UCLA QCBio

Collaboratory

W20: Single Cell RNA-Sequencing Analysis using R

Giovanni Quinones Valdez, PhD

November 25th – 27th, 2024

Workshop Structure

- This is an introductory-level scRNA-Seq workshop.
  - We will learn about the entire pipeline (from cells to data), but we won't expand too much on the details.
- Basic knowledge of R programming is expected.
- Students taking this workshop for credits will be assigned homework and a quiz on the last day (**November 27th**).

W20: Single Cell RNA-Sequencing Analysis using R

Day 1

- Concepts in scRNA-Seq
- Data Exploration
- Quality Control

Day 2

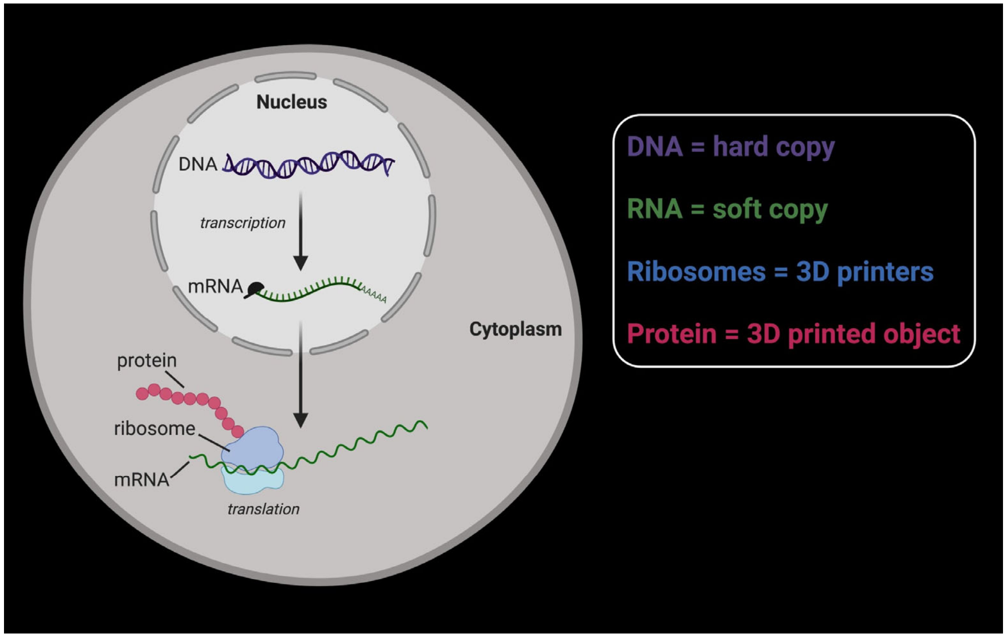
- Data processing
- Clustering and visualization
- Cell annotation

Day 3

- Pseudo-time
- Data integration
- Perspectives

UCLA QCBio
Collaboratory

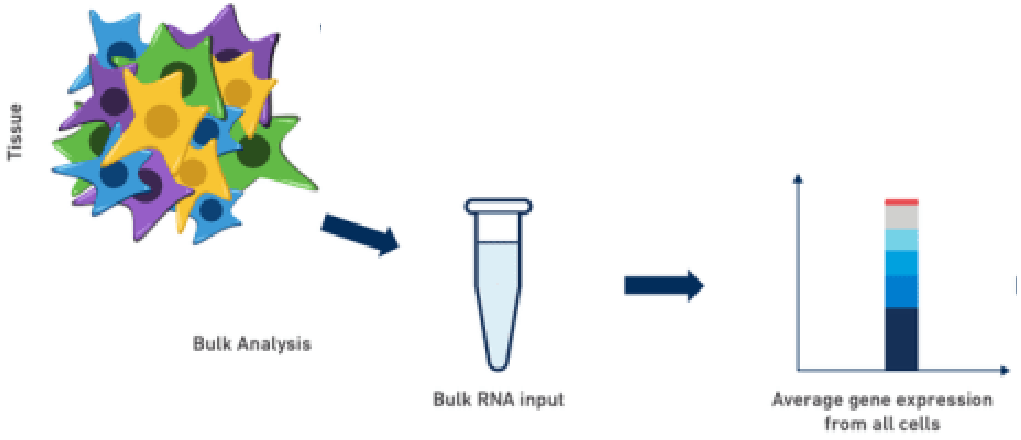
What is RNA-Sequencing?



Sciencein3.com

- RNA sequencing is the “reading” of the RNA molecules present in the cell.
- We can observe where these molecules come from (genes); how many there are (gene expression) and what they look like (variants, RNA processing)

Bulk RNA-Seq vs single cell RNA-Seq



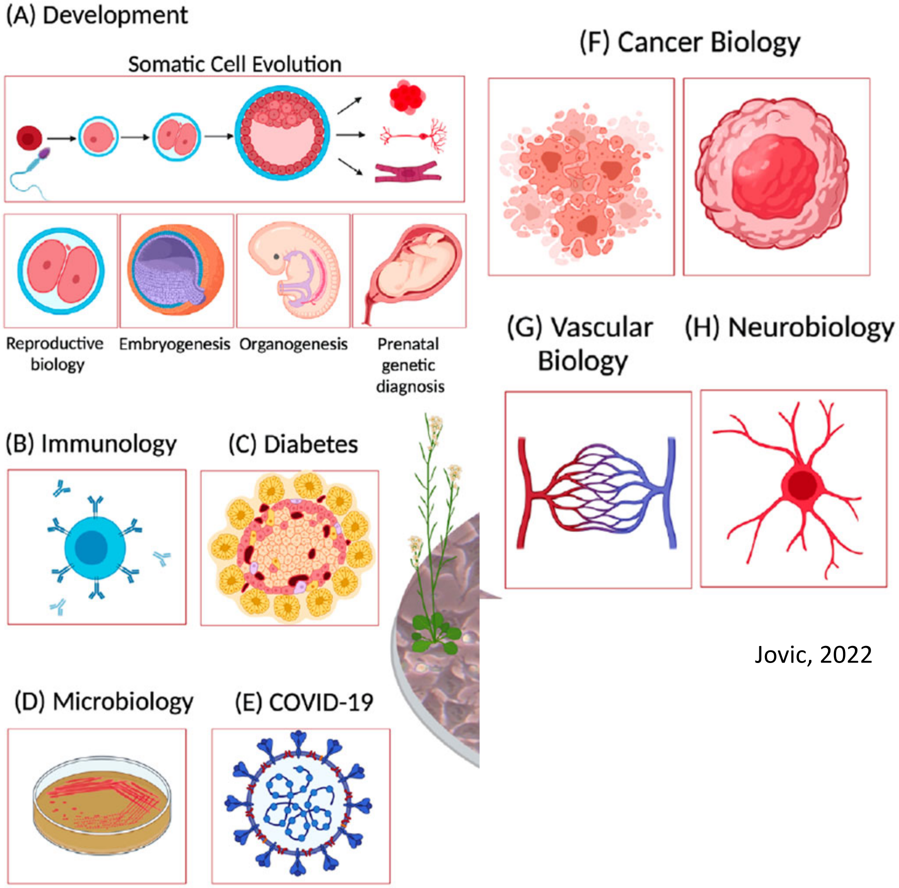
10X Genomics

Bulk RNA-Seq obscures cell-to-cell variability

UCLA QCBio
Collaboratory

Applications

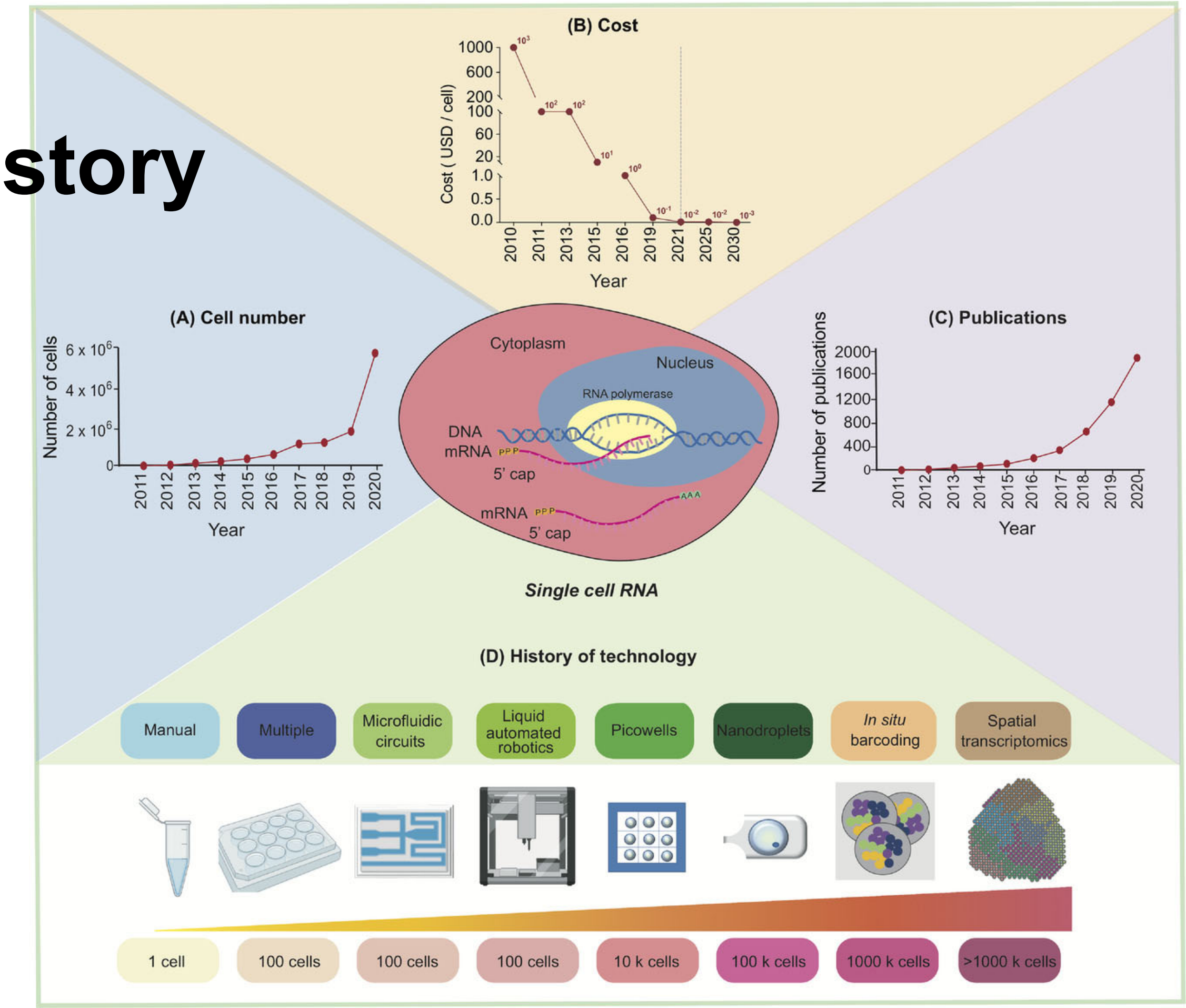
- Study and identify cellular heterogeneity → cell populations within a tissue.
- Discover new cell types.
- Discover new markers and regulatory pathways.
- Reconstruct cellular lineage.



Jovic, 2022

UCLA QCB10 Collaboratory

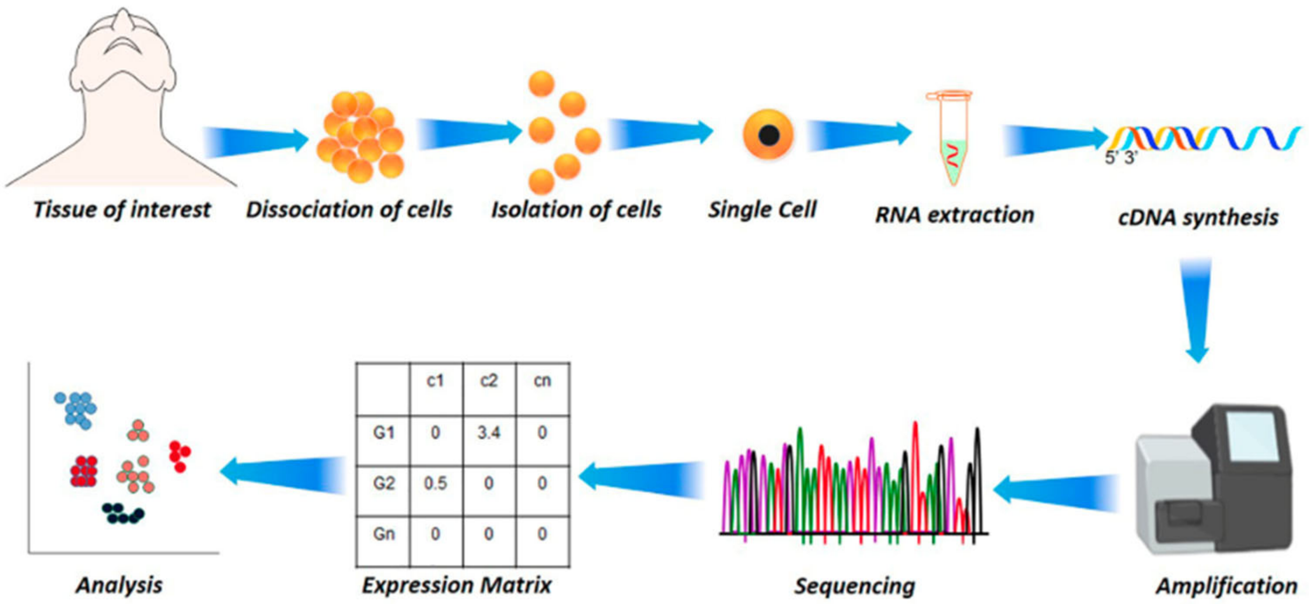
A bit of history



Jovic, 2022

UCLA QCBio Collaboratory

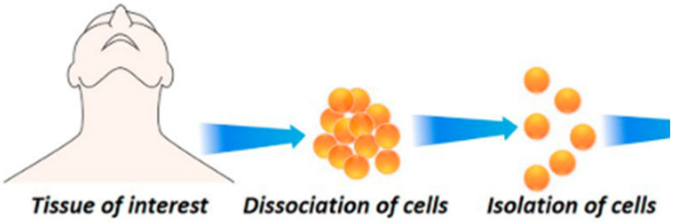
Experimental design and analysis

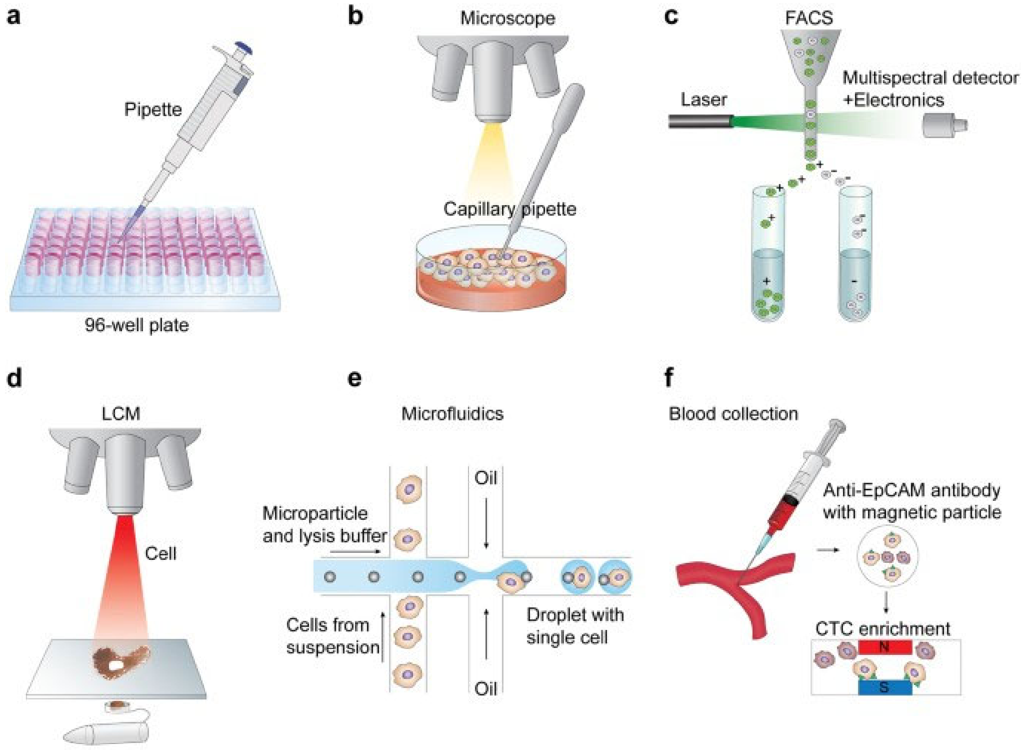


	c1	c2	cn
G1	0	3.4	0
G2	0.5	0	0
Gn	0	0	0

Adil et al (2021)

1. Methods for cell isolation

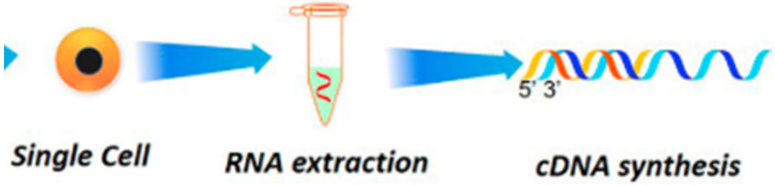




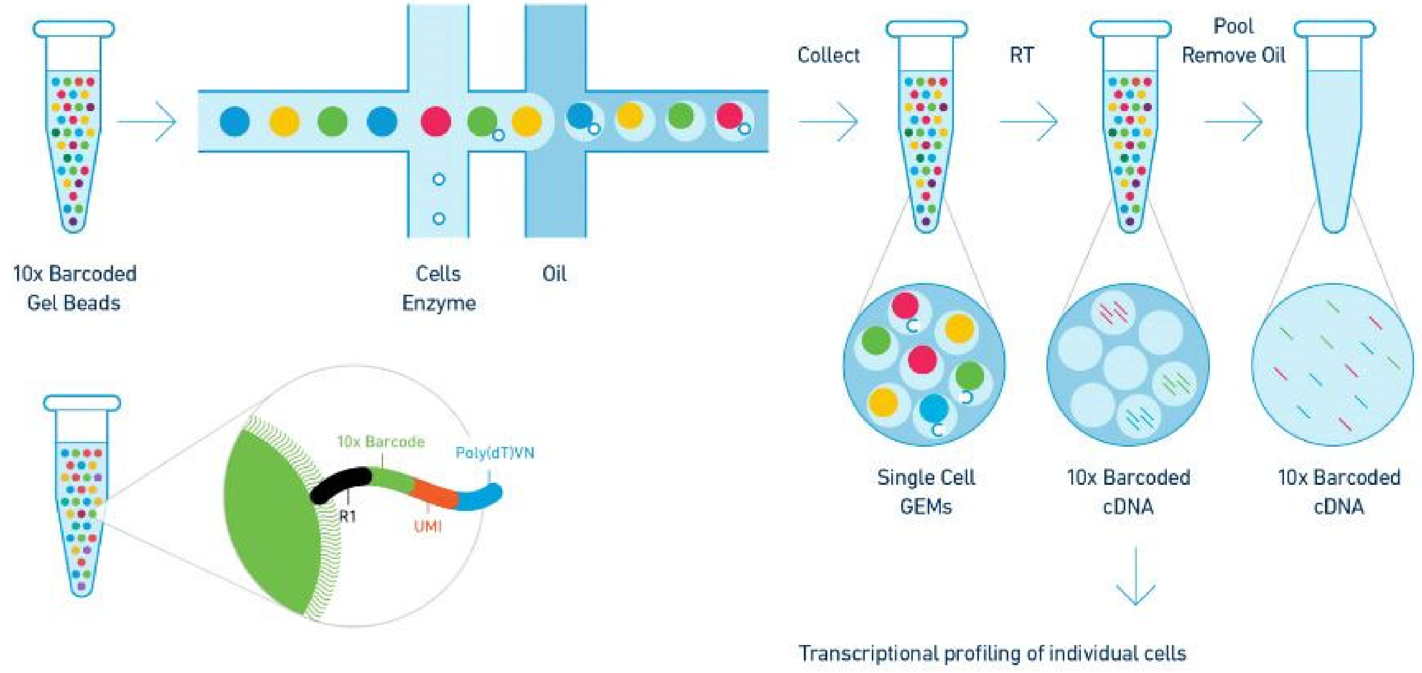
Hwang et al (2018)

UCLA QCBio Collaboratory

2. RNA extraction and cDNA synthesis



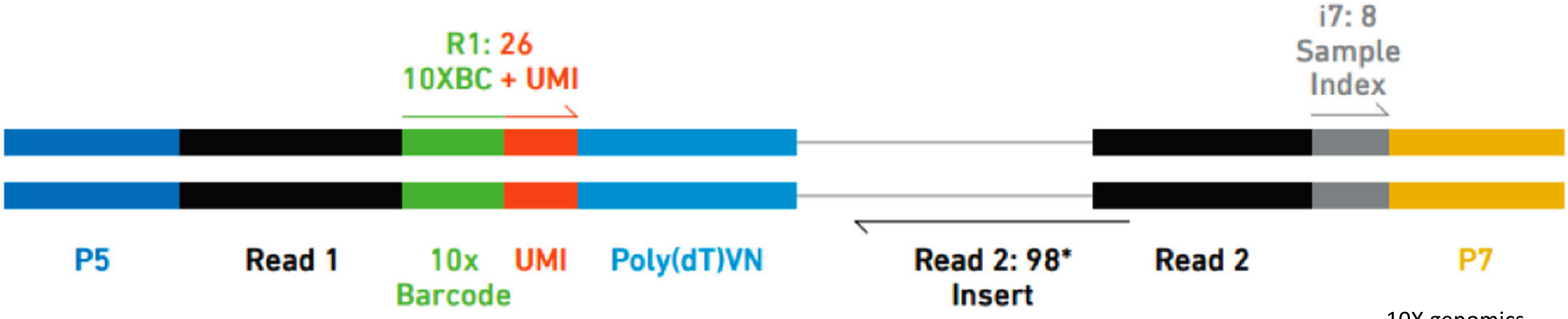
10X Genomics Protocol



10X genomics

UCLA QCBio
Collaboratory

10X Genomics



10X genomics

- **Barcode** = cell's id (length = 16)
- **UMI** = RNA molecule's id (length = 10)
- **Sample index** = Sample's id (length = 8)
- **P5**/**P7** = illumina adapters (necessary for sequencing)

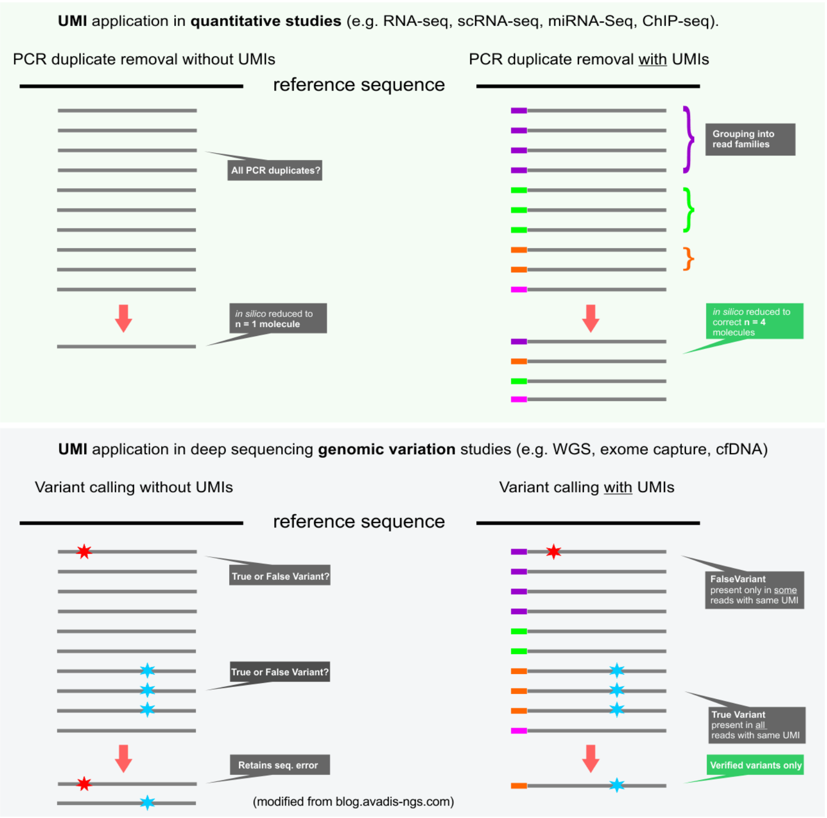
$$unique\ oligos = 4^n$$

UCLA QCBio
Collaboratory

Unique Molecular Identifiers (UMI)

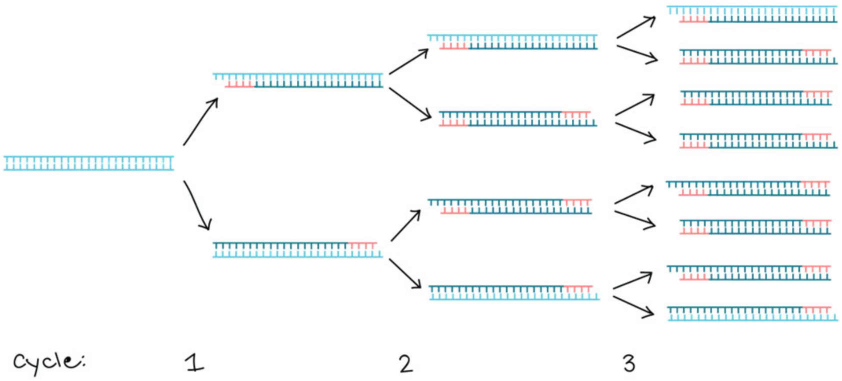
- They serve to remove PCR duplicates.
- They serve to remove in identifying sequencing errors.

Confident analysis of reads sharing the same alignment coordinates.


UMI application in **quantitative studies** (e.g. RNA-seq, scRNA-seq, miRNA-Seq, ChIP-seq).
PCR duplicate removal without UMIs
PCR duplicate removal with UMIs
reference sequence
All PCR duplicates?
Grouping into read families
in silico reduced to **n = 1 molecule**
in silico reduced to correct **n = 4** molecules
UMI application in deep sequencing **genomic variation** studies (e.g. WGS, exome capture, cfDNA)
Variant calling without UMIs
Variant calling with UMIs
reference sequence
True or False Variant?
True or False Variant?
FalseVariant present only in some reads with same UMI
True Variant present in all reads with same UMI
Retains seq. error
Verified variants only

(modified from blog.avadis-ngs.com)

3. Amplification

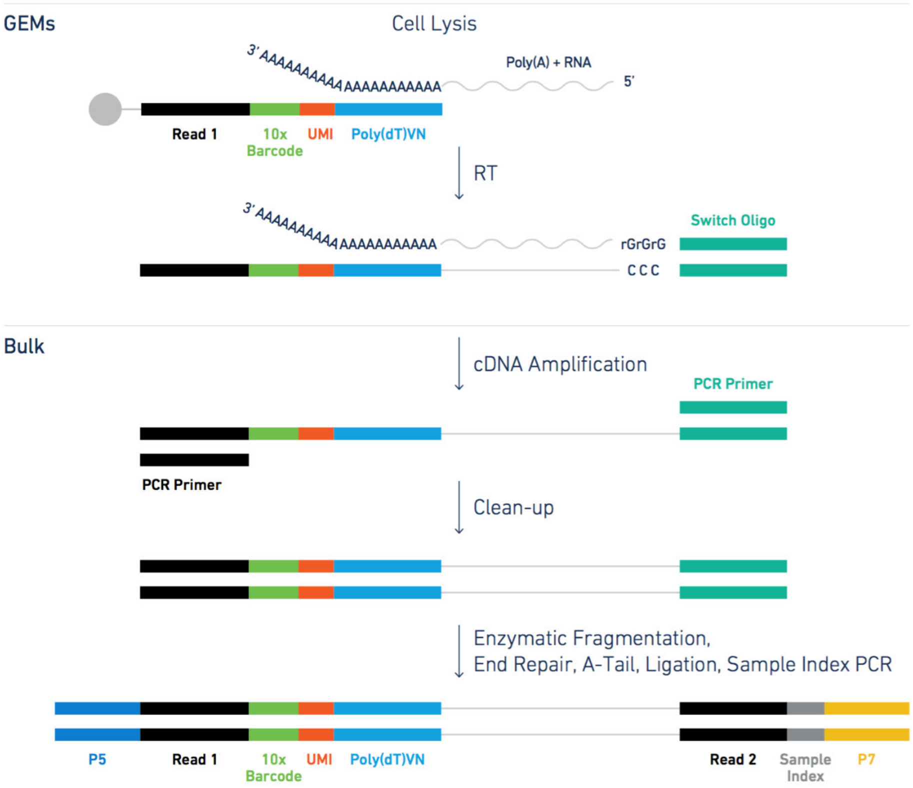


Khan academy

- The amount of RNA required for successful signal detection is 0.1-1.0 μg.
- The amount of RNA present in a single cell is 1-50 pg (2,000 to 1 million times less).
- Not all RNA molecules are captured. Droplet based technologies capture only around 5-8% of RNAs.
- The most-commonly used methods for amplification are:
  - PCR (Polymerase Chain Reaction)
  - IVT (In-vitro transcription)

UCLA QCBio Collaboratory

10X Genomics



10X genomics

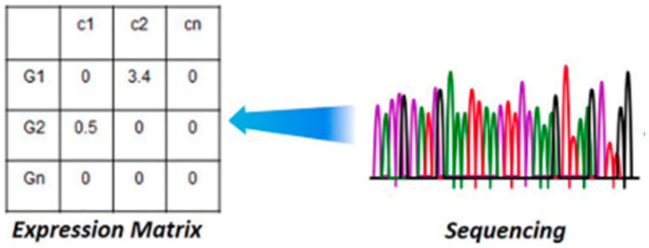
UCLA QCBio Collaboratory

4. Sequencing

Sequencing Machine

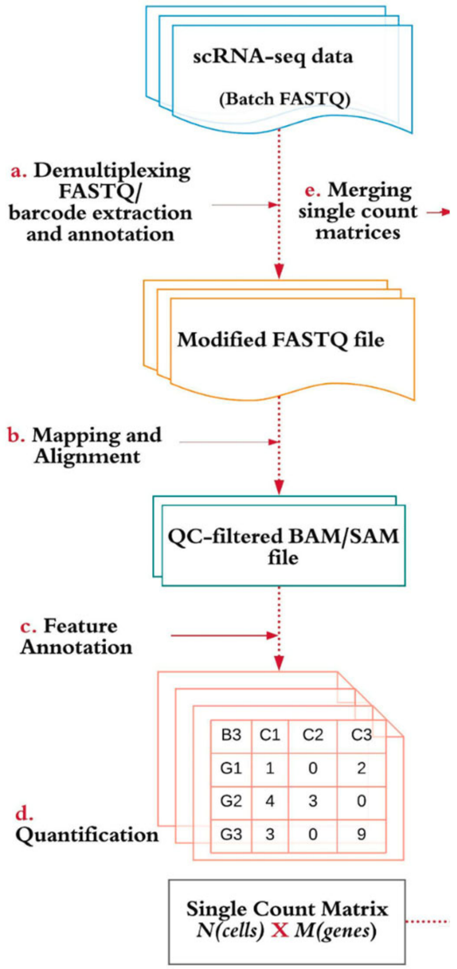
- Sequencing is performed on all the RNAs from all the cells together (multiplexing).
- Each molecule contains labels to indicate their origin (which cell) called barcodes.

5. Raw data processing



	c1	c2	cn
G1	0	3.4	0
G2	0.5	0	0
Gn	0	0	0

- Software like CellRanger and STARsolo automate this step
- RNA-Seq I and RNA-Seq II workshop covers this in more detail!



scRNA-seq data (Batch FASTQ)

a. Demultiplexing FASTQ/ barcode extraction and annotation

e. Merging single count matrices

Modified FASTQ file

b. Mapping and Alignment

QC-filtered BAM/SAM file

c. Feature Annotation

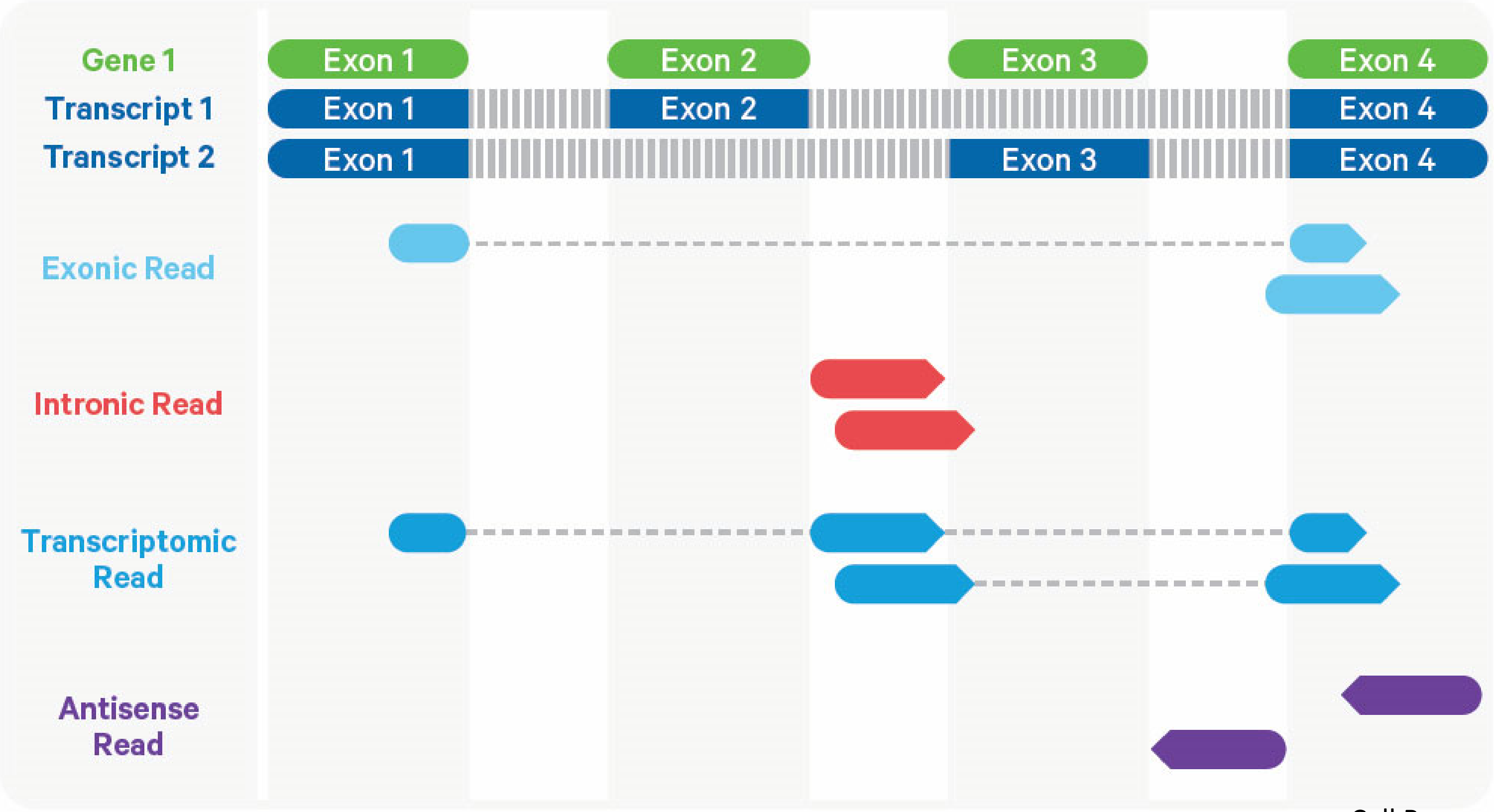
d. Quantification

B3	C1	C2	C3
G1	1	0	2
G2	4	3	0
G3	3	0	9

Single Count Matrix $N(cells) \times M(genes)$

Nayak et al 2021

Read mapping



Cell Ranger

UCLA QCBio Collaboratory

Cell ranger barcode and UMI processing

How many cells
are there in my data?

Are these two different cells?

- **BC1: ATTCCGTTAGCCGACG (100k UMIs linked)**
- **BC2: ATTCCCTTAGCCGACG (30 UMIs linked)**

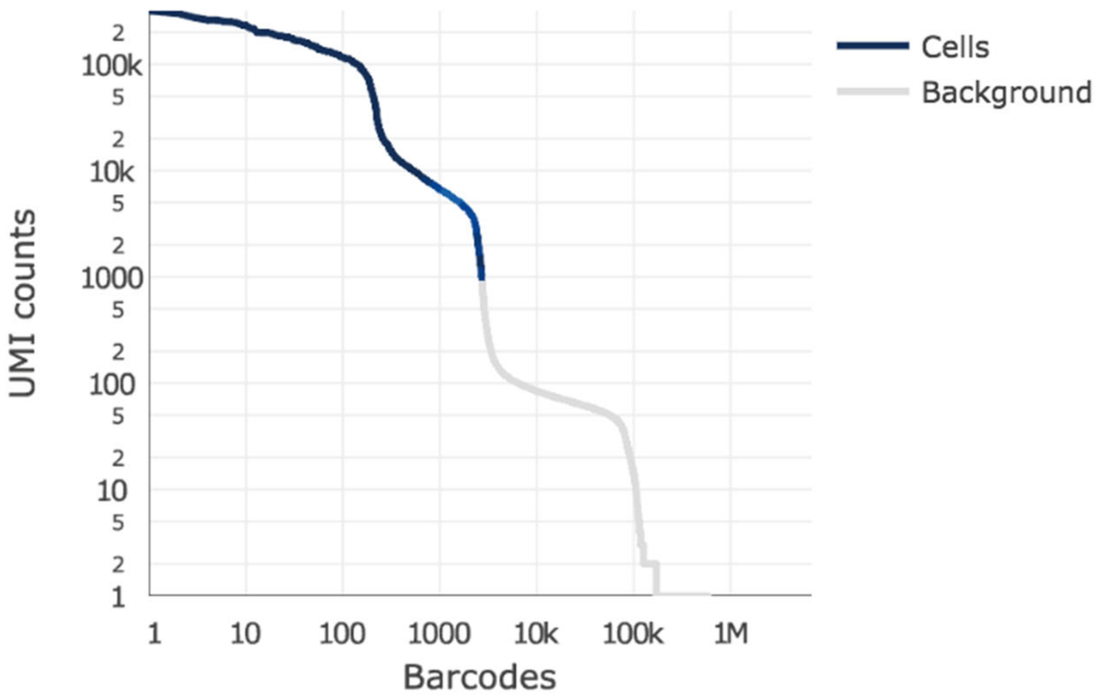
Cell ranger will correct sequences with Hamming distance of 1

UCLA QCBio
Collaboratory

Cell ranger barcode and UMI processing

How many cells are there in my data?

A ‘**whitelist**’ is the list of selected cells use to filter reads downstream



UCLA QCBio Collaboratory

Input files

```
filtered_feature_bc_matrix
 ├── barcodes.tsv.gz
 ├── features.tsv.gz
└── matrix.mtx.gz
```

```
$ gzip -cd filtered_feature_bc_matrix/features.tsv.gz ENSG00000141510     TP53  Gene
Expression
ENSG00000012048     BRCA1 Gene Expression
ENSG00000139687     RB1   Gene Expression
```

F number of features

```
$ gzip -cd filtered_feature_bc_matrices/barcodes.tsv.gz
AAACCCAAGGAGAGTA-1
AAACGCTTCAGCCCAG-1
AAAGAACAGACGACTG-1
```

B number of barcodes

Input files – matrix.mtx

%% MatrixMarket % % comments		
Rows_n	Cols_n	Entries
Row_1	Col_1	M[1,1]
Row_1	Col_2	M[1,2]
Row_2	Col_3	M[2,3]
…	…	…
Row_n	Cols_n	M[n,n]

For Seurat:

- Features (or Genes) are rows
- Barcodes (or Cells) are columns

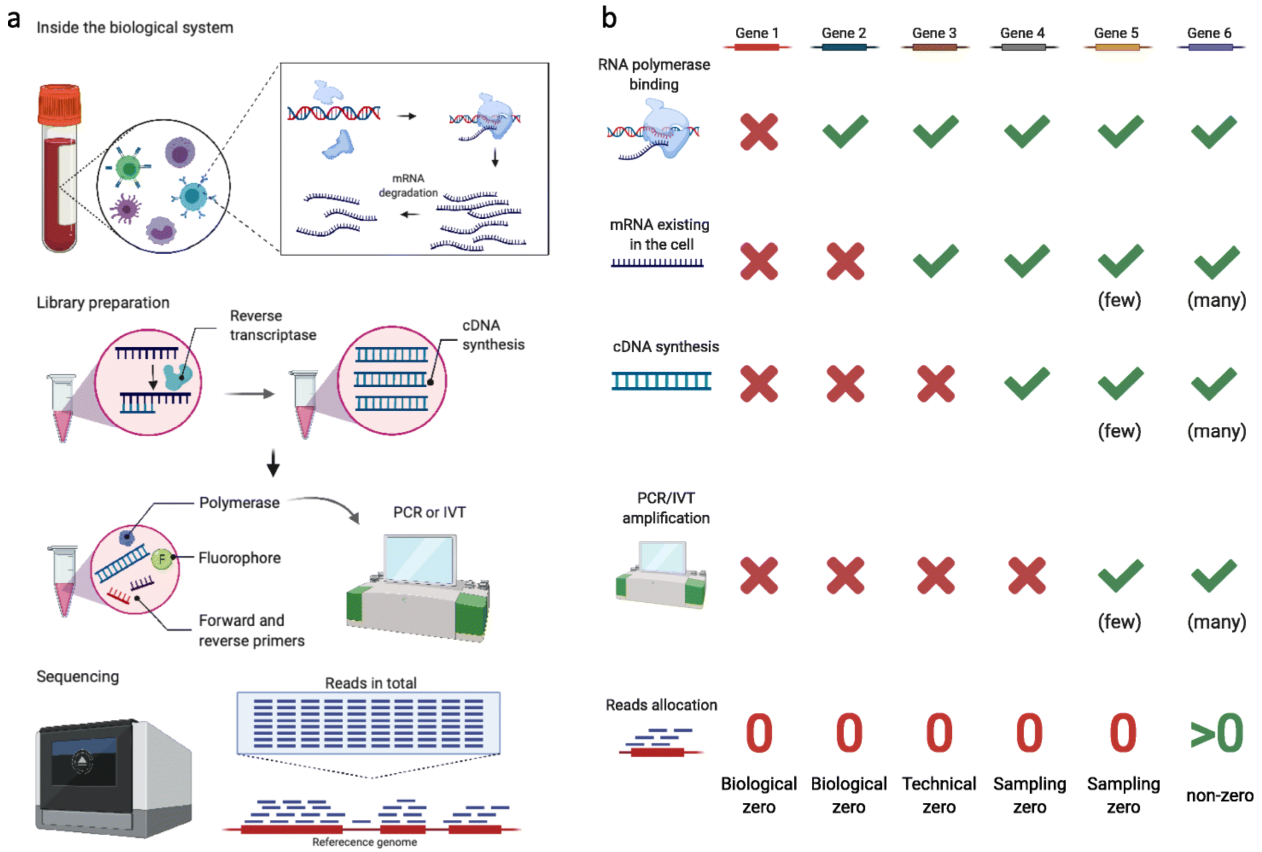
Input files – matrix.mtx

<table>
<tr><td colspan="3">%% MatrixMarket
%
% comments</td></tr>
<tr><td><b>F</b></td><td><b>B</b></td><td>$\leq F \times B$</td></tr>
<tr><td>1</td><td>1</td><td>12</td></tr>
<tr><td>1</td><td>2</td><td>13</td></tr>
<tr><td>2</td><td>1</td><td>3</td></tr>
<tr><td>…</td><td>…</td><td>…</td></tr>
<tr><td>F</td><td>B</td><td>3</td></tr>
</table>

For Seurat:

- Features (or Genes) are rows
- Barcodes (or Cells) are columns

Drop-outs

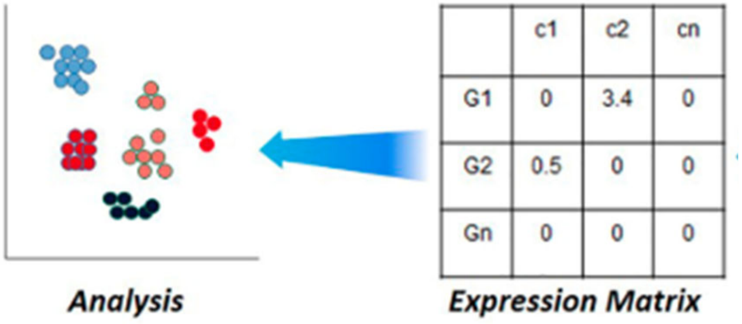


	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6
RNA polymerase binding	✗	✓	✓	✓	✓	✓
mRNA existing in the cell	✗	✗	✓	✓	✓ (few)	✓ (many)
cDNA synthesis	✗	✗	✗	✓	✓ (few)	✓ (many)
PCR/IVT amplification	✗	✗	✗	✗	✓ (few)	✓ (many)
Reads allocation	0 Biological zero	0 Biological zero	0 Technical zero	0 Sampling zero	0 Sampling zero	>0 non-zero

Jian et al 2022

UCLA QCBio Collaboratory

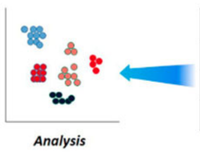
6. Data Analysis



	c1	c2	cn
G1	0	3.4	0
G2	0.5	0	0
Gn	0	0	0

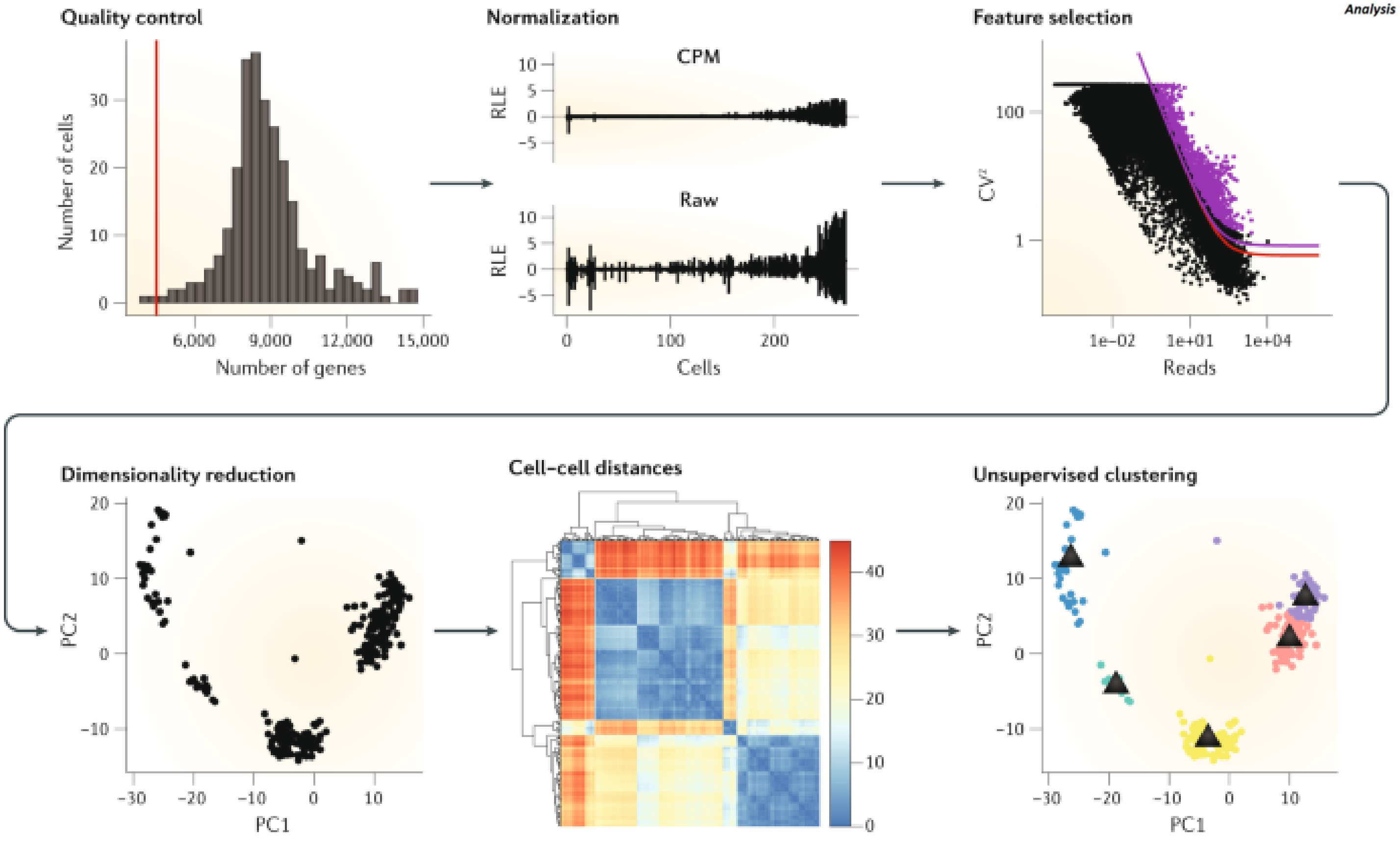
- Using the gene expression data in the form of raw UMI counts to identify cell populations.
- We will use Seurat to perform this analysis

Data analysis pipeline



	c1	c2	cn
G1	0	3.4	0
G2	0.5	0	0
Gn	0	0	0

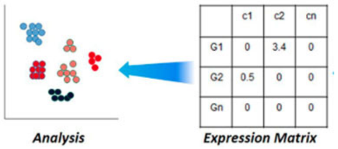
Expression Matrix



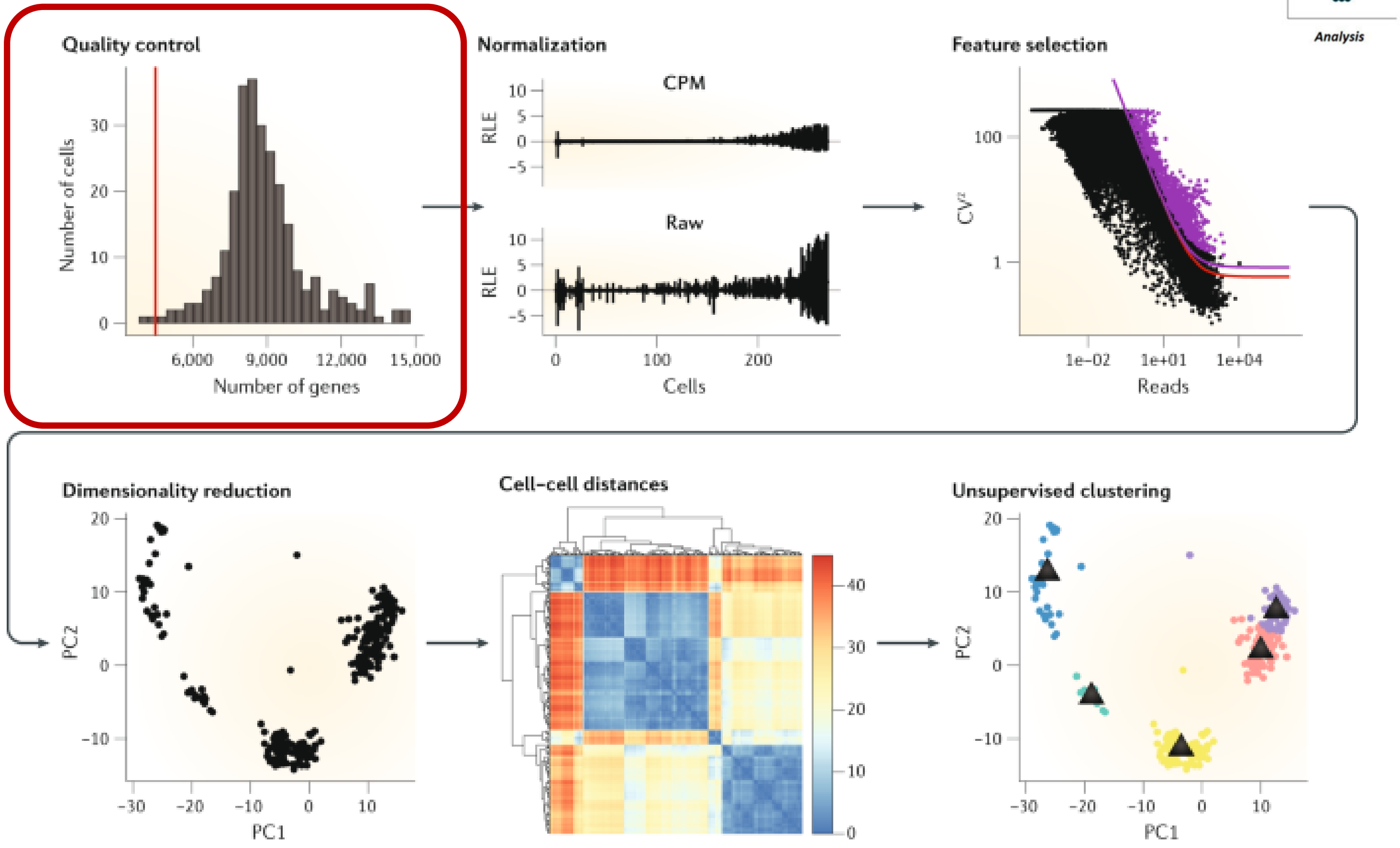
UCLA QCBio

Collaboratory

Quality control



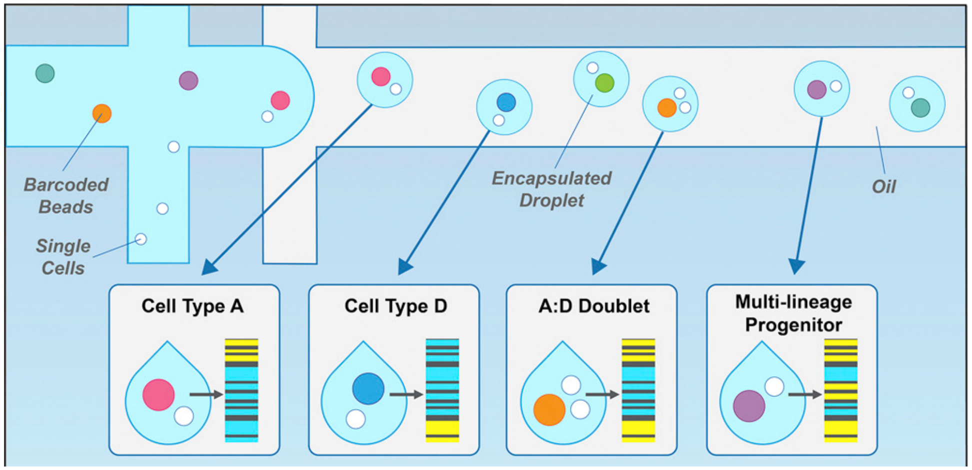
	c1	c2	cn
G1	0	3.4	0
G2	0.5	0	0
Gn	0	0	0



UCLA QCBio Collaboratory

Quality control

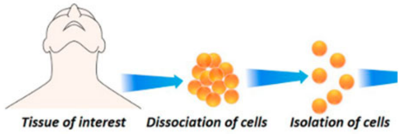
- How many genes do we expect?
    - Too few → empty droplet or low quality of cells
    - Too many → duplets (or multiplets)
- Technical terms:
    - Feature count = number of genes
    - RNA count = number of UMIs

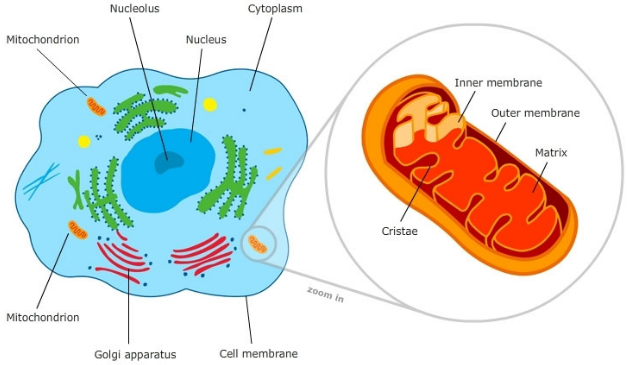


DePasquale et al 2019

Quality control

- Mitochondrial RNA
  - Due to very harsh conditions in tissue dissociation step.
  - Dying cells release their cytoplasmic contents.





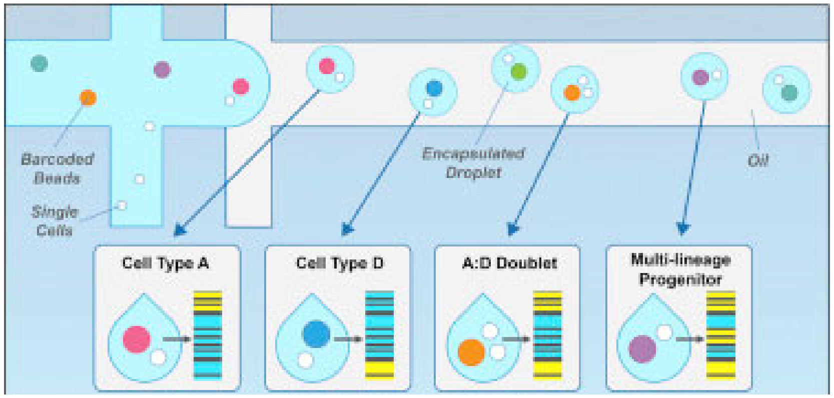
© 2007-2011 The University of Waikato | www.sciencelearn.org.nz

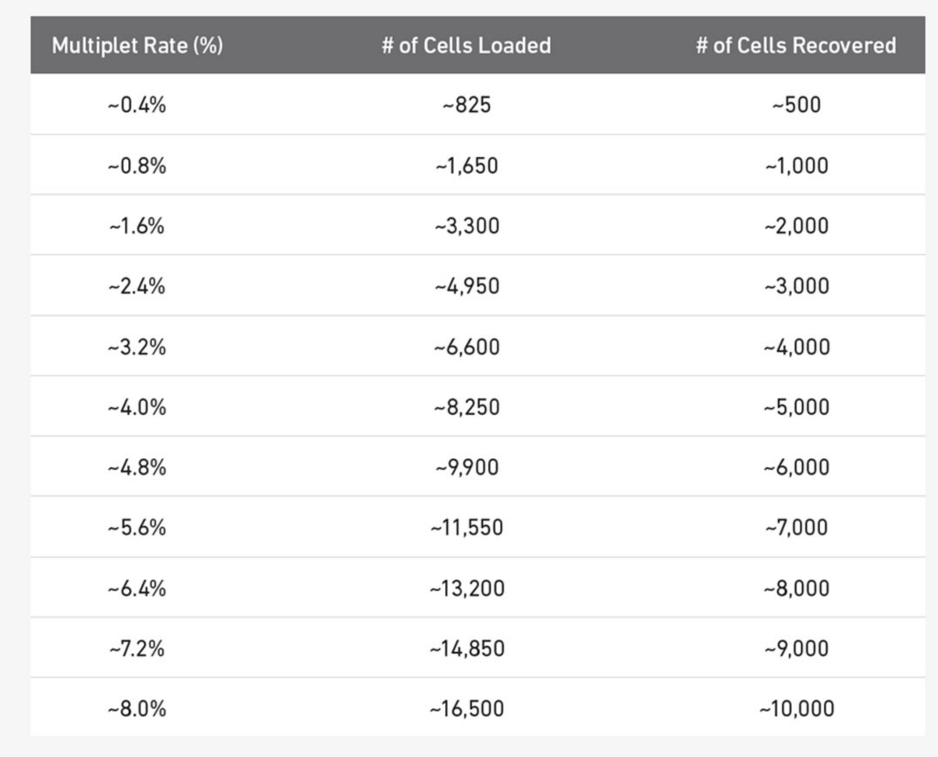
Quality control

- Metrics
  - RNA count (or count depth)
  - Feature count (or gene count)
  - Mitochondria content
- Recommendations
  - Identify and discard outliers
  - Different samples may require different cutoffs.
- Make use or ERCC spike-ins
  - Does the measured expression match the input?

More cells more doublets

- For expression assays (RNA-seq), high troughput can capture up to 20,000 cells per library and up to 16 libraries.
- With higher number of cells, higher the rate of multiplets.





Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~825	~500
~0.8%	~1,650	~1,000
~1.6%	~3,300	~2,000
~2.4%	~4,950	~3,000
~3.2%	~6,600	~4,000
~4.0%	~8,250	~5,000
~4.8%	~9,900	~6,000
~5.6%	~11,550	~7,000
~6.4%	~13,200	~8,000
~7.2%	~14,850	~9,000
~8.0%	~16,500	~10,000

Kahoot time!

- Go to www.kahoot.it

References

- Dong, X., Bacher, R. (2023). *Analysis of Single-Cell RNA-seq Data*. In: Fridley, B., Wang, X. (eds) Statistical Genomics. Methods in Molecular Biology, vol 2629. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-2986-4_6
- Jian Hu, Amelia Schroeder, Kyle Coleman, Chixiang Chen, Benjamin J. Auerbach, Mingyao Li. Statistical and machine learning methods for spatially resolved transcriptomics with histology. *Computational and Structural Biotechnology Journal*. Volume 19. 2021. Pages 3829-3841. ISSN 2001-0370. https://doi.org/10.1016/j.csbj.2021.06.052.
- Jovic, D, Liang, X, Zeng, H, Lin, L, Xu, F, Luo, Y. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med*. 2022; 12:e694. https://doi.org/10.1002/ctm2.694
- Adil Asif, Kumar Vijay, Jan Arif Tasleem, Asger Mohammed. Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. Fron*tiers in Neuroscience.* 15. 2021.10.3389/fnins.2021.591122
- Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, 1–14 (2018). https://doi.org/10.1038/s12276-018-0071-8