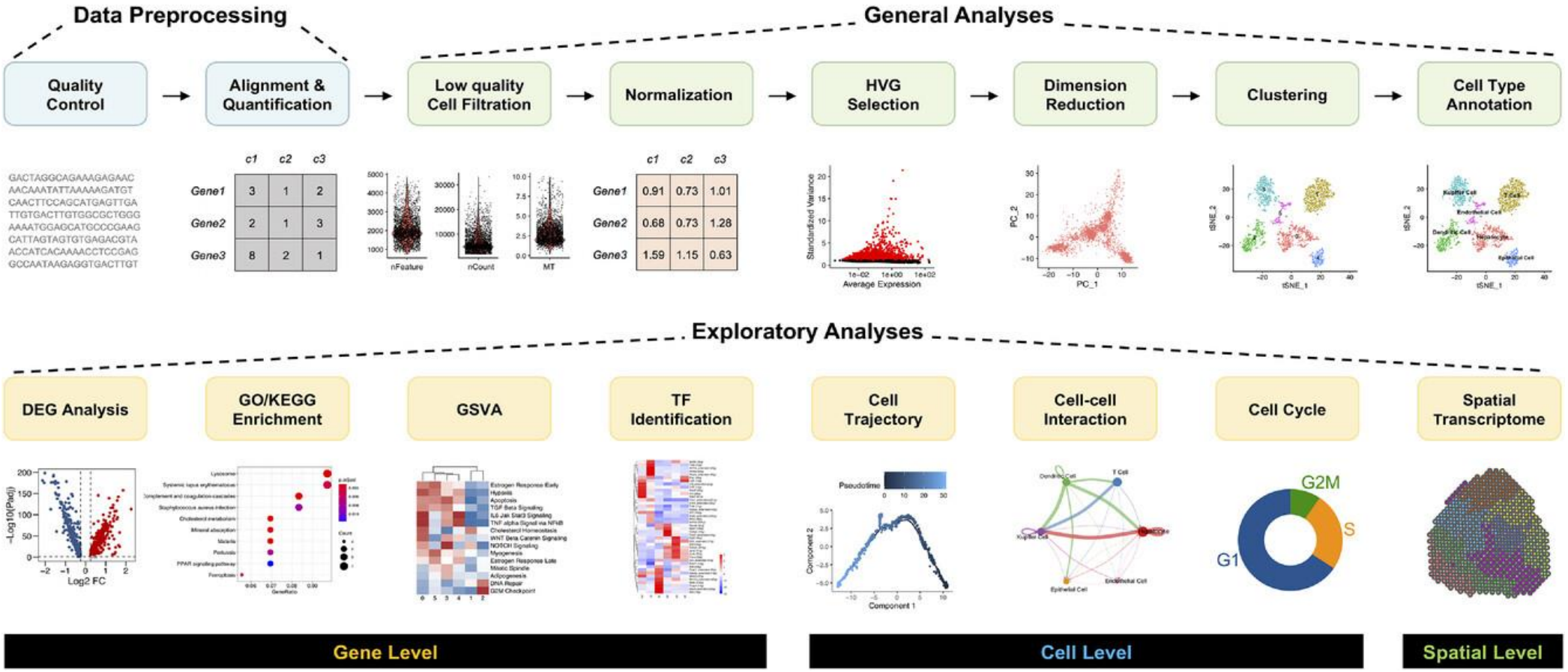
# Single-cell RNAseq Analysis Workshop

# Lecture 4: Beyond Seurat: (More) advanced topics

March 4, 2024

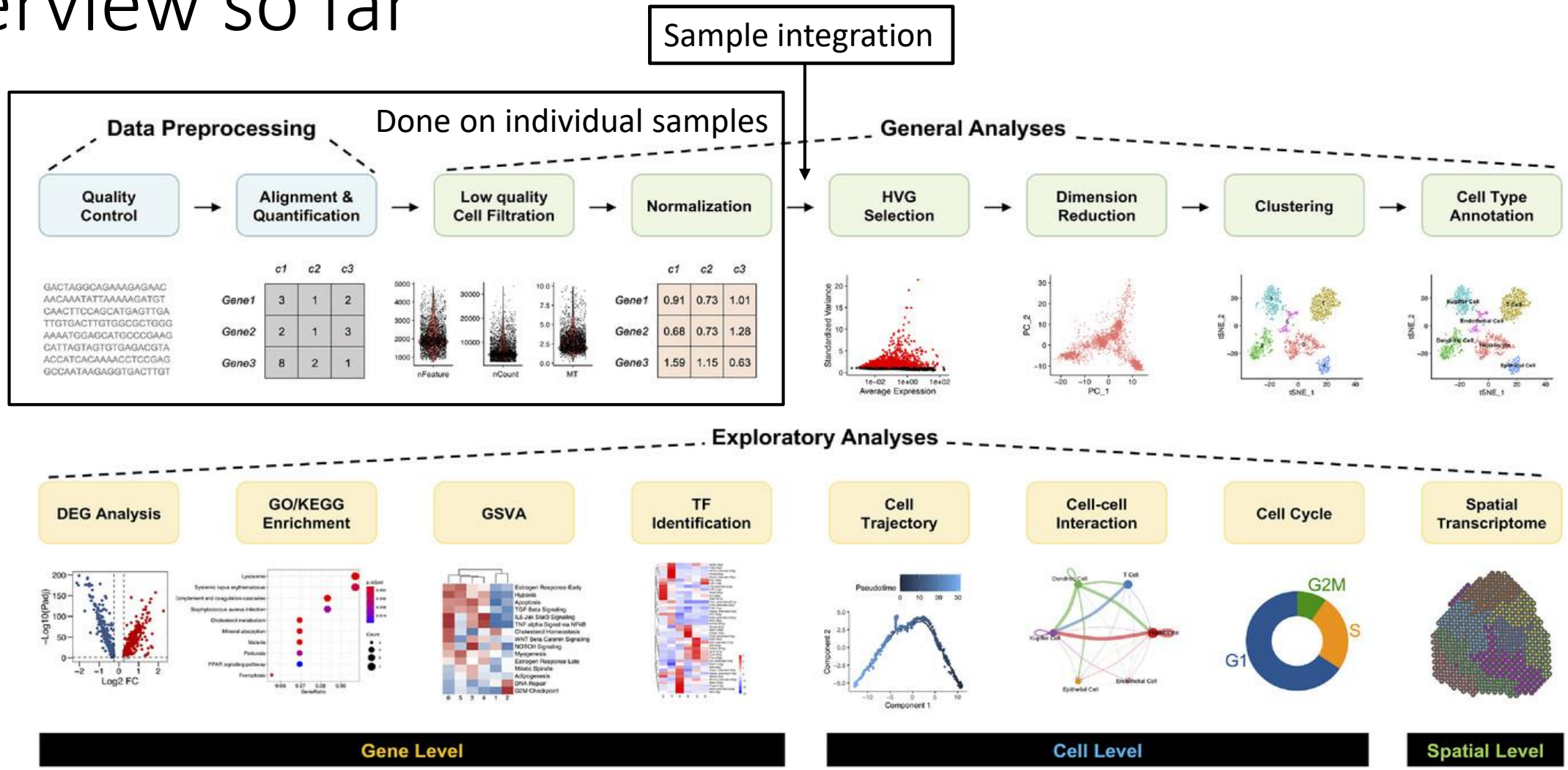
Melissa Hubisz

# Overview so far




Clinical & Translational Med, Volume: 12, Issue: 3, First published: 29 March 2022, DOI: (10.1002/ctm2.694)

# Overview so far

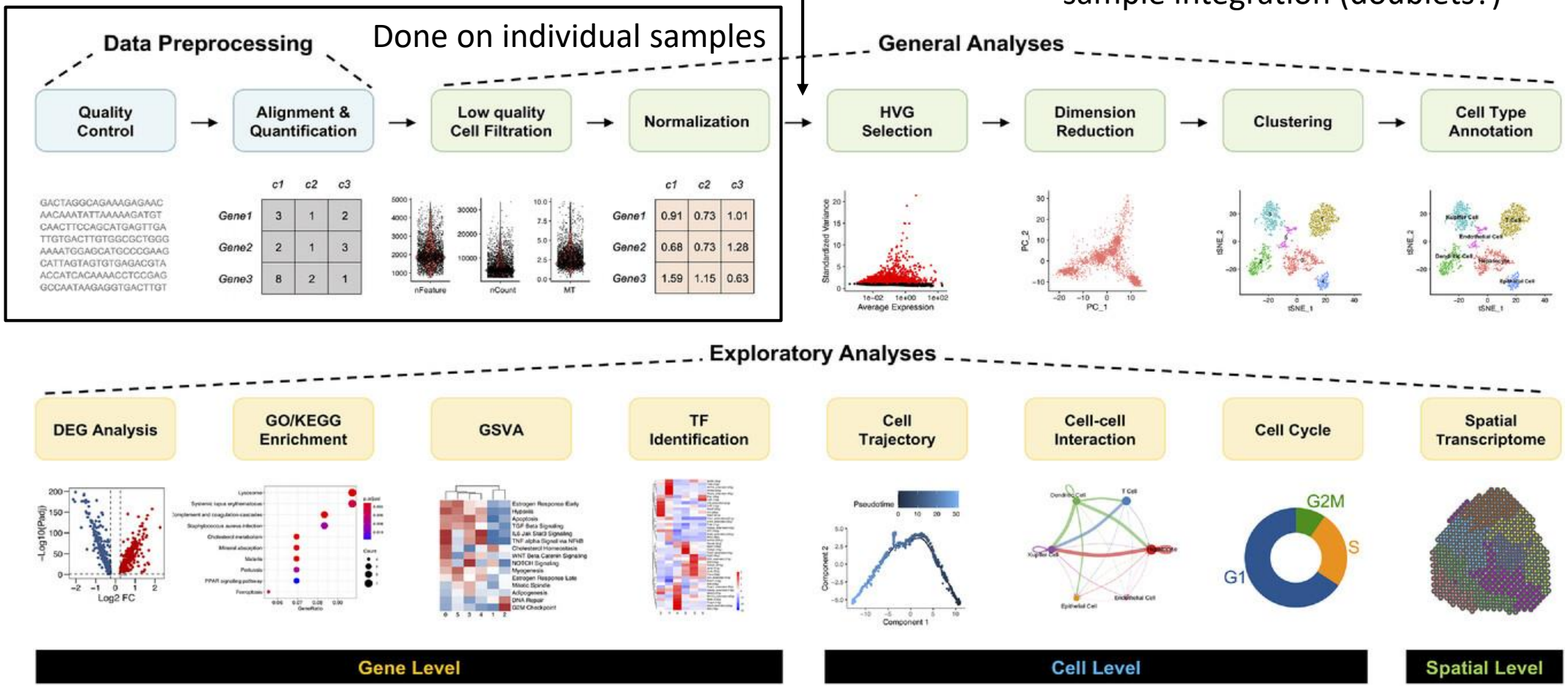



Clinical & Translational Med, Volume: 12, Issue: 3, First published: 29 March 2022, DOI: (10.1002/ctm2.694)

# Overview so far

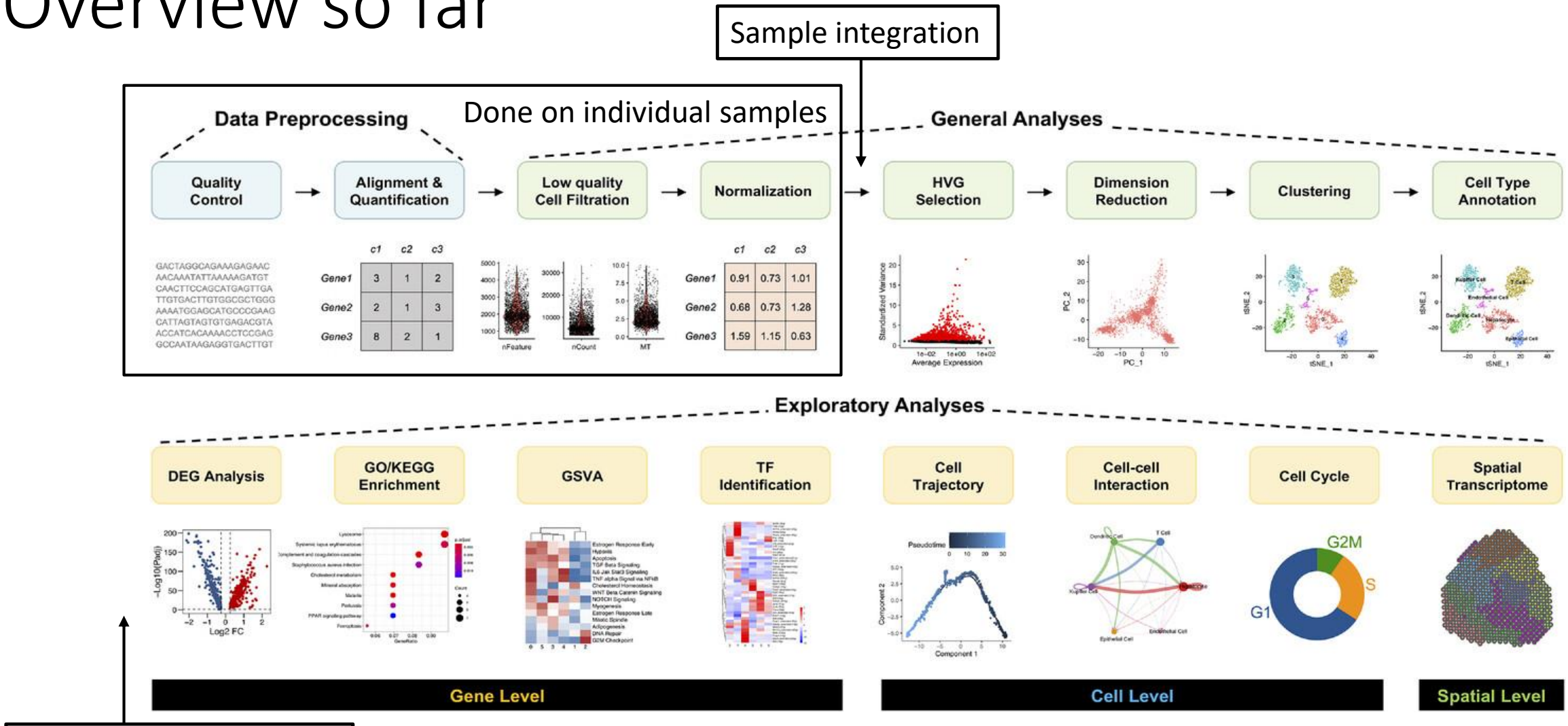
Sample integration

Some quality filtering may be done after sample integration (doublets?)




Clinical & Translational Med, Volume: 12, Issue: 3, First published: 29 March 2022, DOI: (10.1002/ctm2.694)

# Overview so far

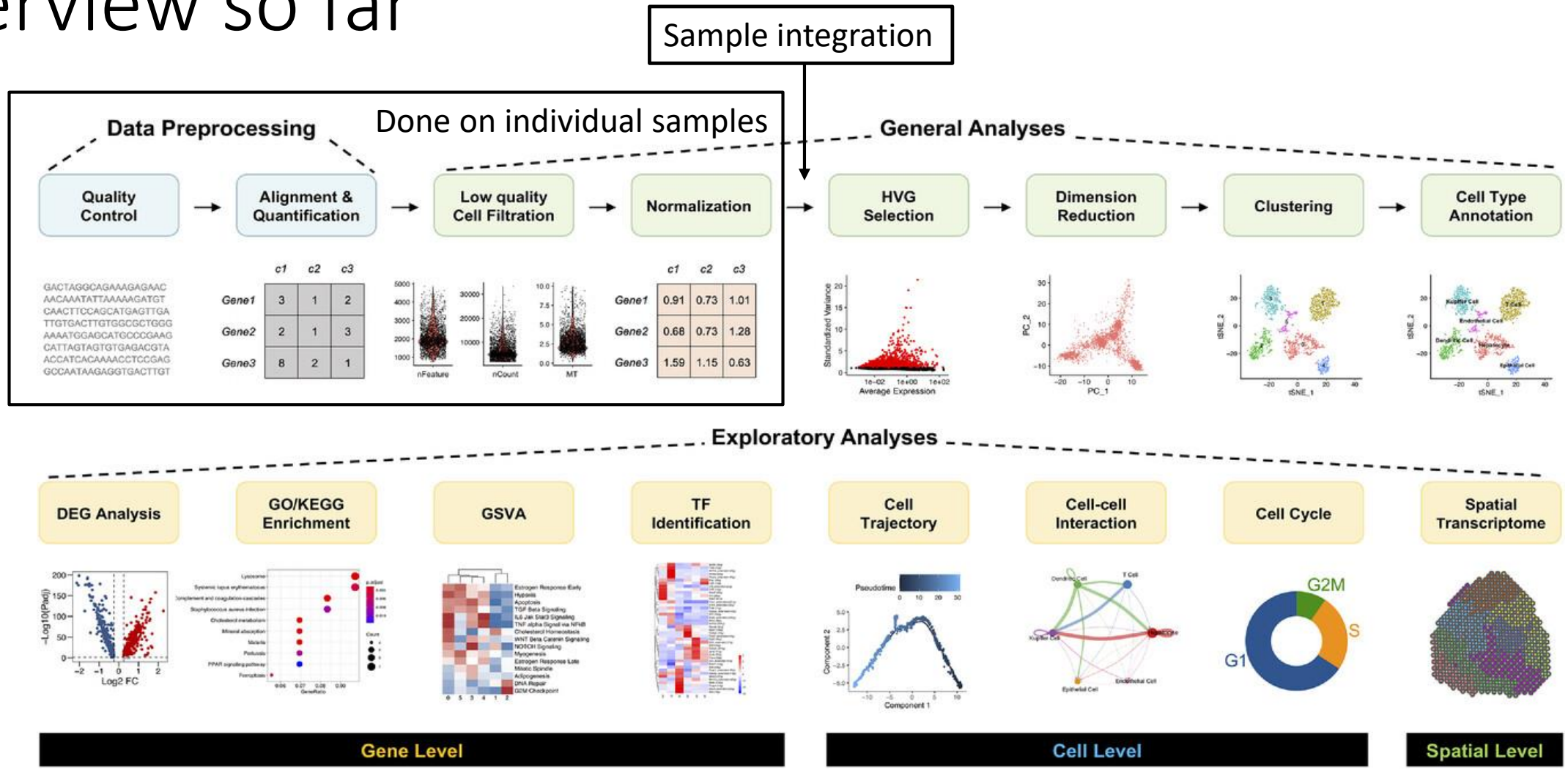



Divide data by cell type

1. Individual sample analysis
2. Merged sample analysis
3. Cell-type specific analysis

Clinical & Translational Med, Volume: 12, Issue: 3, First published: 29 March 2022, DOI: (10.1002/ctm2.694)

# Overview so far

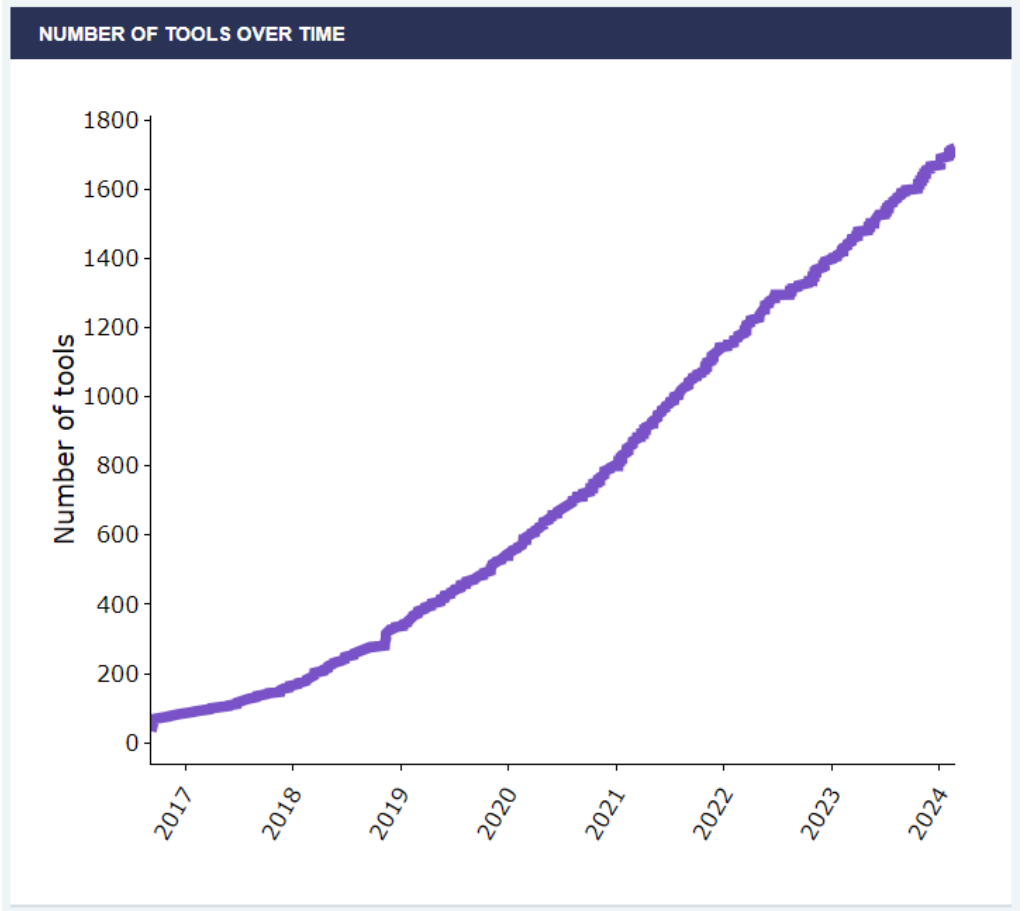


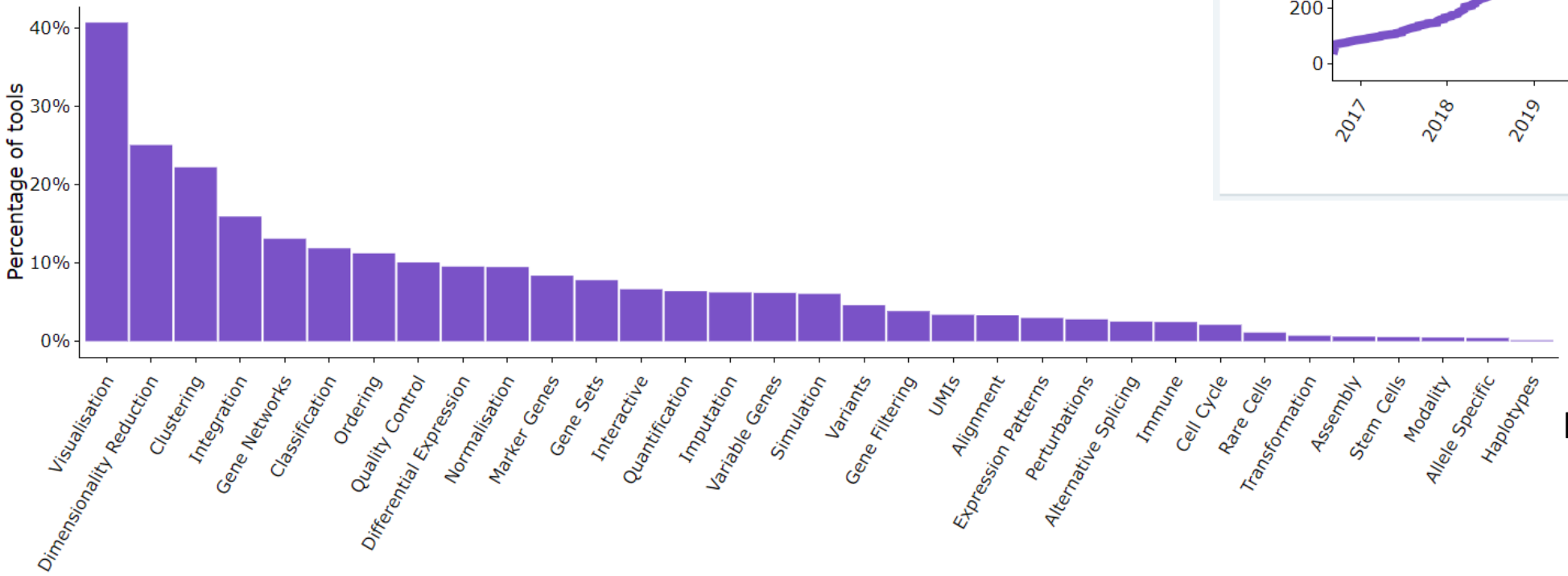

Almost all these topics could be a lecture (or several) in themselves!

