Single cell RNAseq basics

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

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

3 parts – questions after each

Principles

From FASTQ to expression matrices
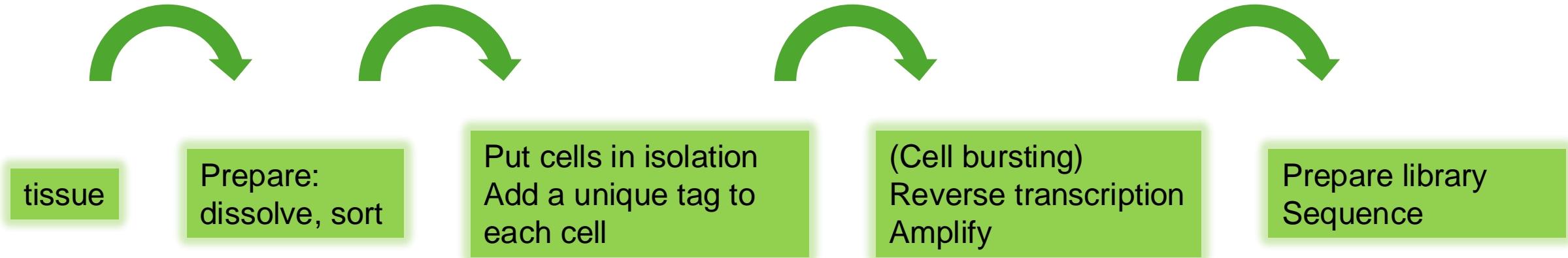
Better expression data

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)

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
+
FFFFFFFFFFFFFFFFFF,FFFFF:FFFFFFFFFFFFFFFFFFF,FFFF,:FFF,:FFFFF,FFFFFFFFFFFFFFFFFFFFFFF: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:
```
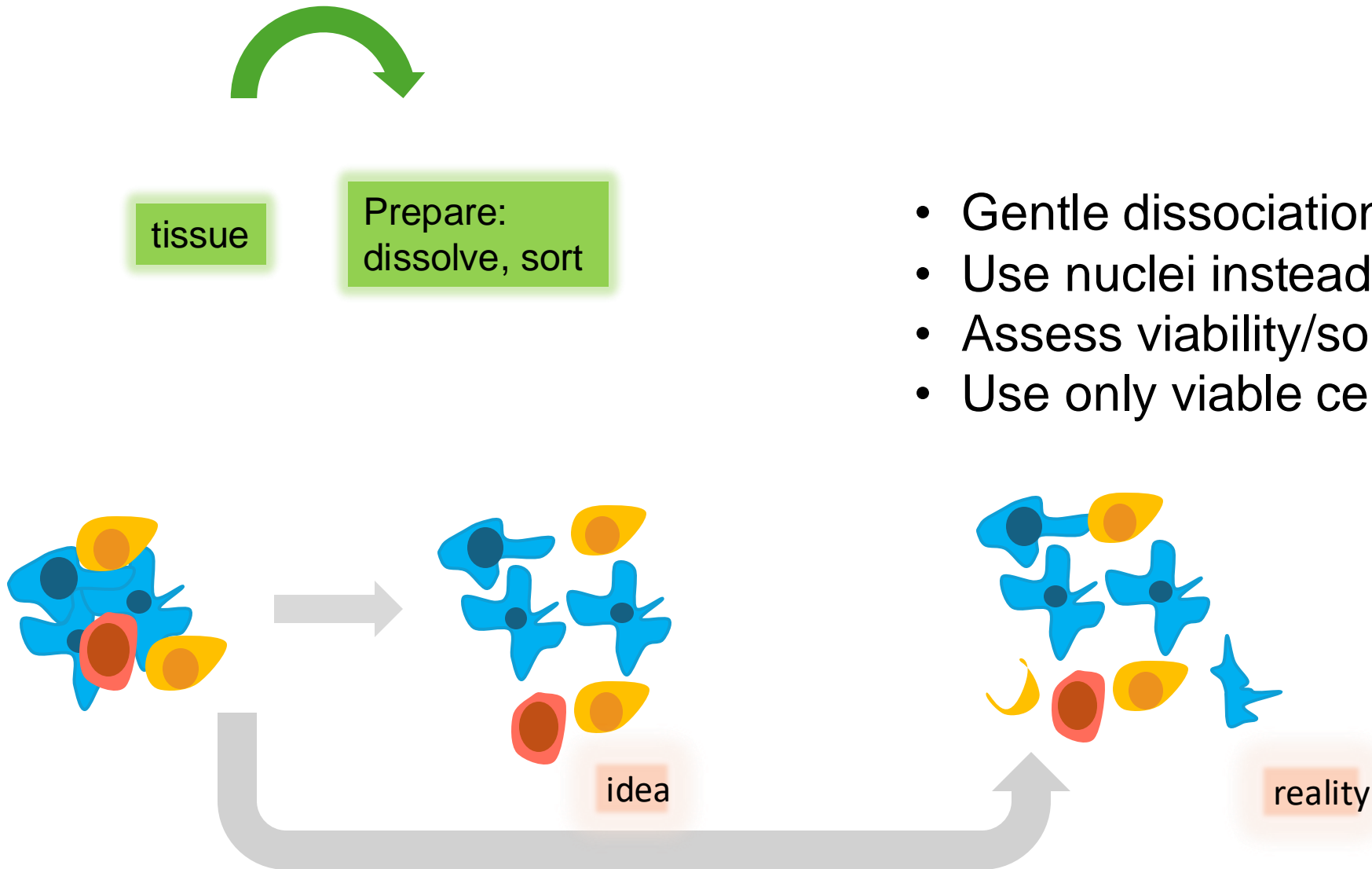
Challenges

- 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
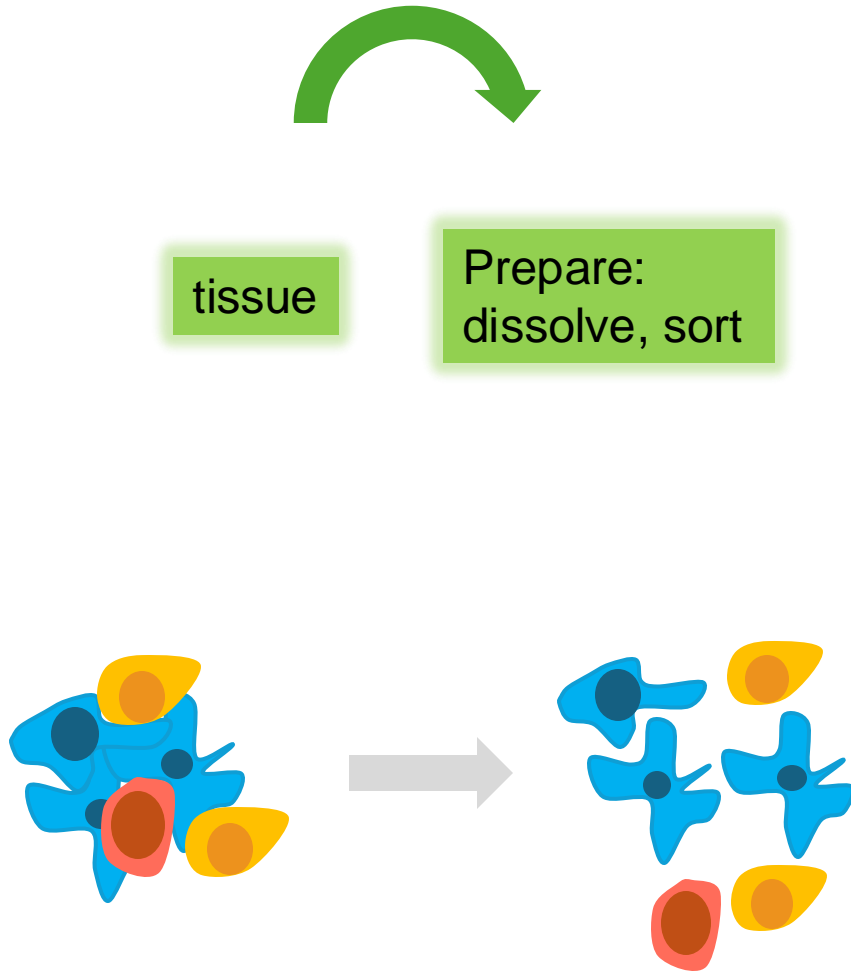
tissue

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

idea

reality

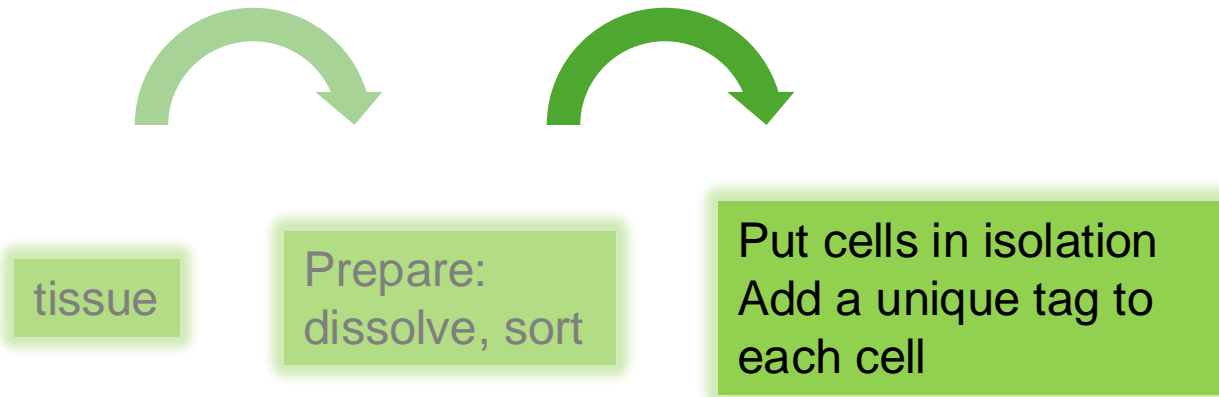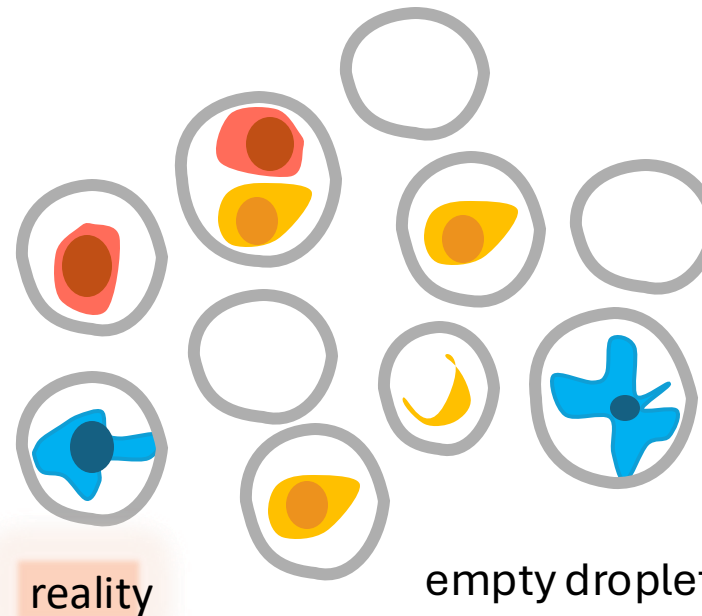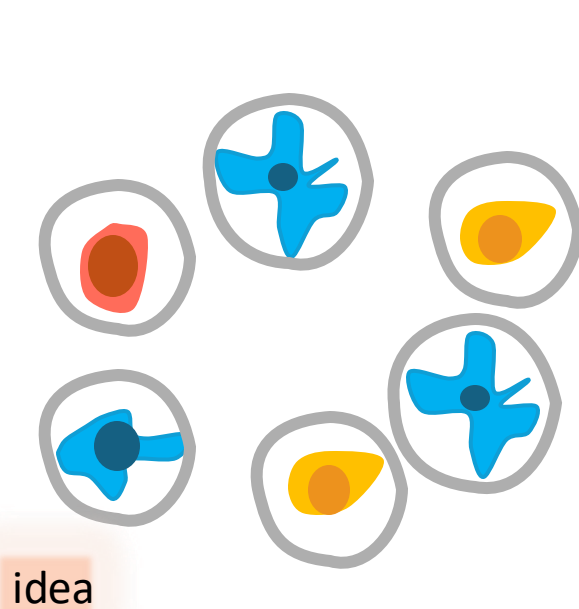And presence of debris is a challenge because...

tissue

Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to
each cell
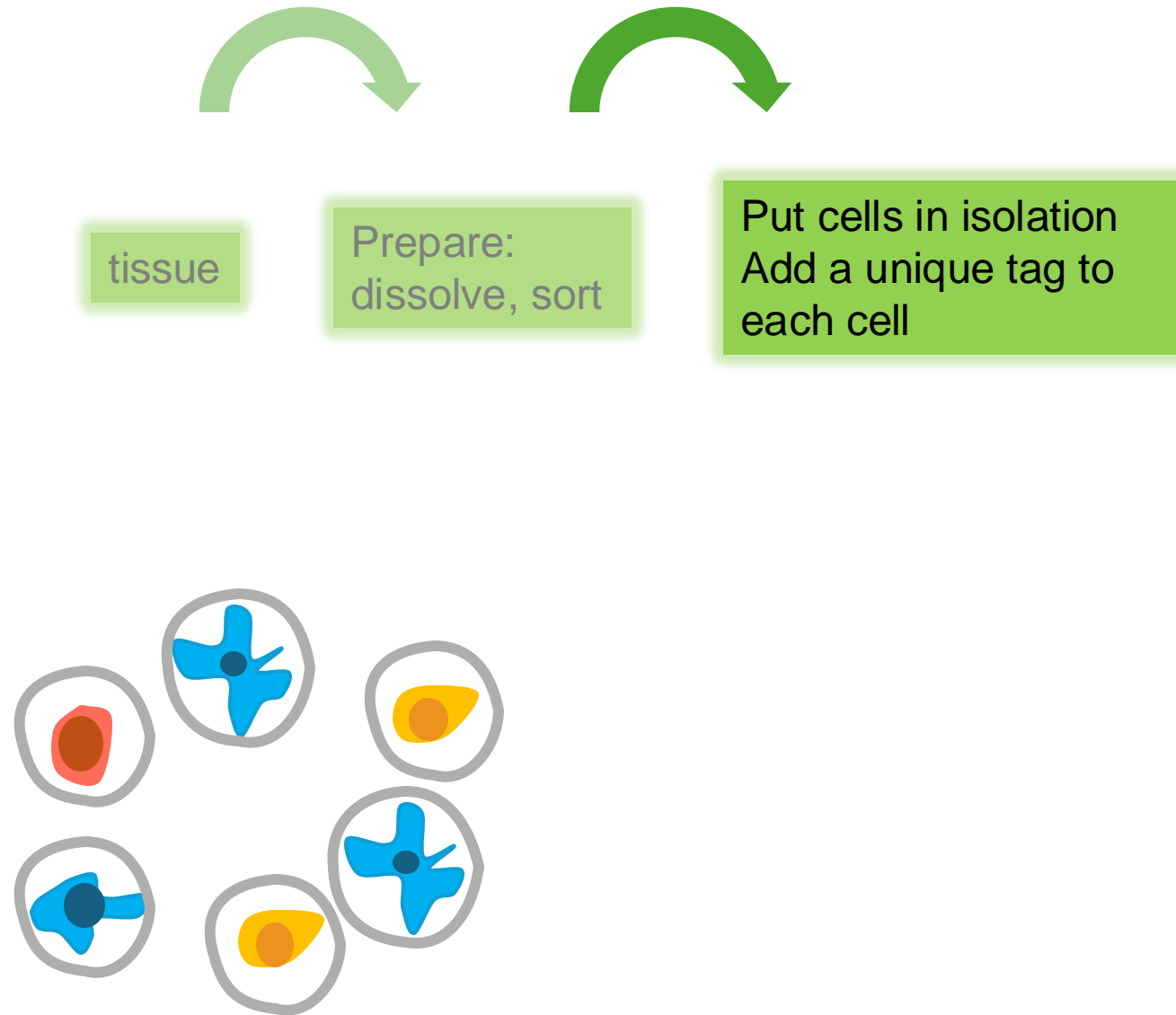
- wells
- microwells (BDRhapsody, Hive)
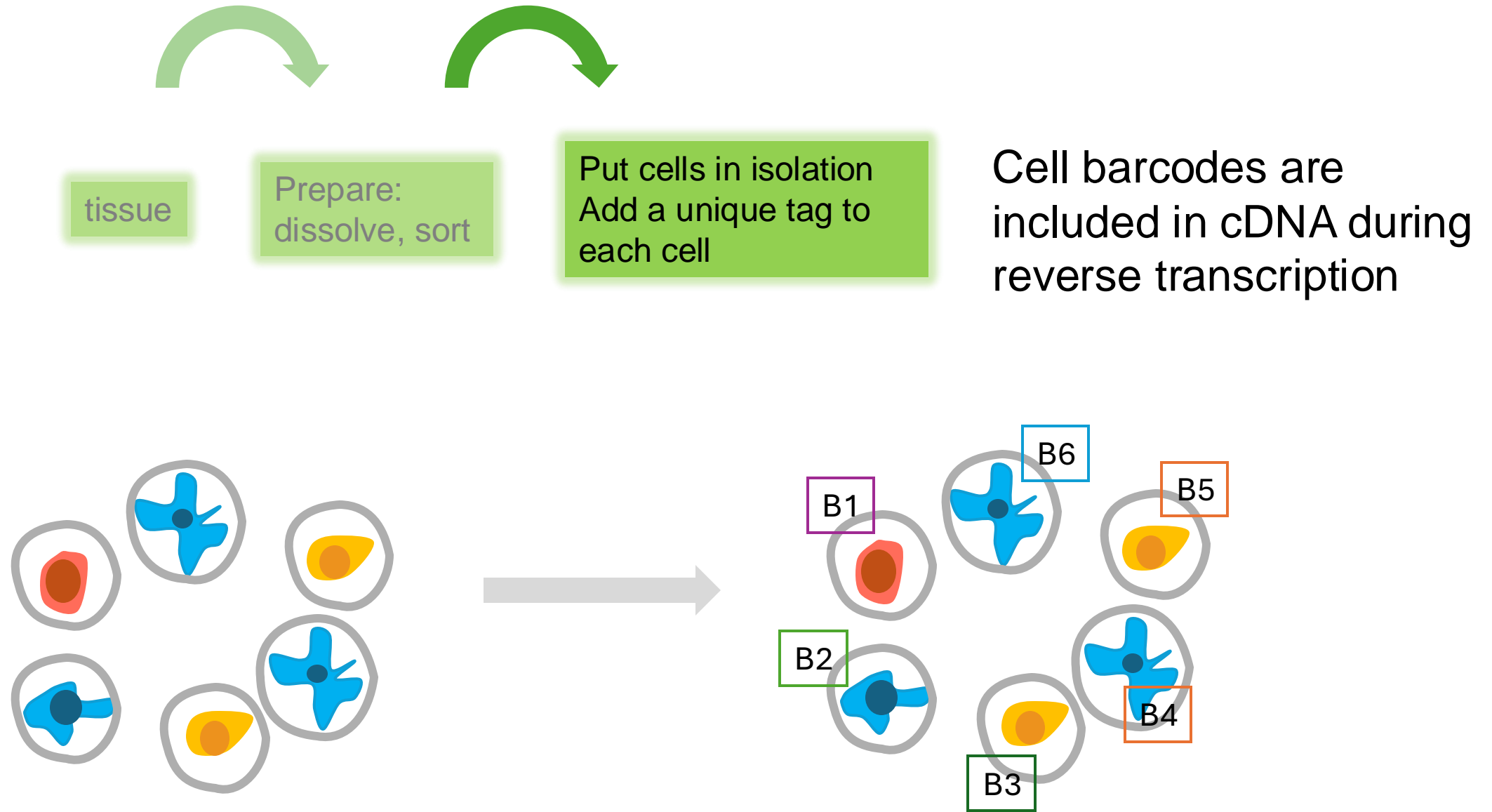- droplets (10x)
- combinatorial tags (PARSE, Scala)

principles

Put cells in isolation
Add a unique tag to
each cell

tissue

Prepare:
dissolve, sort

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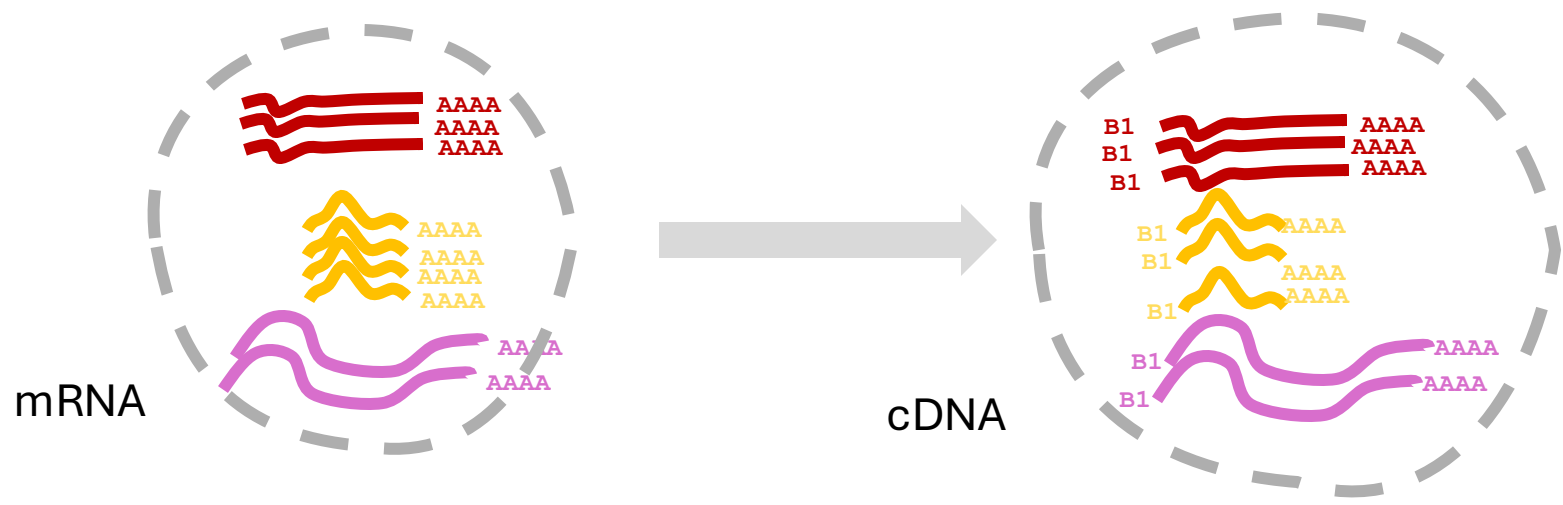- droplets (10x)
- combinatorial tags (PARSE, Scala)

idea

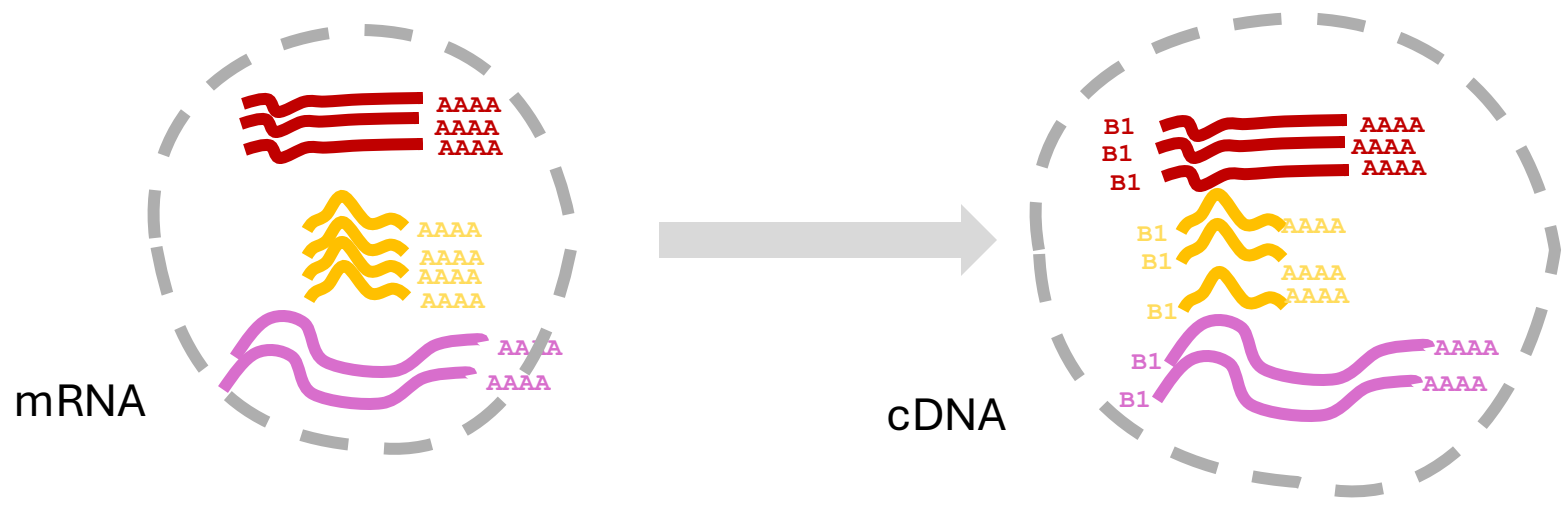reality

empty droplets/wells ← **single cell** → multiplets
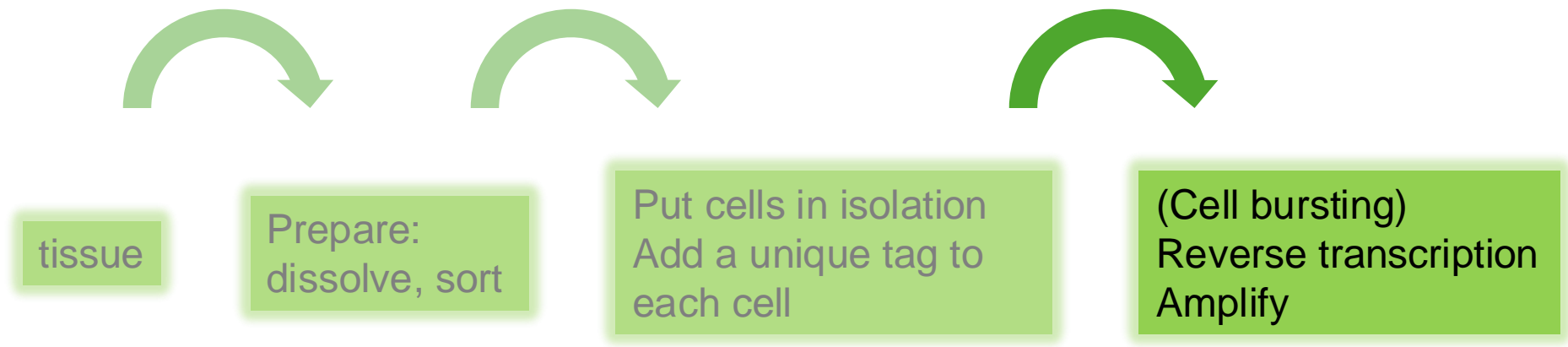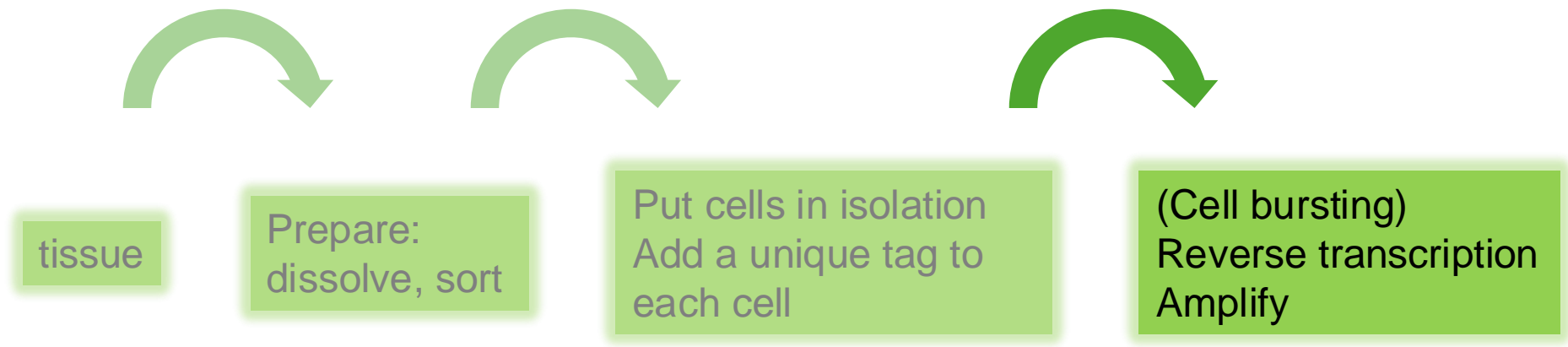
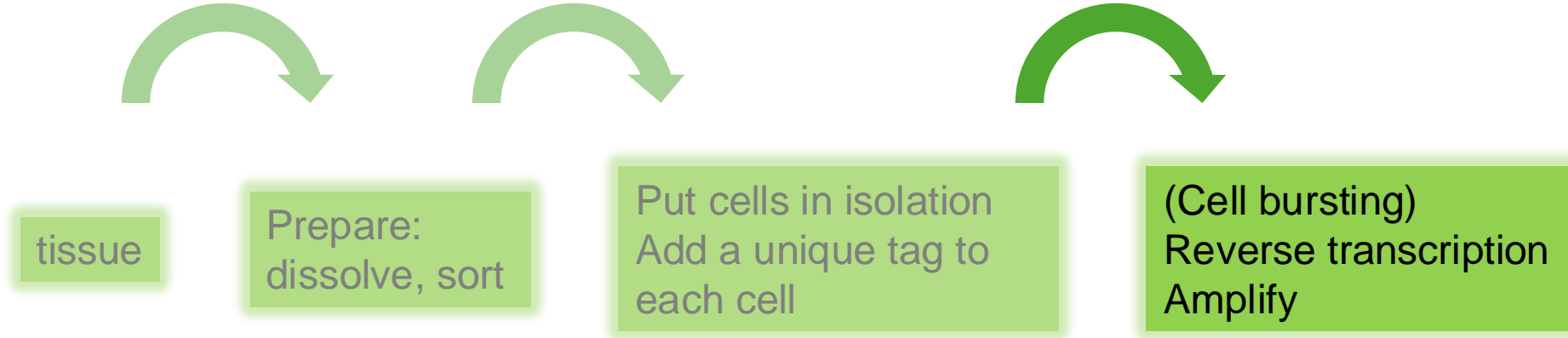tissue → Prepare: dissolve, sort → Put cells in isolation / Add a unique tag to each cell → (Cell bursting) / Reverse transcription / Amplify
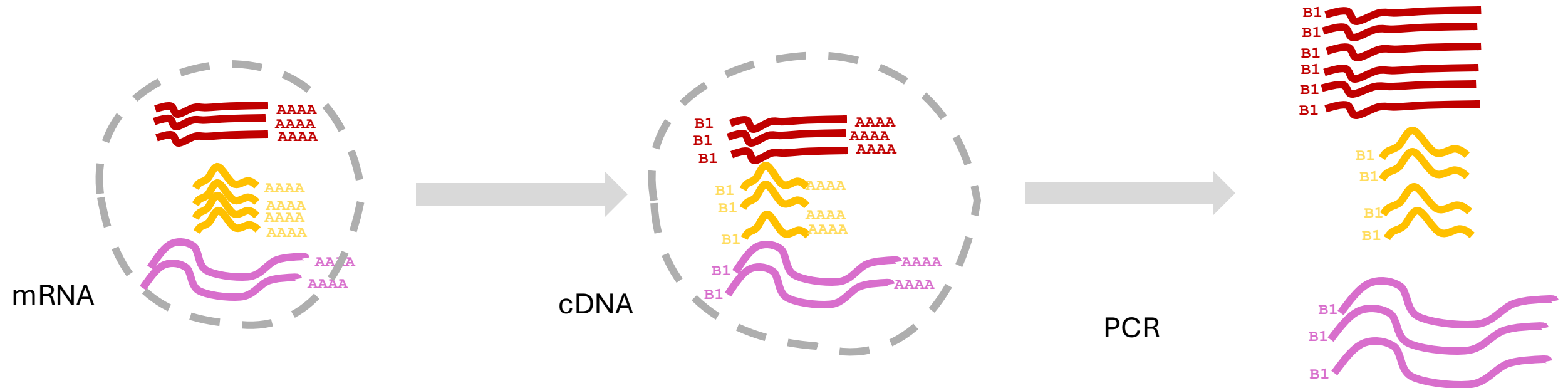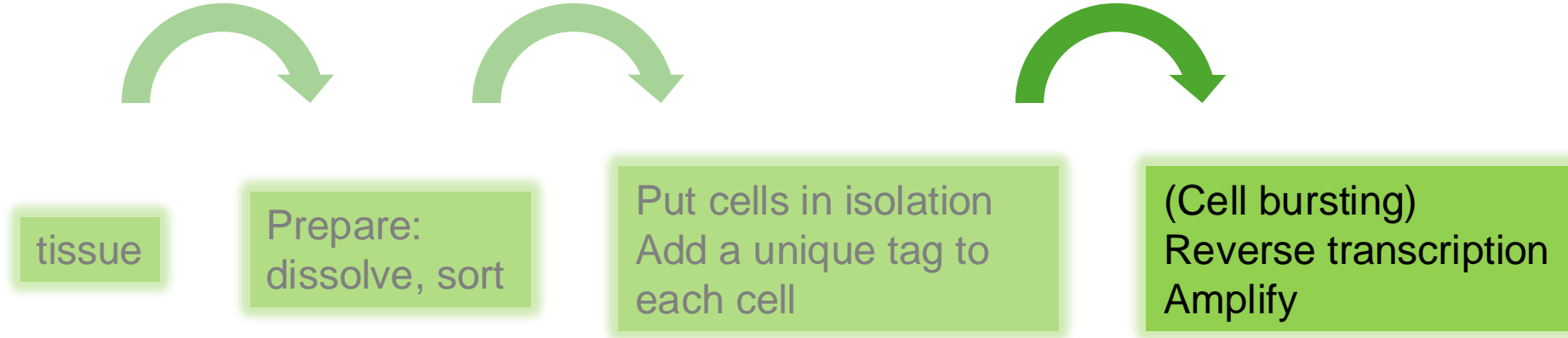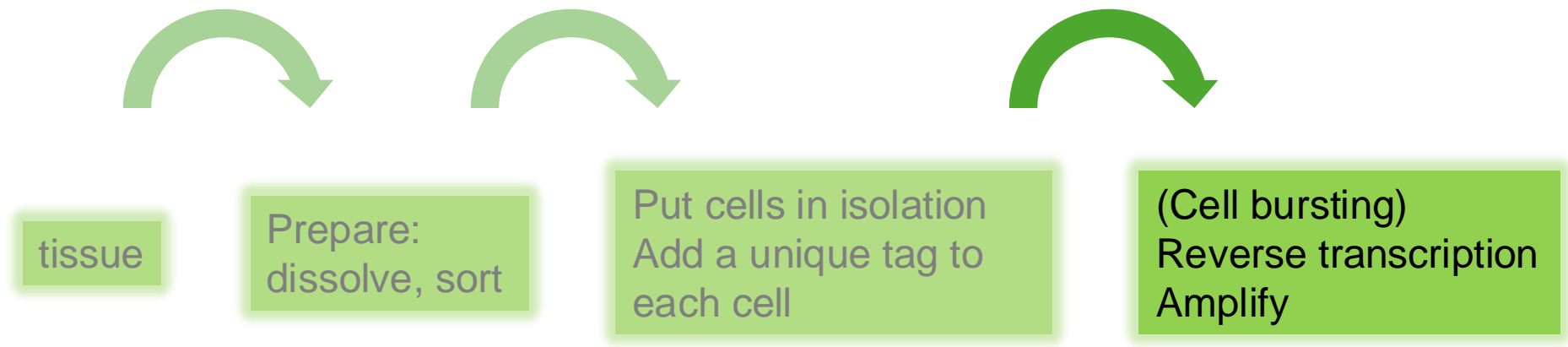
- Drop outs → sparsity of the data
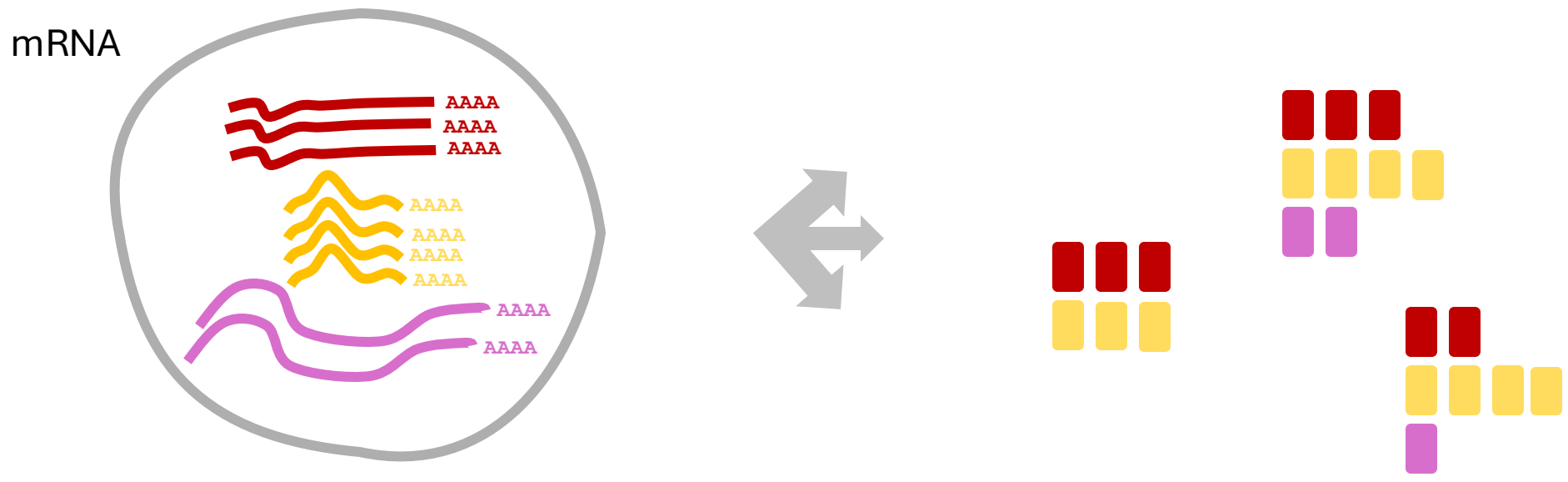
mRNA

tissue

Prepare:
dissolve, sort

Put cells in isolation
Add a unique tag to
each cell

(Cell bursting)
Reverse transcription
Amplify

Prepare library
Sequence

Many platforms

tissue → Prepare: dissolve, sort → Put cells in isolation Add a unique tag to each cell → (Cell bursting) Reverse transcription Amplify → Prepare library Sequence
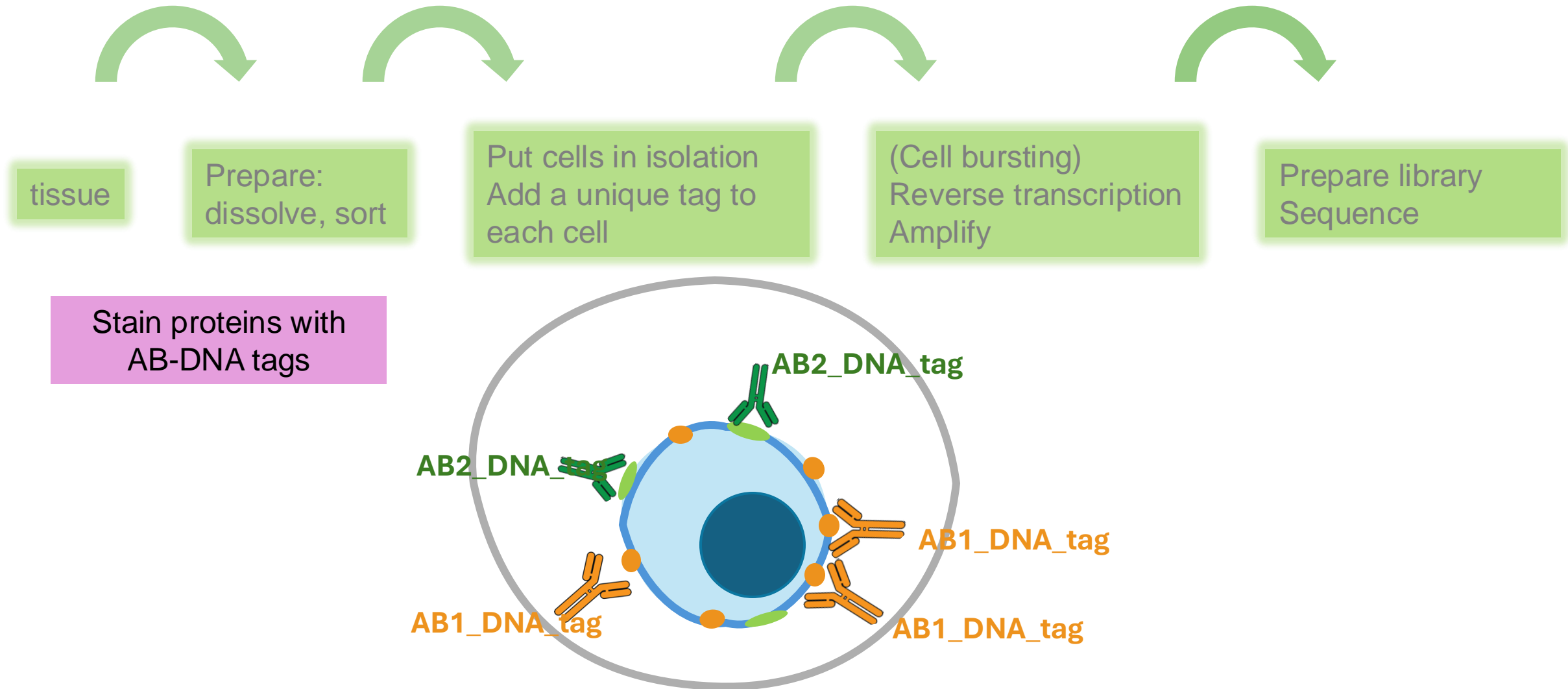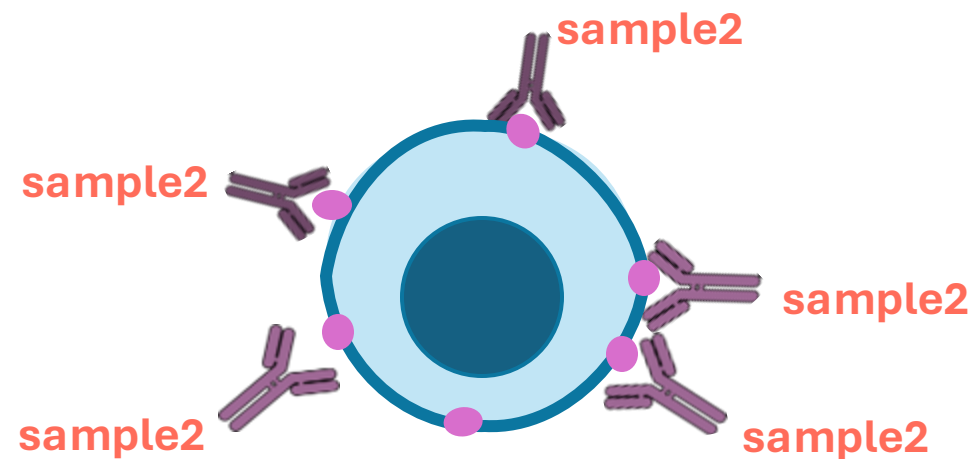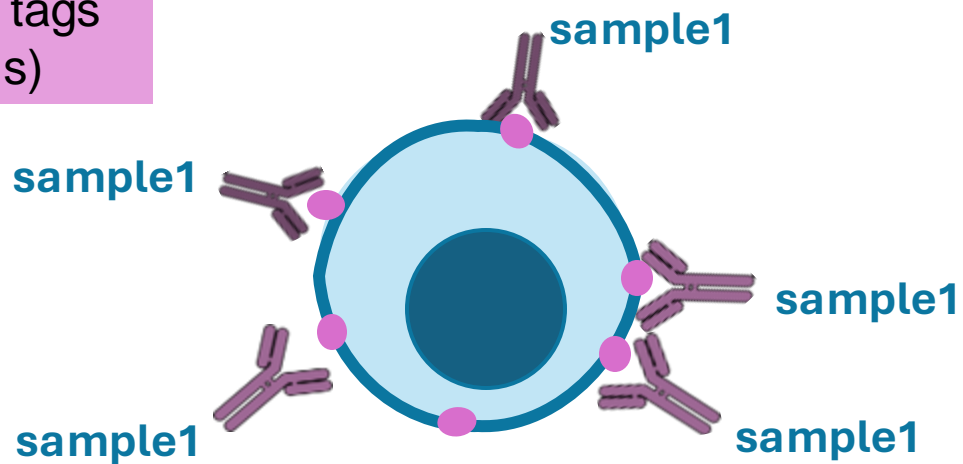
Example of specific platform 10x GENOMICS™

10x Next GEM Technology for Single Cell Partitioning
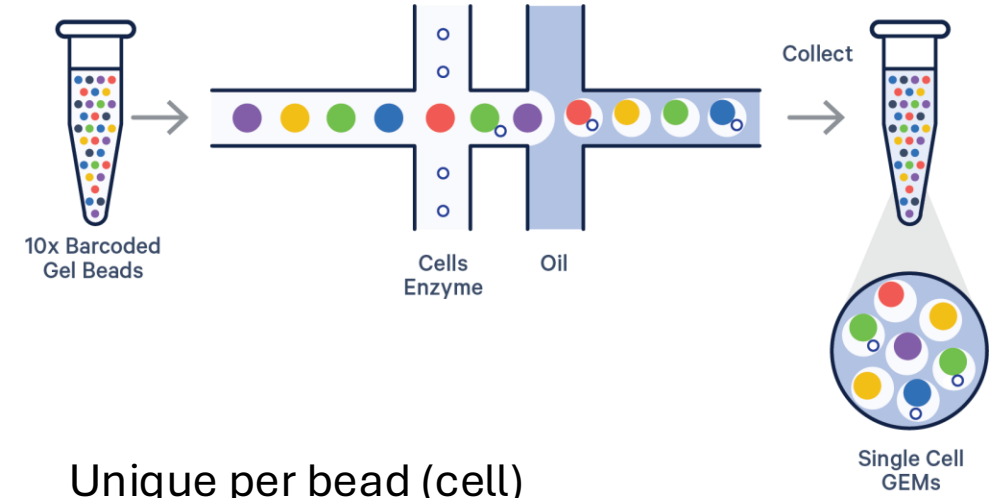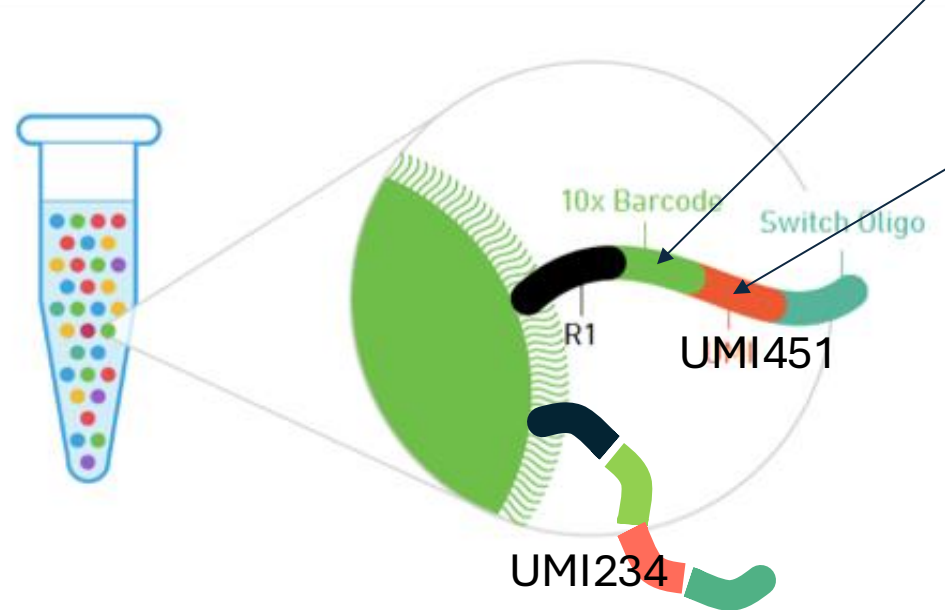
10x Barcoded Gel Beads

Cells Enzyme

Oil

Collect

Single Cell GEMs

Unique per bead (cell)

Unique per oligotail: UMI

10x Barcode

Switch Oligo

R1

UMI451

UMI234

5' GEX + CITEseq

Unique per bead (cell)

Unique per oligotail: UMI

5' GEX

UMI451

UMI234

Inside individual GEMs

Pooled cDNA processed in bulk

GEX
library
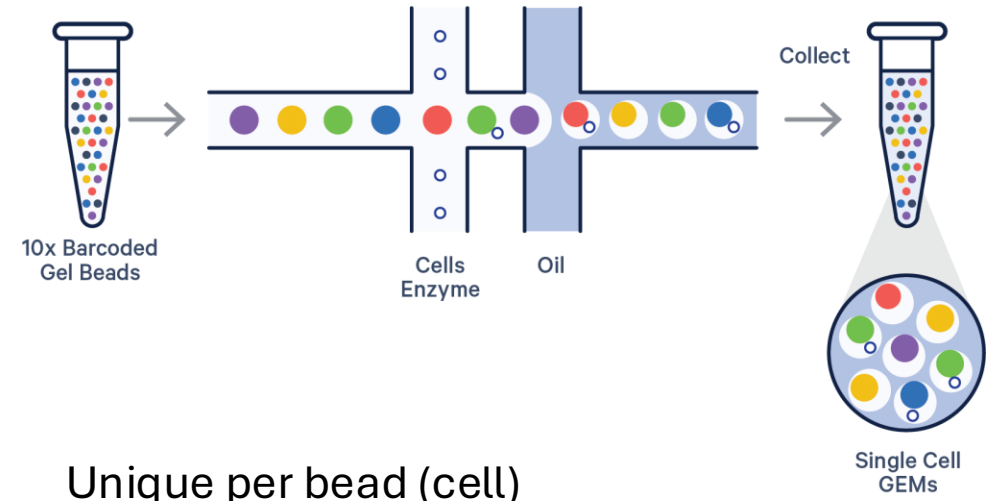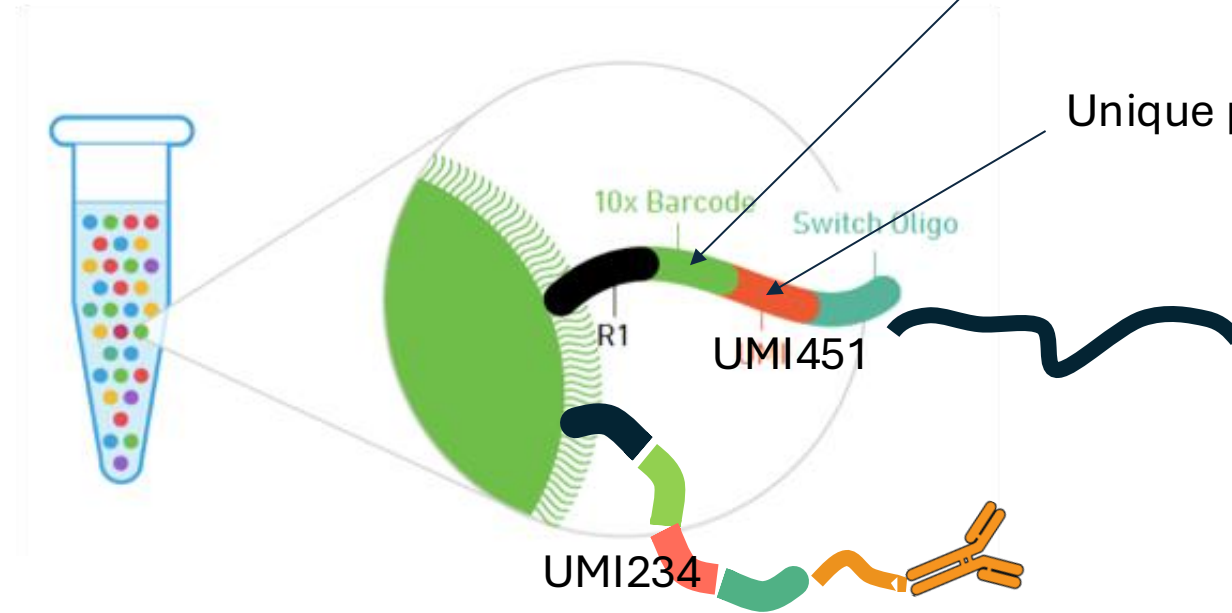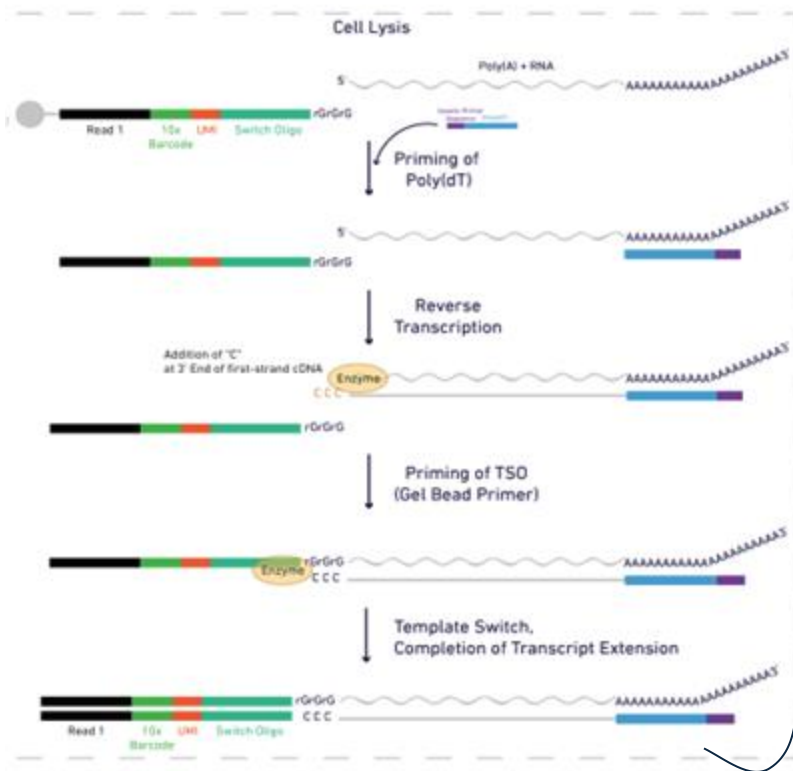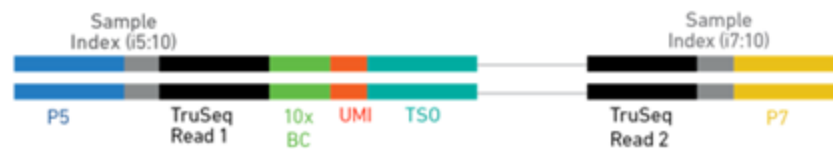
Inside individual GEMs
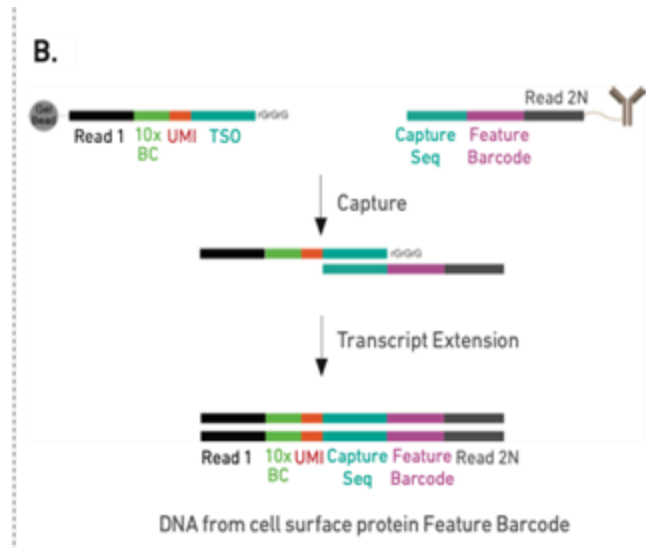
Pooled cDNA processed in bulk

CITEseq library
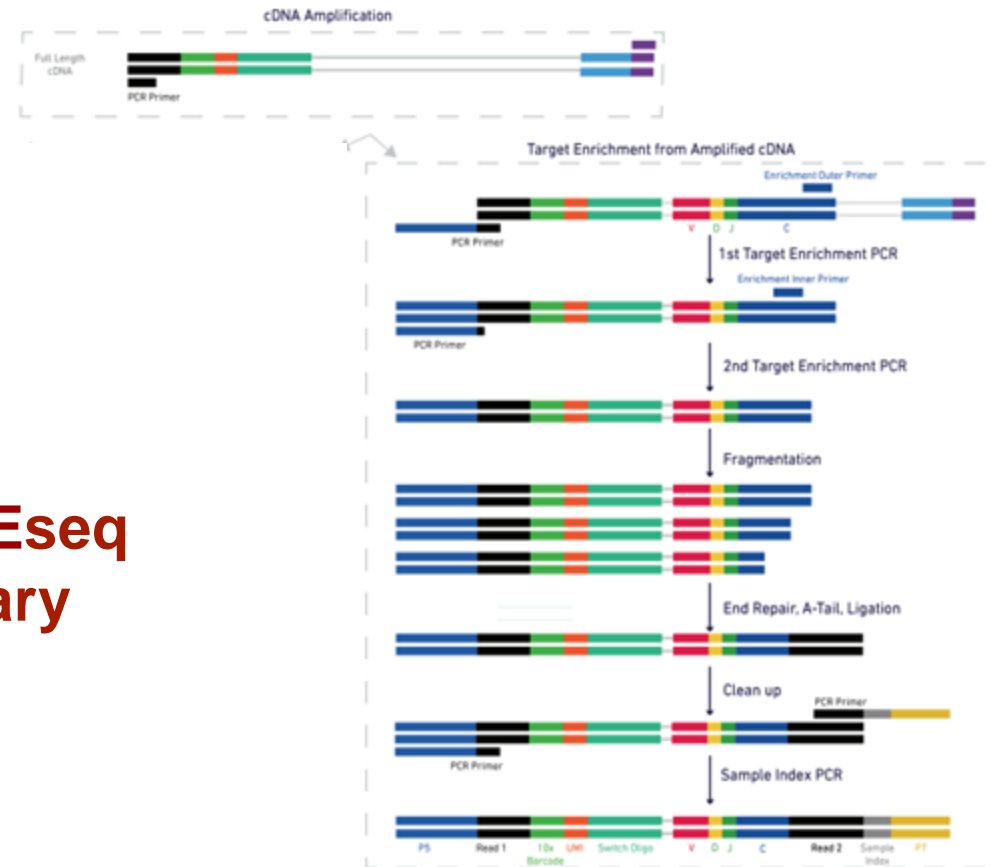
VDJ library

How deep to sequence? ← Depends on the technology & biological questions

scRNAseq

- 10-15K reads per cell
- 50K reads per cell
- Saturation: the bottleneck is the amount of starting RNA
- Trade offs: 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

Pool as much as possible, as soon as possible

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

Use samples' genotypes

treatment 1

treatment 2

sample1

sample1

sample1

sample1

sample1

sample2

sample2

sample2

sample2

sample2
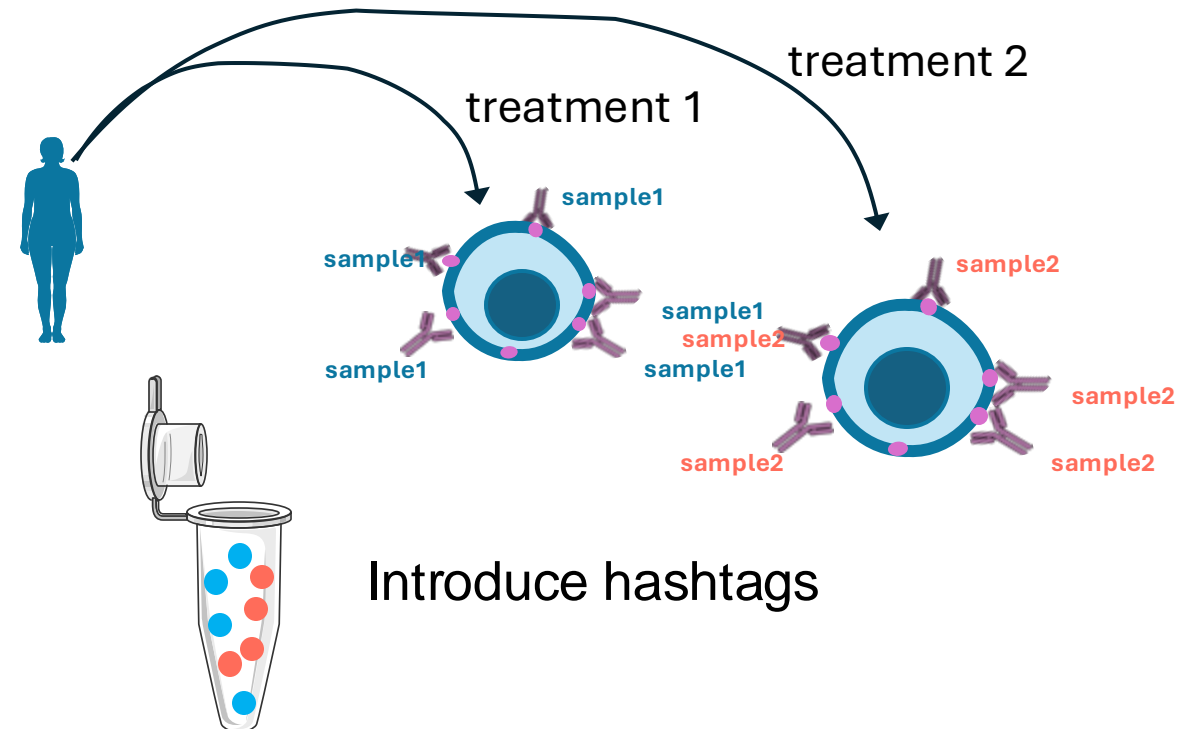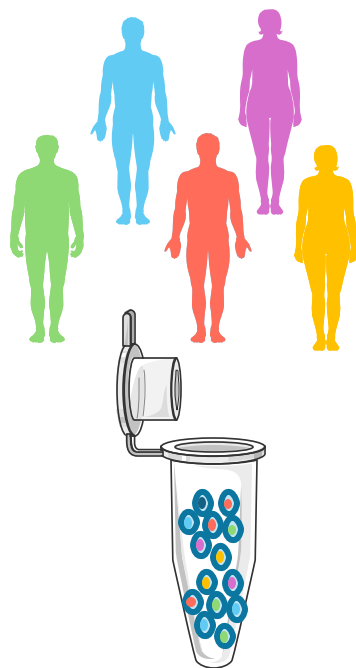
Introduce hashtags

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
+
FFFFFFFFFFFFFFFFFF,FFFFF:FFFFFFFFFFFFFFFFFF,FFFF,:FFF,:FFFFF,FFFFFFFFFFFFFFFFFFFFFF: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:
```
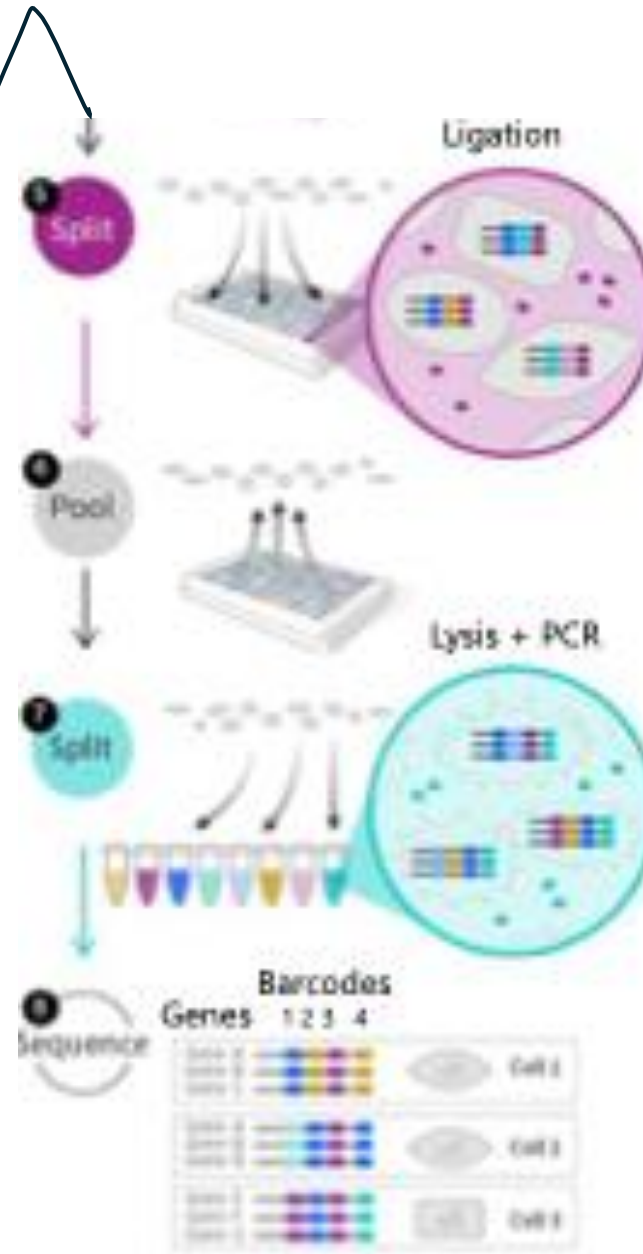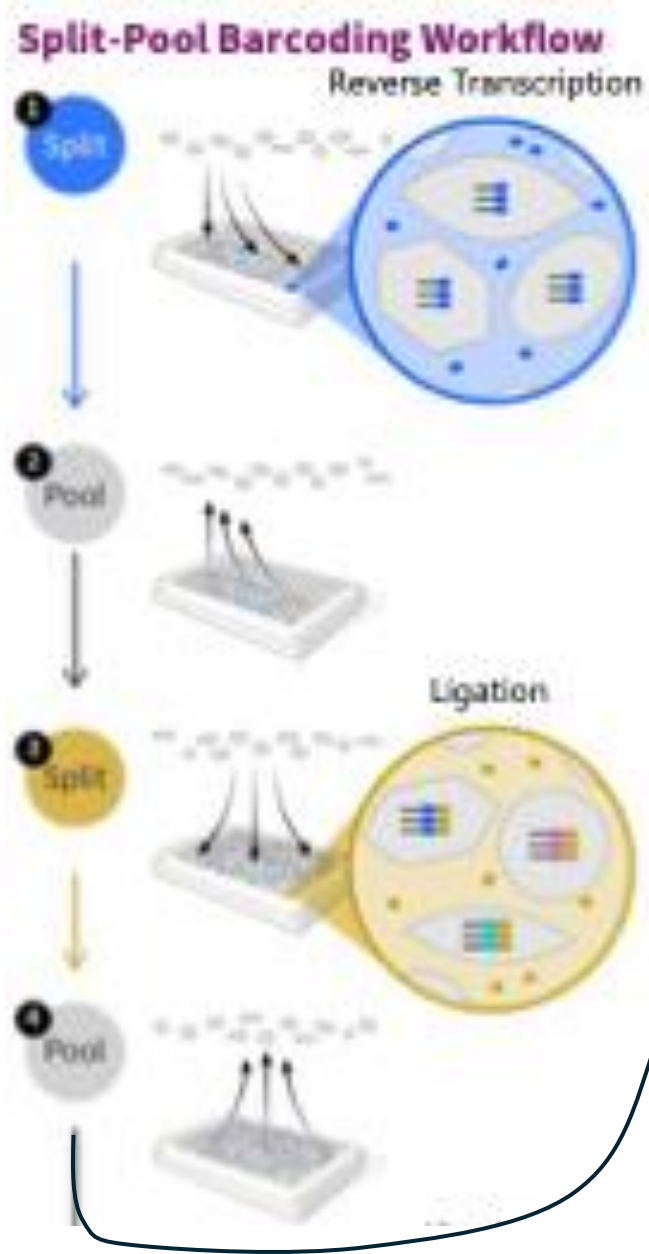
