SC-RNAseq Data Structure and Basic Quality Control

Julieth Andrea López Castiblanco
PhD student in Biotechnology
julalopezcas@unal.edu.co

UNIVERSIDAD NACIONAL DE COLOMBIA

icmt Instituto Colombiano de Medicina Tropical

UNIVERSIDAD DE ANTIOQUIA 1803

Data sources

Repository	Link to resource
NCBI	https://www.ncbi.nlm.nih.gov/
Human Cell Atlas: Data Explorer	https://explore.data.humancellatlas.org/projects
CellXGene Collection	https://cellxgene.cziscience.com/datasets
Single Cell Portal	https://singlecell.broadinstitute.org/single_cell
EBI Single Cell Expression Atlas	https://www.ebi.ac.uk/gxa/sc/home

Data Structure

Data tables

Count matrix

	Cell 1	Cell 2	...	Cell N
Gene 1	0	1	...	0
Gene 2	1	3	...	0
...	...	...	...	...
Gene M	2	2		4

Cells information

	Barcode	Donor	...	Treatment
Cell 1	ACTGTA	D1	...	Drug
Cell 2	TGCATA	D1	...	Control
...	...	...	...	...
Cell N	CCTATA	D6		Drug

Genes information

	ID	Symbol	...	Chromosome
Gene 1	ENSG00000155816	FMN2	...	1
Gene 2	ENSG00000229807	XIST	...	X
...	...	...	...	...
Gene M	ENSG00000139618	BRCA2		13

Log transformation

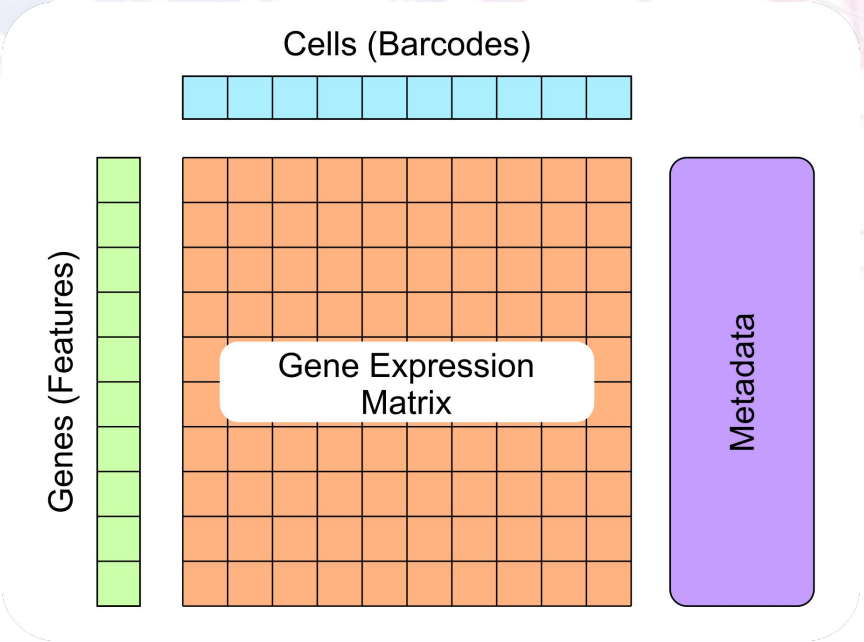
	Cell 1	Cell 2	...	Cell N
Gene 1	0	0.6	...	0
Gene 2	0.3	0.8	...	0
...	...	...	...	...
Gene M	0.35	0.67		2.1

How can we structurate this data?

Single-cell data

Single-cell data consists of 4 main components:

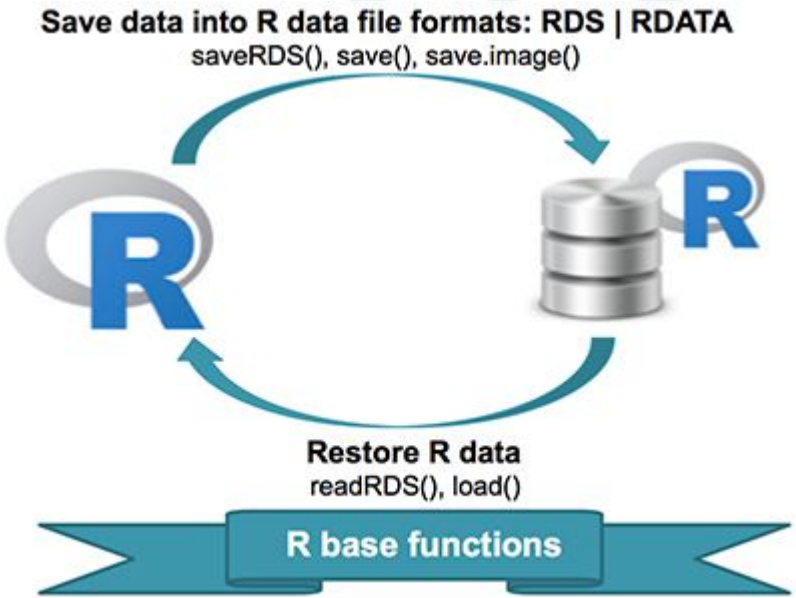
- ➔ Gene expression data
- ➔ Metadata about each genes
- ➔ Metadata about each cell
- ➔ Unstructured metadata about the data collected
  - ◆ Batch/replicate information
  - ◆ Sequencing platform
  - ◆ Data / time
  - ◆ Tissue source



