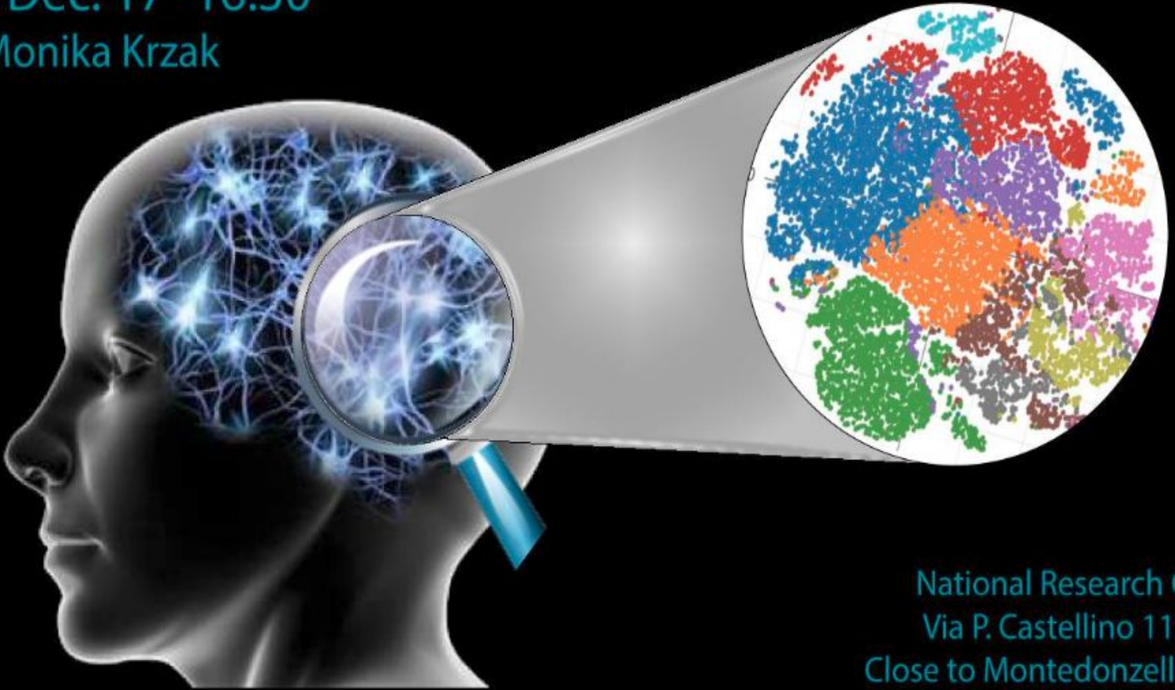
Single Cell Analysis Workflow

Monday, Dec. 17 -16:30

Presents: Monika Krzak



Contact me at:
m.krzak@na.iac.cnr.it

National Research Council - Library
Via P. Castellino 111, 80131, Napoli
Close to Montedonzelli Metro1 Station

Outline

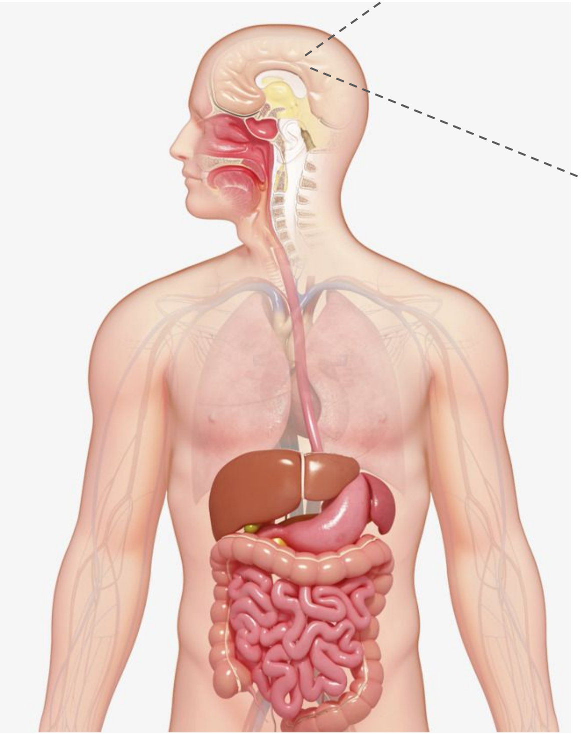
- **What is single-cell RNAseq (scRNAseq) ?**
  - **scRNAseq technology**
  - **ScRNAseq protocols and data types**
  - **Challenges in analyzing scRNAseq data**
  - **Preprocessing scRNAseq data - dealing with noise and dimensionality**
  - **ScRNAseq Applications**
- **Online Materials**
- **Let’s start !**
  - **SingleCellExperiment Object**
  - **Scater Package**
  - **CellDataSet Object**
  - **Monocle Package**

