Single cell RNA sequencing data analysis
Practical exercises

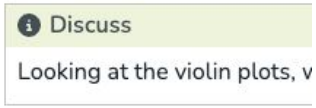
Åsa Björklund
asa.bjorklund@scilifelab.se

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Practicalities

- Work and discuss the exercises in your breakout room.
- Each exercise has a number of Discussion points, take time with your group to talk through them



- TAs will move around to answer questions about the exercises
- If you want a TA to come to your room, just write us a message in slack #exercises

Practicalities

- Last 10 minutes of each exercise will be a summary of that exercise.
- 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/v5 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

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
```

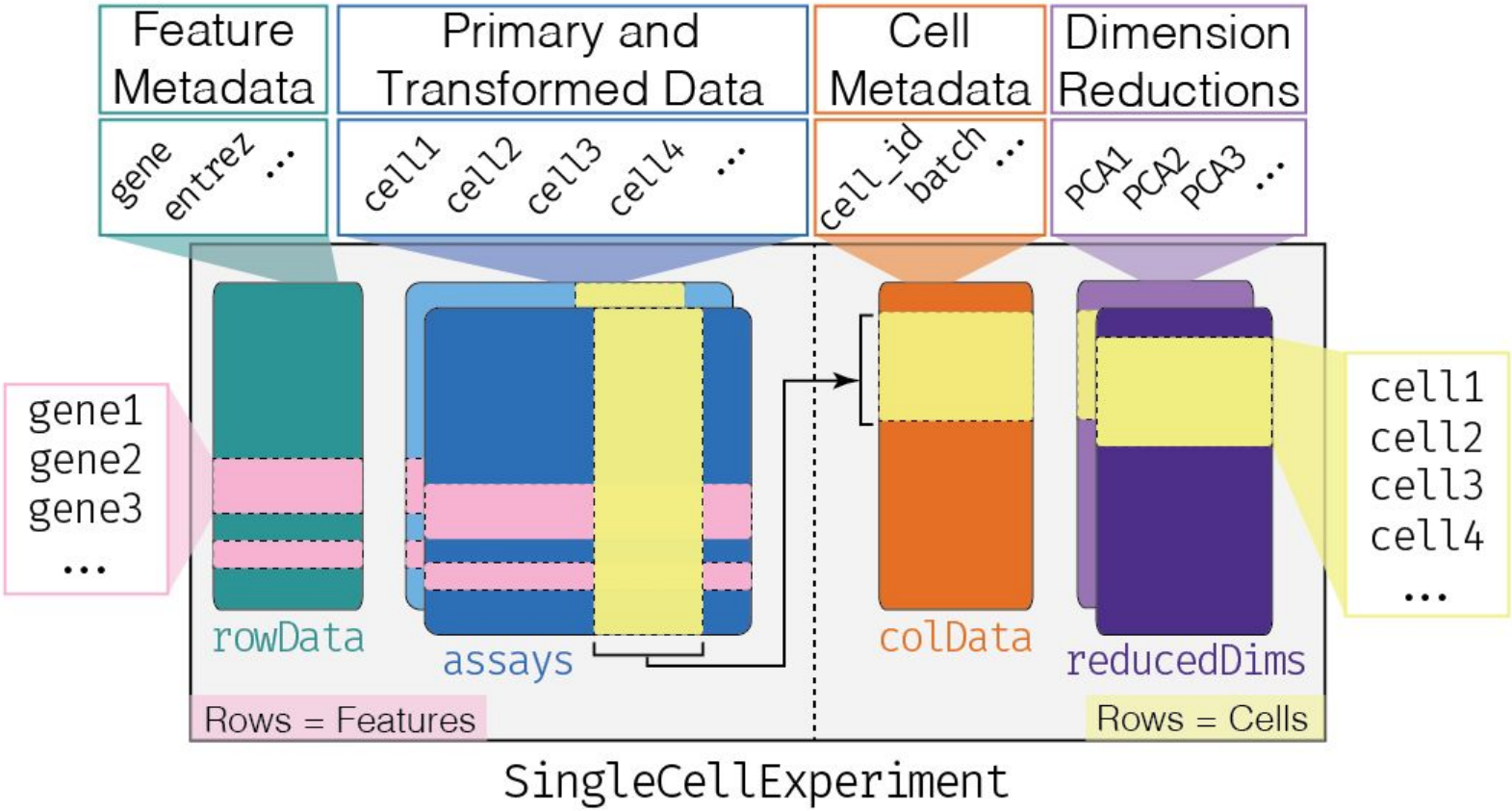
Seurat v5 - Layers

Count/data matrices may be split by sample into multiple layers, or merged into a single matrix.

