



An Introduction to scRNA-seq

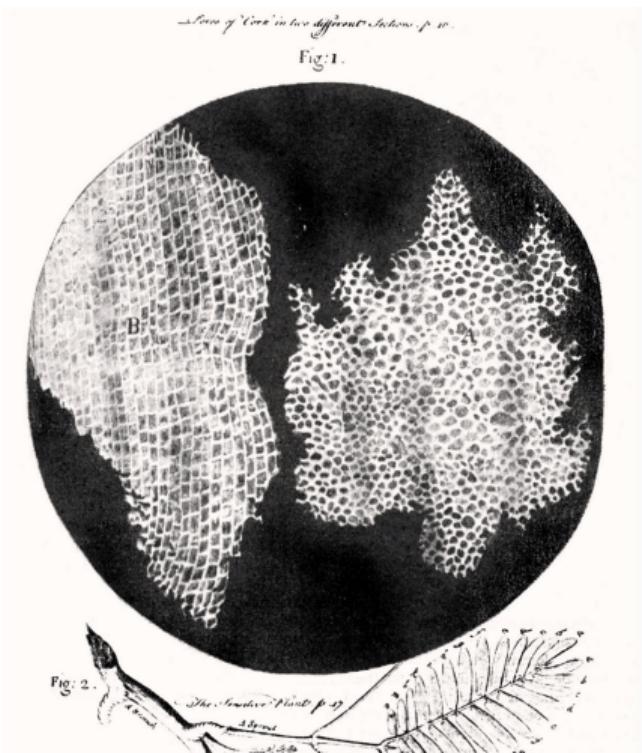
BMHAI Bootcamp

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Director, Center for Biomedical Informatics and Health AI
Rutgers University – New Jersey Medical School

2025-08-04

Cells are Important

- ▶ Fundamental unit of life
- ▶ Autonomous and unique
- ▶ Interactive
- ▶ Dynamic - change over time
- ▶ Evolution occurs on the cellular level



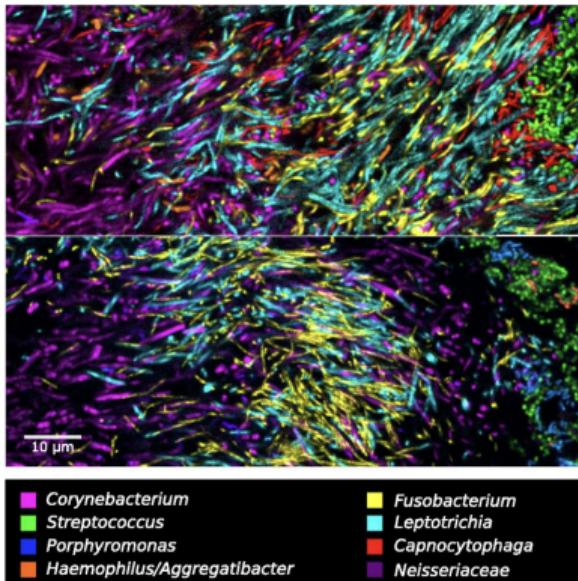
Cells are Diverse

| Type | Prokaryotes | Eukaryotes |
|--------------|---------------------|------------------------|
| Typical size | ~ 1-5 μm | ~ 10-100 μm |
| DNA form | Circular | Linear |
| DNA location | Cytoplasm | Nucleus |
| DNA amount | ~ .3-16 fg | ~3-300,000 fg |
| RNA amount | ~ 5-26,000 fg | ~ 1,000-350,000 fg |

Landenmark HKE, Forgan DH, Cockell CS (2015). An Estimate of the Total DNA in the Biosphere. PLoS Biol 13(6): e1002168.
<https://doi.org/10.1371/journal.pbio.1002168>

Cells are Diverse: Microbial Ecology

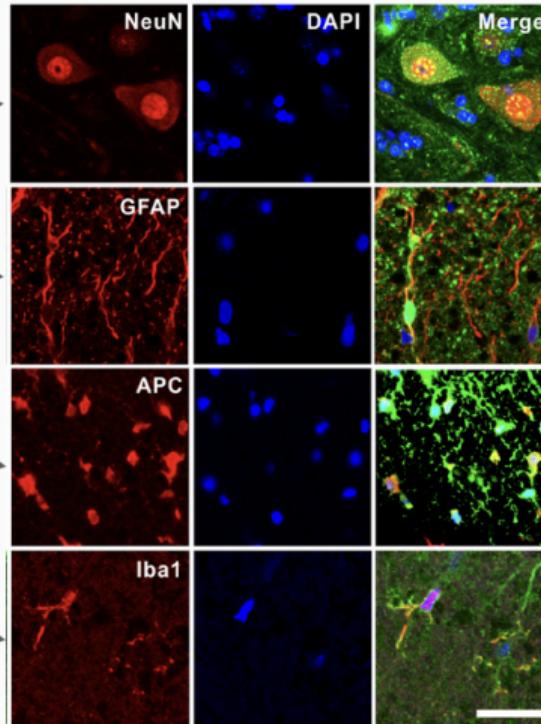
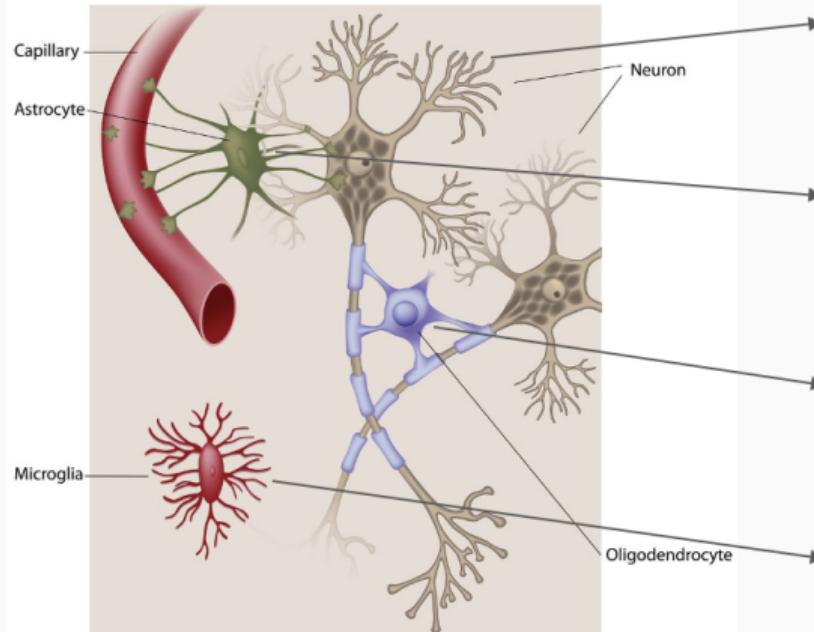
- ▶ Environments on Earth support microbial life
- ▶ Microbes usually work together in a balance
- ▶ Imbalances can disrupt the function of overall ecology
- ▶ Which specific microbes out of millions cause a particular effect?



