
- However, composition bias causes spurious differences between cells
- Early methods developed for bulk RNA-seq are not appropriate for sparse scRNA- seq data.
- The deconvolution method draws information from pools of cells to derive cell- based scaling factors that account for composition bias in scRNA-seq data.

In the demonstration and exercises we will see the effect of deconvolution on the data.