Autumn School for Single Cell-ers

Day 1 - Introduction to scRNAseq

Miguel M. Santos
2025-10-20

Day 1

October 20, 2025 at 9:30 AM — October 20, 2025 at 10:30 AM

9h30 | Opening

October 20, 2025 at 10:30 AM — October 20, 2025 at 11:00 AM

10h30 | Coffee Break

October 20, 2025 at 11:00 AM — October 20, 2025 at 12:30 PM

11h00 | Introductory analysis w/ Miguel Santos

October 20, 2025 at 12:30 PM — October 20, 2025 at 1:30 PM

12h30 | Lunch Break

October 20, 2025 at 1:30 PM — October 20, 2025 at 3:30 PM

13h30 | Introductory analysis tutorial w/ Miguel Santos

October 20, 2025 at 3:30 PM — October 20, 2025 at 4:00 PM

15h30 | Coffee Break

October 20, 2025 at 4:00 PM — October 20, 2025 at 5:30 PM

16h00 | Data preparation and grouping

Github

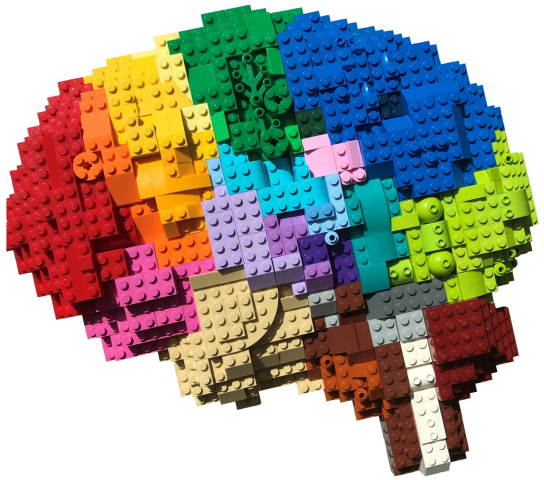
https://github.com/GIMM-BioCode/2025-Autumn-School-for-Single-Cell-ers/tree/main

Datasets

https://drive.google.com/drive/folders/1eKHzXlzMcl2pXIkJnLhivk-zyIrE1Kwn

Whatsapp Group

Building biological insight, one piece at a time



3D organ

bulk RNAseq

single-cell RNAseq

spatial RNAseq

Image credit: @BoXia7

What we will cover in this session

What is single-cell RNA-seq and how does it compare to bulk RNA-seq?

What are some of the typical applications of scRNA-seq?

How are samples typically prepared for scRNA-seq?

What are the differences between some of the most popular protocols and what are their advantages and disadvantages?

What experimental design choices should be considered in scRNA-seq?

What are some of the challenges of scRNA-seq data compared to bulk data?

How to process the data?

- QC, normalisation and scaling, variable features, linear & non linear dimensionality reduction
- Clustering and cell-type identification

RNA-seq - a revolutionary tool for transcriptomics

Profiling of **all transcripts** (in theory) of a given biological sample using **next-generation sequencing**

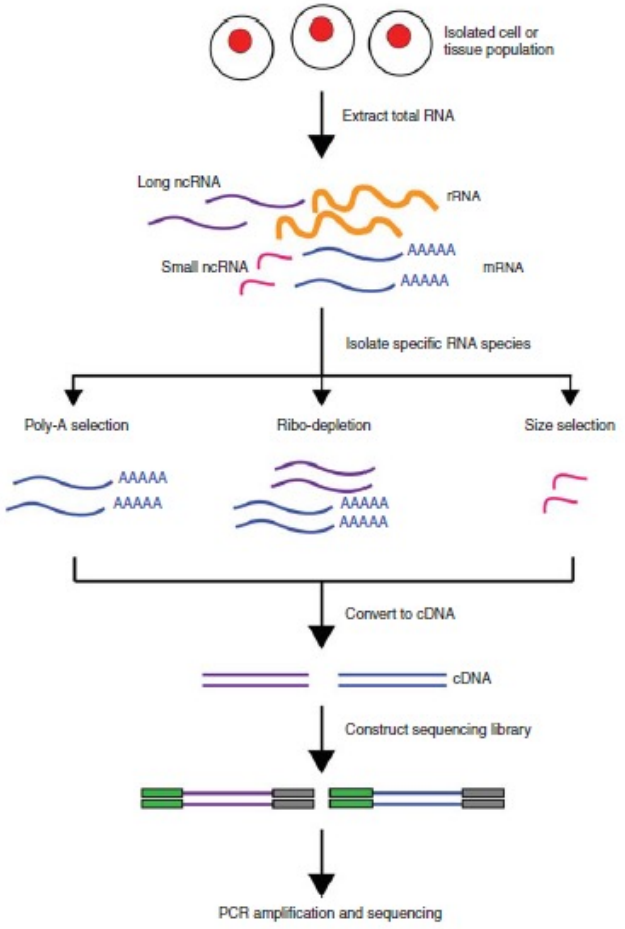
Major breakthrough in late 2000's

Replaced other technologies such as microarrays due to higher coverage and greater resolution

Allowed unbias sampling of all transcripts in a sample, rather than being limited to pre-determined set of transcripts (probes or RT-qPCR)

Initially performed in **bulk** (mixture of cells)

Allowed transcript quantification, isoform detection, annotation of new genes, both in model and non-model organism



Major breakthroughs of (bulk) RNA-seq

Characterise expression signatures between tissues or cell populations in:
1 – Health vs disease,
2 – Wild-type vs mutant
3 – Control vs treated
4 - Comparative Transcriptomics - evolutionary studies, compare tissue samples across different species

Article | Published: 30 May 2008

Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer & Barbara Wold

Nature Methods **5**, 621–628 (2008) | Cite this article

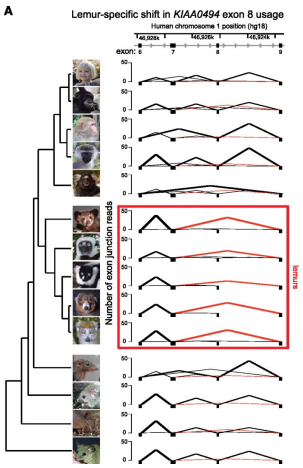
108k Accesses | **12k** Citations | **100** Altmetric | Metrics

REPORTS

The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

UGRAPPA NAGALAKSHMI, ZHONG WANG, KARL WAERN, CHONG SHOU, DEBASISH RAHA, MARK GERSTEIN, AND MICHAEL SNYDER Authors Info & Affiliations

SCIENCE • 6 Jun 2008 • Vol 320, Issue 5881 • pp. 1344-1349 • DOI: 10.1126/science.1158441

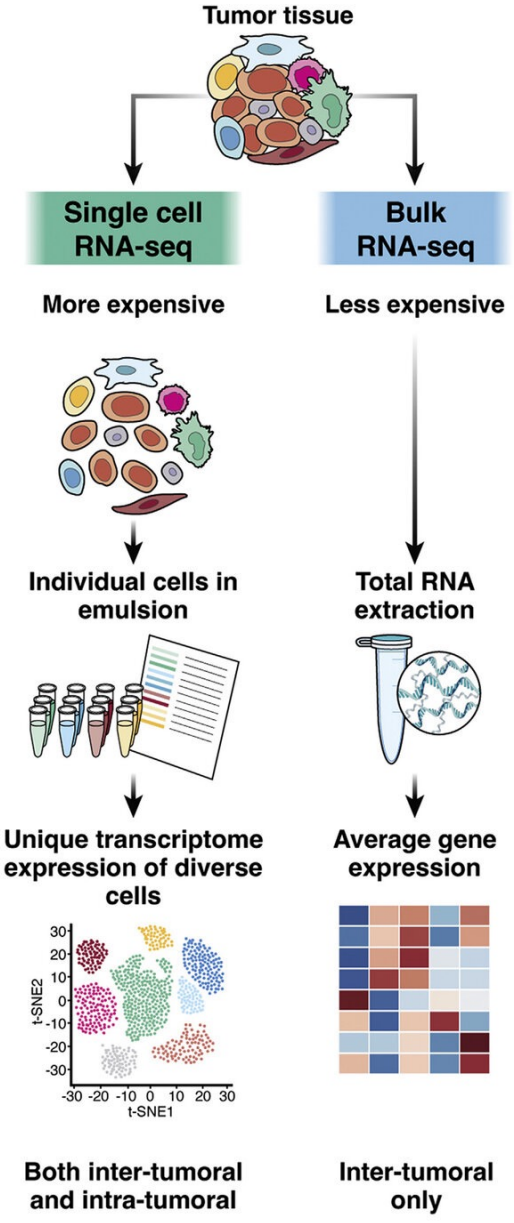


TCGA

Over **20,000 primary cancer** and matched normal samples spanning **33 cancer types**

Bulk RNA-seq does not account for cellular heterogeneity

Cell-specific changes in transcriptome
- Discovery of new or rare cell types
- Differential cell composition between healthy/diseased tissues
- Cell differentiation during development
- One of the most iconic uses: **Cell Atlas**



Average expression level for each gene across cell populations

Editorial | Published: 30 December 2013

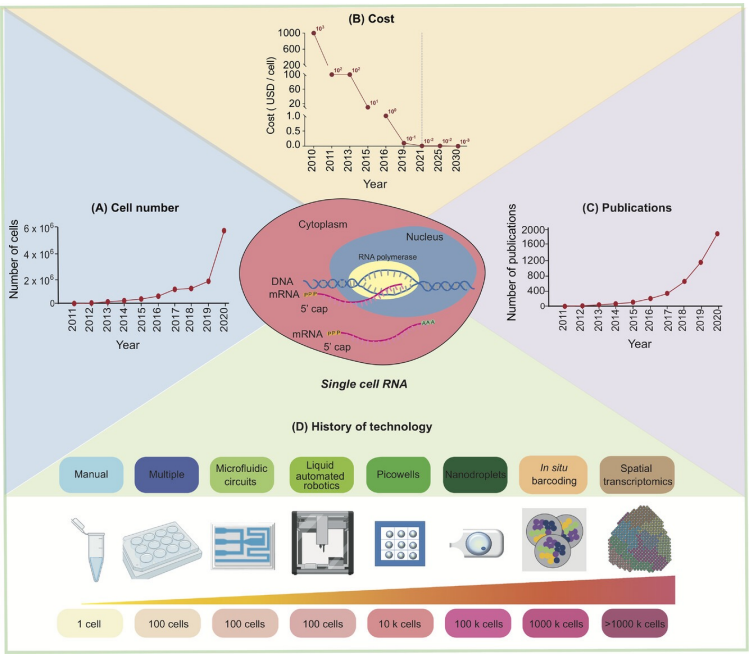
Method of the Year 2013

Nature Methods **11**, 1 (2014) | Cite this article

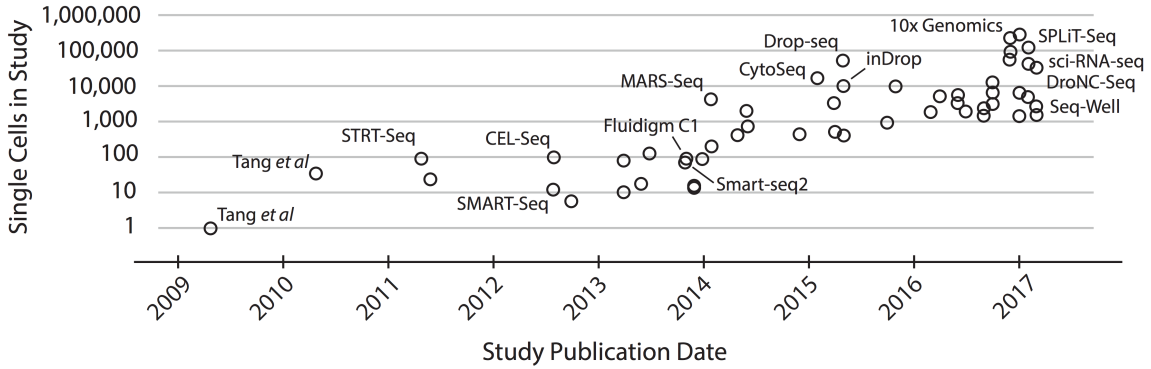
49k Accesses | **138** Citations | **134** Altmetric | Metrics

Methods to sequence the DNA and RNA of single cells are poised to transform many areas of biology and medicine.