Mark Welch, et al. 2016. "Biogeography of a Human Oral Microbiome at the Micron Scale." Proceedings of the National Academy of Sciences of the United States of America 113 (6): E791–800.

Cells are Diverse: Human Brain

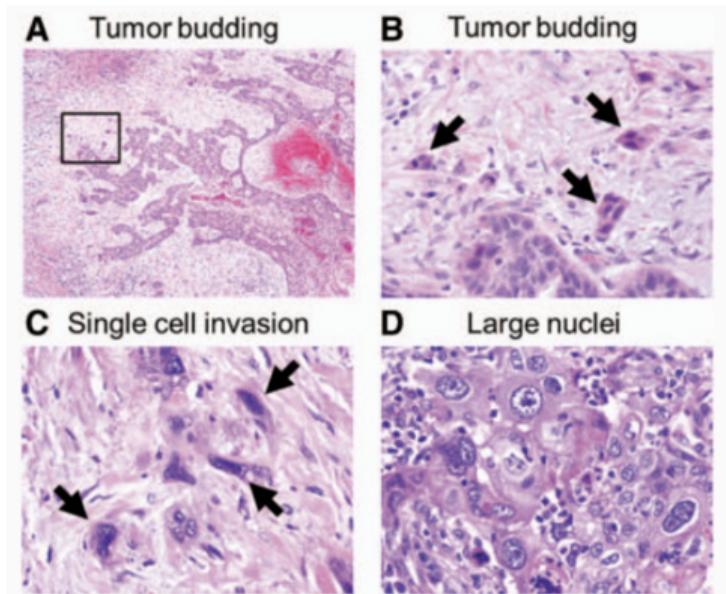
Cells of the Central Nervous System



Glia

Cells are Diverse: Tumors

- ▶ Tumors are seeded by single mutated cells
- ▶ Founder cells divide and further mutate
- ▶ Large tumors undergo angiogenesis
- ▶ Selectively kill cancer cells: cure the cancer
- ▶ But which cells to target?



Tissues are Diverse: Molecular Profiling



Bulk



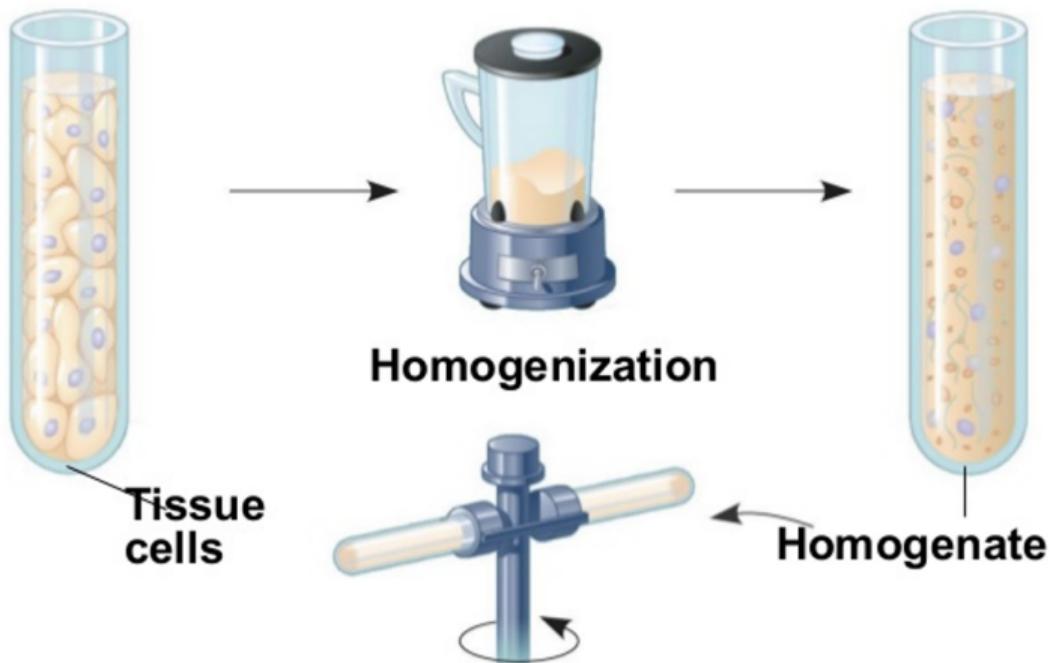
Single Cell



Spatial

The Forest: Tissue Homogenate

LE 6-5A



The Trees: Cells

- ▶ What cell types are in a sample?
- ▶ What are their proportions?
- ▶ How does their transcription differ?
- ▶ Which/how do specific cells respond to stimulus?
- ▶ How do cells develop over time?
- ▶ What is the level of mosaicism in tissues

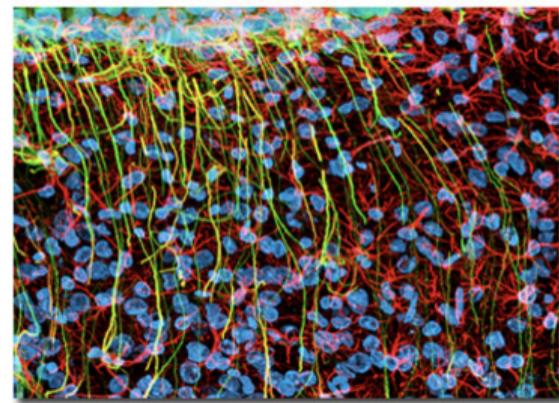
Single Cell Sequencing Workflow

1. Dissociation of tissue, isolation of cells
2. FACS sorting (optional)
3. Nucleic acid extraction and processing
4. Sequencing library prep + sequencing
5. Analysis

