VIII Curso de Férias em Bioquímica, 2024

Mini-curso: SINGLE CELL GENE EXPRESSION PROFILING

Ph.D. Estevao Barcelos

Università degli Studi di Perugia - UNIPG
Universidade Federal do Espirito Santo - UFES

Part I: INTRO SINGLE CELL RNA SEQUENCING

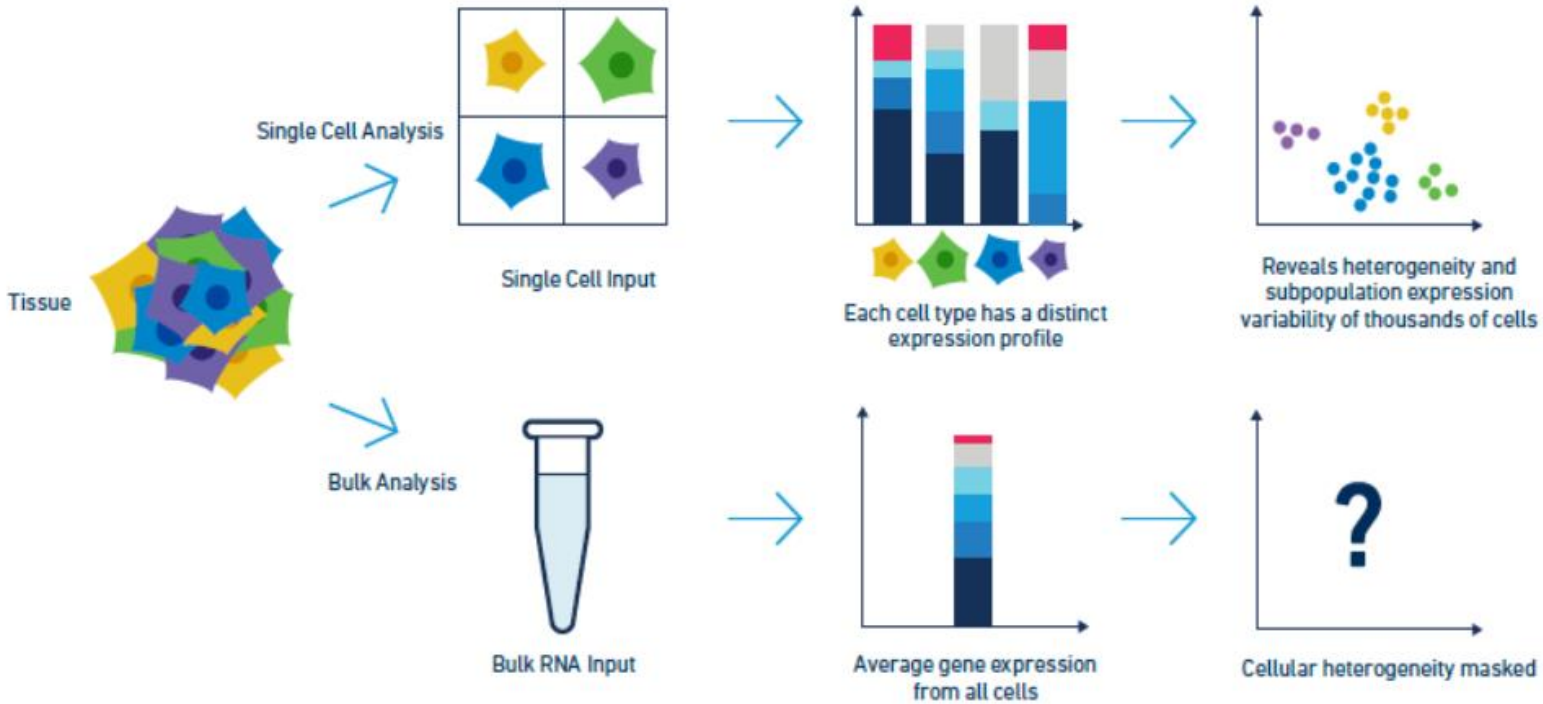
estevaocarlosbarcelos@gmail.com

IMPORTANCE OF SINGLE-CELL MULTIMODAL OMICS

Single-cell sequencing
selected as 2013
Method of the Year

Single-cell multimodal omics
selected as 2019
Method of the Year

IMPORTANCE OF SINGLE-CELL LEVEL ANALYSIS



- Bulk methods measure the average gene expression from all cells.
- Single-cell technologies measure the gene expression profile of each individual cell.

www.10XGenomics.com

Single-cell sequencing and Spatial Transcriptomics Technologies allowed large international initiatives :

HUMAN CELL ATLAS

MISSION

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.

nature

Human BioMolecular Atlas Program

A collection of research articles and related content from the Human BioMolecular Atlas Program describing the distribution of biomolecules across single cells, tissues and organs in the human body.

Outline – Part I

Evolution of single-cell RNA and multi-omics technologies,

Applications of single-cell sequencing technologies,

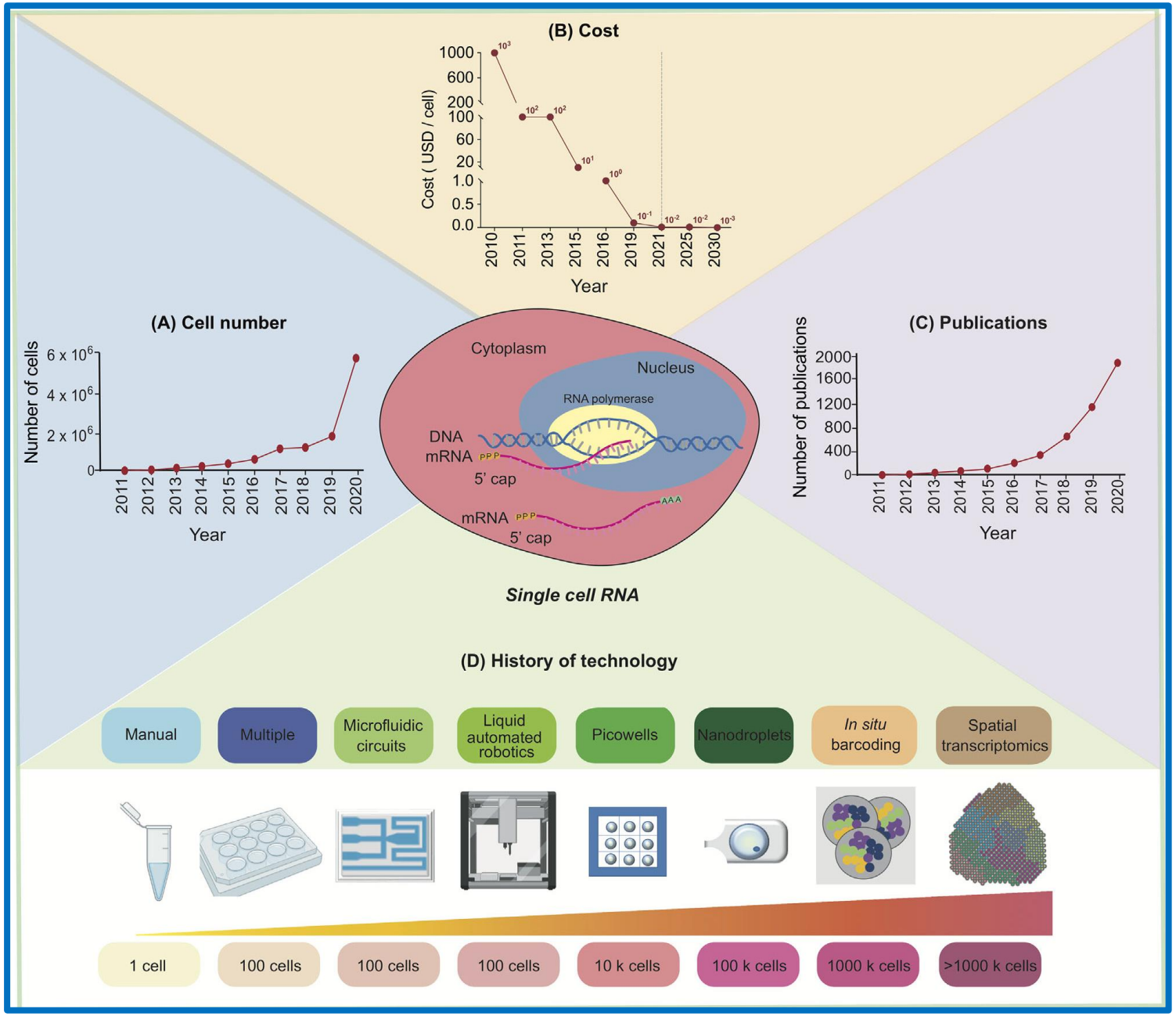
Overview of single cell RNA-seq analysis,

Outline – Part II

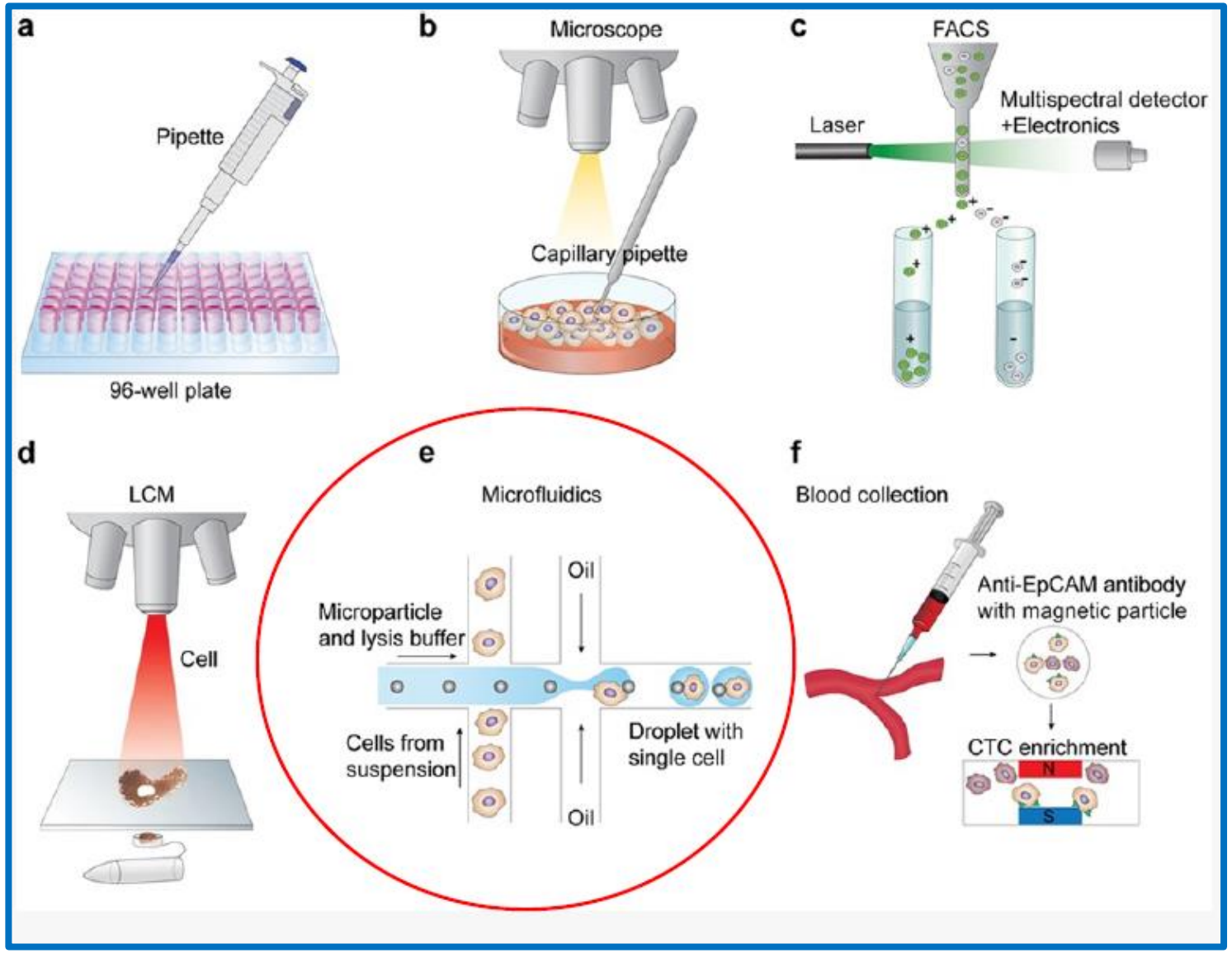
Overview of sc RNA-seq data analysis and Hands-on.

Evolution of single-cell RNA and multi-omics technologies

- **Cell Analysis:**
  - The volume of cells subjected to analysis experienced an increase.
- **Cost Reduction:**
  - There was an exponential reduction in costs (in US dollars).
- **Scientific Output:**
  - The number of published papers demonstrated a notable increase.



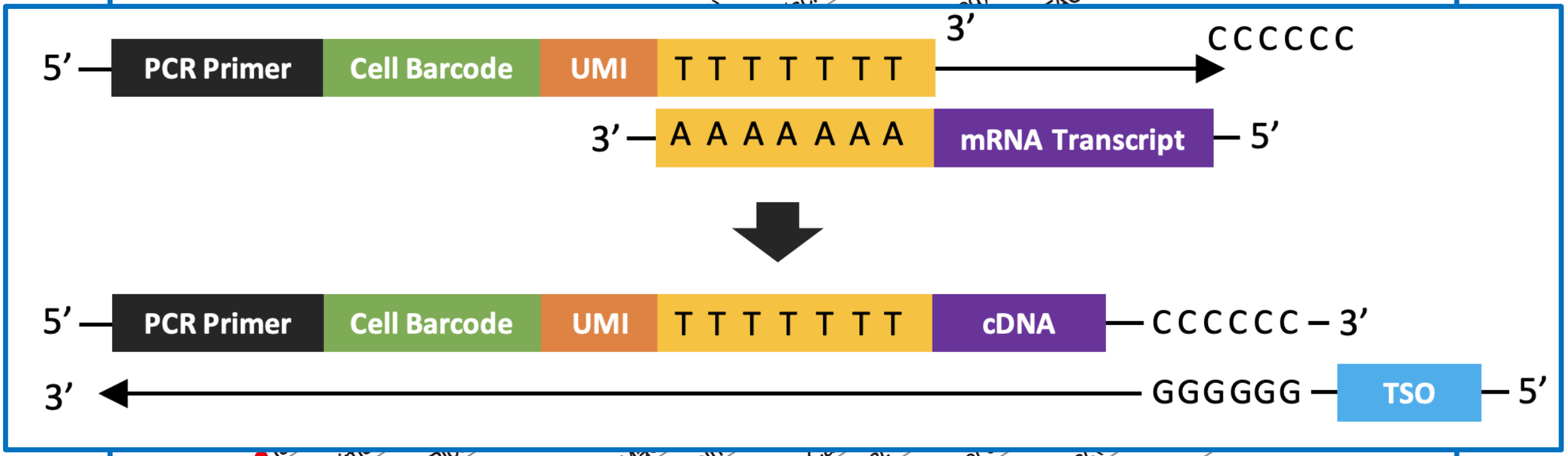
Jovic D et al., Clin Transl Med. 2022.



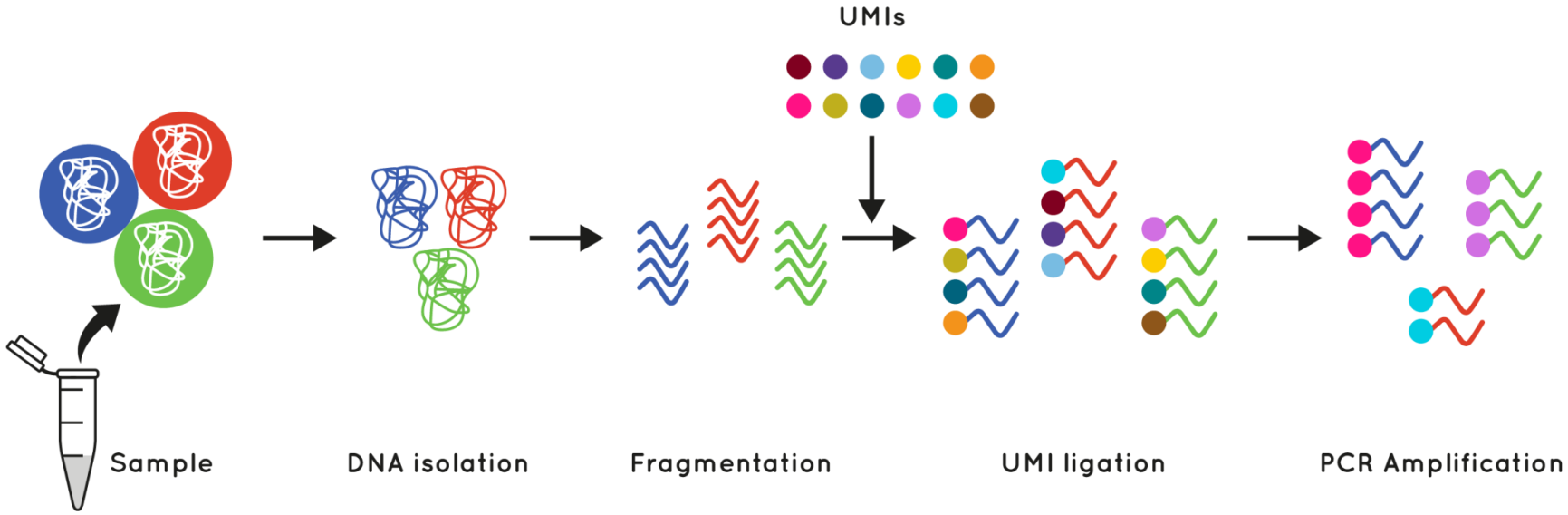
- **The limiting dilution method:**
  - Isolates individual cells, leveraging the statistical distribution of diluted cells.
- **Micromanipulation:**
  - Involves collecting single cells using microscope-guided capillary pipettes.
- **FACS:**
  - Isolates highly purified single cells by tagging cells with fluorescent marker proteins.
- **Laser capture microdissection (LCM):**
  - Utilizes a laser system aided by a computer system to isolate cells from solid samples.
- **Microfluidic technology for single-cell isolation:**
  - Requires nanoliter-sized volumes (cells partitioning in droplets). 10X
- **The CellSearch system:**
  - Enumerates CTCs from patient blood samples by using a magnet conjugated with CTC binding antibodies.

Hwang B et al. Exp Mol Med, 2018.