Clinical & Translational Med, Volume: 12, Issue: 3, First published: 29 March 2022, DOI: (10.1002/ctm2.694)

# Some motivation for today

- scRNA-seq analysis is rapidly developing field
- Big challenge is choosing appropriate software, then installing/using it
- The number of tools is overwhelming, we cannot scratch the surface in teaching them




Source:
https://scrna-tools.org/analysis

scRNA-tools

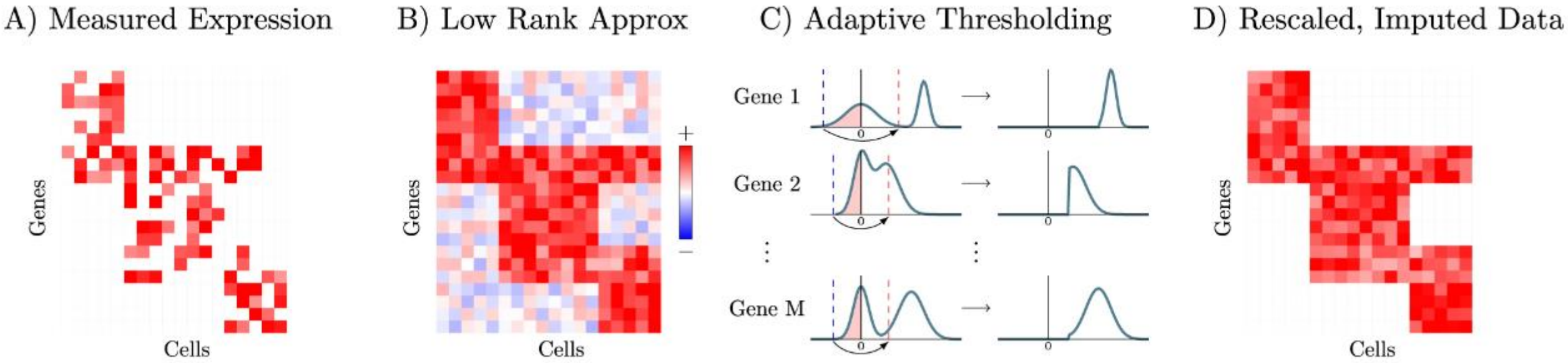
# Plan for today

- Brief review of imputation and trajectory analysis
- Intro to scanpy
- Intro to docker, using jupyter in docker
- Interactive demonstration

# Imputation

- Try to distinguish between **biological** and **technical** zeroes in the data and correct for technical zeroes
- Use correlations in gene expression to estimate unmeasured expression
- Some methods zero-preserving (ARLA), others not (MAGIC, scImpute)
- A bit controversial to use impute-corrected counts in analysis – may cause bias or false signals
- Can be very useful for visualization

# ARLA imputation method



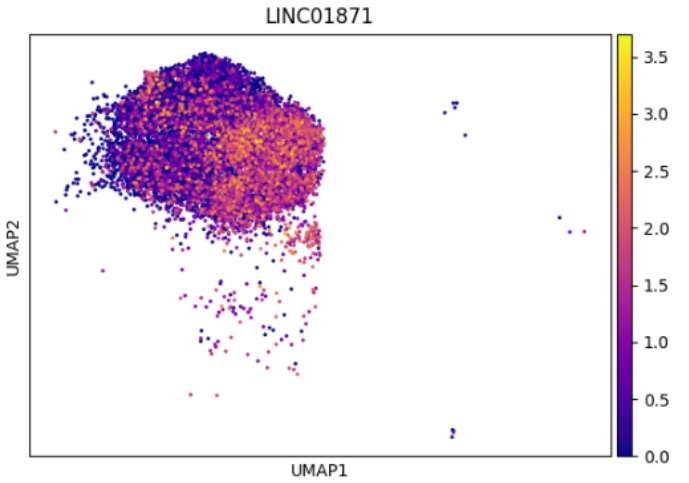

**Zero-preserving imputation of single-cell RNA-seq data**

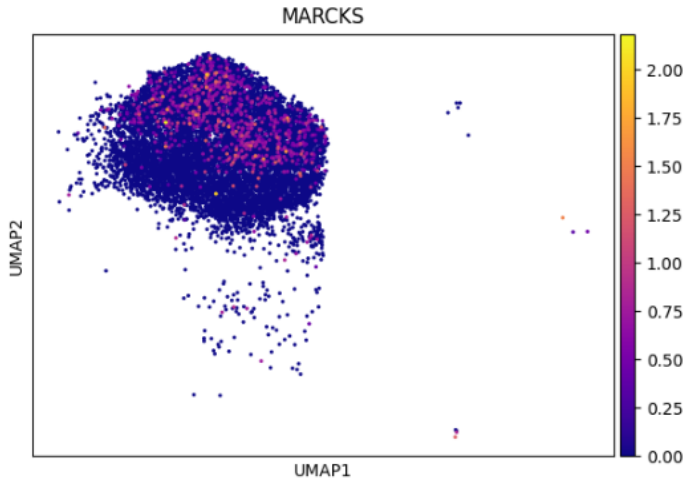
George C. Linderman, Jun Zhao, Manolis Roulis, Piotr Bielecki, Richard A. Flavell, Boaz Nadler & Yuval Kluger

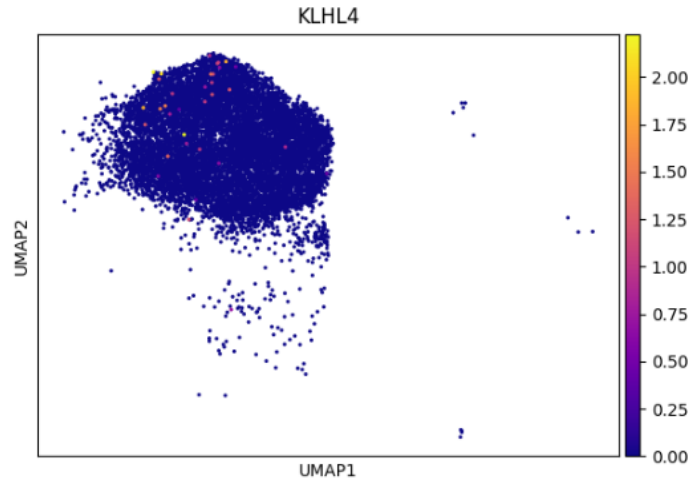
*Nature Communications* **13**, Article number: 192 (2022) | Cite this article

