Introduction to Single-cell Omics Analysis

Joel Graber

Senior Staff Scientist

Director, Comparative Genomics and Data Science Core

MDI Biological Laboratory

Basics of Gene Expression: the Central Dogma

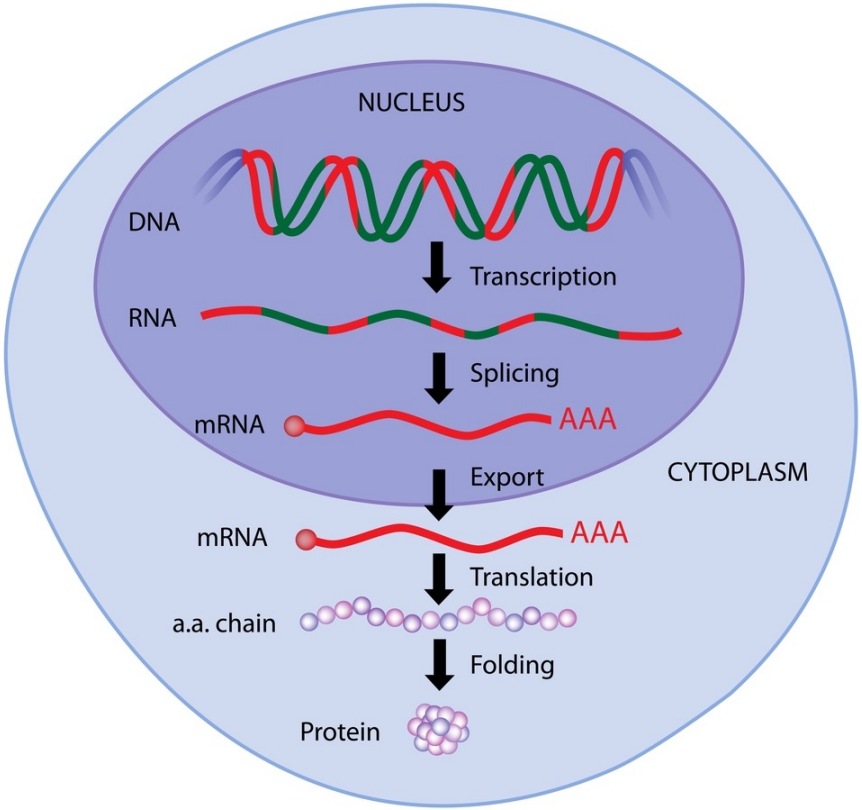


Image Source: https://www.acsh.org/news/2016/12/26/central-dogma-how-your-dna-determines-who-you-are-10645

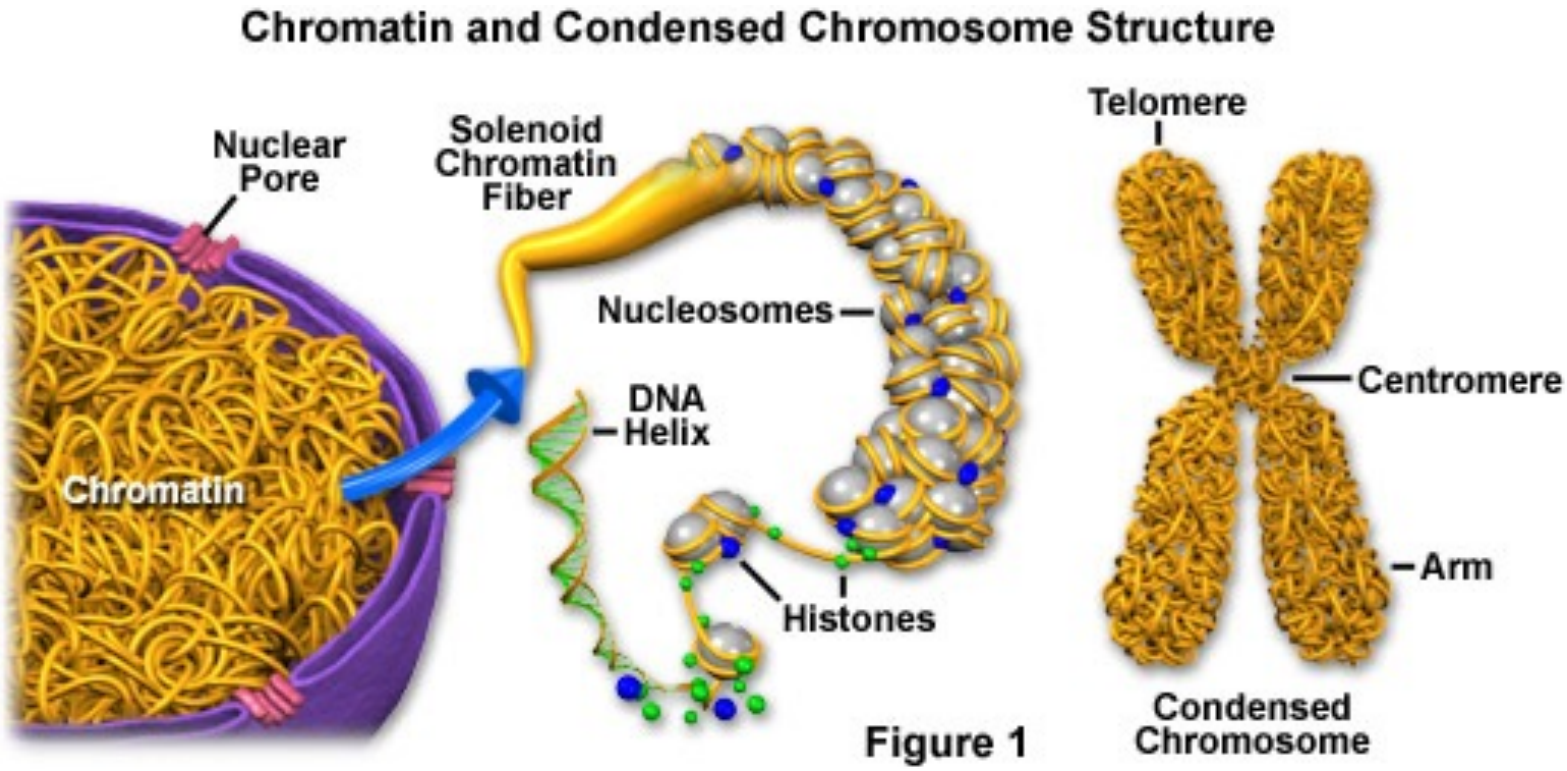


Image Source: http://biogeonerd.blogspot.com/2012/09/mitosis.html

What can we measure?

- Nucleic Acid Sequence
  - hundreds of millions of reads at 100-150 bases in length
  - We mostly sequence as DNA
  - Bulk samples (10^8), single-cell (10^7)
- Proteins/Metabolomics
  - 1000s of fragments per sample

- Question: Why is nucleic acid sequencing easier than protein?
- Answer: because nature did much of the work for us

Sequencing by synthesis

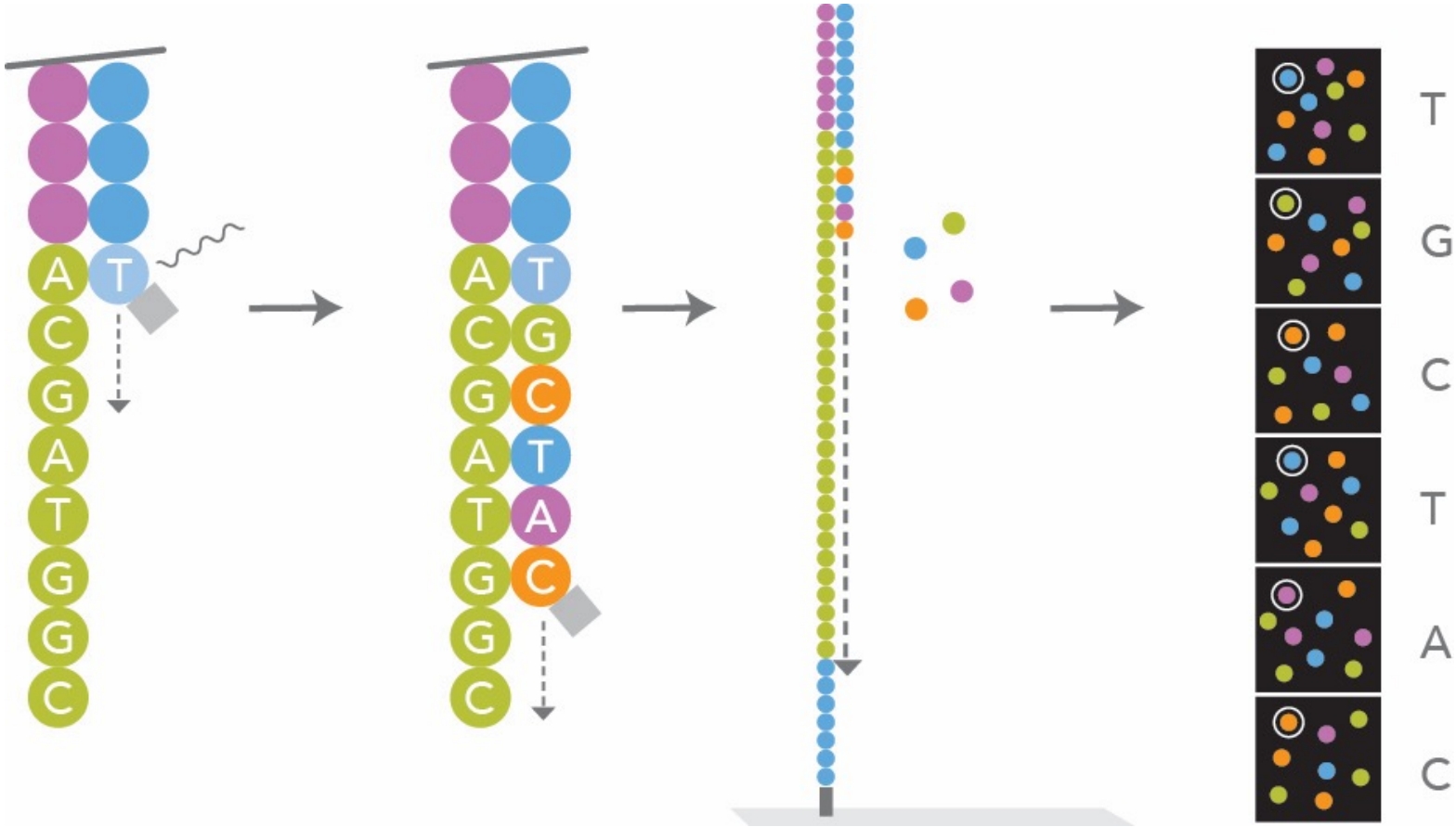


Image source: https://www.lexogen.com/rna-lexicon-next-generation-sequencing/

A conceptual RNAseq experimental workflow

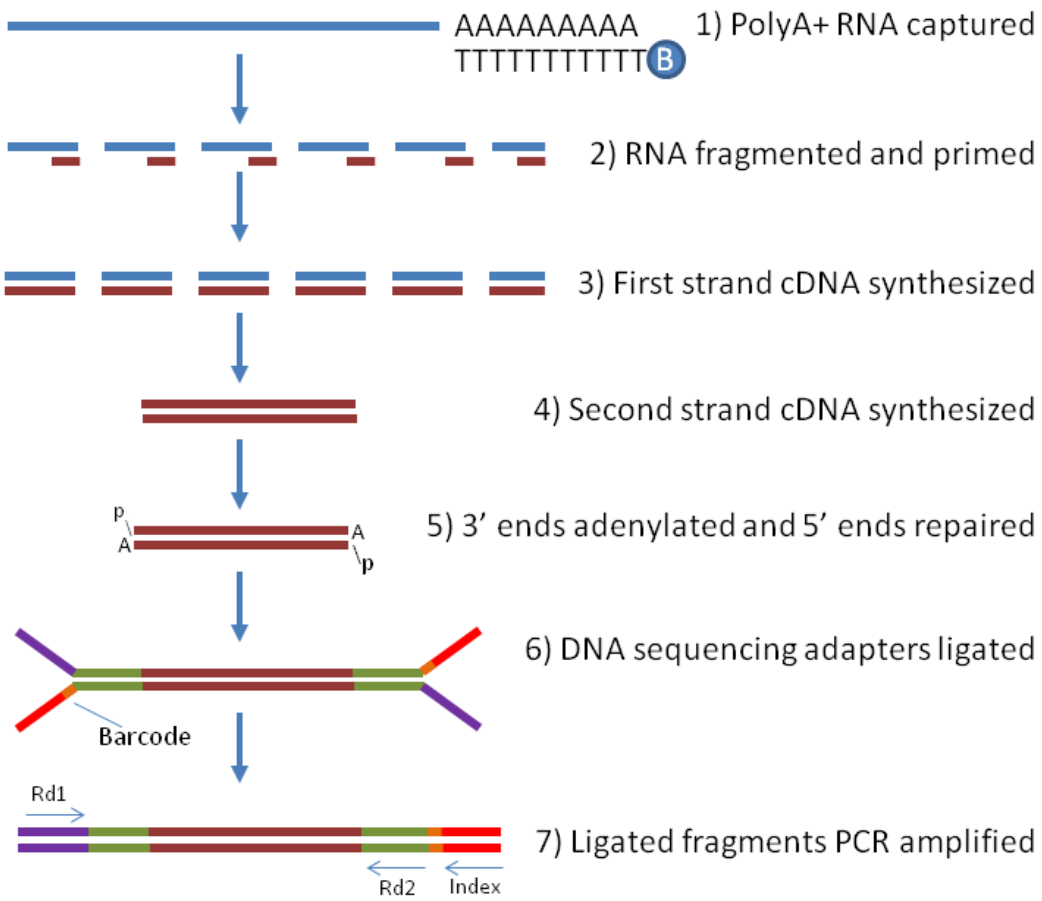
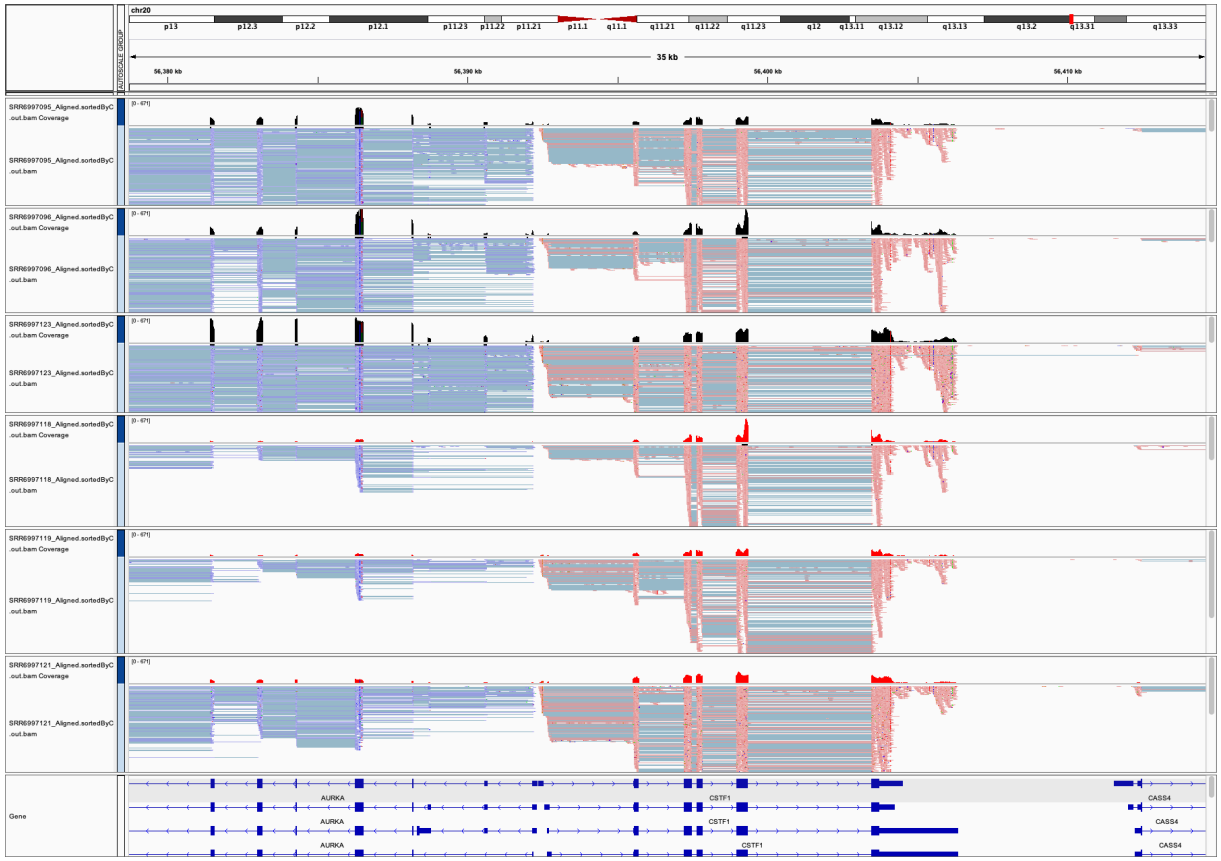


Image source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

RNA-seq analysis: alignment/quantification



After expression is assessed in each sample, they are merged into a “count matrix”

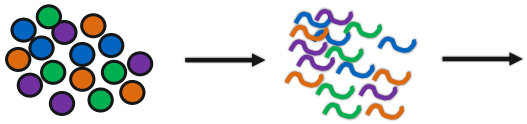
gene_id	SL94881	SL94882	SL94883	SL94884	SL94885	SL94886
ENSDARG00000000001	33	40	36	38	58	40
ENSDARG00000000002	136	126	156	167	170	158
ENSDARG00000000018	319	356	345	368	357	334
ENSDARG00000000019	1174	1390	1430	1356	1130	1237
ENSDARG00000000068	522	468	590	622	506	528
ENSDARG00000000069	1622	1622	1546	1494	1546	1561
ENSDARG00000000086	413	536	474	489	290	476
ENSDARG00000000103	1212	1390	1266	1296	1012	1390
ENSDARG00000000142	118	97	99	110	94	126
…						

Fundamental ideas of transcriptome profiling

- The activity or state of a sample can be defined by a “snapshot” (profile) of molecules
  - In a transcriptome, the molecules are the RNA transcripts
- Perturbations/changes to a system induce systemic (and predictable and reproducible) responses
- Comparison of molecular profiles can
  - delineate the mechanisms/pathways involved
  - Identify and classify related samples

Bulk vs Single Cell RNA-seq (scRNA-seq)

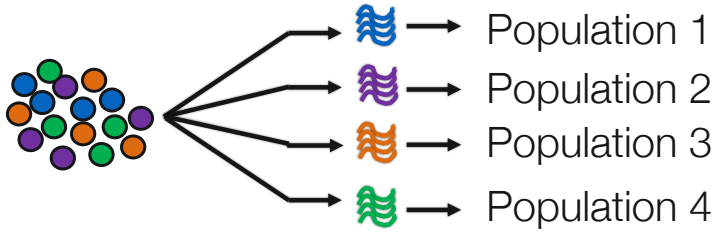
Bulk RNA-seq



average expression level

- comparative transcriptomics
- disease biomarker
- homogenous systems

scRNA-seq



- define heterogeneity
- identify rare cell population
- cell population dynamics

Slide source: https://www.sambuz.com/doc/introduction-to-single-cell-rna-sequencing-ppt-presentation-754293

BULK VS SINGLE CELL RNA-SEQ

Average expression level
- Comparative transcriptomics
- Disease biomarker
- Homogenous systems

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Slide source: https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf

BULK VS SINGLE CELL RNA-SEQ

1. mRNA: TruSeq RNA-Seq (Gold Standard)
   - ~20,000 transcripts
     - More when consider splice variants / isoforms
   - Observe 80-95% of transcripts depending on sequencing depth

2. Low input methods ~3000 cells / well
   - 4000-6000 transcripts per sample
     - Limiting to transcripts observed across all samples
   - Observe 20-60% of the transcriptome

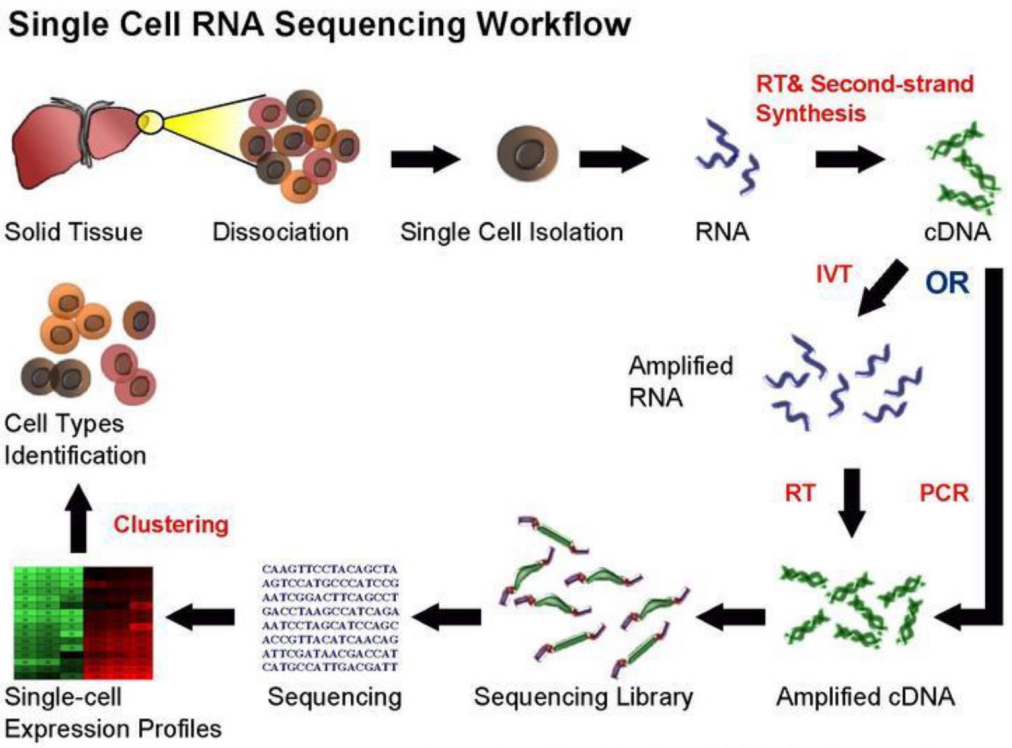
3. Single Cell Methods
   - 200 -10,000 transcripts per cell
   - Observe 10-50% of the transcriptome
   - Many transcripts will show up with zero counts in every cell. (even GAPDH)
   - If you only looked at transcripts observed in all cells numbers drop dramatically.

Source: Sarah Boswell, Harvard Medical School, September 2020

Slide source: https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf

WORKFLOW

Good sample preparation is key to success!



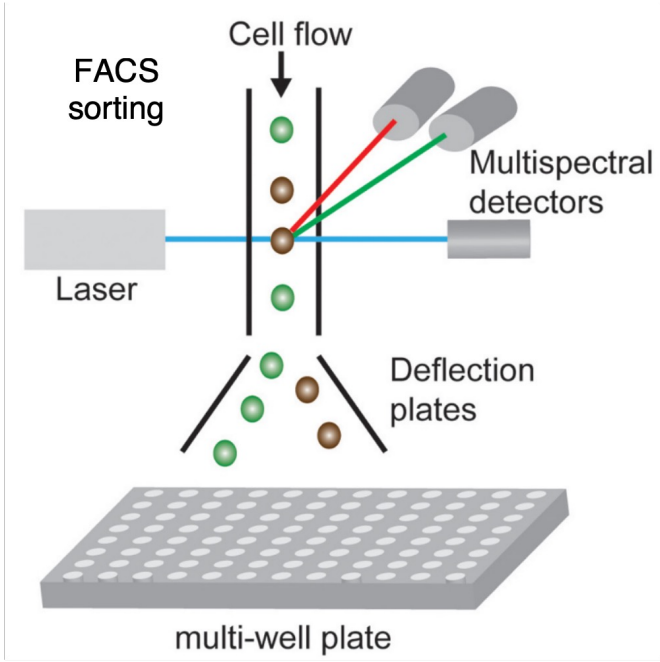
Source: https://en.wikipedia.org/wiki/Single_cell_sequencing

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

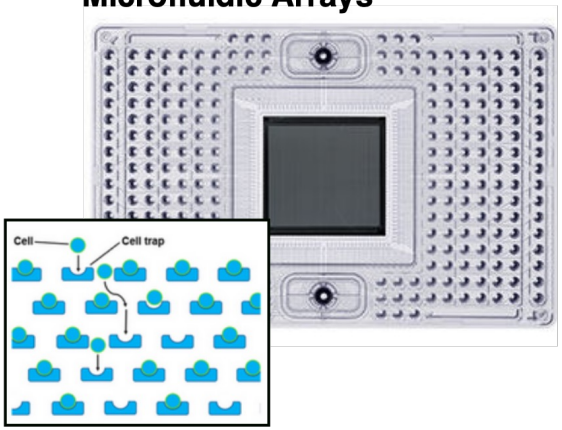
Slide source: https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf

Multiple techniques have been developed

Microtitre Plates



Microfluidic Arrays



Microfluidic Droplets

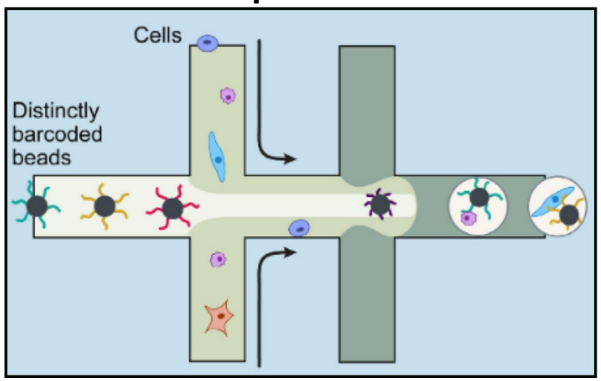


Image source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

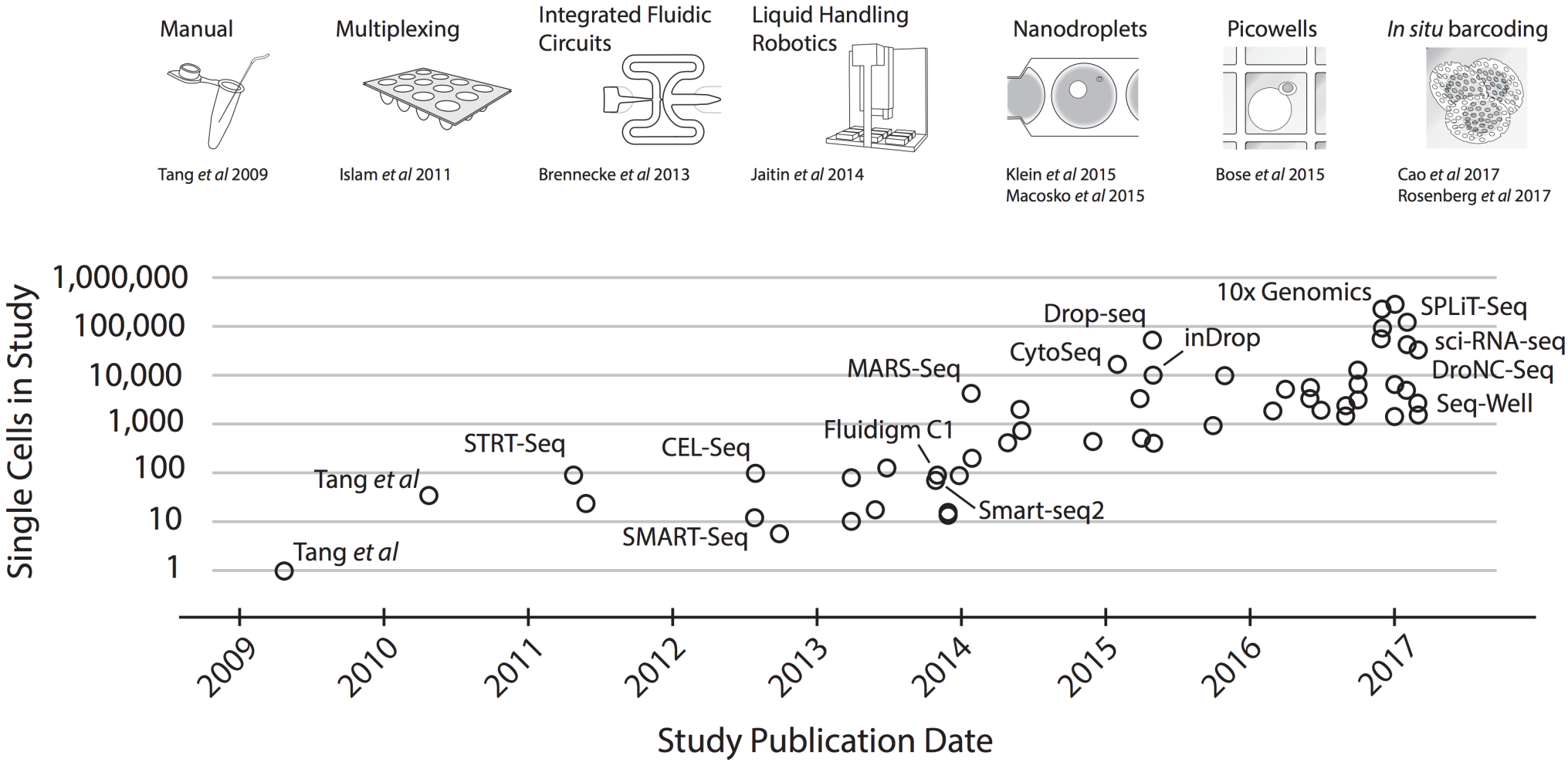
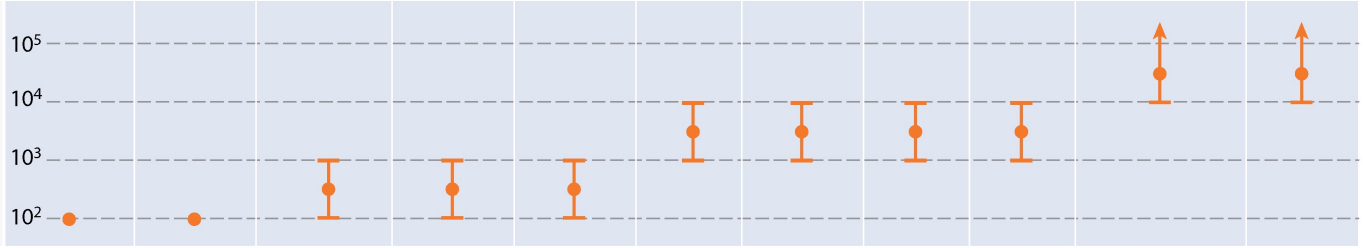


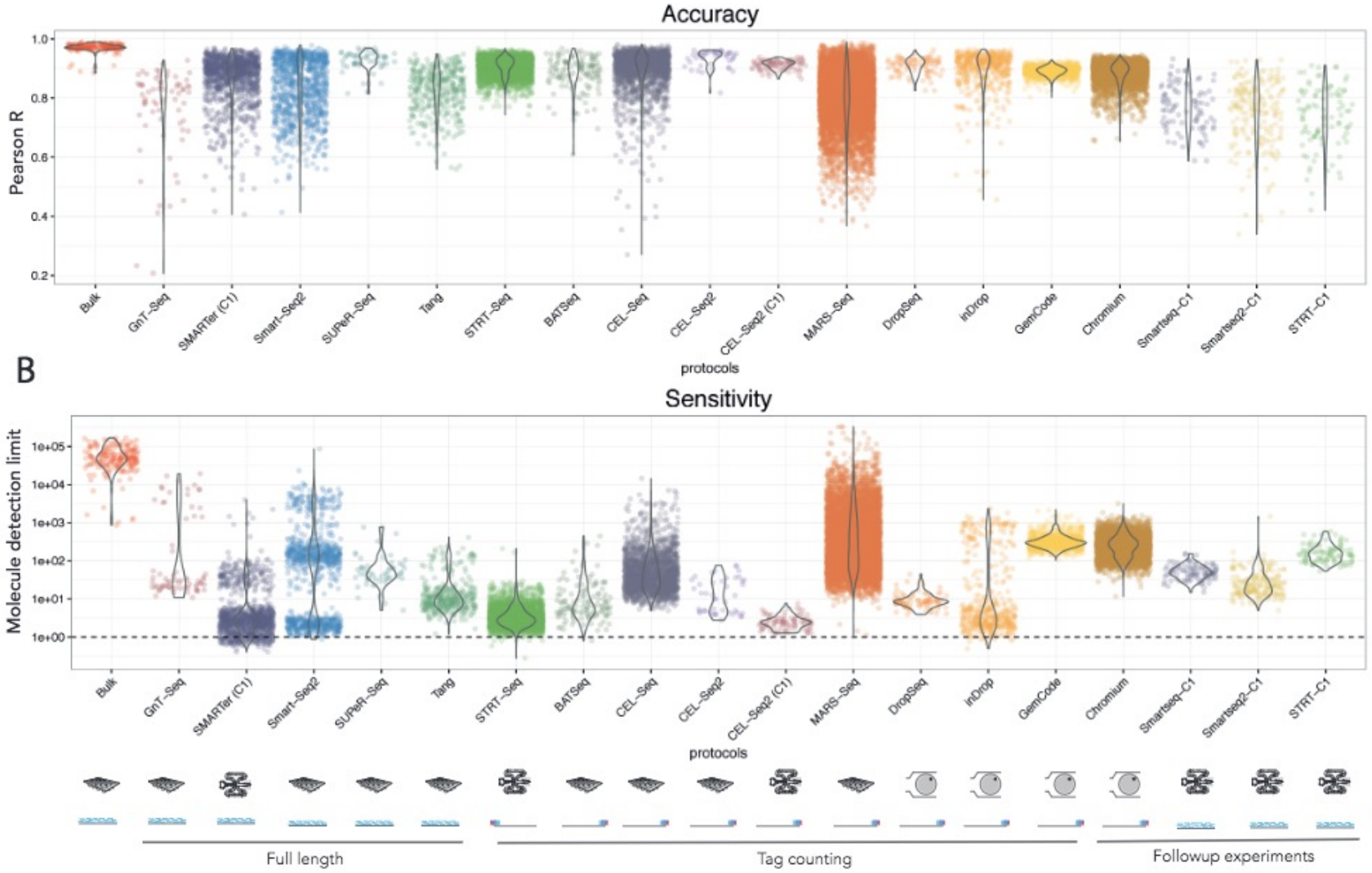
Image source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