Dissociation of Tissue

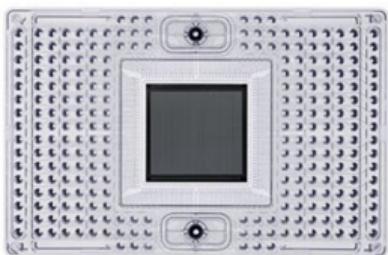
- ▶ Cells in complex tissue are highly intermingled
- ▶ Separate cells without destroying or breaking membranes
- ▶ Complex cellular morphology (e.g. neurons) makes dissociation challenging
- ▶ Can isolate nuclei instead:
 - ▶ Contain DNA/some RNA
 - ▶ Much more input material needed

Rat Brain Hippocampus Sagittal 8-Micrometer Section



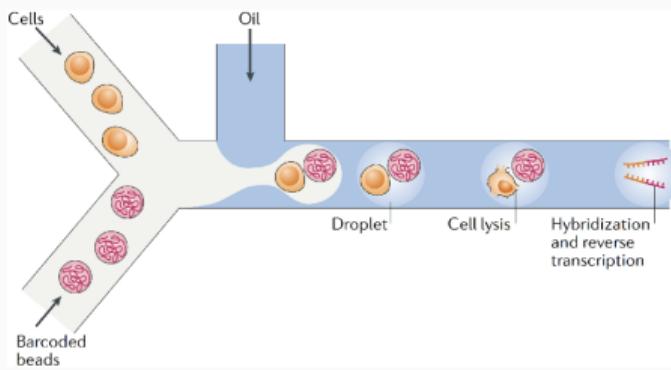
Cell Isolation Techniques

Microfluidics

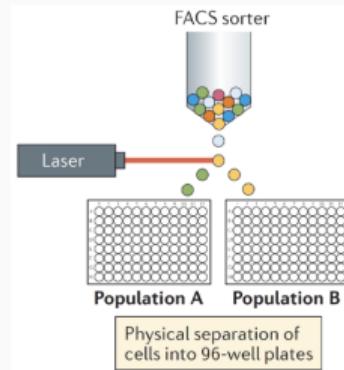


Fluidigm C1
Integrated Fluidic Circuit (IFC)

Droplet Based



Fluorescence Activated Cell Sorting (FACS)

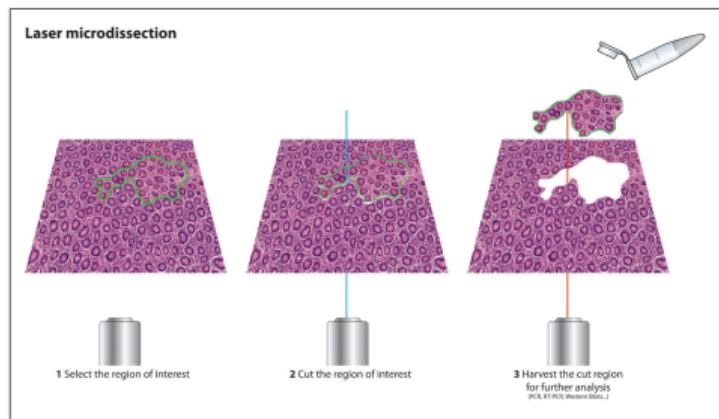


Potter, S. S. 2018.
Papalex, E. Satija, R. 2018.

Some technologies use all three!

Laser Capture Microdissection

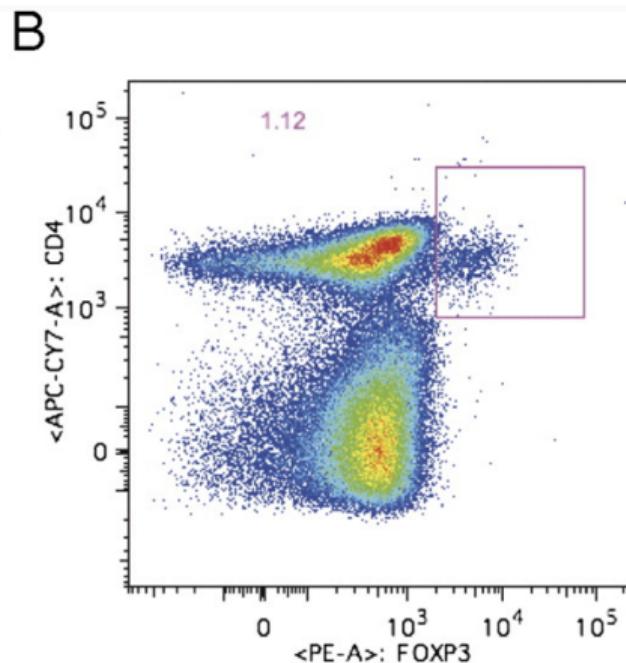
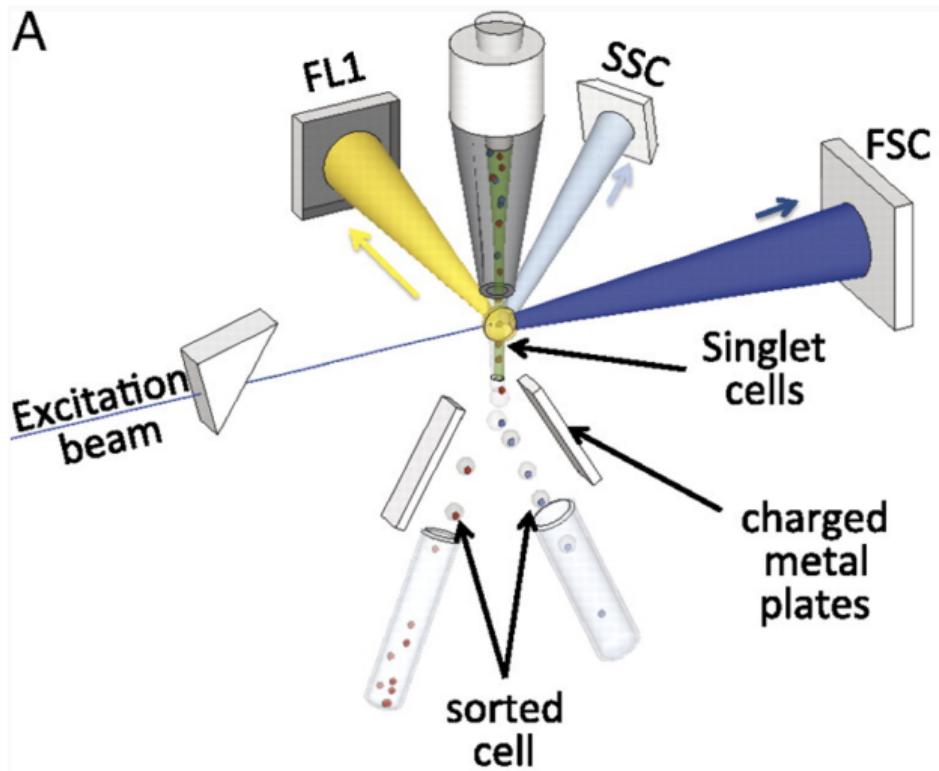
- ▶ Technique for isolating groups of cells *in situ*
- ▶ Low throughput, requires expensive equipment
- ▶ Laser causes damage to tissue and degrades RNA
- ▶ Not generally suitable for single cell sequencing