From tens to millions of cells in just over a decade



Quantum barcoding, 2M cells per run



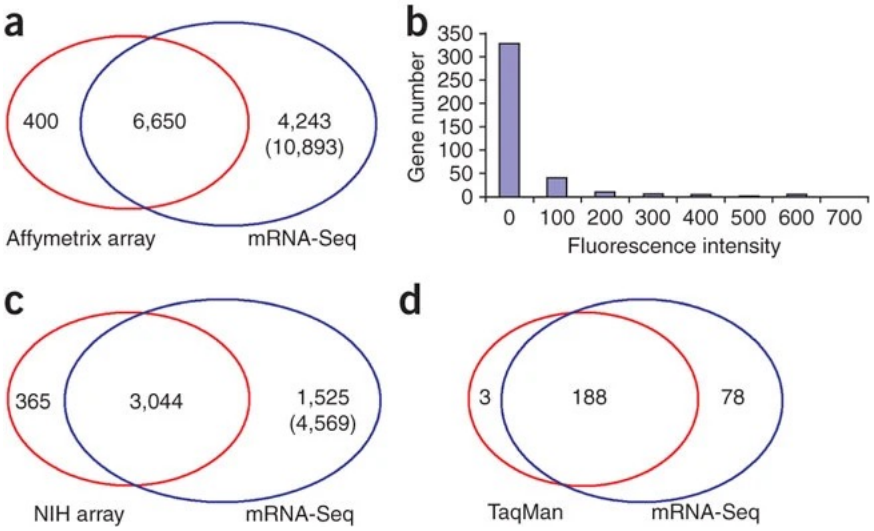
First single cell paper

Article | Published: 06 April 2009

mRNA-Seq whole-transcriptome analysis of a single cell

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao ✉ & M Azim Surani ✉

Nature Methods **6**, 377–382 (2009) | Cite this article



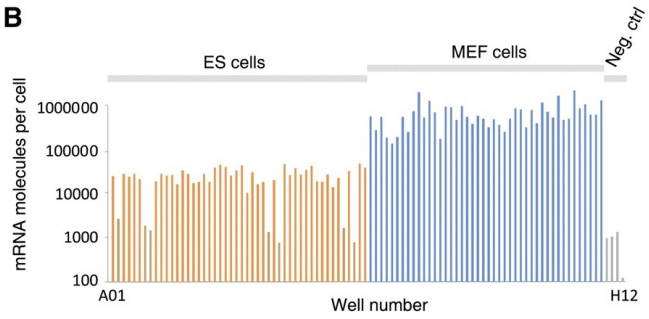
Quantify RNA from very low starting material

But with more genes detected compared to state-of-the-art microarrays

First paper with multiplexing – 96 well plate (2011)

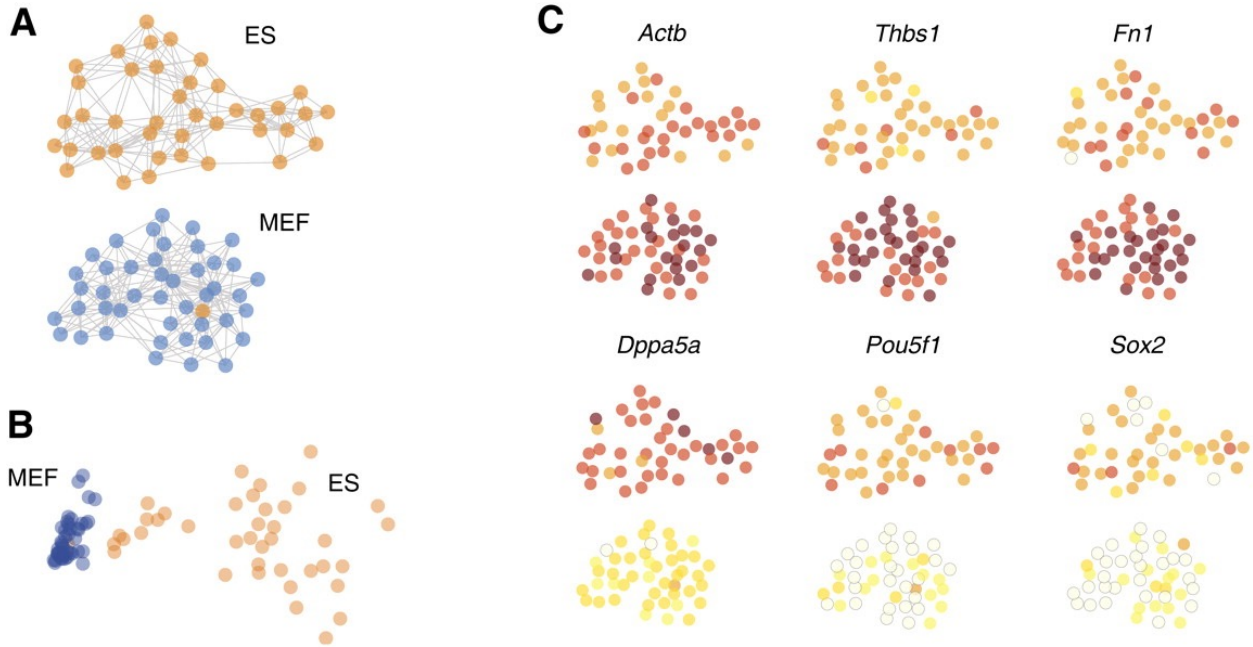
Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq

Saiful Islam[1,4], Una Kjällquist[1,4], Annalena Moliner[2], Pawel Zajac[1], Jian-Bing Fan[3], Peter Lönnerberg[1] and Sten Linnarsson[1,5]



Force directed graph

PCA



92 single cells collected from 2 mouse cell types
Included internal controls in each well, used as normalising factor

Embryonic stem cells
Embryonic fibroblasts

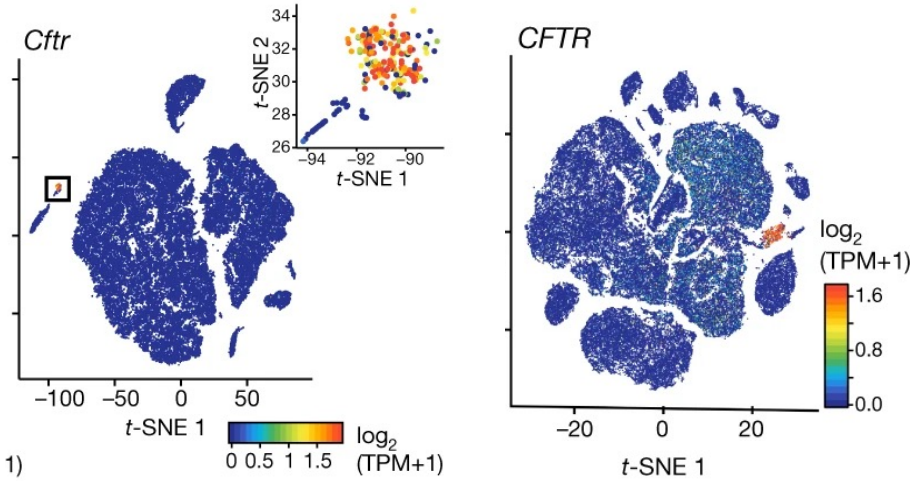
Identification of new cell types – Foxi1+ pulmonary ionocytes

Article | Published: 01 August 2018

A revised airway epithelial hierarchy includes CFTR-expressing ionocytes

Daniel T. Montoro, Adam L. Haber, Moshe Biton, Vladimir Vinarsky, Brian Lin, Susan E. Birket, Feng Yuan, Sijia Chen, Hui Min Leung, Jorge Villoria, Noga Rogel, Grace Burgin, Alexander M. Tsankov, Avinash Waghray, Michal Slyper, Julia Waldman, Lan Nguyen, Danielle Dionne, Orit Rozenblatt-Rosen, Purushothama Rao Tata, Hongmei Mou, Manjunatha Shivaraju, Hermann Bihler, Martin Mense, ... Jayaraj Rajagopal + Show authors

Nature **560**, 319–324 (2018) | Cite this article



1)

< 1% epithelial cells in surface epithelium of mouse trachea

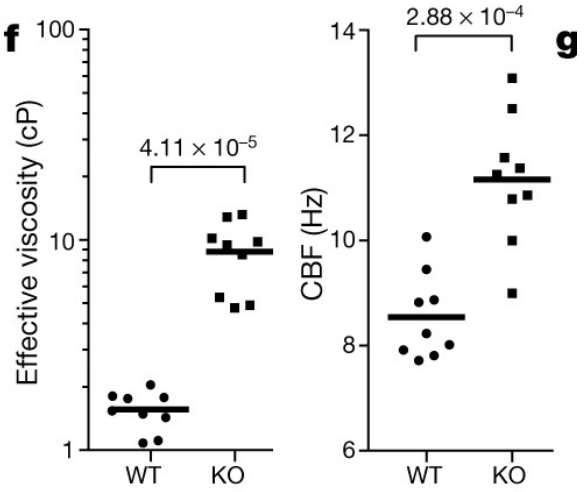
Major source of CFTR (cystic fibrosis transmembrane conductance regulator) in both mouse and humans

Letter | Published: 01 August 2018

A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte

Lindsey W. Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M. Klein & Aron B. Jaffe

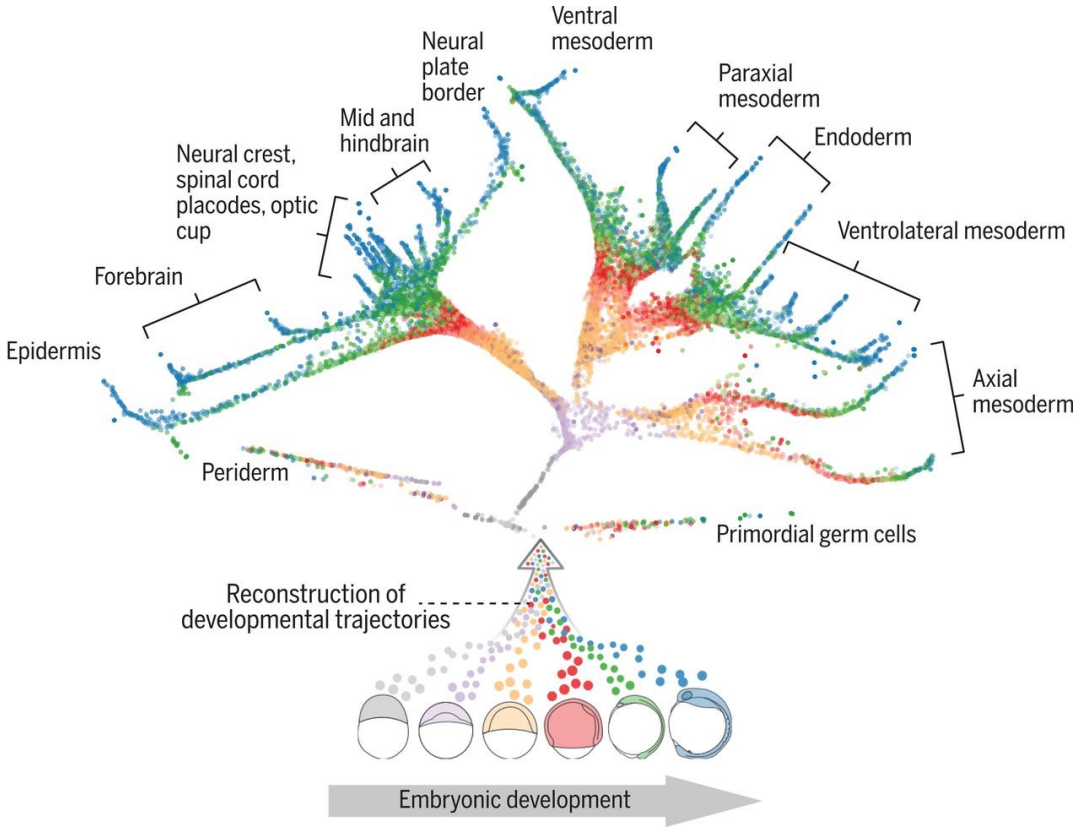
Nature **560**, 377–381 (2018) | Cite this article



KO of *Foxi1* in mouse ionocytes causes loss of *Cftr* expression and **disrupts airway fluid and mucus physiology**, phenotypes that are characteristic of **cystic fibrosis**

Associate cell-type-specific expression programs with key disease genes

Developmental tree of early zebrafish embryogenesis

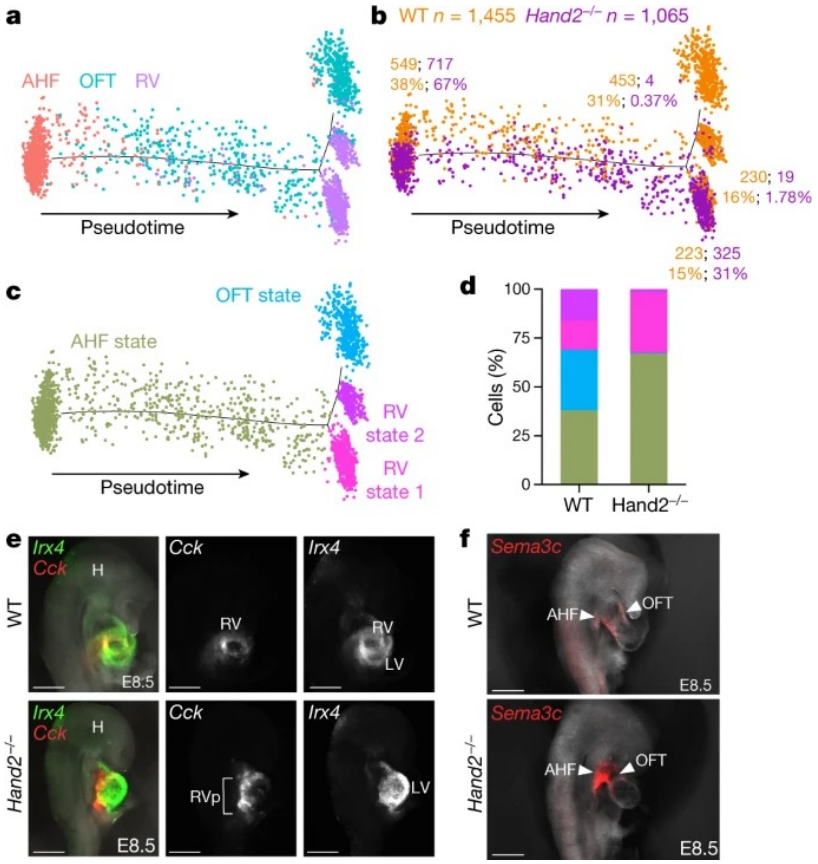


During development, cells acquire **distinct fates** by transitioning through different **transcriptional states**

With scRNA-seq, you can "take snapshots" of a cell during differentiation, which then can be analysed using trajectory inference algorithms