Capabilities, depth, and characteristics vary among approaches

	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

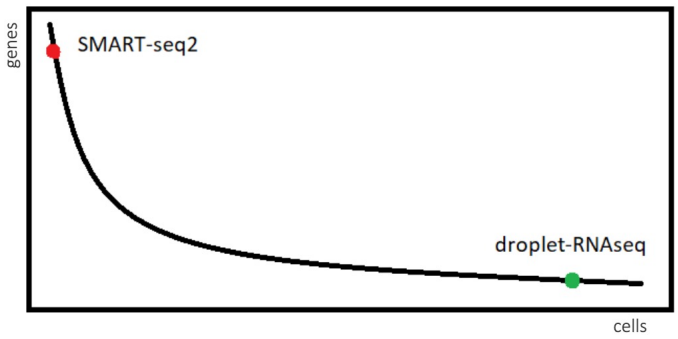
Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html



Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

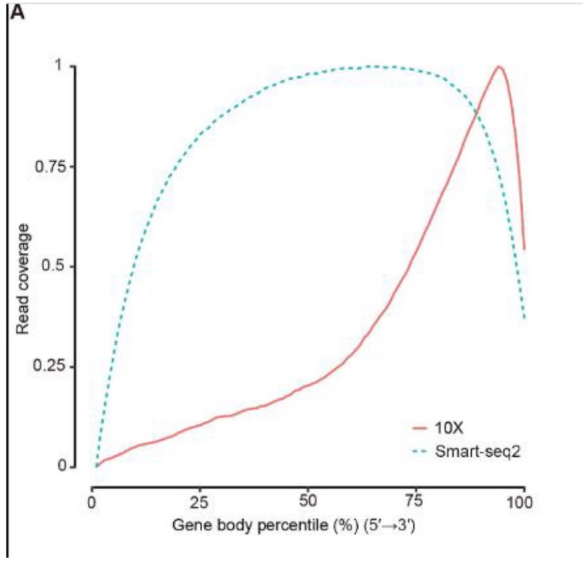
Next Page

MORE CELLS OR MORE GENES?



SMART-seq2
- 100 cells
- Full-length libraries
- 1M reads per cell

Droplet-RNAseq
- 10000 cells
- 50k reads per cell
- 3’/5’ bias



Source: Wang, et al. Genom. Proteom. Bioinform. 19(2), 253-266 (2021).

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

- Required number of cells increases with complexity of the sample.
- Number of reads will depend on biology of sample
- Cell-type classification of a mixed population usually requires lower read depth
- You can always re-sequence your samples.

Slide source: https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf

A conceptual RNAseq experimental workflow

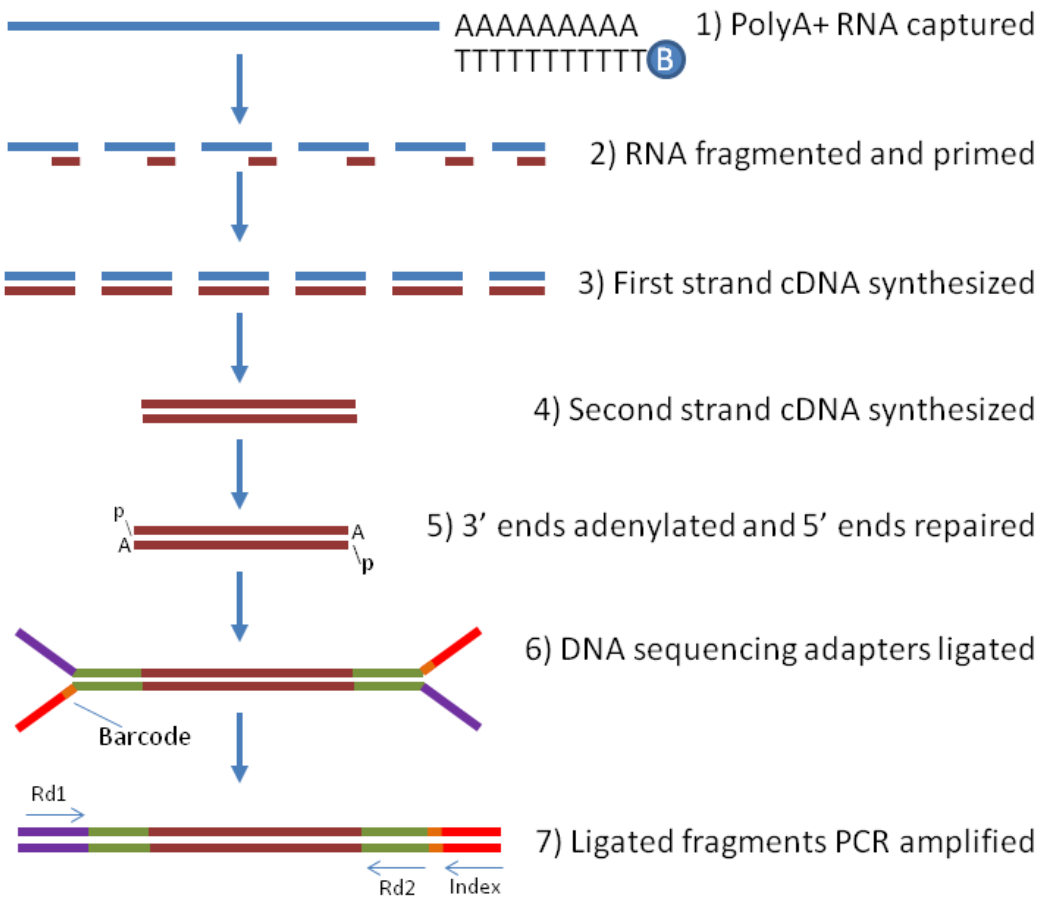
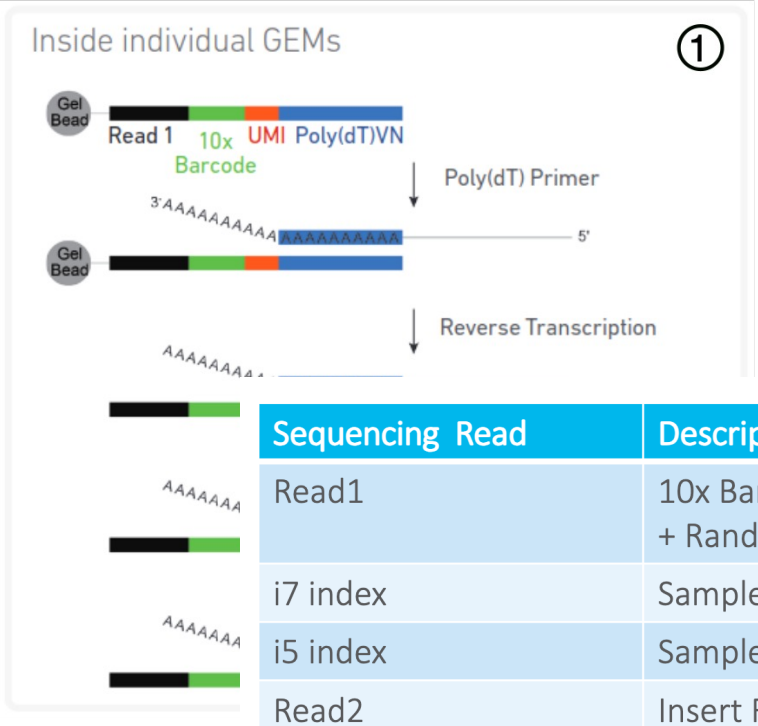
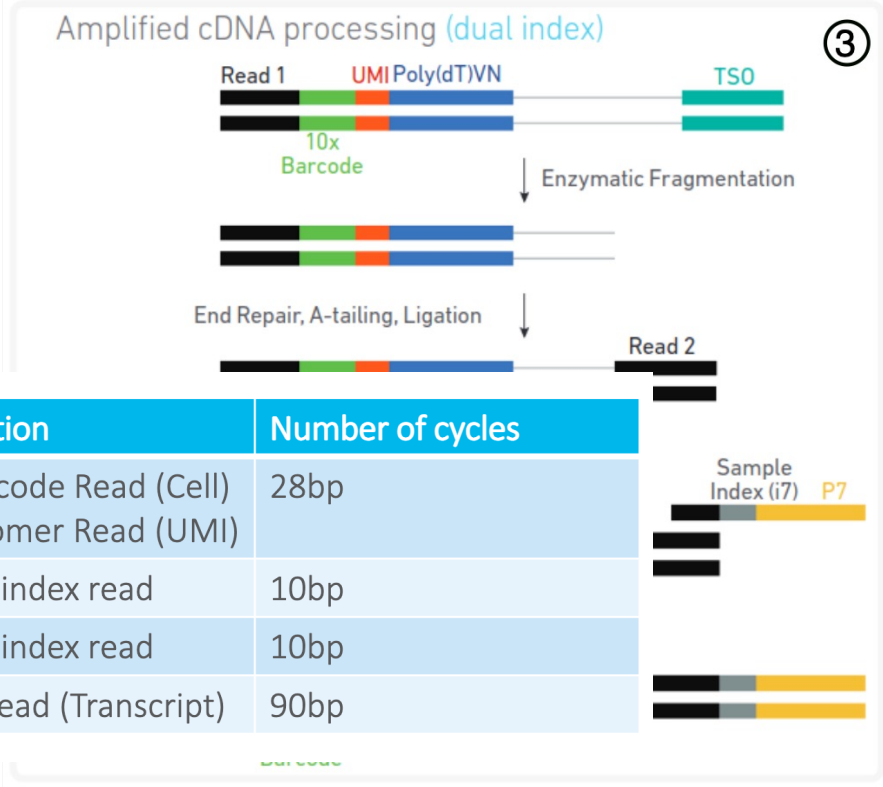
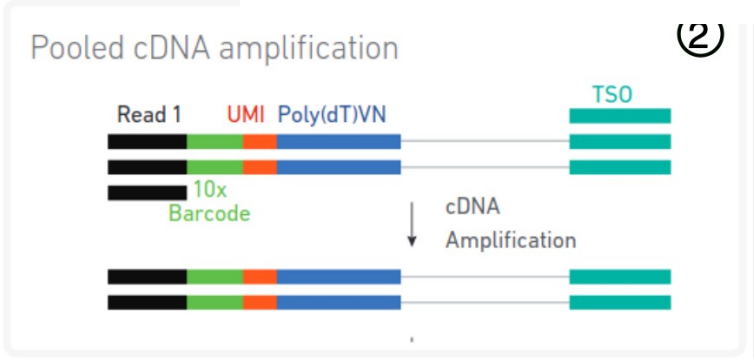


Image source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

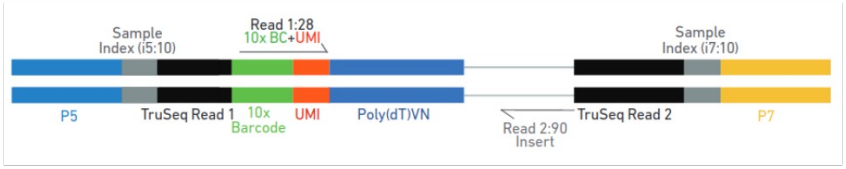




Sequencing Read	Description	Number of cycles
Read1	10x Barcode Read (Cell) + Randomer Read (UMI)	28bp
i7 index	Sample index read	10bp
i5 index	Sample index read	10bp
Read2	Insert Read (Transcript)	90bp

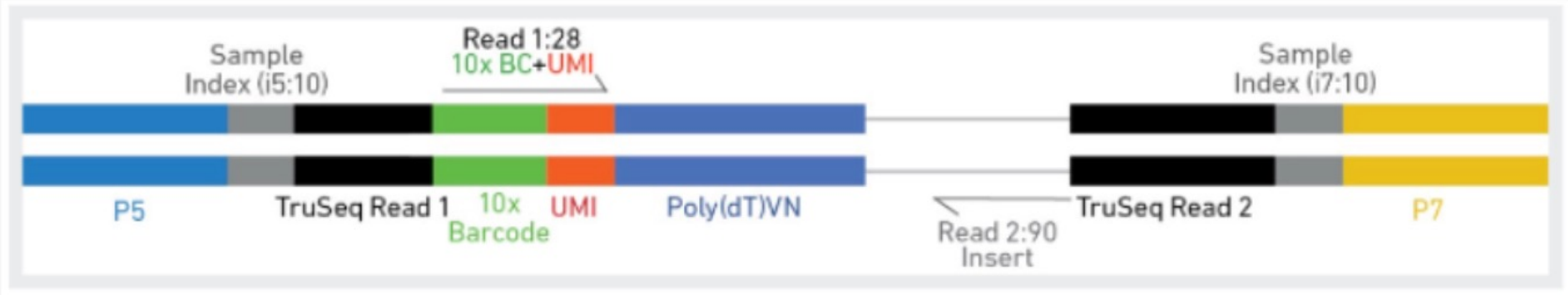


Final library



Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

Single Cell 3' v3.1(Dual Index) Gene Expression Library:



Single Cell 5' v2 Gene Expression Library:

Image Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

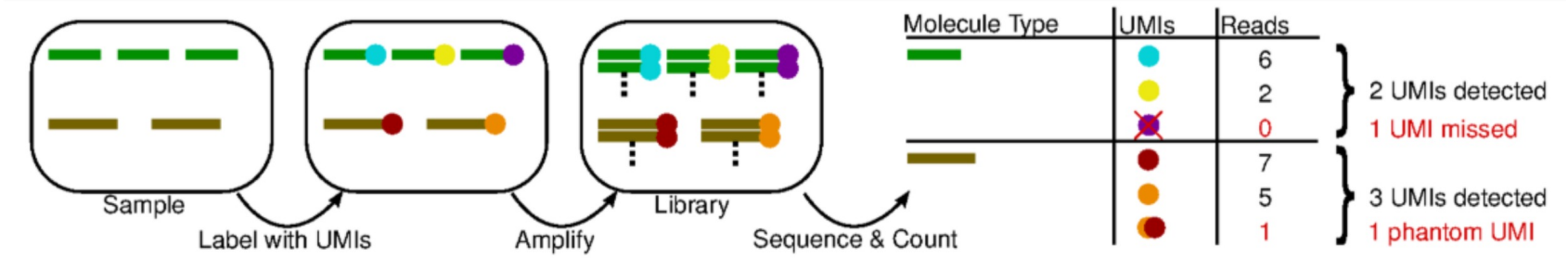
Sequencing is imperfect: Barcodes must be checked and corrected

Chemistry	CB, bp	UMI, bp	Whitelist file
10x Chromium Single Cell 3’ v1	14	10	737K-april-2014_rc.txt
10x Chromium Single Cell 3’ v2	16	10	737K-august-2016.txt
10x Chromium Single Cell 3’ v3	16	12	3M-february-2018.txt
10x Chromium Single Cell 3’ v3.1 (Next GEM)	16	12	3M-february-2018.txt
10x Chromium Single Cell 5’ v1.1	16	10	737K-august-2016.txt
10x Chromium Single Cell 5’ v2 (Next GEM)	16	10	737K-august-2016.txt
10x Chromium Single Cell Multiome	16	12	737K-arc-v1.txt

Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

UMI – UNIQUE MOLECULAR IDENTIFIERS

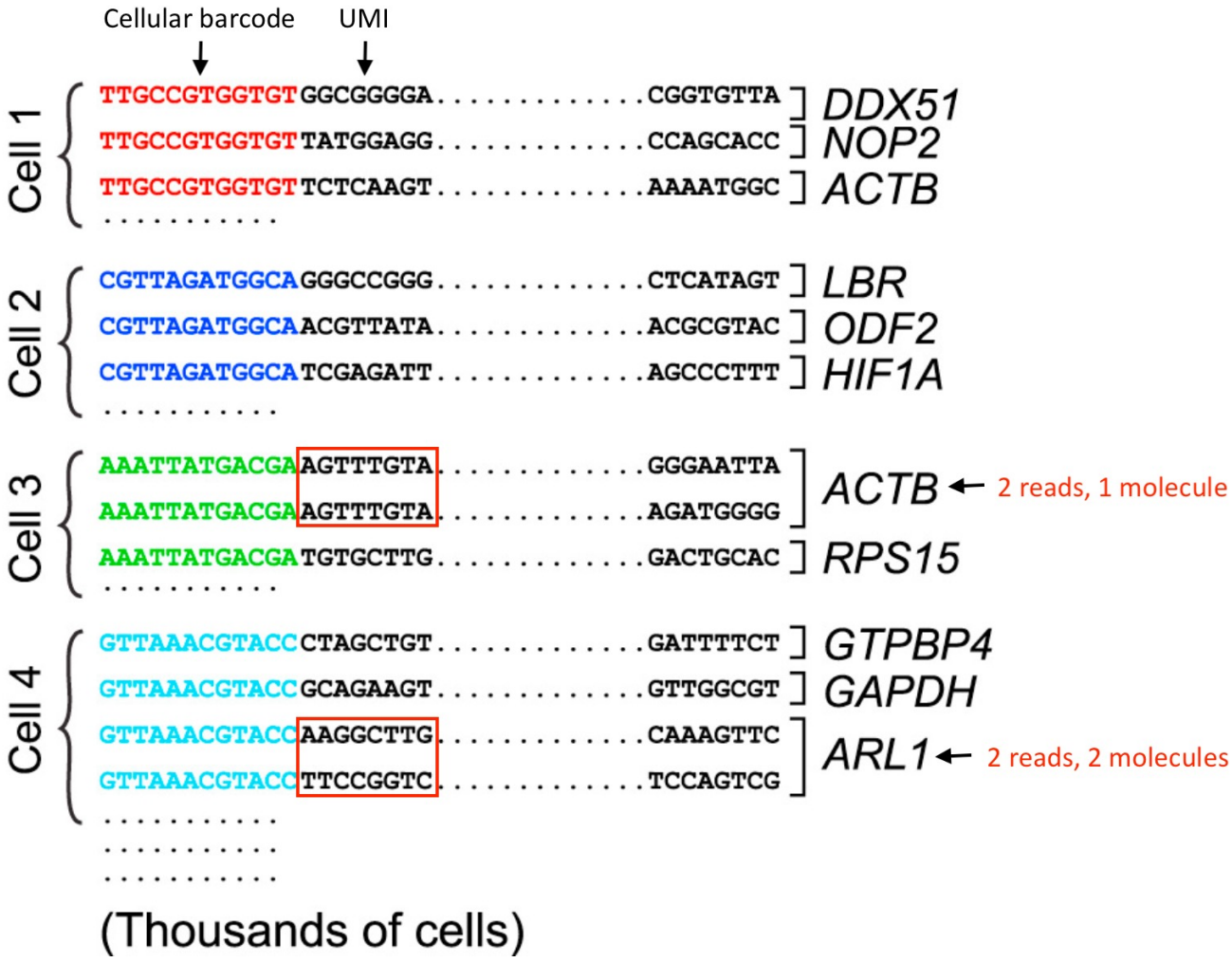
After PCR enrichment, without UMIs, one can not distinguish if multiple copies of a fragment are caused by PCR clones or if they are real biological duplicated.
By using UMIs, PCR clones can be found by searching for non-unique fragment-UMI combinations, which can only be explained by PCR clones.
When performing variant analyses, these falsely overrepresented fragments can result in incorrect calls and thus wrong diagnostic findings



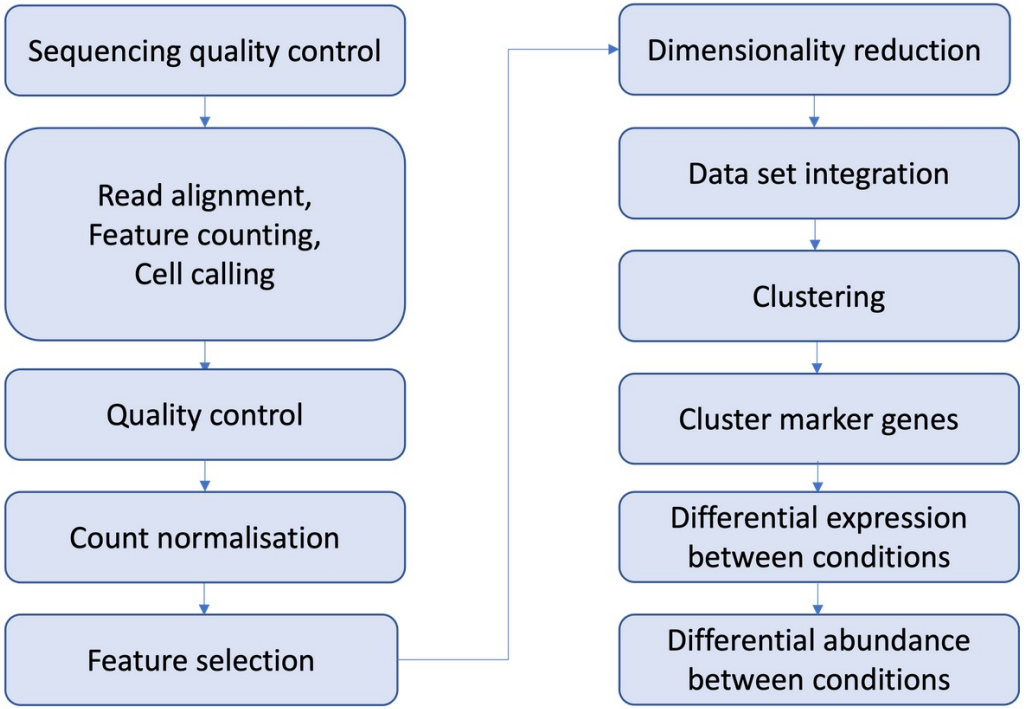
Source: Pflug et al. Bioinformatics (2018)

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Source:https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf


Cellular barcode
UMI
Cell 1
TTGCCGTGGTGT GGCGGGGA ..... CGGTGTTA
TTGCCGTGGTGT TATGGAGG ..... CCAGCACC
TTGCCGTGGTGT TCTCAAGT ..... AAAATGGC
DDX51
NOP2
ACTB
Cell 2
CGTTAGATGGCA GGGCCGGG ..... CTCATAGT
CGTTAGATGGCA ACGTTATA ..... ACGCGTAC
CGTTAGATGGCA TCGAGATT ..... AGCCCTTT
LBR
ODF2
HIF1A
Cell 3
AAATTATGACGA AGTTTGTA ..... GGGAATTA
AAATTATGACGA AGTTTGTA ..... AGATGGGG
AAATTATGACGA TGTGCTTG ..... GACTGCAC
ACTB
2 reads, 1 molecule
RPS15
Cell 4
GTTAAACGTACC CTAGCTGT ..... GATTTTCT
GTTAAACGTACC GCAGAAGT ..... GTTGGCGT
GTTAAACGTACC AAGGCTTG ..... CAAAGTTC
GTTAAACGTACC TTCCGGTC ..... TCCAGTCG
GTPBP4
GAPDH
ARL1
2 reads, 2 molecules
(Thousands of cells)

Typical Computational Workflow



https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept22/

Alignment and assignment depend upon gene knowledge (GFF/GTF)

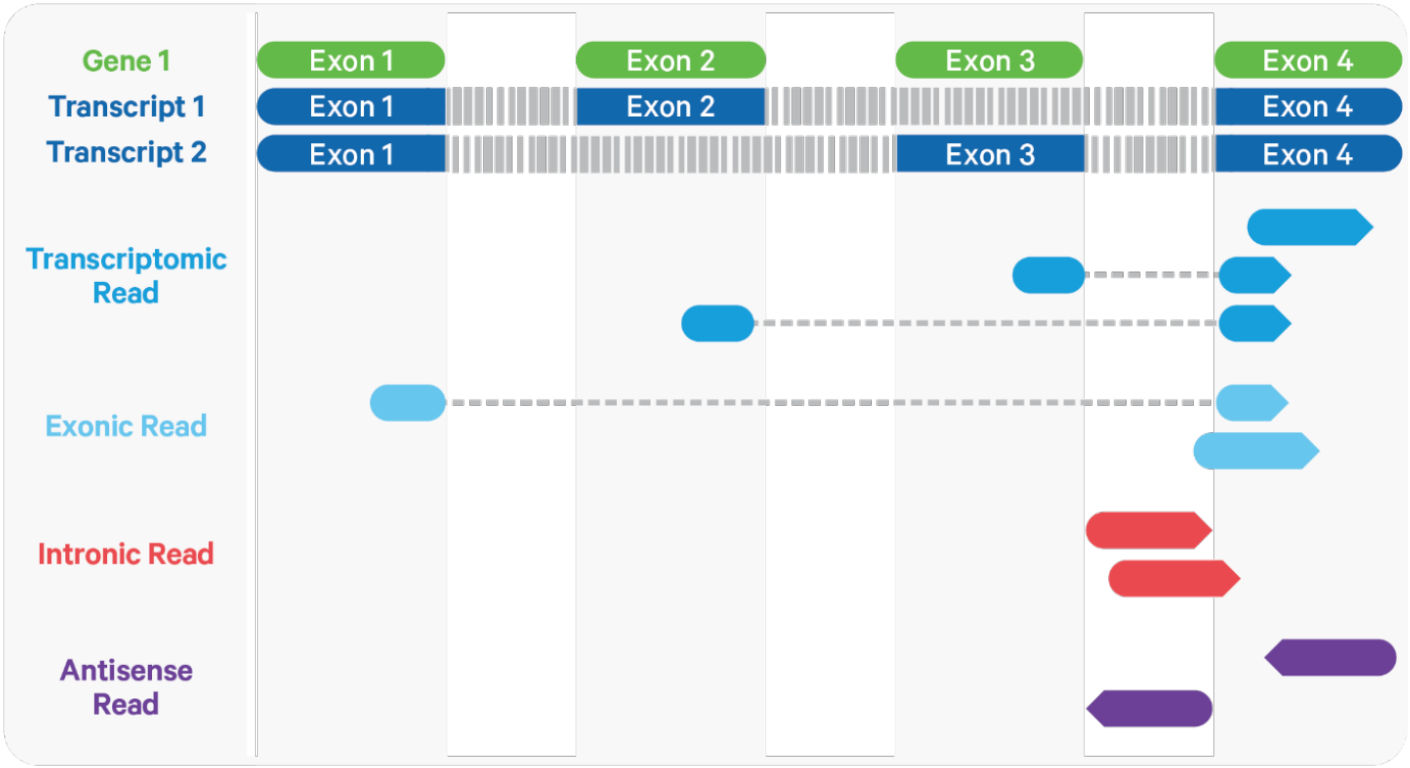


Image Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

Typical Chromium Gene Coverage plots

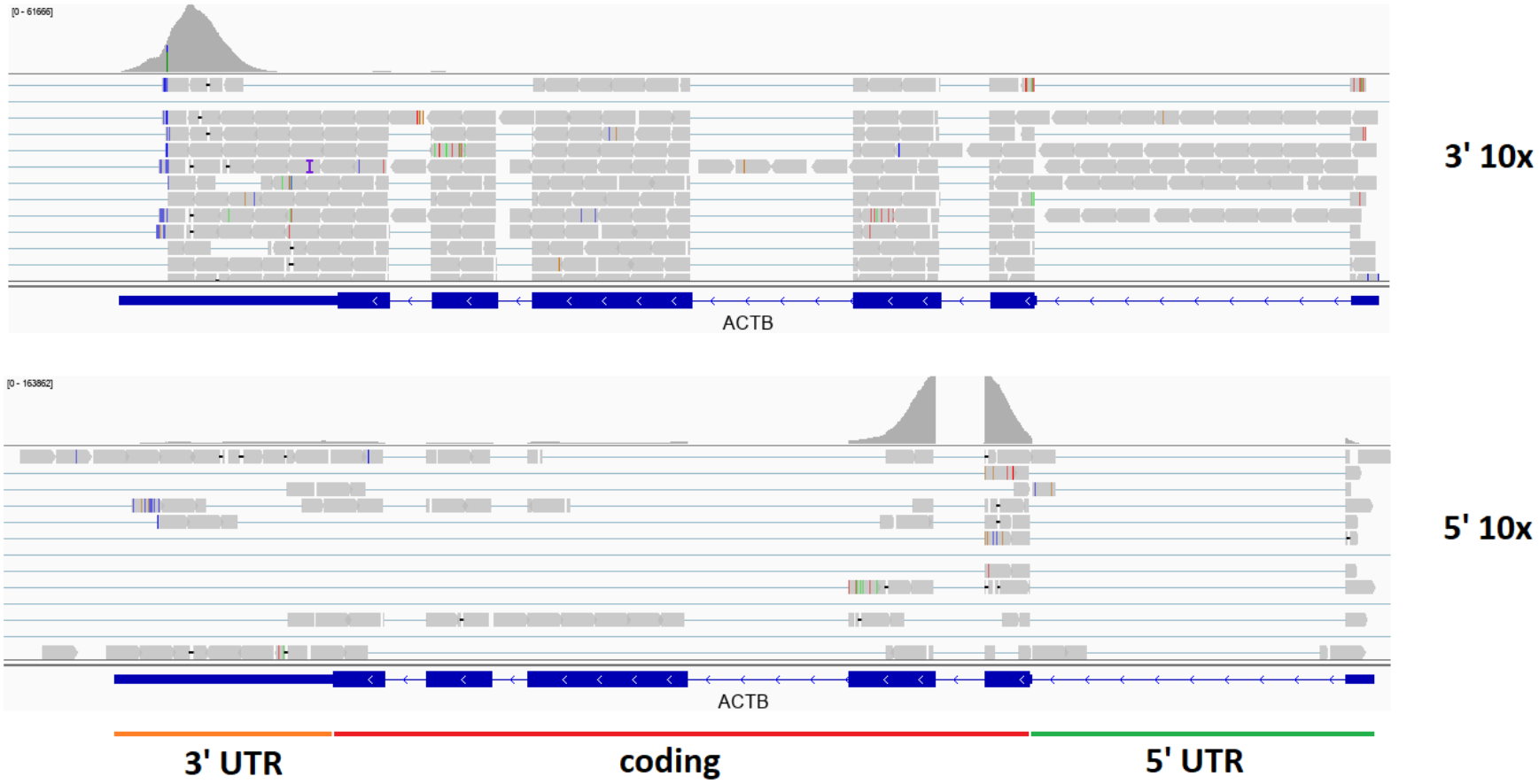


Image Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

Not every droplet is useble

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

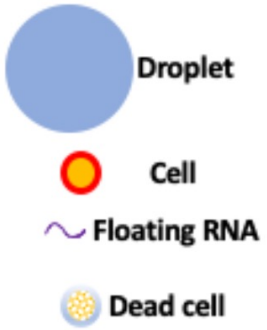
- Low complex transcriptome
- Genes detected much lower than average genes per cell

Droplet with dead cell

- Enriched for mitochondrial genes

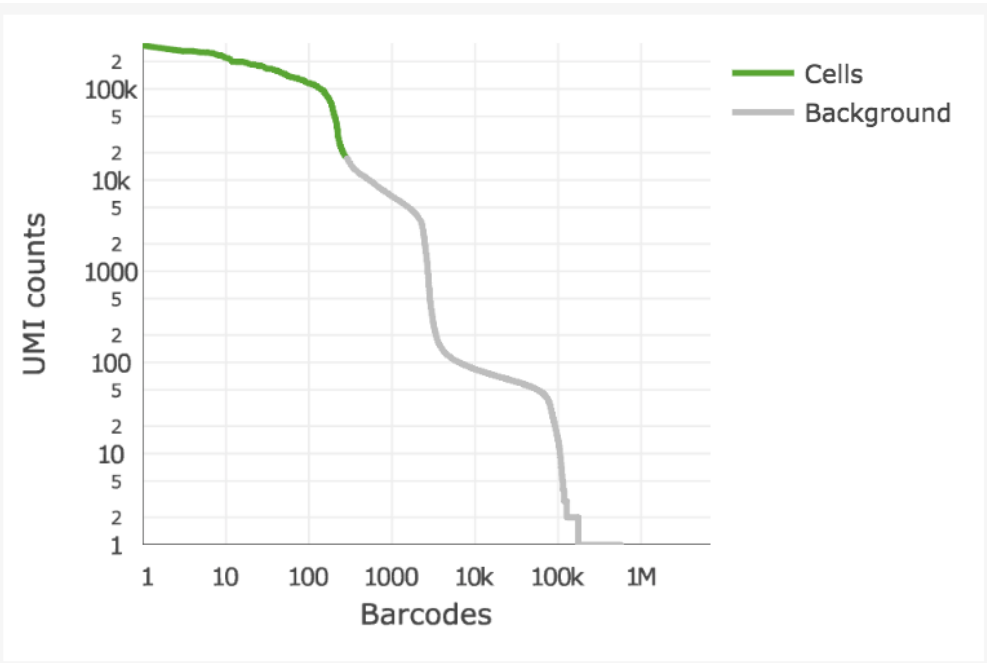
Droplet with multiple cell

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

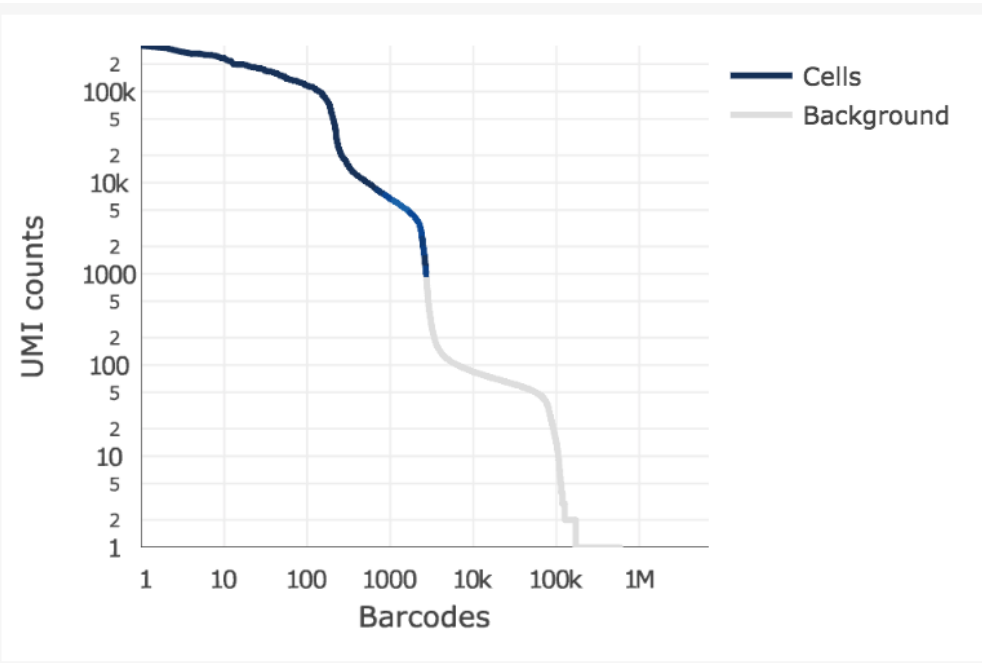


Source:https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf

Processing tools will catch some of the problematic barcodes for you



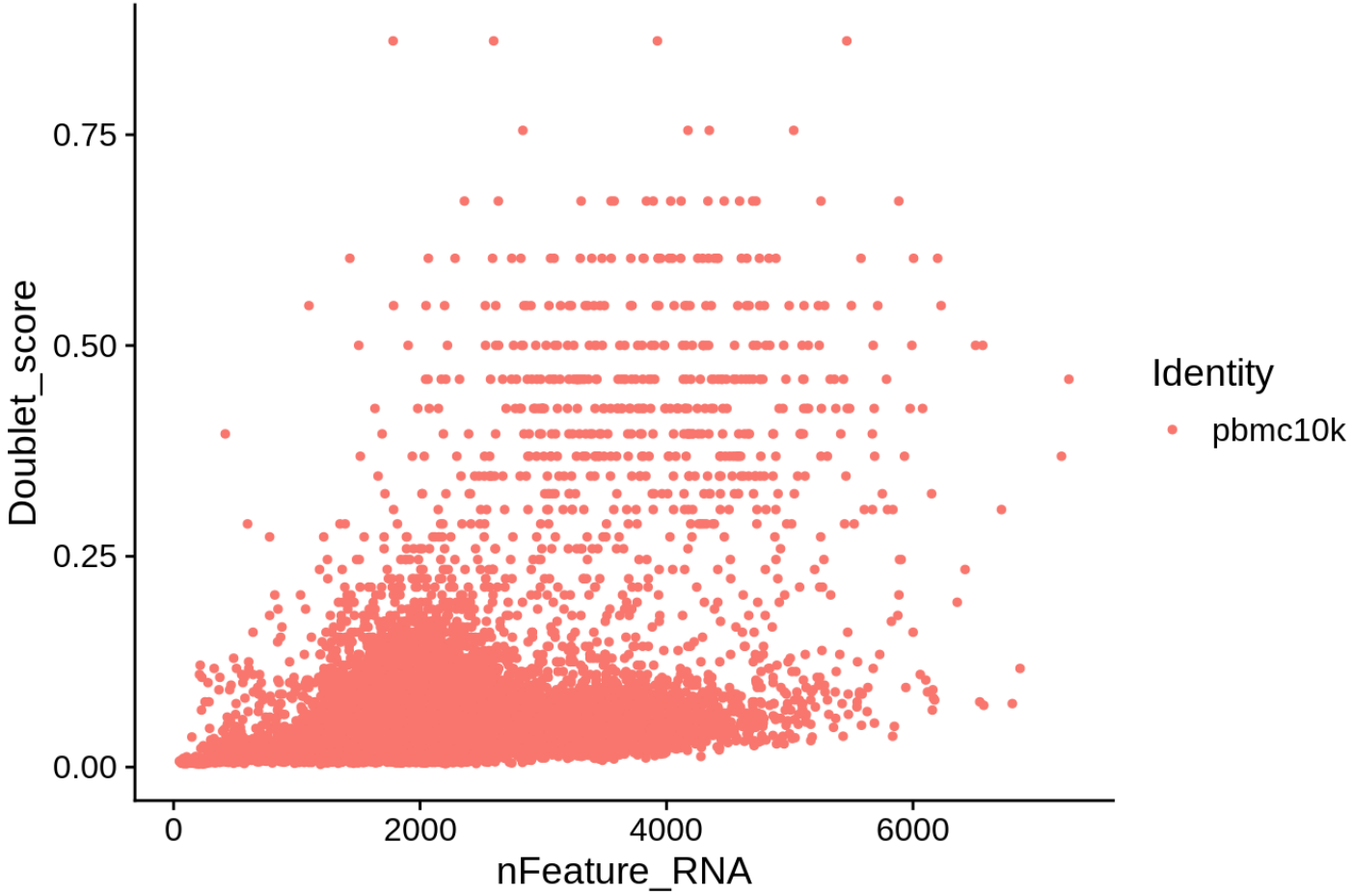
Cell Ranger 2.2



Cell Ranger 3.0

Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

Typical QC data can help ID problem data



Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

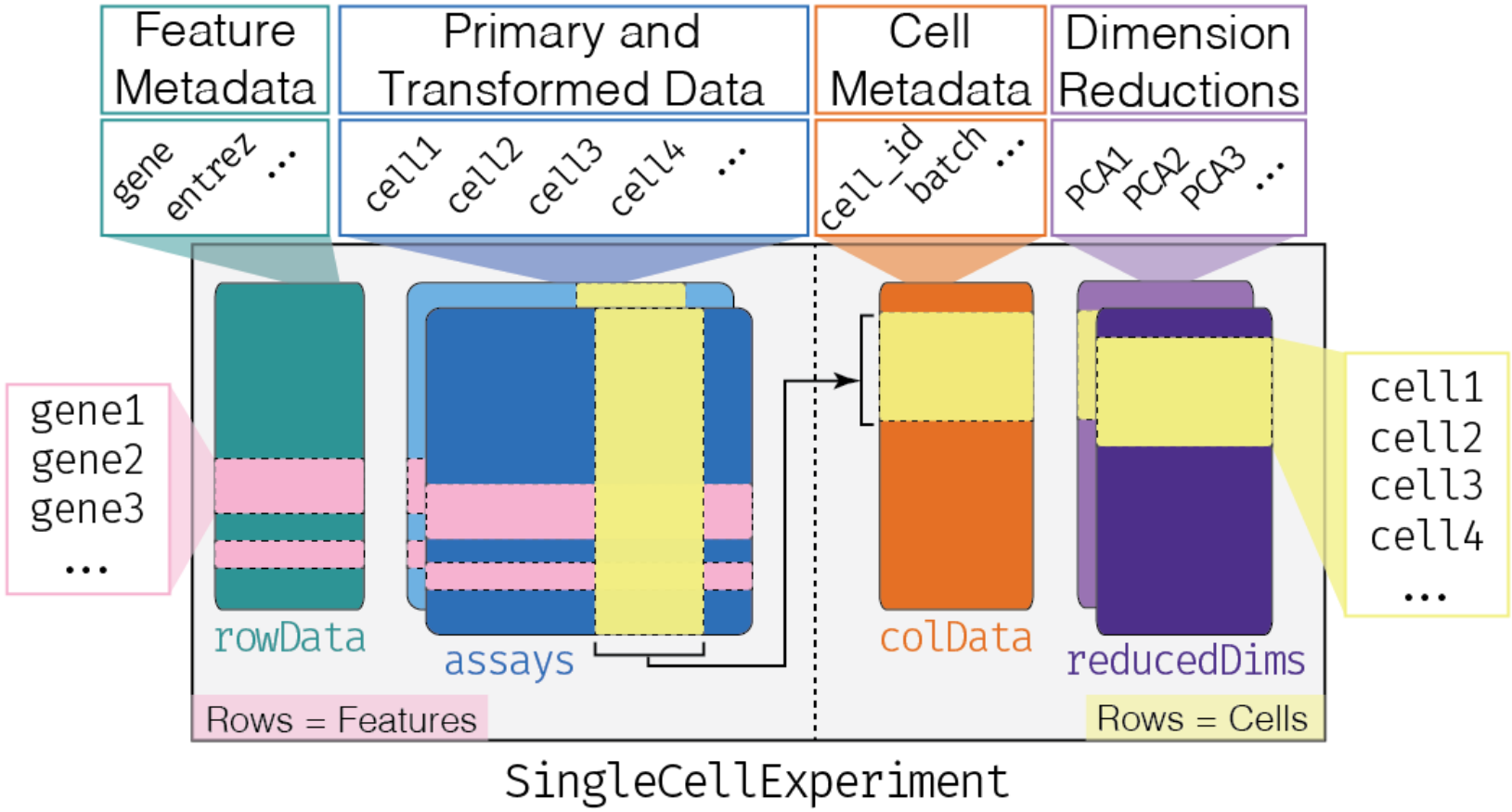
Finally- you have (sparse) count matrix(es)

- CellRanger outputs: gives two output folders raw and filtered
- Each folder has three zipped files
  - features.tsv.gz, barcodes.tsv.gz and matrix.mtx.gz
  - raw_feature_bc_matrix
    - All valid barcodes from GEMs captured in the data
    - Contains about half a million to a million barcodes
    - Most barcodes do not actually contain cells
  - filtered_feature_bc_matrix
    - Excludes barcodes that correspond to this background
    - Contains valid cells according to 10x cell calling algorithm
    - Contains 100s to 1000s of barcodes

```
%h%-$ ls SRR9264343/outs/raw_feature_bc_matrix
barcodes.tsv.gz
features.tsv.gz
matrix.mtx.gz
```


https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept22/

Computation tools have specialized data structures for single-cell analysis



Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

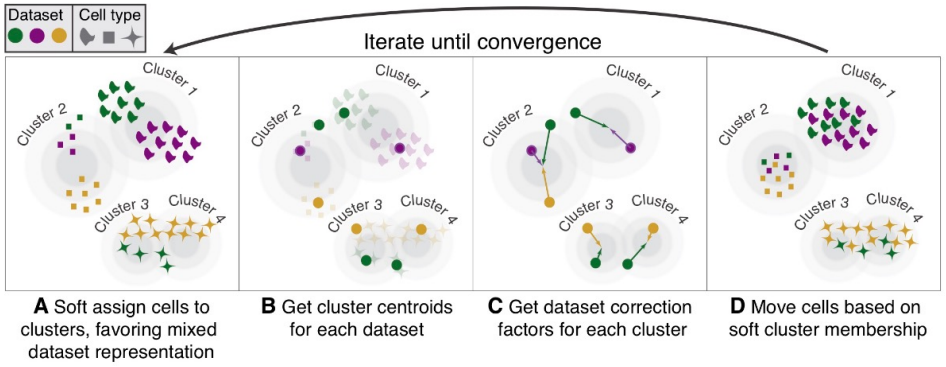
Following QC reduction, issues can still remain

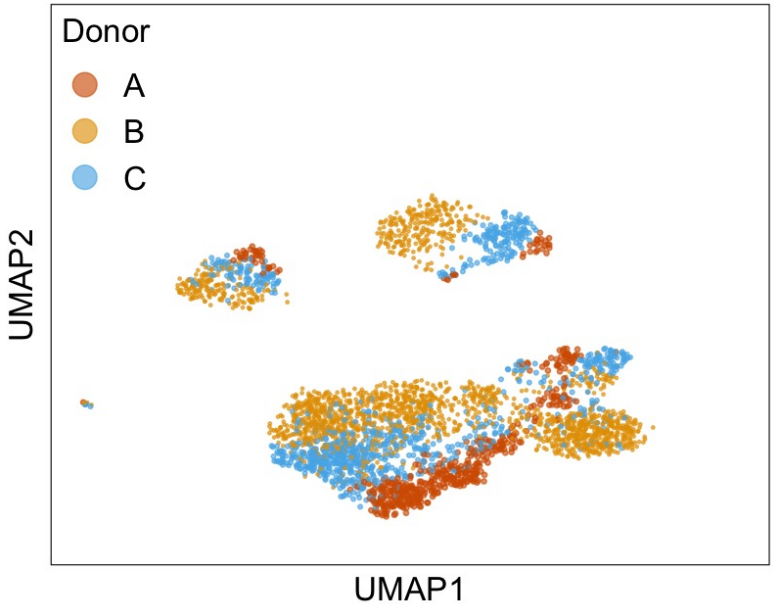
- Common systematic effects that can obscure biological effects
  - Batch effects
  - Cell cycle

Harmony

- **Harmony** : An algorithm that projects cells into clusters based on their cell identity rather than dataset specific conditions.

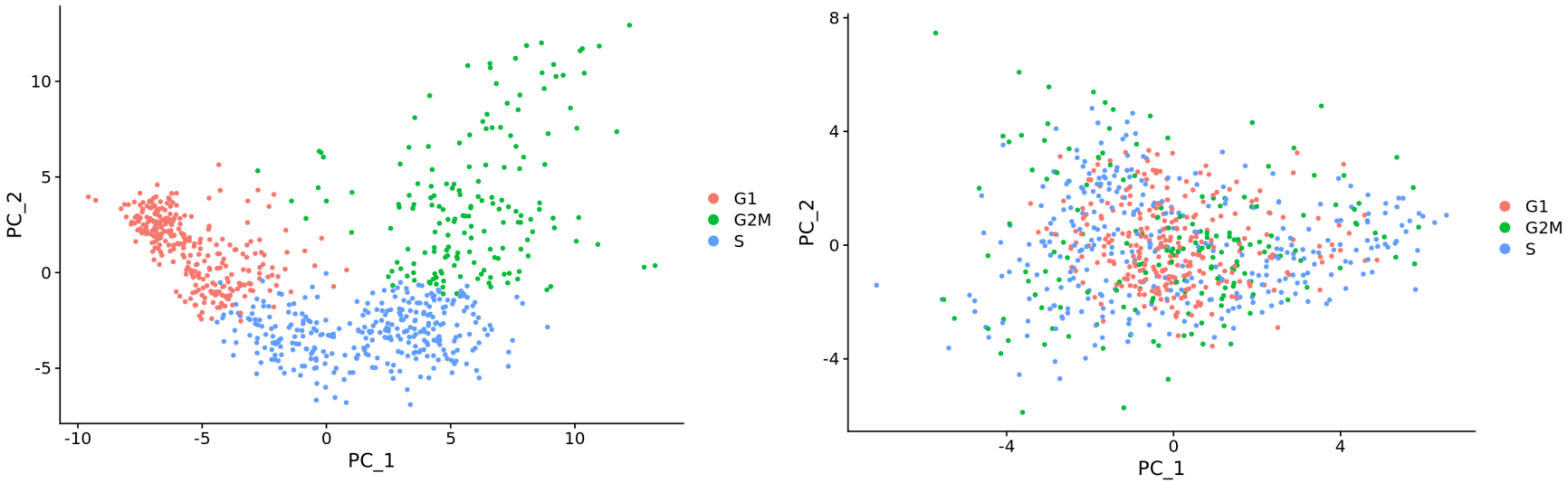
- Harmony applies a transformation to the principal component values. The algorithm then determines if there is a balanced quantity of cells from each batch within the clusters. Each cell is then evaluated to see how much its batch identity influences its PC coordinates. The cells position is corrected by shifting it towards the centroid of its cluster.





Cell cycle has a systematic (but known) measurable effect on expression patterns

- Characteristic genes have been identified
- Cells can be modeled to estimate cell cycle stage

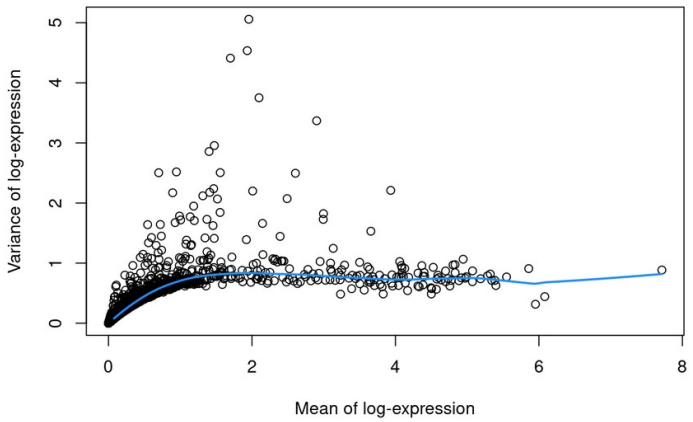


Source: https://satijalab.org/seurat/archive/v3.1/cell_cycle_vignette.html

Which genes should we use for downstream analysis?

Select genes which capture biologically-meaningful variation, while reducing the number of genes which only contribute to technical noise

(Image Source)



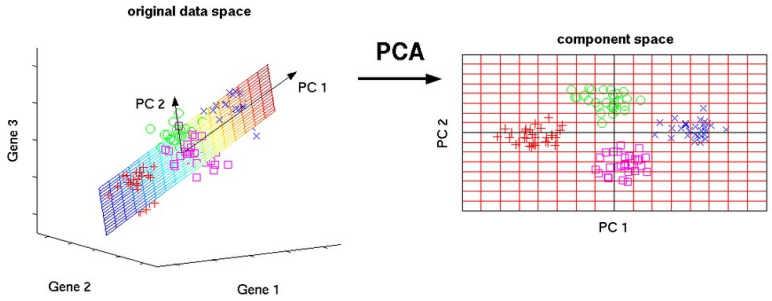
- Model the gene-variance relationship across all genes to define a data-driven "technical variation threshold"
- Select **highly variable genes** (HVGs) for downstream analysis (e.g. PCA and clustering)

Why do high-dimensional data pose a problem?

In single-cell data we typically have thousands of genes across thousands (or millions!) of cells.

- Interpretation/visualisation beyond 2D is hard.
- As we increase the number of dimensions, our data becomes more sparse.
- High computational burden for downstream analysis (such as cell clustering)

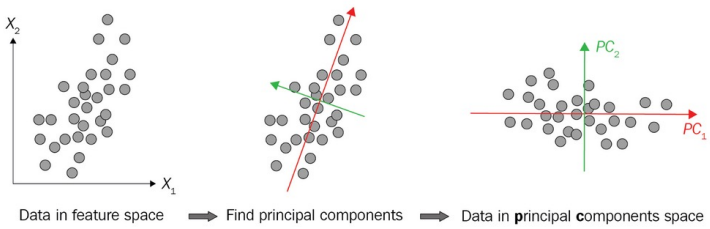
Solution: collapse the number of dimensions to a more manageable number, while preserving information.



(Image source)

All downstream analysis is based on “reduced” data

Principal Components Analysis (PCA)

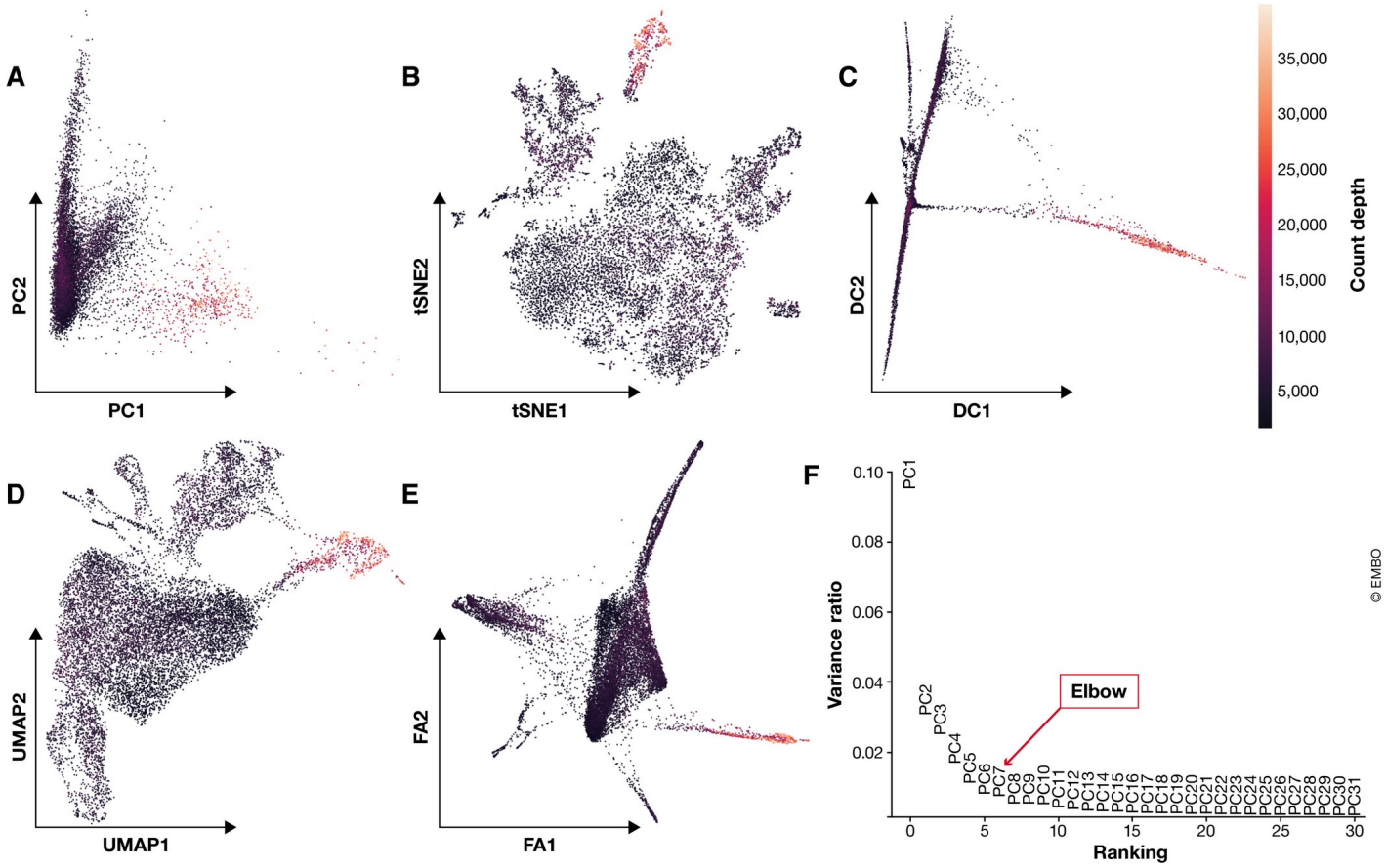


(Image Source)

- It’s a linear algebraic method of dimensionality reduction
- Finds principal components (PCs) of the data
  - Directions where the data is most spread out (highest variance)
  - PC1 explains most of the variance in the data, then PC2, PC3, etc.
  - PCA is primarily a dimension reduction technique, but it is also useful for visualization
  - A good separation of dissimilar objects is provided
  - Preserves the global data structure

https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept22/

Common visualization options



Mol Syst Biol, Volume: 15, Issue: 6, First published: 19 June 2019, DOI: (10.15252/msb.20188746)

Clustering and Biology: What do you want to learn from the experiment?

- Classify cells and discover new cell populations
- Compare gene expression between different cell populations
- Reconstruct developmental 'trajectories' to reveal cell fate decisions of distinct cell subpopulations

Source: http://barc.wi.mit.edu/education/hot_topics/scRNAseq_2020/SingleCellRNAseq2020_4slidesPerPage.pdf

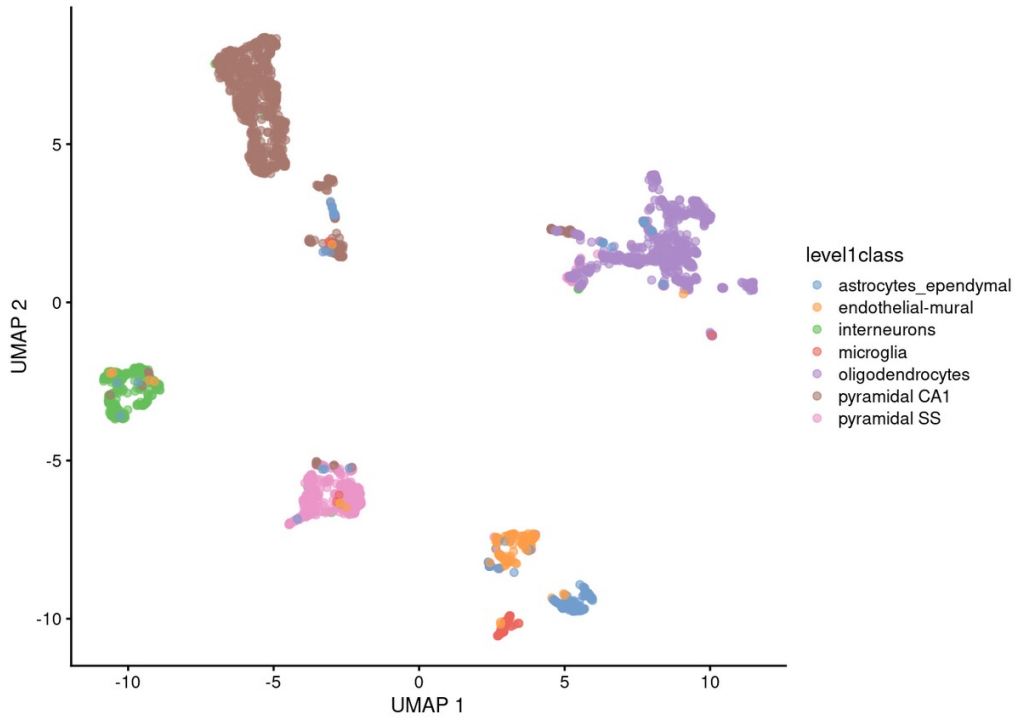
UMAP

Main parameter in UMAP is `n_neighbors` (the **number of neighbours** used to construct the initial graph).

Another common parameter is `min_dist` (**minimum distance** between points)

- Together they determine balance between preserving local vs global structure
- For practical simplicity, we usually only tweak `n_neighbors`, although playing with both parameters can be beneficial

Exploring different number of neighbours that best represent the biological diversity of cells is recommended.



https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept22/

Is there a "correct" clustering?

Clustering, like a microscope, is a tool to explore the data.

We can zoom in and out by changing the resolution of the clustering parameters, and experiment with different clustering algorithms to obtain alternative perspectives on the data.

Asking for an unqualified "best" clustering is akin to asking for the best magnification on a microscope.

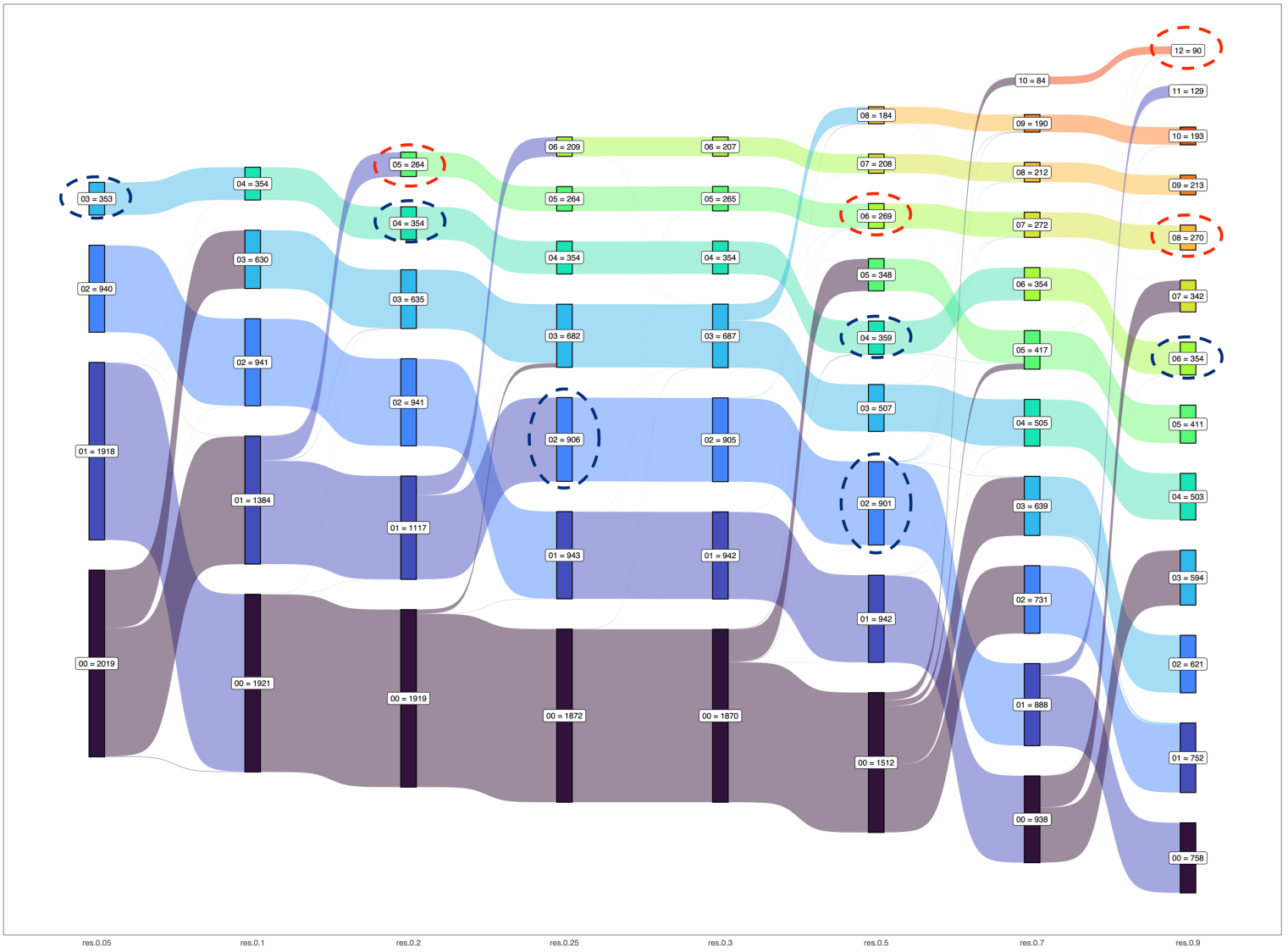
A more relevant question is "how well do the clusters approximate the cell types or states of interest?". Do you want:

- resolution of the major cell types?
- Resolution of subtypes?
- Resolution of different states (e.g., metabolic activity, stress) within those subtypes?

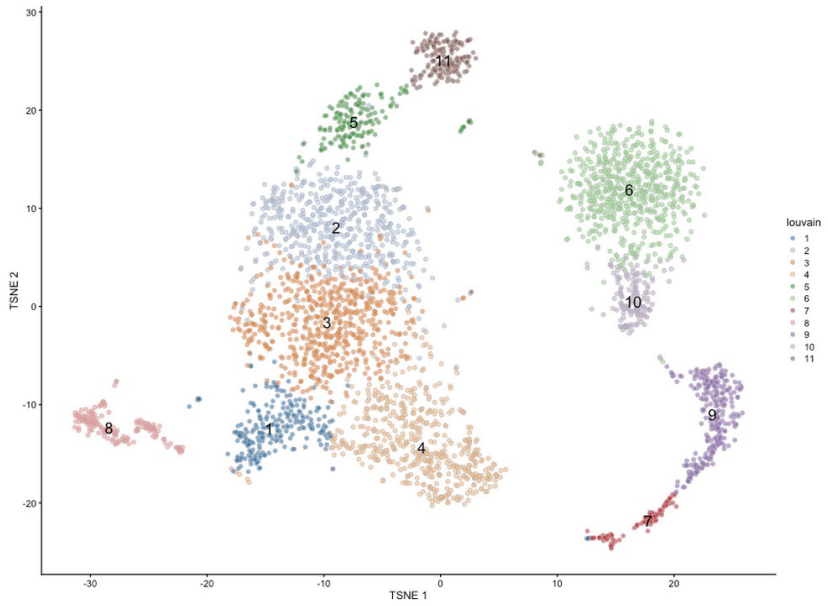
Explore the data, use your biological knowledge!

Image by Les Chatfield from Brighton, England - Fine rotative table Microscope 5, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=32225637

Sankey: Macro Subset Dataset



Identifying Cluster Marker Genes



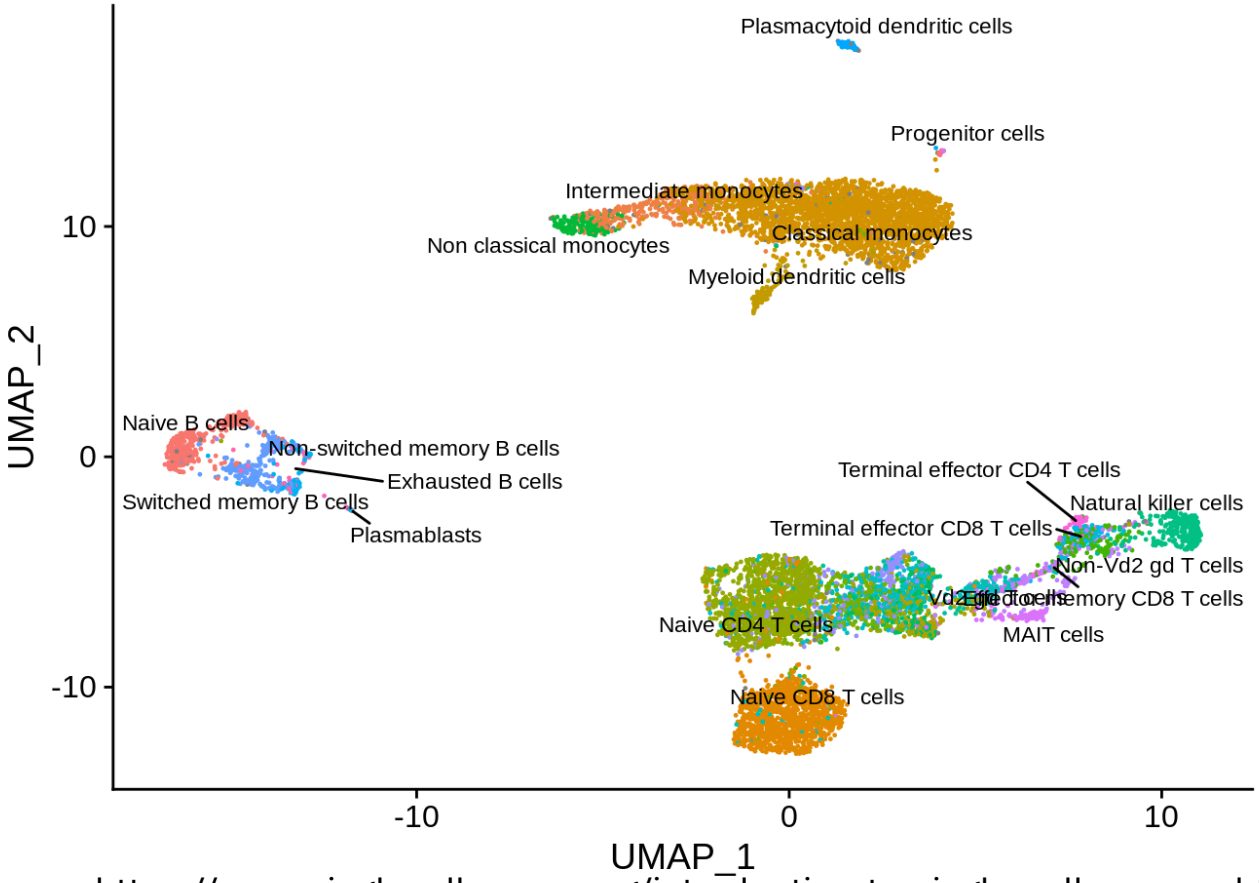
Our goal is to identify genes that are differently expressed between clusters

Calculate effect sizes that capture differences in:

- mean expression level
- rank of expression
- proportion of cells expressing the gene

These are calculated in pairwise cluster comparisons.

Cell identification has generally been based upon known marker genes



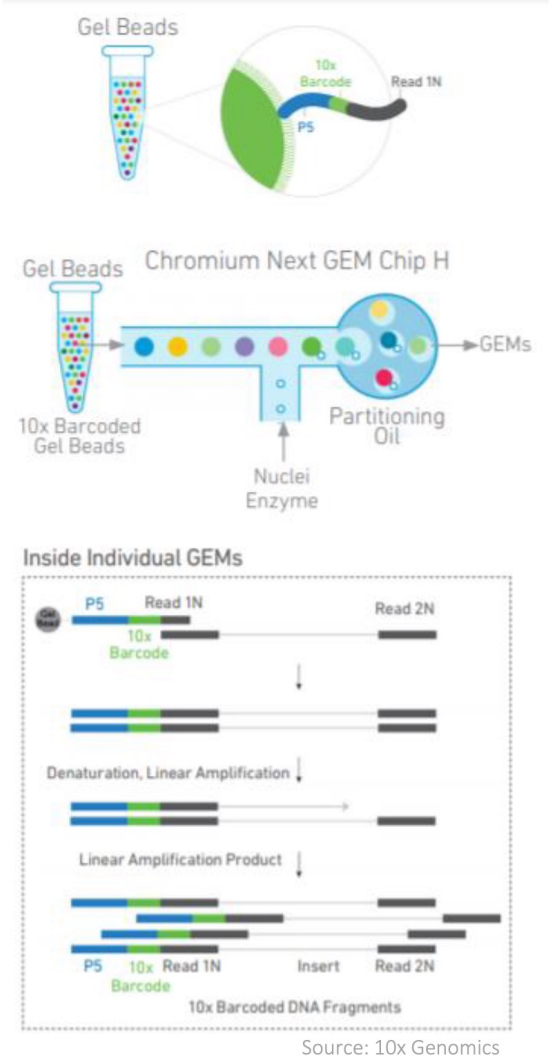
Source: https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html

10X ATAC

Chromium Single Cell ATAC libraries comprise double stranded DNA fragments which begin with P5 and end with P7. Sequencing these libraries produces a standard Illumina® BCL data output folder.

Sequencing Read	Description	Number of cycles
Read1	Insert Sequence 1N	50bp
i7 index	Sample index read	8bp
i5 index	10x Barcode Read (Cell)	16bp
Read2	Insert Sequence 2N (opposite end)	50bp

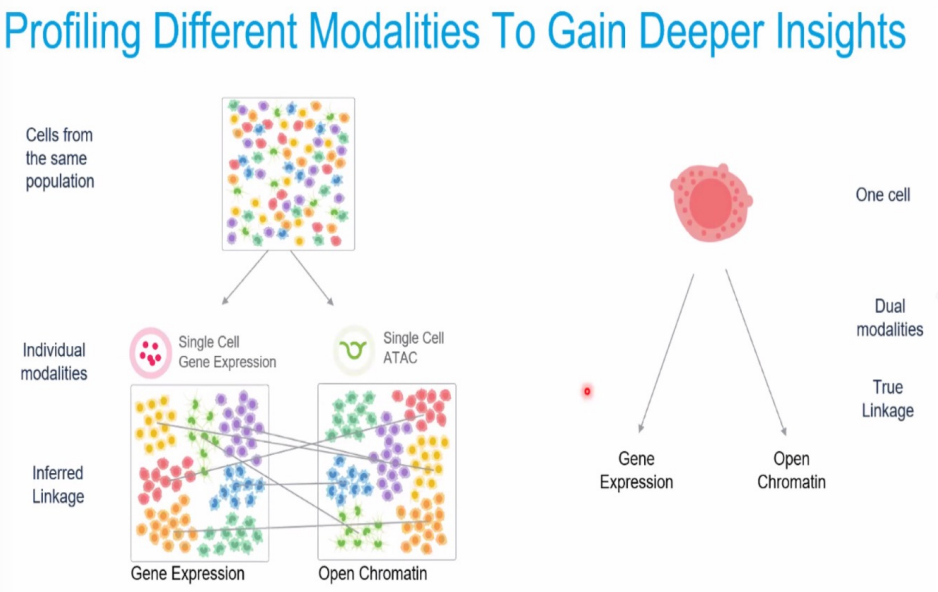
- ASAP-seq is to scATAC-seq what CITE-seq is to scRNA-seq.
- Scale Biosciences – 'pre-indexing of nuclei through tagmentation' = 100k nuclei per 10x channel with low number of doublets

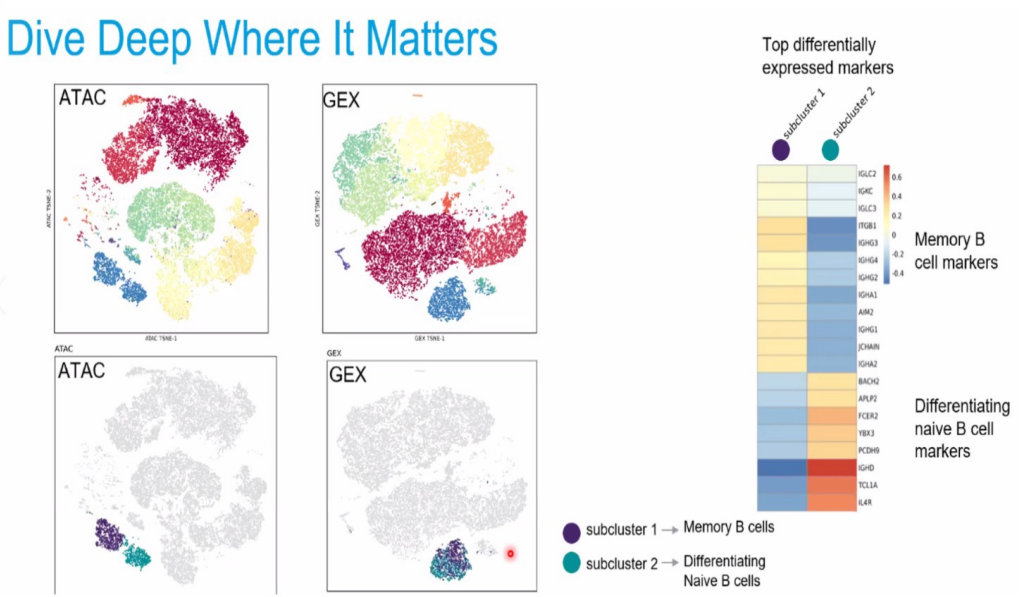


Source: 10x Genomics

Source:https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf

10X MULTIOME (RNA+ATAC)





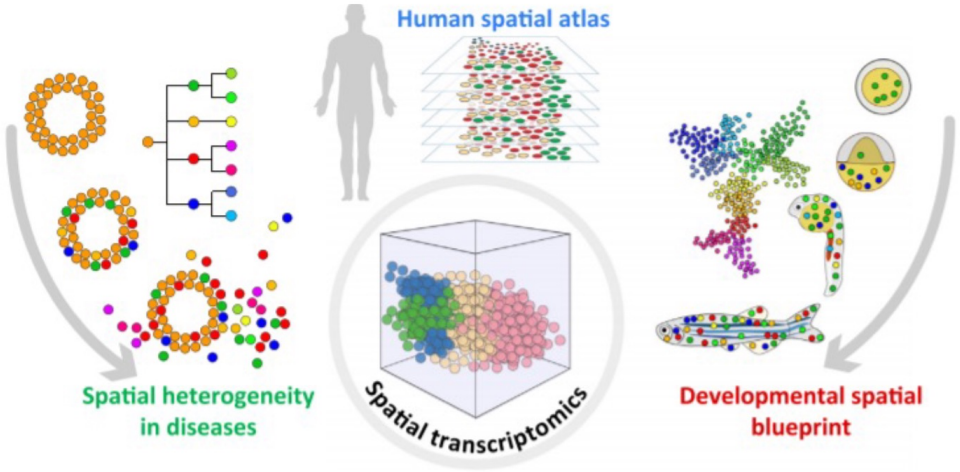
Source: 10x Genomics

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

-TEA-seq (Transcription, Epitopes, and Accessibility) = Multiome with permabilised cells & CITEseq

Source:https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf

SPATIAL TRANSCRIPTOMICS

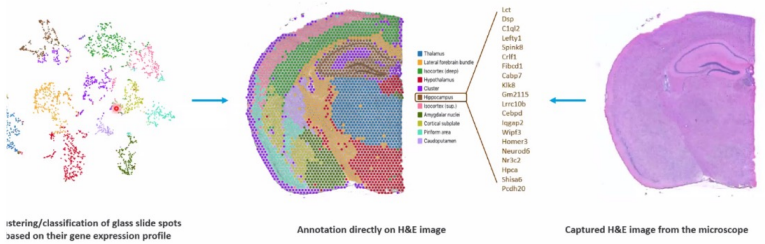


Trends in Biotechnology

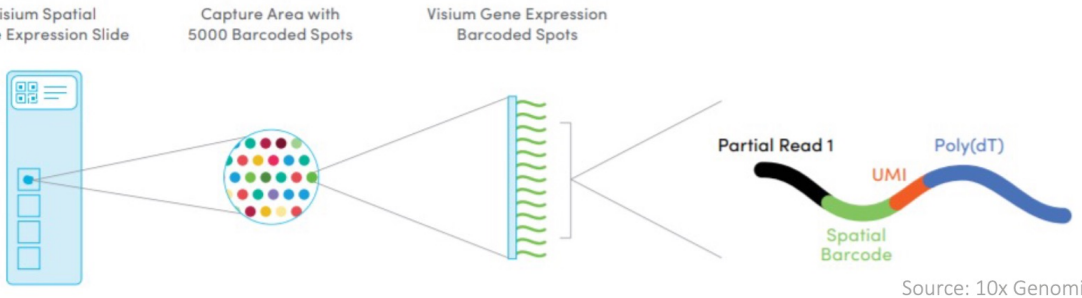
Figure 3. Applications for Spatially Resolved Transcriptomics. Three primary kinds of hot issues can be resolved by spatially resolved transcriptomics: left, discovering spatial heterogeneity of diseases; middle, establishing spatial transcriptome atlases for the human body; and right, delineating an embryonic developmental and spatial blueprint.

Source: Liao et al. Trends in Biotechnology. (2020)

55ul spots -> 1 and 10 cells captured per spot



Source: 10x Genomics

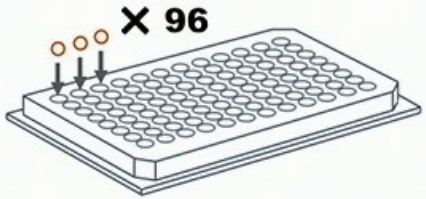


Source: 10x Genomics

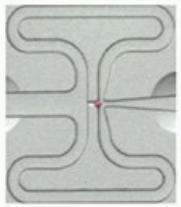
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Source:https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Jan23/UnivCambridge_ScRnaSeqIntro_Base/Slides/01_Introduction.pdf

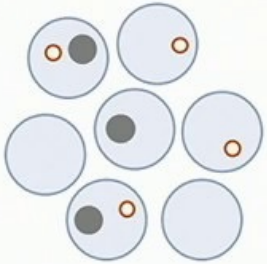
Most single cell methods rely on physical compartmentalization of cells and barcoding primers



Single cell / single well



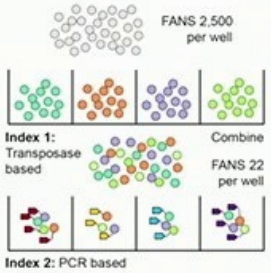
Microfluidics Chip (e.g. Fluidigm)



Droplet-based (e.g. 10X & BioRad)

Microwell-based (e.g. Takara, CelSee)

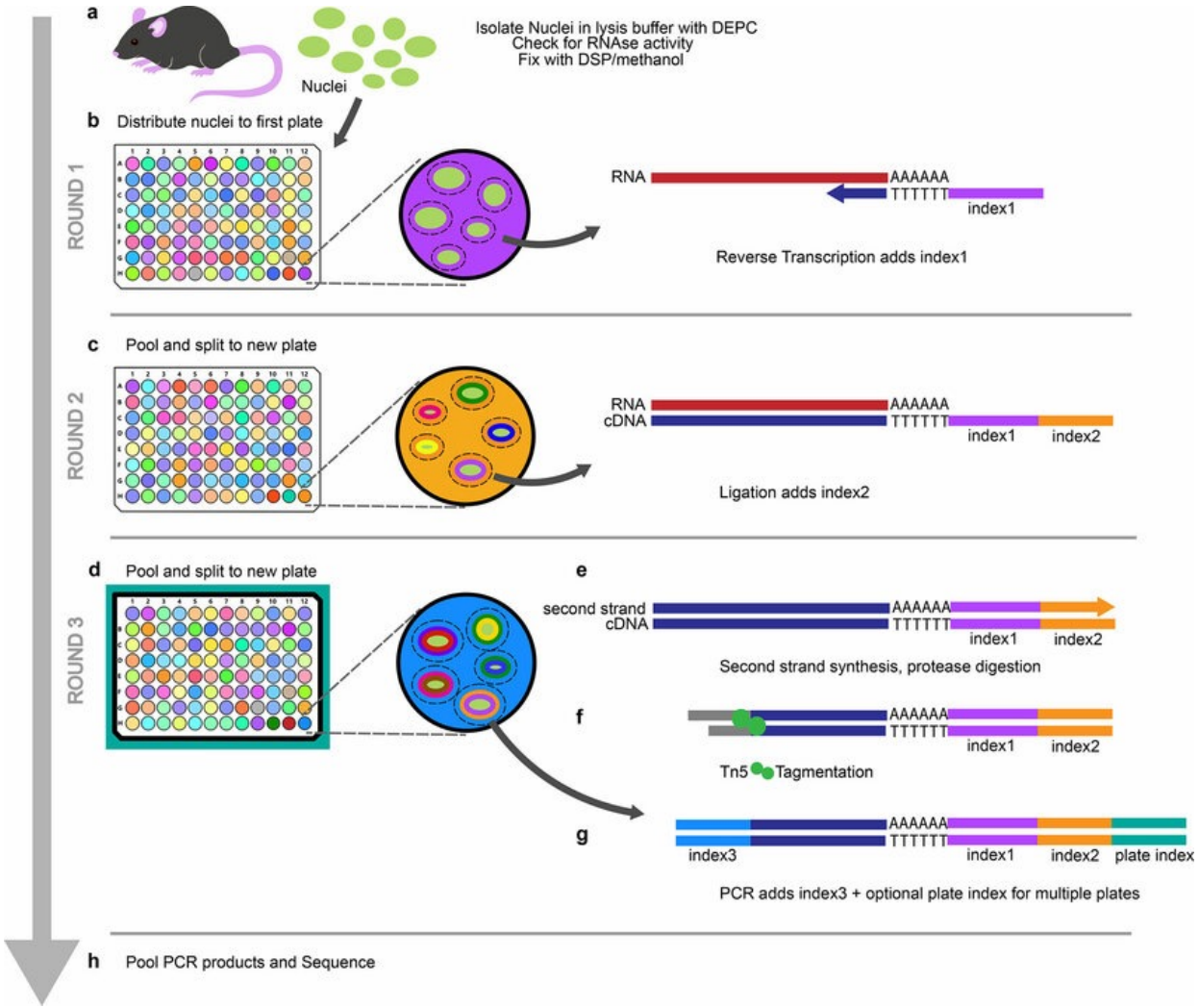
Single-cell combinatorial indexing – an alternative means of high throughput single-cell analysis.



