Single cell RNA sequencing data analysis
Practical exercises

Åsa Björklund

asa.bjorklund@scilifelab.se

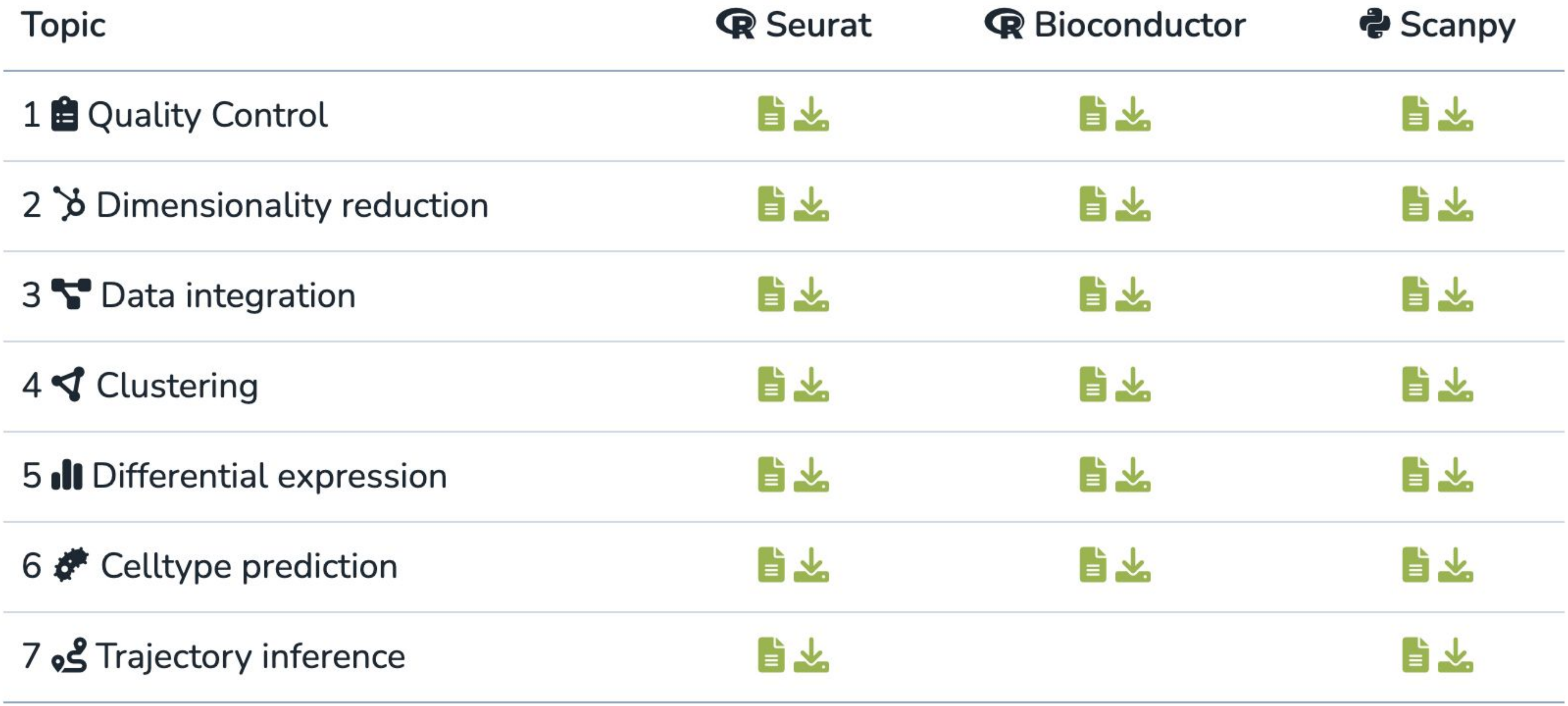
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Practicalities

- Work alone or in pairs as you prefer
- TAs will be around to answer questions about the exercises
- If you finish before hand, please try alternative options in the algorithms we are using. Or try another pipeline.
- If you do not finish on time. Just execute all the code in the notebook so that you can continue with the next step and go back later.