Fluorescence-Activated Cell Sorting (FACS)

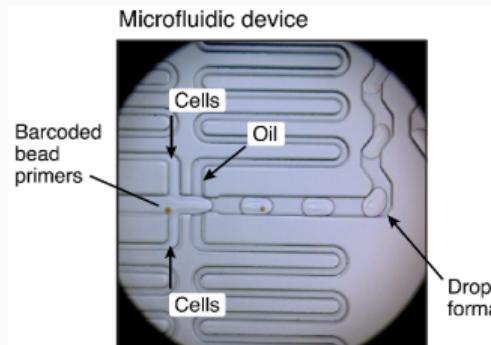
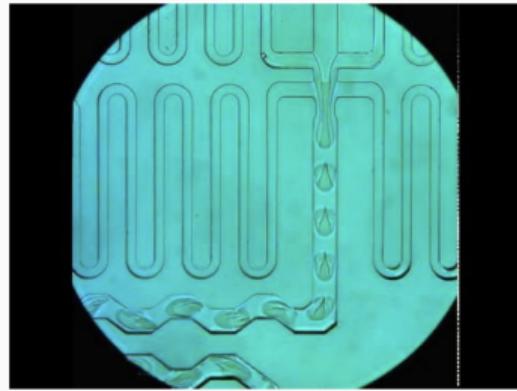
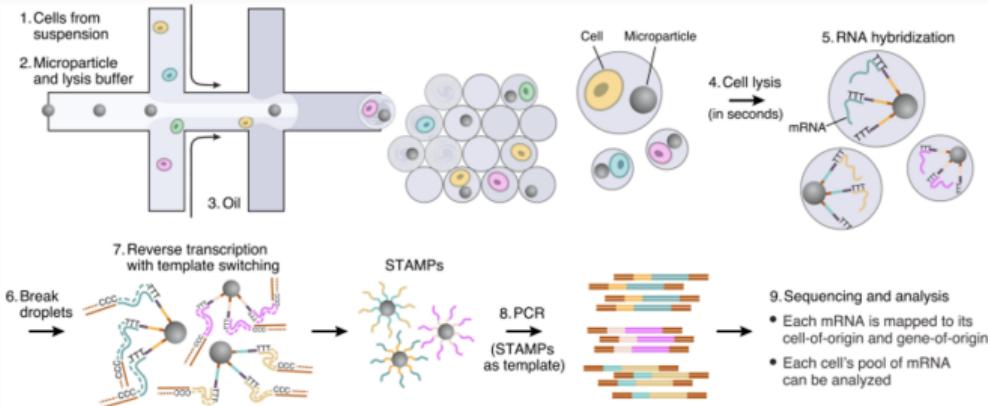
- ▶ Cells with known surface markers are tagged with fluorescent antibodies
- ▶ Tagged cells excited by lasers during flow cytometry
- ▶ Excited and non-excited cells separated and collected
- ▶ Cell type specific populations can be sequenced and studied

Fluorescence-Activated Cell Sorting (FACS)



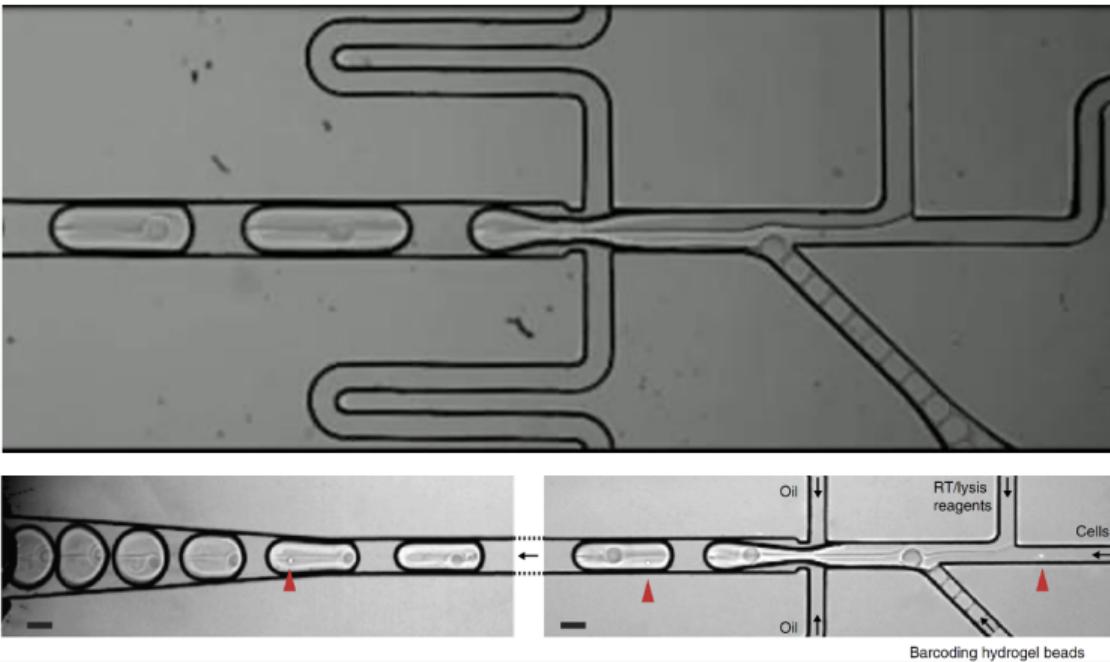
Drop-seq

- Microfluidics used to pair cells and barcodes/reagents into separate oil droplets
- Concentrations carefully controlled to get 1:1 cell/barcode matches in each oil droplet with high statistical confidence

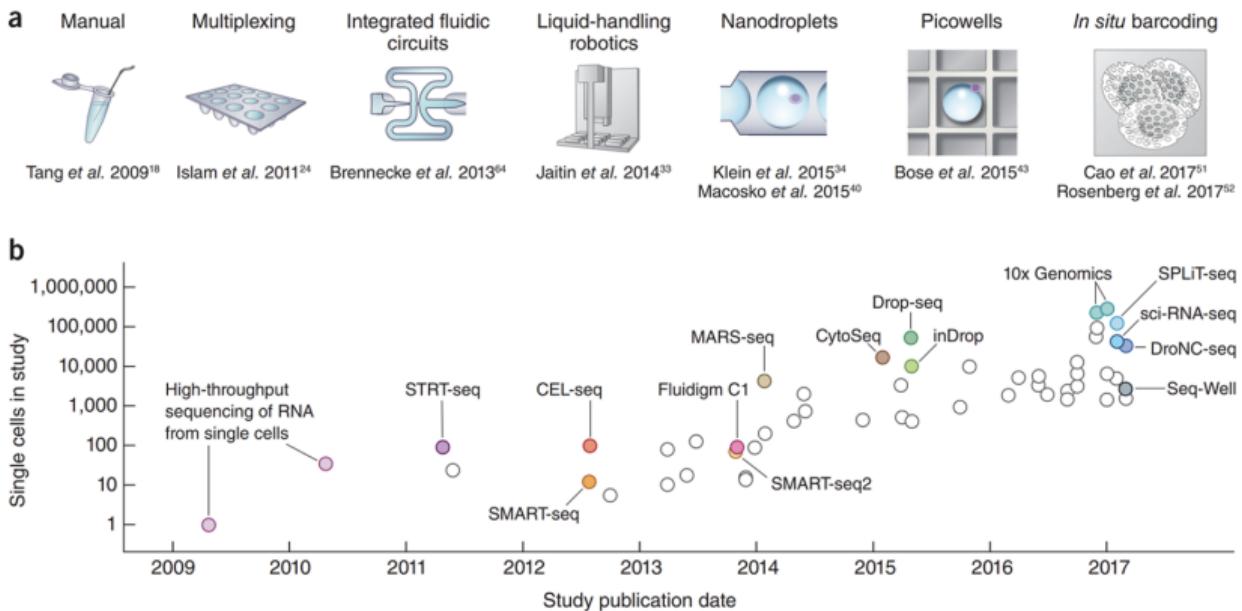


inDrops

Essentially the same strategy Drop-Seq, but uses hydrogel beads



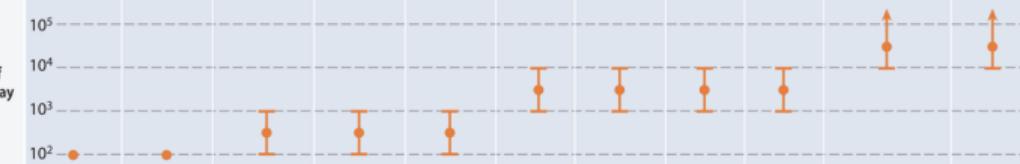
A decade of single cell RNA-seq



[Svensson *et al.* 2018. DOI:10.1038/nprot.2017.149](https://doi.org/10.1038/nprot.2017.149)

Slide by Lior Pachter and Matt Thomson

A decade of single cell RNA-seq

| | 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 | 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 |  | | | | | | | | | | |

Nucleic Acid Extraction + Processing

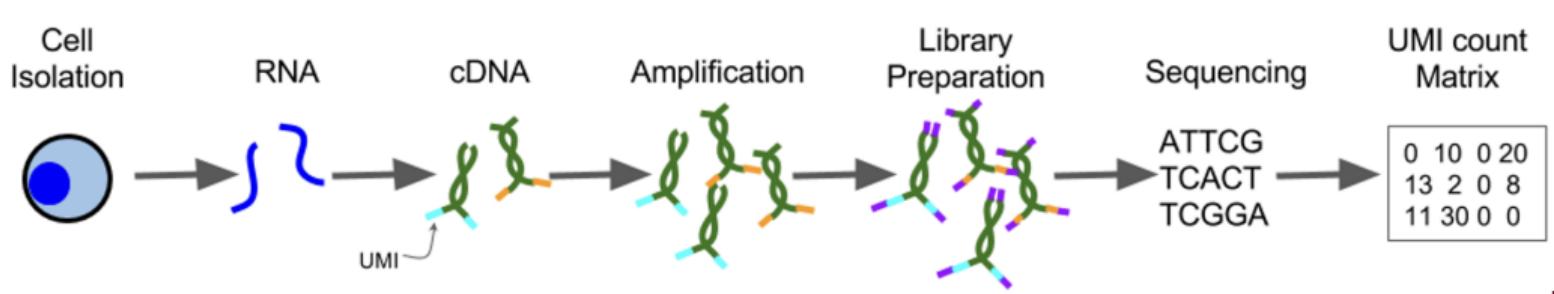
femto to picograms of input material

Each cell is:

- ▶ Assigned a unique DNA barcode
- ▶ Optionally treated with UMIs
- ▶ Amplified by one of:
 - ▶ Reverse transcriptase (RNA)
 - ▶ Multiple displacement amplification (DNA)
 - ▶ In vitro transcription (RNA)

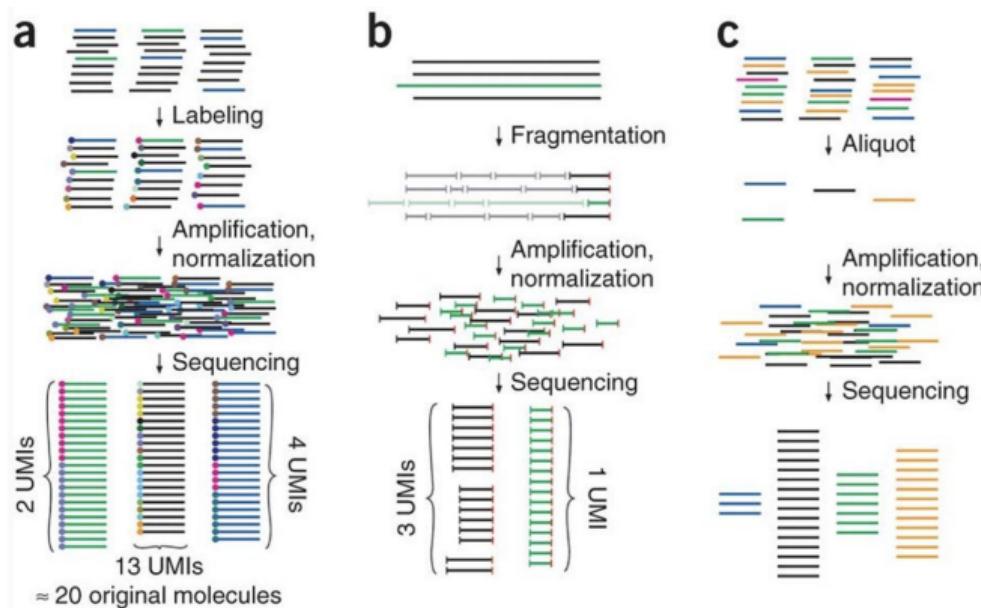
Unique Molecular Identifiers (UMIs)

- ▶ Low input material may cause amplification bias
- ▶ UMIs are sequences that correspond to one fragment
- ▶ Sequenced reads with the same UMI are from the same fragment
- ▶ Unique sequences collapsed/deduplicated for counting



Unique Molecular Identifiers (UMIs)

Strategies for counting individual molecules



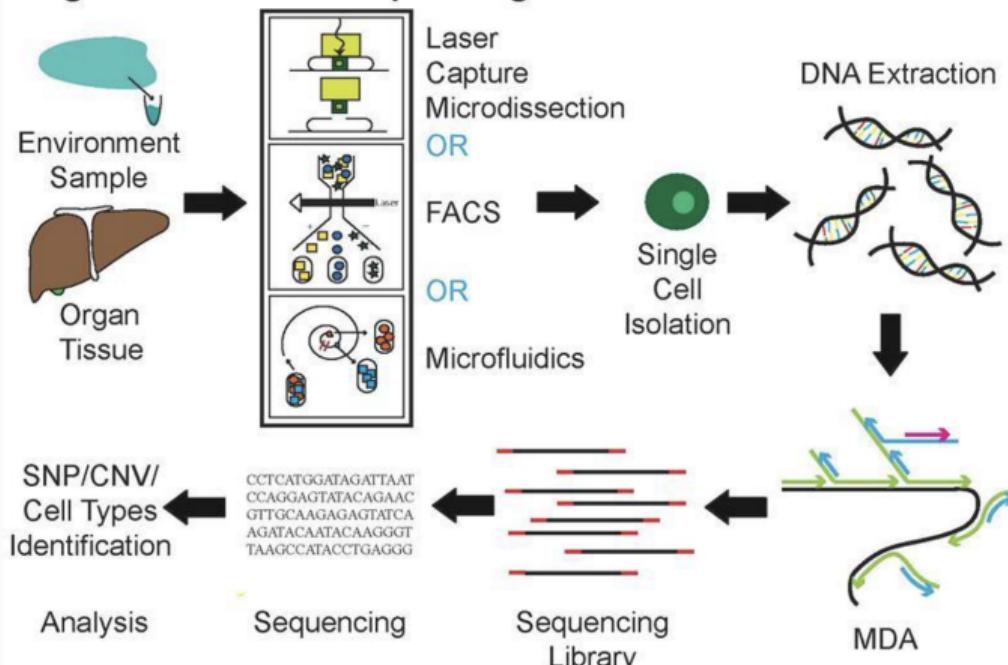
Sequencing Library Prep and Sequencing

- ▶ Previous protocols typically include sequencing primers
- ▶ Sequencing depth = (# of cells) x (required depth):
 - ▶ RNA - 50k paired end reads / cell for cell type classification
 - ▶ RNA - .25M-1M paired reads / cell for transcriptome coverage
 - ▶ DNA - 30-100x per cell
- ▶ e.g. 1000 cell scRNA-Seq = 250M-1B reads per sample!
- ▶ Sequences in one PE fastq file are entirely barcodes
- ▶ Read length > 50bp for annotated genome
- ▶ Single cell sequencing is still *very expensive*

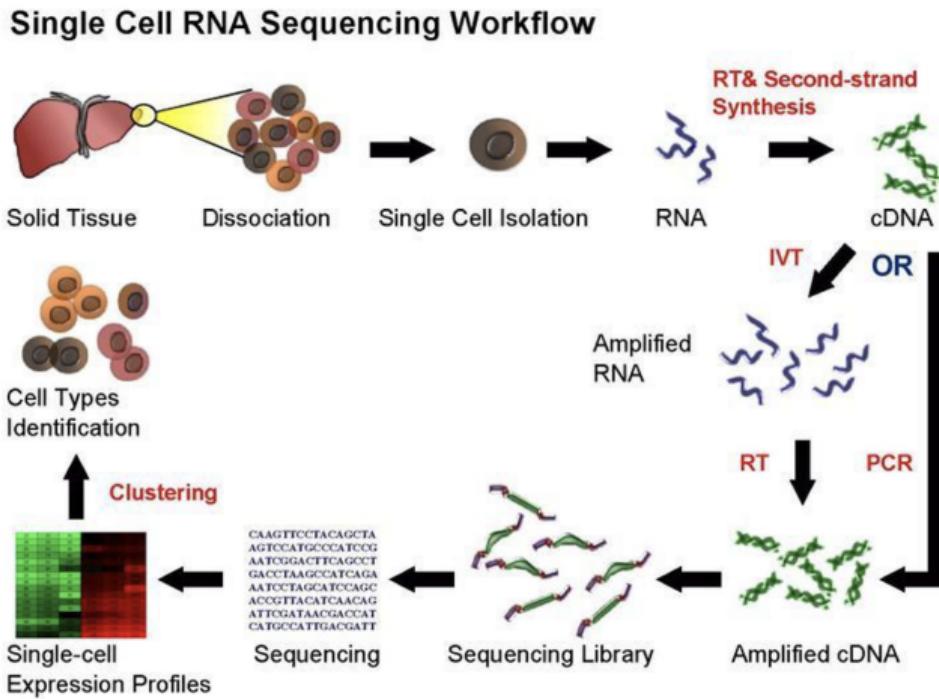
Rizzetto, et al. 2017. "Impact of Sequencing Depth and Read Length on Single Cell RNA Sequencing Data of T Cells." *Scientific Reports* 7 (1): 12781.

scDNA-Seq Workflow

Single Cell Genome Sequencing Workflow

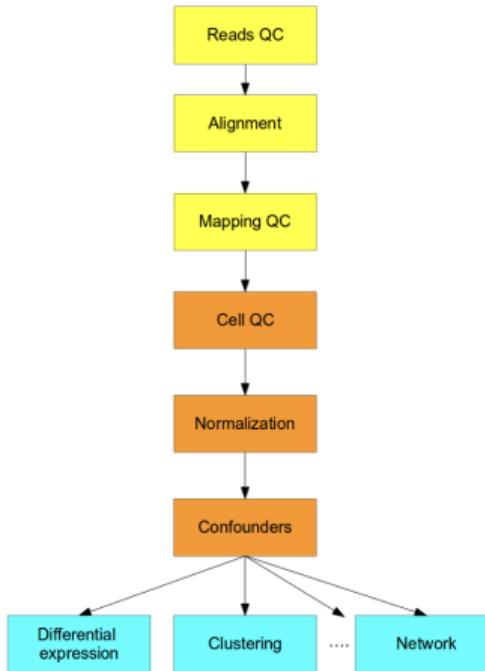


scRNA-seq workflow

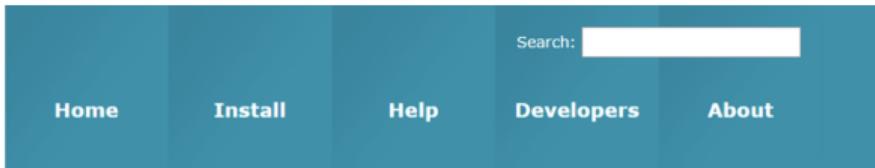


Analysis Overview

1. Sequence QC
 - 1.1 Demultiplex
 - 1.2 UMI Collapsing
2. Alignment
3. Quantification
4. Normalization
5. DE, Clustering, etc



scruff R/Bioconductor package



The Bioconductor header features a logo with three blue DNA helixes above the word "Bioconductor". Below it, the text "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS" is displayed. A search bar is located in the top right corner. The main navigation menu includes "Home", "Install", "Help", "Developers", and "About".

[Home](#) » [Bioconductor 3.8](#) » [Software Packages](#) » scruff

scruff

platforms all rank unknown posts 0 in Bioc < 6 months
build ok updated < 1 month

DOI: [10.1186/B9.bioc.scruff](https://doi.org/10.1186/B9.bioc.scruff)  

Single Cell RNA-Seq UMI Filtering Facilitator (scruff)

Bioconductor version: Release (3.8)

A pipeline which processes single cell RNA-seq (scRNA-seq) reads from CEL-seq and CEL-seq2 protocols. Demultiplex scRNA-seq FASTQ files, align reads to reference genome using Rsubread, and generate UMI filtered count matrix. Also provide visualizations of read alignments and pre- and post-alignment QC metrics.

Author: Zhe Wang [aut, cre], Junming Hu [aut], Joshua Campbell [aut]

Maintainer: Zhe Wang <zhe at bu.edu>

Citation (from within R, enter `citation("scruff")`):

Documentation »

Bioconductor

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [BioC-devel mailing list](#) - for package

Sequence QC

One sample is 100s or 1000s of cells

- ▶ i.e. ~1,000 fastq files per sample
- ▶ May or may not be already demultiplexed by core

UMI-Tools - open source UMI software

Normal fastq processing and QC:

- ▶ Adapter and quality trimming
- ▶ fastqc, multiqc

Alignment and Quantification

STAR+htseq-count, kallisto, salmon, CellRanger

Each sample has a different # of cells

Each cell has the same number of measurements (e.g. genes) = (# of samples) x (# of cells) x (# of genes)

Sparse: most will be zero!

Amarel Submission Script for CellRanger

```
#!/bin/bash
#SBATCH --partition=main
#SBATCH --job-name=cellranger_bam
#SBATCH --array=0-3,5,6
#SBATCH --cpus-per-task=30
#SBATCH --mem=100G
#SBATCH --time=24:00:00

# path to fastq files
FASTQPATH=/scratch/$USER/tmp/awsbucket/fastqs/

# get sample to process
INDEX=$($SLURM_ARRAY_TASK_ID)
INPUT=(${ls -d $FASTQPATH*R1_001.fastq.gz})
FASTQ=(${echo ${INPUT[$INDEX]##*/} | cut -d_ -f1-1})

# Path to cellranger
crpath=/projects/f_wj183_1/apps/cellranger-8.0.1/
```

Amarel Submission Script for CellRanger

```
# path to reference library
refpath=/projects/f_wj183_1/reflib/2024_cellranger/refdata-gex-GRCh38-2024-A/
cd $FASTQPATH

# load python
module load python/3.8.2

$crpath./cellranger count --id=$FASTQ \
  --create-bam=true \ # true or false, necessary
  --sample=$FASTQ \ # prefix of files to align
  --fastqs=$FASTQPATH \
  --localcores=30 \
  --localmem=100 \
  --chemistry=SC3Pv2 \ # optional
  --transcriptome=$refpath
```

The Counts Matrix

- ▶ Counts matrix contains either:
 - ▶ Read counts or
 - ▶ UMI counts if used
- ▶ Each cell has:
 - ▶ Total number of counts (col. sum, “library size”)
 - ▶ Number of non-zero genes
- ▶ Each gene has:
 - ▶ number of non-zero cells
 - ▶ Non-zero mean/variance
- ▶ Matrix is sparse: many zeros
- ▶ Zeros may be:
 - ▶ Cell lacks gene
 - ▶ A “drop-out”: gene present but was missed by qPCR

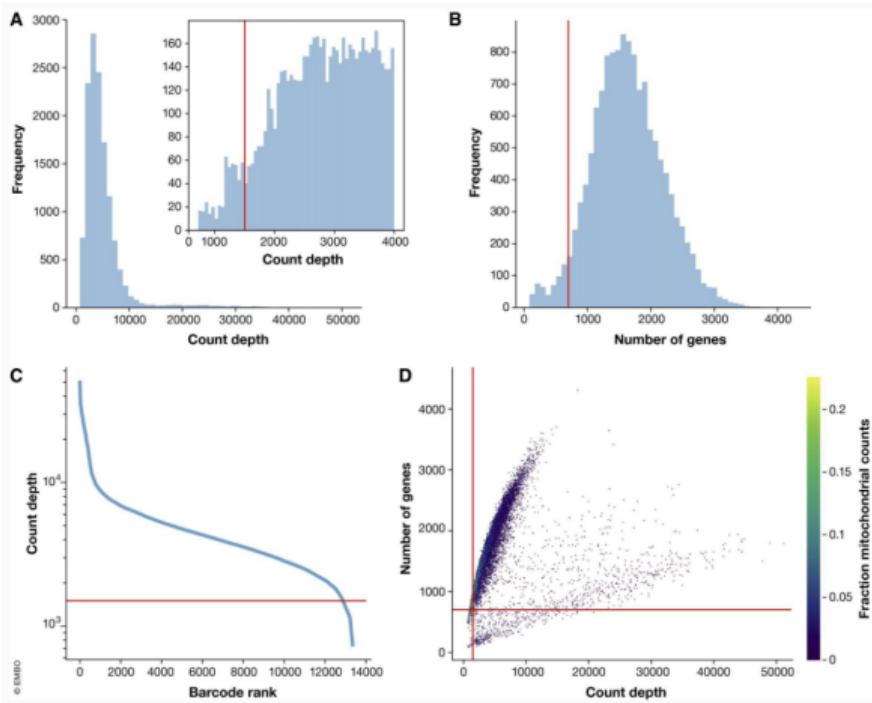
The Counts Matrix

| | cell1 | cell2 | cell3 | cell4 | cell5 | cell6 | ... | cellM |
|-------|-------|-------|-------|-------|-------|-------|-----|-------|
| gene1 | 93 | 25 | 0 | 0 | 3335 | 0 | | 82 |
| gene2 | 5 | 2 | 0 | 3 | 1252 | 0 | | 12 |
| gene3 | 0 | 0 | 0 | 0 | 0 | 0 | | 0 |
| gene4 | 98 | 21 | 1 | 1 | 5318 | 0 | | 75 |
| gene5 | 0 | 0 | 513 | 0 | 0 | 325 | | 135 |
| gene6 | 0 | 0 | 113 | 0 | 1 | 497 | | 255 |
| gene7 | 3 | 0 | 0 | 0 | 6 | 0 | | 0 |
| ... | | | | | | | | |
| geneN | 68 | 52 | 0 | 2 | 4313 | | | 63 |

Cell Quality Control

- ▶ Consider metrics jointly and threshold by outliers in distribution
- ▶ Most common cell QC metrics are:
 - ▶ Count depth
 - ▶ Number of genes
 - ▶ Fraction of mitochondrial genes

Cell Quality Control



Cell QC metrics should be considered together

Example 1:

Cell with high % of MT genes may represent:

- ▶ Cell subtypes highly engaged in respiratory processes (muscle, fat, etc.)
- ▶ Broken cells where the cytoplasmic mRNA has escaped, leaving only the MT RNA
- ▶ Cell with low count depth, few genes, and high % of MT genes may be a better marker of true “broken” cell events

Cell QC metrics should be considered together

Example 2:

Cell with abnormally high count depth and large amount of genes may represent:

- ▶ Doublets (droplets or events containing multiple cells)
- ▶ Larger cells (cell size can correlate with counts)
- ▶ Newly developed tools specifically handle doublets and attempt to resolve these cases (Scrublet, Doublet Finder, etc)

Cell QC metrics should be considered together

Example 3:

Cell with low counts and low amount of genes may represent:

- ▶ “Empty” droplets or events
- ▶ Quiescent cell populations

Additional QC Checks

- ▶ Remove genes that are only expressed in a few cells
 - ▶ This naturally sets a limit on the size of cell clusters you can recover
 - ▶ i.e. Removing genes expressed in fewer than 20 cells will make it difficult to detect clusters smaller than this
- ▶ Account for ambient gene expression (contaminating mRNA from lysed cells prior to library construction)

Filtering Cells and Genes

Many measurements

- ▶ e.g. 30k genes \times 1ks of cells

Some cells are uninformative, e.g.:

- ▶ Very few reads, few genes detected
- ▶ Two cells sequenced together (i.e. doublets)

Some genes are uninformative:

- ▶ Low # reads, low variance across all cells
- ▶ Too few cells express gene (e.g. < 10 of 10,000 cells nonzero)

Must filter genes *and* cells to reduce noise

Filtering the Counts Matrix

Cells might also be filtered:

- ▶ Very few or zero counts (cell4)
 - ▶ Empty well?
 - ▶ Ambient expression?
 - ▶ Quiescent cell?
- ▶ Very many counts (cell5)
 - ▶ Possible “doublet” of same cell type
- ▶ Inconsistent expression pattern (cellM)
 - ▶ Possible “doublet” of different cell types

Doublet: two cells with same cell barcode

Filtering the Counts Matrix

| | cell1 | cell2 | cell3 | cell4 | <i>cell5</i> | cell6 | ... | cellM |
|-------|-------|-------|-------|--------------|--------------|-------|-----|-------|
| gene1 | 93 | 25 | 0 | 0 | 3335 | 0 | | 82 |
| gene2 | 5 | 2 | 0 | 3 | 1252 | 0 | | 12 |
| gene3 | 0 | 0 | 0 | 0 | <i>0</i> | 0 | | 0 |
| gene4 | 98 | 21 | 1 | 1 | <i>5318</i> | 0 | | 75 |
| gene5 | 0 | 0 | 513 | 0 | <i>0</i> | 325 | | 135 |
| gene6 | 0 | 0 | 113 | 0 | 1 | 497 | | 255 |
| gene7 | 3 | 0 | 0 | 0 | 6 | 0 | | 0 |
| ... | | | | | | | | |
| geneN | 68 | 52 | 0 | 2 | <i>4313</i> | | | 63 |

Session info

```
sessionInfo()

## R version 4.5.1 (2025-06-13)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sequoia 15.5
##
## Matrix products: default
## BLAS:    /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRblas.0.dylib
## LAPACK:  /Library/Frameworks/R.framework/Versions/4.5-arm64/Resources/lib/libRlapack.dylib;  LAPACK version 3.12.1
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats      graphics   grDevices  utils       datasets   methods    base
##
## loaded via a namespace (and not attached):
## [1] compiler_4.5.1    fastmap_1.2.0    cli_3.6.5      tools_4.5.1
## [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10    rmarkdown_2.29
## [9] knitr_1.50       xfun_0.52       digest_0.6.37   rlang_1.1.6
## [13] evaluate_1.0.4
```