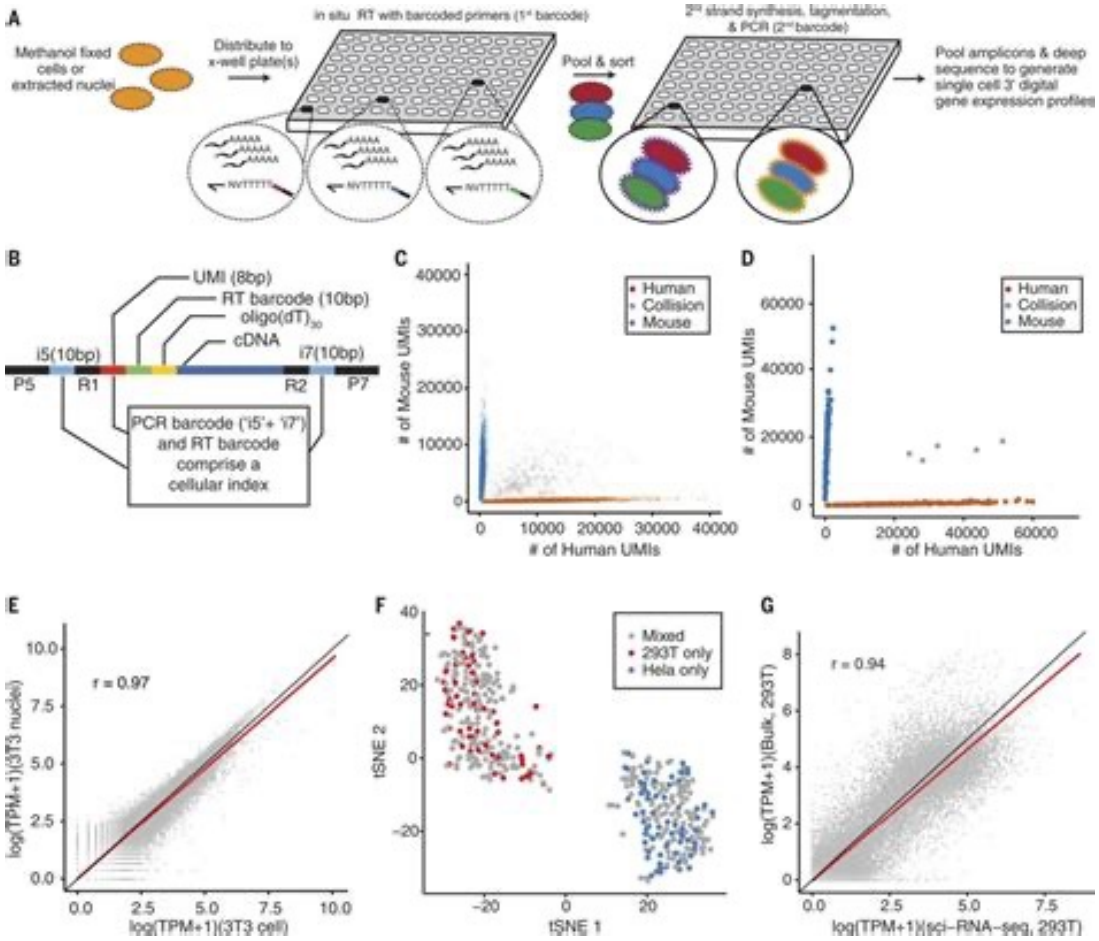
- Combinatorial scaling of throughput
- No specialized equipment required
- Adaptable to a variety or properties
- Inexpensive to perform

OHSU

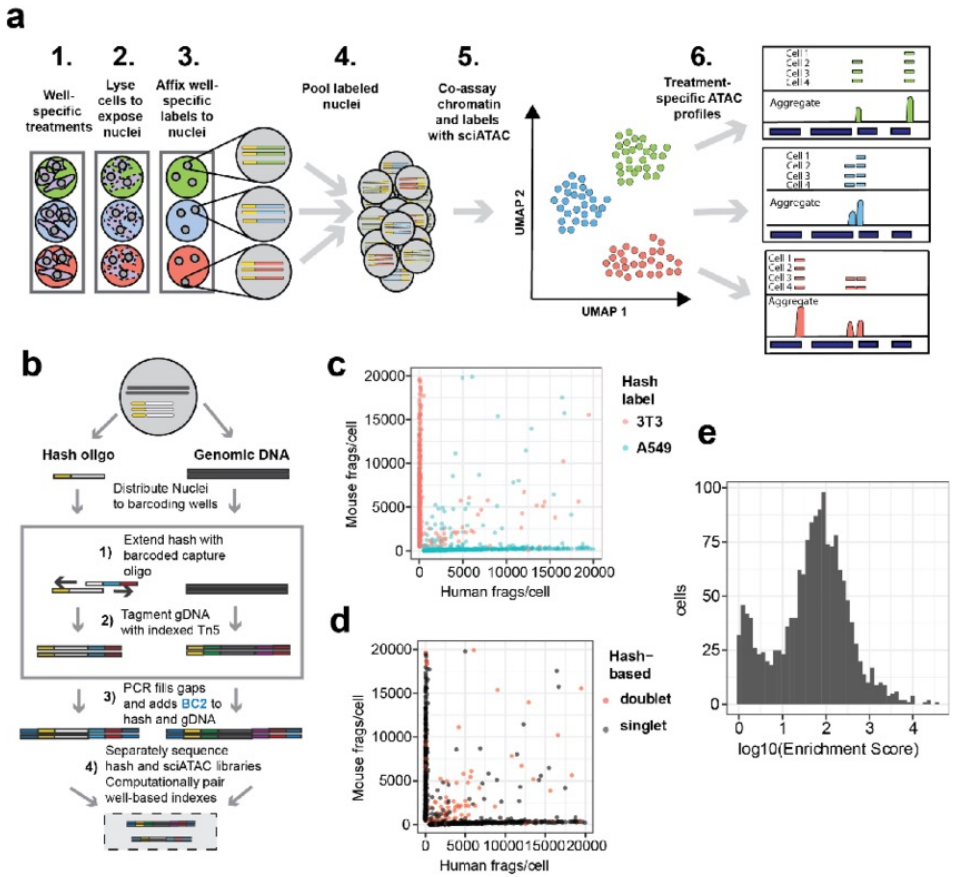
Combinatorial indexing allows for great cost reduction per cell



Combinatorial indexing allows for great cost reduction per cell



Nuclear Oligo Hashing allows both sample labeling and impoved normalization




Questions?

WHAT PLATFORM SHOULD I USE?

Choose protocol based on:

- Throughput (number of cells per reaction)
- Sample of origin
- Cost / Labour / Time limitations
- Gene body coverage: 5'/ 3' biased or full-length?
- UMI vs no-UMI
- Sequencing depth per cell

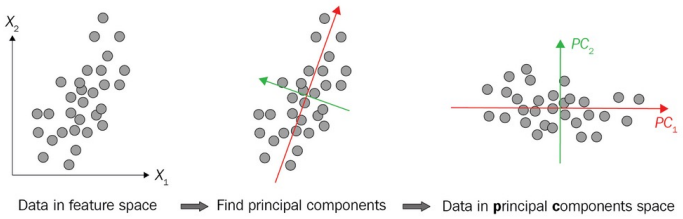
Examples:

- If you sample is fairly homogeneous – bulk RNAseq
- If your sample is limited in cell number – plate-based method
- If you want re-annotate the transcriptome and discover new isoforms – full-length coverage (SMART-seq2, seqWell)
- If you are looking to classify all cell types in a diverse tissue - high throughput
- If you have only archival human samples – nuclei isolation or 10x fixed RNA profiling

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

All downstream analysis is based on PCA reduction of data

Principal Components Analysis (PCA)



(Image Source)

- When data is very highly-dimensional, we can select the most important PCs only, and use them for downstream analysis (e.g. clustering cells)
  - This reduces the dimensionality of the data from ~20,000 genes to maybe 20-50 PCs
  - Each PC represents a robust 'metagene' that combines information across a correlated gene set
- Prior to PCA we scale the data so that genes have equal weight in downstream analysis and highly expressed genes don't dominate

https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept22/

Making a graph

Nearest-Neighbour (NN) graph:

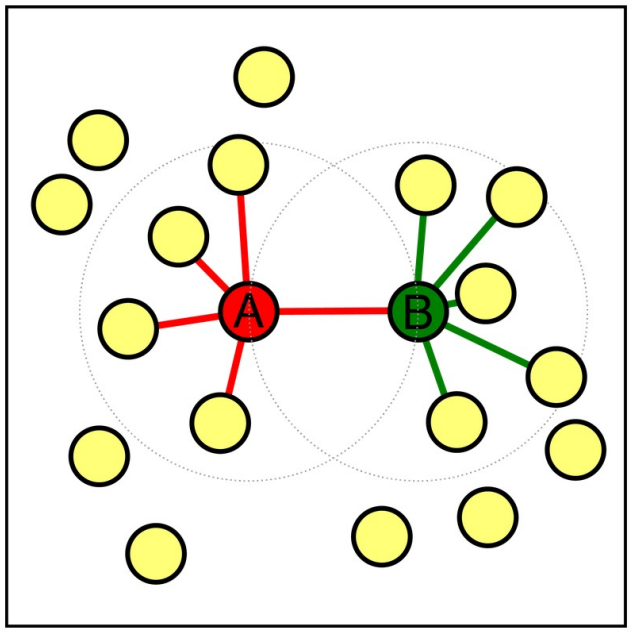
- cells as nodes
- their similarity as edges

In a NN graph two nodes (cells), say A and B, are connected by an edge if:

- the distance between them (in e.g. principal component space) is amongst the **k** smallest distances (here k = 5) from A to other cells, (**K**NN)

or

- In a **shared**-NN graph (**S**NN) two cells are connected by an edge if any of their nearest neighbors are shared (n.b. in Seurat this is different)



Once edges have been defined, they can be weighted. By default the weights are calculated using the 'rank' method which relates to the highest ranking of their shared neighbours.

https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept22/

Identifying communities/clusters - Louvain

Nodes are also first assigned their own community.

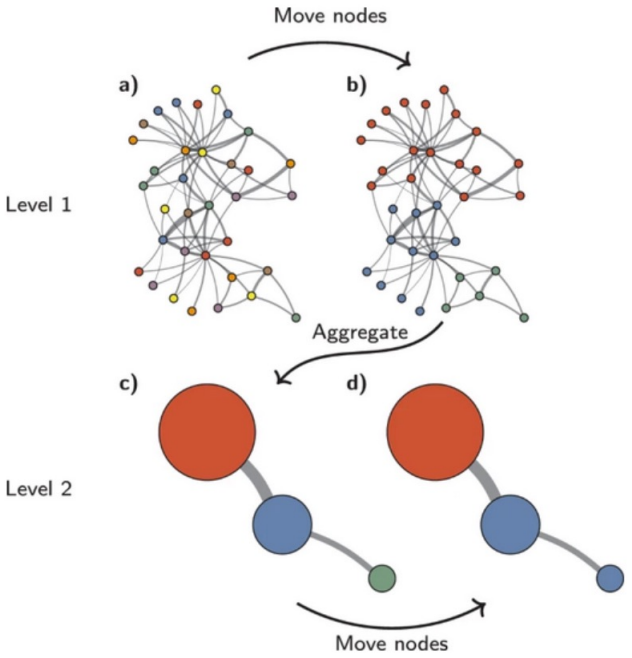
Two-step iterations:

- nodes are re-assigned one at a time to the community for which they increase modularity the most,
- a new, 'aggregate' network is built where nodes are the communities formed in the previous step.

This is repeated until modularity stops increasing.

(Blondel et al, Fast unfolding of communities in large networks)

(Traag et al, From Louvain to Leiden: guaranteeing well-connected communities)



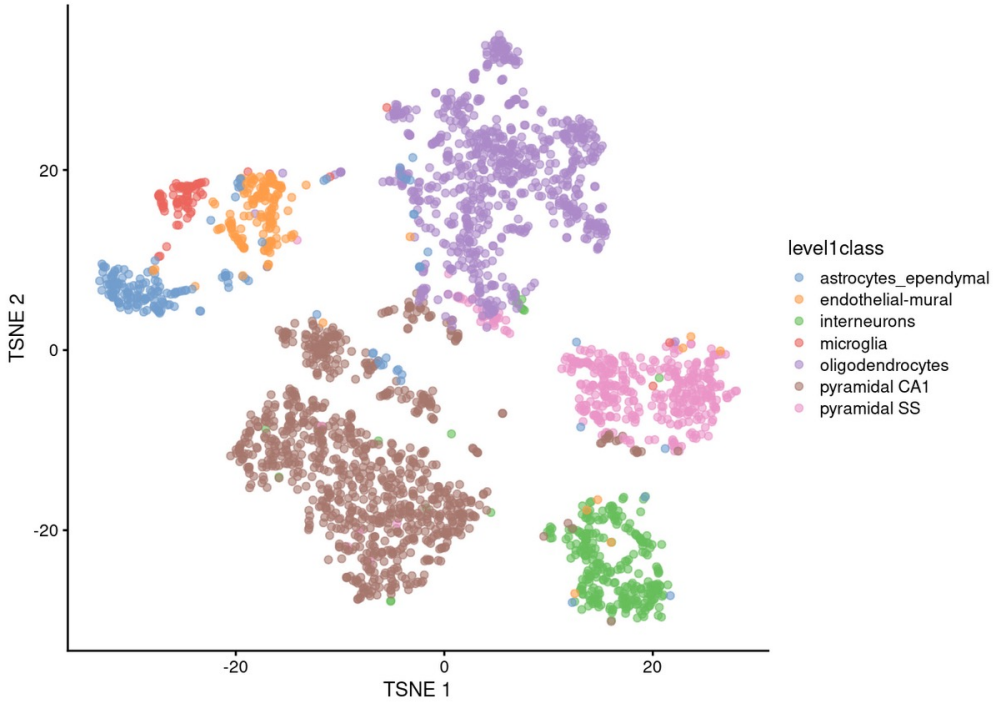
https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept22/

t-SNE

Main parameter in t-SNE is the **perplexity** (~ number of neighbours each point is "attracted" to)

- Balance between preserving local vs global structure
- Higher values usually result in more compact clusters
- But too high can lead to overlap of clusters, making them harder to distinguish

Exploring different perplexity values that best represent the biological diversity of cells is recommended.



https://bioinformatics-core-shared-training.github.io/SingleCell_RNASeq_Sept22/