https://nbisweden.github.io/workshop-scRNAseq/home_contents.html

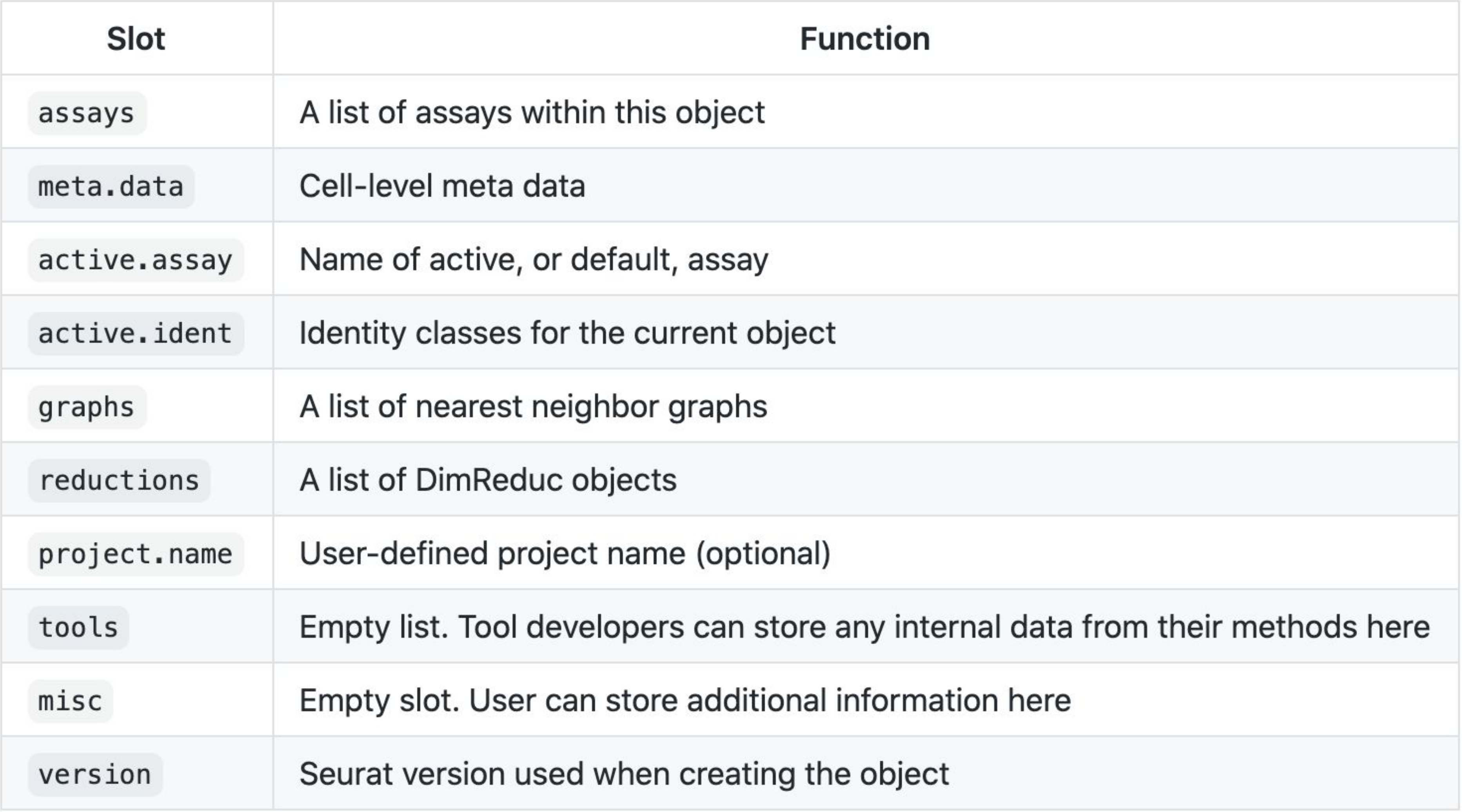


Topic	Seurat	Bioconductor	Scanpy
1 Quality Control	📄 ⬇	📄 ⬇	📄 ⬇
2 Dimensionality reduction	📄 ⬇	📄 ⬇	📄 ⬇
3 Data integration	📄 ⬇	📄 ⬇	📄 ⬇
4 Clustering	📄 ⬇	📄 ⬇	📄 ⬇
5 Differential expression	📄 ⬇	📄 ⬇	📄 ⬇
6 Celltype prediction	📄 ⬇	📄 ⬇	📄 ⬇
7 Trajectory inference	📄 ⬇		📄 ⬇