Reconstruct complex developmental trajectories from single-cell transcriptomes

Developmental errors behind congenital disease



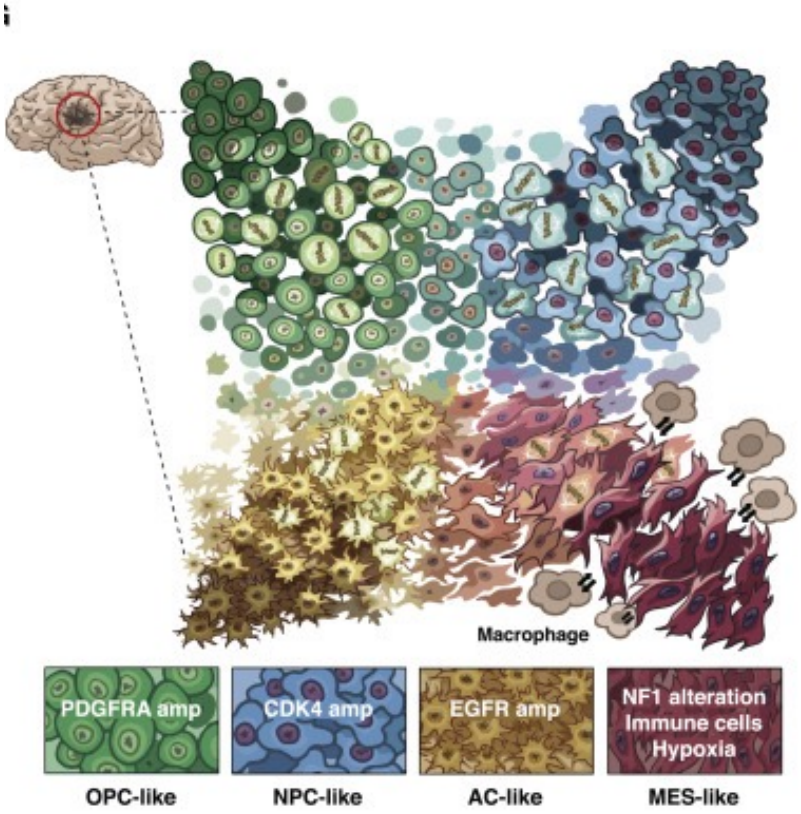
The disruption of cardiac progenitor cells results in congenital heart defects

Lineage-specifying transcription factor Hand2 determined outflow tract cells

Loss of Hand2 led to improper differentiation of cells and, thus, a disrupted cardiac development

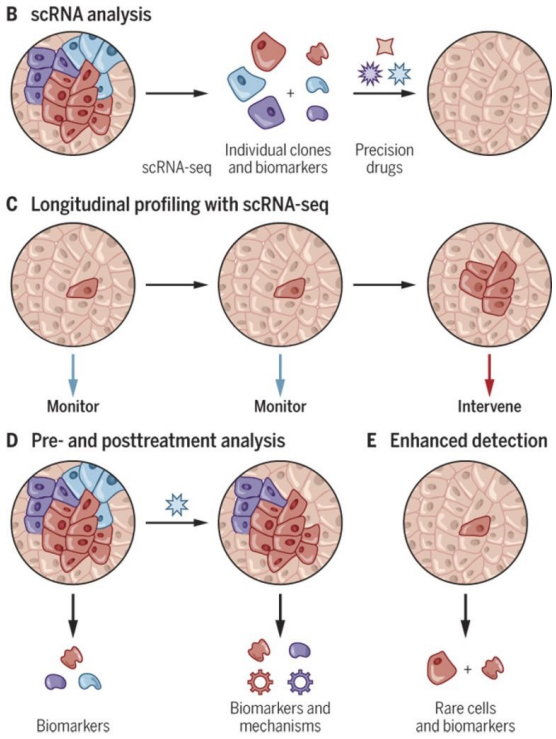
Mechanisms underlying developmental disorders

Cellular states of glioblastoma and their genetic and micro-environmental determinants



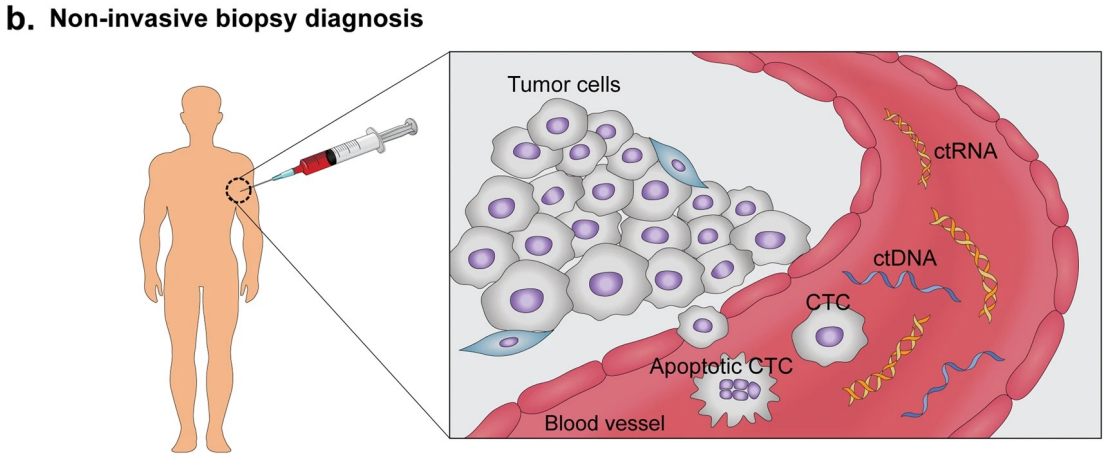
Malignant cells in glioblastoma exist in **four main cellular states** that recapitulate distinct neural cell types, are influenced by the tumour **microenvironment**, and exhibit **plasticity**

Each state is driven by distinct mutations



Intra-tumour heterogeneity, can have implications for treatment and monitoring of patients

Analysis of liquid biopsies to improve diagnosis, monitoring and treatment



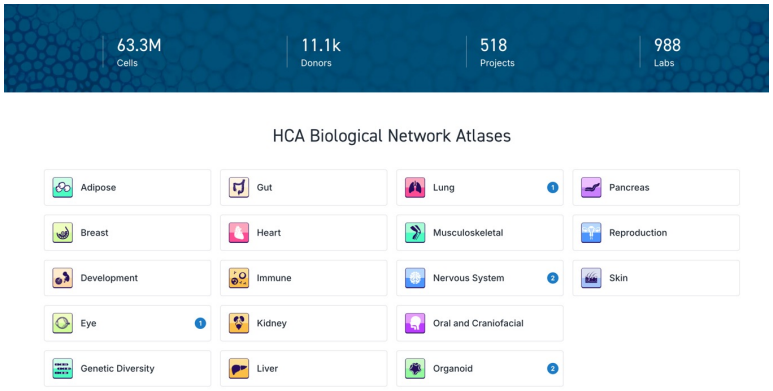
Circulating tumour cells
Cell-free DNA

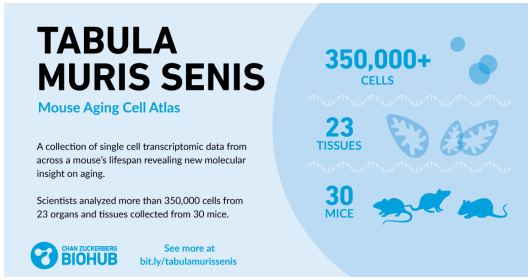
Clinical diagnostics, assessment of treatment response and disease progression

Cell Atlases: comprehensive compendium of cells in an organism

HUMAN CELL ATLAS

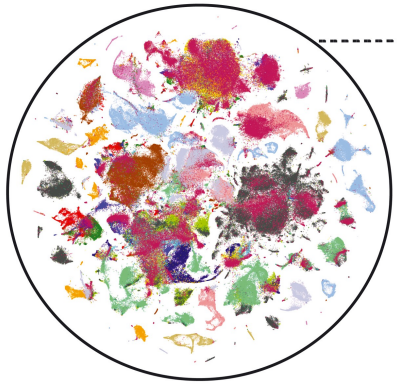
"To create comprehensive reference maps of **all human cells**—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease."

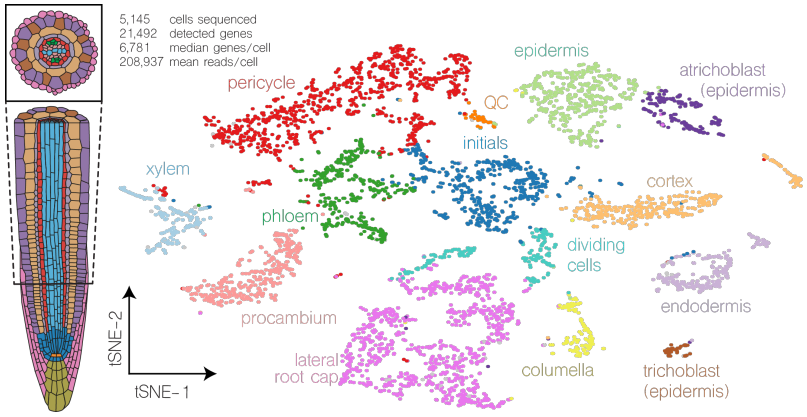




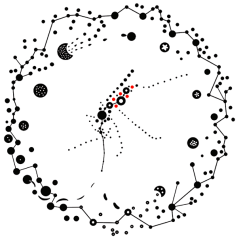
Fly cell atlas

Adult fly
17 tissues
580,000 cells
>250 cell types



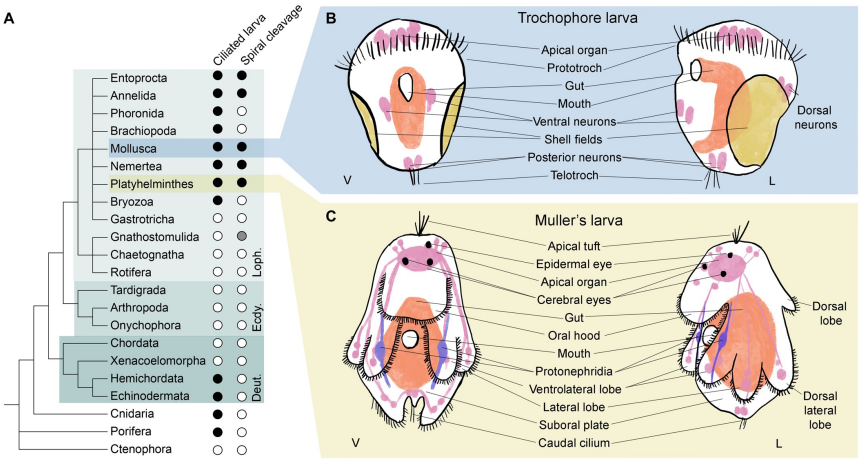


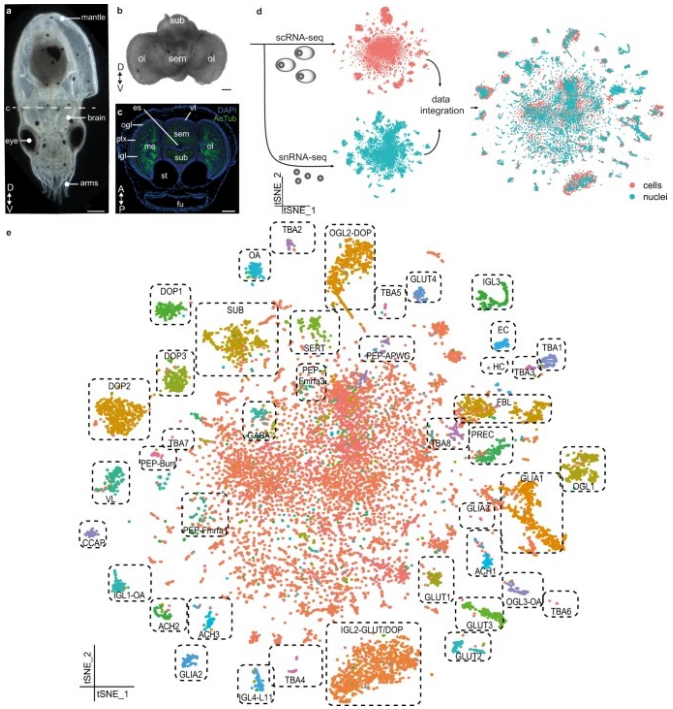
Malaria Cell Atlas

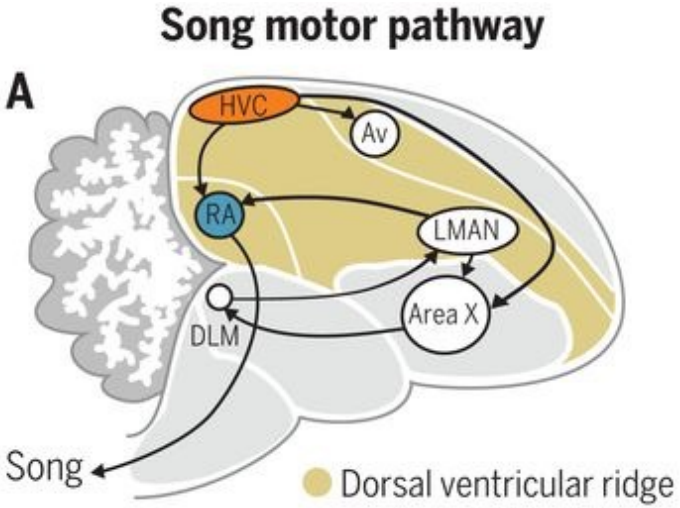


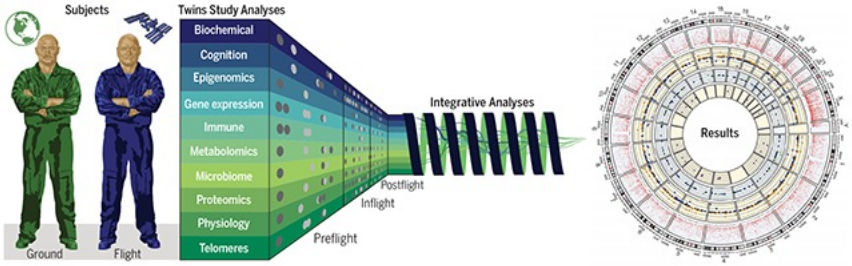
Non model-organisms!