# Imputation visualization example

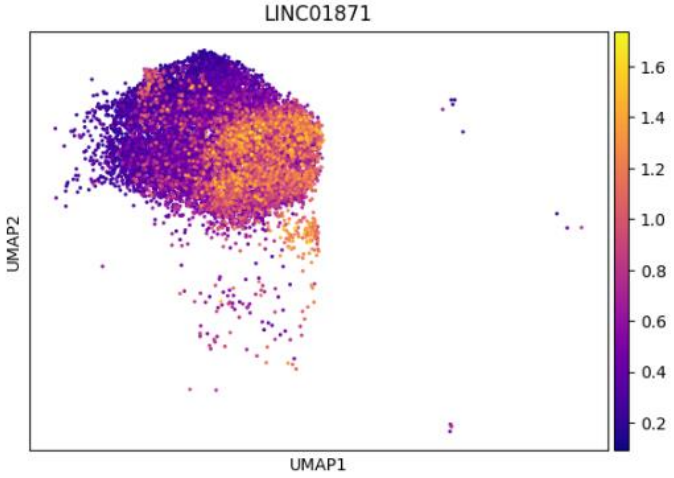
Without imputation

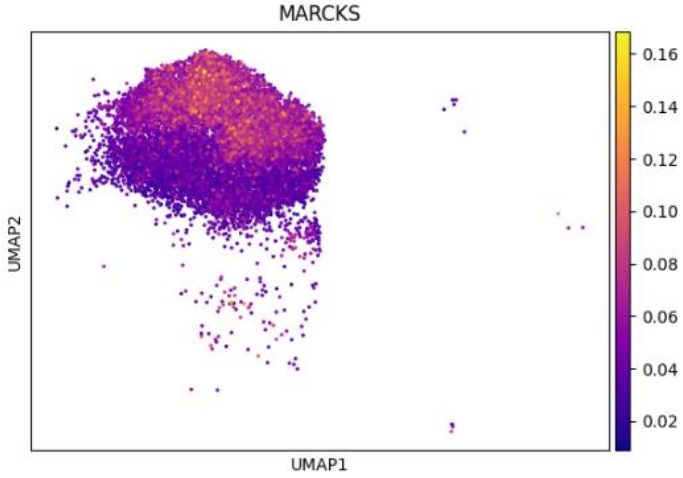


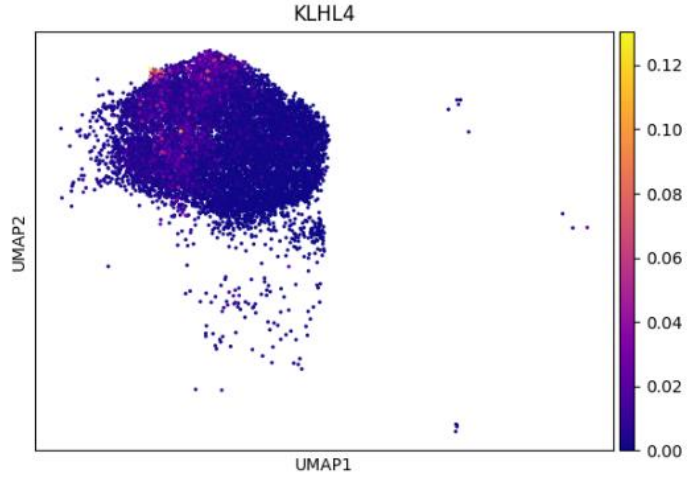




With ‘magic‘ imputation

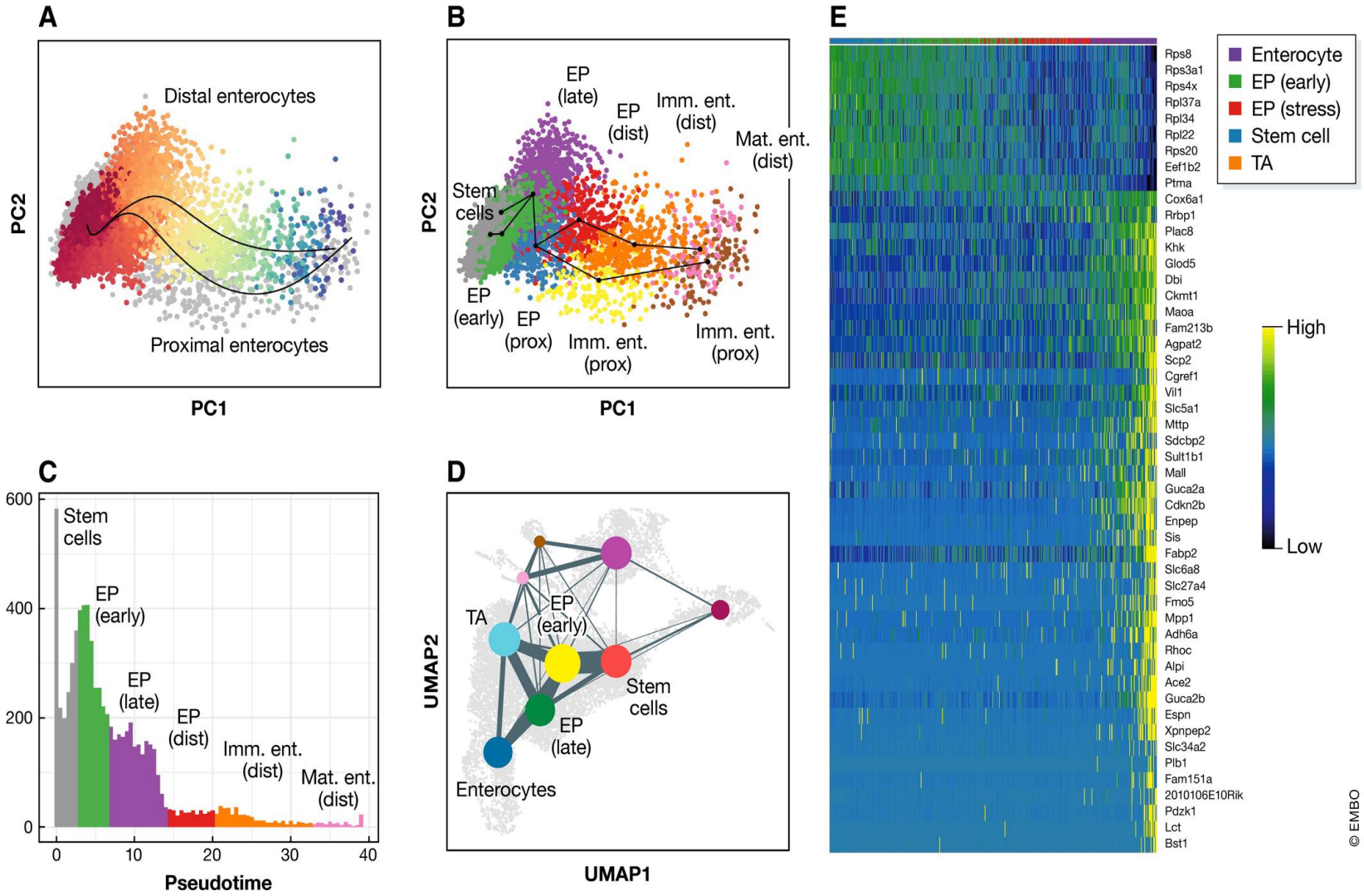


# Trajectory inference

- Discrete classification of cells (clusters) not always appropriate
- Assign cells to “pseudotime” – progress through a dynamic process
- Study how cells evolve from one state to another, when/how cell fate decisions are made

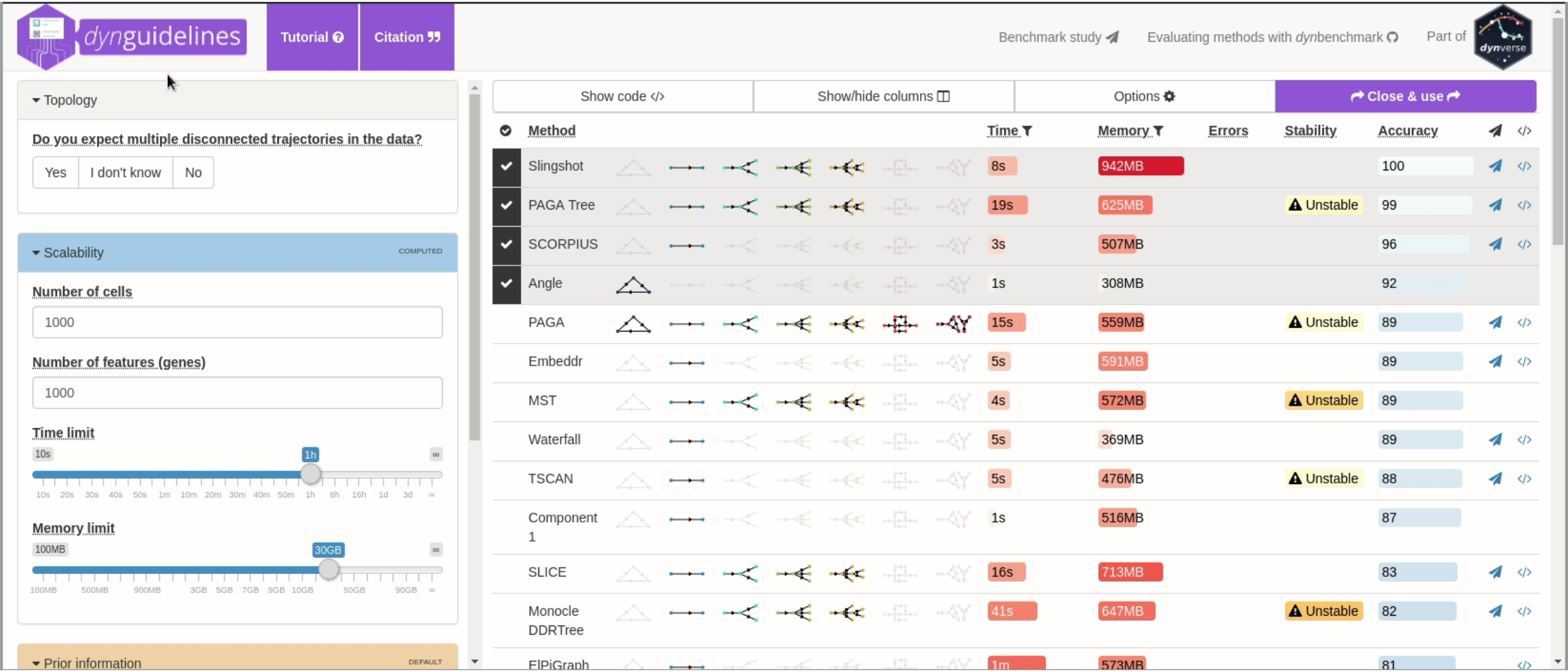





Molecular Systems Biology, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

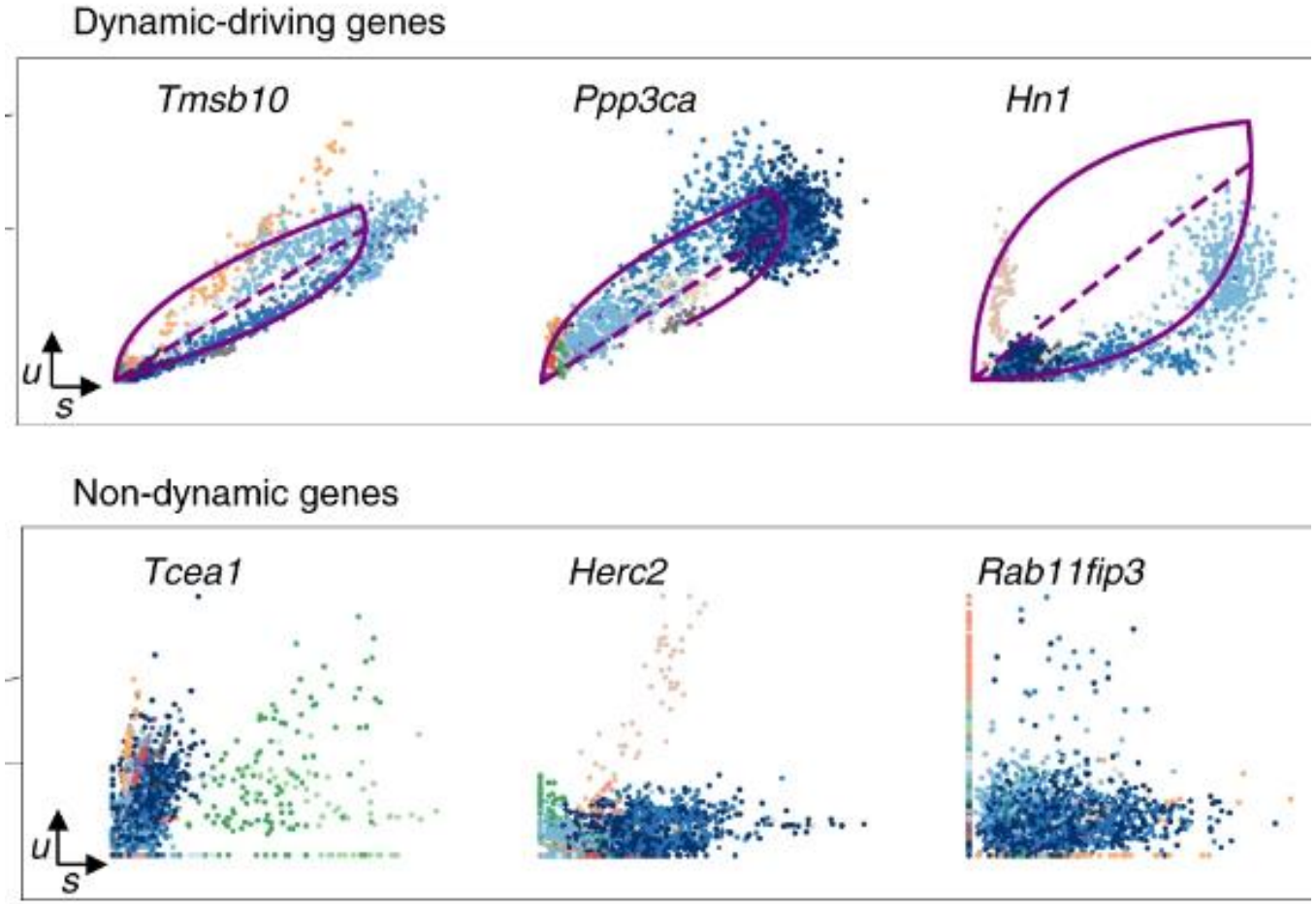
# Many trajectory inference methods!

https://github.com/dynverse/dynguidelines.git: interactive interface to help you choose from > 60 methods



# RNA velocity

- Uses differences in rates of spliced vs unspliced counts to determine if RNA production is increasing or decreasing.
- Need to parse cellranger output bam files to get spliced/unspliced counts, using velocyto software (https://velocyto.org/velocyto.py/index.html), this produces .loom file
- Then package scVelo models RNA velocity dynamics



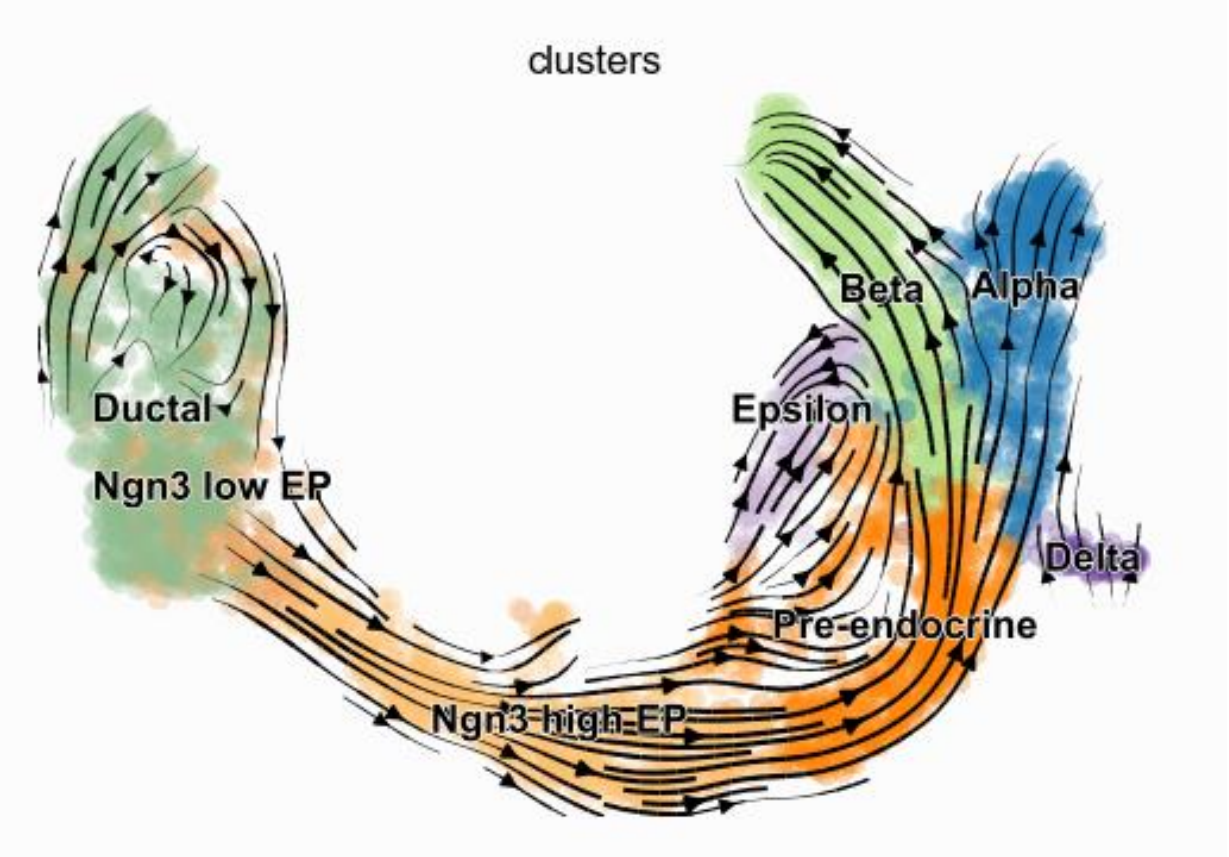

**Generalizing RNA velocity to transient cell states through dynamical modeling**

Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf & Fabian J. Theis

*Nature Biotechnology* **38**, 1408–1414 (2020) | Cite this article

# RNA velocity

- Uses differences in rates of spliced vs unspliced counts to determine if RNA production is increasing or decreasing.
- Need to parse cellranger output bam files to get spliced/unspliced counts, using velocyto software (https://velocyto.org/velocyto.py/index.html), this produces .loom file
- Then package scVelo models RNA velocity dynamics

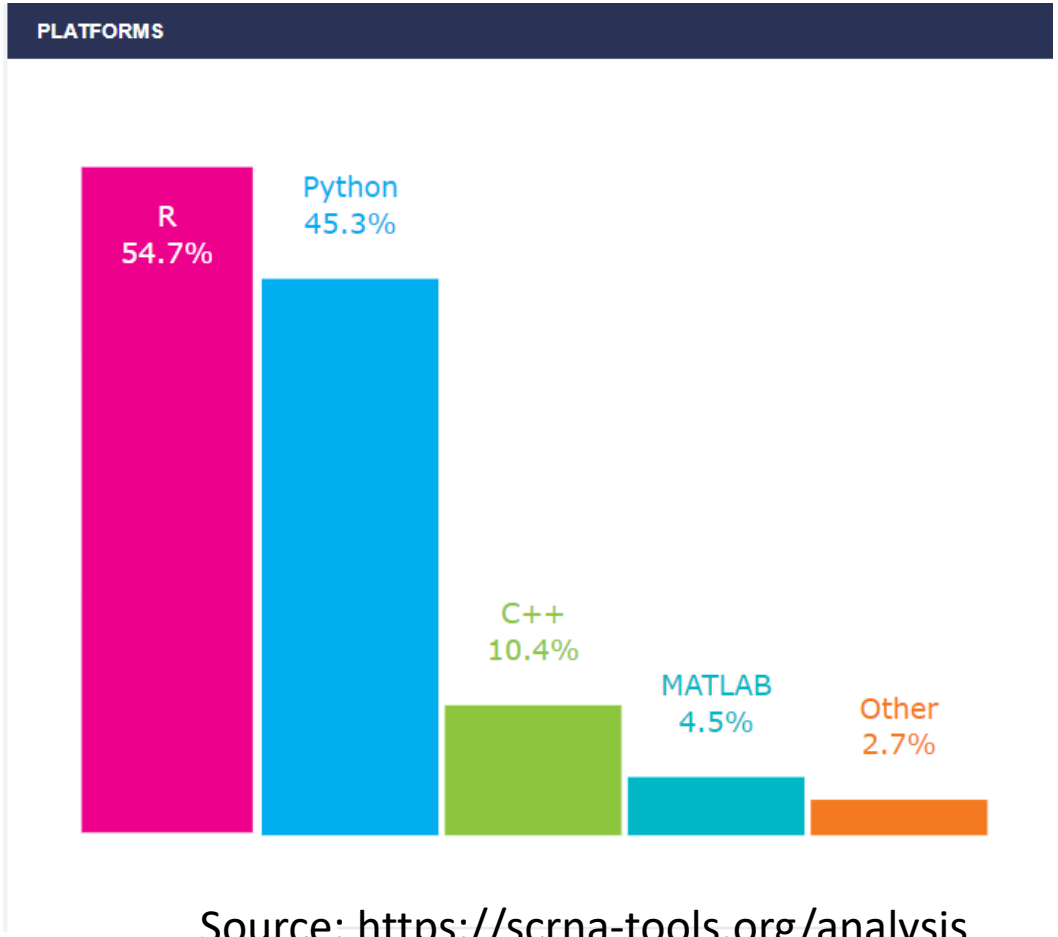



**Generalizing RNA velocity to transient cell states through dynamical modeling**

Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf & Fabian J. Theis

*Nature Biotechnology* **38**, 1408–1414 (2020) | Cite this article

# Why learn scanpy?




Source: https://scrna-tools.org/analysis

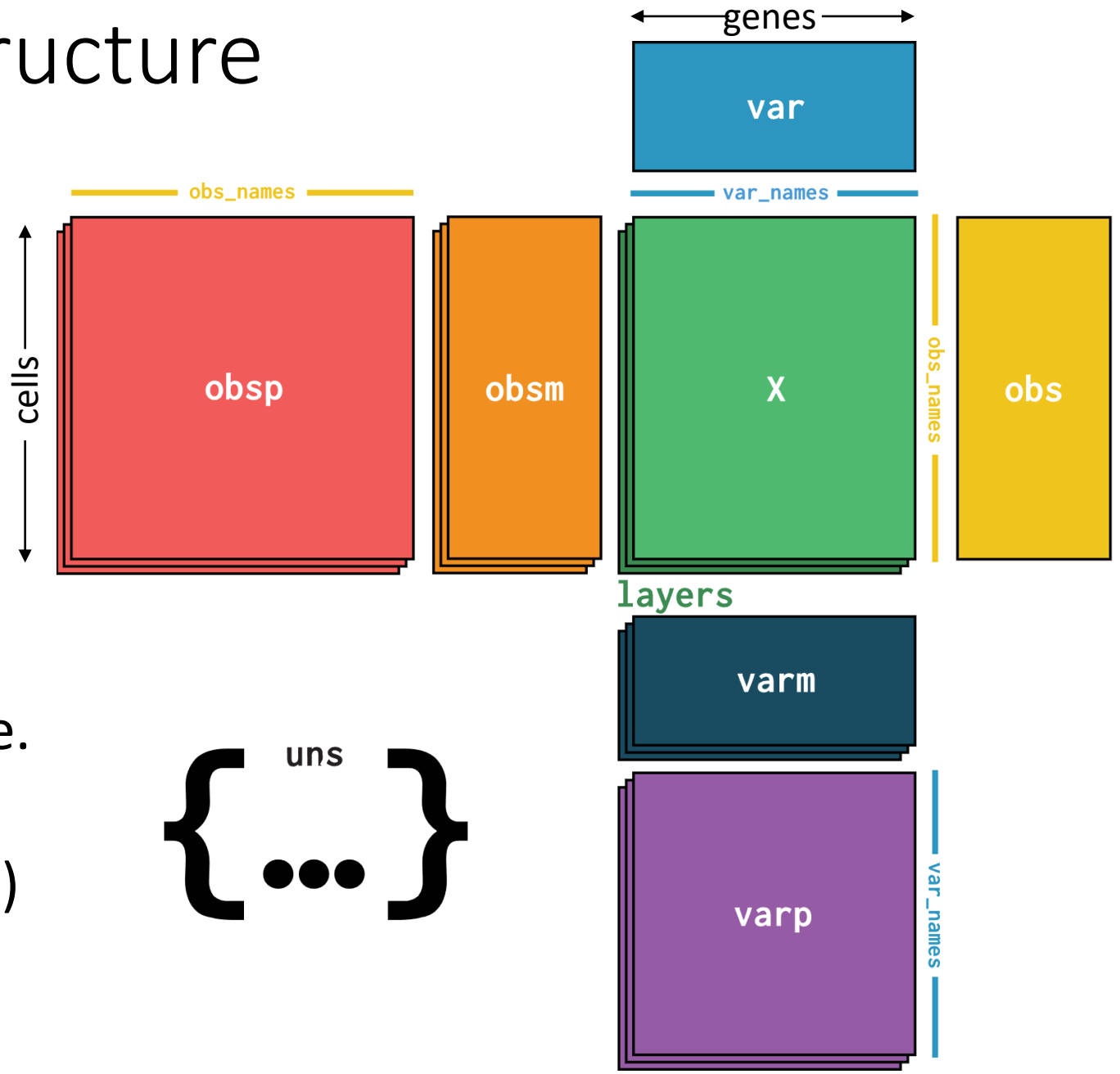
scRNA-tools

- R/Seurat is very popular… but python/scanpy is close behind
- If you want to try new tools being published, you will *need* to use both
- You may prefer python to R!