```
Layers() # List existing layers
JoinLayers() # Merge all layers
split() # split into layers by a factor
```

It is important to keep in mind what layers you have, as some functions will behave differently.

SingleCellExperiment (SCE) objects

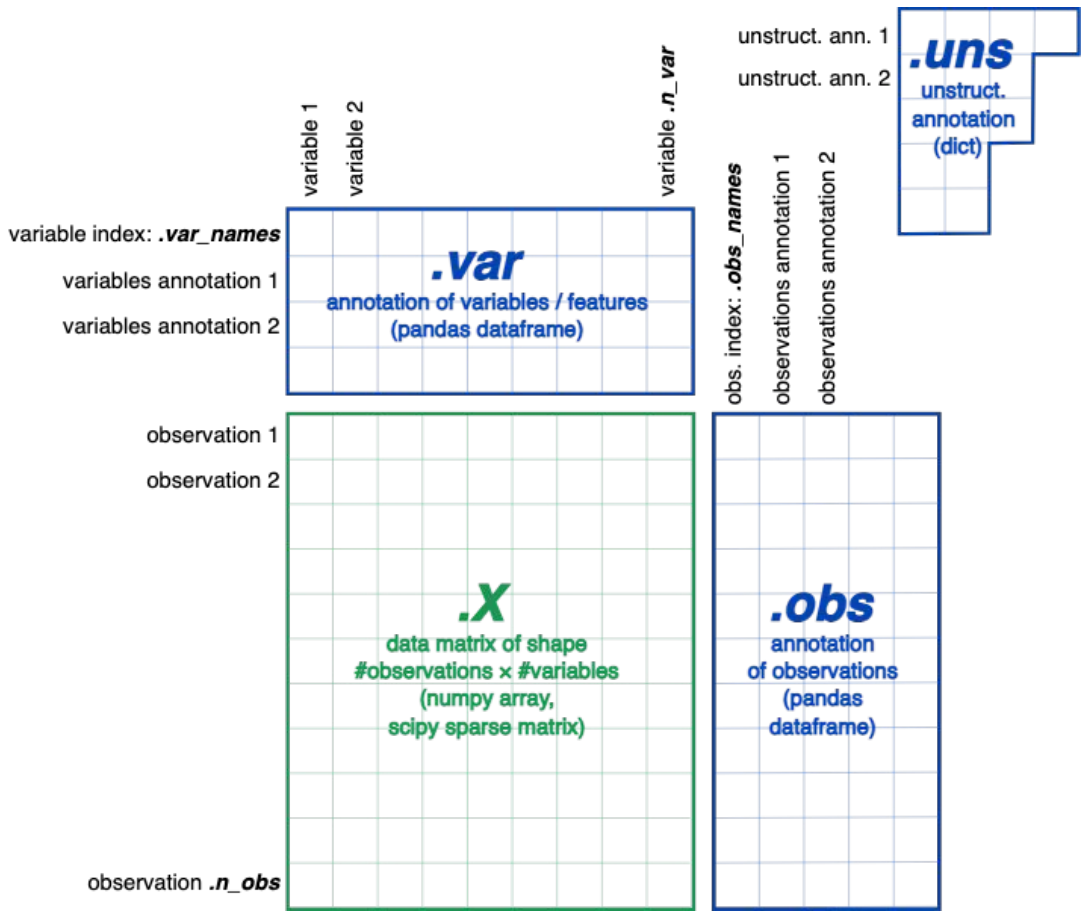


https://bioconductor.org/books/3.13/OSCA.intro/the-singlecellexperiment-class.html

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

AnnData (Scanpy) objects



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

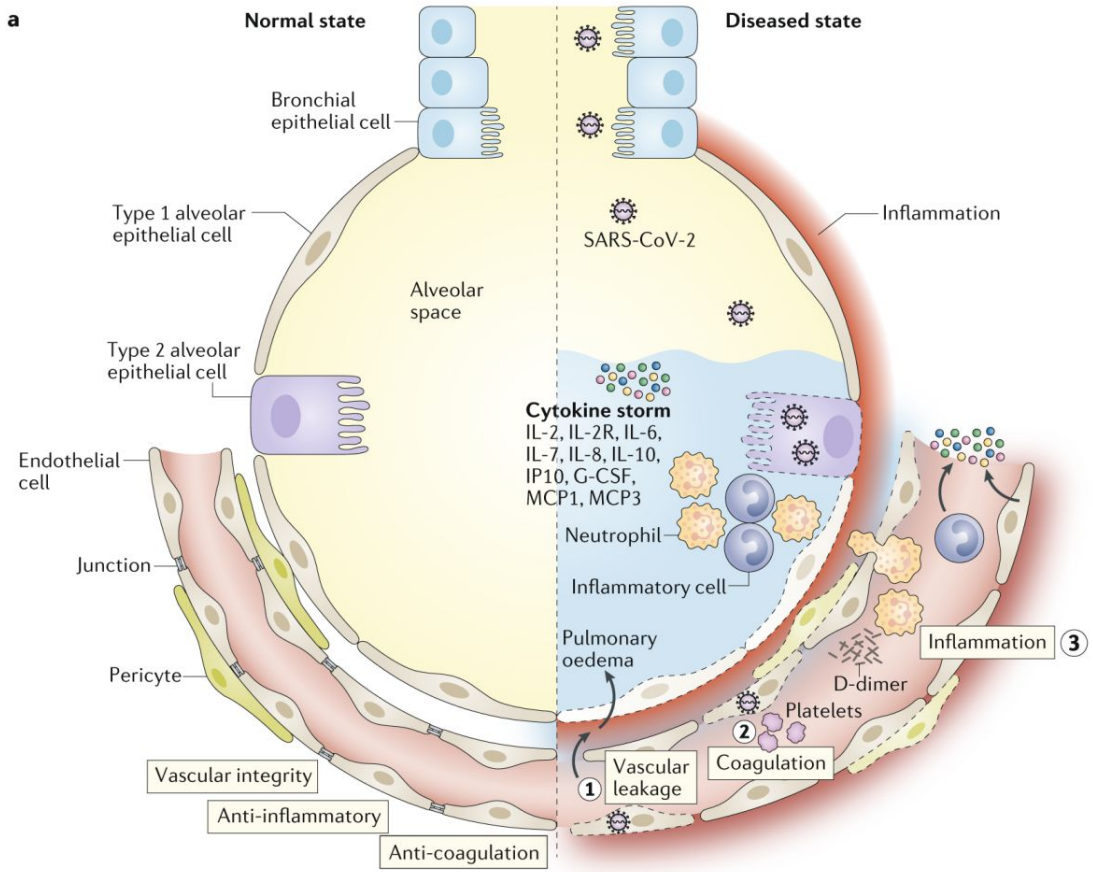
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

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

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

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

SciLifeLab

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

- An environment with all necessary tools have been prepared for you in Docker containers
- Computations run on Scilifelab serve cluster
- You work interactively in Rstudio IDE or JupyterLab in your browser https://nbisweden.github.io/workshop-scRNAseq/other/scilifelab-serve.html

Containers - Docker

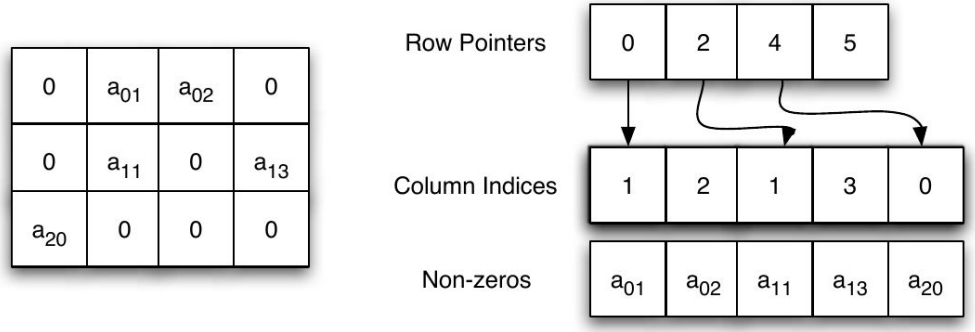
- Instructions on running labs locally: https://nbisweden.github.io/workshop-scRNAseq/other/containers.html
- You can install the conda environment that is in the containers with environment files in (obs for now only works in linux without hacking yourself): workshop-scRNAseq/tree/master/containers/conda