Astronauts, octopus and more



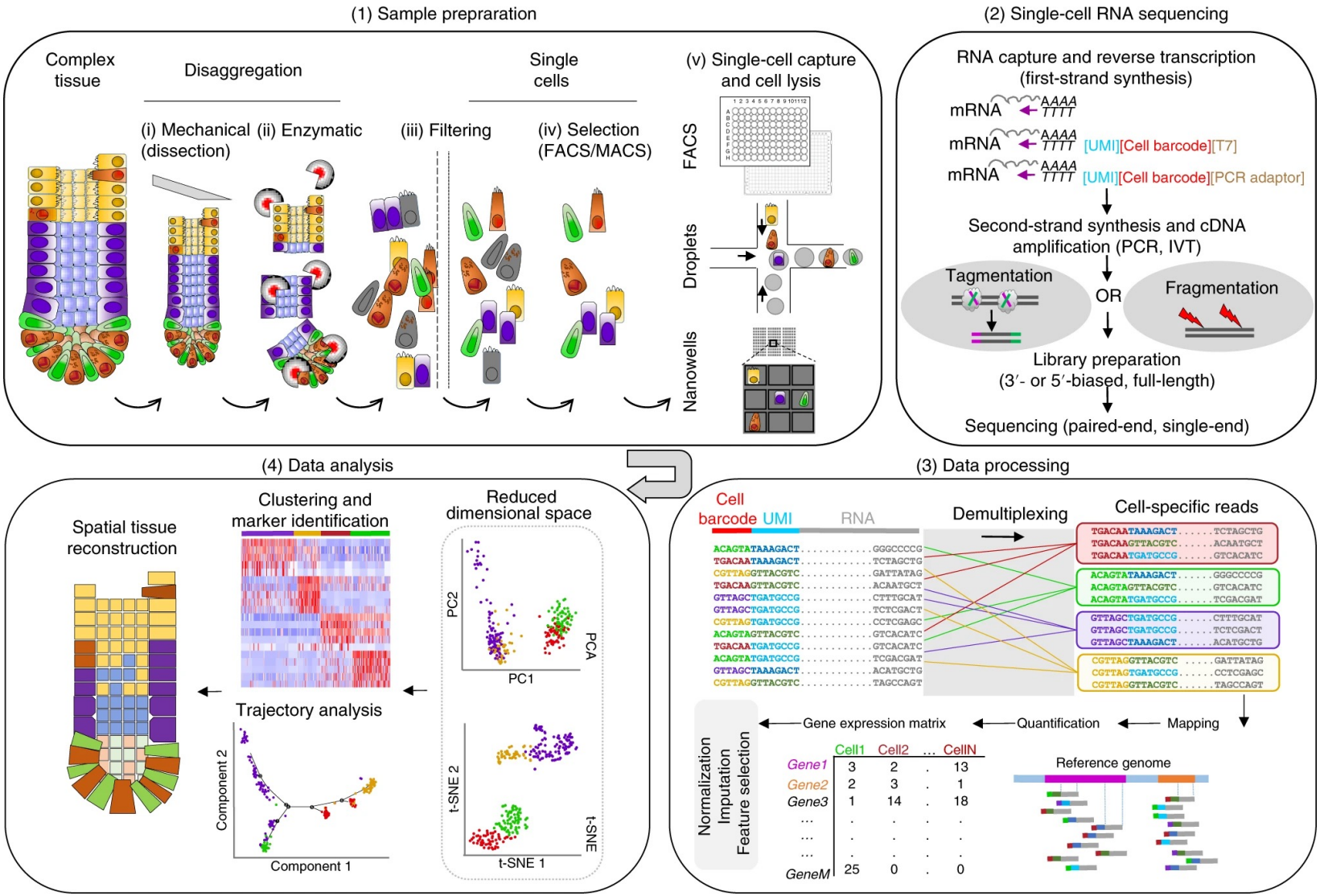




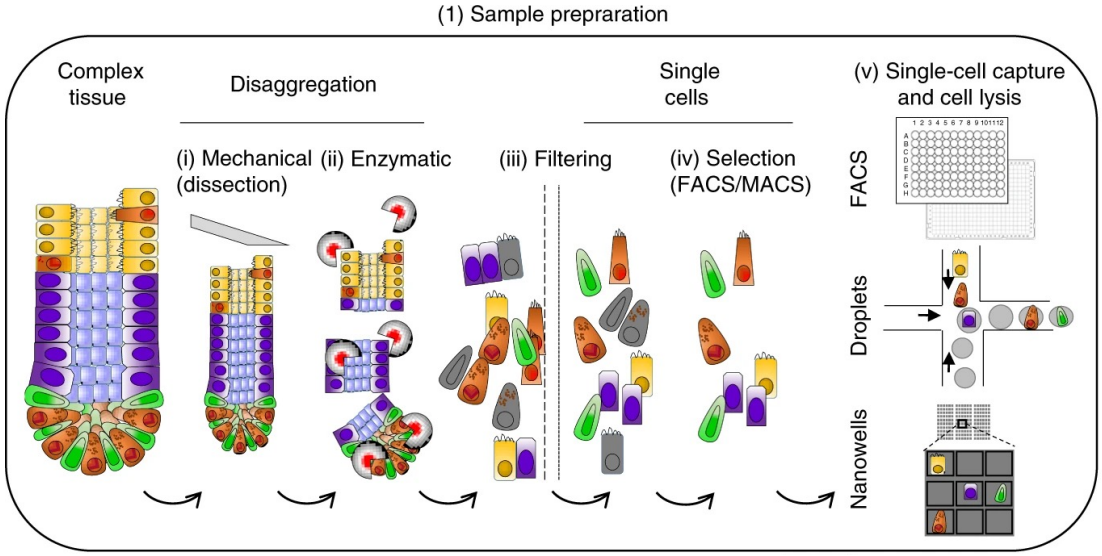


NASA twin study

The single-cell RNAseq pipeline

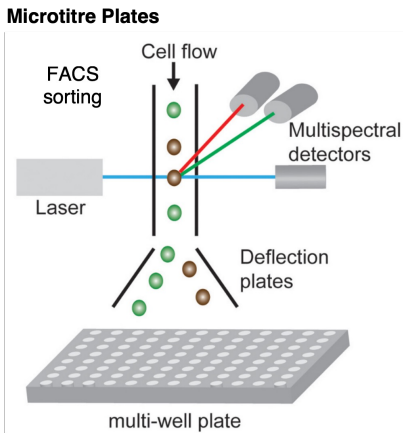


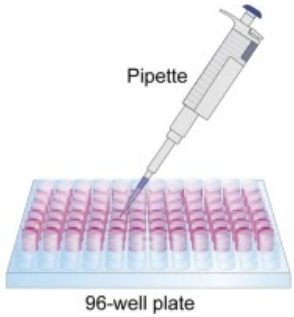
Single cell-capture



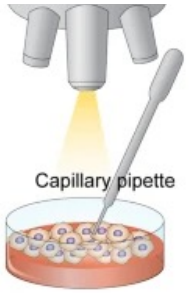
Each single cell is captured in **an isolated reaction mixture**, of which all **transcripts from one single cell will be uniquely barcoded** after converted into complementary DNAs (cDNA)

Cell isolation – plate based

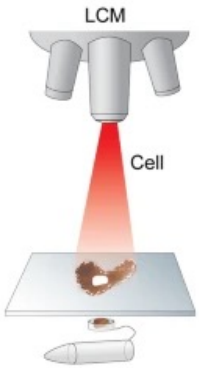




Serial dilution



Microdissection



Laser capture microdissection

Advantages

Imaging before library preparation, providing an additional data modality

Identify and discard doublets or dying cells

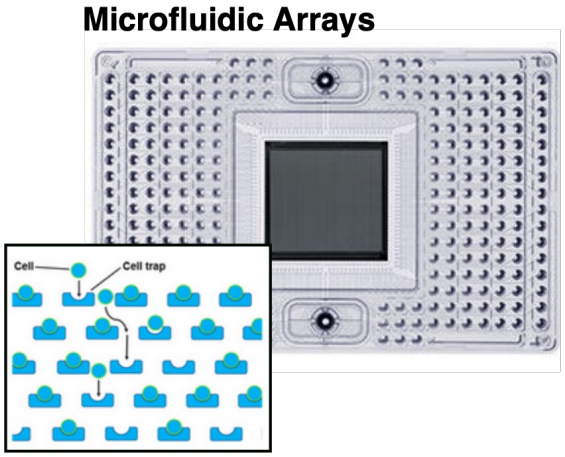
Highly selective, precise isolation

Disadvantages

Low throughput

Requires a large quantity of cells and may cause cell damage due to high flow rate

Cell isolation – microfluidic arrays



Advantages

Higher throughput compared to microwell

Integrated system

Low sample volume requirement

Disadvantages

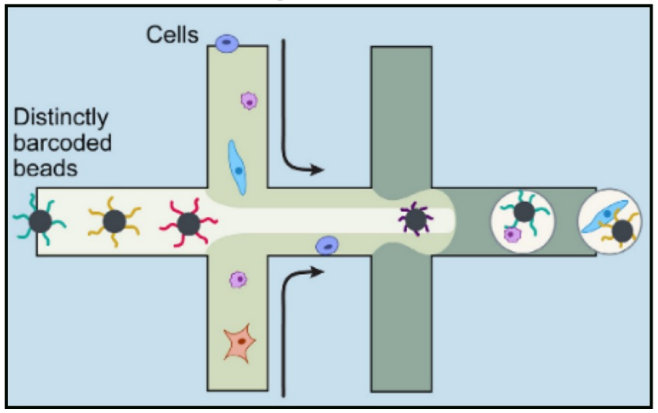
More expensive

Only 10% of cells are captured

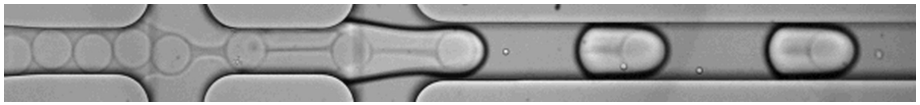
Expensive, cellular stress

Cell isolation – droplet base

Microfluidic Droplets



Encapsulation of cells in droplets with barcoded beads



Advantages

Highest throughput

Cost-effective

Automated

Low sample volume requirement

Disadvantages

Smaller cell libraries

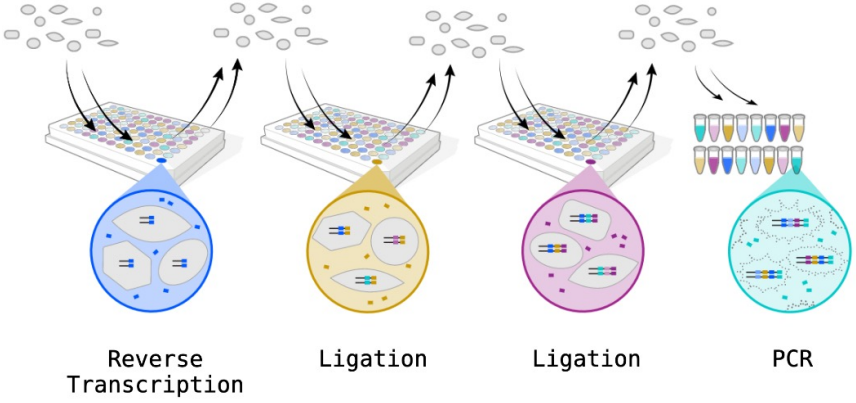
Specialised and expensive equipment

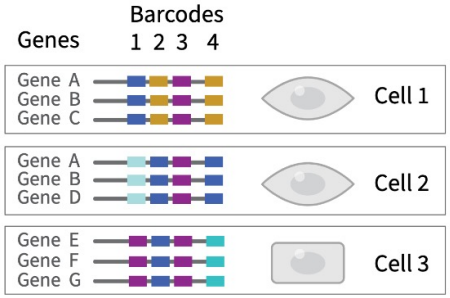
Ambient RNA - RNA molecules that are freely floating in the cell lysate due to the breakdown of dead or dying cells before the droplets are separated