Timeline of technical progress of single cell analysis

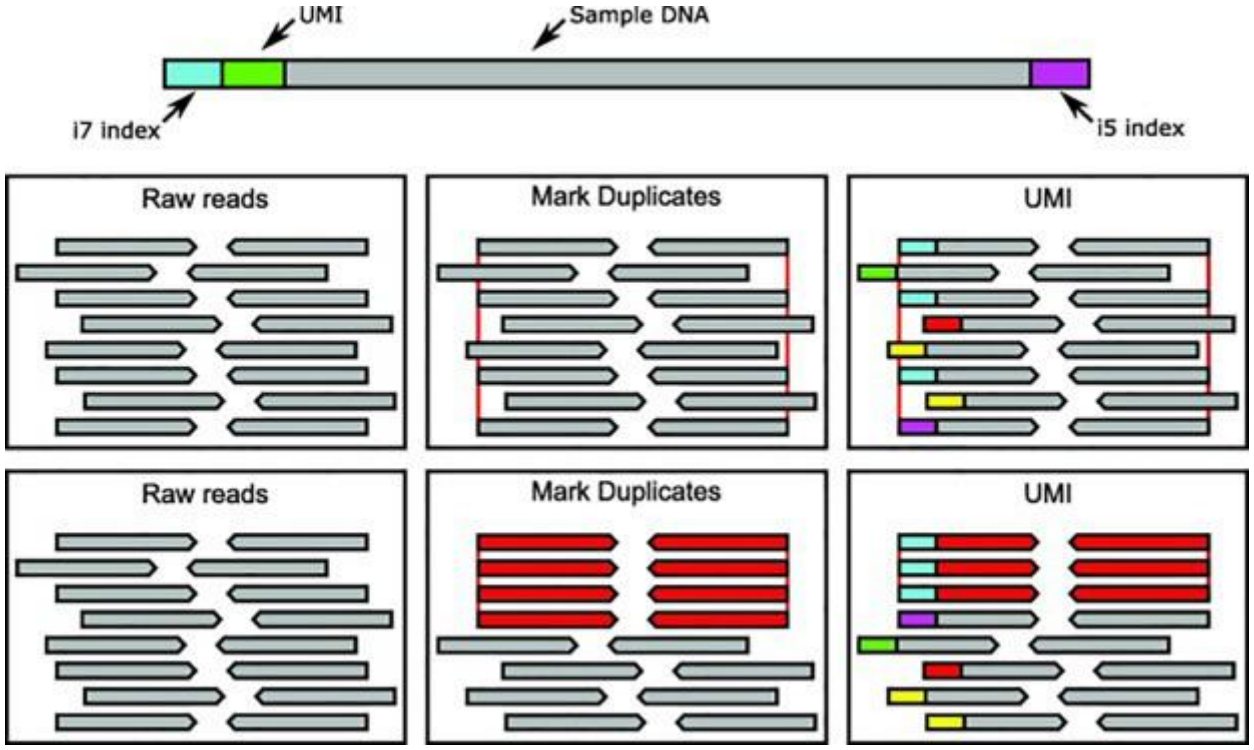


Asada K et al., Biomedicines. 2021.



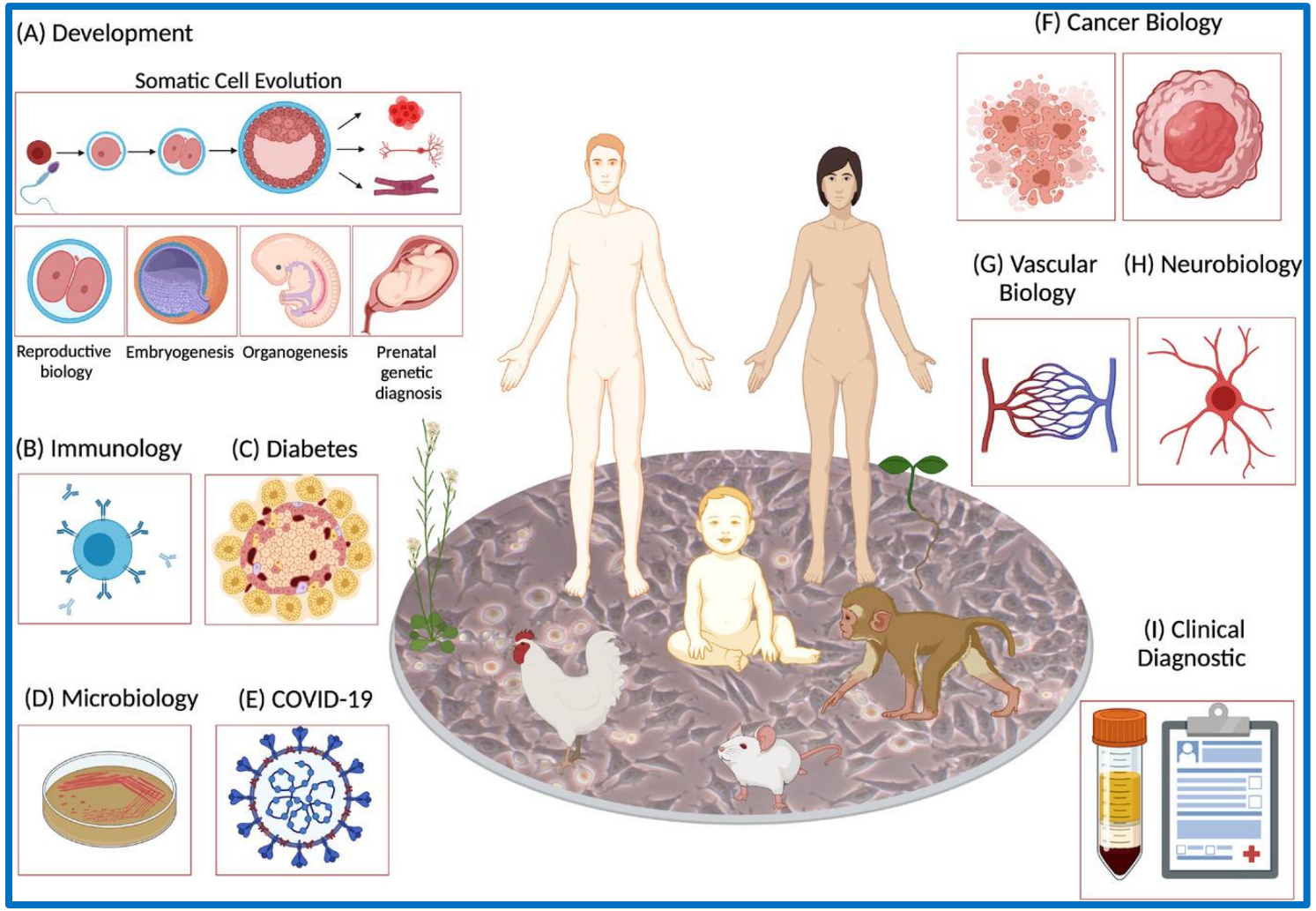
So, what are UMIs and why are they useful?

https://nonacus.com/blog-unique-molecular-identifiers-unmask-low-frequency-variants/

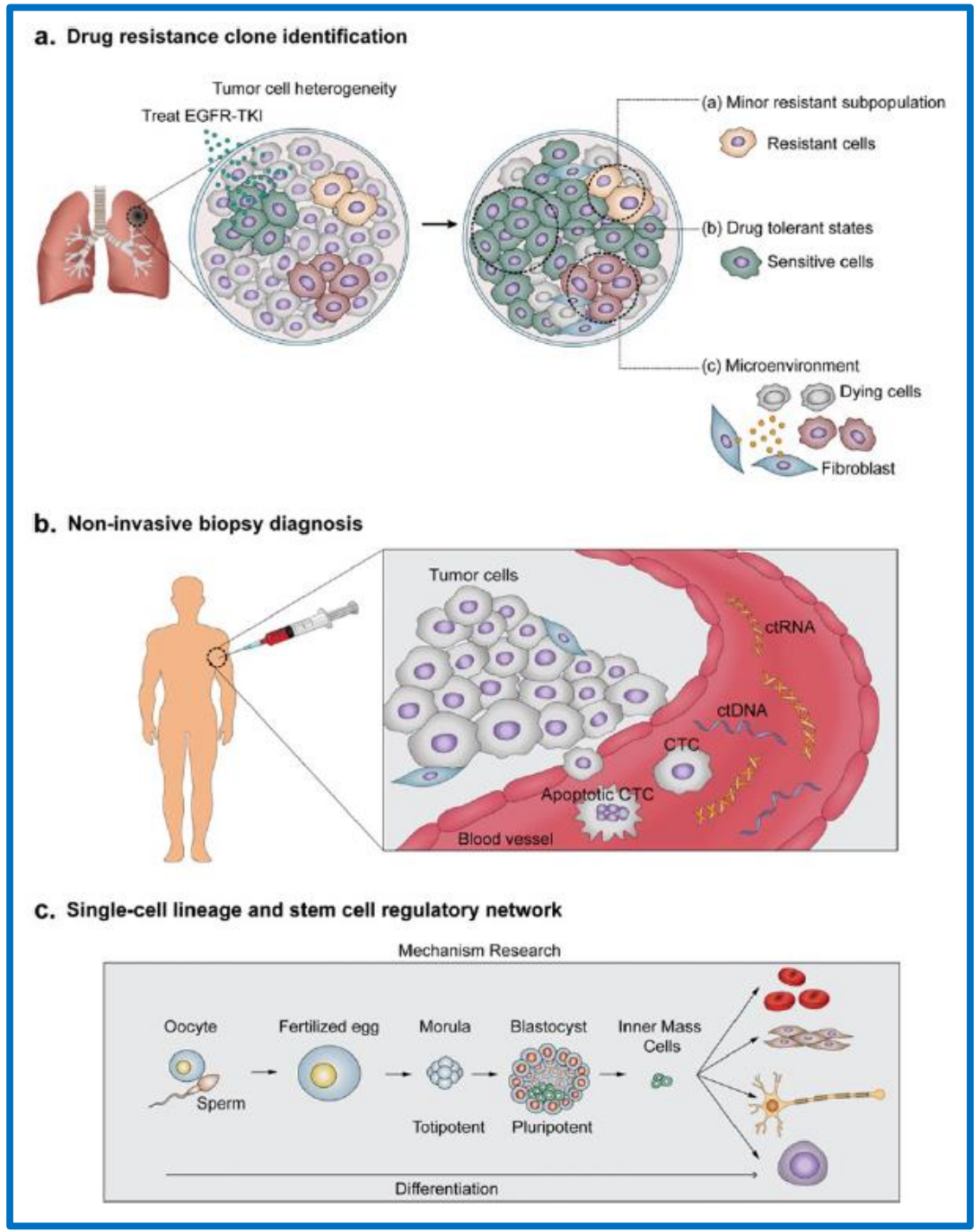


Applications of single-cell sequencing technologies

- Single-cell RNA sequencing has been employed in different species (**humans, animals, plants**) to improve understanding of normal and disease models.
- Many single-cell RNA sequencing (scRNA-seq) methods are focused on understanding (A) development, (B) immunology, (C) diabetes, (D) microbiology, (E) SARS-CoV-2, (F) cancer biology, (G) vascular biology (H) neurobiology and (I) clinical diagnostics.



Jovic D et al., Clin Transl Med. 2022.



Applications of single-cell sequencing technologies

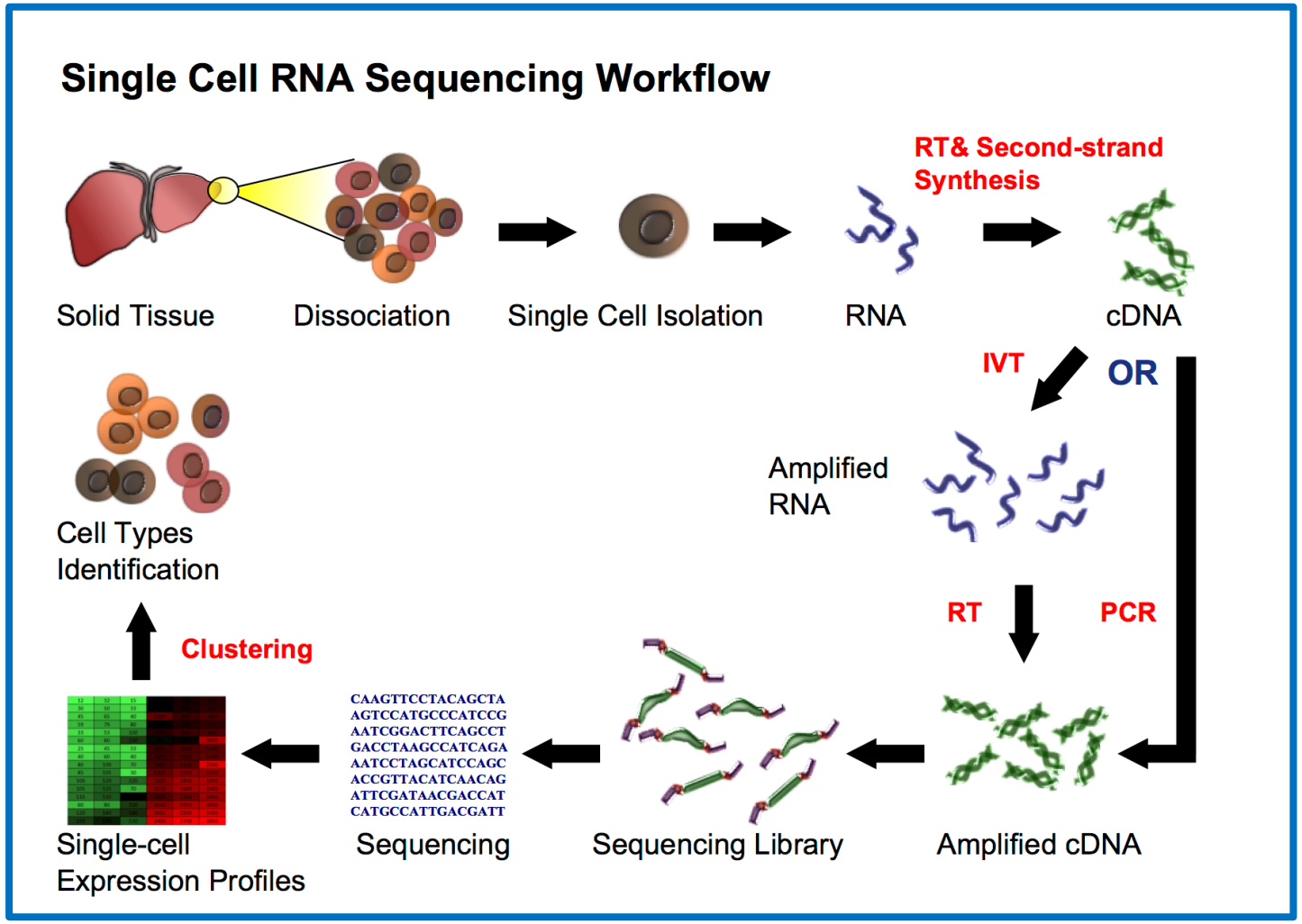
- **A) Intratumor heterogeneity:**
  - Identifying subgroups based on responsiveness in various contexts.
- **B) Liquid biopsy:**
  - Novel insights into biomarker characterization.
- **C) Lineage Information from the early developmental stage:**
  - Novel differential markers.

Hwang B, et al. Exp Mol Med. 2018

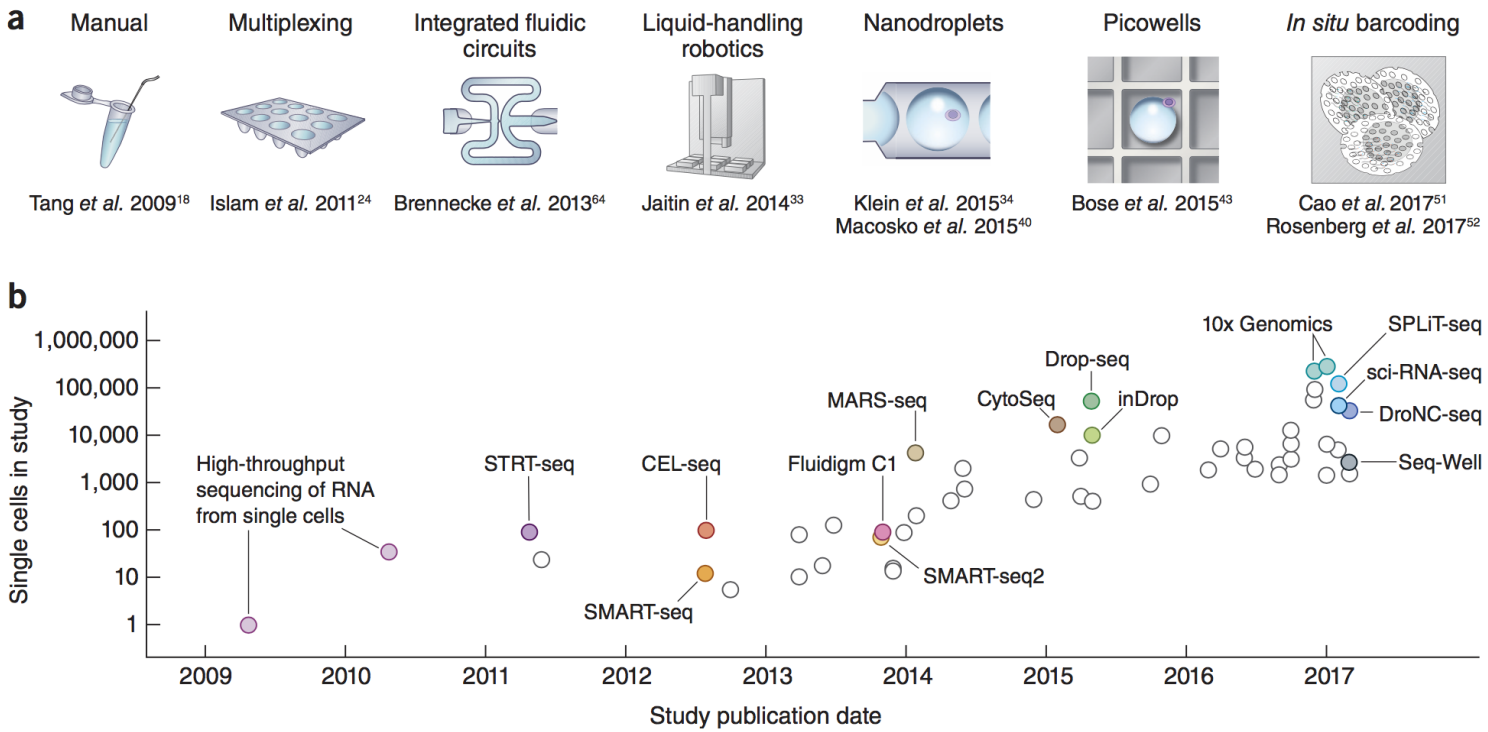
Overview of single cell RNA-seq analysis

- **Main steps of Single-cell RNA-seq:**
  - Tissues/samples dissociation,
  - Partitioning of single cells,
  - Single-cells RNA capture,
  - cDNA synthesis,
  - Library preparation,
  - Sequencing,
  - Data Analysis.