Which tools can we use to give this structure to the data?

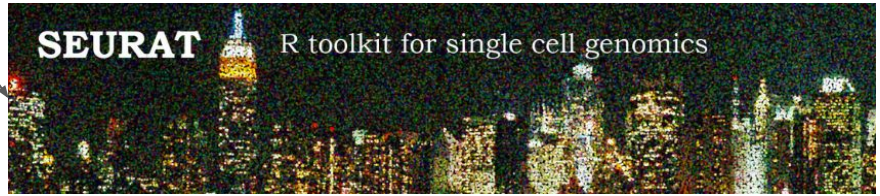
https://training.galaxyproject.org/training-material/topics/single-cell/tutorials/scrna-data-formats/slides-plain.html

Which R-based classes can we use?



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

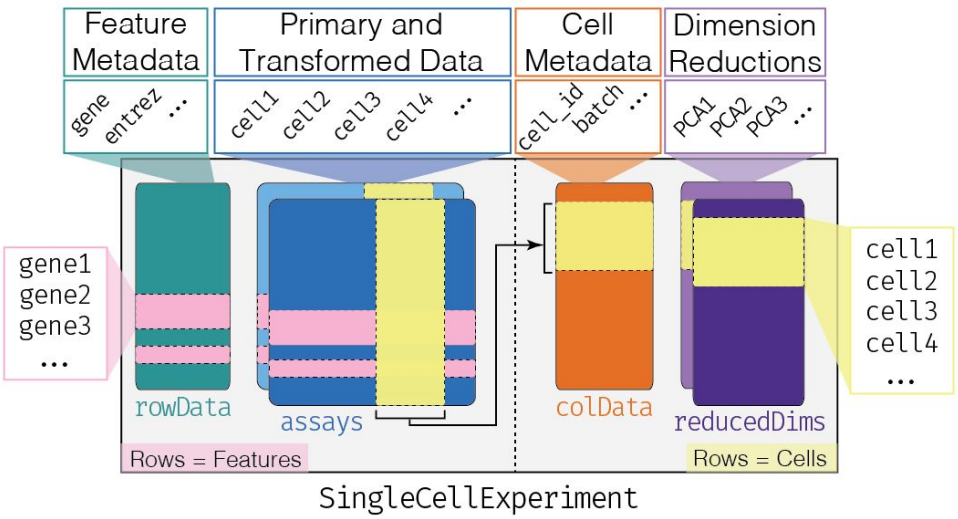
SingleCellExperiment



SingleCellExperiment Object (Bioconductor)

This **Bioconductor class** implements a data structure that stores all aspects of single-cell data and allow to manipulate them in a synchronized manner.

- ➔ Primary data
  - ◆ Count matrix
  - ◆ Transformed data
- ➔ Feature metadata
  - ◆ Transcript length
  - ◆ Gene symbol
- ➔ Cell metadata
- ➔ Dimension reduction
  - ◆ PCA, tSNE, etc
- ➔ Other study metadata
  - ◆ Batch/replicate information
  - ◆ Sequencing platform
  - ◆ Data / time
  - ◆ Tissue source