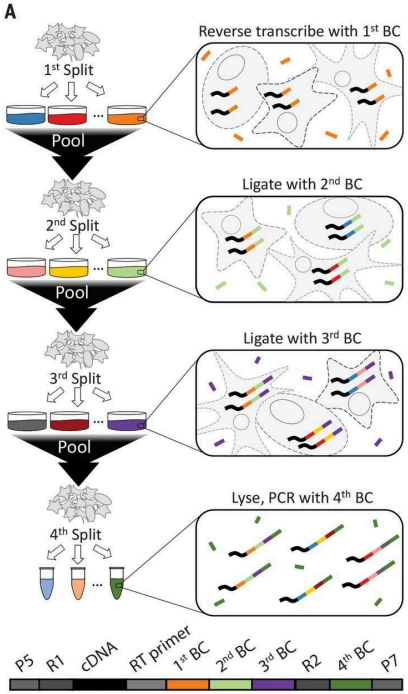
Instrument free cell isolation

"Split-pooling" - Combinatorial barcoding

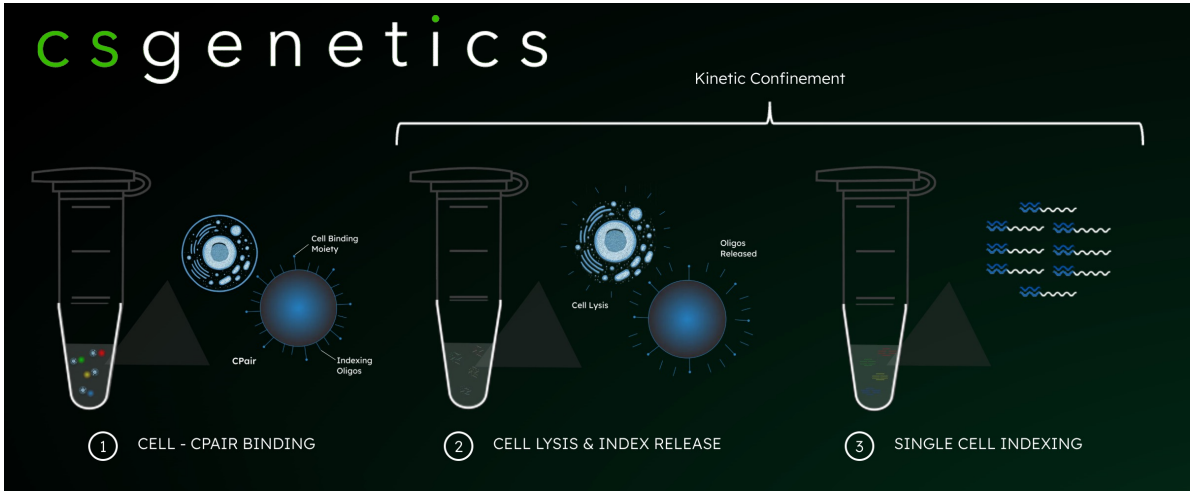
Parse Evercode™ Split Pool Combinatorial Barcoding







Kinetic Confinement



Single-nuclei isolation

Frozen tissue, no intact cells

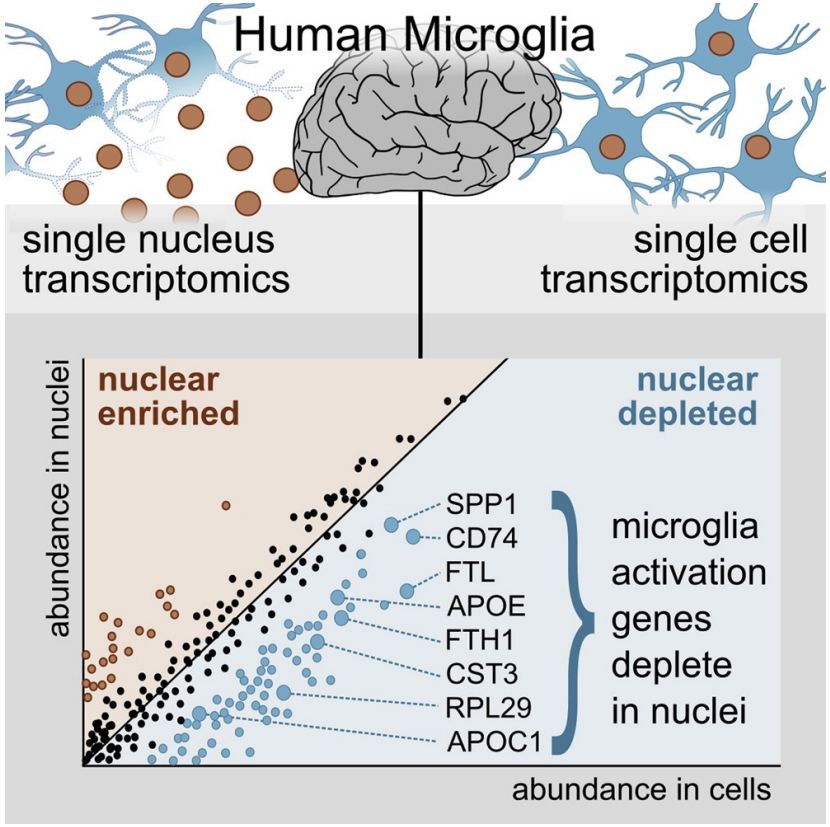
Oddly shaped cells, hard to encapsulate in microfluidics

RNA content and gene diversity lower than whole cell, but possible to identify cell types

Multiome: scRNA + scATAC

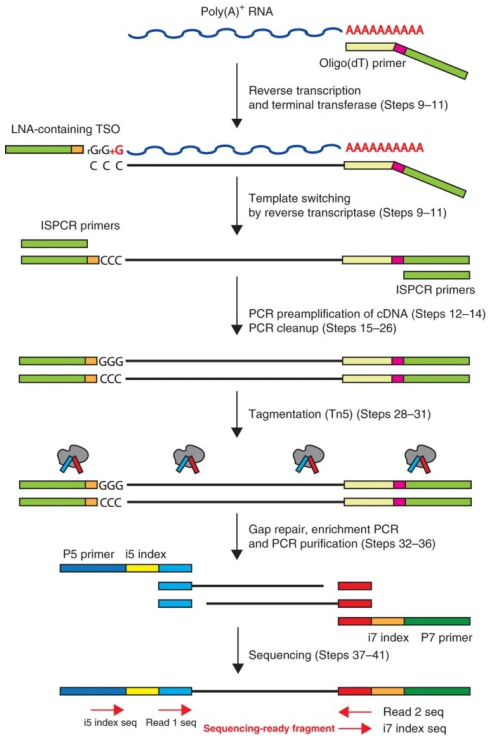
Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans

Nicola Thrupp [1,2] · Carlo Sala Frigerio [1,2,3] · Leen Wolfs [1,2] · … · Renzo Mancuso [1,2] · Bart de Strooper [1,2,3,6] · Mark Fiers [1,2,3] … Show more



Transcript quantification: full length vs tag-based

Smart-seq2



Terminal transferase adds Cs to cDNA, providing basis for template switching - **the process used in single-cell RNA sequencing to capture and convert a cell's mRNA into a cDNA library**

TSO (template switch oligo) hybridizes to untemplated C nucleotides added during reverse transcription and adds a common 5' sequence to full length cDNA that is used for downstream cDNA amplification

Sequencing libraries are prepared by **tagmentation**, which simultaneously **fragments and indexes** the cells

As sequencing libraries contain multiple fragments from different parts of each expressed gene, the likelihood of **capturing individual genes** increases

Extensive liquid handling, ~1000 cells

Splice variants and alternative transcripts, SNPs and fusions

Transcript quantification: full length vs tag-based

Only 5’ or 3’ end of transcript is sequenced

The miniscule amounts of starting material requires the use of **multiple rounds of PCR**, leading to potentially very **large amplification biases.**

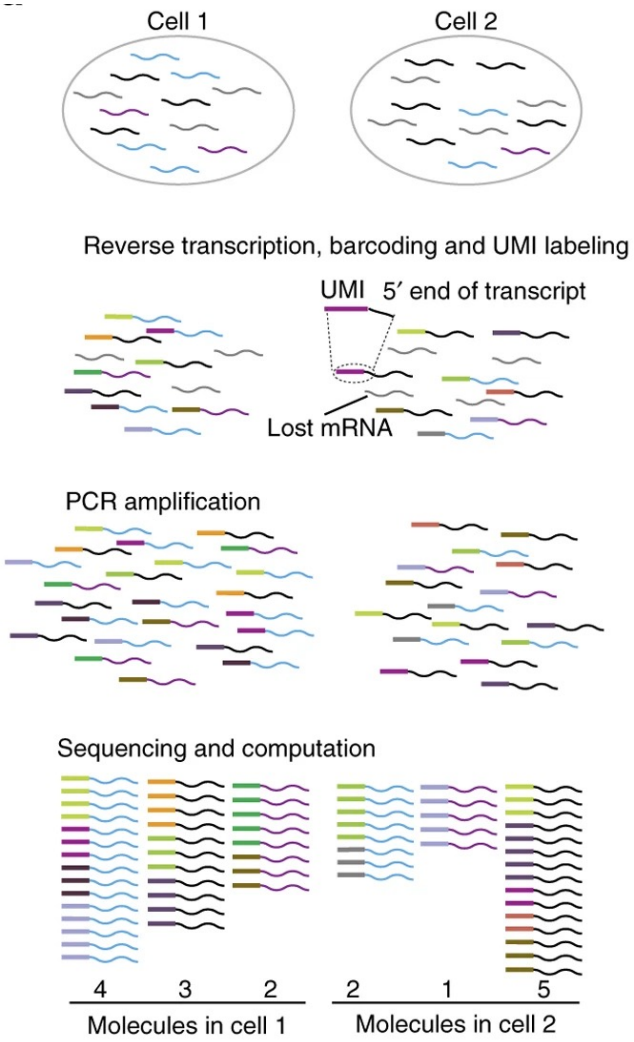
UMI – Unique Molecular Identifier

Very diverse set of oligonucleotides with random sequence

Molecules with the same UMI were derived from the same transcript, and can be removed computationally

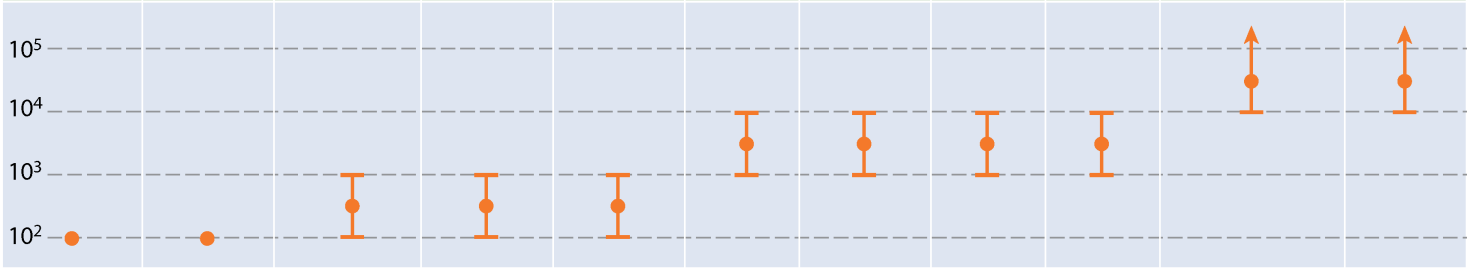
TCRseq: identify and track specific T cells and their clones, based on sequencing of T cell receptors.

Requires 5’ sequencing, as highly variable region is located at 5’ of rearranged transcript



Drop-seq
Chromium (10x Genomics)

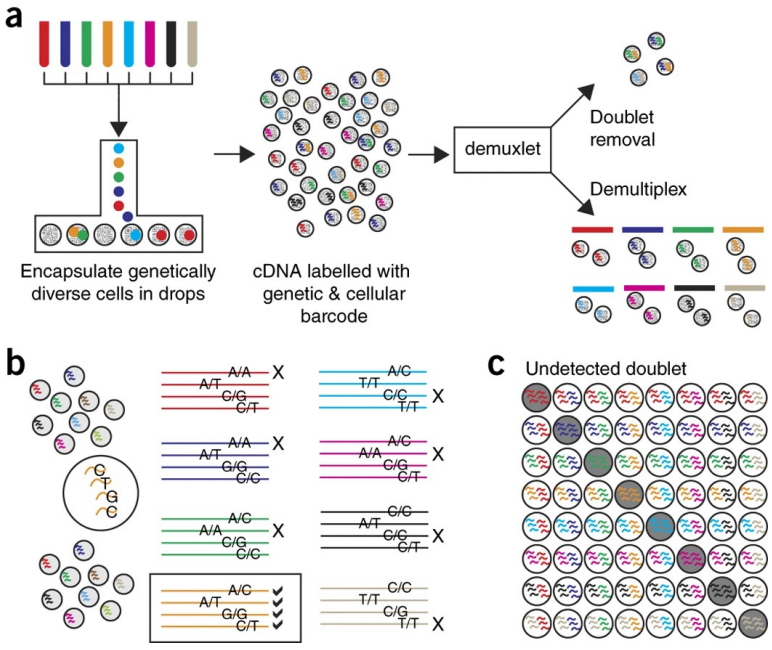
Comparison technologies

	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3'	5'	3'	3'	3'	3'	3'	3'	3'	3'
Number of cells per assay											

Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

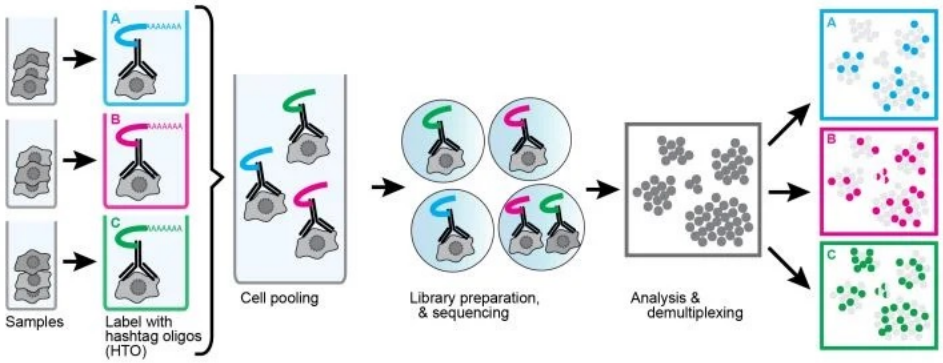
Experimental design – multiplexing

Genotype based



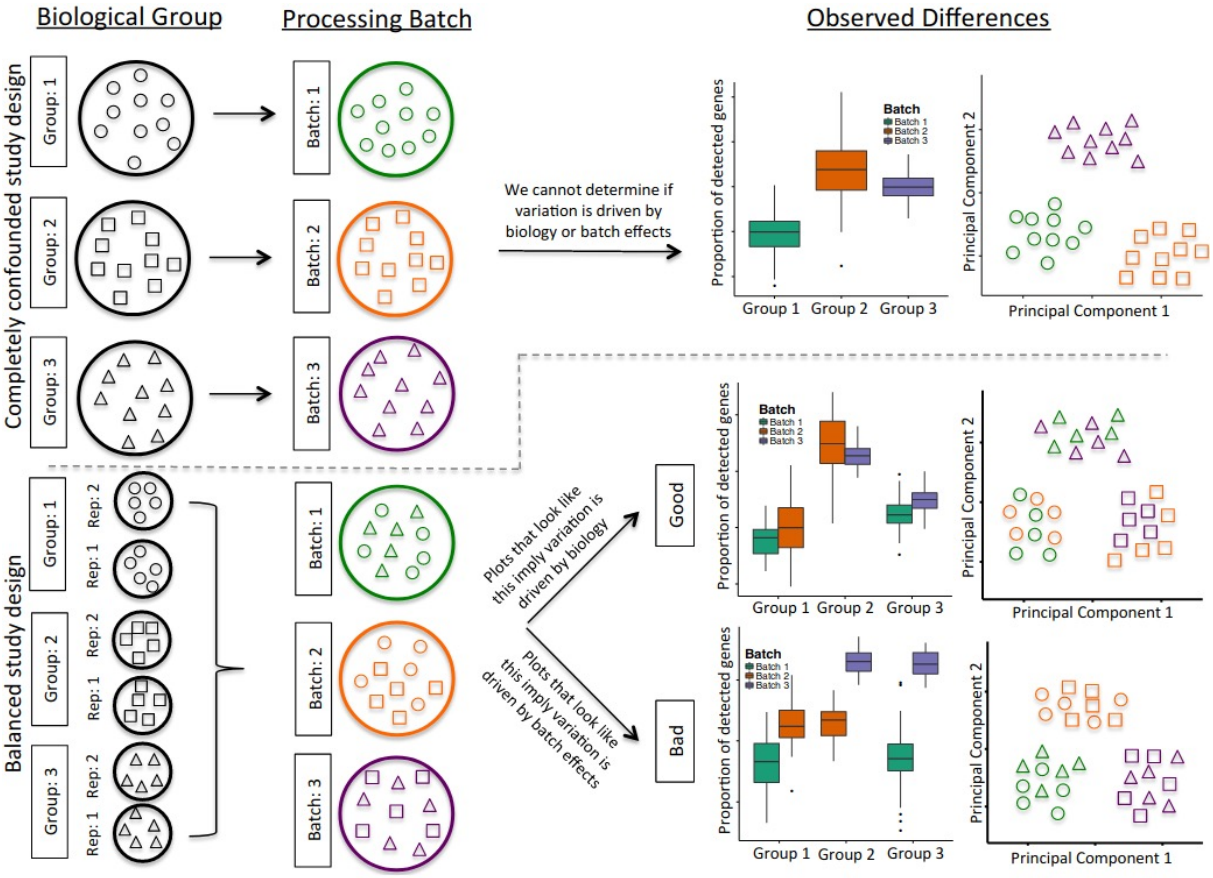
Harnesses **natural genetic variation** to determine the sample identity of each droplet containing a single cell (singlet) and detect droplets containing two cells (doublets)

Hashtags



Oligonucleotide-conjugated antibodies against markers that are ubiquitously expressed across different cell types

Experimental design – batch effects



scRNA (data) challenges

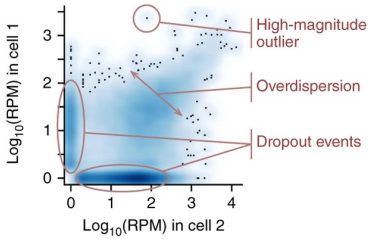
Each data point represents single cell, and there is no way of having “biological replicates” at the single-cell level: **each cell is unique and impossible to replicate**

Instead, cells can be clustered by their similarity, and comparisons can then be done across groups of similar cells (as we shall see later in the course)

Very **low starting amounts of transcripts** since the RNA comes from one cell only.
- Amplification bias
- Noisy

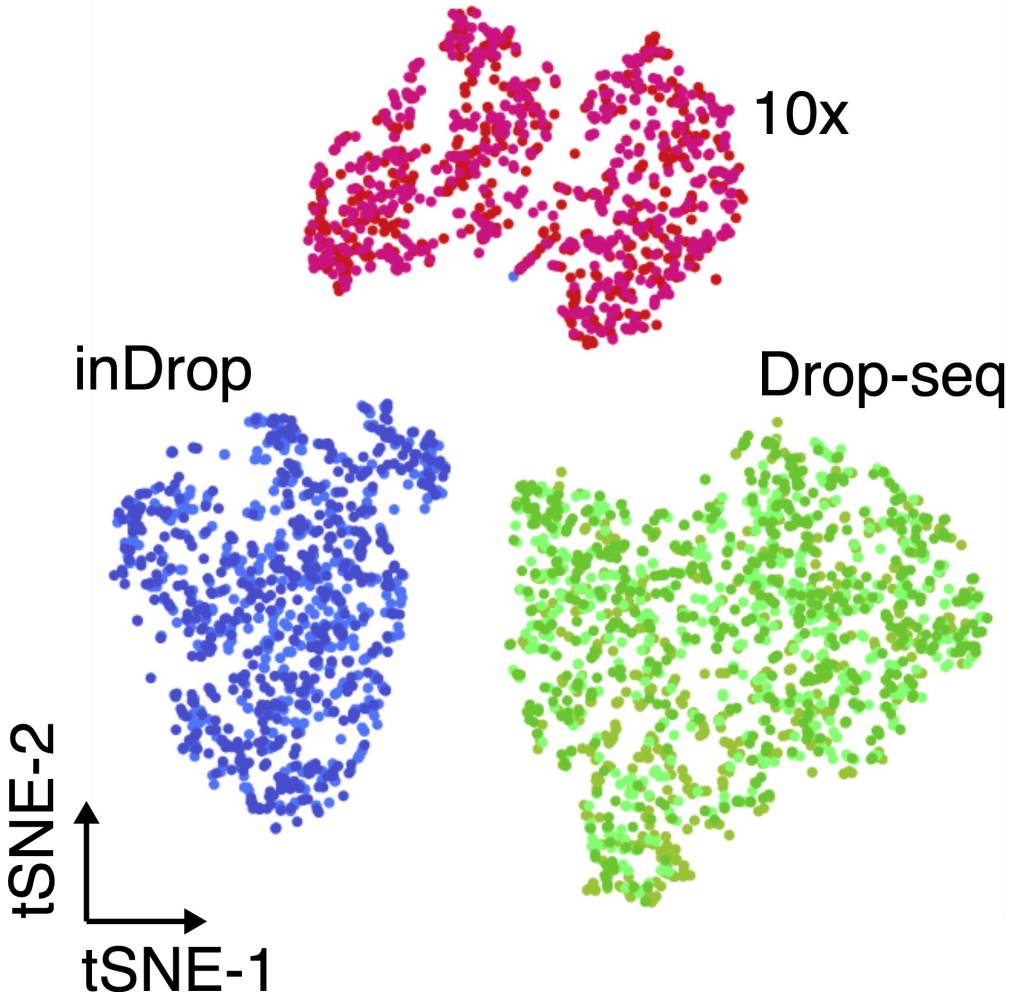
Sparsity: not all transcripts present in a single cell can be captured; therefore, the gene expression matrices in these cases contain many zeros

1 – Real zero: gene not expressed in that cell
2 – Dropout: gene expressed, but not detected in that cell

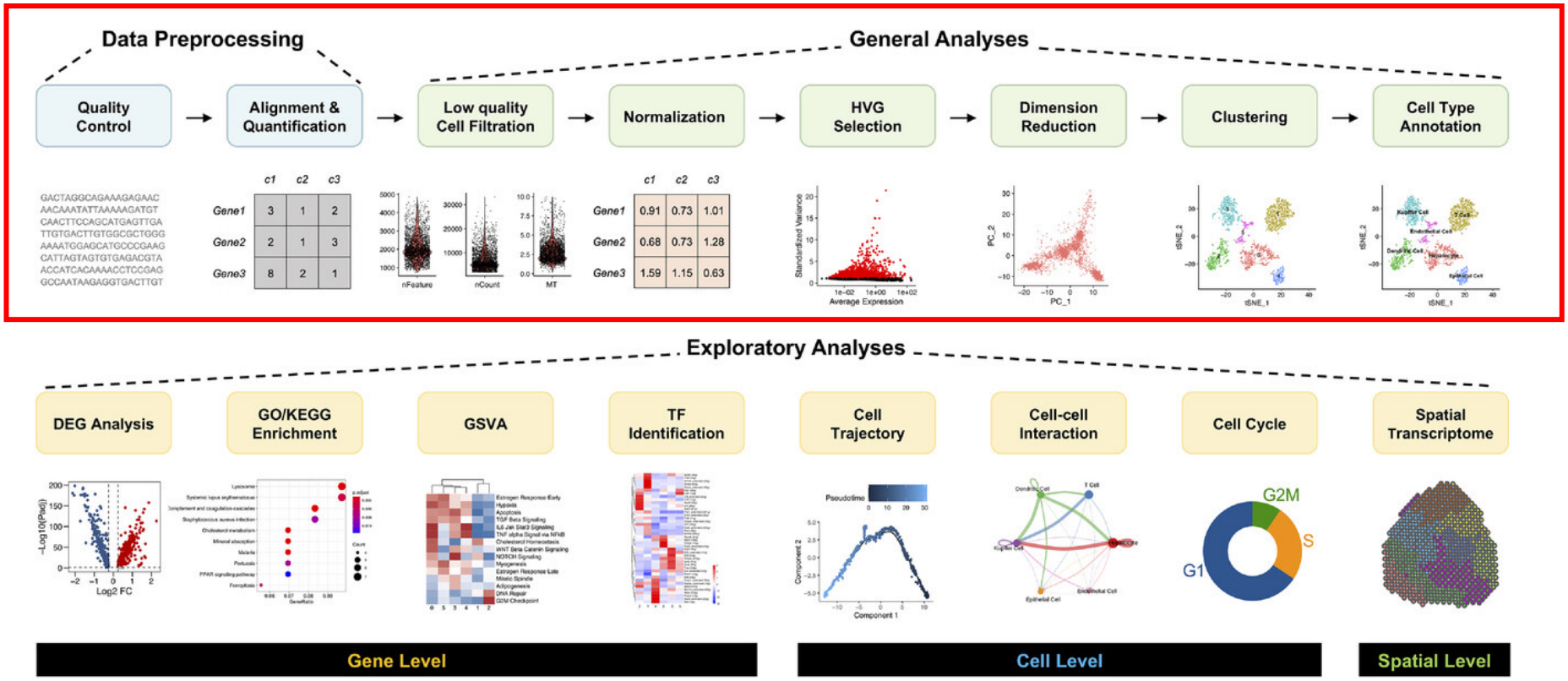


Cell to cell variation that is not always biological: uneven PCR amplification, gene dropouts (gene detected in one cell but absent from another)

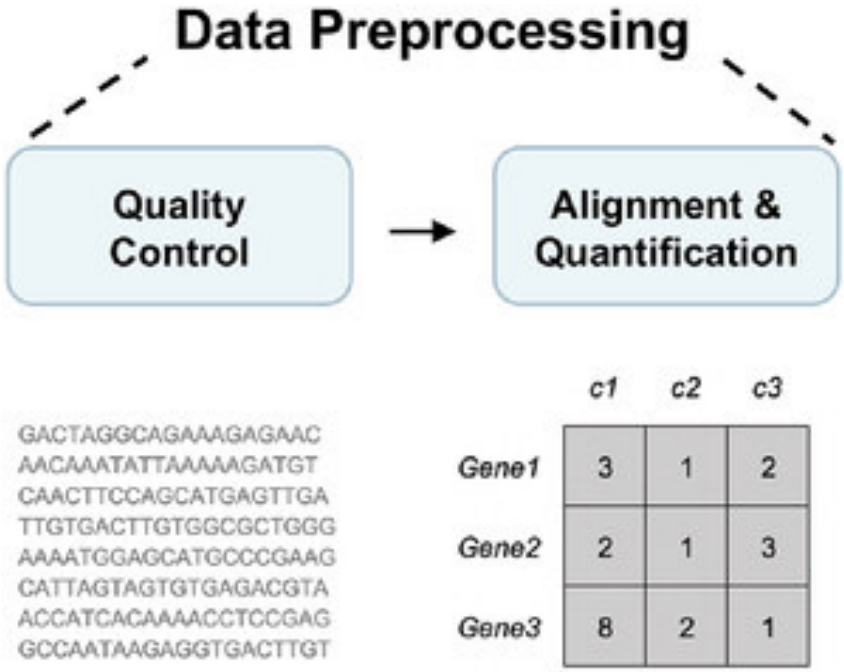
scRNA (data) challenges



scRNA analysis



scRNA analysis – Alignment and quantification



QC of raw reads - FASTQC, MultiQC (not covered in this workshop)

Most sequencing companies or core facilities provide these (or similar) reports

Align reads to the genome
Annotate reads to features (genes)
Quantify gene expression

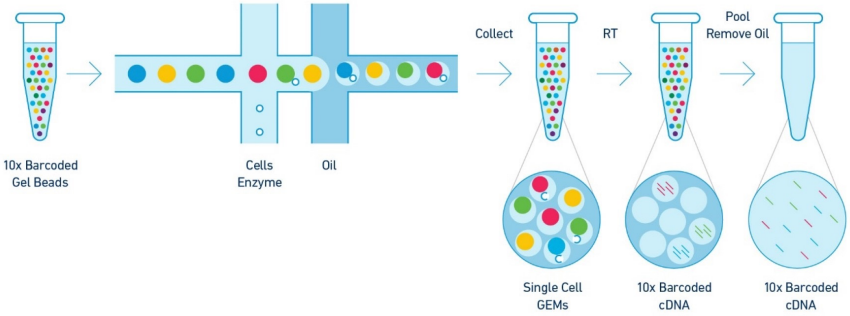
10x Chromium

Article | Open access | Published: 16 January 2017

Massively parallel digital transcriptional profiling of single cells

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, ... Jason H. Bielas + Show authors