https://ouyanglab.com/singlecell/intro.html



Single-cell sequencing Platforms



10x GENOMICS

Droplet-Based Platforms

Chromium Controller

Chromium Connect

Chromium X

BD

Microwell-based Platform

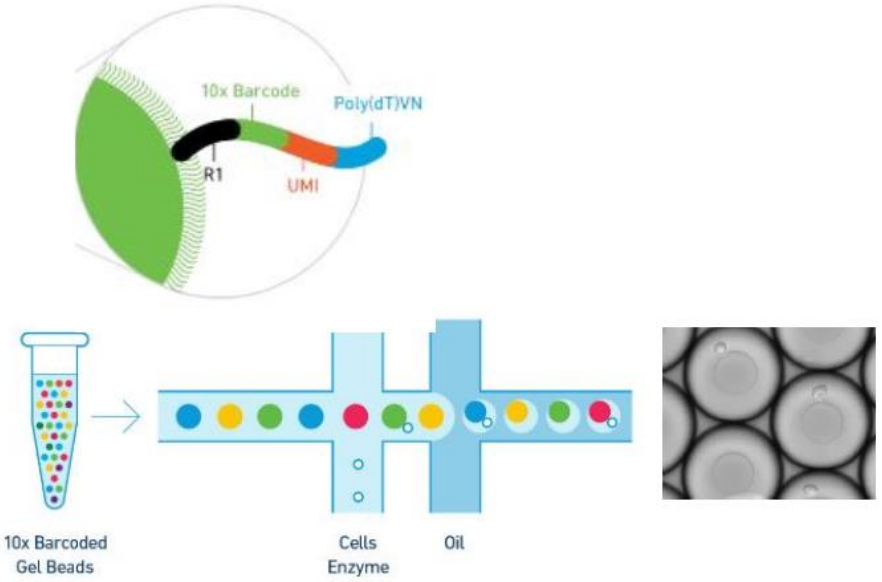
BD Rhapsody

Svennson et al. 2018

Isolation of single cells in droplets or wells with functionalized beads

10x GENOMICS

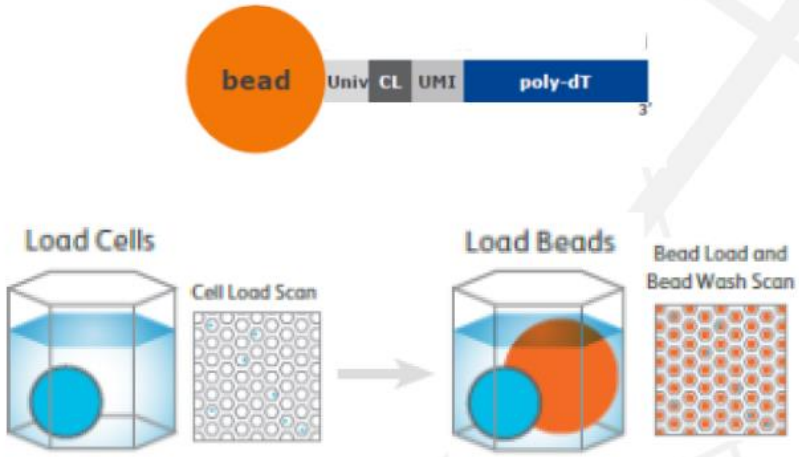
Droplet-Based Platforms



Barcoded Gel Beads mixed with cells, enzymes, oil to form single-cell droplets

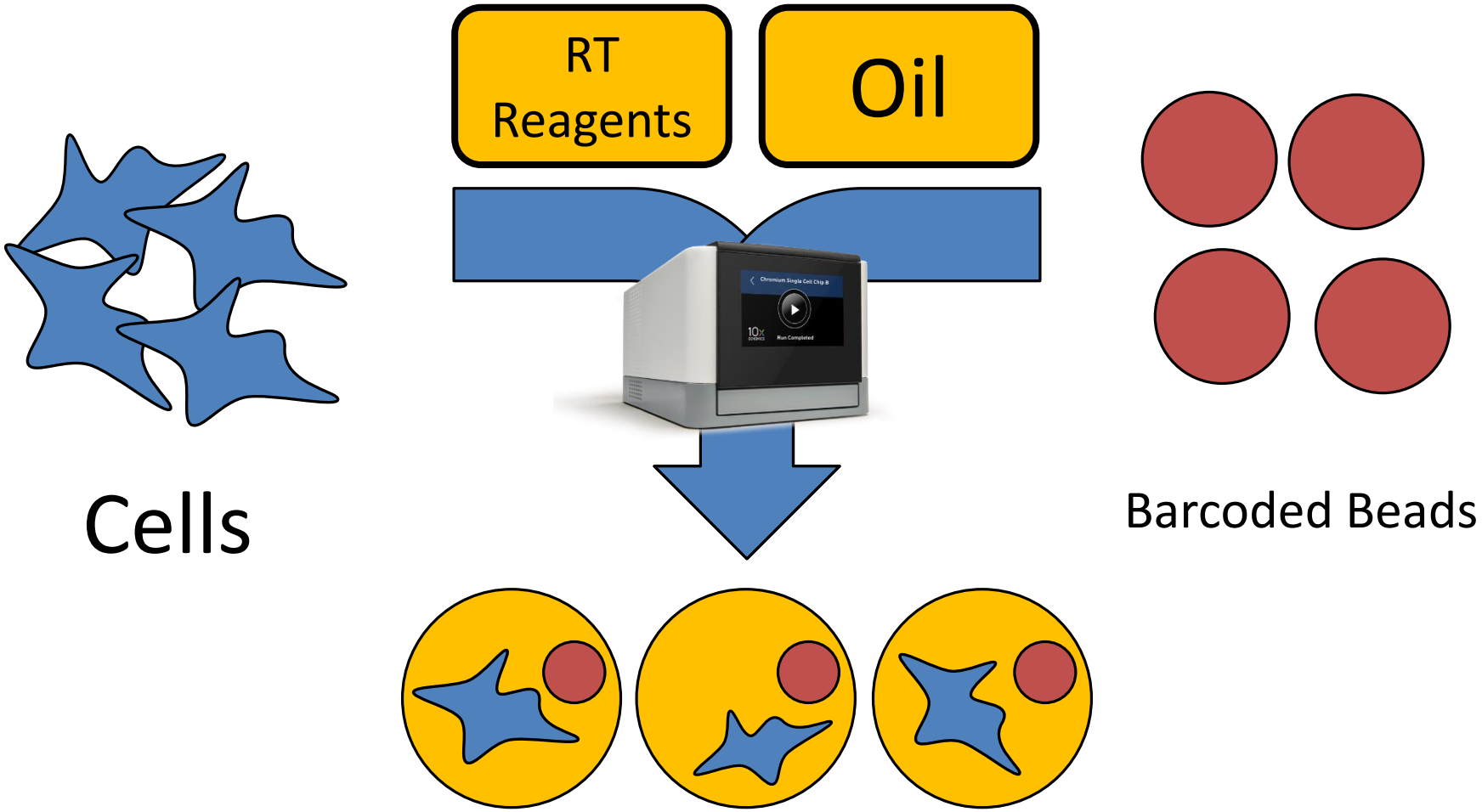
Microwell-based Platform

BD

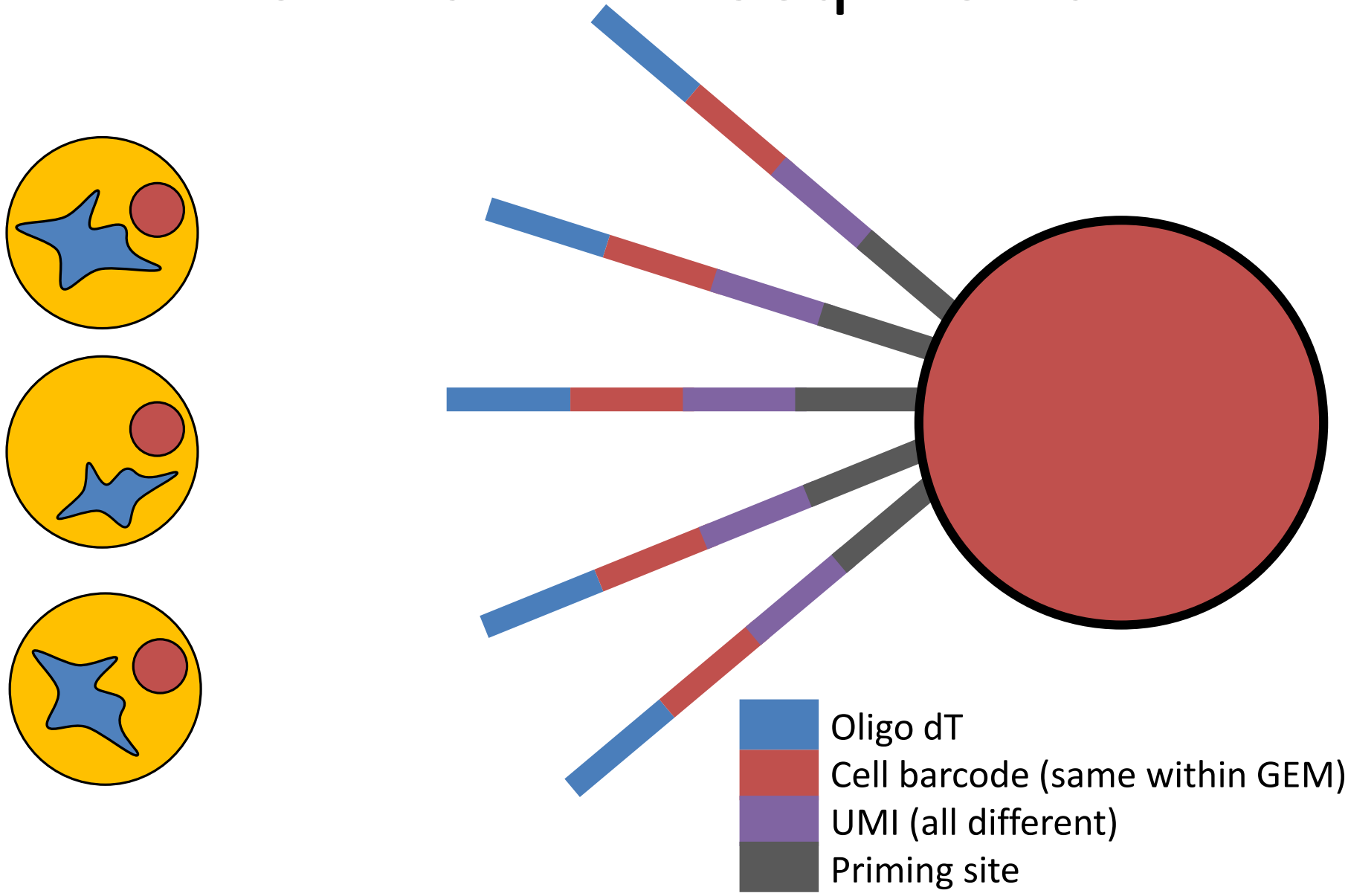


Single cells and single barcoded beads deposited into microwells

How 10X RNA-Seq Works

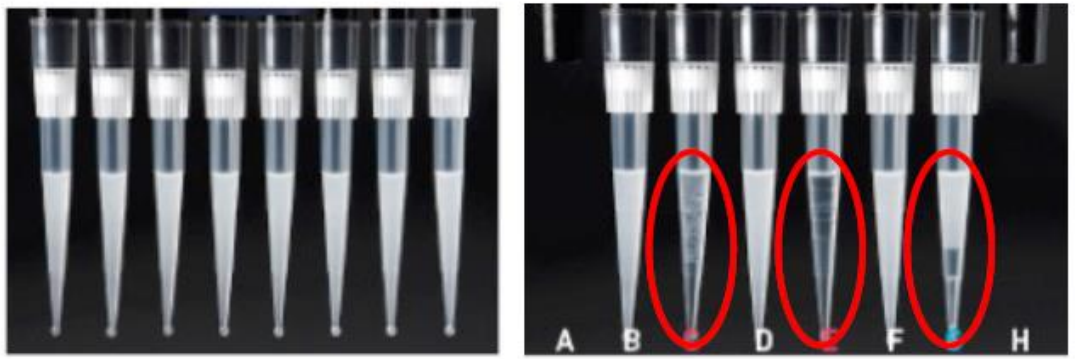


How 10X RNA-Seq Works



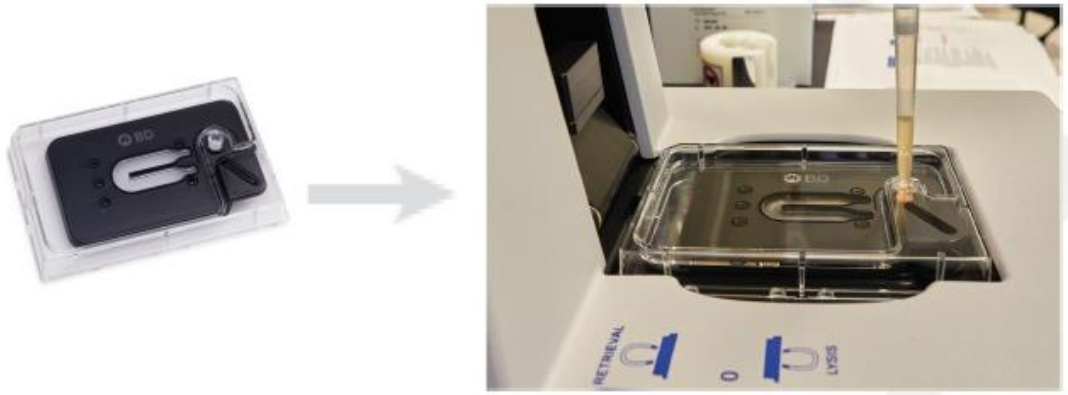
Capture of single cells with functionalized beads

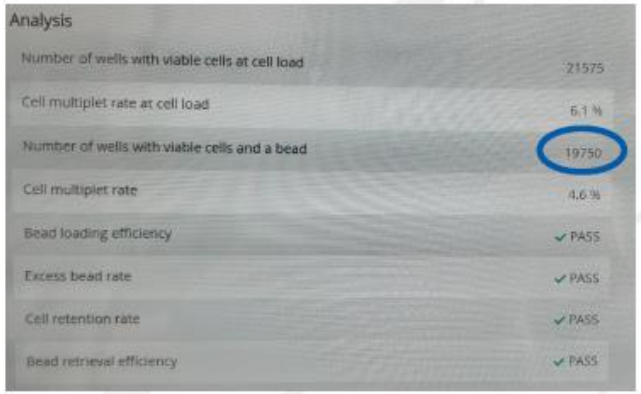
10x GENOMICS



Visual inspection of emulsion

BD



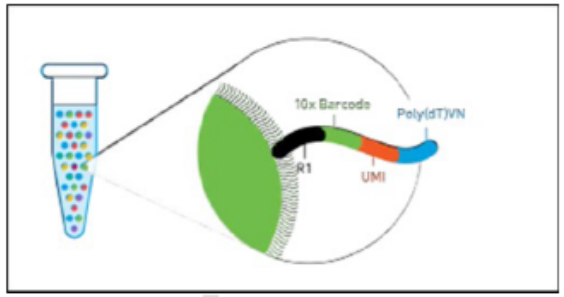


Indication on the number of captured cells

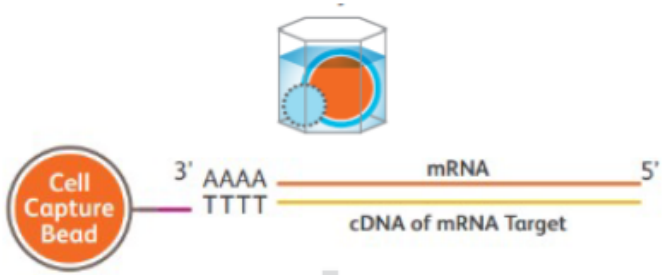
SYNTHESIS OF cDNA

Into a droplet and into a microwell:

1. cells are lysed and release mRNA
2. mRNA is captured by the oligodT of the functionalized beads
3. Retro-Transcription is performed to obtain single-cells cDNAs



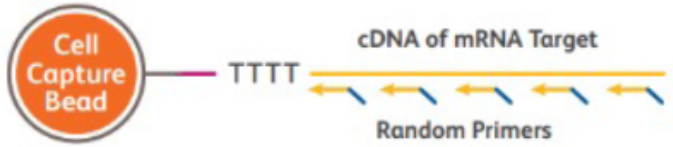
RT and template switch



Beads are magnetically retrieved and pooled

RT and RPE

Random Priming and Extension (RPE)



NGS LIBRARY PREPARATION

10x GENOMICS

BD

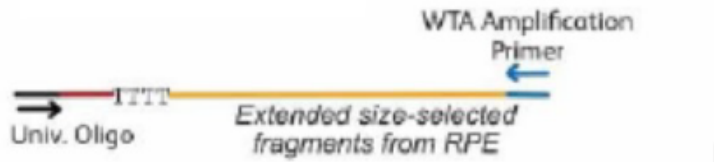
cDNA fragmentation

end repair, A tailing, ligation

index PCR

final library

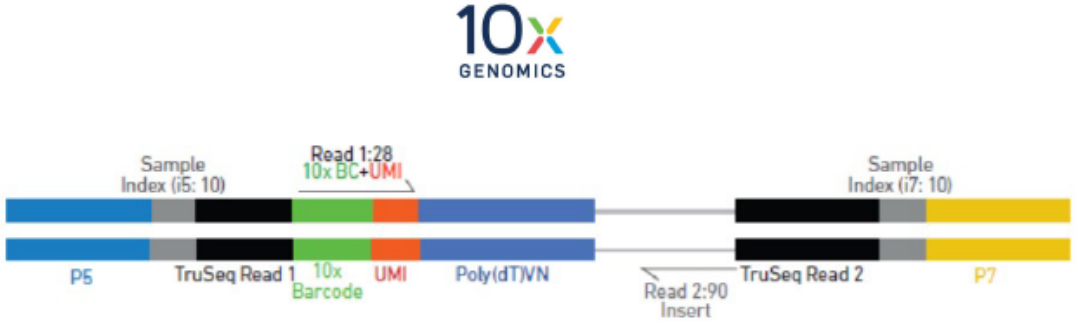
Random Priming and Extension PCR



index PCR

final library

SEQUENCING



- Cell Barcode -> identification of the Cell
- RNA insert -> identification of transcript
- UMI -> transcritpts count and normalization

Nextseq 2000

Novaseq 6000

10x and BD libraries are sequenced paired-ends on high-throughput Illumina sequencers