Three main toolkits for analysing single cell data:

- Seurat:
  - R based, centered around Seurat objects.
  - Mainly developed for droplet based data
  - Easy to use, recommended for R beginners
  - Cons: uses a LOT of memory
- Bioconductor:
  - R based, centered around SingleCellExperiment objects
  - Has more different statistical methods
  - Can handle spike-ins
  - Cons: More complicated than Seurat to run.
- Scanpy:
  - Python based
  - Handles large datasets better. More and more development here.
  - Cons: Does not have all the functionality of the R based tools.

Seurat v4 object



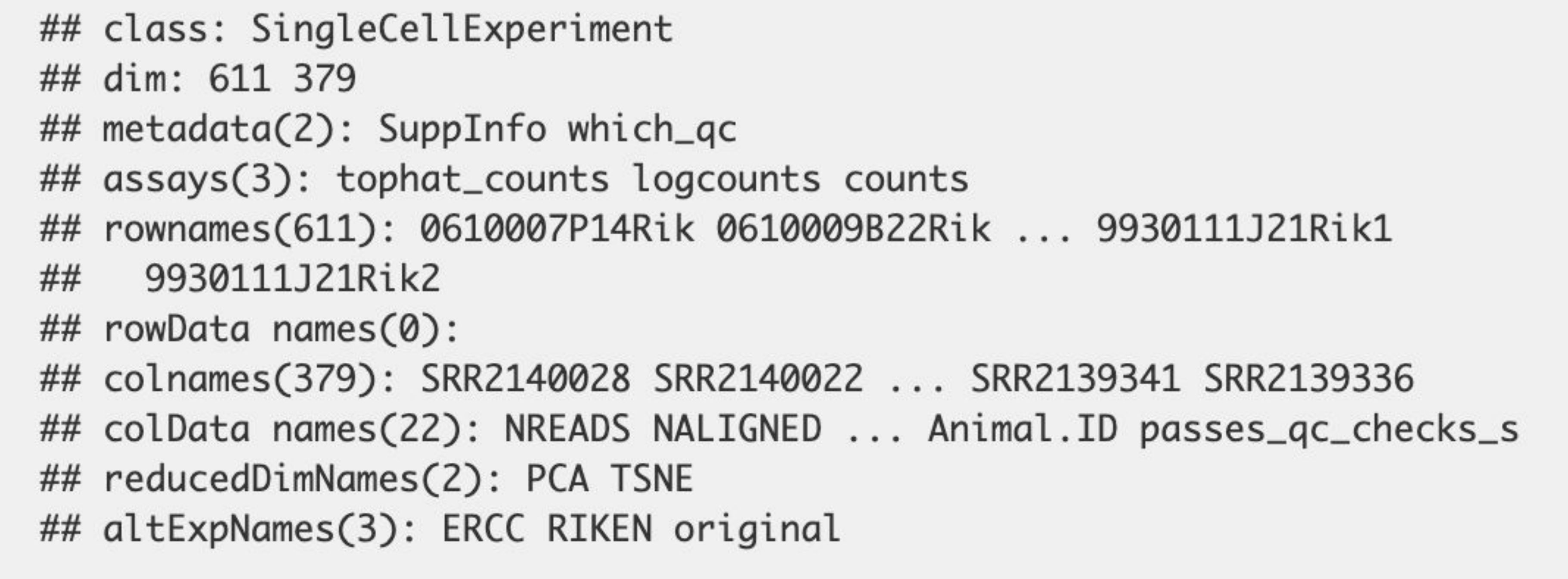
Slot	Function
`assays`	A list of assays within this object
`meta.data`	Cell-level meta data
`active.assay`	Name of active, or default, assay
`active.ident`	Identity classes for the current object
`graphs`	A list of nearest neighbor graphs
`reductions`	A list of DimReduc objects
`project.name`	User-defined project name (optional)
`tools`	Empty list. Tool developers can store any internal data from their methods here
`misc`	Empty slot. User can store additional information here
`version`	Seurat version used when creating the object

https://github.com/satijalab/seurat/wiki/Seurat

Retrieve data from Seurat

```
GetAssayData() # Get expression matrices
Embeddings() # Get reduced dimension components
VariableFeatures() # Get HVGs
Idents() # Get cell identities
Loadings() # Get PCA loadings
FetchData() # Get any column by name
Assays() # List existing assays
Reductions() # List existing reductions
```

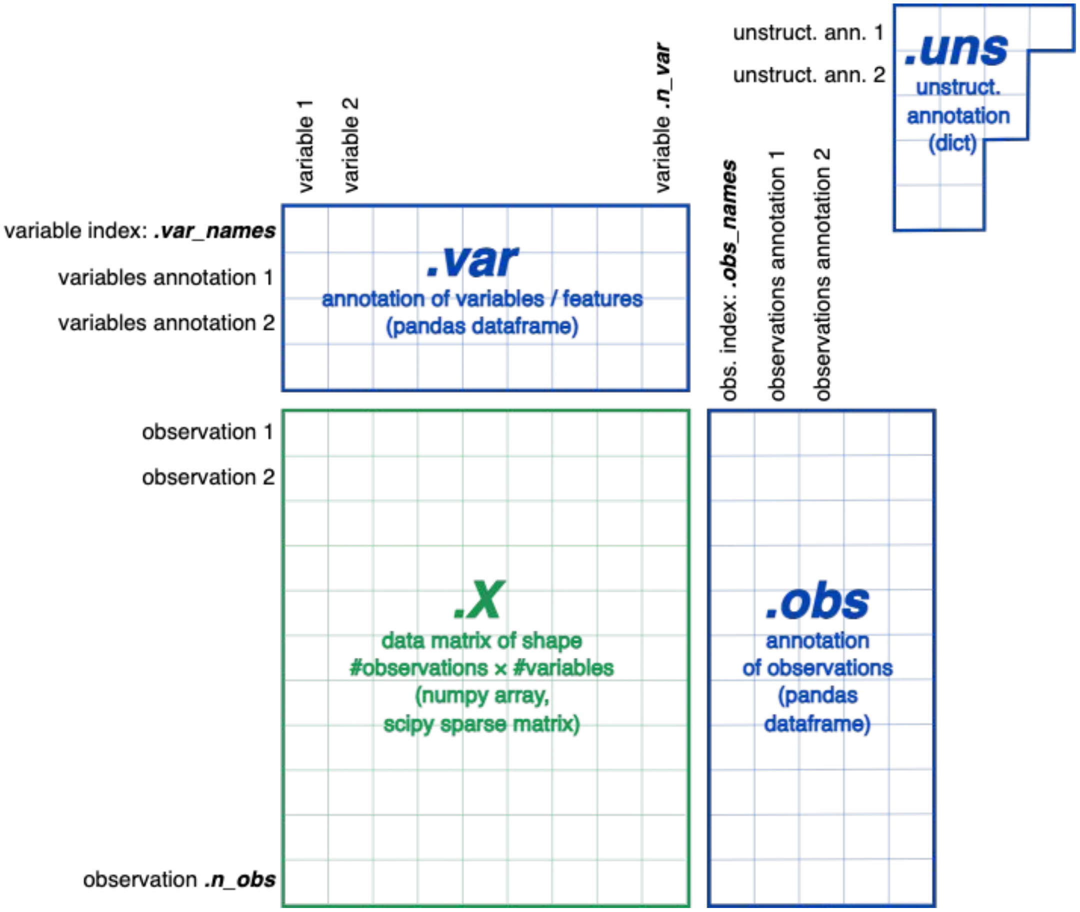
SingleCellExperiment (SCE) objects



```
## class: SingleCellExperiment
## dim: 611 379
## metadata(2): SuppInfo which_qc
## assays(3): tophat_counts logcounts counts
## rownames(611): 0610007P14Rik 0610009B22Rik ... 9930111J21Rik1
##   9930111J21Rik2
## rowData names(0):
## colnames(379): SRR2140028 SRR2140022 ... SRR2139341 SRR2139336
## colData names(22): NREADS NALIGNED ... Animal.ID passes_qc_checks_s
## reducedDimNames(2): PCA TSNE
## altExpNames(3): ERCC RIKEN original
```

https://bioconductor.org/packages/release/bioc/vignettes/SingleCellExperiment/inst/doc/intro.html

AnnData (Scanpy) objects


unstruct. ann. 1
unstruct. ann. 2
.uns
unstruct. annotation (dict)
variable 1
variable 2
variable .n_var
variable index: .var_names
variables annotation 1
variables annotation 2
.var
annotation of variables / features (pandas dataframe)
obs. index: .obs_names
observations annotation 1
observations annotation 2
observation 1
observation 2
observation .n_obs
.X
data matrix of shape #observations × #variables (numpy array, scipy sparse matrix)
.obs
annotation of observations (pandas dataframe)

https://anndata.readthedocs.io/en/latest/anndata.AnnData.html

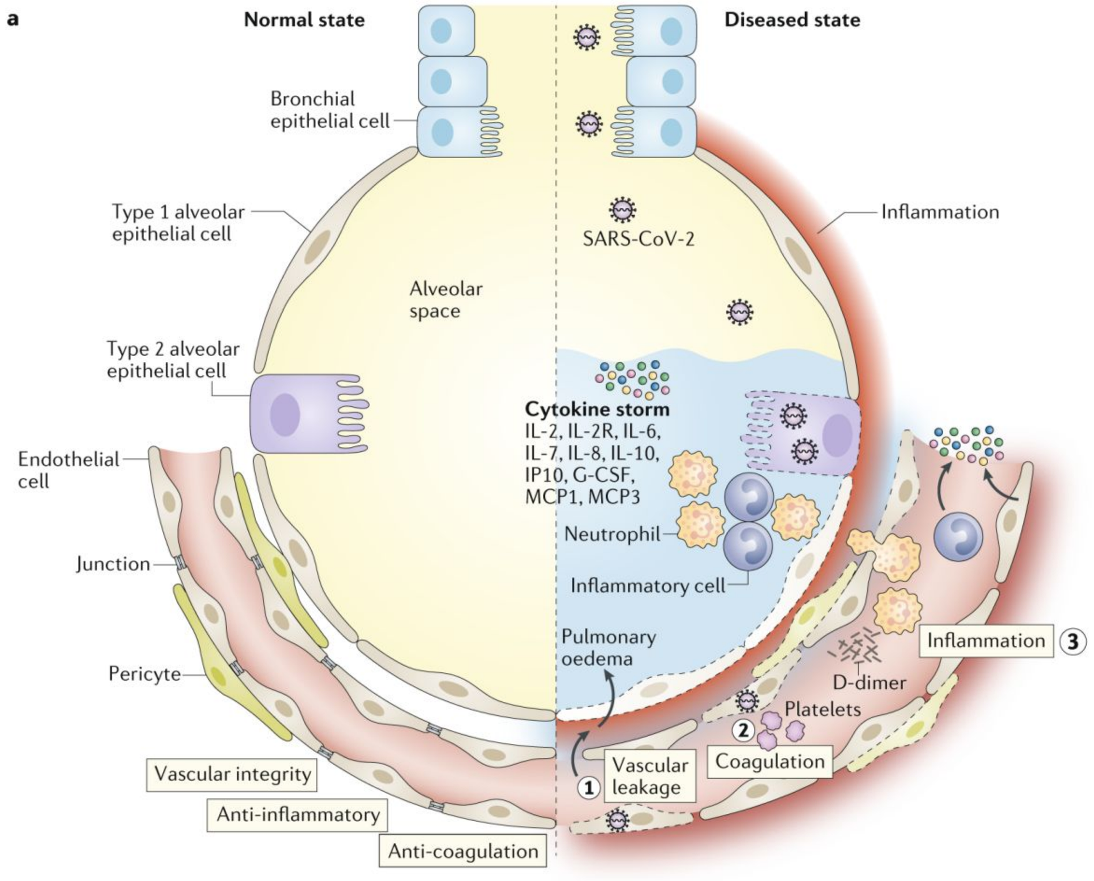
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

What to chose?

- It is recommended that you go through all the steps with one pipeline as each exercise depends on saved objects from the previous step.
- Everyone works in very different pace. Focus on one of the pipelines first. If you have time left over, you can also try out the other ones.

The datasets – Covid-19 PBMCs



Teuwen et al (2020) *Nat reviews Immunology*

Elderly patients usually develop severe lung inflammation and lung dysfunction.

Many cell types orchestrate the immune response to the virus.

Their relative contribution at the single-cell resolution is still unclear

GOAL: Which cell types and genes are altered when comparing blood immune cells from healthy vs disease?

The datasets – Covid-19 PBMCs

- Data from paper: ”Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19” Lee et al. Sci Immuno
- We have selected 4 controls and 4 severe covid samples and subsampled to 1500 cells per subject for computational speed/memory.
- ST and trajectory lab will be with other datasets.

Containers - Docker/Singularity

- An environment with all necessary tools have been prepared for you (Docker/Singularity)
- Computations run on Uppmax cluster
- You work interactively in Rstudio IDE or JupyterLab in your browser
- Detailed instructions on running labs: https://nbisweden.github.io/workshop-scRNAseq/other/containers.html

The code:

- All code for the exercises is available as Quarto documents (.qmd), or jupyter notebooks, in the folder: **workshop-scRNAseq/compiled/labs/**

- Please report to us if you find any errors in the code!
  - Slack channel #**exercises**
  - An Issue on the github page

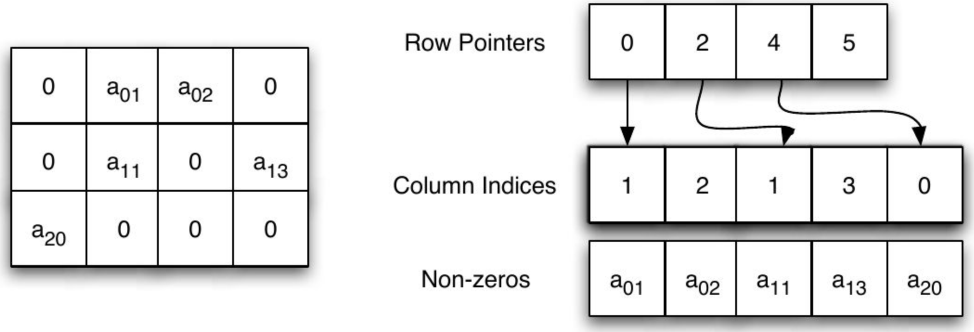
- We may find bugs and update the code – in that case, update your git repo.

Reproducible coding

- You should always be able to find and recreate the results.
  - Scripts should be able to run from input files to create the output.
  - Never work with saved R sessions!
- Name your scripts with relevant names so you can find them 2 years later ☺️
- Always backup code – good idea to use github that also gives you version control.

Sparse vs dense matrices

- scRNAseq data is large matrices with many zeros -> perfect for sparse matrices.
- Only has representation of non-zero value and its positions.
- In R – need package Matrix for any matrix operations. Seurat uses `dgCMatrix` format.
- In python - scipy.sparse, normally `csr_matrix`



0	a_{01}	a_{02}	0
0	a_{11}	0	a_{13}
a_{20}	0	0	0

Row Pointers	0	2	4	5	
Column Indices	1	2	1	3	0
Non-zeros	a_{01}	a_{02}	a_{11}	a_{13}	a_{20}

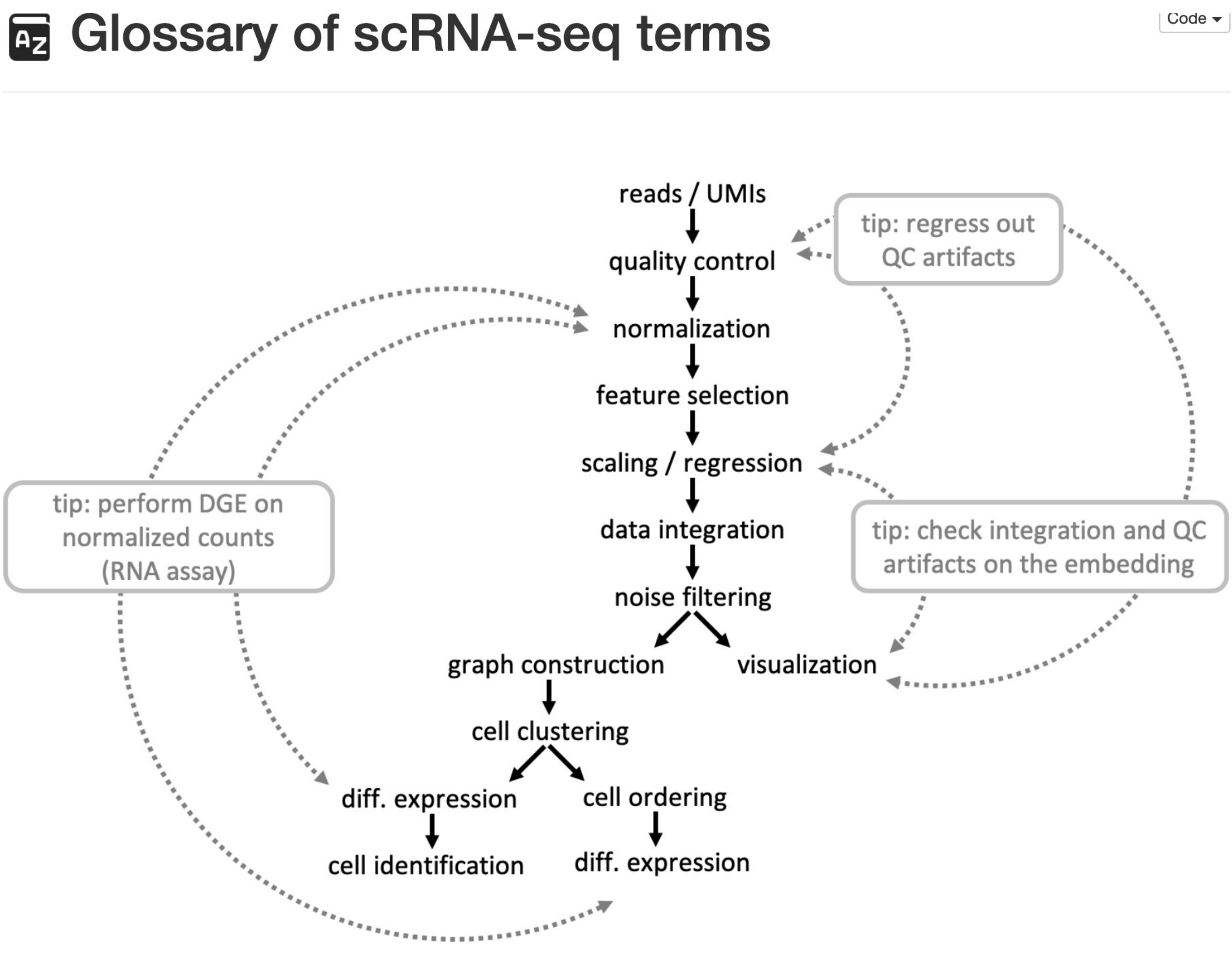
Memory issues

- scRNAseq datasets are often large, think about how you code. Avoid duplicating objects!
- Remove unused matrices and clear memory with gc().
- Try to keep your matrices sparse!
- If you still have issues with memory in R, test setting e.g. R_MAX_VSIZE=70Gb in the .Renviron file. Default is 16Gb. (check FAQ section)
- In Seurat – can use DietSeurat() function to remove assays, data slots etc.

Troubleshooting

- Slack channel - #**exercises** or just raise your hand
- It is important that you learn how to troubleshoot yourselves.
  - Look at your error messages, perhaps the answer is there?
  - If not – Google is your best friend! Forums like Seqanswers, Stackexchange, Bioconductor support forum, specific forums (or github issues) for each package may have the answer.
- TAs are there to answer any questions and give suggestions, but we may not always have the answer.

NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN



Glossary of scRNA-seq terms

reads / UMIs

quality control

normalization

feature selection

scaling / regression

data integration

noise filtering

graph construction

visualization

cell clustering

diff. expression

cell ordering

cell identification

diff. expression

tip: regress out QC artifacts

tip: perform DGE on normalized counts (RNA assay)

tip: check integration and QC artifacts on the embedding

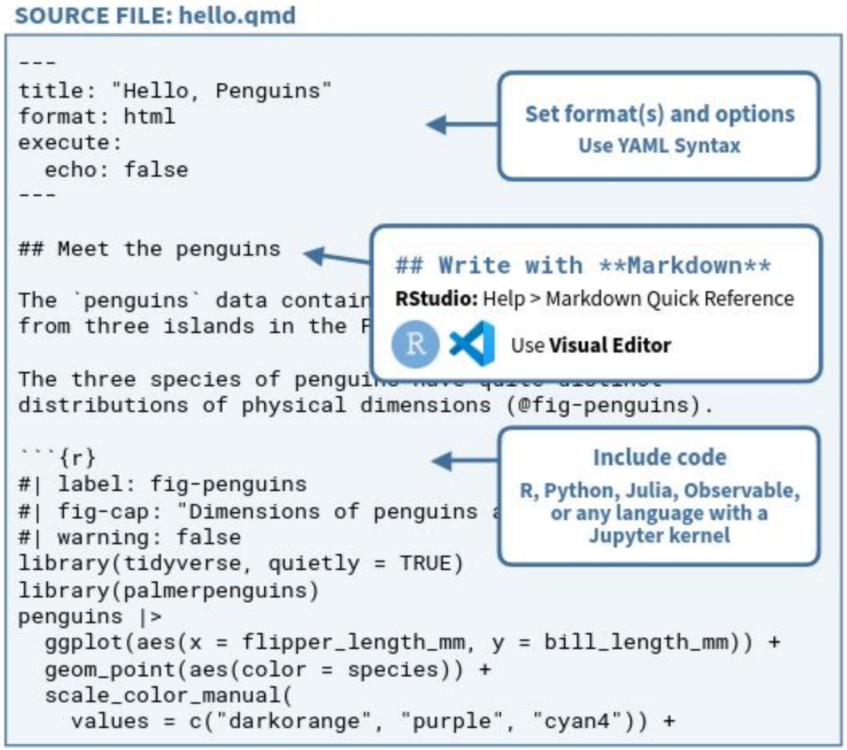
https://nbisweden.github.io/single-cell-pbl/glossary_of_terms_single_cell.html

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Quarto (.qmd)

- Complete reports with both text, code and plots.
- 3 main parts:
  - **Yaml header** – specify output formats and config.
  - **Code chunks** – all code, define output styles for plots and code evaluation
  - **Markdown text** – follows markdown syntax to produce headers and text.



SOURCE FILE: hello.qmd

~~~
---
title: "Hello, Penguins"
format: html
execute:
  echo: false
---

## Meet the penguins

The `penguins` data contain
from three islands in the P

The three species of penguins have quite distinct
distributions of physical dimensions (@fig-penguins).

```{r}
#| label: fig-penguins
#| fig-cap: "Dimensions of penguins a
#| warning: false
library(tidyverse, quietly = TRUE)
library(palmerpenguins)
penguins |>
  ggplot(aes(x = flipper_length_mm, y = bill_length_mm)) +
  geom_point(aes(color = species)) +
  scale_color_manual(
    values = c("darkorange", "purple", "cyan4")) +
~~~

Set format(s) and options
Use YAML Syntax
## Write with **Markdown**
RStudio: Help > Markdown Quick Reference
Use Visual Editor
Include code
R, Python, Julia, Observable, or any language with a Jupyter kernel

https://rstudio.github.io/cheatsheets/quarto.pdf

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

UPPMAX

Make sure you are asking for compute resources only once

```
squeue -A naiss2023-22-1345 | sort -k 4
```

Your username must be listed only once

Make sure you have some space on your home directory

Singularity containers mount your home directory, so there is chance of conflict with existing RStudio/Python configuration

In Rstudio server, disable auto save history

Demonstration