Nature Communications **8**, Article number: 14049 (2017) | Cite this article



Popular high-throughput method

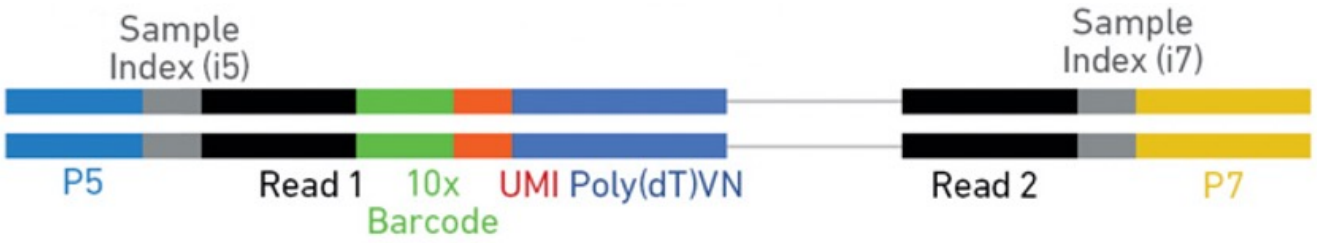
GEM – Gel bead in emulsion

UMIs

Up to 60k cells per well

Cells are delivered at limiting dilution, such that the majority (90-99%) of GEMs contains no cell, while the remainder largely contain a single cell

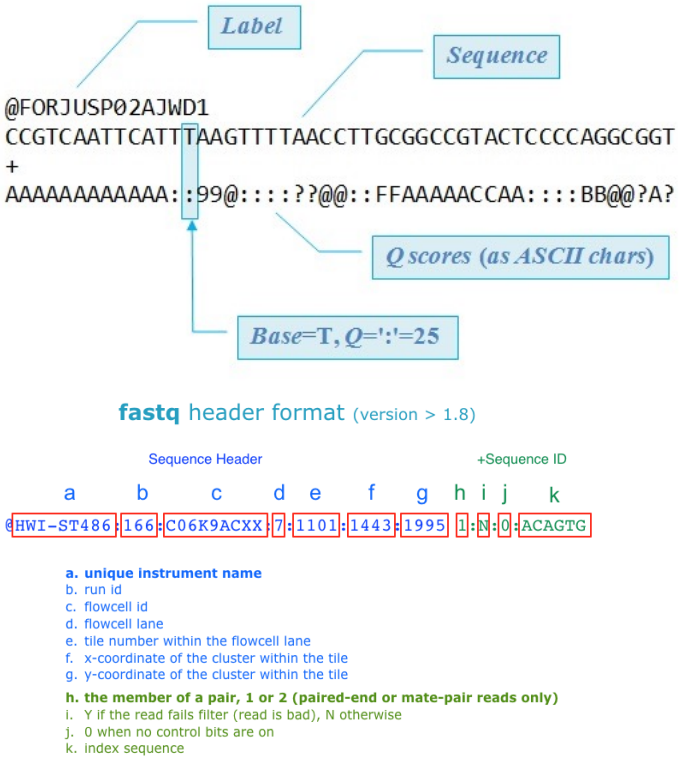
10x library structure



Sample index: identifies the library, with one or two indexes
10x barcode: identifies the cell
UMI: identifies the transcript
Insert: the transcript molecule

FastQ files: standard text-based file format used to store biological sequence data from NGS

I1: I7 sample index
I2: I5 sample index if present (dual indexing only)
R1: 10x barcode + UMI
R2: insert sequence



Alignment, quantification and sample demultiplexing with CellRanger

What is Cell Ranger?

A set of analysis pipelines that perform sample demultiplexing, barcode processing, single cell 3' and 5' gene counting, V(D)J transcript sequence assembly and annotation, and Feature Barcode analysis from single cell data.

Download Cell Ranger | Use Cloud Analysis | Get Started with Cell Ranger

A command-line software suite to quantify (and analyse) data from 10x Genomics scRNA-seq/snRNA-seq

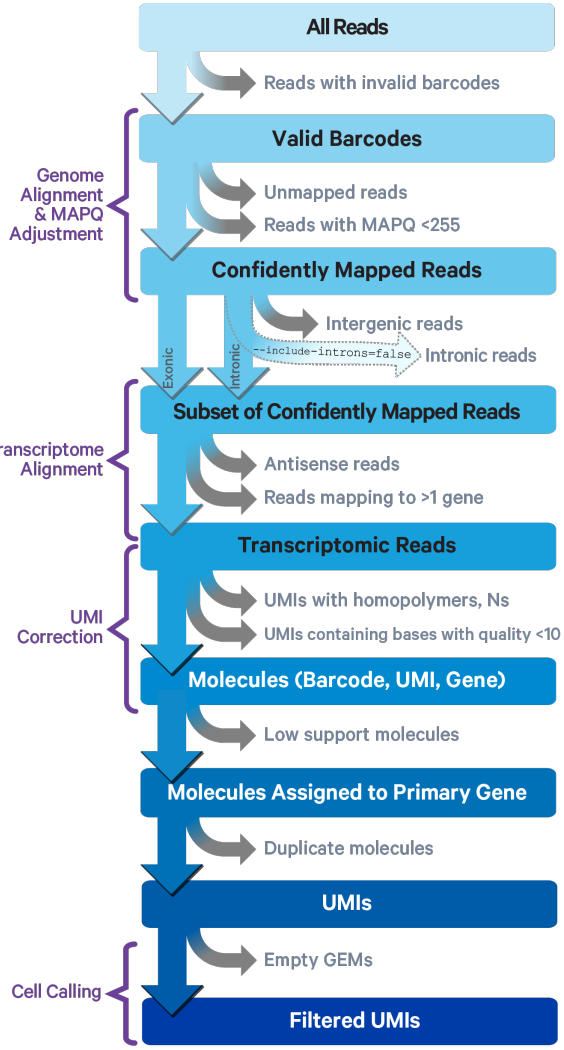
Based on STAR for alignment (there are other mapping algorithms, not covered here)

Call cells, i.e. filter the raw matrix to remove droplets that do not contain cells

Generate a very useful report in html format, which will provide some QC metrics and an initial look at the data

Generate a “cloupe” file, which can be opened using the 10x Loupe Browser software to further explore the data

Computationally very intense, **you will not be able to run it on your laptops**. You will need access to, for example, a high-performance computing (HPC) cluster, a server or other cloud-based computational resource with sufficient power - talk to your local IT support.



https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

CellRanger count – preparing fastq files

Requires fastq file names to follow a convention:

```
<SampleName>_S<SampleNumber>_L00<Lane>_<Read>_001.fastq.gz
```

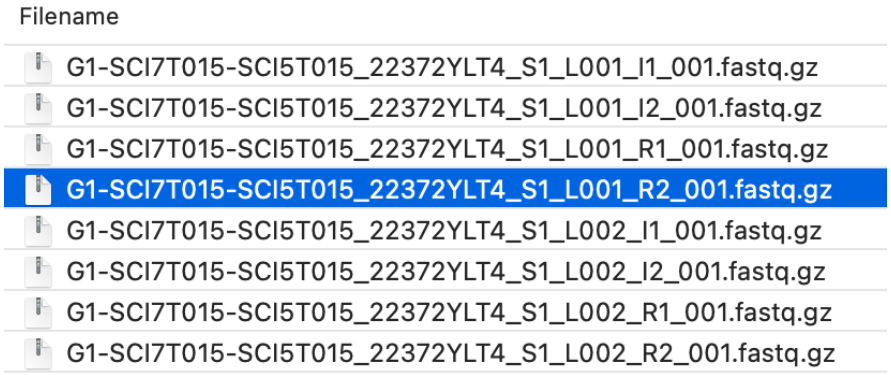
<SampleName> - An identifier for the sample, this is what Cell Ranger uses to determine which fastq files to combine into a single sample.

<SampleNumber> - This is the sample number based on the order that samples were listed in the sample sheet used when running bcl2fastq. This is not important for Cell Ranger, other than it indicates the end of the Sample Name, you can set all your samples to S1.

<Lane> - The lane number. If your sequencing was run across multiple lanes, then you may have multiple sets of fastqs for a single sample with different lane numbers.

<Read> - The read type: **R1** for Read 1, **R2** for Read 2, and index reads are **I1** and **I2**.

001 - The last segment is always 001.



FastQ files
I1: I7 sample index
I2: I5 sample index if present (dual indexing only)
R1: 10x barcode + UMI
R2: insert sequence

Reference transcriptome

1 - From 10x website (human, mourse or human + mouse)
https://www.10xgenomics.com/support/software/cell-ranger/downloads#reference-downloads

2 – Build custom reference with `cellranger mkref`

```
cellranger mkref \
  --fasta={GENOME FASTA} \
  --genes={ANNOTATION GTF} \
  --genome={OUTPUT FOLDER FOR INDEX} \
  --nthreads={CPUS}
```

GENOME FASTA is a file containing the reference genome in FASTA format

ANNOTATION GTF is a file containing the transcript annotation file in GTF format

OUTPUT FOLDER FOR INDEX is a name for the output folder containing the new reference package (you do not need to create this folder)

CPUS - Is the number of CPUs we would like CellRanger to use. The more CPUs CellRanger can use, the faster the job (up to a point).

cellranger count – aligning reads and generating a count matrix

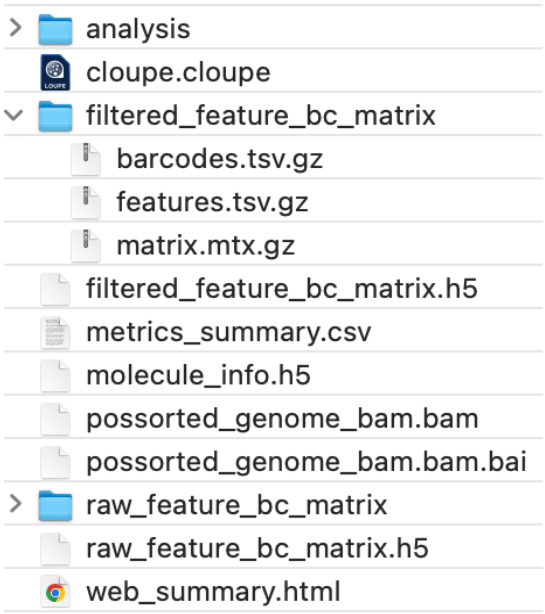
The minimum information require to run cellranger count is:
--id - A sample ID. This is used for naming the outputs
--transcriptome - the directory containing the Cell Ranger reference
--fastqs - the directory containing the fastq files

This will process all fastq files in the --fastqs directory into a single sample. If you have multiple samples in a single directory then you need to add:
--sample - the SampleName from the fastq files

In addition, Cell Ranger is very computationally intensive, you will usually be wanting to run it on a high-performance cluster or server. It will greedily attempt to use all resources it can find available, and so it is advisable to set limits to the resources it will use:
--localcores - the number of processors Cell Ranger should use
--localmem - the amount of memory, in Gigabytes, Cell Ranger should use.
A complete command would look like this

```
cellranger count --id={OUTPUT_SAMPLE_NAME} \
                 --transcriptome={DIRECTORY_WITH_REFERENCE} \
                 --fastqs={DIRECTORY_WITH_FASTQ_FILES} \
                 --sample={NAME_OF_SAMPLE_IN_FASTQ_FILES} \
                 --localcores={NUMBER_OF_CPUS} \
                 --localmem={RAM_MEMORY}
```

cellranger count outputs



analysis - The results of clustering and differential expression analysis on the clusters. These are used in the *web_summary.html* report.
cloupe.cloupe - a cloupe file for loading into the 10x loupe browser
filtered_feature_bc_matrix - The filtered count matrix directory
filtered_feature_bc_matrix.h5 - The filtered count matrix as an HDF5 file
metrics_summary.csv - summary metrics from the analysis
molecule_info.h5 - per-molecule read information as an HDF5 file
possorted_genome_bam.bam - The aligned reads in bam format
possorted_genome_bam.bam.bai - The bam index
raw_feature_bc_matrix - The raw count matrix directory
raw_feature_bc_matrix.h5 - The raw count matrix as an HDF5 file
web_summary.html - The summary report

The two count matrix directories each contain 3 files:
barcodes.tsv.gz - The cell barcodes detected; these correspond to the columns of the count matrix
features.tsv.gz - The features detected. In this cases gene ids. These correspond to the rows of the count matrix.
matrix.mtx.gz - the count of unique UMIs for each gene in each cell.

The count matrix directories and their corresponding HDF5 files contain the same information, they are just alternative formats for use depending on the tools that are used for analysis. In this course we will be loading the contents of the count matrix directories into R.

The filtered count matrix only contains droplets that have been called as cells by Cell Ranger.

There are also many intermediate and log files/directories that will not be of immediate interest.

cellranger count web summary

The Cell Ranger summary report - web_summary.html - is a very useful first port of call for assessing the quality of the data.

The first tab, **Summary**, contains various QC metrics about sequencing quality, mapping quality and cell calling.

The second tab, **Analysis**, provides some basic analysis of droplets that Cell Ranger has identified as potentially containing cells, including some clustering and gene expression analysis. This report is interactive, allowing you to some very basic data exploration.

The report itself contains brief explanations of the contents of each section. These can be accessed by clicking the question mark icon in the header of the section. A more comprehensive explanation of the various components of the report can be found on the 10x website.