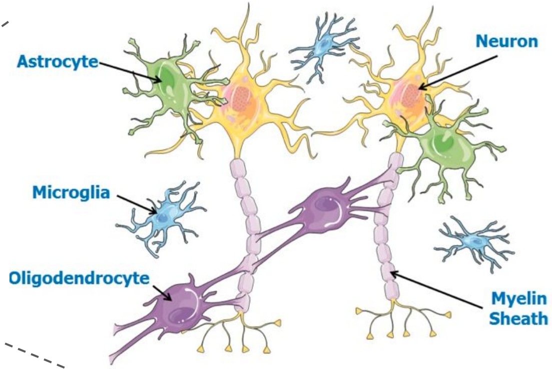
AIM

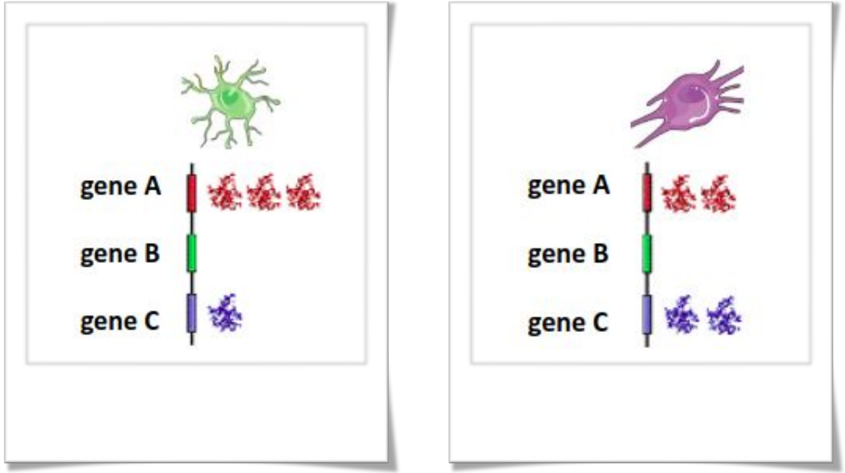
Useful tools and functions for analysis of scRNAseq data

