Preprocessing

Normalization and scaling

Bastien Job, Gustave Roussy, Villejuif

Nathalie Lehmann, ADLIN

École de bioinformatique AVIESAN-IFB-INSERM 2023

Organisation of the scRNA-seq course

- From raw count matrix to normalised matrix
  - Filtering low quality droplets
  - Filtering dead cells
  - Filtering doublets
- Data normalization
  - Why do we need to normalize the data ?
  - What are the methods available ?
  - Regression of biological biaises

Why do we need to normalize our data ?

We need to remove **technical biases** in order to…

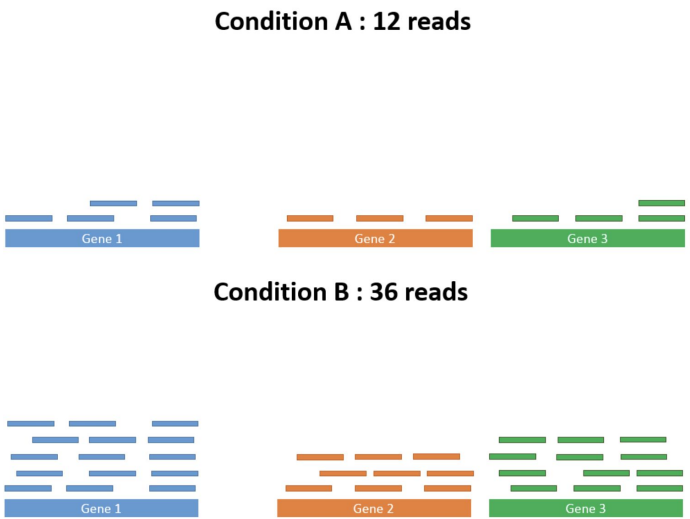
- To be able to compare **cells**

Images from https://www.biostars.org/p/349881/

Why do we need to normalize our data ?

We need to remove **technical biases** in order to…

- To be able to compare **cells**

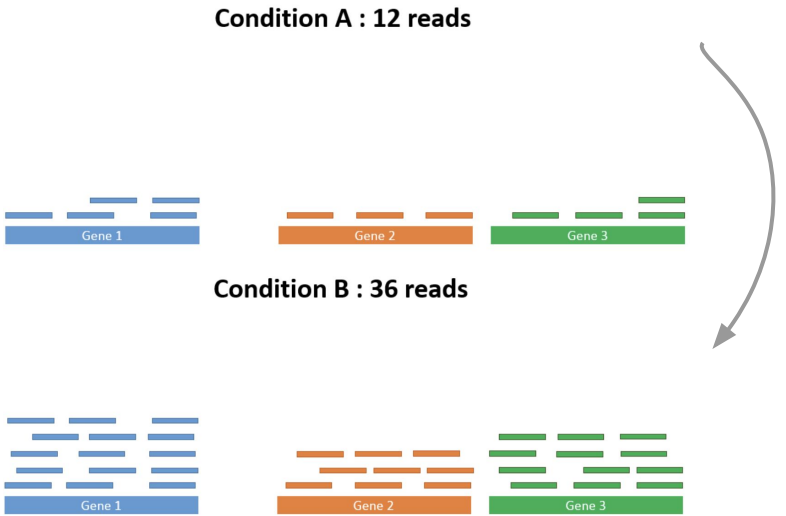


Images from https://www.biostars.org/p/349881/

Why do we need to normalize our data ?

We need to remove **technical biases** in order to…

- To be able to compare **cells**



The 2 libraries have the **same RNA composition**.

But the condition B has 3 times more reads than the condition A.

We need to correct for differences in **library size**.

Images from https://www.biostars.org/p/349881/

Why do we need to normalize our data ?

We need to remove **technical biases** in order to…

- To be able to compare **cells**

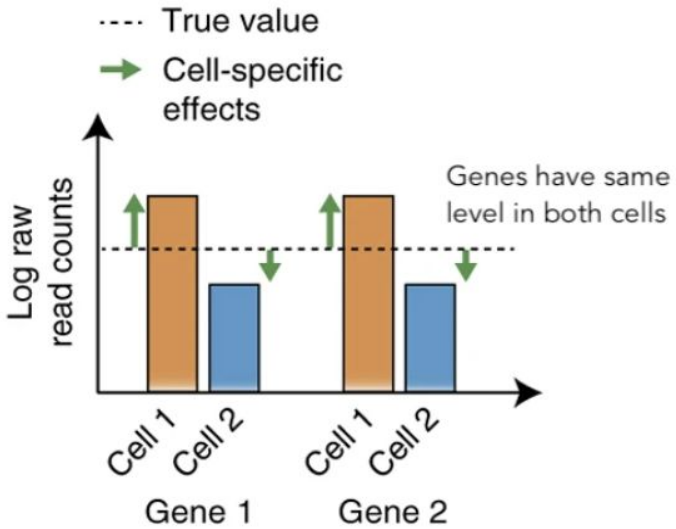


Image adapted from *Vallejos et al., 2017*

Why do we need to normalize our data ?

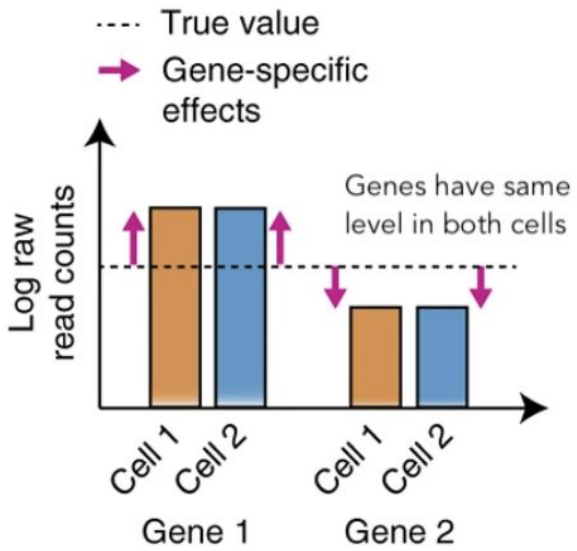
We need to remove **technical biases** in order to…

- To be able to compare **genes**

Why do we need to normalize our data ?

We need to remove **technical biases** in order to…

- To be able to compare **genes**



Examples of biological biases that you may want to correct :

- Amplification
- RNA capture efficiency
- Gene length
- GC content

Image adapted from *Vallejos et al., 2017*

Plenty of normalization approaches for bulk RNA-seq

- TPM
- CPM

- RPKM
- FPKM

- Global scaling (eg: Upper Quartile)

- Size factors calculation (eg: estimation of library sampling depth) :
  - DESeq2
  - edgeR

- ...

Plenty of normalization approaches for bulk RNA-seq

- TPM
- CPM

- RPKM
- FPKM

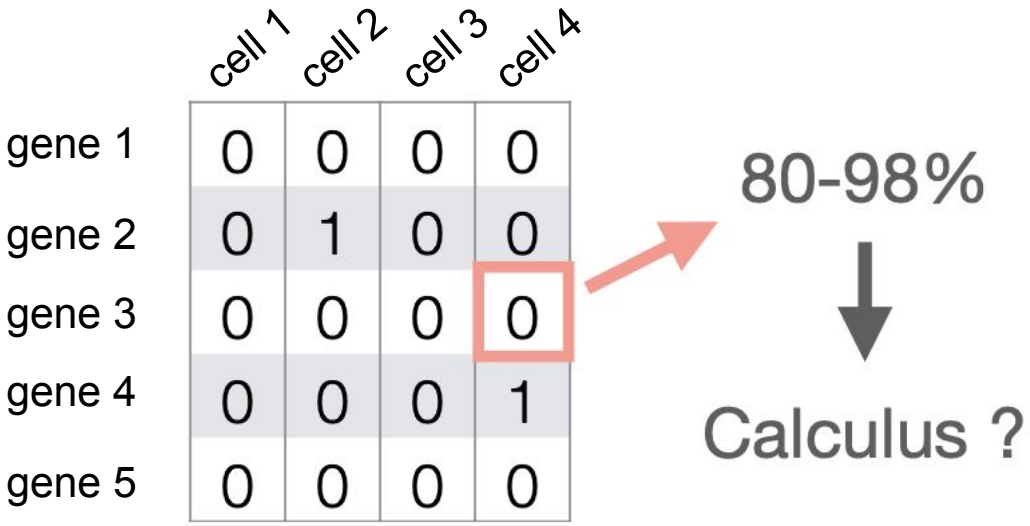
These methods do not apply to single-cell data (or partially)

- Global scaling (eg: Upper Quartile)

- Size factors calculation (eg: estimation of library sampling depth) :
  - DESeq2
  - edgeR

- ...

This is mostly due to the sparsity of the single-cell data



A **sparse matrix** is a matrix filled with a LOT of zeros

Solutions to normalize single-cell data

- Rough solution : **global log-normalization** / Z-scoring

- **Scaling by factors** :
  a. Cells which expression profiles behave very similarly are put together in small groups (pools)
  b. This reduces the number of zeros
  c. We are able to estimate the normalization factor of each pool.
  d. The same operation is repeated many times, slightly changing the groups
  e. In the end we are able to estimate the normalization factor of each cell.

- *Variance stabilization (**sctransform** in Seurat) : clever, but complicated*

Acknowledgements

- Some illustrations/slide were created by Marine Aglave