Not every droplet is usable

A single happy cell in a droplet is ideal

- Complex transcriptome
- Average number of genes detected

Empty droplet: No cell in a droplet

- No genes detected

Droplet with ambient RNA

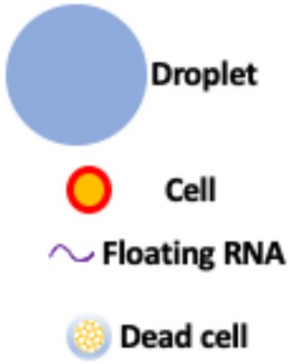
- Low complex transcriptome
- Genes detected much lower than average genes per cell

Droplet with dead cell

- Enriched for mitochondrial genes

Droplet with multiple cell

- Very complex transcriptome
- Genes detected much higher than average genes per cell



There are several methods for doublet detection, ambient RNA correction, etc, which are not covered in the workshop. We will use the filtered matrix from CellRanger.

Method | Open access | Published: 22 March 2019

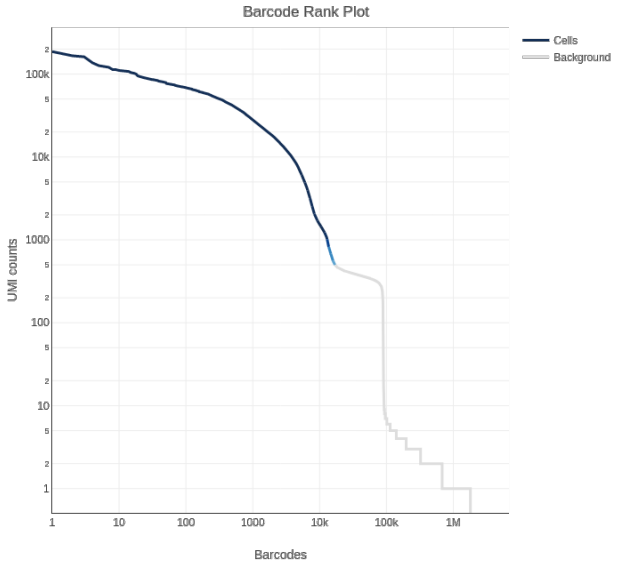
EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data

Aaron T. L. Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree & John C. Marioni

Genome Biology **20**, Article number: 63 (2019) | Cite this article

41k Accesses | **961** Citations | **43** Altmetric | Metrics

cellranger count web summary

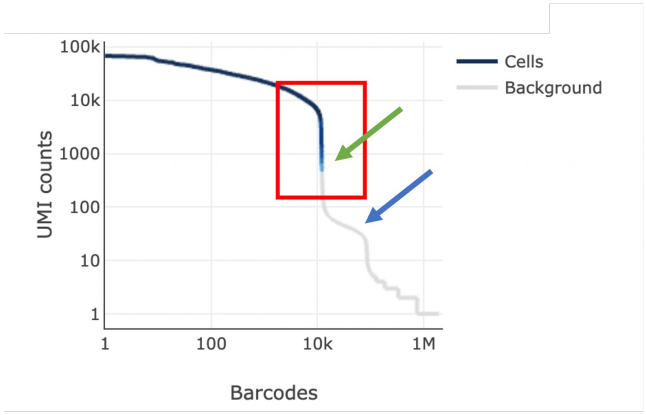


UMI counts for all droplets

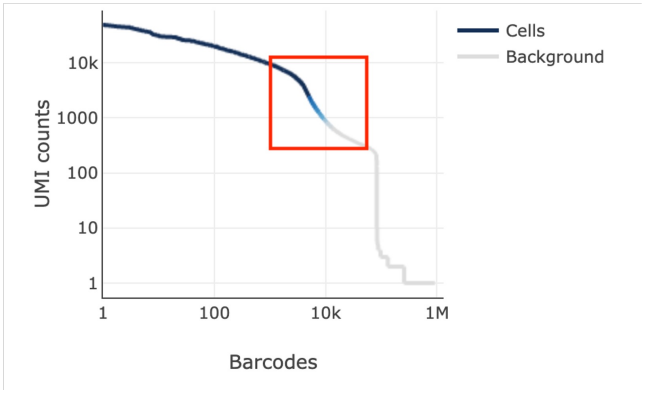
Droplets have been ranked by UMI count and coloured according to whether they have been called as cells or not.

The light blue region, indicates a region where some droplets have been called as cells, but other have not - depending on the outcome of the RNA profiling step of the algorithm.

Characteristic Steep Cliff



Lack of Steep Cliff



Single-cell vocabulary alert

Cell = barcode
Transcript = UMI

What are the counts?
Remember that for each read we have a sample barcode, a cell barcode, and a UMI.

In addition we now have location on the genome and a corresponding gene annotation.

The count for any gene for any cell is the number of unique reads detected for the combination:
sample barcode + cell barcode + UMI + gene ID

Thus we avoid counting PCR duplicate reads as these will have the same UMI. For this reason we more commonly talk about the “UMI count” rather than the “Read count”; the “Read count” could be higher than the “UMI count”

Activity – reading cell ranger output

Open web summary and explore it.

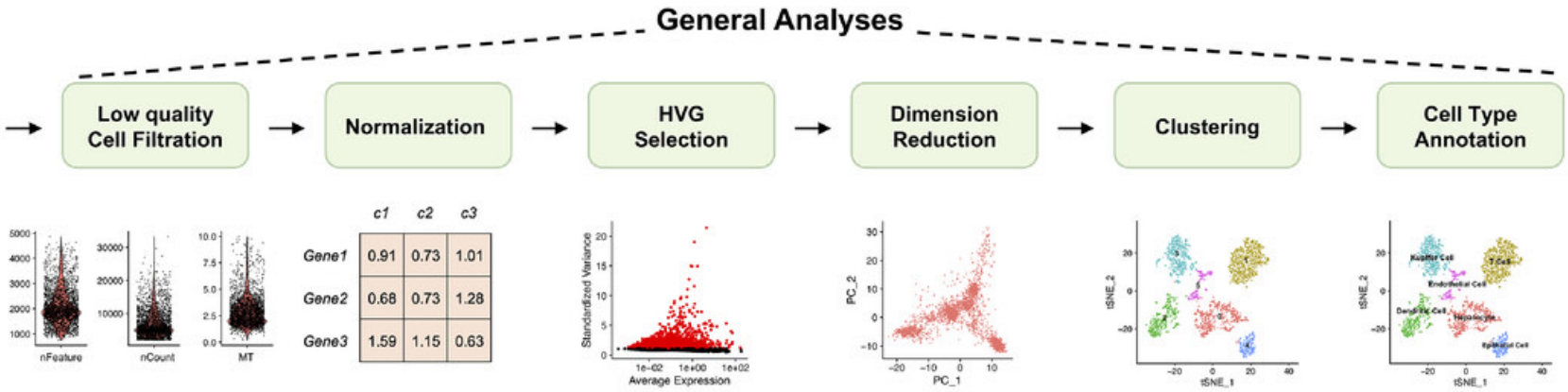
Which metrics are the most important to assess the quality of your data? Why?

If we have time, I can quickly show you how a Loupe file looks like.
Or you can try to download Loupe and open the file yourself
https://drive.google.com/drive/folders/1eKHzXlzMcl2pXlkJnLhivk-zyIrE1Kwn

https://www.10xgenomics.com/support/software/loupe-browser/downloads

https://www.10xgenomics.com/analysis-guides/quality-assessment-using-the-cell-ranger-web-summary

Downstream analysis – from QC to cell types



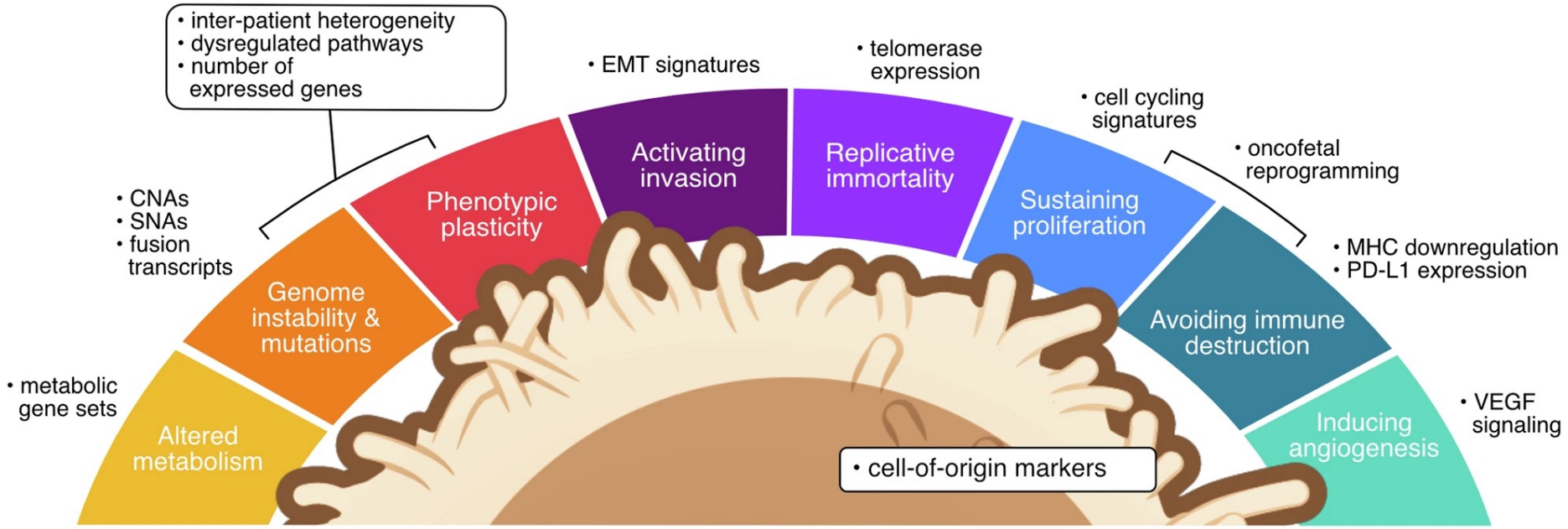
Important notes:

- There is no such thing as “one size fits all” approach in scRNAseq analysis

- During most steps of the analysis you will have to make decisions that will impact the following steps (filtering cut-offs, number of principal components, number of variable features, resolution of cell clusters, etc.)

- Knowing the biology behind your data is important.
  - Which cells types do you expect will be present?
  - How heterogeneous is your cell population?
  - Are these cells under any stress?
  - etc

Challenges on distinguishing malignant from non-malignant cells



Some tumour are characterised by profound genomic rearrangements

Inter-patient heterogeneity

If one is not careful setting up thresholds during QC, these cells might be filtered out

Downstream analysis – from QC to cell types


```r
library(Seurat)

# load data and create Seurat object
pbmc.data = Read10X(data.dir =
"../data/pbmc3k/filtered_gene_bc_matrices/hg19/")
pbmc = CreateSeuratObject(counts = pbmc.data, min.cells =
3, min.features = 200)

# normalise data
pbmc = NormalizeData(pbmc, normalization.method =
"LogNormalize", scale.factor = 10000)

# find highly variable genes
pbmc = FindVariableFeatures(pbmc, selection.method = "vst",
nfeatures = 2000)

# scale data
pbmc = ScaleData(pbmc, features = rownames(pbmc))

# linear dimensionality reduction
pbmc = RunPCA(pbmc, features = VariableFeatures(object =
pbmc))

# clustering
pbmc = FindNeighbors(pbmc, dims = 1:40)
pbmc = FindClusters(pbmc, resolution = 0.5)

# non-linear dimensionality reduction
pbmc = RunUMAP(pbmc, dims = 1:10)

# find marker genes
pbmc.mk = FindAllMarkers(pbmc, only.pos = TRUE, min.pct =
0.25, logfc.threshold = 0.25)
```


```python
import scanpy as sc

# load data and create AnnData (scanpy) object
adata =
sc.read_10x_mtx('data/filtered_gene_bc_matrices/hg19/',
var_names='gene_symbols')

# normalise data
sc.pp.normalize_total(adata, target_sum=1e4)
sc.pp.log1p(adata)

# find highly variable genes
sc.pp.highly_variable_genes(adata, min_mean=0.0125,
max_mean=3, min_disp=0.5)

# scale data
sc.pp.scale(adata, max_value=10)

# linear dimensionality reduction
sc.tl.pca(adata, svd_solver='arpack')

# clustering
sc.pp.neighbors(adata, n_neighbors=10, n_pcs=40)
sc.tl.leiden(adata)

# non-linear dimensionality reduction
sc.tl.umap(adata)

# find marker genes
sc.tl.rank_genes_groups(adata, 'leiden', method='t-test')
```

Practical exercises

`Analysis_intro.Rmd`: from count matrices to cell types

`Batch_correction.Rmd`: merging and integration of cells from 2 different donors (extra, have a look in case you have time)

https://github.com/GIMM-BioCode/2025-Autumn-School-for-Single-Cell-ers

References & further reading

https://www.nature.com/articles/nrg2484
https://cshprotocols.cshlp.org/content/early/2015/04/11/pdb.top084970.abstract
https://www.nature.com/articles/nmeth0708-585
https://www.nature.com/articles/nmeth.1226
https://www.nature.com/articles/ng.2764
https://www.science.org/doi/10.1126/science.1158441
https://www.nature.com/articles/s41576-019-0150-2
https://www.accurascience.com/blogs_3_0.html
https://www.humancellatlas.org/
https://arxiv.org/abs/1704.01379
https://onlinelibrary.wiley.com/doi/10.1002/ctm2.694
https://genomicsinform.biomedcentral.com/articles/10.1186/s44342-025-00044-5#Bib1
https://www.nature.com/articles/s41586-018-0393-7
https://www.science.org/doi/10.1126/science.aar3131
https://www.sciencedirect.com/science/article/pii/S0092867419306877?via%3Dihub
https://www.10xgenomics.com/analysis-guides/introduction-to-ambient-rna-correction
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y
https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm