

Current Trends in Bioinformatics

A brief introduction to analysing scRNA-seq data

Dave Tang

2025-11-17

Why study scRNA-seq?

If you ask ChatGPT:

Why should people be interested in single cell RNA sequencing? Output just the top 5 reasons as concise bullet points.

- **Reveals cellular heterogeneity:** uncovers differences between individual cells that bulk RNA-seq averages away.
- **Identifies rare cell types/states:** detects transient or low-frequency populations important in development and disease.
- **Maps differentiation and lineage trajectories:** reconstructs how cells change over time or respond to stimuli.
- **Improves understanding of disease mechanisms:** pinpoints which specific cell populations drive pathology.
- **Guides targeted therapeutics:** enables cell-type-specific drug discovery and biomarker identification.



Tip

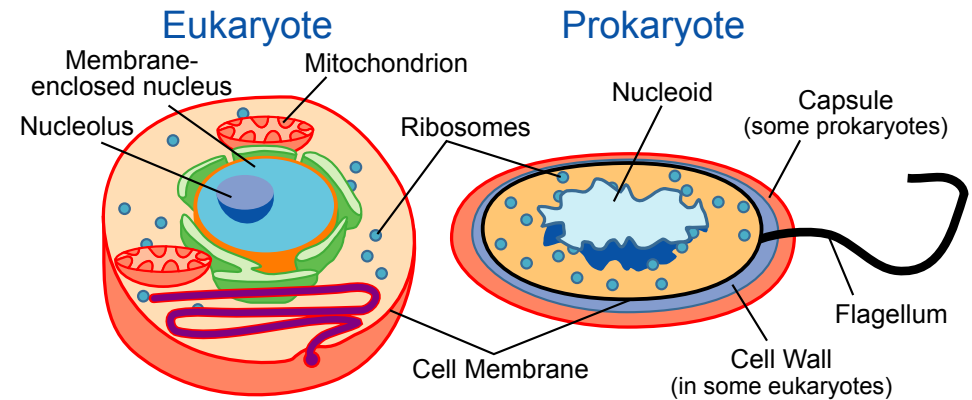
While this is all true, the answer is more fundamental!

Why study cells?

The late Sydney Brenner, who won the **Nobel Prize in 2002** for discoveries concerning genetic regulation of organ development and programmed cell death, said the following in their Nobel Lecture on **December 8, 2002**:

So genocentric has modern biology become that we have forgotten that the **real units of function and structure in an organism are cells and not genes.**

The genome has gives us the inventory of gene loci and we must now get on to the discovery of the actions of their products and how these are integrated in the physiology of cells.



If you have taken biology in school and/or university, you will remember that **cells are the basic units of life, so understanding how they work is fundamental to understanding biology.**

Nothing smaller is alive; while cells contain many complex parts (organelles, proteins, DNA), none of these components can survive or function independently as living entities.

Why study scRNA-seq?

In the same Nobel Lecture in 2002, Sydney Brenner also said:

First, we will need to define all of the non-contingent states of gene expression in an organism, which is **proposed as the correct way of defining a cell type**.

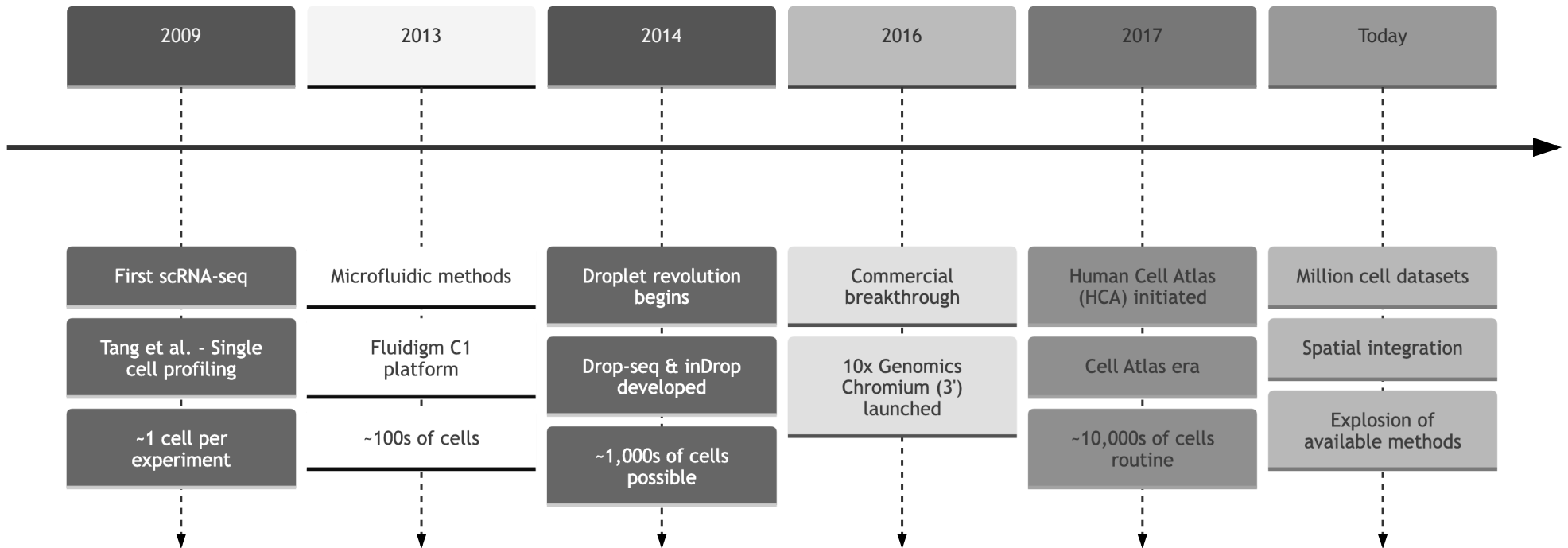
- By “non-contingent states” Brenner meant states or conditions that are not dependent on something else; they exist or are true independently, without requiring other conditions to be met first.
- **We should study cells by their gene expression because that is one way of defining cell types and one way to study them.**

CellMap is seen as a map in many dimensions; it is at once **a map of the cells in the organism onto which are projected the map of instantiations, as well as a map of the molecules in the cell**.


From Sequences and consequences ([Brenner 2010](#)):

I first thought about Cellmap in late 1999 walking on the beach at La Jolla Shores and wrote it as a project in December 2000, but nothing more came of it.

Evolution of single cell RNA



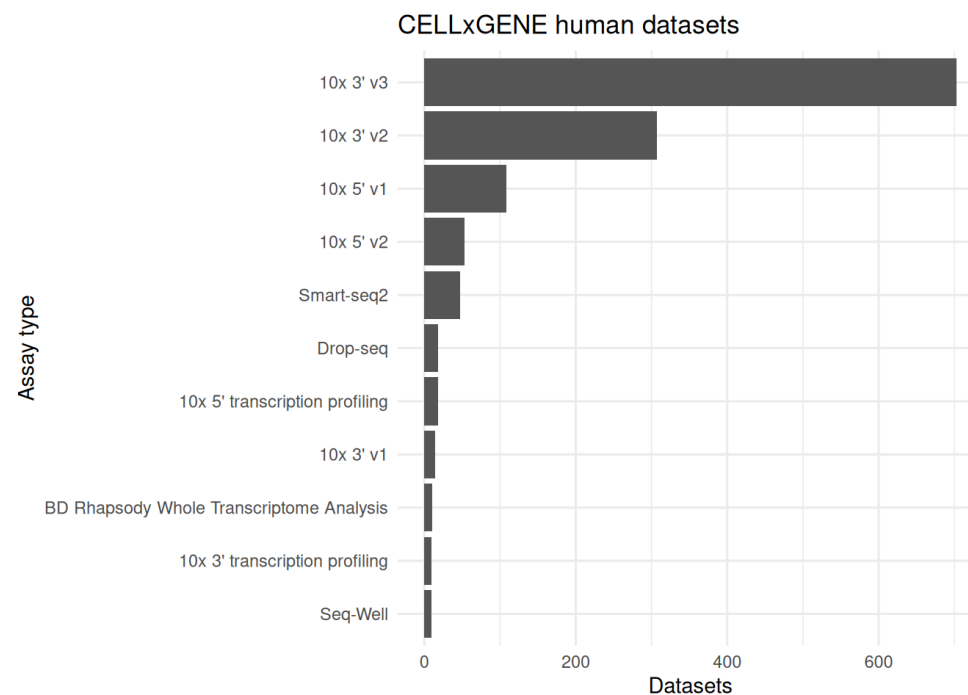
CELLxGENE



Discover the mechanisms of human health

Download and visually explore data to understand the functionality of human tissues at the cellular level with Chan Zuckerberg CELL by GENE Discover (CZ CELLxGENE Discover).

UNIQUE CELLS	DATASETS	CELL TYPES
140.1M	1994	1071



* CELLxGENE human dataset numbers are based on the latest stable Census release: 2025-01-30.

- **CELLxGENE** data is under a **CC-BY 4.0 license**, meaning users can download, share, and use their data without restriction beyond providing attribution to the original data contributor(s).
- There are **Formatting Requirements**, which means datasets have consistent metadata.
- Provides efficient **computational tooling** to access, query, and analyse single-cell RNA data.

Single-cell best practices

Expert Recommendation | Published: 31 March 2023

Best practices for single-cell analysis across modalities

[Lukas Heumos](#), [Anna C. Schaar](#), [Christopher Lance](#), [Anastasia Litinetskaya](#), [Felix Drost](#), [Luke Zappia](#),
[Malte D. Lücken](#), [Daniel C. Strobl](#), [Juan Henao](#), [Fabiola Curion](#), [Single-cell Best Practices Consortium](#),
[Herbert B. Schiller](#) & [Fabian J. Theis](#) 

[Nature Reviews Genetics](#) **24**, 550–572 (2023) | [Cite this article](#)

272k Accesses | **803** Citations | **358** Altmetric | [Metrics](#)

- Heumos et al. (2023) is available as a [Single-cell best practices book](#) that:
 - Aims to guide both beginners and experienced professionals in best practices of single-cell sequencing analysis.
 - The recommendations presented are grounded in external benchmarks and reviews wherever possible, ensuring the approaches taught are both effective and reliable.



Tip

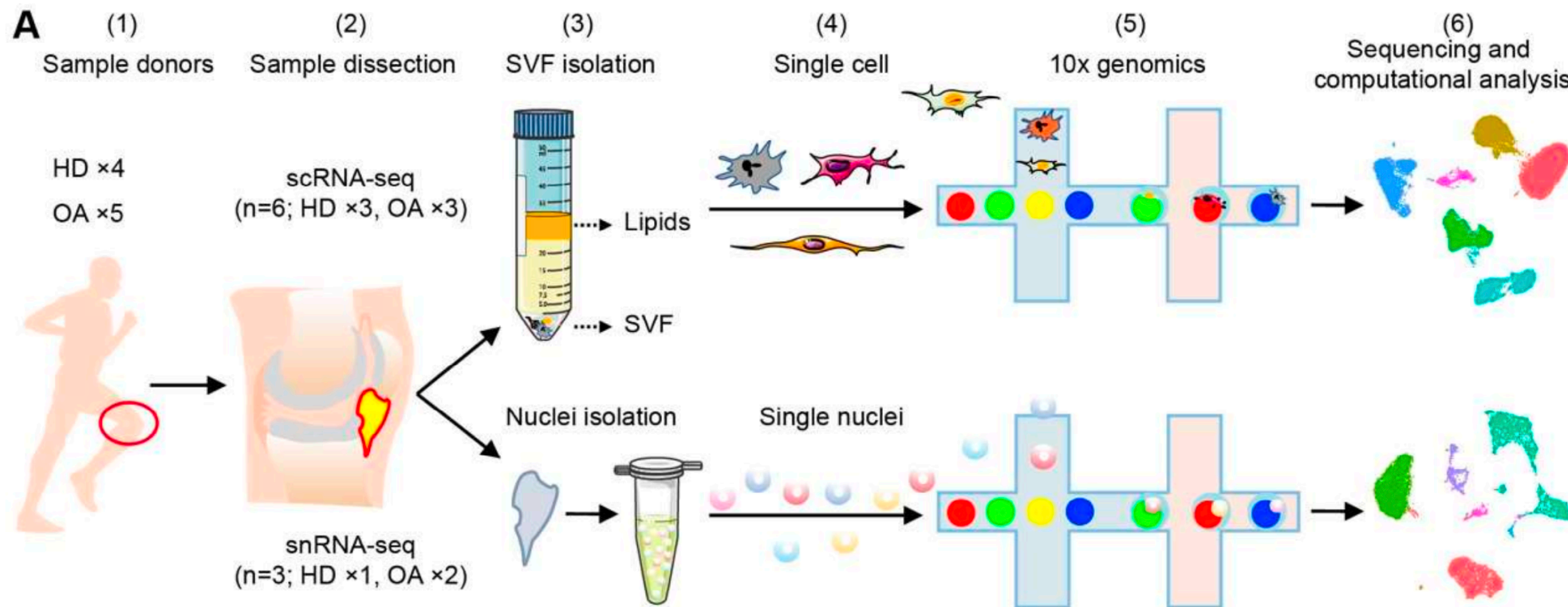
If you want to analyse single cell data, refer to and read the [Single-cell best practices](#) first!

Single-cell atlas of human infrapatellar

- Single cell and single nuclei transcriptomic profiling were performed on infrapatellar fat pad (IPFP) and synovium cells from four healthy donors and five OA patients ([Tang et al. 2024](#)).
- Cell Ranger (Version 6.0.2) was used to demultiplex the reads, which was followed by extraction of the cell barcodes and unique molecular identifiers (UMIs). Filtered matrix files are available for download.
- Raw FASTQ files were also downloaded using [nf-core/fetchngs](#) and will be processed using Cell Ranger 9.0.1 with automatic cell typing.

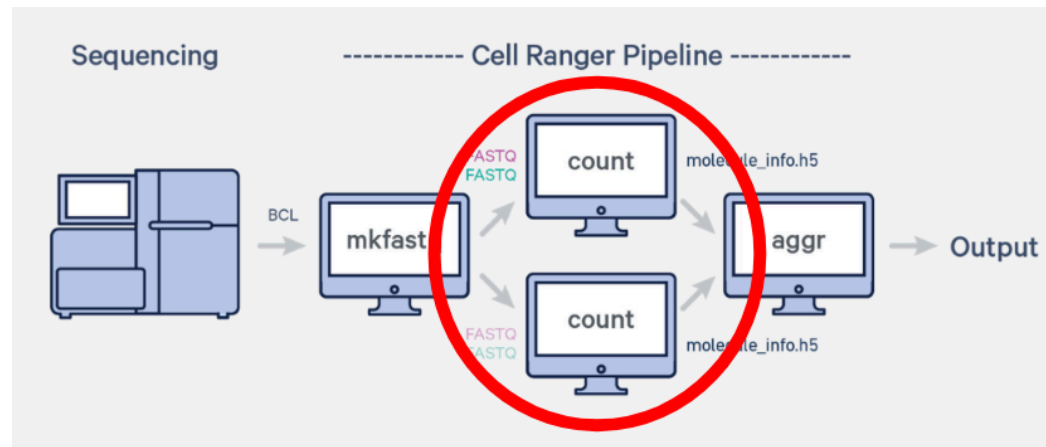
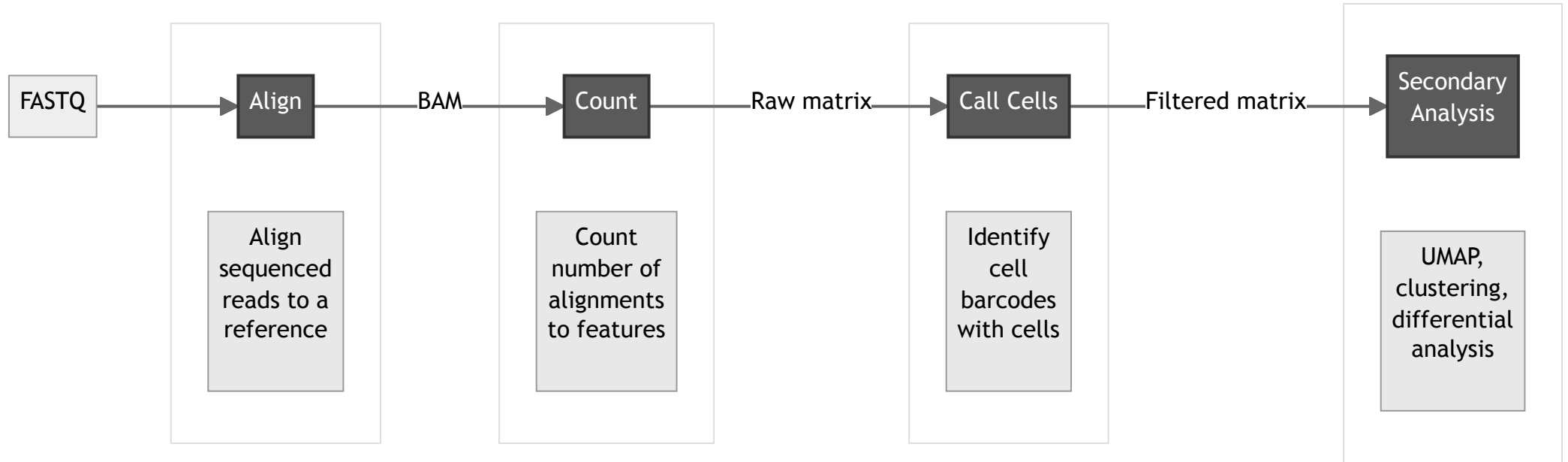
sample_id	sample_name	tissue	cell_type	state
GSM6685302	sc_healthy_1	joint	IPFP/synovium cell	healthy
GSM6685303	sc_healthy_2	joint	IPFP/synovium cell	healthy
GSM6685304	sc_healthy_3	joint	IPFP/synovium cell	healthy
GSM6685305	sc_OA_1	joint	IPFP/synovium cell	osteoarthritis
GSM6685306	sc_OA_2	joint	IPFP/synovium cell	osteoarthritis
GSM6685307	sc_OA_3	joint	IPFP/synovium cell	osteoarthritis