The sequencing depth should be at least 50’000 reads per cell

Part II

nature reviews genetics
https://doi.org/10.1038/s41576-023-00586-w

Expert recommendation

Check for updates

Best practices for single-cell analysis across modalities

Lukas Heumos[1,2,3,28], Anna C. Schaar[1,4,5,28], Christopher Lance[1,6], Anastasia Litinetskaya[1,4], Felix Drost[1,3], Luke Zappia[1,4], Malte D. Lücken[1,7], Daniel C. Strobl[1,3,8,9], Juan Henao[1], Fabiola Curion[1,4], Single-cell Best Practices Consortium*, Herbert B. Schiller[2] & Fabian J. Theis[1,3,4,5]

nature protocols
REVIEW ARTICLE
https://doi.org/10.1038/s41596-020-00409-w

Check for updates

Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data

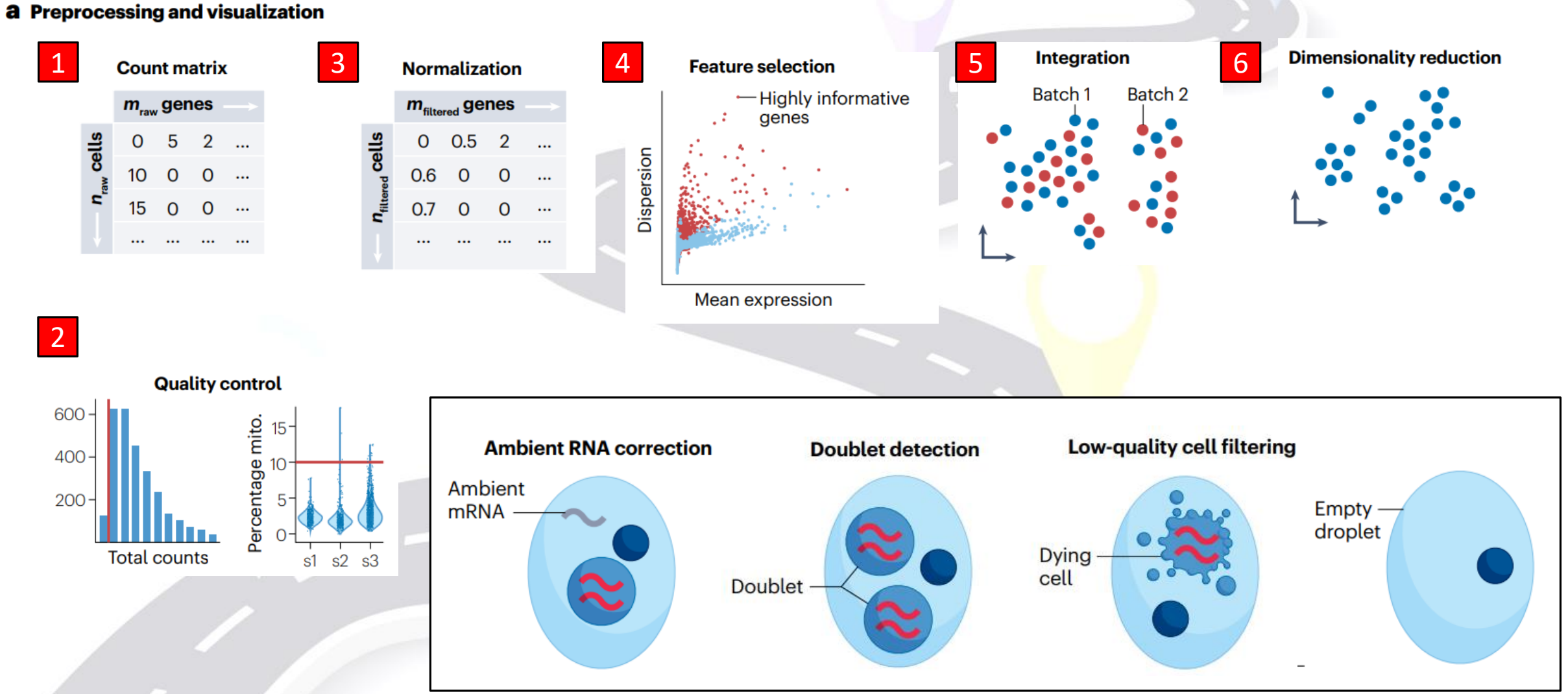
Tallulah S. Andrews[1], Vladimir Yu Kiselev[1], Davis McCarthy[2,3] and Martin Hemberg[1]

Central Component of scRNA-seq analysis

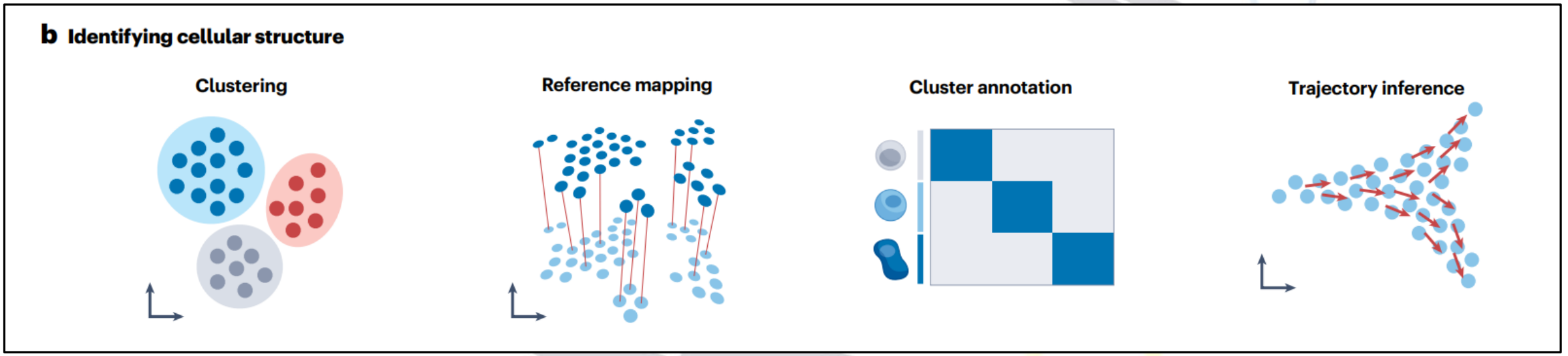
EXPRESSION MATRIX

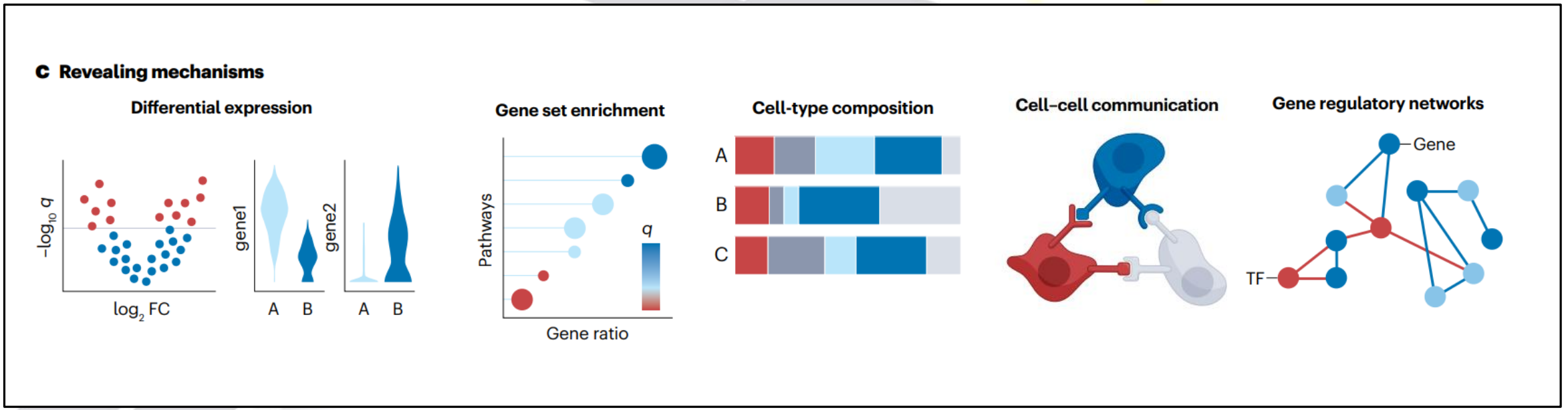
scRNA-seq data sets contain systematic and random noise (such as from poor-quality cells) that obscures the biological signal. Preprocessing of scRNA-seq data attempts to remove these confounding sources of variation.

Roadmap


a Preprocessing and visualization
1
Count matrix
m_{raw} genes
n_{raw} cells
0 5 2 ...
10 0 0 ...
15 0 0 ...
... ...
2
Quality control
600
400
200
Total counts
Percentage mito.
15
10
5
0
s1 s2 s3
3
Normalization
$m_{filtered}$ genes
$n_{filtered}$ cells
0 0.5 2 ...
0.6 0 0 ...
0.7 0 0 ...
... ...
4
Feature selection
Highly informative genes
Dispersion
Mean expression
5
Integration
Batch 1
Batch 2
6
Dimensionality reduction
Ambient RNA correction
Ambient mRNA
Doublet detection
Doublet
Low-quality cell filtering
Dying cell
Empty droplet

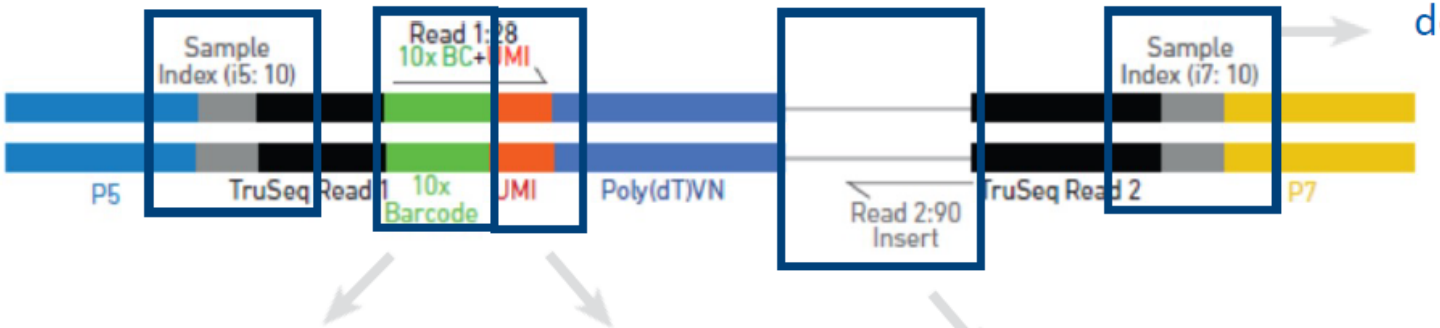
Roadmap


b Identifying cellular structure
Clustering
Reference mapping
Cluster annotation
Trajectory inference


c Revealing mechanisms
Differential expression
−log10 q
log2 FC
gene1
gene2
A
B
A
B
Gene set enrichment
Pathways
q
Gene ratio
Cell-type composition
A
B
C
Cell–cell communication
Gene regulatory networks
Gene
TF

1

PRELIMINARY DATA ANALYSIS



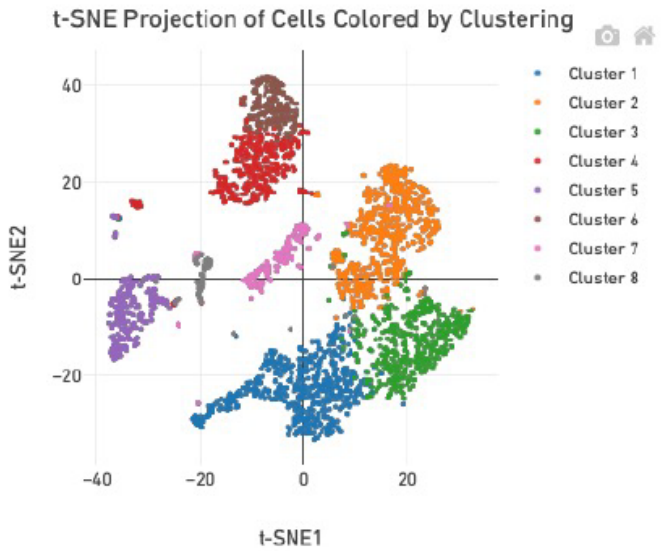
- Pipelines available: Cell Ranger (10x) and Seven Bridges (BD)

- Demultiplexing-alignment-raw count matrix

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

1

PRELIMINARY DATA ANALYSIS



Sequencing Depth	Minimum 20,000 read pairs per cell
Sequencing Type	Paired-end, single indexing
Sequencing Read	Recommended Number of Cycles
Read 1	28 cycles
i7 Index	8 cycles
i5 Index	0 cycles
Read 2	91 cycles

18,312
Estimated Number of Cells

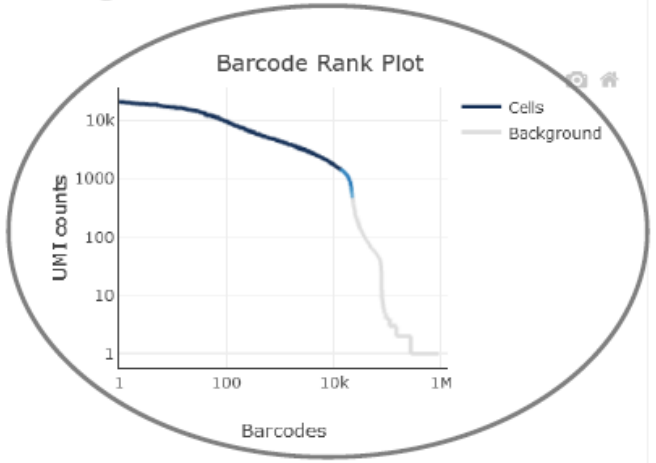
6,293
Mean Reads per Cell

863
Median Genes per Cell

Sequencing

Number of Reads	115,233,874
Number of Short Reads Skipped	0
Valid Barcodes	97.9%
Valid UMIs	99.9%
Sequencing Saturation	24.7%
Q30 Bases in Barcode	96.3%
Q30 Bases in RNA Read	94.4%
Q30 Bases in UMI	96.5%

Cells



Estimated Number of Cells	18,312
Fraction Reads in Cells	78.5%
Mean Reads per Cell	6,293
Median UMI Counts per Cell	1,840
Median Genes per Cell	863
Total Genes Detected	25,840

Web Summary obtained with Cell Ranger

- Check that targeted cells near the estimated number of cells
- Check that the barcode rank plot has a steep drop-off
- Check the fraction reads in cells
- Check the sequencing saturation and median genes per cell

HUMAN TECHNOPOLE

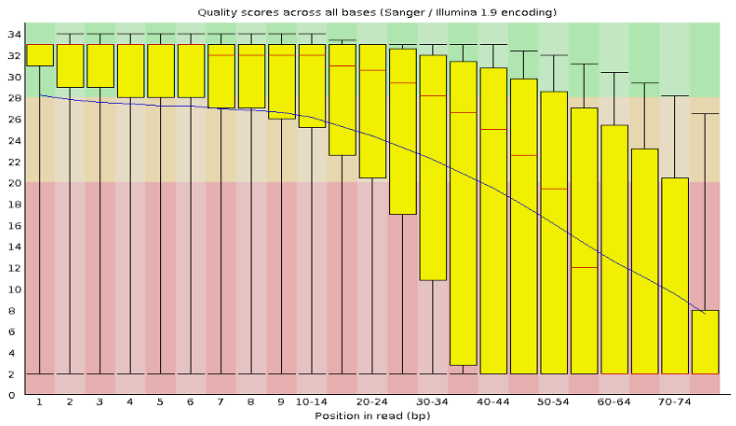
1

How many cells do you have?

- Start by looking at the quality of the base calls in the barcodes
- Bad calls will lead to inaccurate cell assignments

Estimated Number of Cells

15,894

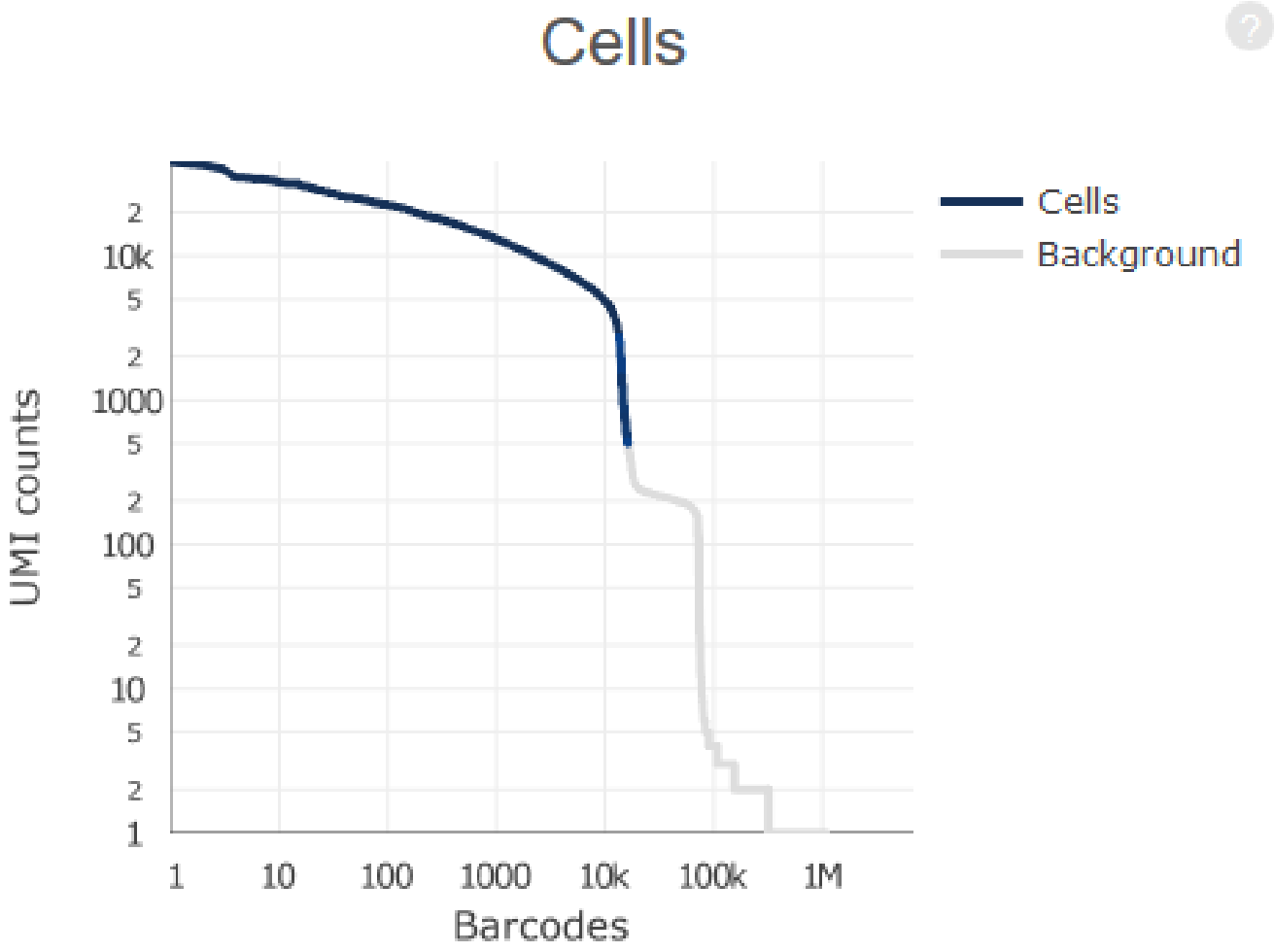


Sequencing

Number of Reads	180,878,636
Valid Barcodes	98.1%
Sequencing Saturation	10.3%
Q30 Bases in Barcode	98.4%
Q30 Bases in RNA Read	82.7%
Q30 Bases in UMI	98.7%

1

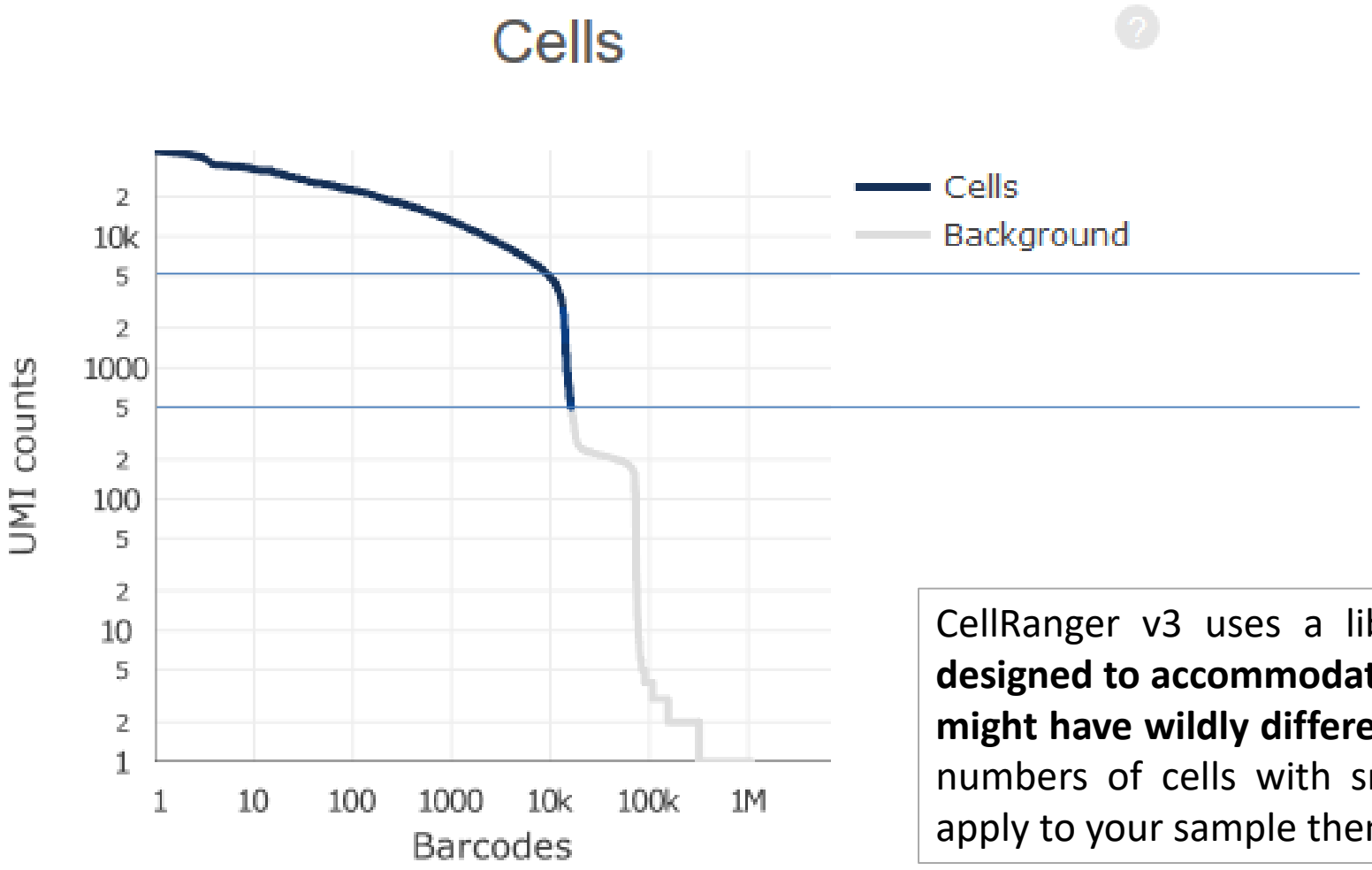
How many cells do you have?



- Plot of UMIs (reads) per cell vs number of cells
- Blue region was called as valid cells
- Grey region is considered noise
- Both axes are log scale!!!

1

How many cells do you have?



5000 reads per cell. 10k cells

500 reads per cell. 15k cells

CellRanger v3 uses a liberal cutoff to define cells. **This was designed to accommodate (normally cancer) samples where cells might have wildly different amounts of RNA.** It will include large numbers of cells with small numbers of UMIs. If this doesn't apply to your sample then this **will over-predict** valid cells.

Major scRNA Package Systems

https://satijalab.org/seurat/

https://cole-trapnell-lab.github.io/monocle3/

Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R

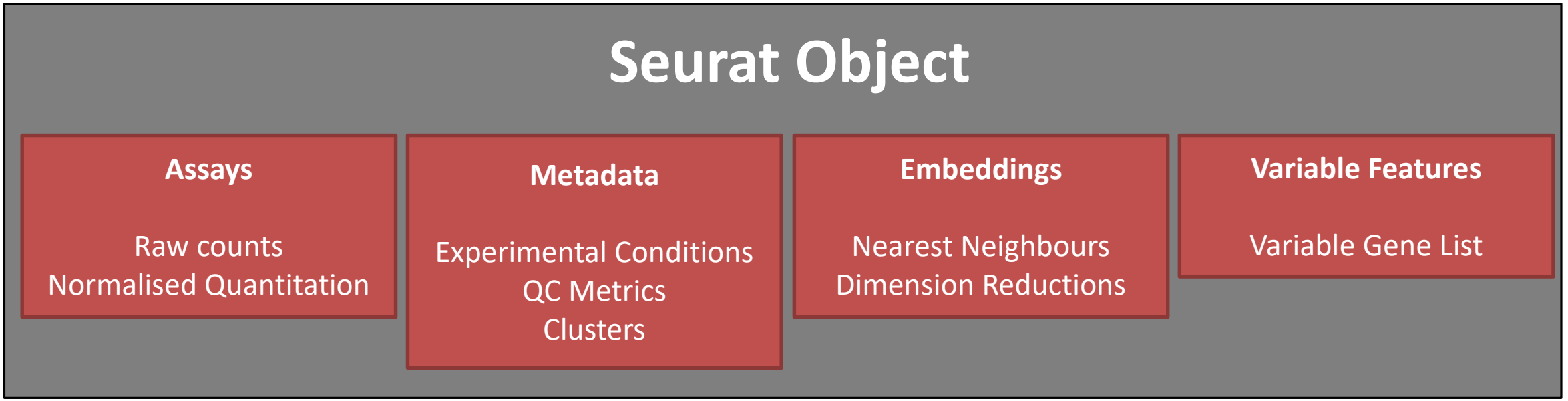
Davis J McCarthy ✉, Kieran R Campbell, Aaron T L Lun, Quin F Wills

https://bioconductor.org/packages/release/bioc/html/scater.html

https://scanpy.readthedocs.io/en/stable/

Seurat Data Structure

- Single object holds all data
  - Build from text table or 10X output (feature matrix h5 or raw matrix)



Seurat Methods

- Data Parsing
  - `Read10X`
  - `Read10X_h5*`
  - `CreateSeuratObject`
- Data Normalisation
  - `NormalizeData`
  - `ScaleData`
- Graphics
  - Violin Plot – metadata or expression (`VlnPlot`)
  - Feature plot (`FeatureScatter`)
  - Projection Plot (`DimPlot, DimHeatmap`)
- Dimension reduction
  - `RunPCA`
  - `RunTSNE`
  - `RunUMAP`
- Statistics
  - Select Variable Genes `FindVariableFeatures`
  - Build nearest neighbour graph `FindNeighbors`
  - Build graph based cell clusters `FindClusters`
  - Find genes to classify clusters (multiple tests) `FindMarkers`

*Requires installing the `hdf5r` package

Reading Data

```
Read10X("../filtered_feature_bc_matrix/") -> data

Read10X_h5("raw_feature_bc_matrix.h5") -> data

CreateSeuratObject(
      counts=data,
      project="course",
      min.cells = 3,
      min.features=200
) -> data
```

Reading Data

Barcodes.tsv

cell_ids
AAACATACAACCAC-1
AAACATTGAGCTAC-1
AAACATTGATCAGC-1
AAACCGTGCTTCCG-1
AAACCGTGTATGCG-1
AAACGCACTGGTAC-1
AAACGCTGACCAGT-1
AAACGCTGGTTCTT-1
AAACGCTGTAGCCA-1
AAACGCTGTTTCTG-1
AAACTTGAAAAACG-1

Features.tsv

ENSG00000243485	MIR1302-10
ENSG00000237613	FAM138A
ENSG00000186092	OR4F5
ENSG00000238009	RP11-34P13.7
ENSG00000239945	RP11-34P13.8
ENSG00000237683	AL627309.1
ENSG00000239906	RP11-34P13.14
ENSG00000241599	RP11-34P13.9
ENSG00000228463	AP006222.2
ENSG00000237094	RP4-669L17.10
ENSG00000235249	OR4F29
ENSG00000236601	RP4-669L17.2

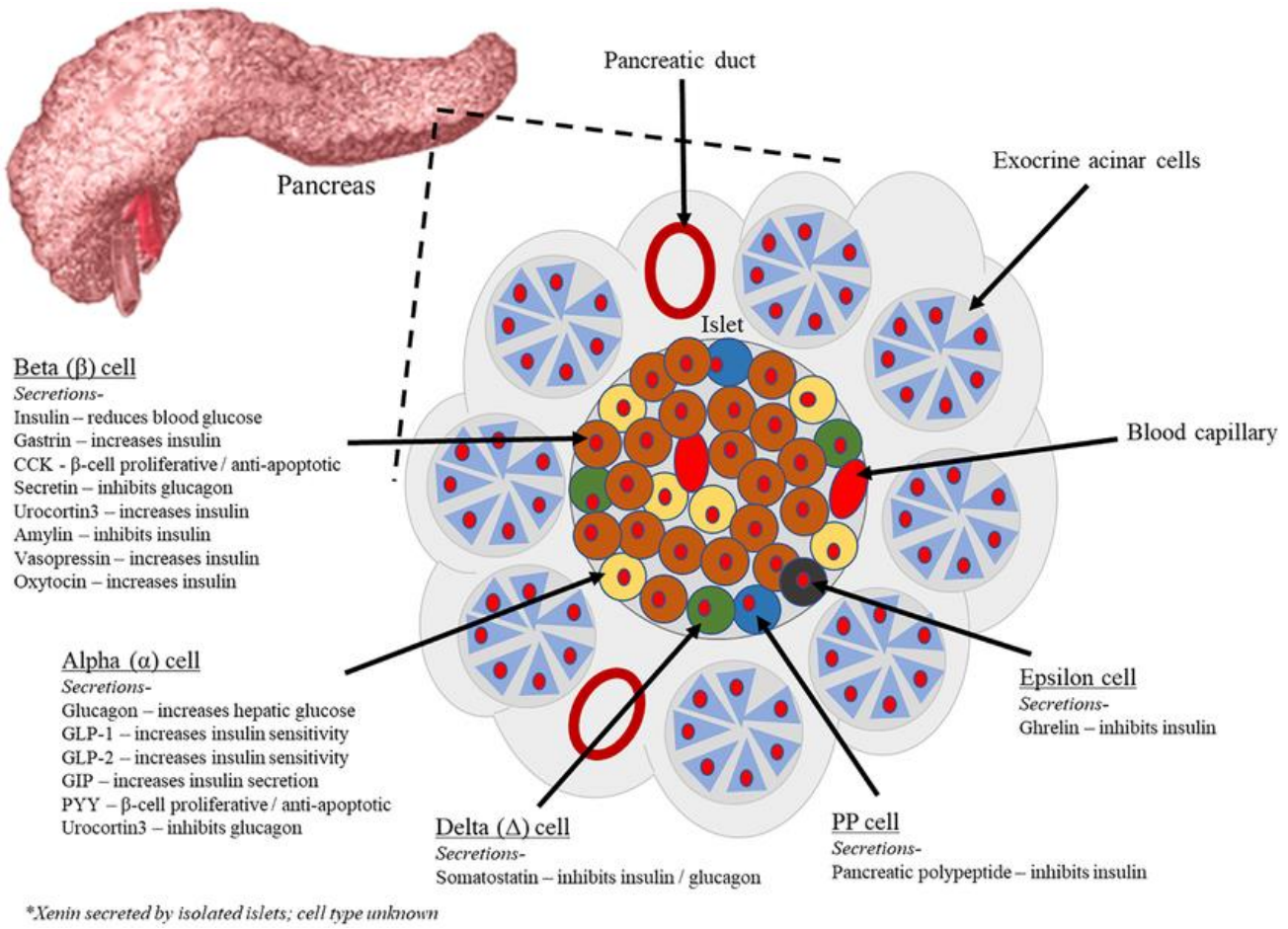
Matrix.mtx

0	0
0	0
2	0
0	0
0	0
0	3
0	2
40	0
0	05
0	2
0	0

- `expression_matrix <- ReadMtx( mtx = "GSM6619605_matrix.mtx.gz", features = "GSM6619605_features.tsv.gz", cells = "GSM6619605_barcodes.tsv.gz")seurat_object <-`
- `CreateSeuratObject(counts = expression_matrix)`

Dataset panc8

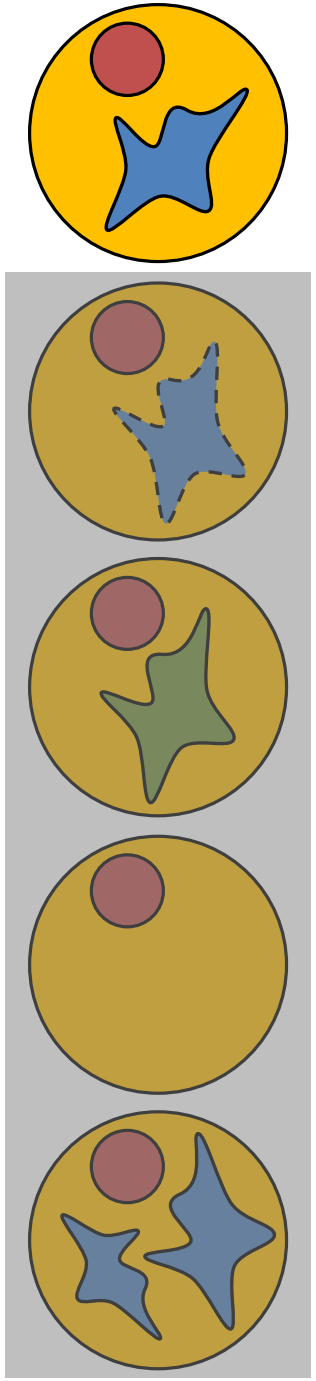
- **Human pancreatic islet cell** datasets produced across four technologies, CelSeq (GSE81076) CelSeq2 (GSE85241), Fluidigm C1 (GSE86469), and SMART-Seq2 (E-MTAB-5061).



Seurat Metadata

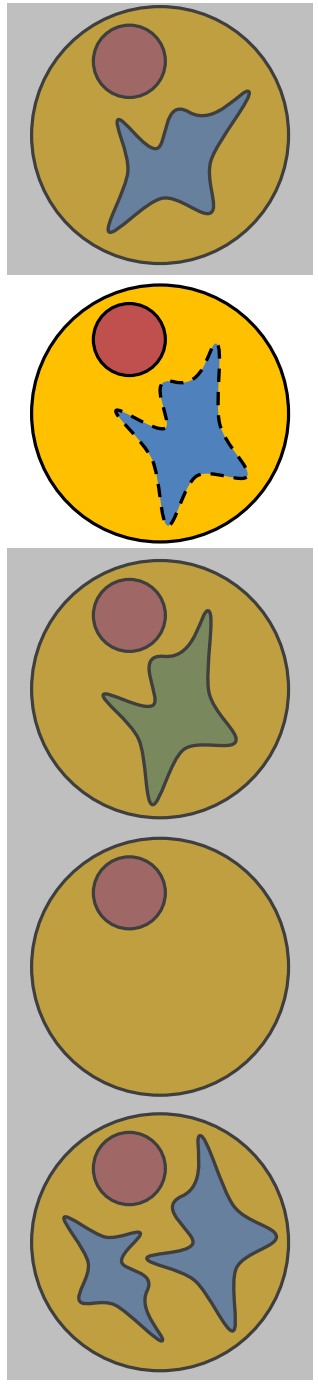
```
> head(panc8[[]])
 orig.ident nCount_RNA nFeature_RNA   tech replicate
D101_5         D101   4615.810         1986 celseq    celseq
D101_7         D101  29001.563         4209 celseq    celseq
D101_10        D101   6707.857         2408 celseq    celseq
D101_13        D101   8797.224         2964 celseq    celseq
D101_14        D101   5032.558         2264 celseq    celseq
D101_17        D101  13474.866         3982 celseq    celseq

> table(panc8$tech)
 celseq    celseq2 fluidigmc1     indrop  smartseq2
      1004       2285        638       8569       2394
```

QC – What problems are likely?



- Lysed cells
- Dead or dying cells
- Empty GEMs
- Double (or more) occupied GEMs
- Cells in different cell cycle stages

Lysed Cells

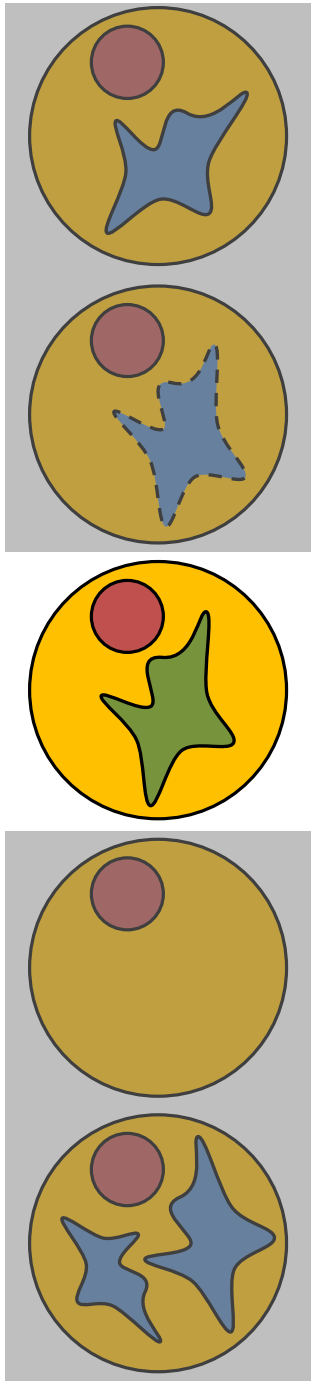


- Outer membrane is ruptured – cytoplasmic RNAs leak out
  - Loss of mature RNA, increase in pre-mRNA
    - Higher proportion mapping to introns
    - Loss of 3’ sequencing bias
  - Increase in nuclear RNAs
    - *MALAT1* is an easy marker to use
  - Increase in Membrane associated transcripts
    - MS4A family
    - IL7R
    - Complement C3

Dead or Dying Cells

- Cells undergoing apoptosis have very different transcriptomes
  - Lower total RNA production
  - **Huge upregulation of mitochondrial transcription**
  - Upregulation of caspases

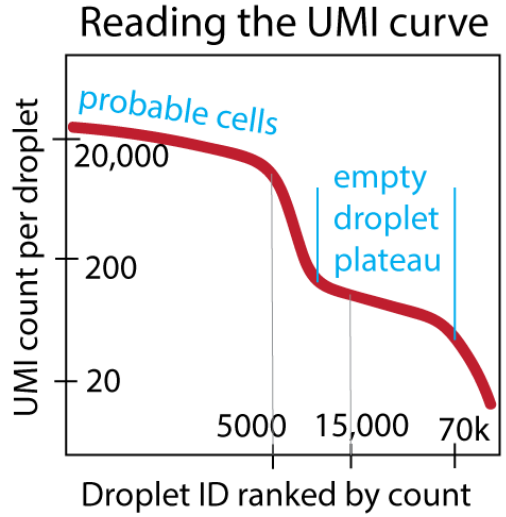
- Degraded transcripts are short
  - Read through into template switching oligo (seen earlier)



Empty GEMs and Ambient RNA

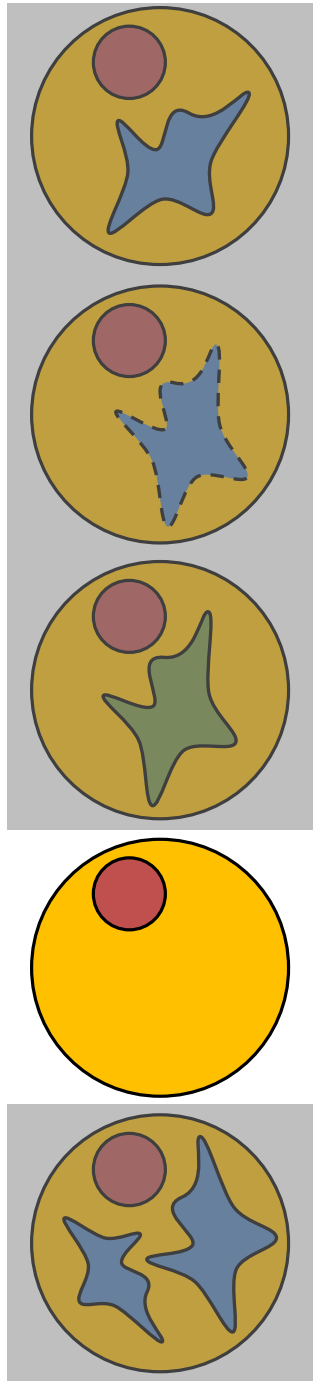
- GEMs containing no cell will still produce some sequence
  - Background RNA in the flow medium
  - Will be worse with higher numbers of lysed cells

- Cellbender: software package for eliminating technical artifacts.



Cellbender

Ambient RNA

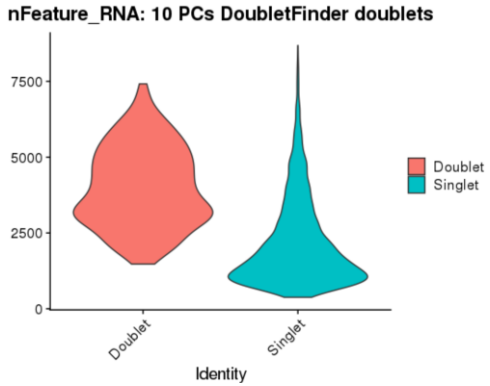


Double occupied GEMs

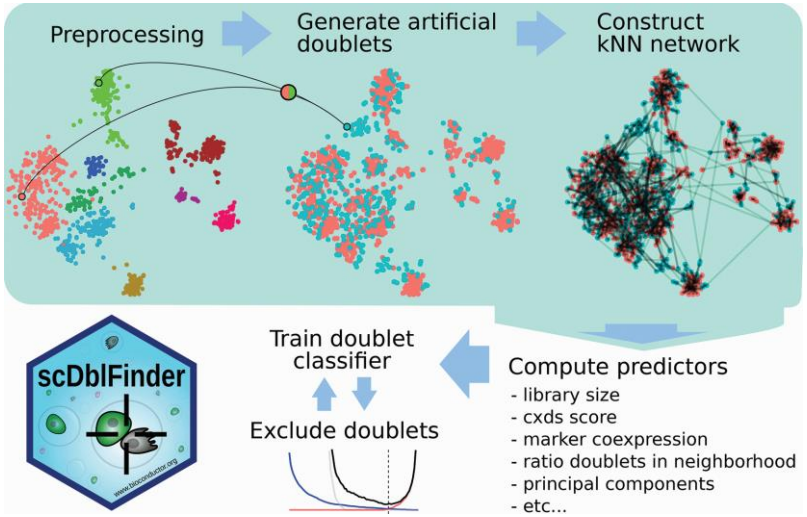
- Will get a mixed signal from two different cells.

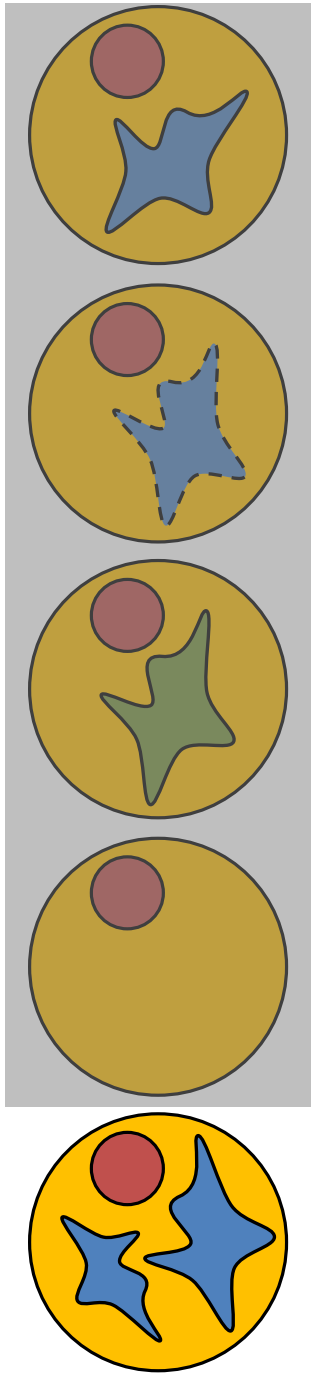
- Not as obvious a signal as empty GEMs
  - Greater diversity
  - More UMIs per cell
  - Intermediate clustering

DoubletFinder



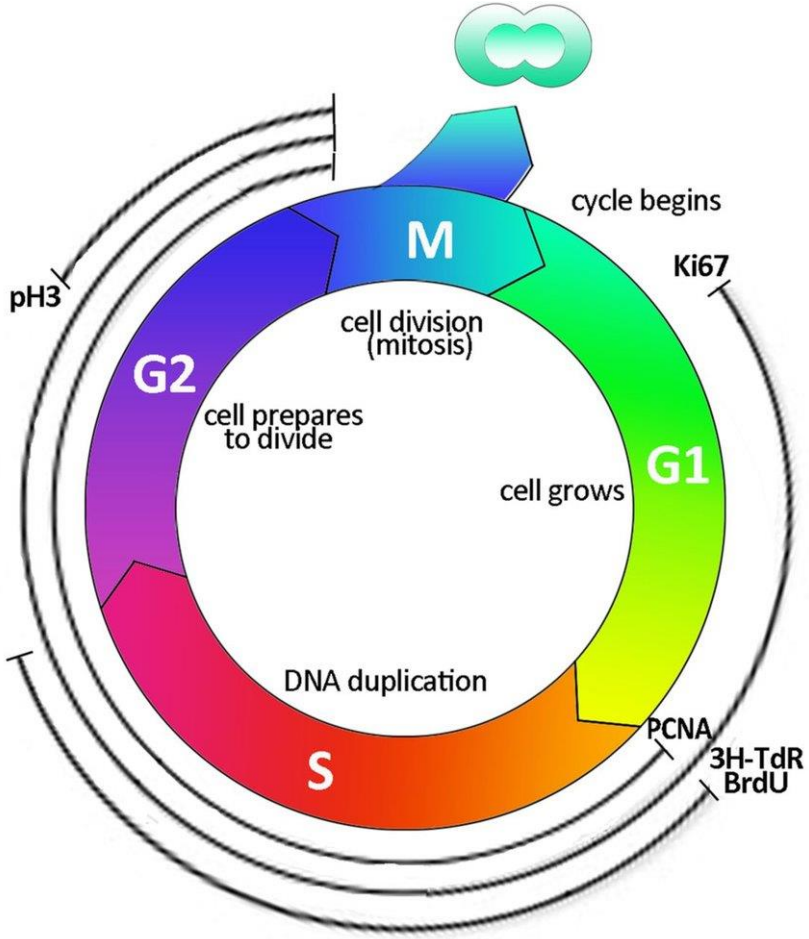
scDblFinder





Cell Cycle Variation

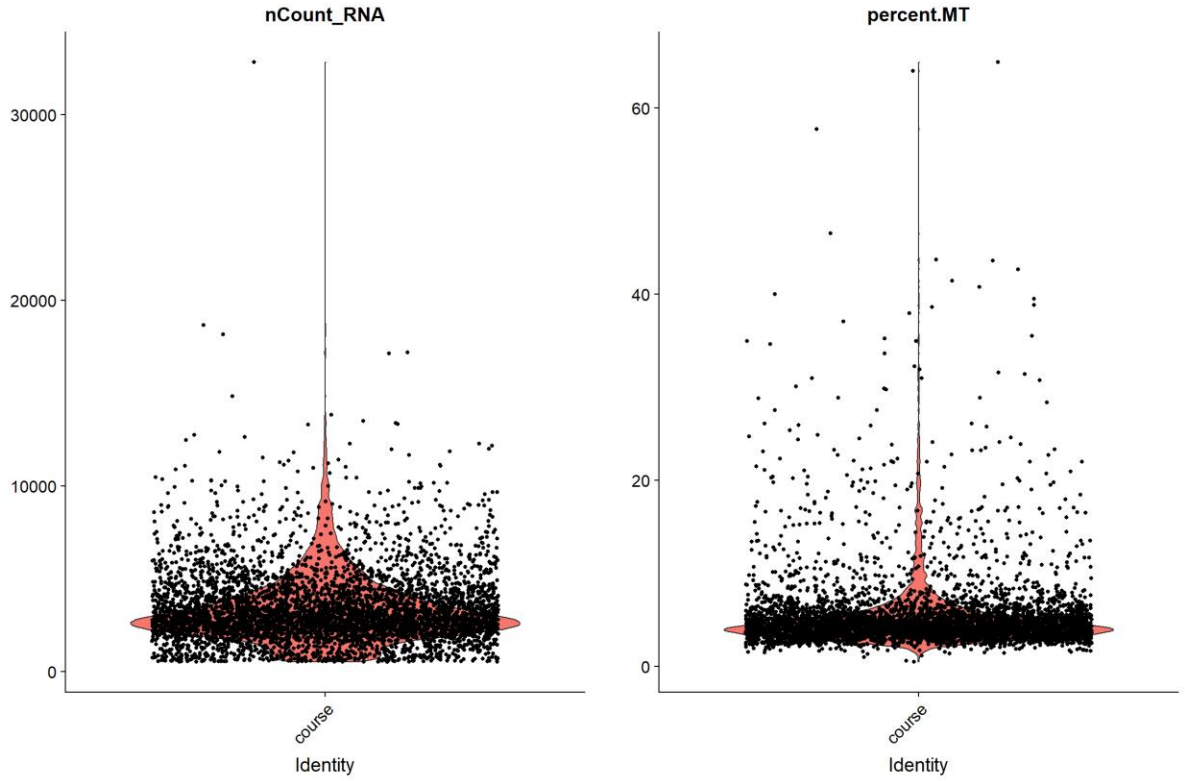
- Cells in different stages of the cell cycle have quite different expression profiles
  - Use genes which classify different phases to classify cells in different phases
  - Exclude unusual cells
  - Attempt to include cell cycle as a factor during quantitation / differential expression



Lombardi, et al. (2015). 10.1093/ndt/gfv262.

QC and Cell Filtering

- Standard QC Measures
  - Number of observed genes per cell
  - Number of reads per cell
  - Relationship between the two
- Calculated QC Measures
  - Amount of mitochondrial reads
  - Amount of ribosomal reads
  - Marker genes (eg *MALAT1*)
  - Cell cycle

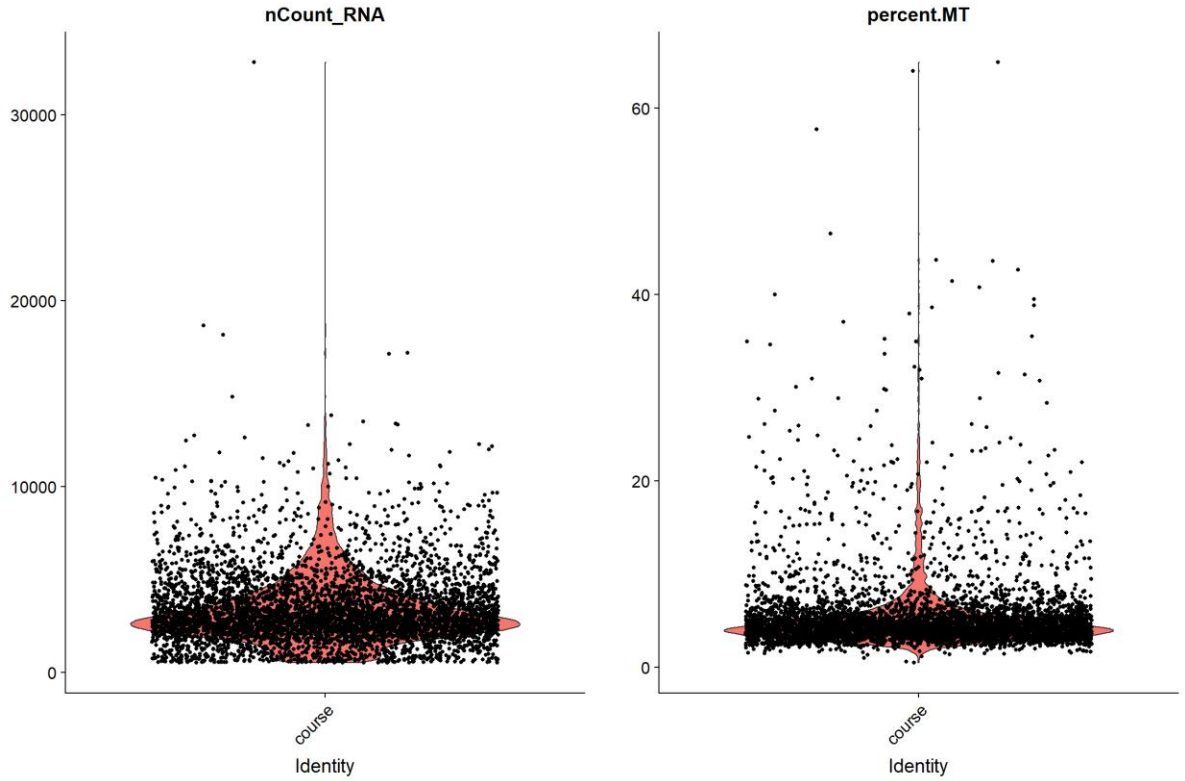


2

QC and Cell Filtering

```
PercentageFeatureSet(
      data,
      pattern="^MT-“
) -> data$percent.MT
```

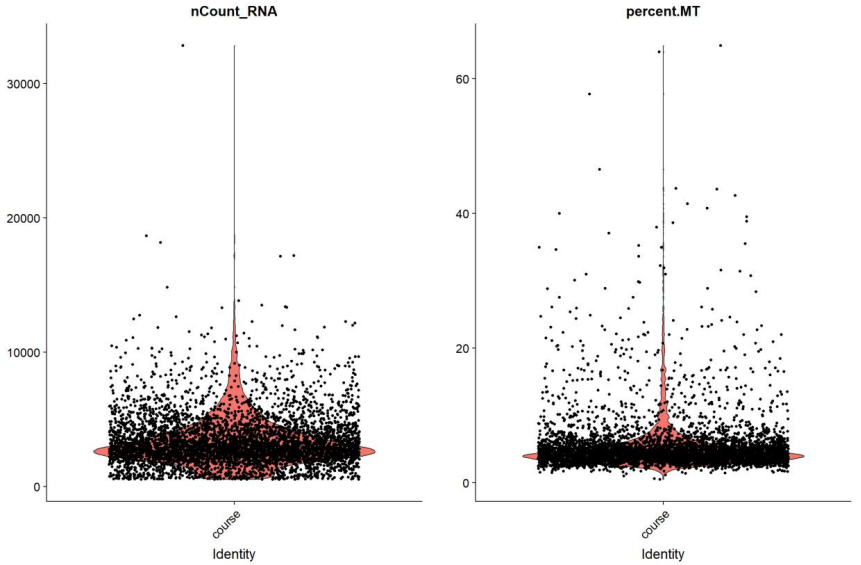
```
apply(
  data@assays$RNA@counts,
  2,
  function(x)(100*max(x))/sum(x)
) -> data$Percent.Largest.Gene
```



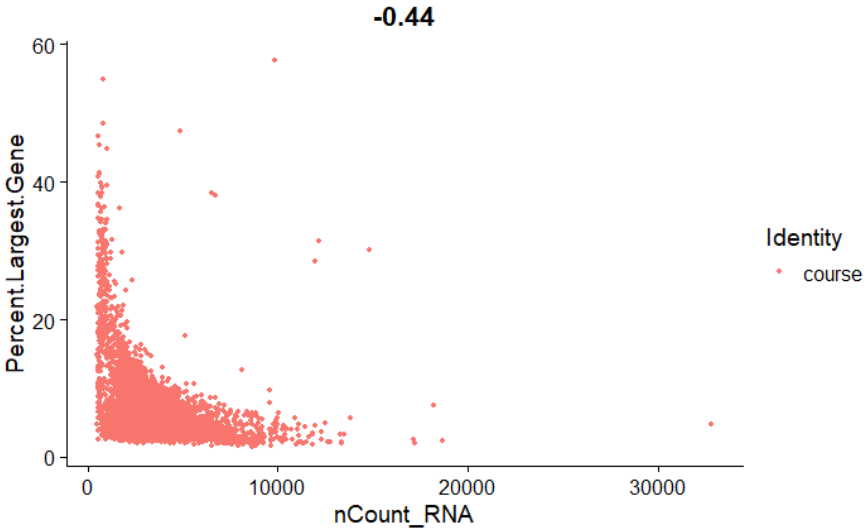
2

QC and Cell Filtering

```
VlnPlot(
        data,
        features=c("nCount_RNA","percent.MT")
)
```



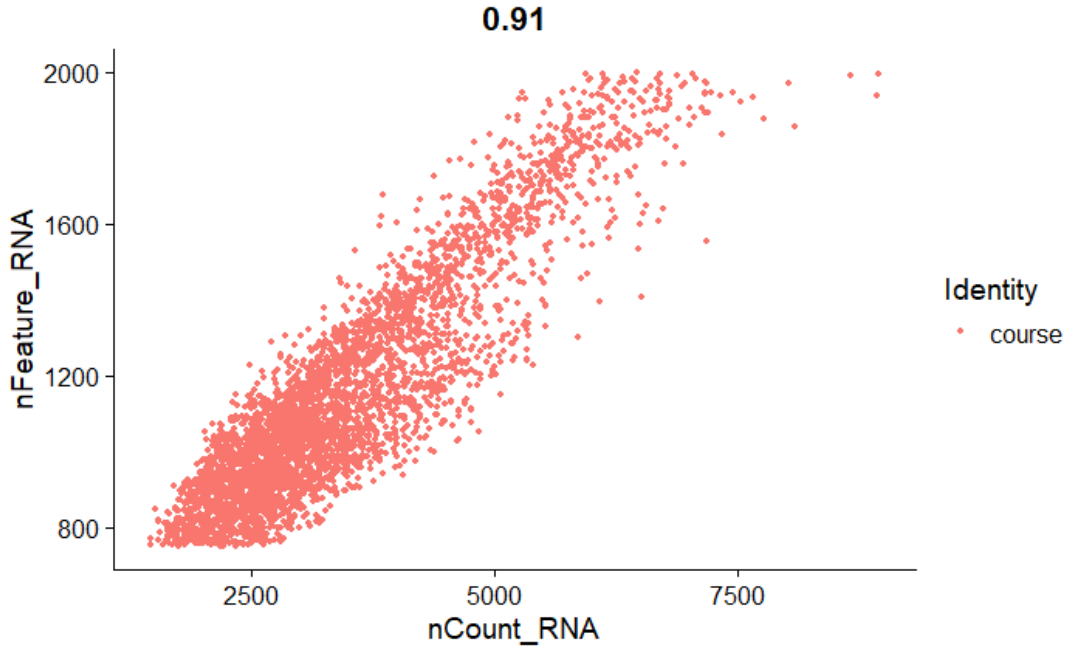

```
FeatureScatter(
        data,
        feature1 = "nCount_RNA",
        feature2 = "Percent.Largest.Gene“
)
```



2

QC and Cell Filtering

```
subset(
  data,
  nFeature_RNA>750 &
    nFeature_RNA < 2000 &
    percent.MT < 10 &
    Percent.Largest.Gene < 20
) -> data
```



Count Normalisation and Scaling

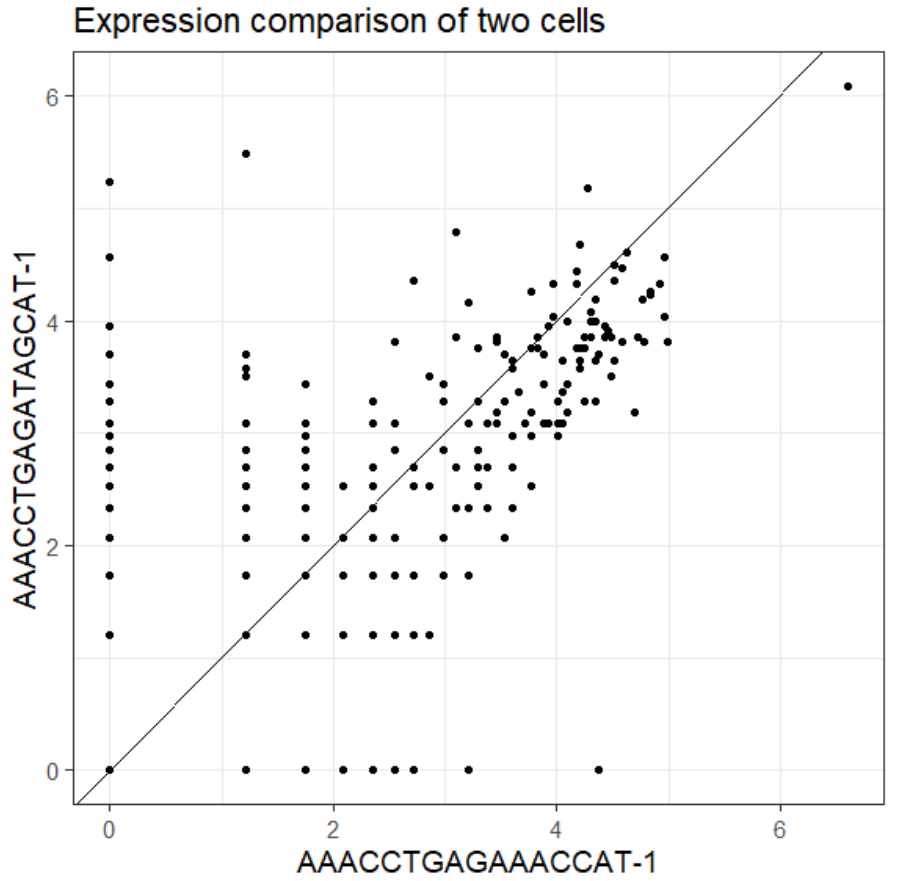
- Raw counts are biased by total reads per cell
- Counts are more stable on a log scale
- Standard normalisation is just log reads per 10,000 reads
- Can use an additional centring step which may help
  - Similar to size factor normalisation in conventional RNA-Seq

- For PCA counts scale each gene's expression to a z-score
  - Can also use this step to try to regress out unwanted effects

3

Count Normalisation and Scaling

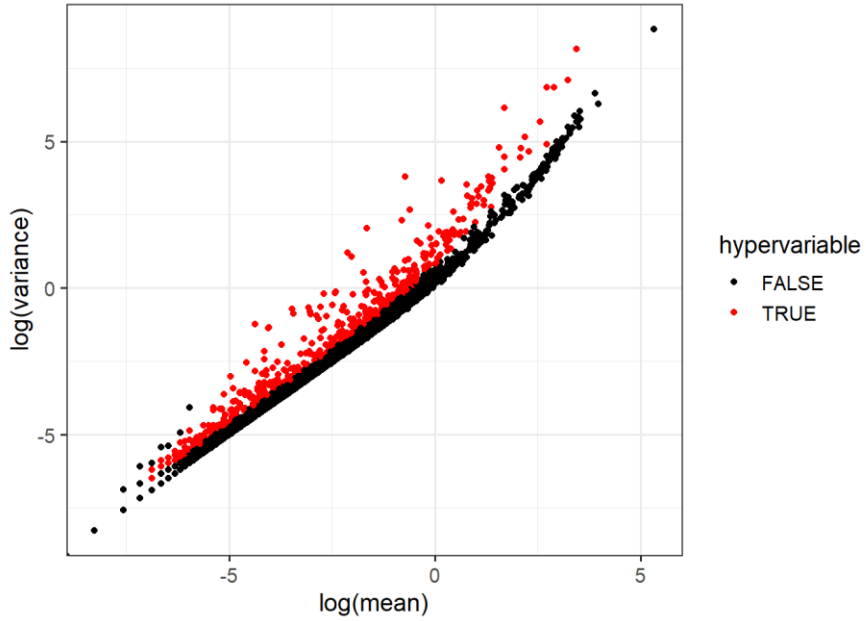
```
NormalizeData(
    data,
    normalization.method = "CLR"
) -> data


ScaleData(
    data,
    features=rownames(data)
) -> data
```



Variable Feature Selection

- Selects a subset of genes to use for downstream analysis
- Identify genes with an unusual amount of variability
- Link the variability with the expression level to find variation which is high in the context of the expression level
- Keep only the most variable genes

```
FindVariableFeatures(
  data,
  selection.method = "vst",
  nfeatures=500
) -> data
```



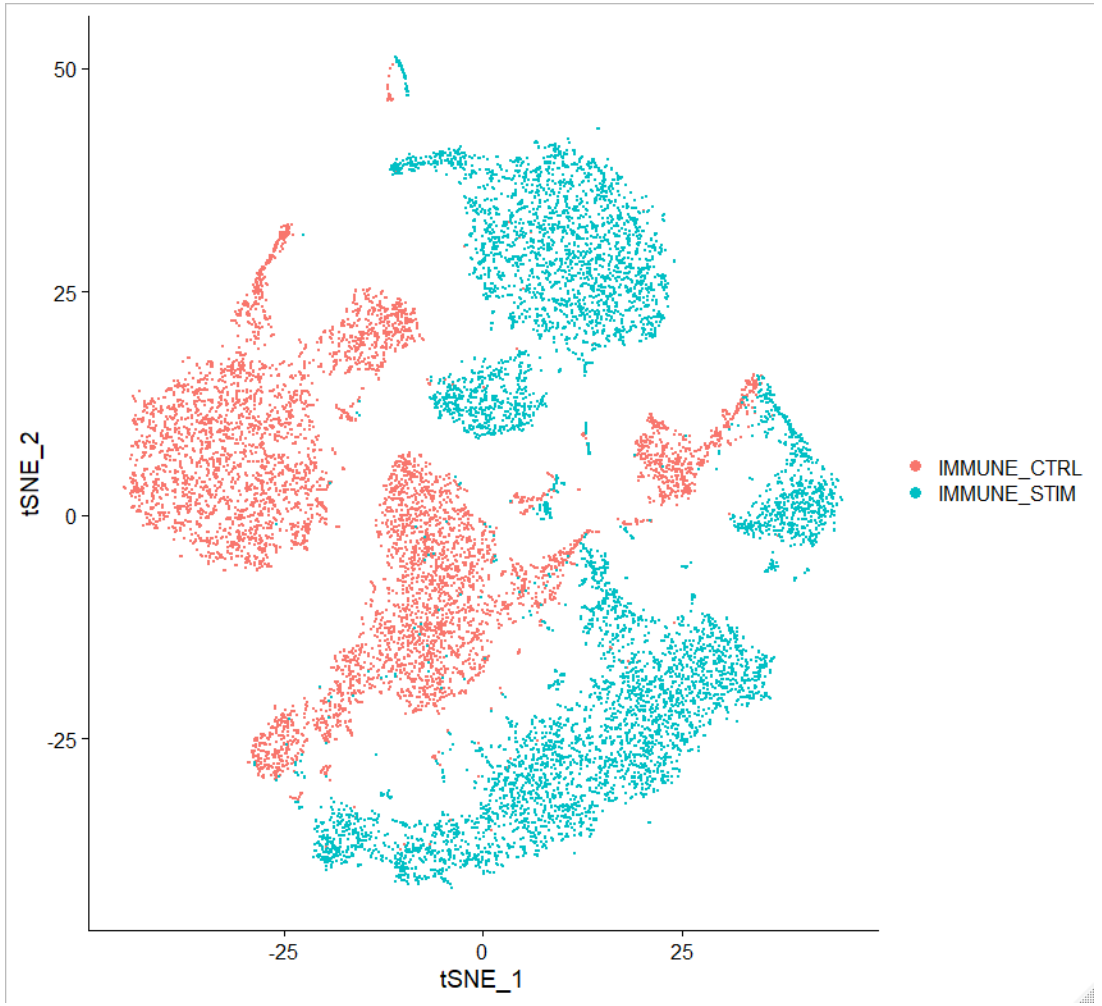
Integrating Multiple Runs

- When multiple runs are combined (eg Unstim and Stim), the batch differences between the runs can overwhelm the biological differences

- Raw comparisons can therefore miss changes between what are actually matched subgroups

Raw merged runs

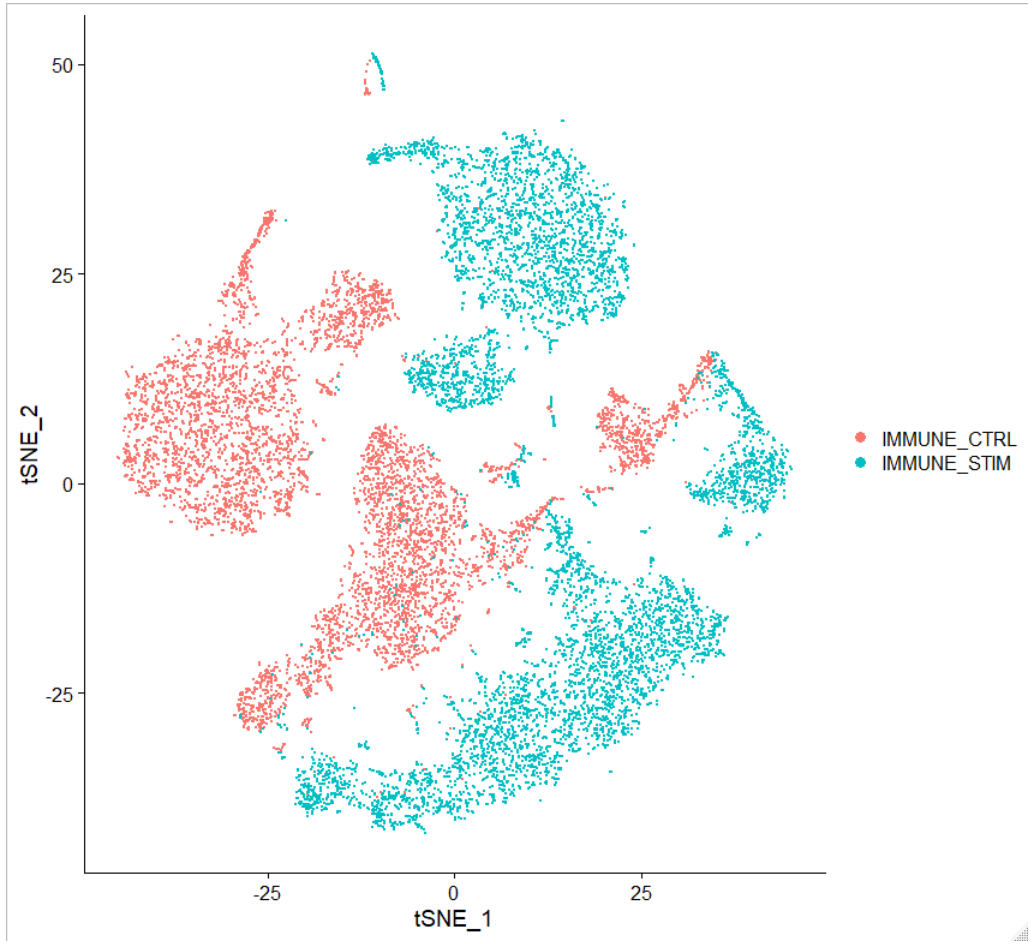
- Two PBMC populations run at different times
- tSNE spread coloured by library
- Little to no overlap between cell populations



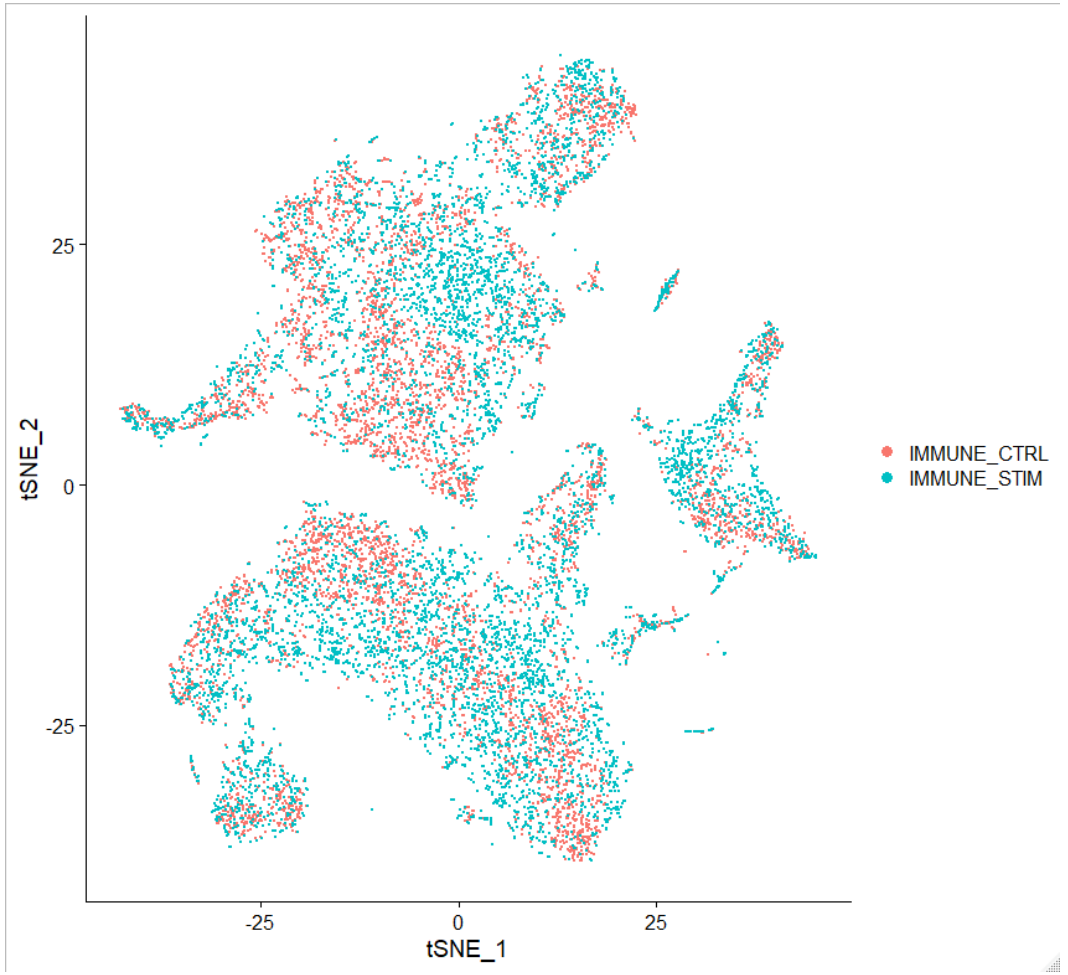
Anchoring Runs

- Method to try to re-align different runs
- Uses mutual nearest neighbour searches between runs to pair up cells
- Uses pairs to align the dimension reduction plots

Anchoring Runs

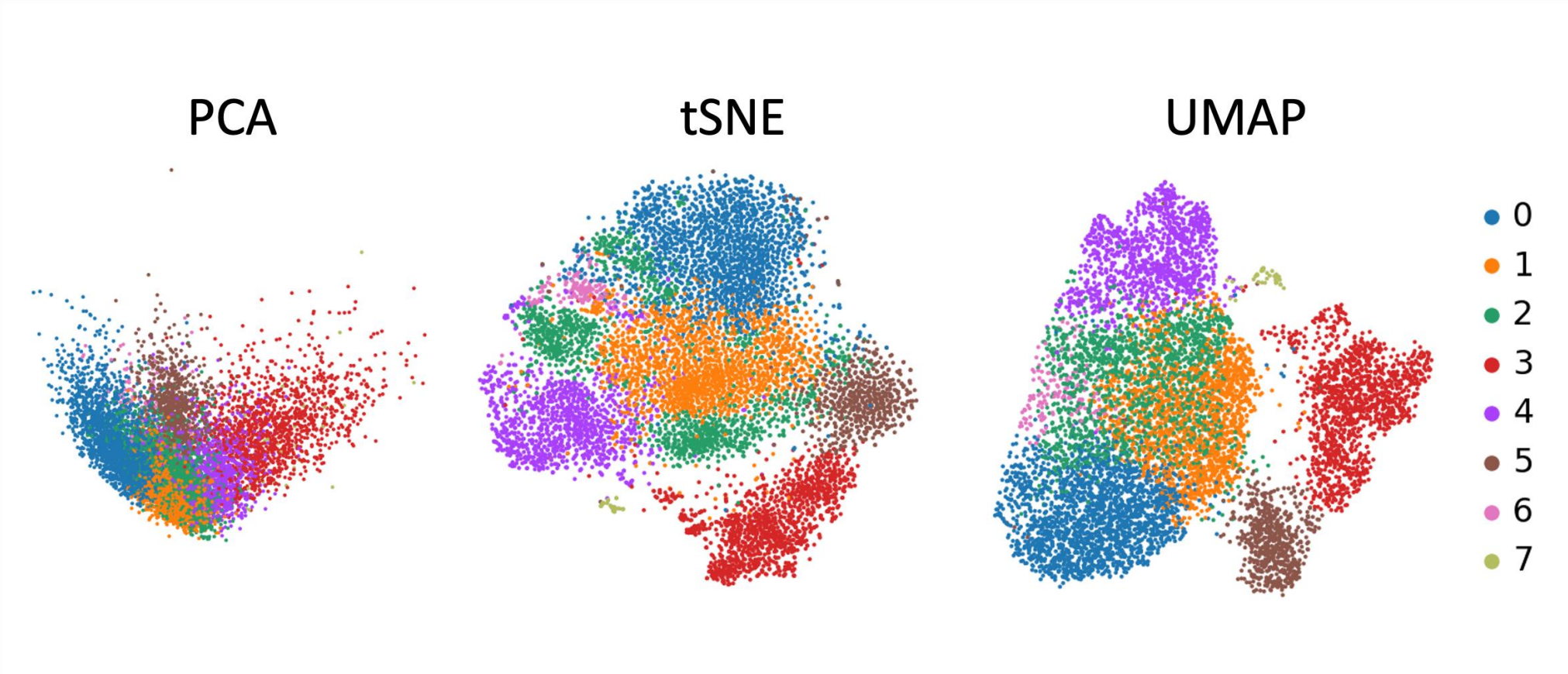


Raw



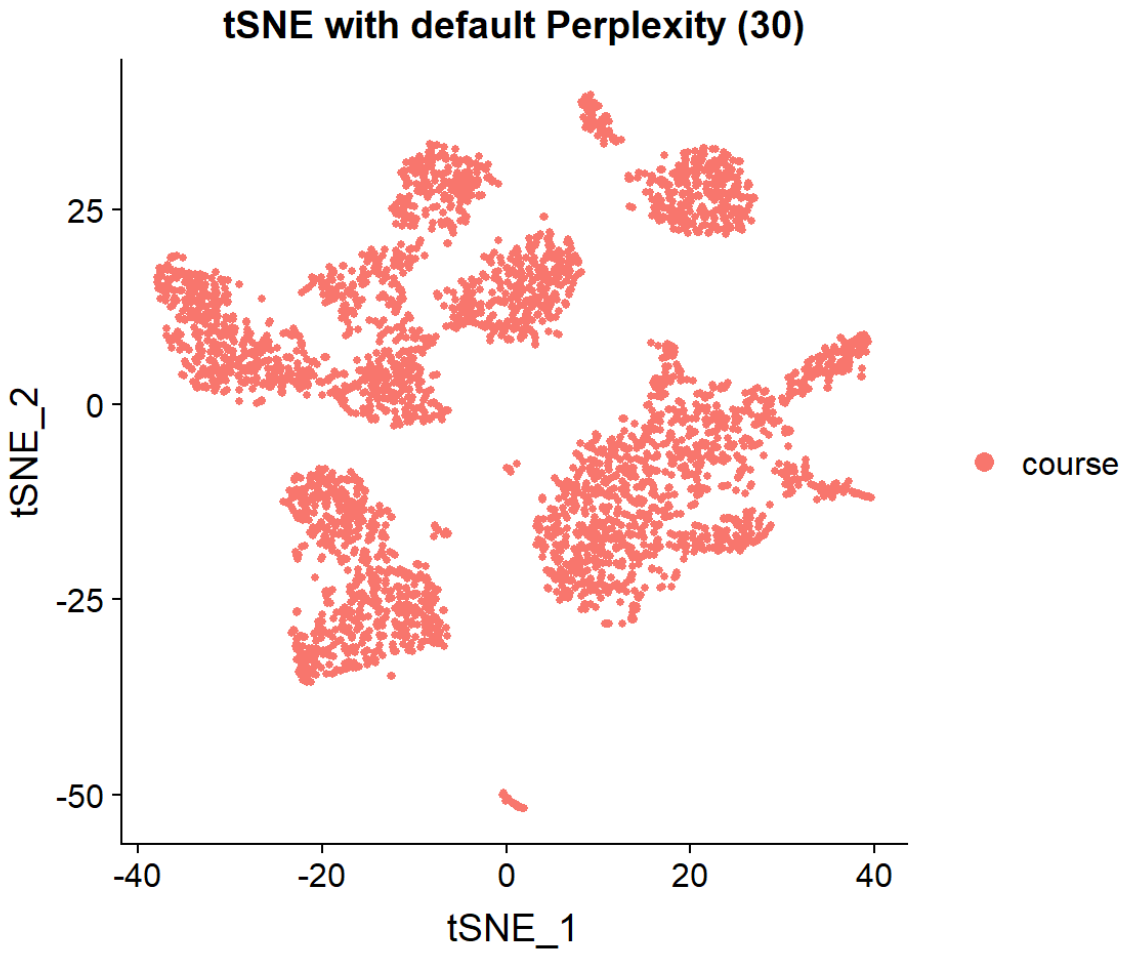
Anchored

Dimensionality Reduction



Dimensionality Reduction

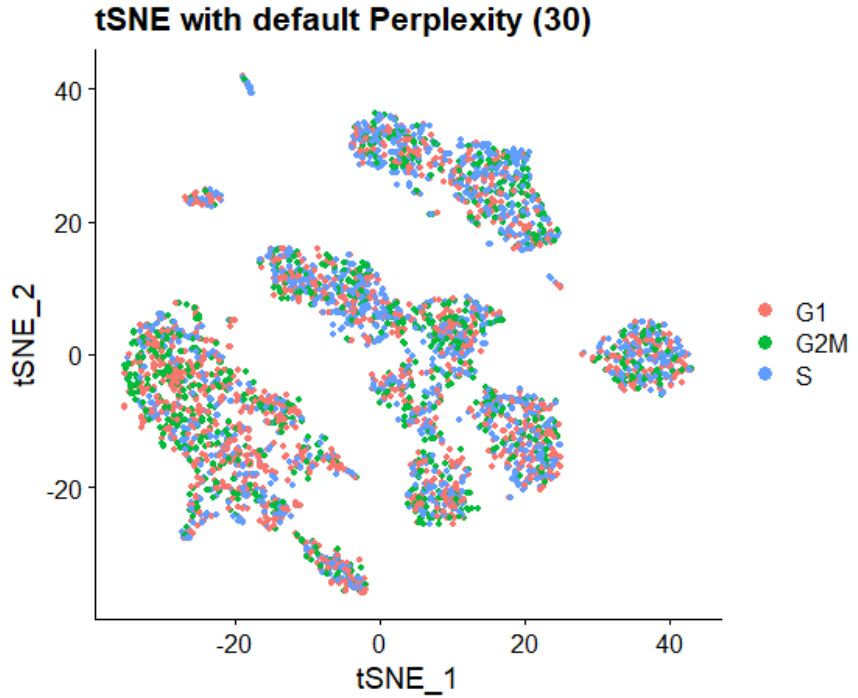
- Start with PCA on the normalised, filtered (both cells and genes), scaled data
- Scree / Elbow plot to decide how many PCs are informative
- Pass only the interesting PCs to subsequent tSNE or UMAP reduction to get down to 2 dimensions

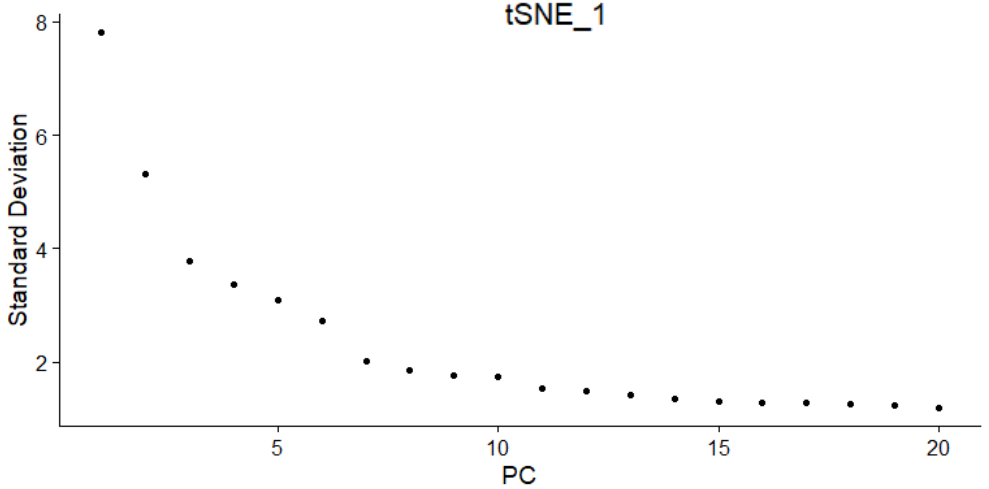


Dimensionality Reduction

```
RunPCA(
      data,
      features=VariableFeatures(data)
) -> data
```

```
RunTSNE(
   data,
   dims=1:15,
   seed.use = saved.seed,
   perplexity=30
) -> data
```



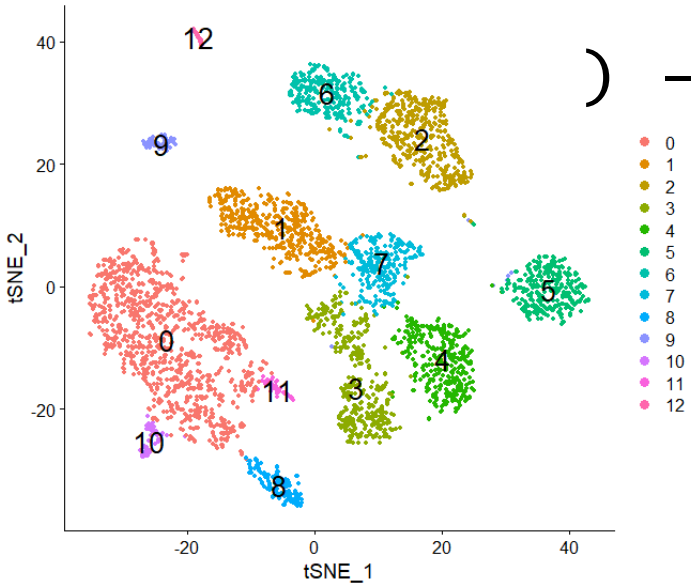


Defining clusters

- Construct nearest neighbour graph (can specify how many neighbours)
  - Constructed from PCA
  - Normally use the same number of dimensions as for tSNE/UMAP

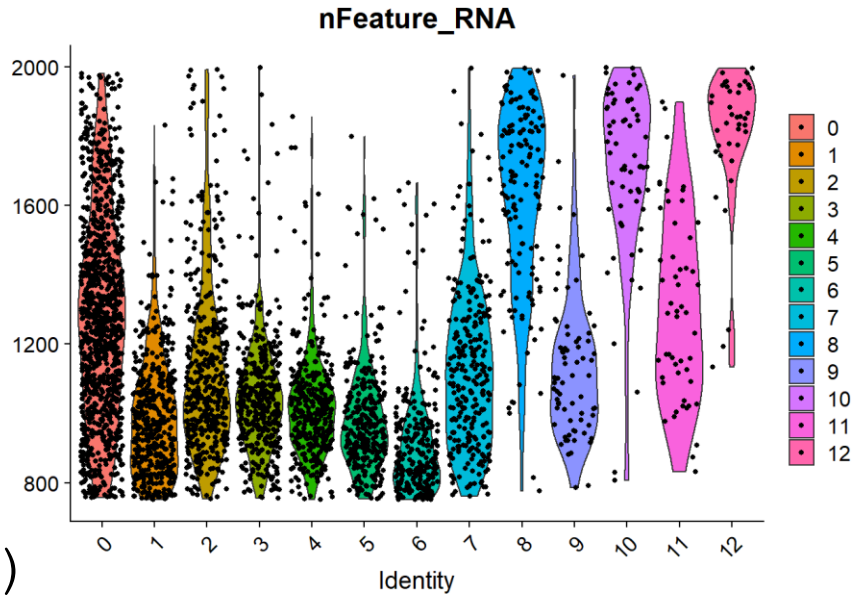
- Find local clusters
  - All cells are classified
  - Graph Based Clustering (Louvain method)
  - Resolution defines granularity

```
FindNeighbors(
      data,
      dims=1:15
) -> data

FindClusters(
      data,
      resolution = 0.5
) -> data
```



Comparing Properties of Clusters

- We want to know that clusters are occurring because of biological changes, not technical differences
- We can plot out the aggregate QC metrics for clusters
  - Read/Gene counts
  - Mitochondrion
  - MALAT1



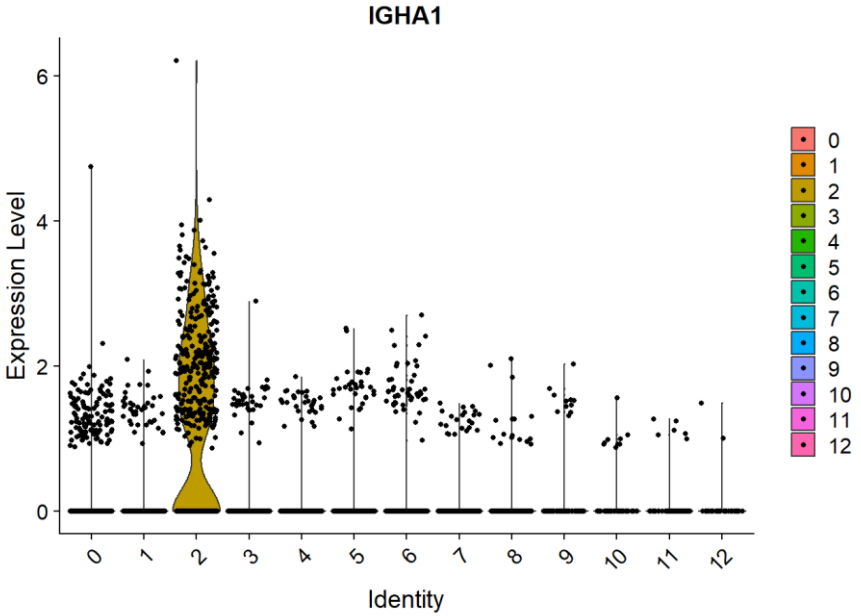

```
VlnPlot(data,features="nFeature_RNA")
```

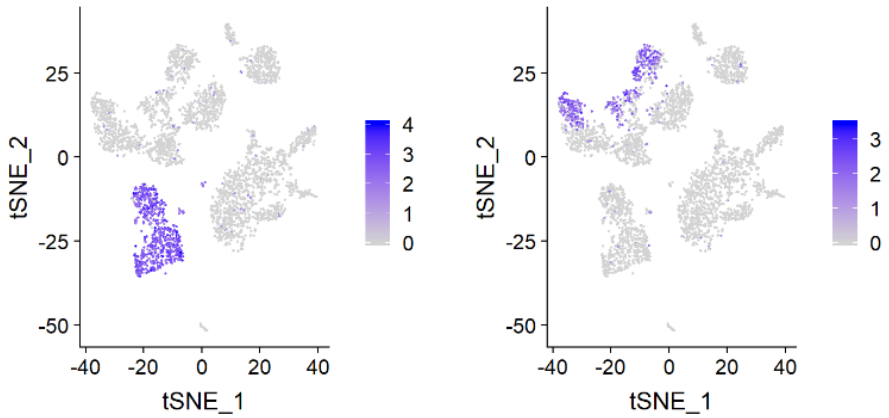
Statistical analysis of differences between clusters

- Different types of hits
  - Quantitatively significant between clusters
  - Qualitatively different (predictive) of cluster membership
- Different type of markers
  - Global: Distinguish one cluster from all of the rest of the data
  - Local: Distinguish one cluster from another defined set of clusters
- Often filter genes based on coverage in the set or the size of groups
- Several choices of method to identify genes

Statistical analysis of differences between clusters

- Non-parametric
  - Wilcox rank sum test
- Parametric
  - T-test
  - Negative binomial (eg DESeq)
- Classification
  - ROC analysis
- Specialised
  - MAST

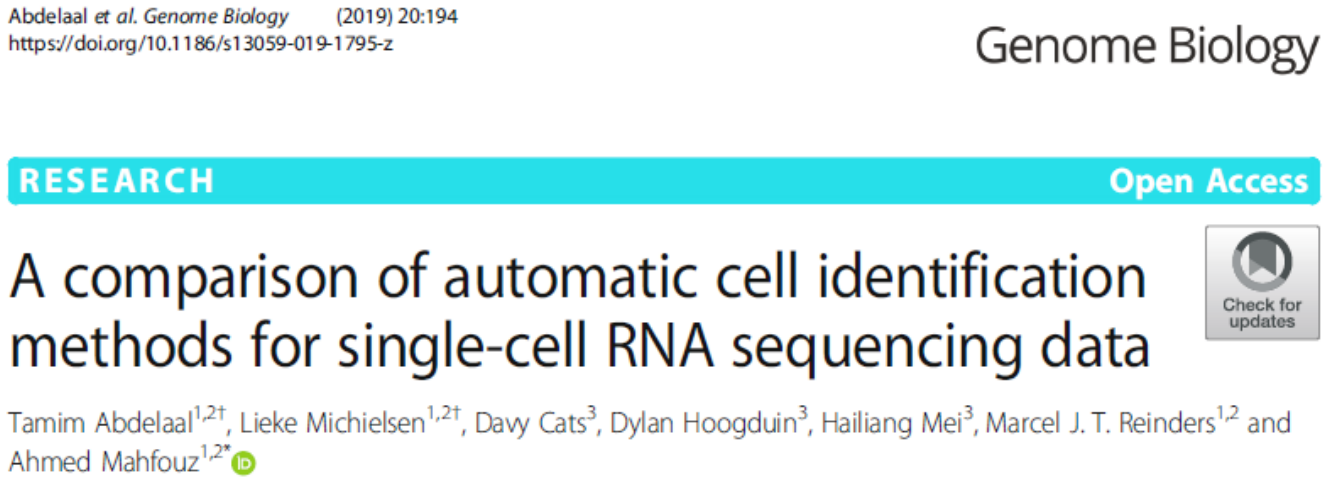





```
FindMarkers(
      data,
      ident.1 = 2,
      ident.2 = 6,
      test.use = "roc",
      only.pos = TRUE
)
```

Automated Cell Assignment

- Can automatically assign cell identities to clusters
- Need a source of marker genes
  - Result of a previous run/experiment
  - Publicly available data
  - Biggest hurdle
- Many packages to do this
  - SCINA
  - SingleR


Abdelaal *et al. Genome Biology* (2019) 20:194
https://doi.org/10.1186/s13059-019-1795-z

Genome Biology

RESEARCH Open Access

A comparison of automatic cell identification methods for single-cell RNA sequencing data

Tamim Abdelaal[1,2†], Lieke Michielsen[1,2†], Davy Cats[3], Dylan Hoogduin[3], Hailiang Mei[3], Marcel J. T. Reinders[1,2] and Ahmed Mahfouz[1,2*]

Hands on