NOT: Golden standard analysis pipeline

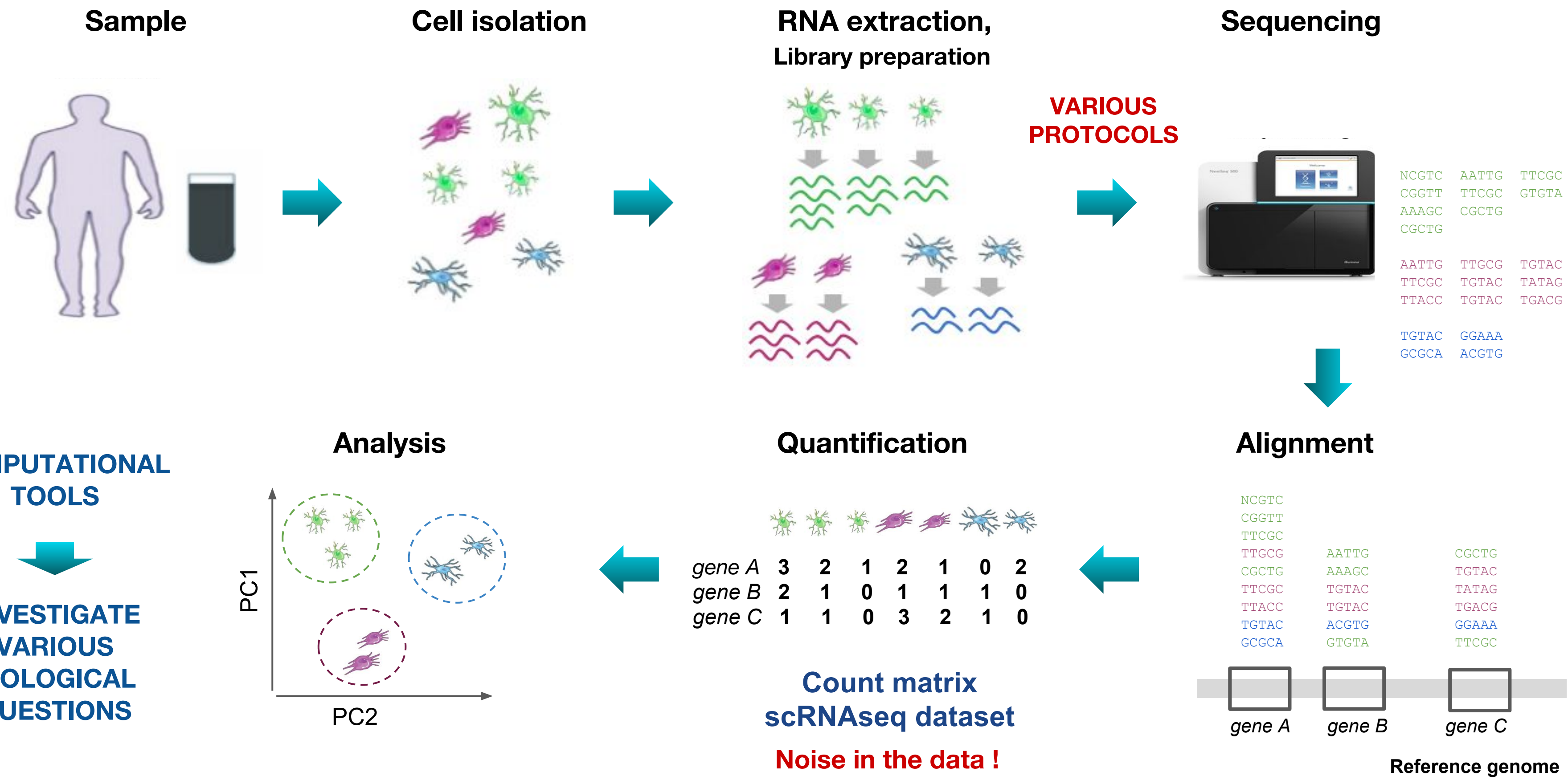
Motivation




Astrocyte
Neuron
Microglia
Oligodendrocyte
Myelin Sheath


gene A
gene B
gene C
gene A
gene B
gene C

Single-cell RNAseq



gene A	3	2	1	2	1	0	2
gene B	2	1	0	1	1	1	0
gene C	1	1	0	3	2	1	0

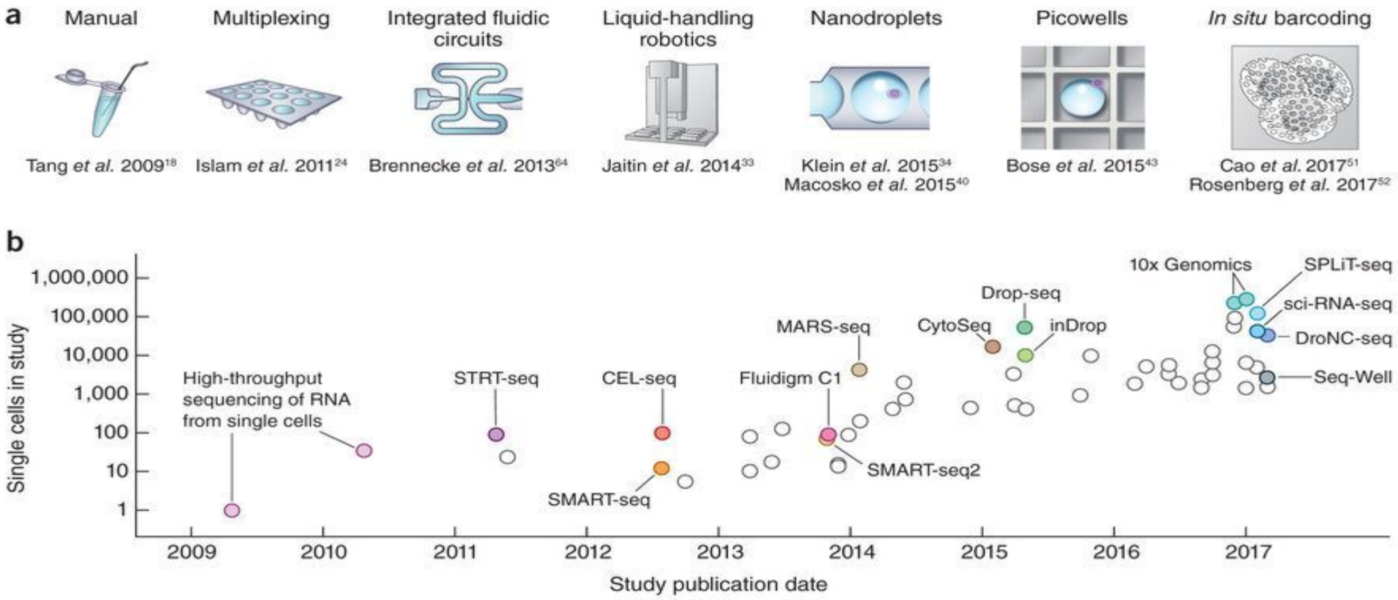
Count matrix
scRNAseq dataset

Noise in the data !

COMPUTATIONAL TOOLS

INVESTIGATE VARIOUS BIOLOGICAL QUESTIONS

ScRNAseq protocols



Full-length

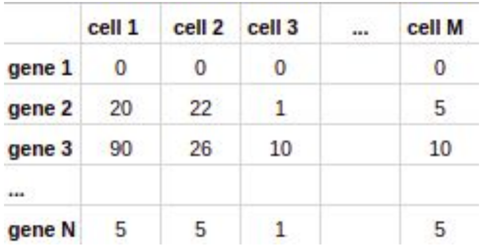
Read counts

Tag-based