Single-cell atlas of human infrapatellar

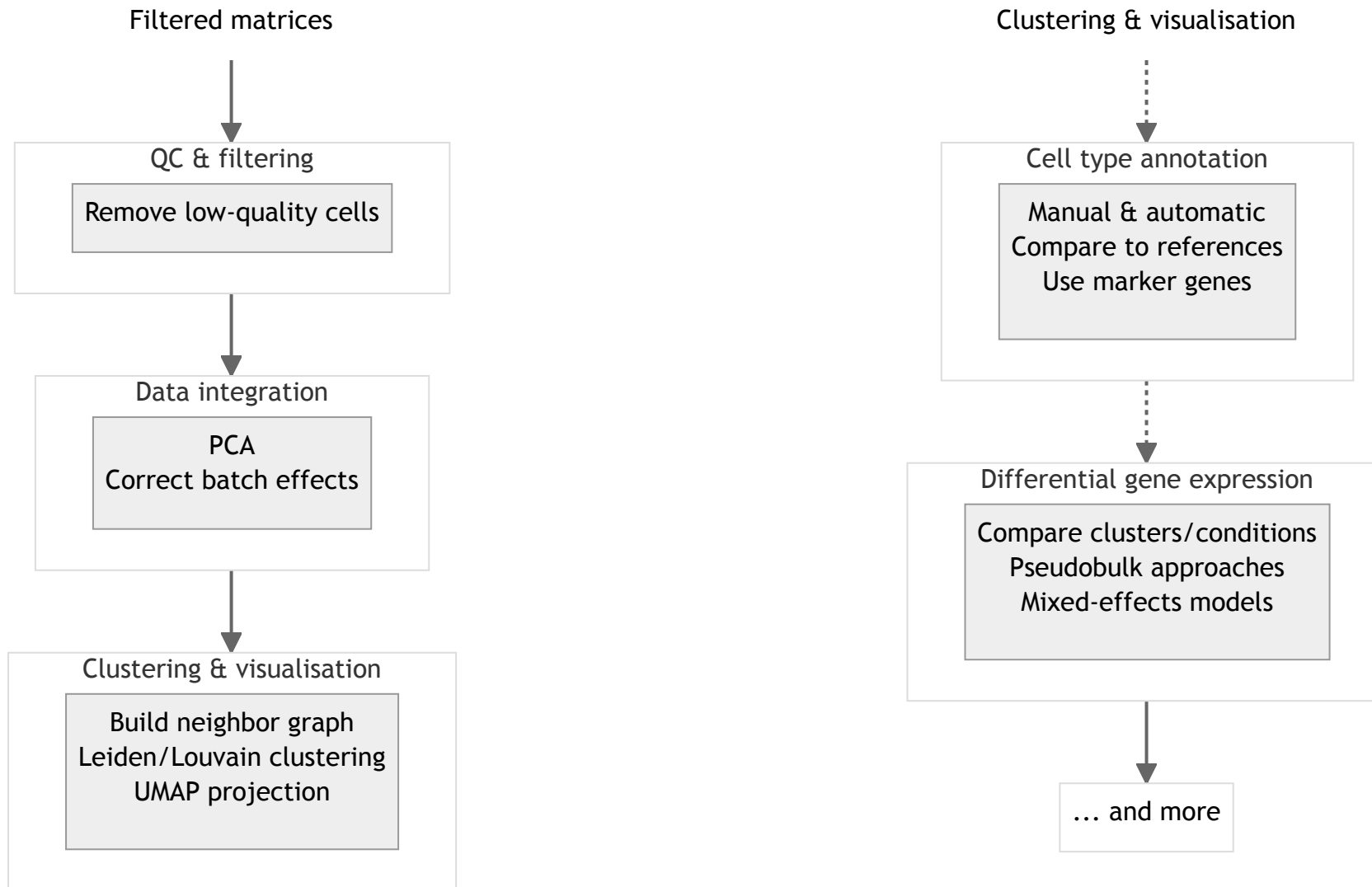


- Schematic workflow of the experimental strategy. IPFP and synovium biospecimens were dissected freshly for stromal vascular fraction (SVF) isolation or frozen immediately for nuclei isolation. Cells or nucleus isolated from IPFP and synovium were subjected to droplet-based scRNA-seq or snRNA-seq, respectively.

Cell Ranger pipeline



Secondary Analysis



Analysis frameworks

From [Single-cell best practices](#):

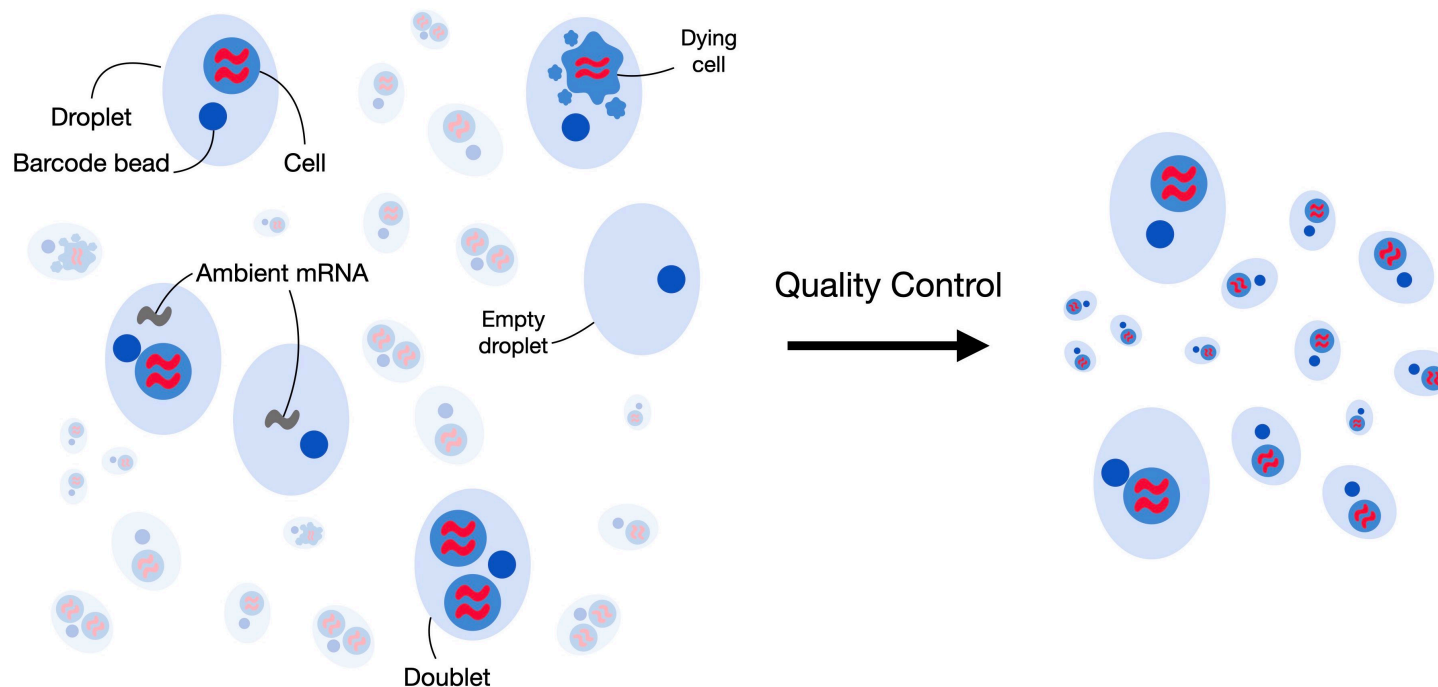
There are three main ecosystems for single-cell analysis, the Bioconductor and Seurat ecosystems in R and the Python-based scverse ecosystem.

A common question from new analysts is which ecosystem they should focus on learning and using? While it makes sense to focus on one to start with, and a successful standard analysis can be performed in any ecosystem, we promote the idea that competent analysts should **be familiar with all three ecosystems and comfortable moving between them.**

This approach allows analysts to use the best-performing tools and methods regardless of how they were implemented.

- Python: **Scanpy** as the main analysis framework and **AnnData** as the main data structure; see [scverse](#).
- R: **Seurat**, which was one of the first packages for analysing scRNA-seq data
- R: Bioconductor has many packages and data structures supporting scRNA-seq analysis; see [Orchestrating Single-Cell Analysis with Bioconductor](#).

Quality Control

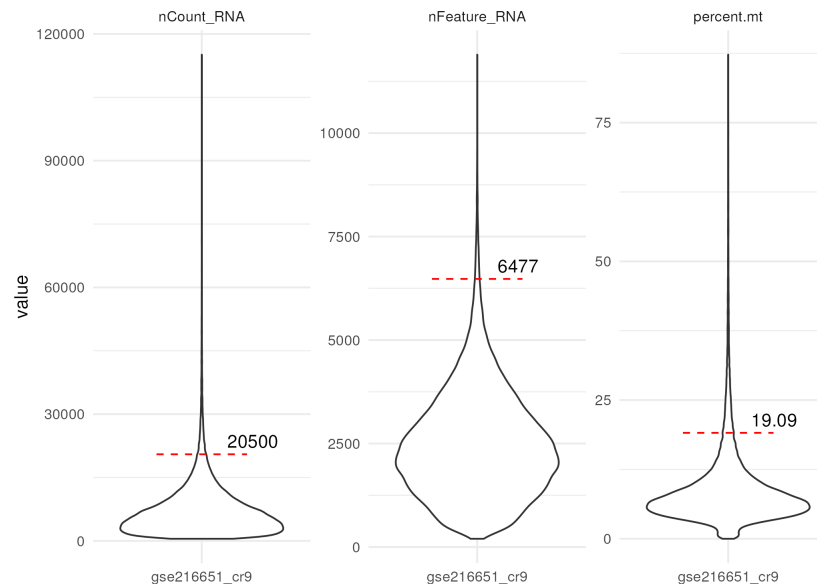


Goal

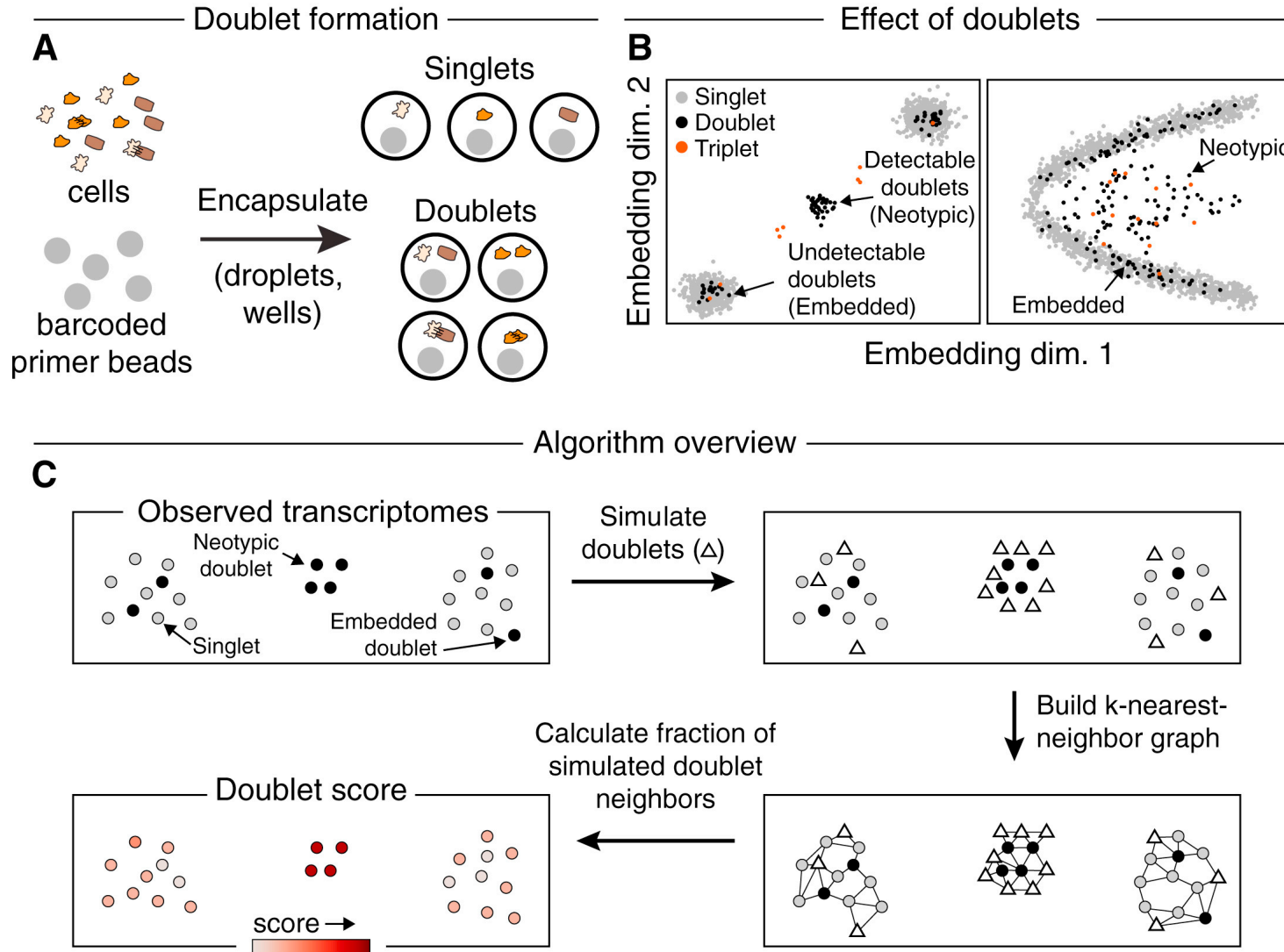
Remove damaged or dying cells (low-quality cells), empty droplets without a cell, droplets with multiple cells (doublets) before conducting the downstream analyses!

Quality Control metrics

- Three metrics are very useful for filtering out problematic droplets: 1) **nCount_RNA** total expression of a cell, 2) **nFeature_RNA** total number of genes detected in a cell, and 3) **percent.mt** contribution of mitochondrial genes to total expression.
 - Low-quality barcodes or empty droplets will often have very few genes.
 - Doublets may exhibit an aberrantly high gene count.
- The percentage mitochondrial can be indicative of low-quality/dying cells due to extensive mitochondrial contamination.
- **Single-cell best practices** recommends the use of the Median Absolute Deviation (MAD), shown as red dashed lines.
- The actual filtering used in the paper was: “genes>6000, genes<200, or >25% genes mapping to mitochondrial genome”.



Doublet detection



Normalisation

- In practice **log-normalisation**, which scales counts to a common size (e.g., 10,000 counts/cell) and performs log-transformation, works well.
 - It is the default in many pipelines.
 - Works well for most applications.
 - Fast and interpretable.
- Choosing a method that is compatible with different downstream tasks is important when building a workflow.
- The normalisation method used in Tang et al. (2024) was a more sophisticated approach called SCTransform.



Tip

Different cells have different total counts and the goal is to make cells comparable.

Feature selection and dim

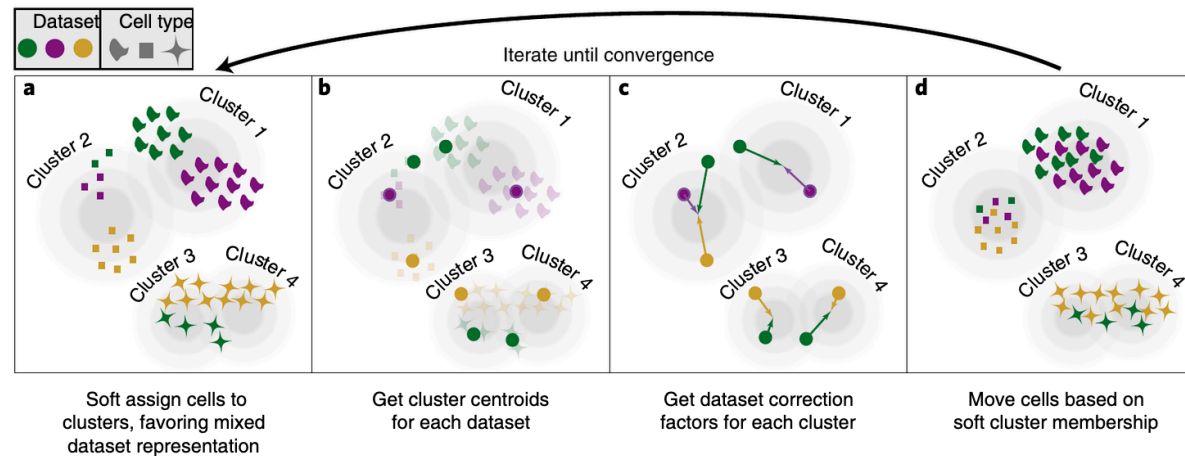
- An scRNA-seq count matrix (32,815 genes by ~ 10,000 cells) might be 80–95% zeros, sometimes even more, i.e., a sparse matrix.

$$\begin{array}{c}
 \text{Gene 1} \\
 \vdots \\
 \text{Gene m}
 \end{array}
 \begin{array}{c}
 \text{Cell 1} \quad \cdots \quad \text{Cell n} \\
 \left[\begin{array}{cccc}
 x_{11} & x_{12} & \cdots & x_{1n} \\
 x_{21} & x_{22} & \cdots & x_{2n} \\
 \vdots & \vdots & \ddots & \vdots \\
 x_{m1} & x_{m2} & \cdots & x_{mn}
 \end{array} \right]
 \end{array}$$

- scRNA-seq data is sparse because **cells express only a subset of genes AND the technology fails to detect many low-abundance transcripts.**
- Most genes are uninformative and the standard is to analyse only Highly Variable Genes (Seurat's default is 2,000); this is also known as feature selection.
- Although 2,000 genes is much less, it is still computational intensive when there are many cells and Principal Component Analysis (PCA) is typically conducted to “compress” the data to 50 principal components (Seurat's default), i.e., this is known as a **low-dimensional embedding of the cells.**

Data integration

- Harmony ([Korsunsky et al. 2019](#)) is a computational method for integrating multiple single-cell datasets (e.g., different batches, donors, technologies) by correcting for batch (or other covariate) effects.
 - Harmony uses and corrects the lower-dimensional embedding, i.e., PCA, rather than using or modifying raw gene expression.
- A benchmark of batch correction methods used in scRNA-seq showed that Harmony is the only method that consistently performs well ([Antonsson and Melsted 2025](#)).

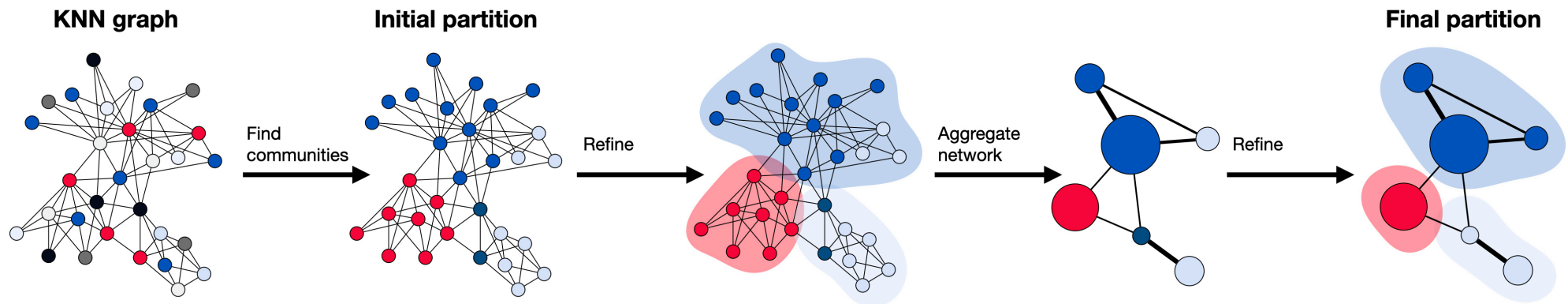


Tip

This Harmony-corrected embedding is what is used in later steps! Actual gene expression data, like the normalised counts, is not used!

Unsupervised clustering

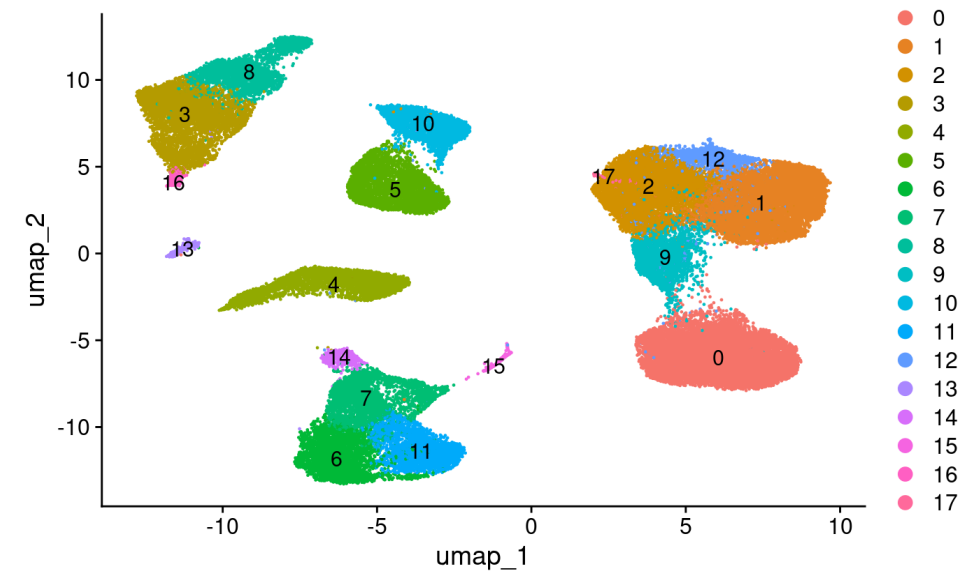
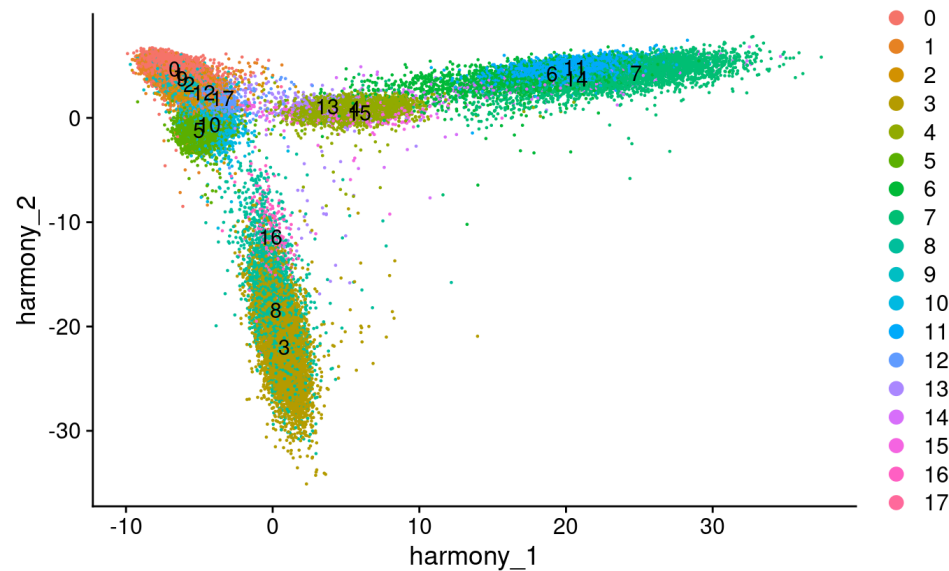
- Unsupervised clustering in scRNA-seq is a useful approach to identify distinct cell populations or states without prior knowledge of their identities.
- A k-nearest neighbor (kNN) graph is constructed, which essentially connects cells that have a similar expression pattern of genes.
 - Reminder, the Harmony-corrected embedding is used!



Tip

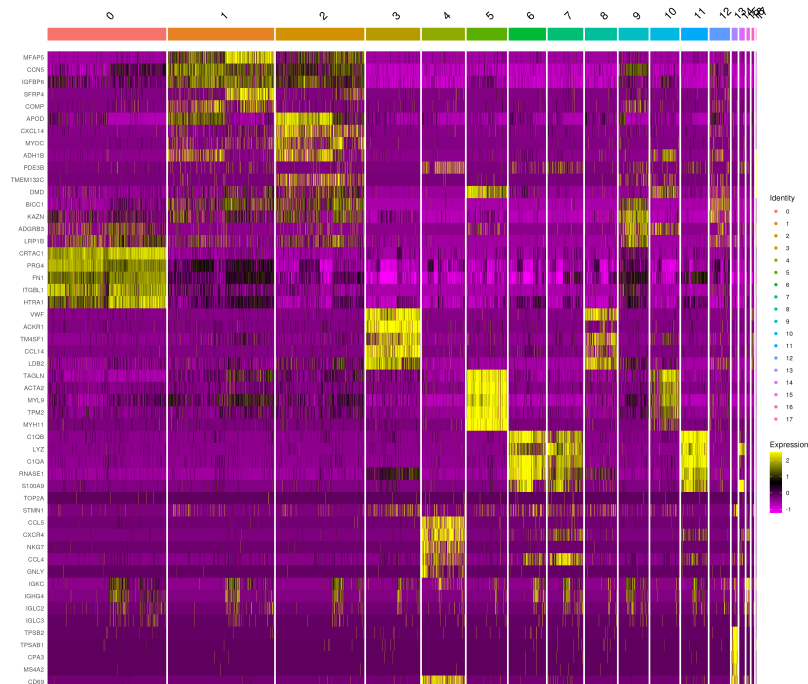
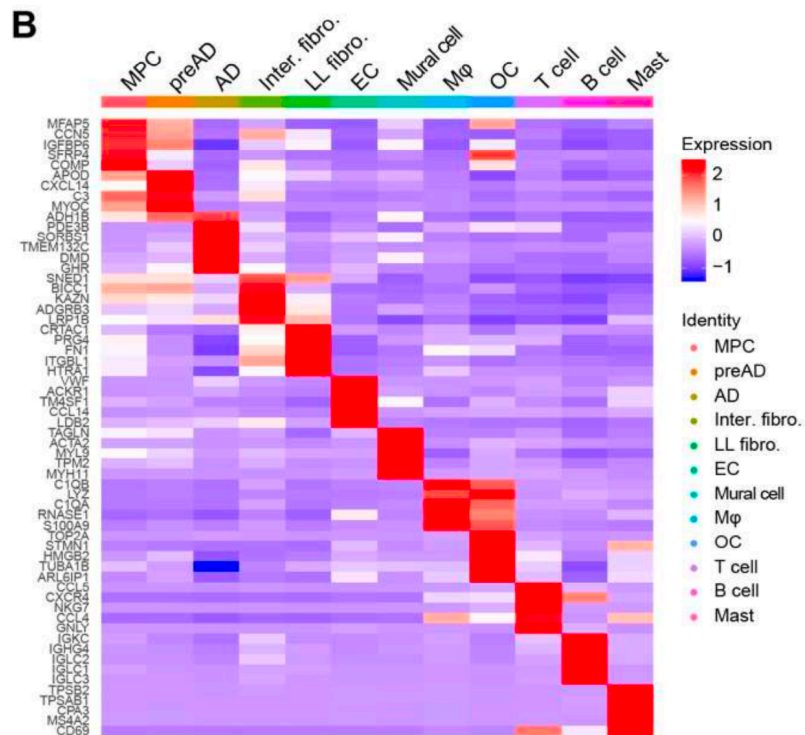
These communities represent groups of cells with similar transcriptional profiles and are often referred to as **cell clusters**.

Non-linear dimensionality reduction



- Uniform Manifold Approximation and Projection (UMAP) is widely used in scRNA-seq analysis for **visualisation and exploratory purposes**.
- UMAP aims to uncover the underlying structure of high-dimensional data (using the Harmony-corrected embedding) by mapping cells into a lower-dimensional space to help identify patterns, clusters, or trajectories.
 - Local relationships are usually preserved, ensuring that cells with highly similar gene expression profiles remain close to each other in the low-dimensional representation. **However, the global structures are often distorted.**

Cell type annotation



Tip

Cell clusters don't mean much until marker genes (or automatic cell typing) identify what they are!

Take-home messages

1. Cells are the basic units of life, so understanding how they work is fundamental to understanding biology.
 - One way to classify cells is by their gene expression.
2. Follow the Single-cell best practices unless there's a good reason not to.
3. For single cell analysis, it's not R versus Python but it's R + Python.
4. Quality control is critical - garbage in, garbage out.
5. It is essential to correct for batch effects; using Harmony is one way.
6. PCA is for analysis; UMAP is for visualisation.
7. Stay updated - the single cell genomics field is rapidly evolving!



References

- Antonsson, Sindri Emmanúel, and Páll Melsted. 2025. “Batch Correction Methods Used in Single-Cell RNA Sequencing Analyses Are Often Poorly Calibrated.” *Genome Research* 35 (8): 1832–41.
- Brenner, Sydney. 2010. “Sequences and Consequences.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1537): 207–12.
- Heumos, Lukas, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, et al. 2023. “Best Practices for Single-Cell Analysis Across Modalities.” *Nature Reviews Genetics* 24 (8): 550–72.
- Korsunsky, Ilya, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. 2019. “Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony.” *Nature Methods* 16 (12): 1289–96.
- Tang, Su'an, Lutian Yao, Jianzhao Ruan, Jingliang Kang, Yumei Cao, Xiaoyu Nie, Weiren Lan, et al. 2024. “Single-Cell Atlas of Human Infrapatellar Fat Pad and Synovium Implicates APOE Signaling in Osteoarthritis Pathology.” *Science Translational Medicine* 16 (731): eadf4590.

Wolock, Samuel L, Romain Lopez, and Allon M Klein. 2019. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data.” *Cell Systems* 8 (4): 281–91.