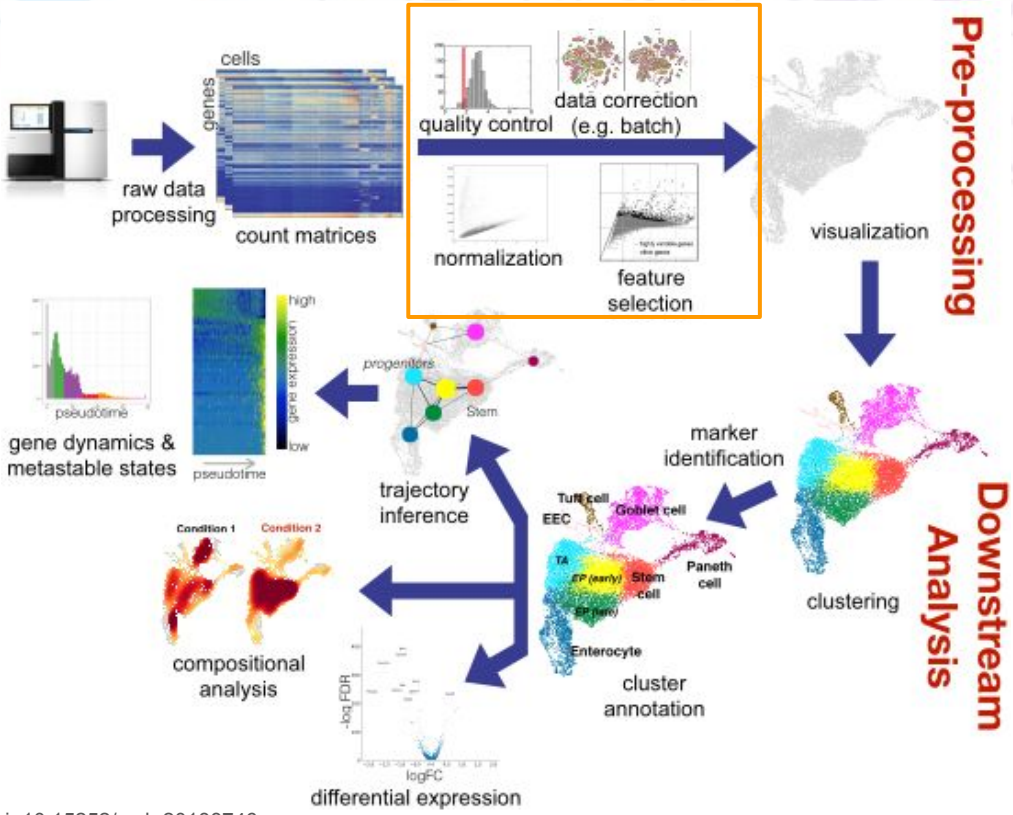
https://bioconductor.org/books/3.13/OSCA.intro/the-singlecellexperiment-class.html

Seurat Object

```
Formal class 'Seurat' [package "SeuratObject"] with 13 slots
  ..@ assays     :List of 1
  .. ..$ RNA:Formal class 'Assay' [package "SeuratObject"] with 8 slots
  .. .. .. ..@ counts                :Formal class 'dgCMatrix' [package "Matrix"] with 6 slots
  .. .. .. .. .. ..@ i               : int [1:2282976] 29 73 80 148 163 184 186 227 229 230 ...
  .. .. .. .. .. ..@ p               : int [1:2701] 0 779 2131 3260 4220 4741 5522 6304 7094 7626 ...
  .. .. .. .. .. ..@ Dim             : int [1:2] 13714 2700
  .. .. .. .. .. ..@ Dimnames:List of 2
  .. .. .. .. .. .. ..$ : chr [1:13714] "AL627309.1" "AP006222.2" "RP11-206L10.2" "RP11-206L10.9" ...
  .. .. .. .. .. .. ..$ : chr [1:2700] "AAACATACAACCAC-1" "AAACATTGAGCTAC-1" "AAACATTGATCAGC-1" "AAACCGTGCTTCCG-1" ...
  .. .. .. .. .. ..@ x               : num [1:2282976] 1 1 2 1 1 1 1 4 1 1 1 ...
  .. .. .. .. .. ..@ factors : list()
  .. .. .. ..@ meta.features:'data.frame':    13714 obs. of  0 variables
  .. .. .. ..@ misc                  : list()
  ..@ meta.data   :'data.frame':    2700 obs. of  3 variables:
  .. ..$ orig.ident  : Factor w/ 1 level "pbmc3k": 1 1 1 1 1 1 1 1 1 1 ...
  .. ..$ nCount_RNA  : num [1:2700] 2419 4903 3147 2639 980 ...
  .. ..$ nFeature_RNA: int [1:2700] 779 1352 1129 960 521 781 782 790 532 550 ...
  ..@ active.assay: chr "RNA"
  ..@ active.ident: Factor w/ 1 level "pbmc3k": 1 1 1 1 1 1 1 1 1 1 ...
  .. ..- attr(*, "names")= chr [1:2700] "AAACATACAACCAC-1" "AAACATTGAGCTAC-1" "AAACATTGATCAGC-1"
"AAACCGTGCTTCCG-1" ...
  ..@ graphs    : list()
  ..@ neighbors   : list()
  ..@ reductions  : list()
```

https://github.com/satijalab/seurat/wiki/Seurat - Object description

Now, we can start working with the data



https://github.com/theislab/single-cell-tutorial; doi: 10.15252/msb.20188746

Basic Quality Control

Motivation

There are cells featuring one or more of the next characteristics:

⚠️ Low total counts
⚠️ Few expressed genes
⚠️ High proportion of reads coming from mitochondria

What happen if we do not apply quality control? ❌

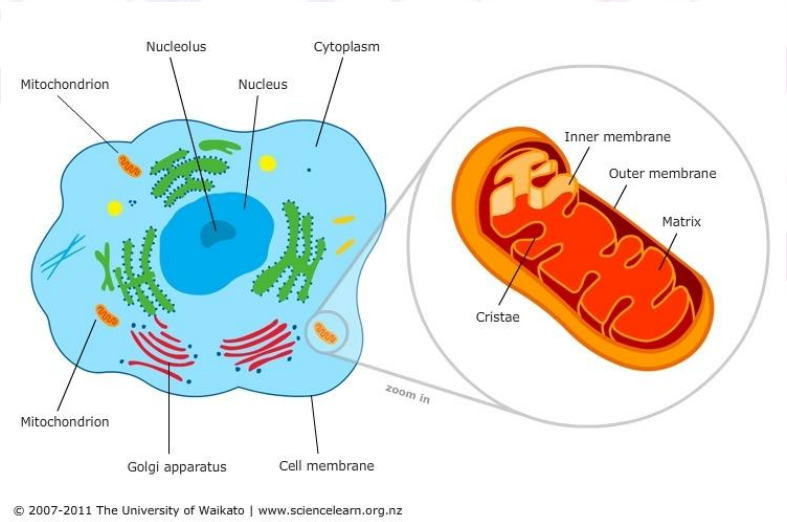
- Distinct cluster(s) **complicating interpretation** of results
- Distortion of **population heterogeneity**
- **Artificial** 'upregulation' of certain genes

Luecken et al., 2019

About mitochondrial content

Why having into account **mitochondrial RNA?**

- ➔ Due to very harsh conditions in tissue dissociation step, dying cells release their cytoplasmic contents
- ➔ High mitochondrial contamination means low-quality cells



What can we do?

➔ First metrics

- ◆ RNA count (or count depth, number of UMIs)
- ◆ Feature count (or gene count)
- ◆ Mitochondrial content

→

- Low-quality cells or empty droplets will often have very few genes
- Cell doublets or multiplets may exhibit an aberrantly high gene count

➔ Recommendations

- ◆ Identify and discard outliers
- ◆ Different samples may require different cutoffs

What do we achieve?

The consequences of applying quality control are:

- ➔ **Sufficient data quality** for downstream analysis
  - ◆ Cannot be determined a priori
  - ◆ Iterative quality control
- ➔ **Enhance interpretation**
  - ◆ Relevant for datasets containing heterogeneous cell populations (low-quality or outlier cells can be misinterpreted)
  - ◆ Results are reflective of **biological variability** rather than technical artifacts

Luecken et al., 2019

Quality control software options

Seurat 4.0.4 | Install | Get started | Vignettes | Extensions | FAQ | News | Reference | Archive

Official release of Seurat 4.0

Bioinformatics, 33(8), 2017, 1179–1186
doi: 10.1093/bioinformatics/btw777
Advance Access Publication Date: 14 January 2017
Original Paper

OXFORD

Gene expression

Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R

Davis J. McCarthy[1,2,3,4], Kieran R. Campbell[2,4], Aaron T. L. Lun[5] and Quin F. Wills[2,6]

Cell Systems
CellPress
Volume 8, Issue 4, 24 April 2019, Pages 329-337.e4

Brief Report

DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors

Ask Copilot: Save time, read 10X faster with AI

Save | Related Papers | Conclusions | Points Discussed | Evidence/Examples Used | Purpose | Biases or Limitations | Summarize

Christopher S. McGinnis [1], Lyndsay M. Murrow [1], Zev J. Gartner [1 2 3 4]

Wolf et al. Genome Biology (2018) 19:15
https://doi.org/10.1186/s13059-017-1382-0
Genome Biology

SOFTWARE | Open Access

SCANPY: large-scale single-cell gene expression data analysis

F. Alexander Wolf[1*], Philipp Angerer[1] and Fabian J. Theis[1,2*]

Thanks!

@JuliethLopz
julalopezcas@unal.edu.co
https://www.linkedin.com/in/julieth-andrea-lopez/

PROJECT JAGUAR

CHAN ZUCKERBERG INITIATIVE

HUMAN CELL ATLAS

UNIVERSIDAD NACIONAL DE COLOMBIA

icmt Instituto Colombiano de Medicina Tropical

UNIVERSIDAD DE ANTIOQUIA 1803

wellcome connecting science