- Basic analysis / preprocessing can be done on either platform.
- Exercises today will include basic analysis in scanpy instead of seurat

# Scanpy uses AnnData structure

- Rows are cells
  - adata.obs contains cell metadata
- Columns are genes/features
  - adata.var contains gene annotations
- Counts matrix stored in adata.X
- Alternative count matrix in adata.layers[“logX”]
- Projections stored in adata.obsm, i.e. adata.obsm[“X_umap”]
- Unstructured data (e.g., parameters) stored in adata.uns



# Seurat/scanpy object comparison

## Slots

| Slot | Function | Scanpy equivalent |
|---|---|---|
| `assays` | A list of assays within this object | layers |
| `meta.data` | Cell-level meta data | obs |
| `active.assay` | Name of active, or default, assay | X |
| `active.ident` | Identity classes for the current object | |
| `graphs` | A list of nearest neighbor graphs | obsp[‘neighbors’] |
| `reductions` | A list of DimReduc objects | obsm |
| `project.name` | User-defined project name (optional) | |
| `tools` | Empty list. Tool developers can store any internal data fr | |
| `misc` | Empty slot. User can store additional information here | uns |
| `version` | Seurat version used when creating the object | |

# Scanpy – Single-Cell Analysis in Python

Scanpy is a scalable toolkit for analyzing single-cell gene expression data built jointly with anndata. It includes preprocessing, visualization, clustering, trajectory inference and differential expression testing. The Python-based implementation efficiently deals with datasets of more than one million cells.

- Discuss usage on Discourse and development on GitHub.
- Get started by browsing tutorials, usage principles or the main API.
- Follow changes in the release notes.
- Find tools that harmonize well with anndata & Scanpy via the external API and the ecosystem page.
- Check out our contributing guide for development practices.
- Consider citing Genome Biology (2018) along with original references.

### Key Contributors

anndata graph | scanpy graph | ❋ = maintainer

- Isaac Virshup: lead developer since 2019 ❋
- Gökcen Eraslan: developer, diverse contributions ❋
- Sergei Rybakov: developer, diverse contributions ❋
- Fidel Ramirez: developer, plotting ❋
- Giovanni Palla: developer, spatial data
- Malte Luecken: developer, community & forum
- Lukas Heumos: developer, diverse contributions
- Philipp Angerer: developer, software quality, initial anndata conception ❋
- Alex Wolf: lead developer 2016-2019, initial anndata & scanpy conception
- Fabian Theis & lab: enabling guidance, support and environment

## News

### Scanpy hits 100 contributors! 2022-03-31

100 people have contributed to Scanpy's source code!

Of course, contributions to the project are not limited to direct modification of the source code. Many others have improved the project by building on top of it, participating in development discussions, helping others with usage, or by showing off what it's helped them accomplish.

Thanks to all our contributors for making this project possible!

### New community channels 2022-03-31

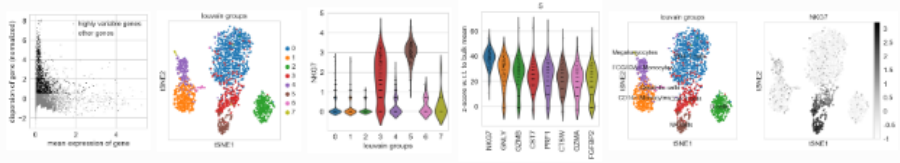
We've moved our forums and have a new publicly available chat!

- Our discourse forum has migrated to a joint scverse forum (discourse.scverse.org).
- Our private developer Slack has been replaced by a public Zulip chat (scverse.zulipchat.com).

# Tutorials

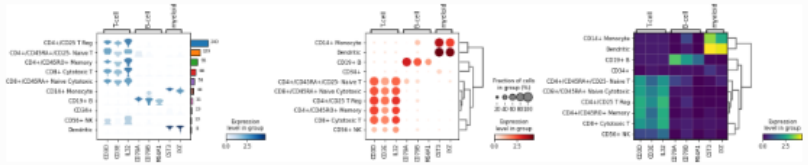
## Clustering

For getting started, we recommend Scanpy's reimplementation → tutorial: pbmc3k of Seurat's [^cite_satija15] clustering tutorial for 3k PBMCs from 10x Genomics, containing preprocessing, clustering and the identification of cell types via known marker genes.



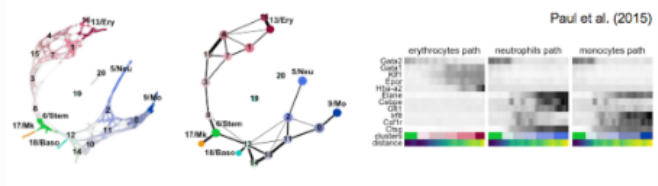
## Visualization

This tutorial shows how to visually explore genes using scanpy. → tutorial: plotting/core



## Trajectory inference

Get started with the following example for hematopoiesis for data of [^cite_paul15]: → tutorial: paga-paul15

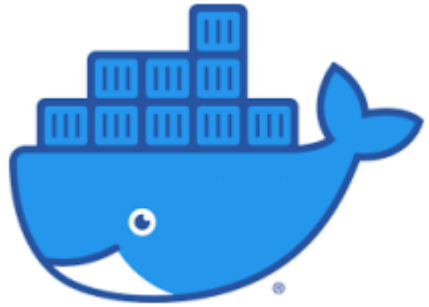



More examples for trajectory inference on complex datasets can be found in the PAGA repository [^cite_wolf19], for instance, multi-resolution analyses of

# Move objects between Seurat/scanpy?

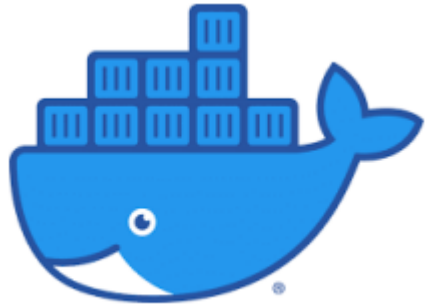
- It is surprisingly difficult..
- There are several tutorials and packages addressing this issue, but none work too well
- You can write scanpy object to h5ad file, R library rhdf5 can parse these files
- More difficult to read Seurat object in python
- Easiest solution: write the matrix/data frames that you need to disk, rather than writing Seurat object

# Docker – overview



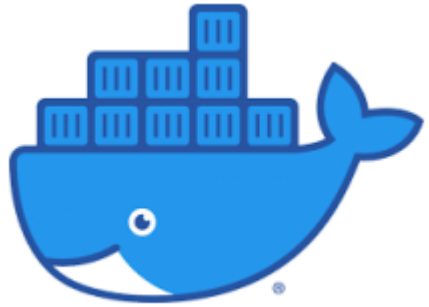
- Basically, docker is a way to run virtual linux machine on the server
- Fully encapsulated
  - Except for directories and ports mapped between docker and the server
- A docker ‘image’ is a copy of an operating system, with software installed
- We have prepared an image for you for this workshop, with lots of scRNA-seq packages installed
- A docker ‘container’ is a virtual machine running an ‘instance’ of the image

# Docker – why??



- Docker (or similar tool) is *essential* for research reproducibility. You can share your docker environment with other researchers
- You can test software in a container without breaking your working environment
- Docker provides a linux ‘playground’ where you can make mistakes. It is completely isolated from the host machine
- You are ‘root’ inside docker container, can install any software you want

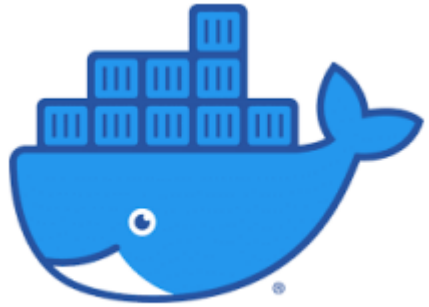
# Docker at BioHPC



- At BioHPC we use the command ‘docker1’ instead of ‘docker’.
- docker1 is a script that calls docker, after making sure that your command cannot harm other user’s data
  - You can only mount directories you own in /workdir, /local, or /home2

- We have beginning and advanced docker workshops (free) on our webpage https://biohpc.cornell.edu/workshops.aspx
  - Basic Docker and Singularity
  - Using docker in BioHPC cloud

# Docker commands reference



- docker1 pull: pull images from Docker hub (image reposito
- docker1 images: show available images
- docker1 run [options] imageName : start a container from an image
- docker1 ps: Show running containers
- docker1 exec [options] imageID command: run a command inside a container
- docker1 stop containerID: stop a container
- docker1 rm containerID: remove a container
- docker1 rmi imageID: remove an image

- docker1 [command] –help: show options for command

# Week 4 exercise: scanpy analysis

Instructions here:
https://github.com/bixBeta/scRNA-WS24/blob/main/Lessons/Week4.md

Thank you! We still have office hours this week if you need help or have questions.

Please fill out our survey when we send it out ☺