-

The code:

- All code for the exercises is available as Quarto documents (.qmd), or jupyter notebooks, in the folder: **workshop-scRNAseq/compiled/labs/**
- A copy script is included in the containers that will copy all the code for you.
- 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, rerun the copy script.

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`



NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

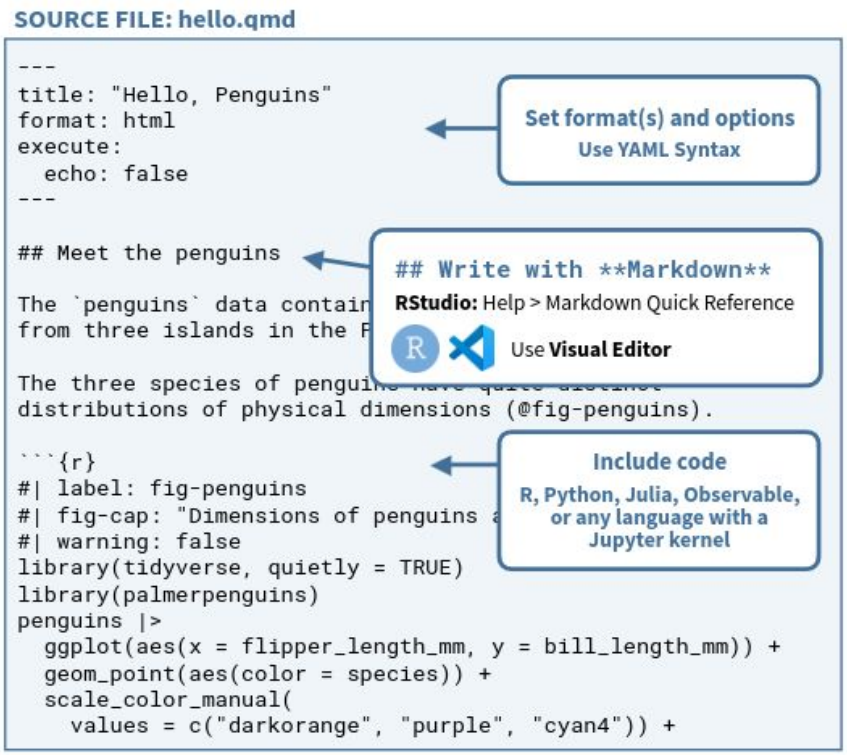
SciLifeLab

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.

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.



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

Demonstration

List of breakout rooms