UMI counts

ScRNAseq data types

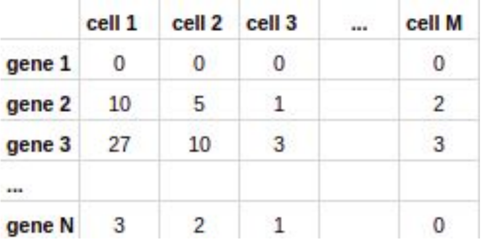
Read counts



	cell 1	cell 2	cell 3	...	cell M
gene 1	0	0	0		0
gene 2	20	22	1		5
gene 3	90	26	10		10
...					
gene N	5	5	1		5

bigger counts

UMI counts



	cell 1	cell 2	cell 3	...	cell M
gene 1	0	0	0		0
gene 2	10	5	1		2
gene 3	27	10	3		3
...					
gene N	3	2	1		0

smaller counts

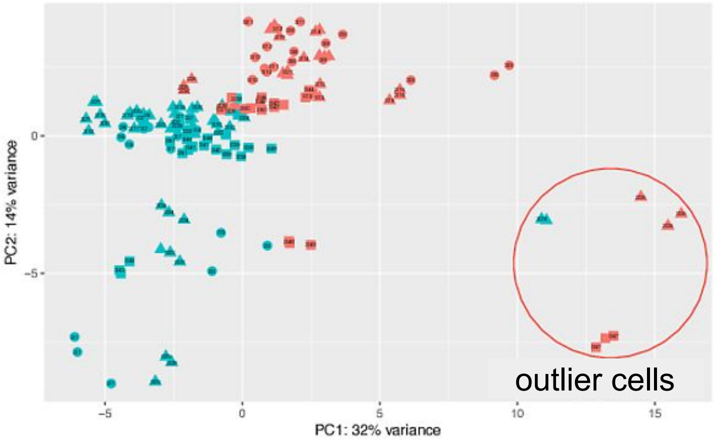
Both data types has different characteristics and contain different source of noise

Not all computational methods are suitable for both data types !

Challenges in analyzing scRNAseq data

Challenges are posed by:

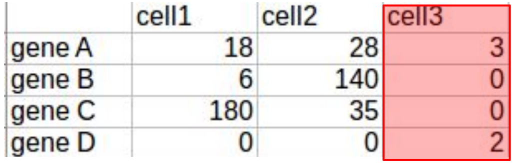
- Technical and biological factors
- Dropouts - missing information about genes expression
- Outliers
- High-dimensionality
  - nr genes: ~ thousands
  - nr cells: ~ hundreds / thousands



Preprocessing scRNAseq data

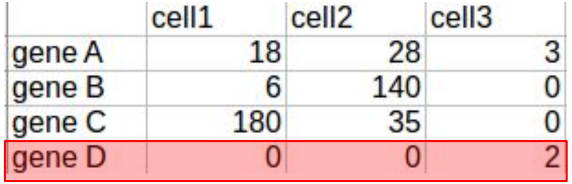
For removing noise

FILTERING
low quality cells



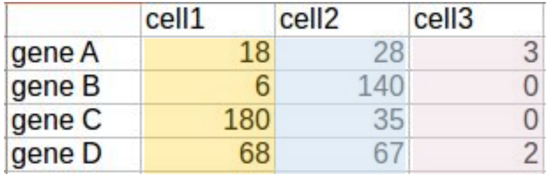
	cell1	cell2	cell3
gene A	18	28	3
gene B	6	140	0
gene C	180	35	0
gene D	0	0	2

FILTERING
lowly expressed genes



	cell1	cell2	cell3
gene A	18	28	3
gene B	6	140	0
gene C	180	35	0
gene D	0	0	2

NORMALIZATION



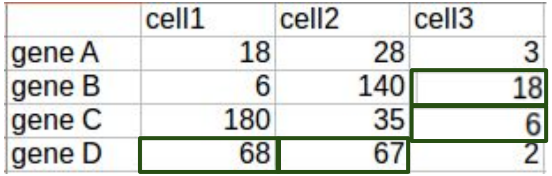
	cell1	cell2	cell3
gene A	18	28	3
gene B	6	140	0
gene C	180	35	0
gene D	68	67	2

* scaling factor 1

* scaling factor 2

* scaling factor 3

IMPUTATION - optional



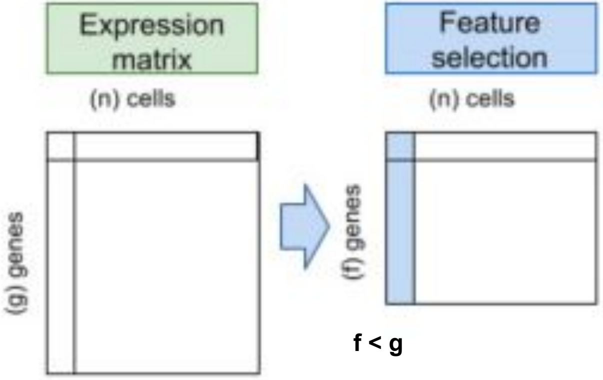
	cell1	cell2	cell3
gene A	18	28	3
gene B	6	140	18
gene C	180	35	6
gene D	68	67	2

Preprocessing scRNAseq data

For dealing with dimensionality

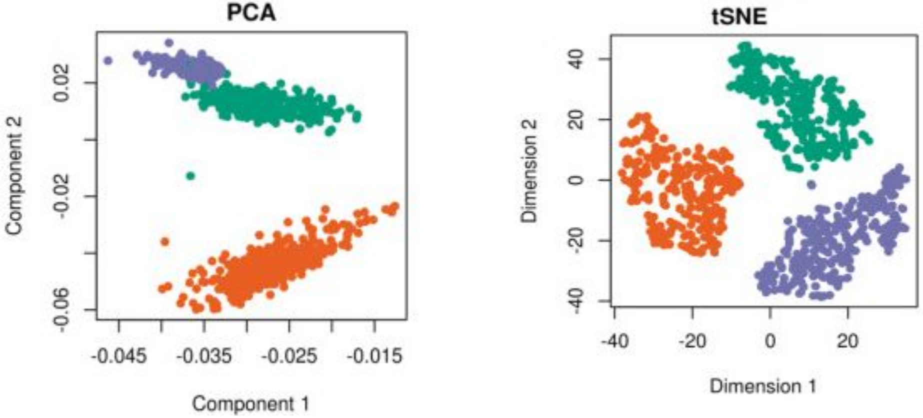
FEATURE SELECTION

Highly variable genes



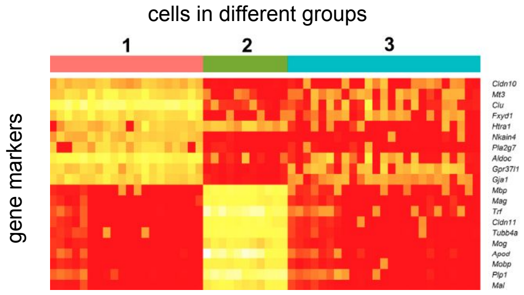
f < g

DIMENSION REDUCTION

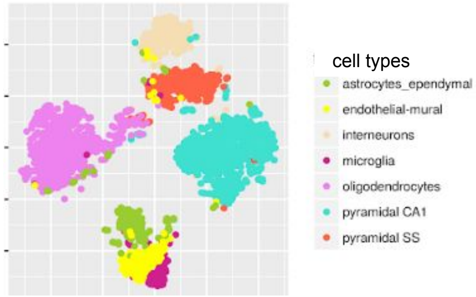


ScRNAseq Applications

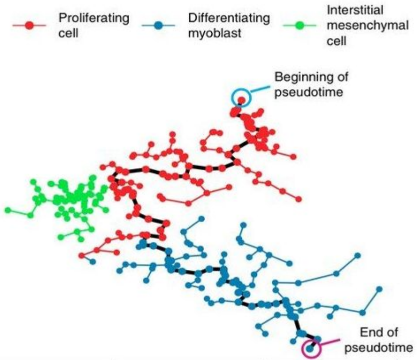
DIFFERENTIAL EXPRESSION



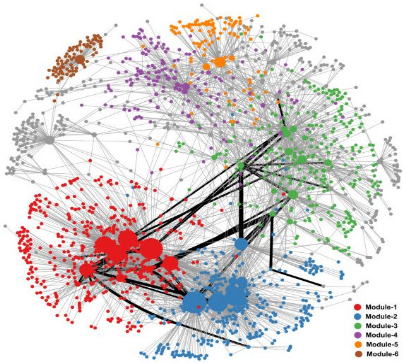
CELL POPULATION DETECTION



TRACKING STAGES OF PROCESS



GENE REGULATORY NETWORKS



Online Materials

For this course:

My github materials:

https://github.com/mkrzak/Single_Cell_Analysis_Workflow

Additional materials:

Single cell Workflows:

Marioni Workflow link

Risso Workflow link

Online single cell course :

https://hemberg-lab.github.io/scRNA.seq.course/

List of softwares for scRNAseq data analysis:

https://github.com/seandavi/awesome-single-cell

Up-to-date articles:

http://academickarma.org/theme/singlecell_rna_sequencing

Let’s start !

Requirements:

- R and RStudio

- Bioconductor packages:

  SingleCellExperiment

  Class for storing data from single cell experiments

  Scater

  Tools for quality control and visualization of scRNA-seq data

  Monocle

  Package for downstream analysis of scRNAseq data

SingleCellExperiment

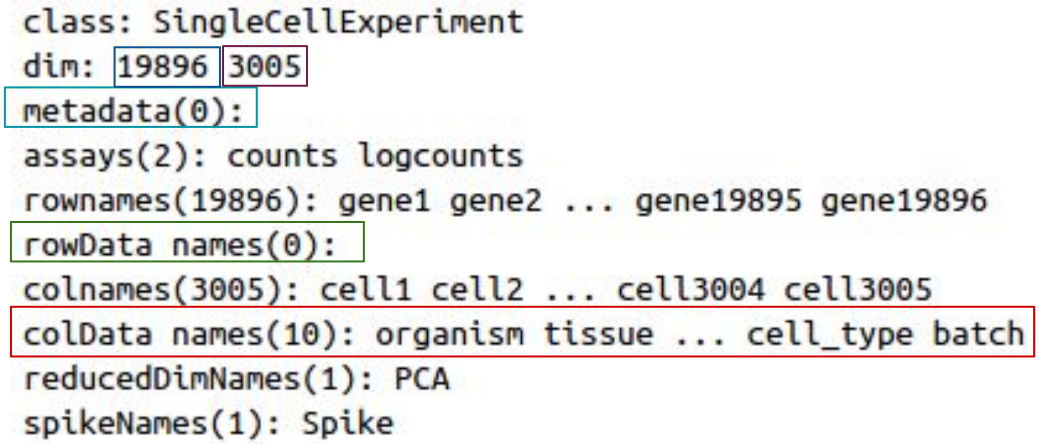
SingleCellExperiment(assays = list(counts = count_matrix, colData = cell_info, rowData = gene_info)

# genes # cells

Info about experiment

Info about features

Info about cells



```
class: SingleCellExperiment
dim: 19896 3005
metadata(0):
assays(2): counts logcounts
rownames(19896): gene1 gene2 ... gene19895 gene19896
rowData names(0):
colnames(3005): cell1 cell2 ... cell3004 cell3005
colData names(10): organism tissue ... cell_type batch
reducedDimNames(1): PCA
spikeNames(1): Spike
```

Scater

Quality control:

calculateQCMetrics()
isOutlier()

Useful for Filtering:

calcAverage()

Normalization:

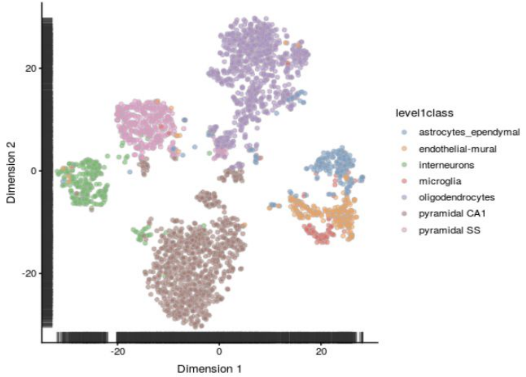
calculateCPM()
calculateFPKM()

Visualization:

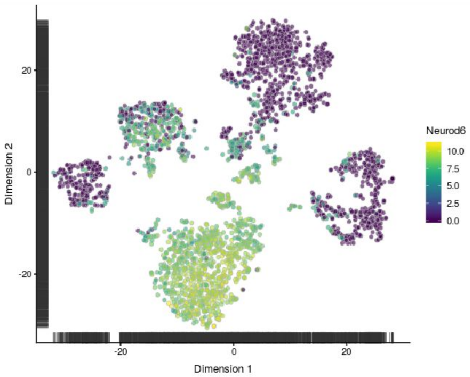
runPCA() / plotPCA()
runTSNE() / plotTSNE()

plotExpression()
plotHighestExprs()
plotExplanatoryVariables()

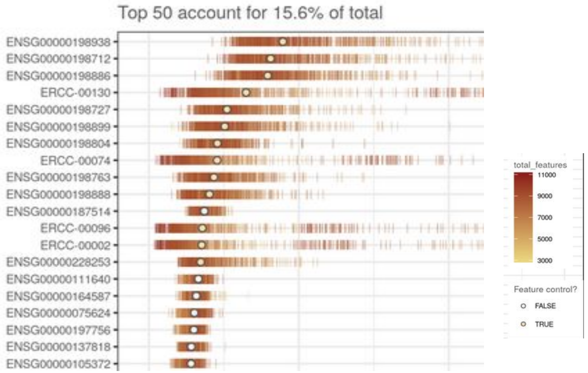
plotTSNE()



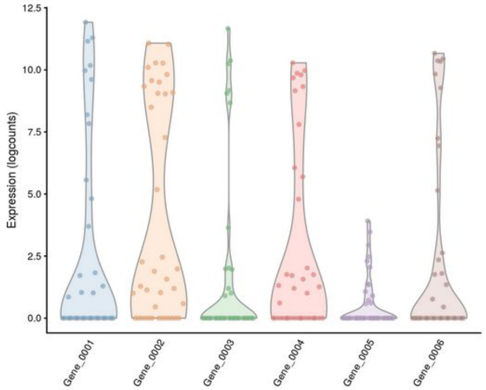
plotTSNE(colour_by="marker_gene")



plotHighestExprs()



plotExpression()



CellDataSet

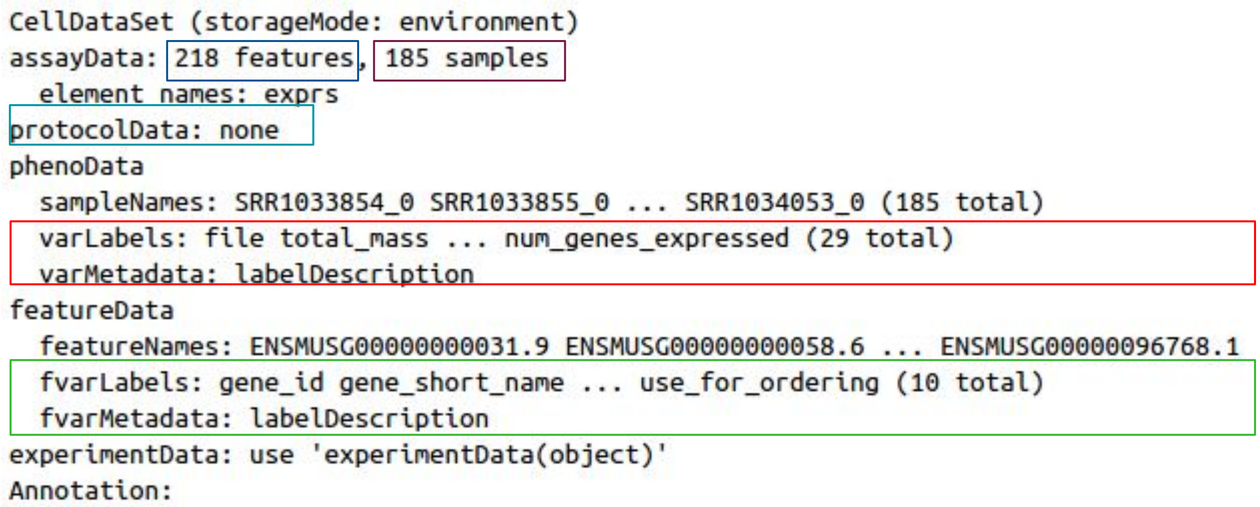
newCellDataSet(cellData = count_matrix, phenoData = cell_info, featureData = gene_info)

# genes # cells

Info about experiment

Info about cells

Info about features



```
CellDataSet (storageMode: environment)
assayData: 218 features, 185 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: SRR1033854_0 SRR1033855_0 ... SRR1034053_0 (185 total)
  varLabels: file total_mass ... num_genes_expressed (29 total)
  varMetadata: labelDescription
featureData
  featureNames: ENSMUSG00000000031.9 ENSMUSG00000000058.6 ... ENSMUSG00000096768.1
  fvarLabels: gene_id gene_short_name ... use_for_ordering (10 total)
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:
```

Monocle

Useful for Filtering:

estimateDispersions() (BiocGenerics)

plot_ordering_genes()

Useful for Normaliation:

estimateSizeFactors() (BiocGenerics)

Dimension reduction:

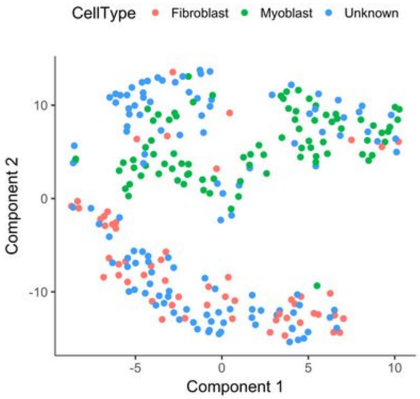
reduceDimension()

Cell population Detection:

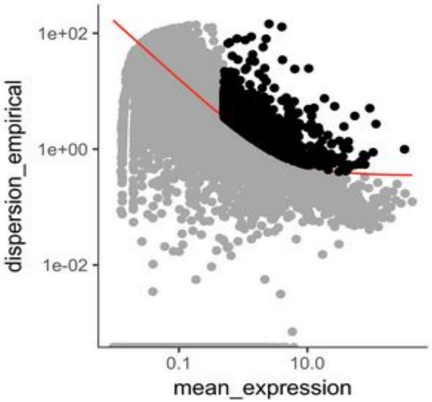
clusterCells()

plot_cell_clusters()

plot_cell_clusters()



plot_ordering_genes()



Merry Christmas
and
Happy New Year !