Single cell RNAseq basics

Anna Lorenc (Ania)
al16@sanger.ac.uk

wellcome sanger institute

Aims for this lecture

- Recap of the technology
- Understand challenges & ways to solve them
- Steps from sequenced reads to cell-level data
- From cell-level data to better cell level data

Principles

From FASTQ to expression matrices

Better expression data

3 parts – questions after each

principles

We want to know expression of ~~all~~ many genes in ~~all~~ many cells

...to understand how cells and tissues differ between each other

...to understand how cells and tissues change in a perturbation (experiment, disease, development)

principles

We want to know expression of many genes in many cells

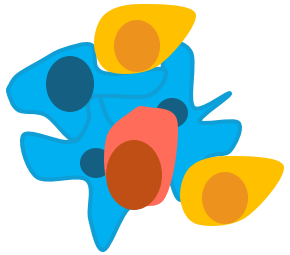
tissue

Prepare: dissolve, sort

Put cells in isolation
Add a unique tag to each cell

(Cell bursting)
Reverse transcription
Amplify

Prepare library
Sequence



```
@A00815:607:H5V32DRX3:1:2101:1108:1000 2:N:0:ACAATGTGAA+TAACGGTACG
GAACCACTGAGCACAAGTTTCTTCATCGTTCCTCAGATTCAGTAACATTATTAATTTTAGACAATCCCCGTGAAGGCCAATTCATCAGTG
+
FF:FFFF:F,:FFFFF,F,:F,FFF,FFFFFFFFF,,,,:F::,FFFFFFF,FF,FF,FF::,:FFF,,F,:FFF:F,FF,F:F:FF,F,
@A00815:607:H5V32DRX3:1:2101:1181:1000 2:N:0:AAAATGTGAA+TAACGGTACG
CGTCAAGGCAGGGCCTCGTTCTTGCTGGGCACCAAGAGCAGATGACATATATAGCACAGTGCCTCCCCCAGGACAGGAAGATGAGGCTAG
+
F,FFFF,FF,:FF,FF,,,F:F:,,FFF,F:FF:FFF,F,F,,,FFF,,FFF:,:F:FFF,FF,FFF:::F:F:,FF,,FF,,,FFFF:,
@A00815:607:H5V32DRX3:1:2101:1597:1000 2:N:0:ACAATGTGAA+TAACGGTACG
GAAGTGAGGGATGCTGAGGGCCGGGACAAGCTATCGGACTGTCTGCTGCCATCGGTAATGAGTCTCAGTAGACCTGGAACGTCACCTCGC
+
FFFFFFFFFFFFFFFFF,FFFFF:FFFFFFFFFFFFFFFFFF,FFFF,:FFF,:FFFFF,FFFFFFFFFFFFFFFFFFFFF:FFFFFF:
@A00815:607:H5V32DRX3:1:2101:1615:1000 2:N:0:ACAATGTGAA+TAACGGTACG
ATGTGACTATAGGCTCATAGCCATCTCATTATGCAAAATGTATTCATACTGTCTTTGTATGTCTCAATAGTCCTCCAGATATACGGCGGT
+
,,,:FFFF,,FFFF,FFFF,F:F,F,FF,FF,:FFF:,F,:F,FFF,:FF,::,::F,F,FF,,F,F,FFFFFF,FF,,F,F,F,:,,F:
```

Challenges

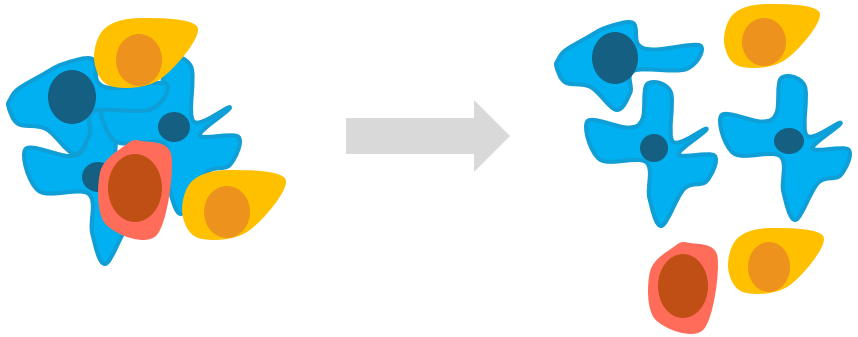
- very little RNA in each cell
- PCR steps – amplification bias
- separate cells so we can measure each cell separately
  - → cells difficult to separate, fragile
  - → empty droplets/multiplets
- cells dying/bursting during the procedure
- very wide dynamic range of expression of genes
- not enough of cells of a specific type

principles

tissue

Prepare:
dissolve, sort

- Gentle dissociation protocols
- Use nuclei instead of cells
- Assess viability/sort
- Use only viable cells

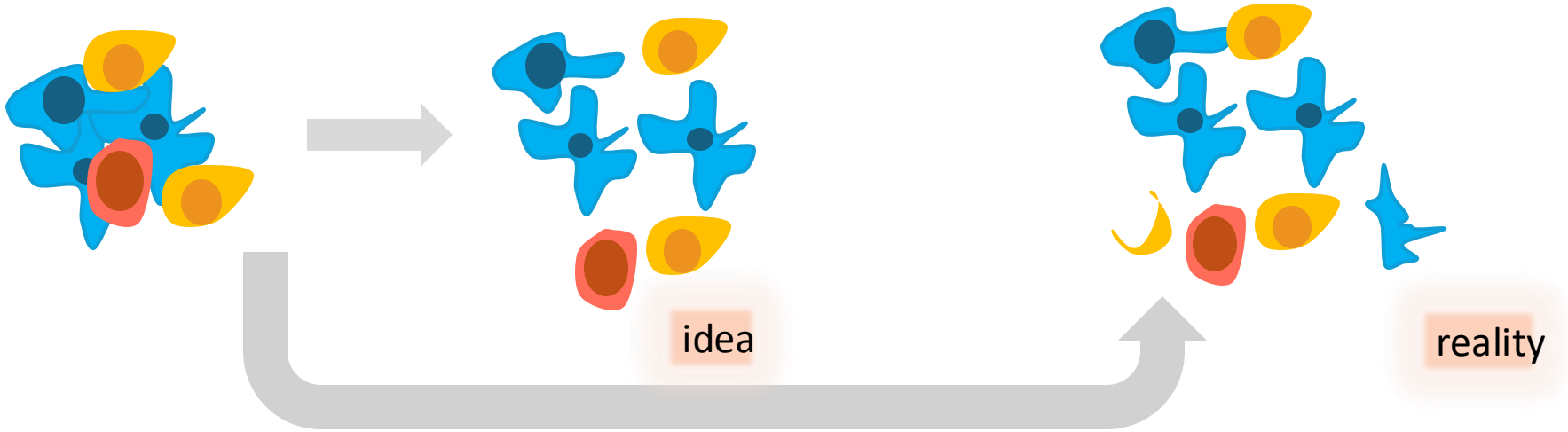


principles

tissue

Prepare:
dissolve, sort

- Gentle dissociation protocols
- Use nuclei instead of cells
- Assess viability/sort
- Use only viable cells



And presence of debris is a challenge because...

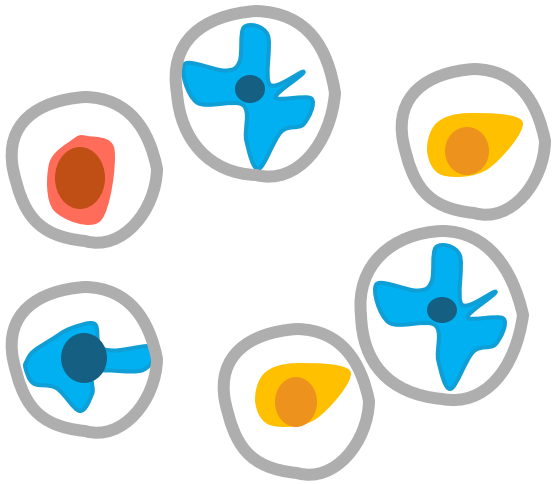
principles

tissue

Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to each cell

- wells
- microwells (BDRhapsody, Hive)
- droplets (10x)
- combinatorial tags (PARSE, Scala)



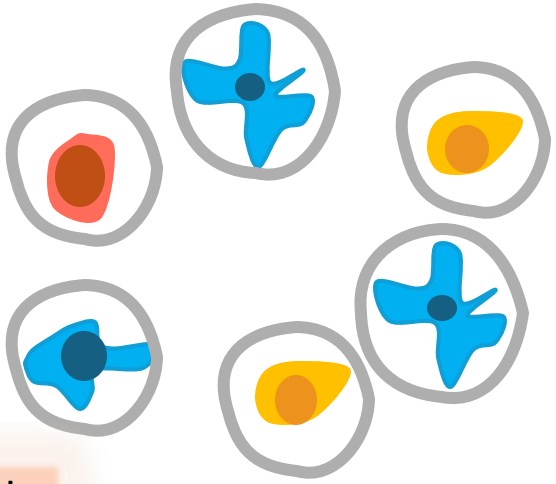
principles

tissue

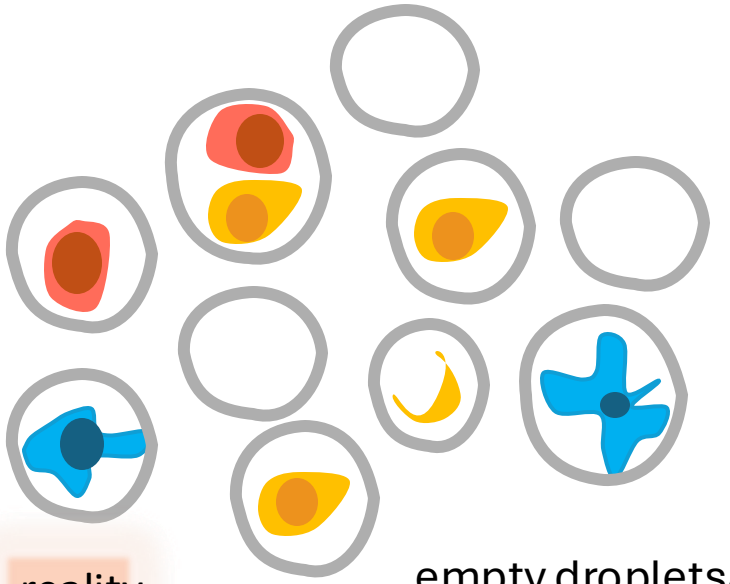
Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to each cell

- wells
- microwells (BDRhapsody, Hive)
- droplets (10x)
- combinatorial tags (PARSE, Scala)



idea



reality

empty droplets/wells ← **single cell** → multiplets

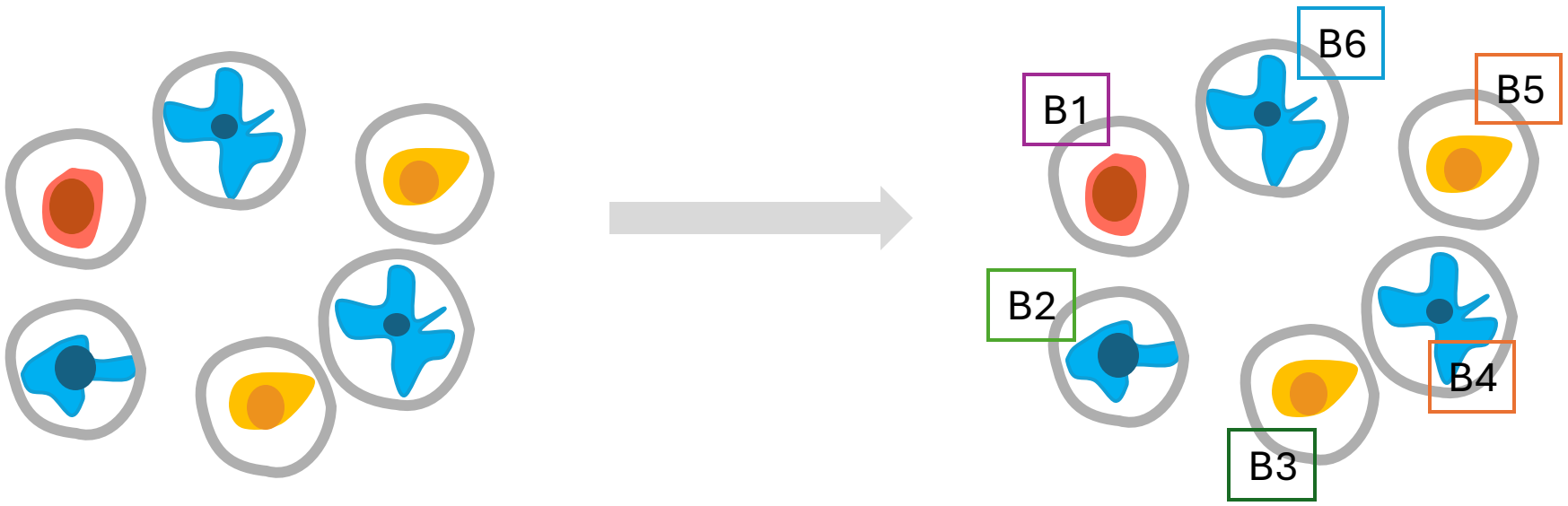
principles

tissue

Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to each cell

Cell barcodes are included in cDNA during reverse transcription



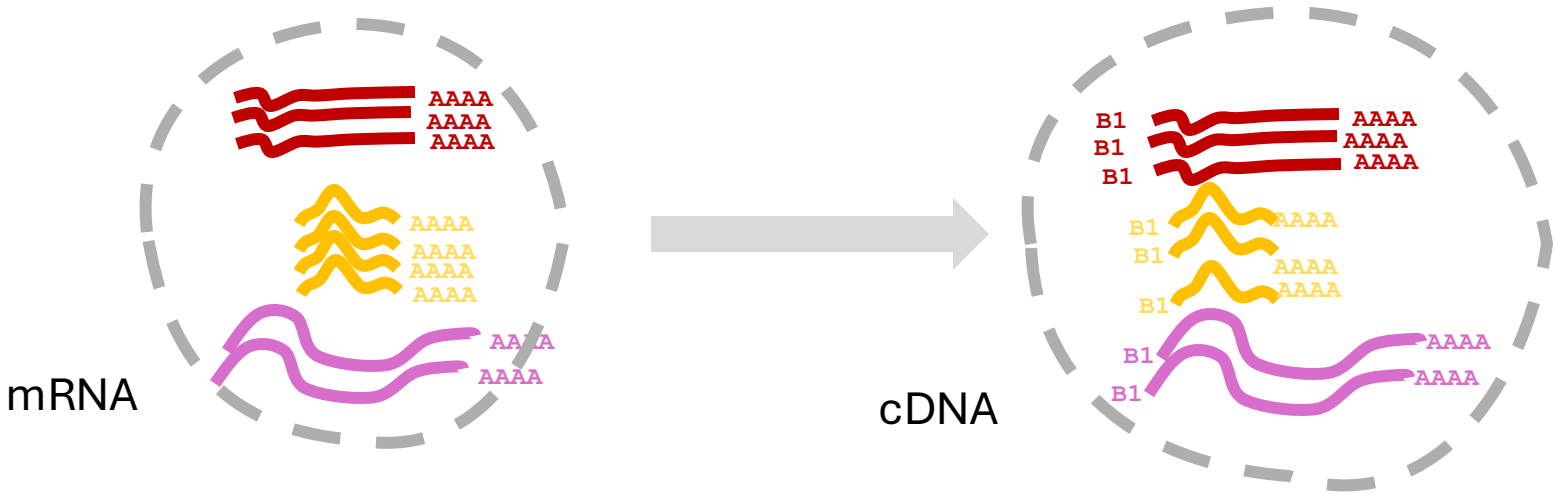
principles

tissue

Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to each cell

(Cell bursting)
Reverse transcription
Amplify



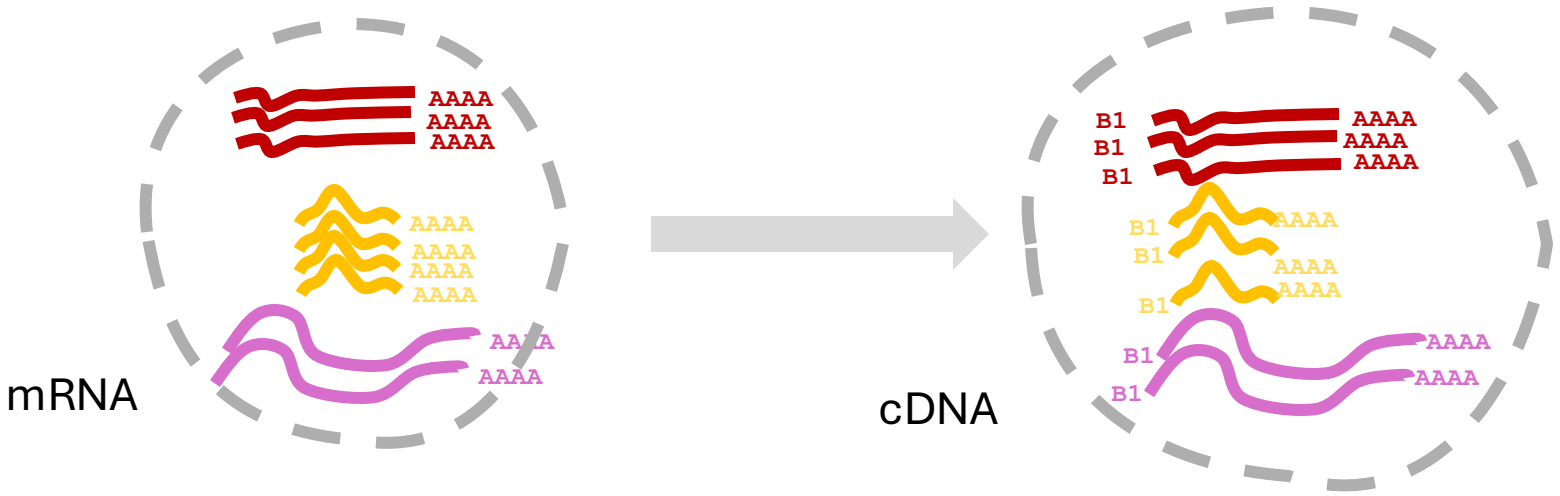
principles

tissue

Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to each cell

(Cell bursting)
Reverse transcription
Amplify



principles

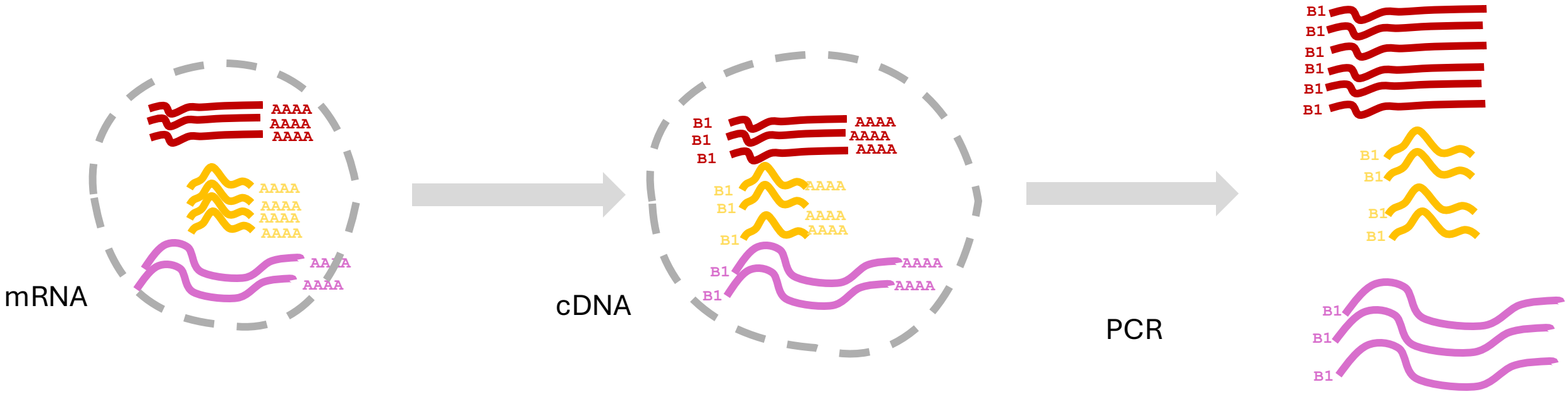
tissue

Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to each cell

(Cell bursting)
Reverse transcription
Amplify

- Each cell has very little mRNA → amplification
- PCR bias → UMI included when cDNA constructed



principles

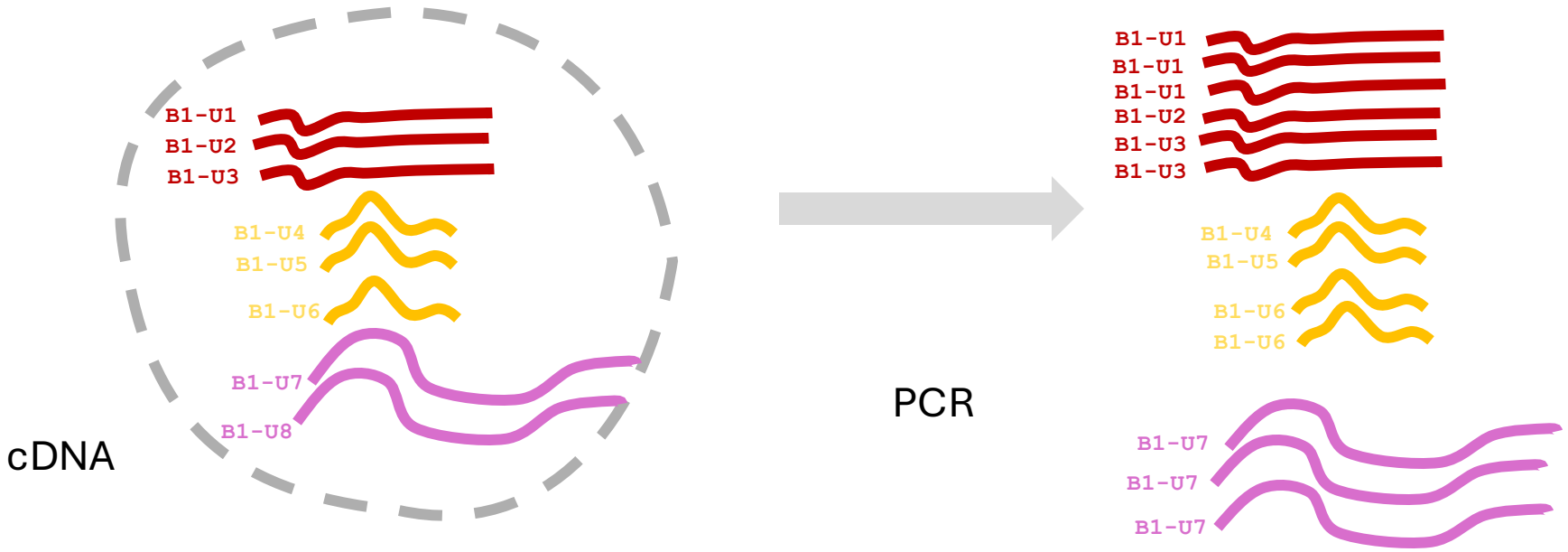
tissue

Prepare: dissolve, sort

Put cells in isolation
Add a unique tag to each cell

(Cell bursting)
Reverse transcription
Amplify

- Each cell has very little mRNA → amplification
- PCR bias → UMI included when cDNA constructed



principles

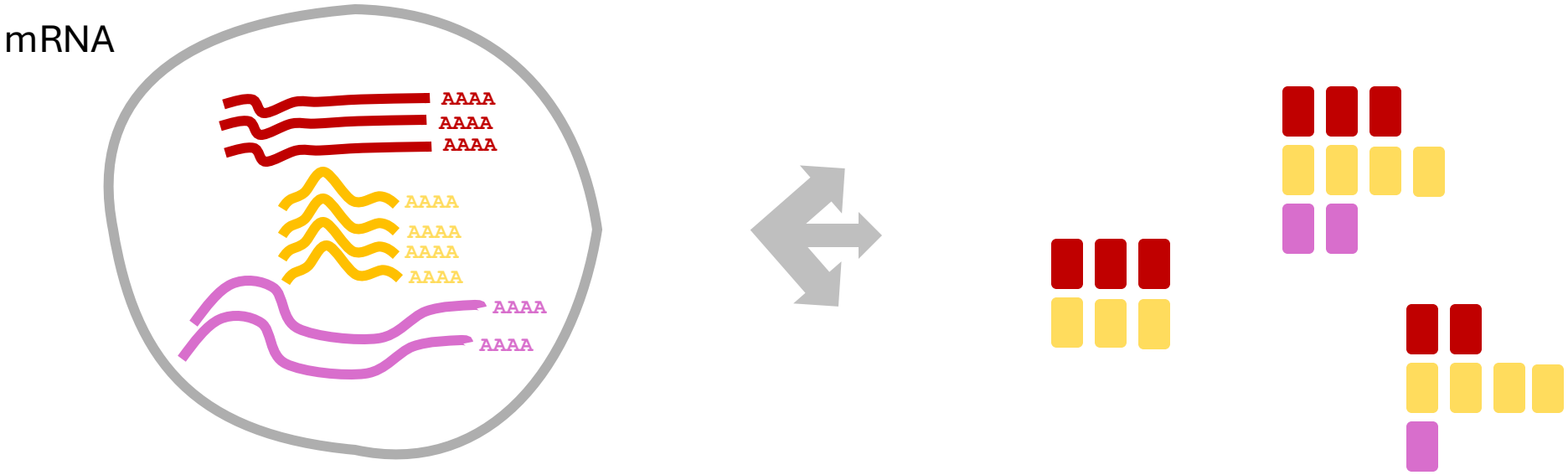
tissue

Prepare: dissolve, sort

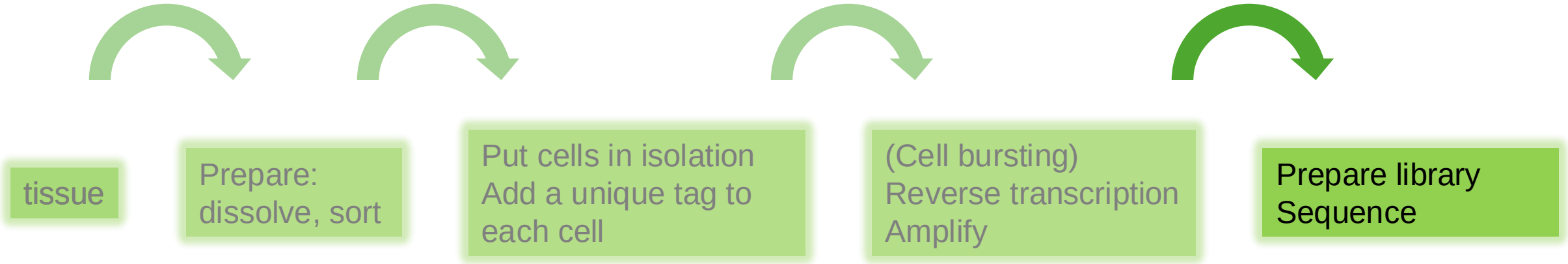
Put cells in isolation
Add a unique tag to each cell

(Cell bursting)
Reverse transcription
Amplify

- Drop outs → sparsity of the data



principles



Many platforms

principles

tissue

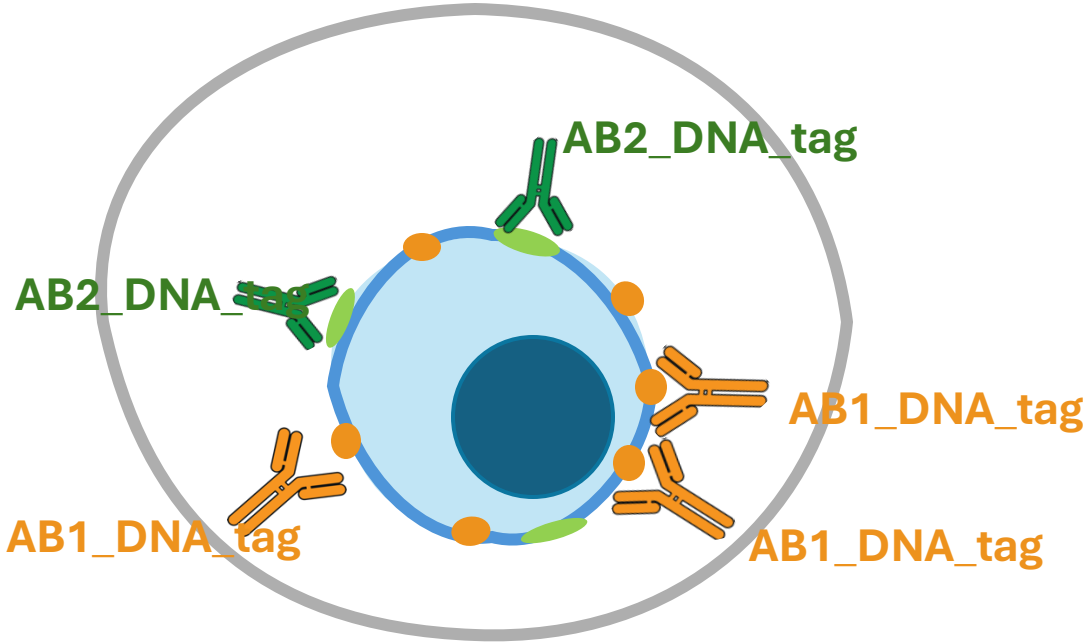
Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to each cell

(Cell bursting)
Reverse transcription
Amplify

Prepare library
Sequence

Stain proteins with AB-DNA tags



principles

tissue

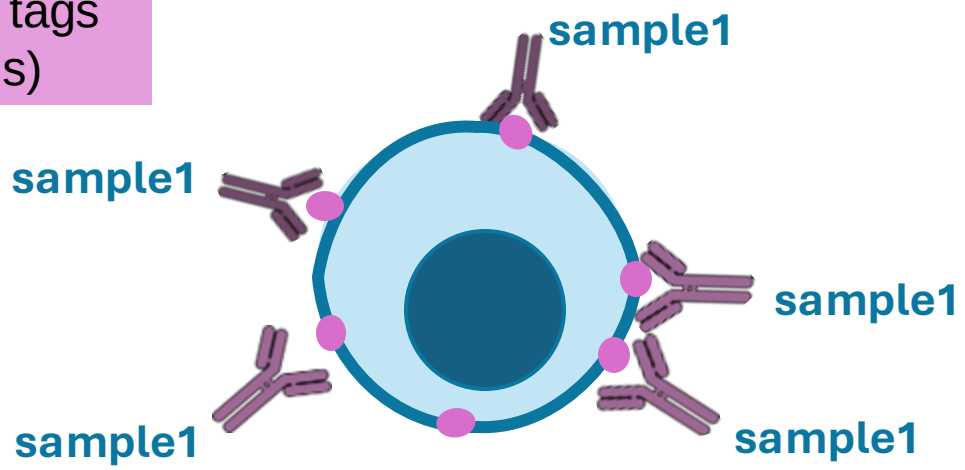
Prepare:
dissolve, sort

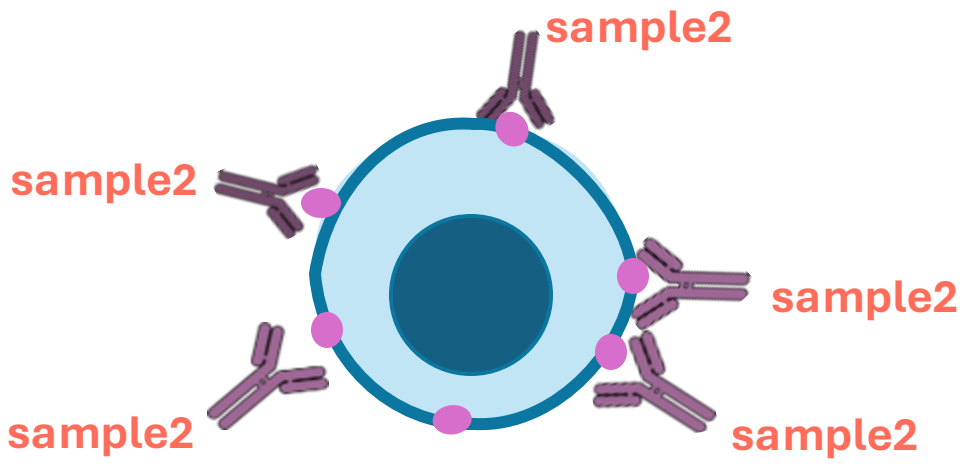
Put cells in isolation
Add a unique tag to each cell

(Cell bursting)
Reverse transcription
Amplify

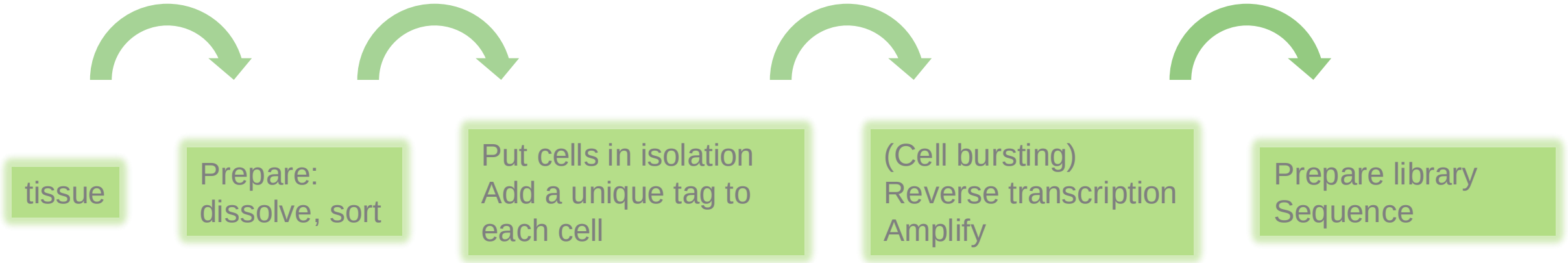
Prepare library
Sequence

Sample tags
(#tags)





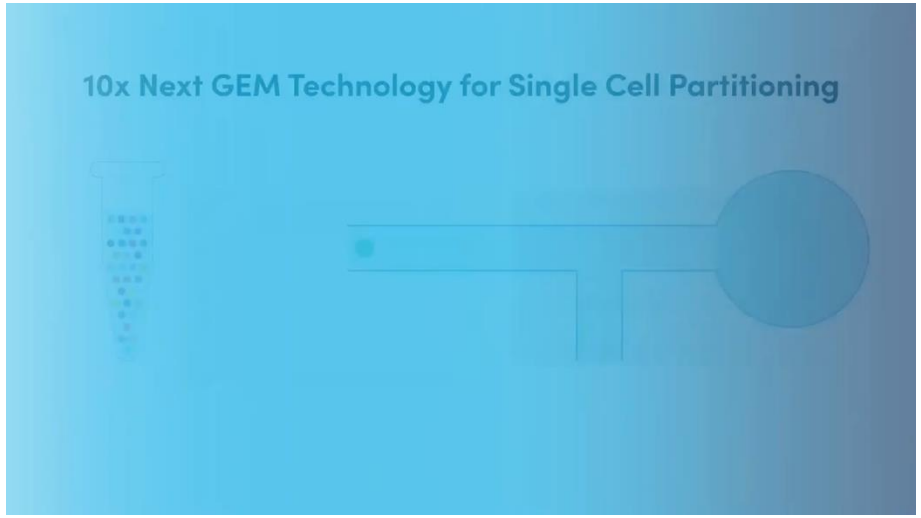
principles

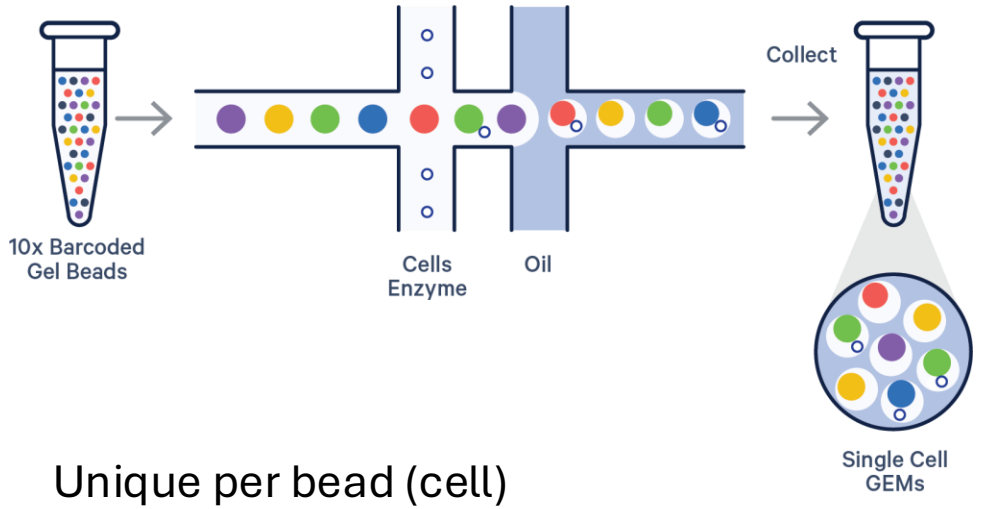


Example of specific platform

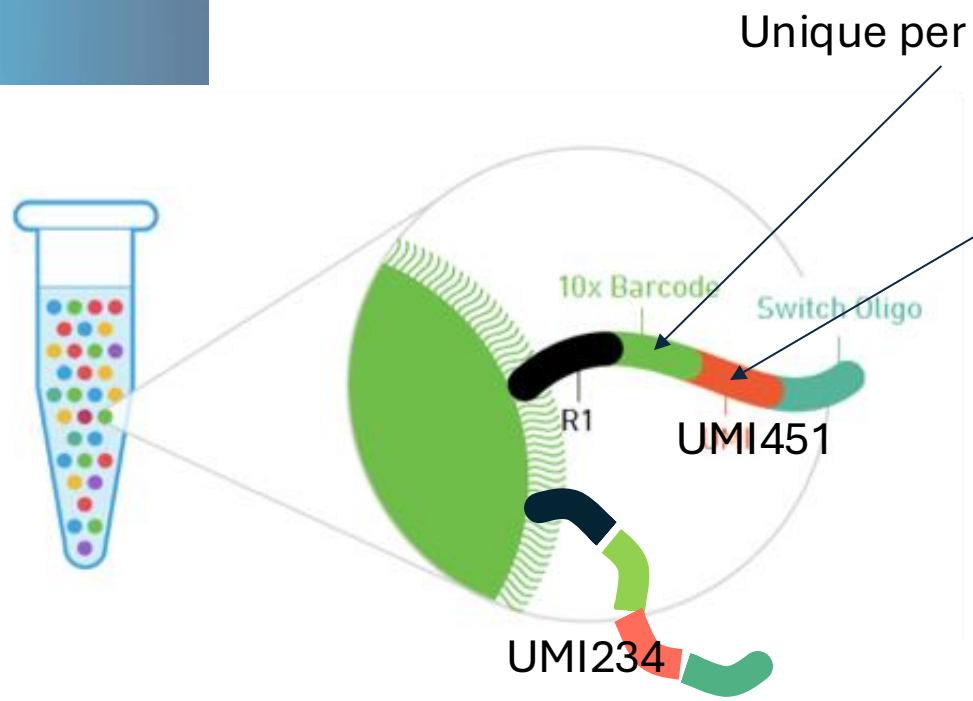
10x GENOMICS™

principles 10x

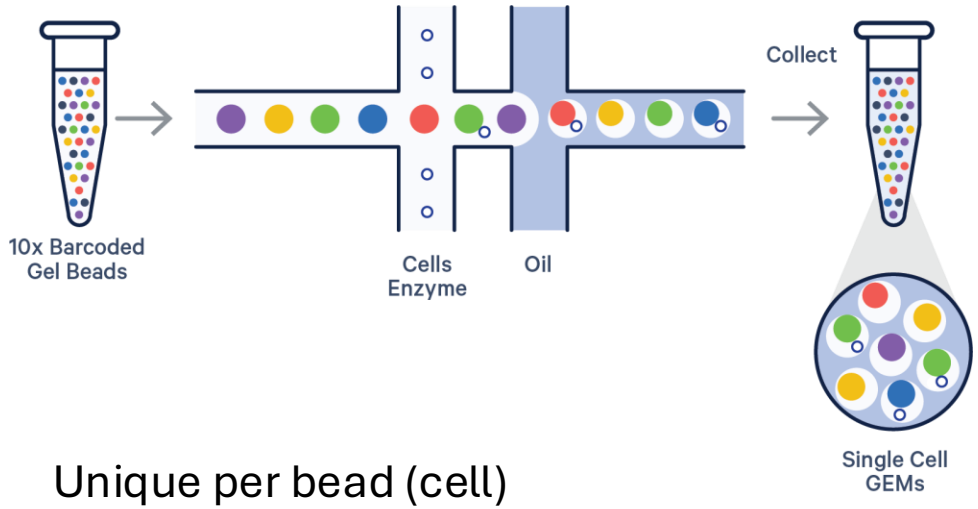




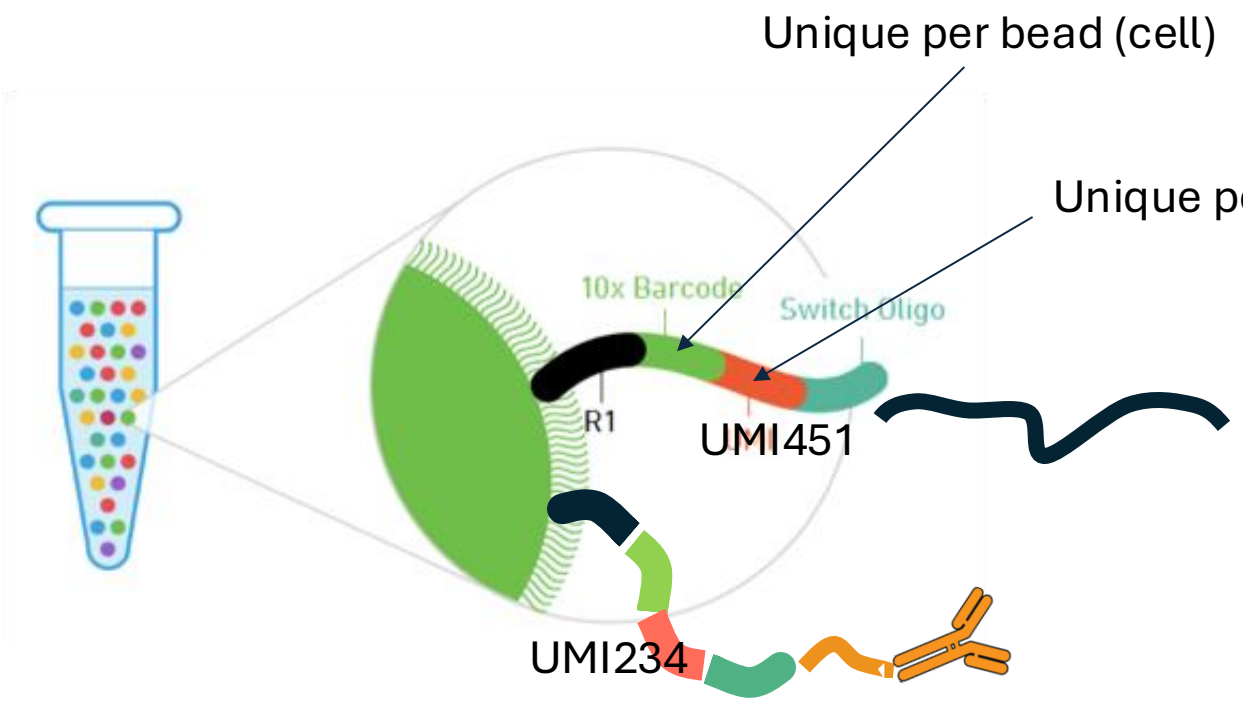
5’ GEX +
CITEseq



principles 10x

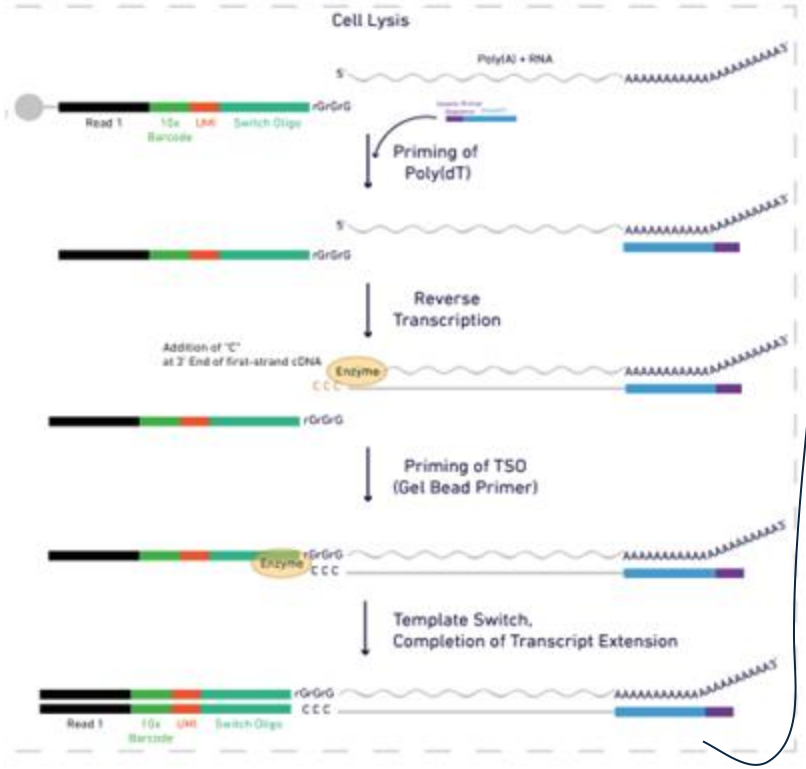


5’ GEX

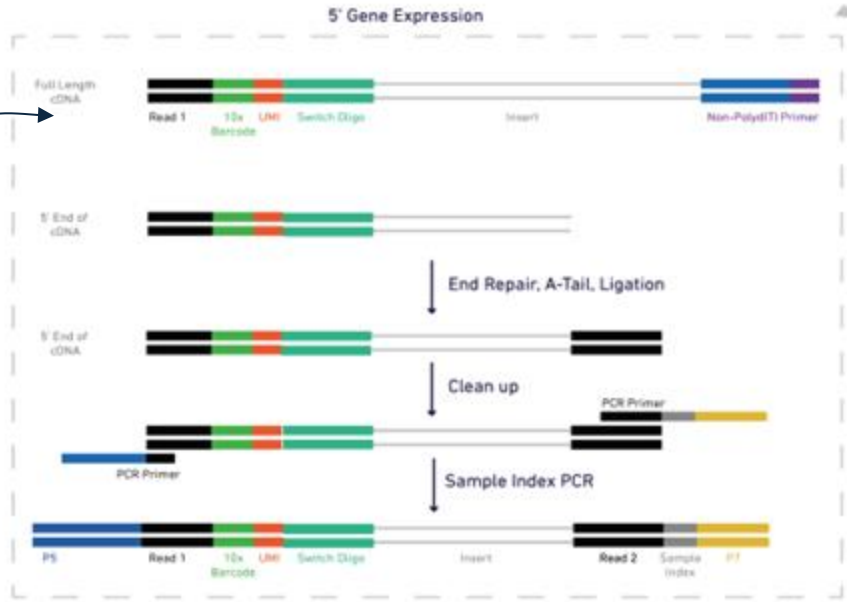


principles 10x

Inside individual GEMs



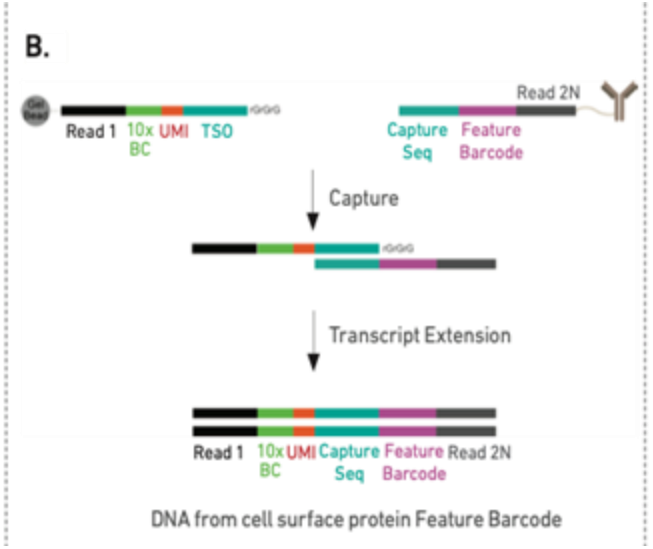
Pooled cDNA processed in bulk



GEX library

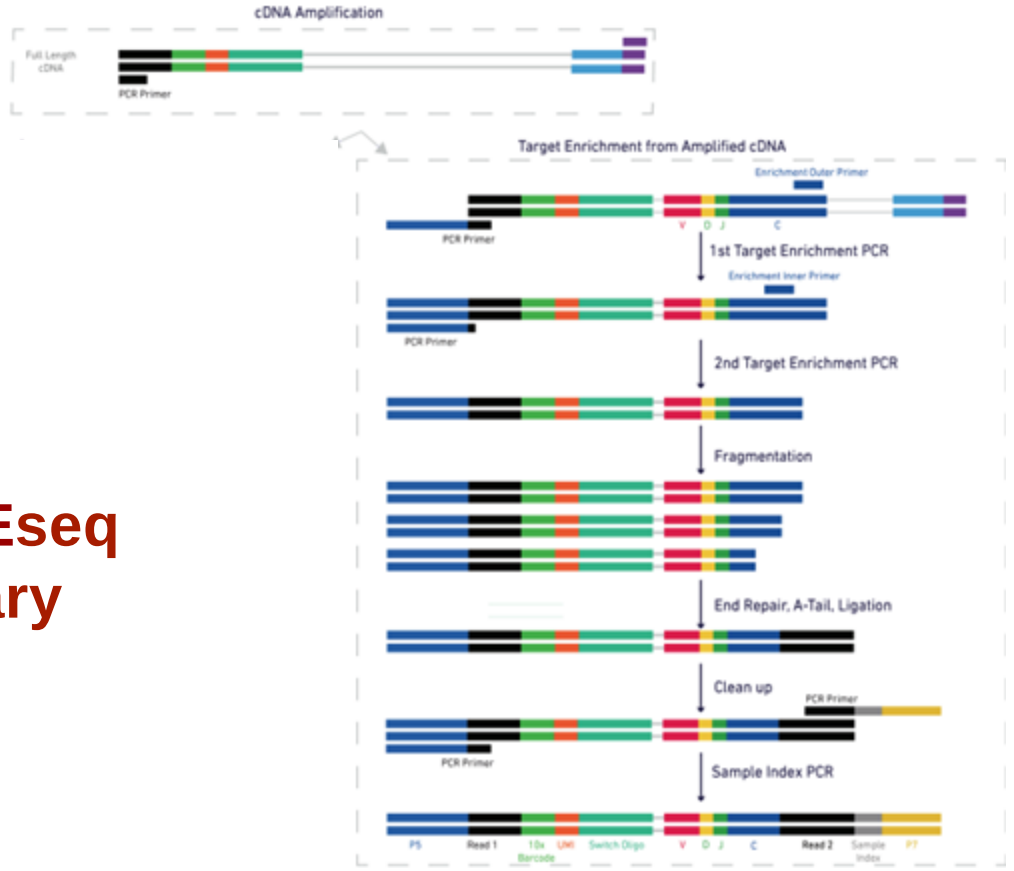
principles 10x

Inside individual GEMs



CITEseq library

Pooled cDNA processed in bulk



VDJ library

10x GENOMICS™

principles

How much to sequence? ← Depends on the technology & desired resolution

scRNAseq

- 10-15K reads per cell
- 50K reads per cell
- Saturation: the bottleneck is the amount of starting RNA
- Optimise: more transcripts, more cells, more biological replicates

CITEseq
cell hashtags

- Less complexity
- Small pool of possible sequences
- Few types of reads per cell (depends on the number of Abs)

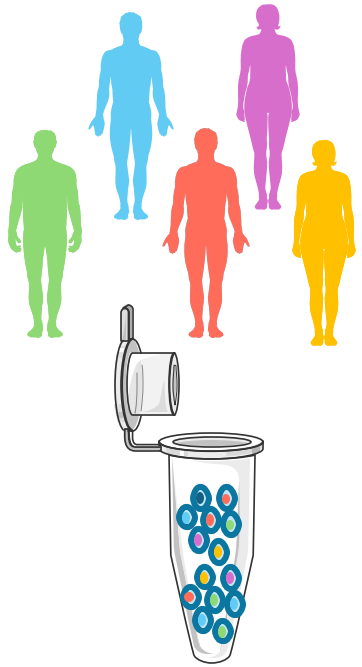
questions

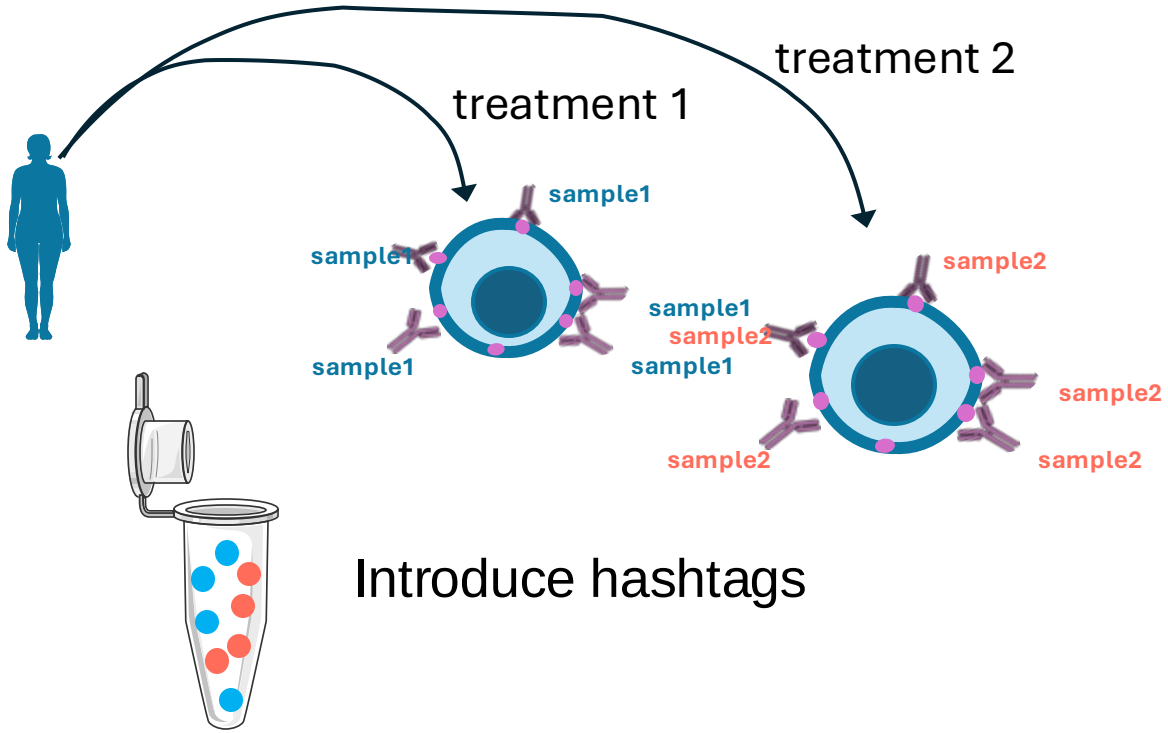
principles

Pool as much as possible, as soon as possible

- Reduce batch effect
- Identify multiplets
- Pool samples to reach minimal cell numbers

Use samples' genotypes

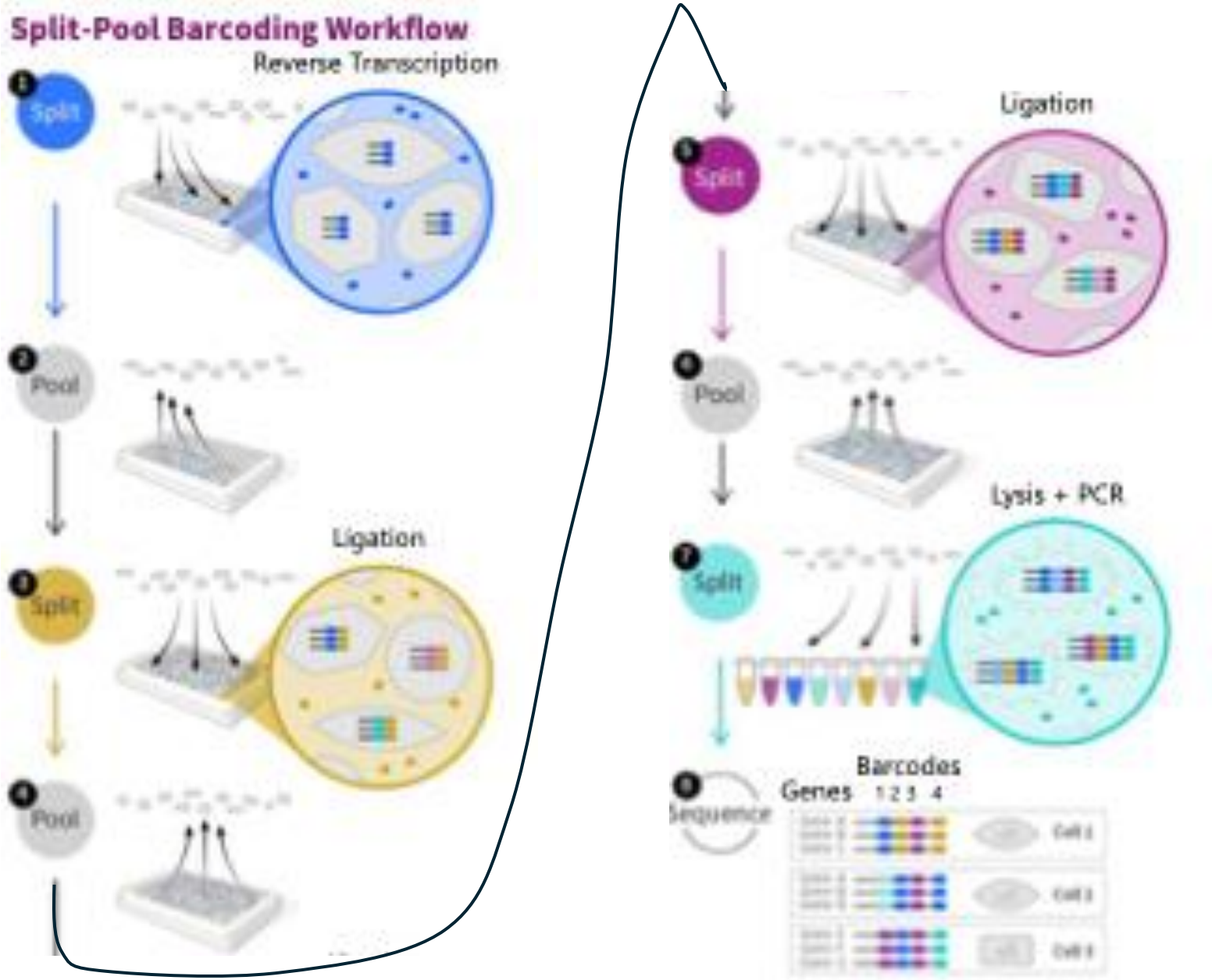




Introduce hashtags

principles

Fixed cells



PARSE
SCALE Bio

From FASTQ to expression matrices

FASTQ files

```
@A00815:607:H5V32DRX3:1:2101:1108:1000 2:N:0:ACAATGTGAA+TAACGGTACG
GAACCACTGAGCACAAGTTTCTTCATCGTTCCTCAGATTCAGTAACATTATTAATTTTAGACAATCCCCGTGAAGGCCAATTCATCAGTG
+
FF:FFFF:F,:FFFFF,F,:F,FFF,FFFFFFFFF,,,,:F::,FFFFFFF,FF,FF,FF::,:FFF,,F,:FFF:F,FF,F:F:FF,F,
@A00815:607:H5V32DRX3:1:2101:1181:1000 2:N:0:AAAATGTGAA+TAACGGTACG
CGTCAAGGCAGGGCCTCGTTCTTGCTGGGCACCAAGAGCAGATGACATATATAGCACAGTGCCTCCCCCAGGACAGGAAGATGAGGCTAG
+
F,FFFF,FF,:FF,FF,,,F:F:,,FFF,F:FF:FFF,F,F,,,FFF,,FFF:,:F:FFF,FF,FFF:::F:F:,FF,,FF,,,FFFF:,
@A00815:607:H5V32DRX3:1:2101:1597:1000 2:N:0:ACAATGTGAA+TAACGGTACG
GAAGTGAGGGATGCTGAGGGCCGGGACAAGCTATCGGACTGTCTGCTGCCATCGGTAATGAGTCTCAGTAGACCTGGAACGTCACCTCGC
+
FFFFFFFFFFFFFFFFF,FFFFF:FFFFFFFFFFFFFFFFF,FFFF,:FFF,:FFFFF,FFFFFFFFFFFFFFFFFFFFF:FFFFFF:
@A00815:607:H5V32DRX3:1:2101:1615:1000 2:N:0:ACAATGTGAA+TAACGGTACG
ATGTGACTATAGGCTCATAGCCATCTCATTATGCAAAATGTATTCATACTGTCTTTGTATGTCTCAATAGTCCTCCAGATATACGGCGGT
+
,,,:FFFF,,FFFF,FFFF,F:F,F,FF,FF,:FFF:,F,:F,FFF,:FF,::,::F,F,FF,,F,F,FFFFFF,FF,,F,F,F,:,,F:
```

	cell1	cell2	cell3
gene1			
gene2			
gene3			

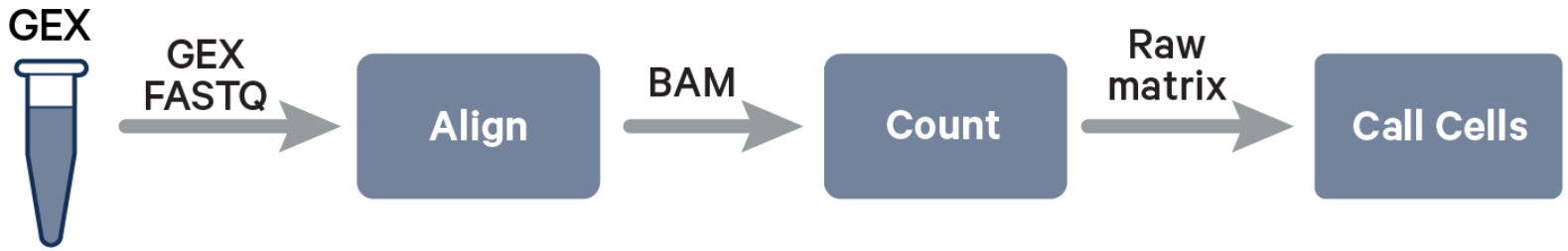
	cell1	cell2	cell3
abs1			
abs2			
abs3			

	cell1	cell2	cell3
#1			
#2			
#3			

- Done by proprietary software/free alternatives available
- Basis for any downstream analysis
- Some steps might be improved with additional community-developed software

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

From FASTQ to expression matrices



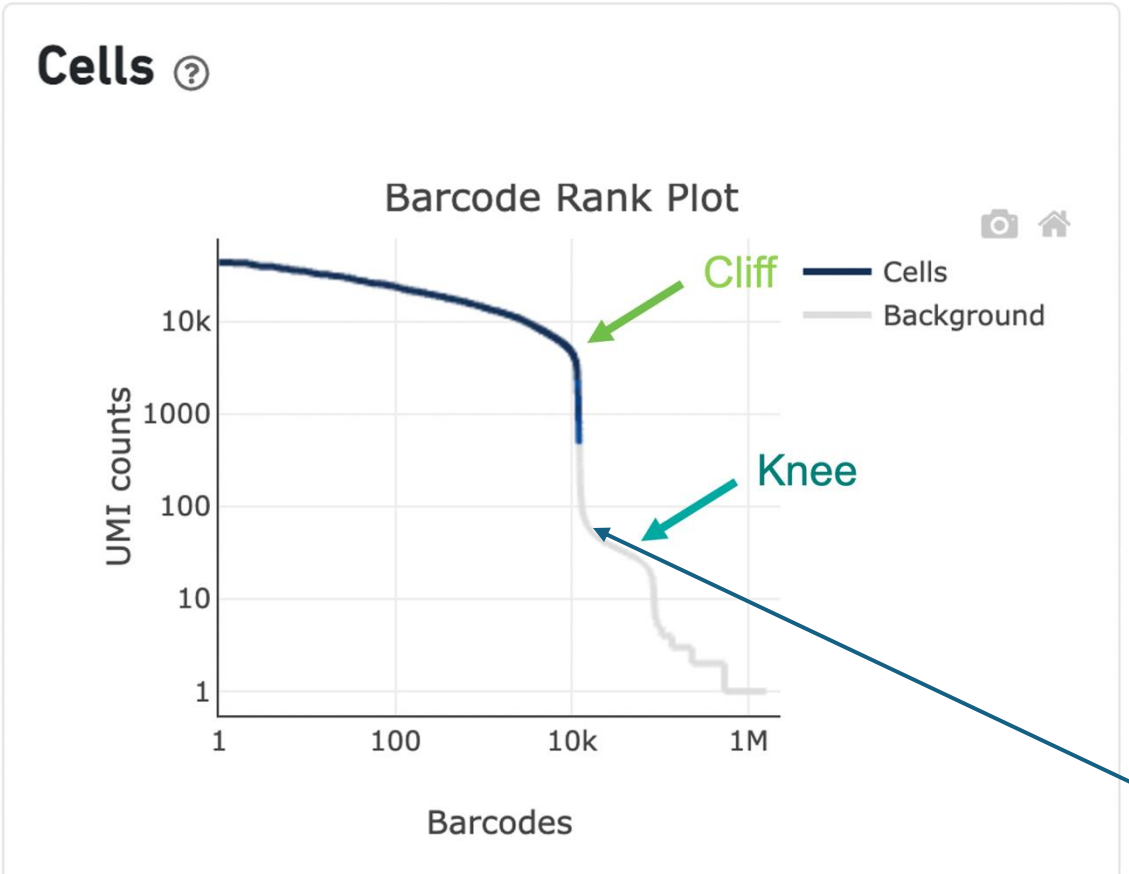
- Reads with valid barcodes
- Map to a genome
- Exclude/annotate intergenic reads
- Align to a transcriptome
- Transcriptomic reads
- UMI correction
- exclude low support molecules/resolve duplicated molecules (barcode+UMI+gene)

	barc1	barc2	barc3
gene1			
gene2			
gene3			

Cell Ranger

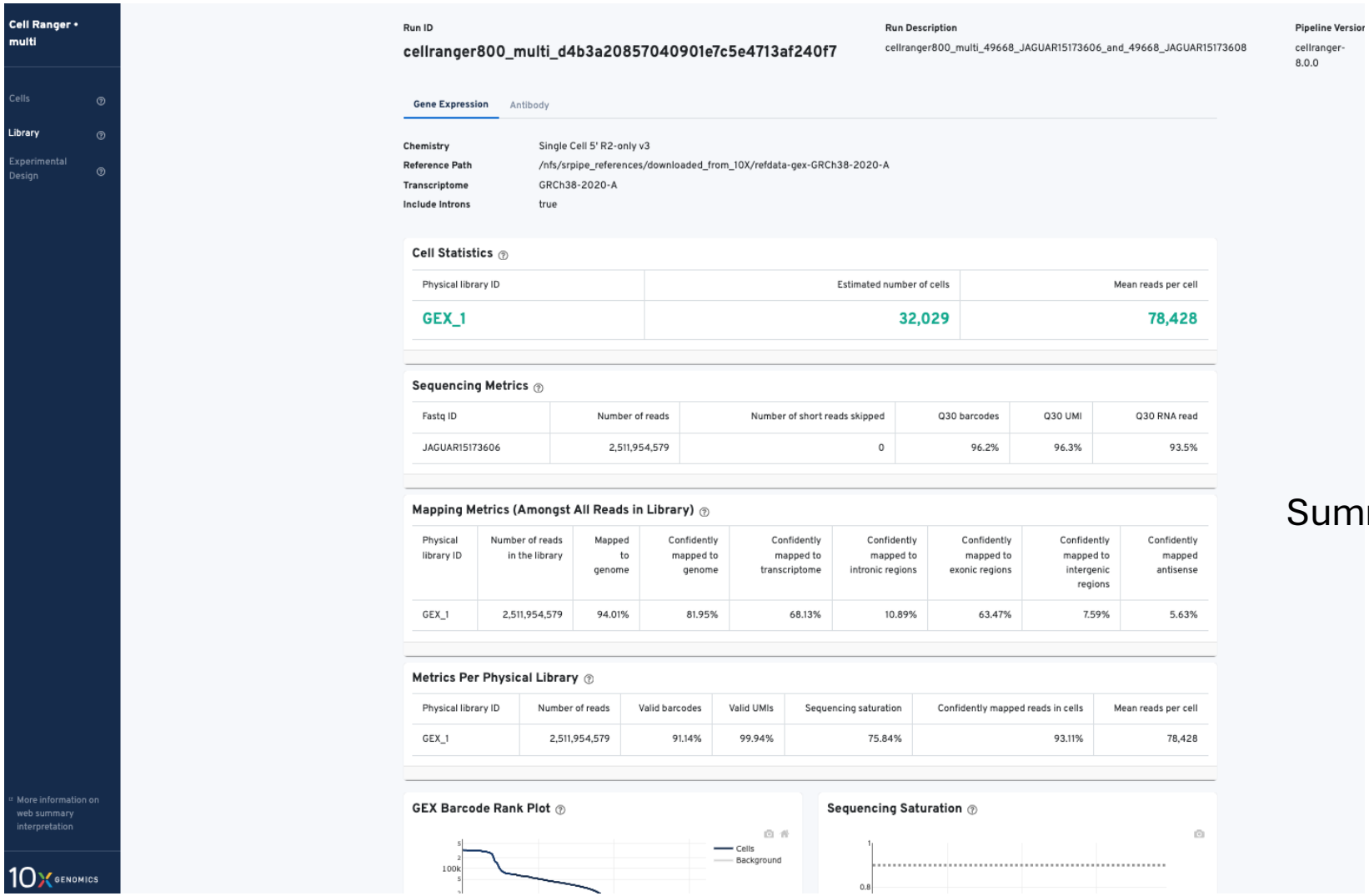
	barc1	barc2	barc3
gene1			
gene2			
gene3			
Total			

Is barcode a cell or an empty droplet?
→ Order barcodes by UMI counts
→ Barcode rank plot – identify the cutoff



Cell Ranger

From FASTQ to expression matrices



Cell Ranger • multi

Cells
Library
Experimental Design

More information on web summary interpretation

10x GENOMICS

Run ID
cellranger800_multi_d4b3a20857040901e7c5e4713af240f7

Run Description
cellranger800_multi_49668_JAGUAR15173606_and_49668_JAGUAR15173608

Pipeline Version
cellranger-8.0.0

Gene Expression | Antibody

Chemistry	Single Cell 5' R2-only v3
Reference Path	/nfs/srpipe_references/downloaded_from_10X/refdata-gex-GRCh38-2020-A
Transcriptome	GRCh38-2020-A
Include Introns	true

Cell Statistics

Physical library ID	Estimated number of cells	Mean reads per cell
GEX_1	32,029	78,428

Sequencing Metrics

Fastq ID	Number of reads	Number of short reads skipped	Q30 barcodes	Q30 UMI	Q30 RNA read
JAGUAR15173606	2,511,954,579	0	96.2%	96.3%	93.5%

Mapping Metrics (Amongst All Reads in Library)

Physical library ID	Number of reads in the library	Mapped to genome	Confidently mapped to genome	Confidently mapped to transcriptome	Confidently mapped to intronic regions	Confidently mapped to exonic regions	Confidently mapped to intergenic regions	Confidently mapped antisense
GEX_1	2,511,954,579	94.01%	81.95%	68.13%	10.89%	63.47%	7.59%	5.63%

Metrics Per Physical Library

Physical library ID	Number of reads	Valid barcodes	Valid UMIs	Sequencing saturation	Confidently mapped reads in cells	Mean reads per cell
GEX_1	2,511,954,579	91.14%	99.94%	75.84%	93.11%	78,428

GEX Barcode Rank Plot

Sequencing Saturation

Summary of reads/transcripts/cells...

Expression matrices

	cell1	cell2	cell3
gene1			
gene2			
gene3			

	cell1	cell2	cell3
#1			
#2			
#3			

	cell1	cell2	cell3
abs1			
abs2			
abs3			

Not ready for comparing gene expression etc. yet!

Demultiplexing – if cell hashtags
Demultiplexing by genotypes

Cell-level QC and removal of bad quality cells

Additional cleanup – removal of ambient RNA
Alternative cell calling
Additional cleanup of abs-based data

questions

Collab notebooks

Better expression data

Demultiplexing of cell hashing

- Within Cell Ranger
- Might be done also from Seurat/scanpy environment
- Per barcode signal from all possible hashtags
- Winning hashtag per cell/if too mixed → a multiplet

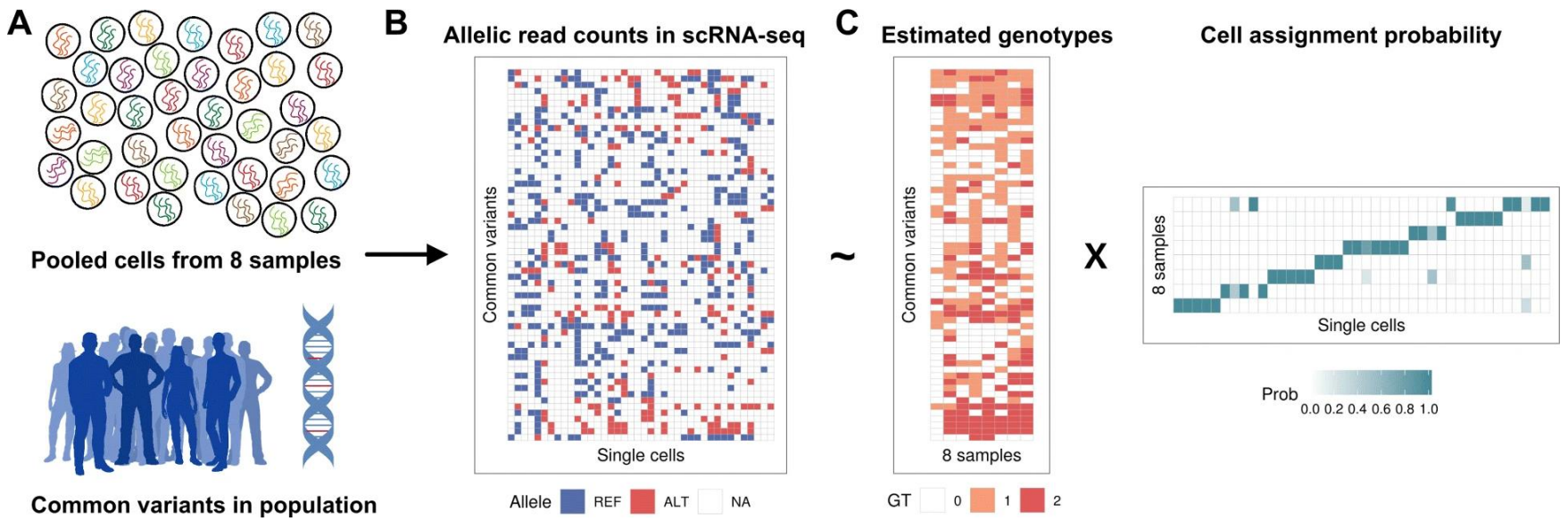
	#tag1	#tag2	#tag3	#tag4
Cell1	**4071**	4	5	1
Cell2	3	**2380**	16	2
Cell3	**1341**	5	21	7
Cell4	5	4	**5434**	8
Cell5	4	2	6	**1474**
Cell6	4	**1203**	3	**1020**

Seurat

Scanpy

CellRanger

Better expression data



Huang et al. *Genome Biol* **20**, 273 (2019). https://doi.org/10.1186/s13059-019-1865-2

Demultiplexing by genotypes

If genotypes available +++

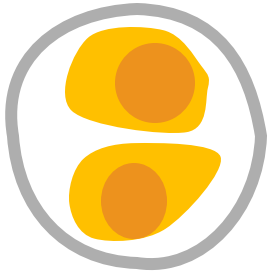
For each barcode: SNP calling from BAM files
Minimal input: a reference list of SNPs (eg 1KG data), expected number of individuals

Demuxafy

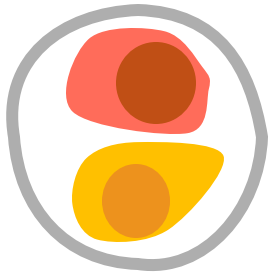
Neavin, D., Senabouth, A., Arora, H. *et al. Demuxafy*: improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. *Genome Biol* **25**, 94 (2024)

Better expression data

Multiplets



← Homotypic and heterotypic →



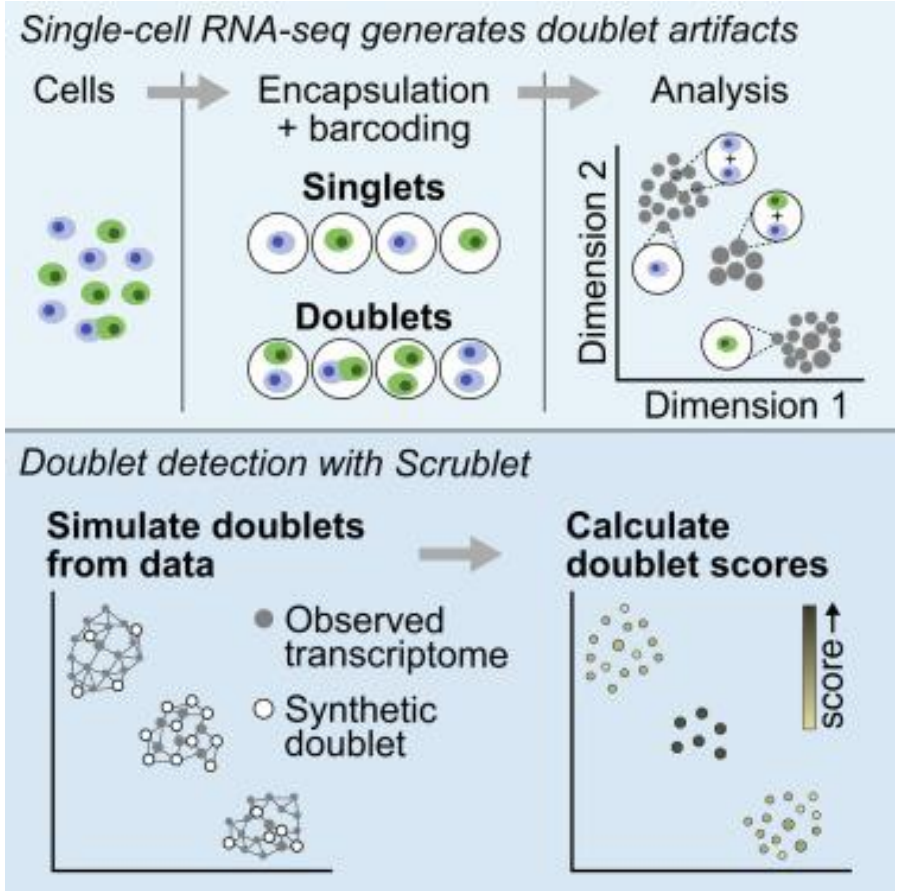
https://www.spreadshirt.co.uk/

How to identify

- Genotypes
- Hashtags
- Extremely high number of expressed genes/UMI count
- Impossible biology
  - TCRs
  - Mutually exclusive proteins/genes CD3, CD19, CD14, CD16

Better expression data

Modelling artificial doublets from the actual experiment data, using this to define doublets



Demuxafy

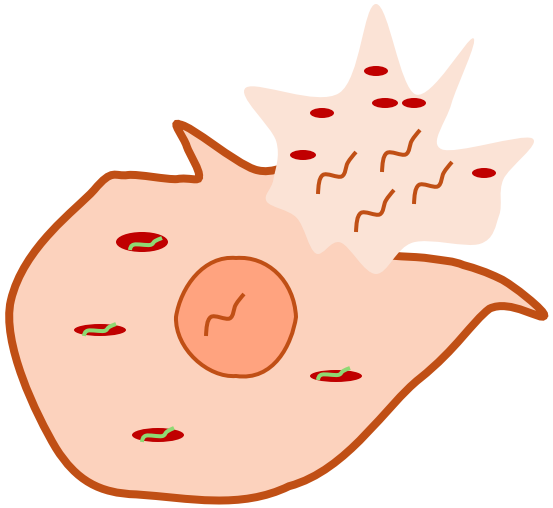
Seurat

Scanpy

Standalone software

Better expression data

Damaged, early burst cells

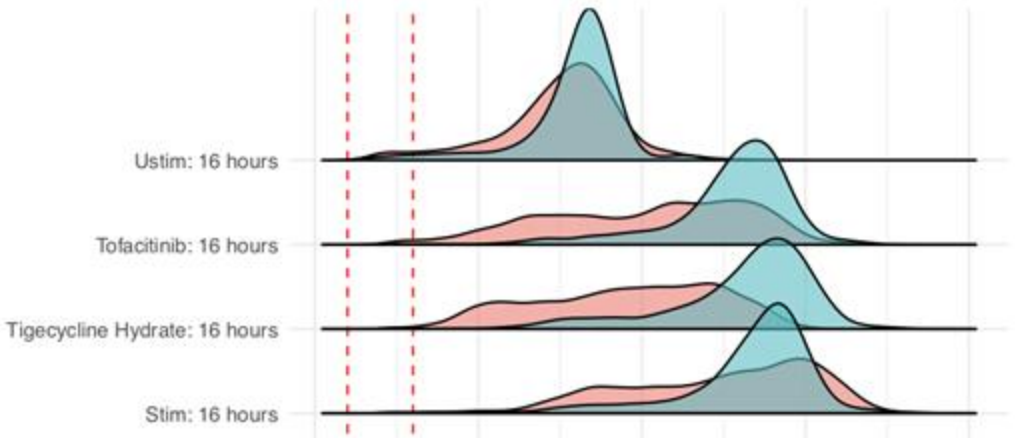


Relatively

- More of mt transcripts
- Less of cytoplasmic transcripts (ribosomal RNA)
- More of nuclear transcripts (PBMCs: lncRNA MALAT)
- Lower number of detected genes
- Lower total number of transcripts

Better expression data

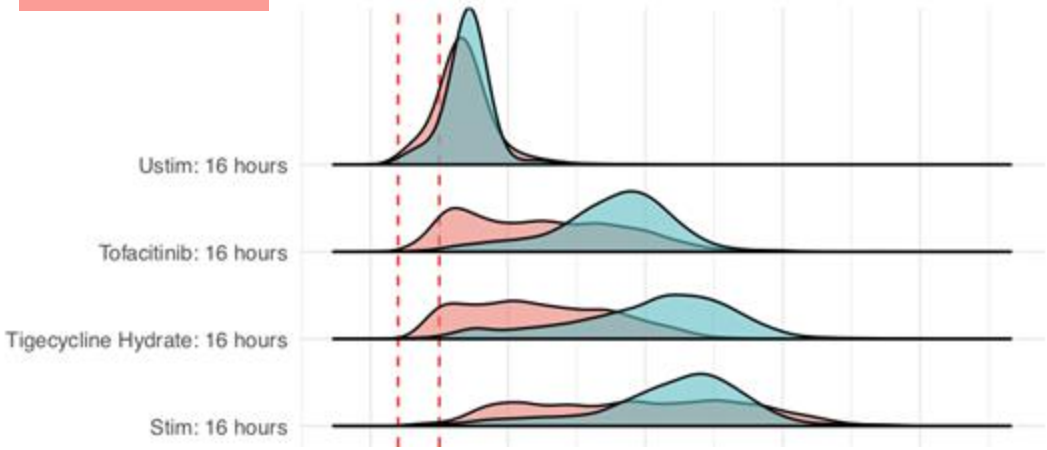
Expression intensity (UMI/cell)

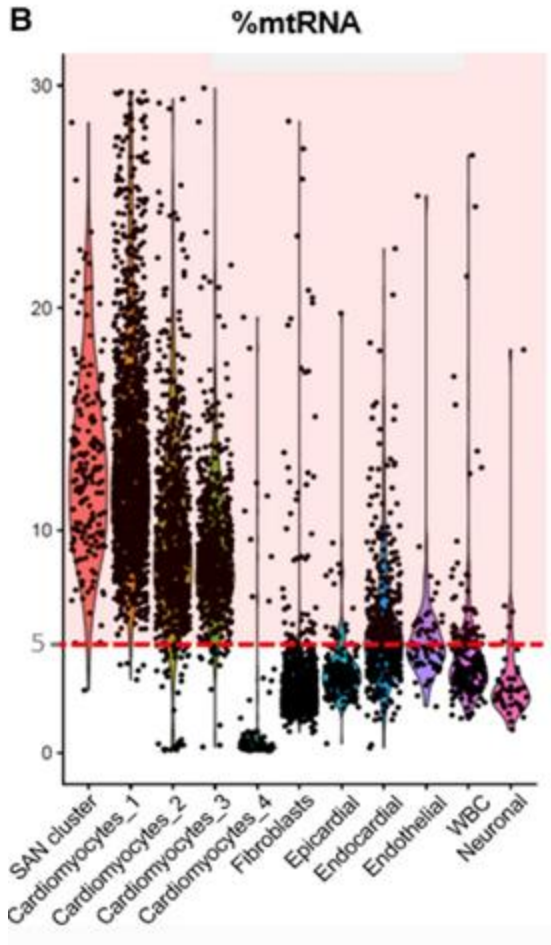


CD4 naive

CD4 mem

Expressed genes (genes/cell)





AM Gallow, 2021

plots: Ziying Ke

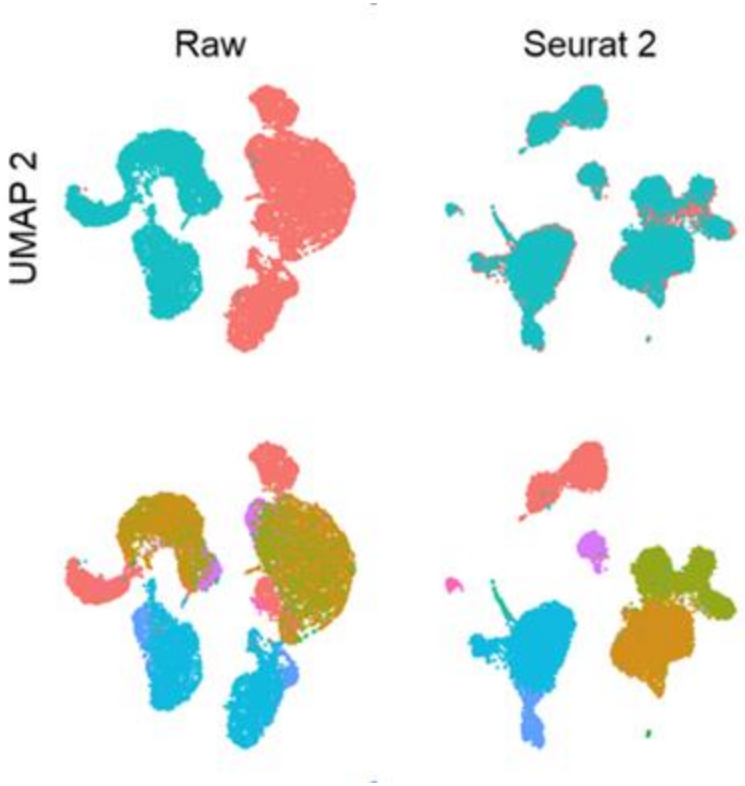
questions

end

Batch effects

VIABILITY!

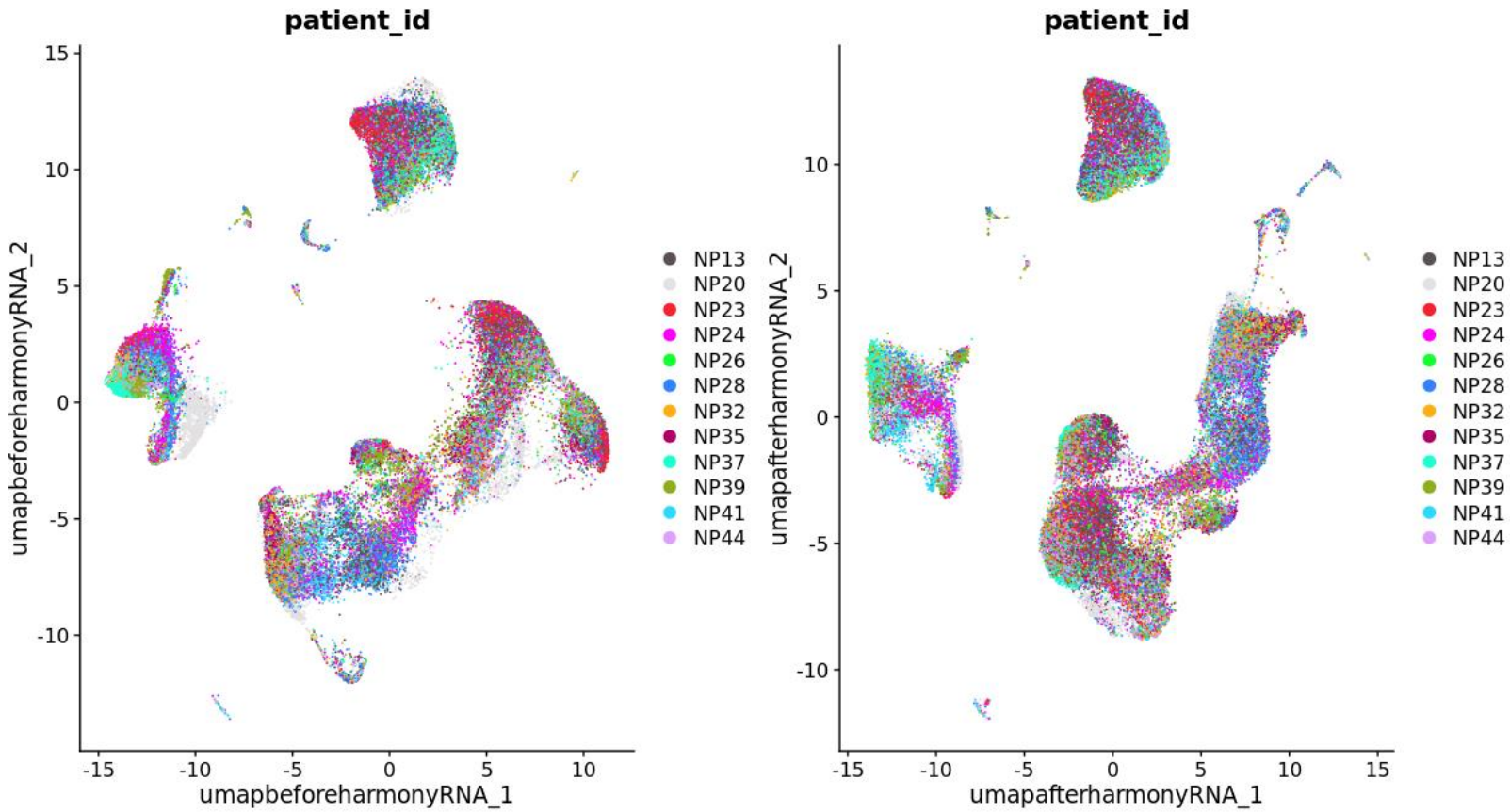
Cells cluster by sample



Cells cluster by celltype

Tran et al. Genome Biology (2020)

Seurat/Scanpy/etc, many methods


patient_id
umapbeforeharmonyRNA_2
umapbeforeharmonyRNA_1
15
10
5
0
-5
-10
-15
-10
-5
0
5
10
NP13
NP20
NP23
NP24
NP26
NP28
NP32
NP35
NP37
NP39
NP41
NP44
patient_id
umapafterharmonyRNA_2
umapafterharmonyRNA_1
10
5
0
-5
-10
-15
-10
-5
0
5
10
15
NP13
NP20
NP23
NP24
NP26
NP28
NP32
NP35
NP37
NP39
NP41
NP44

QC and filtering → post-CellRanger

Goals:
Exclusion of non-cells
Exclusion of dying cells
Exclusion of doublets

Debris - high background
VIABILITY!

Iterative process: QC filter → normalise →UMAP, annotate → redo
QC filter→ normalise → UMAP, annotate

Seurat/Scanpy/etc, software for doublet detection

UMAP, cell annotation, DGE

Goals:
Representation of relationships between cells in 2D
Identification of similar cells
Cell annotation
Downstream analysis

DimRed and clustering are dominated by the genes which vary most between cells → one can use just these genes

Genes to exclude from the analysis: TCRs, BCRs

Seurat/Scanpy/etc