- Done by proprietary software/free alternatives available
- Basis for any downstream analysis
- Some steps might be improved with additional community-developed software

	cell1	cell2	cell3
gene1			
gene2			
gene3			

	cell1	cell2	cell3
abs1			
abs2			
abs3			

	cell1	cell2	cell3
#1			
#2			
#3			

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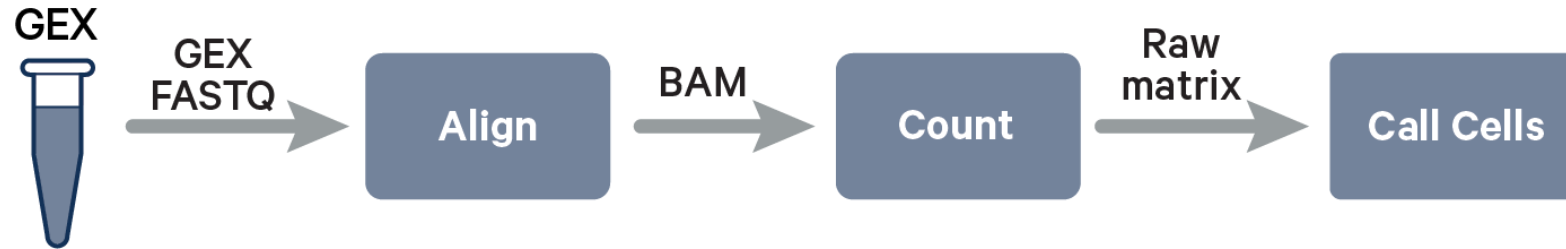
Is barcode a cell or an empty droplet?
→ Order barcodes by UMI counts
→ Barcode rank plot – identify the cutoff

	barc1	barc2	barc3
gene1			
gene2			
gene3			
Total			

reality

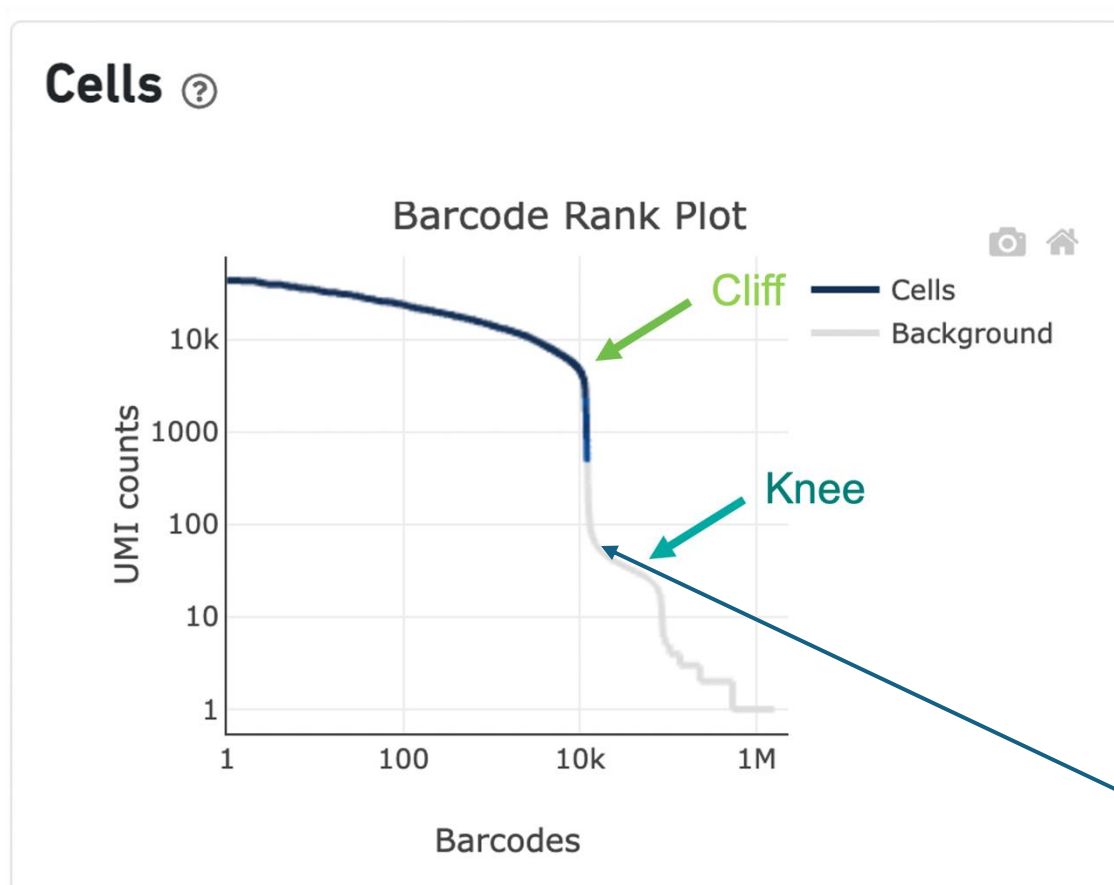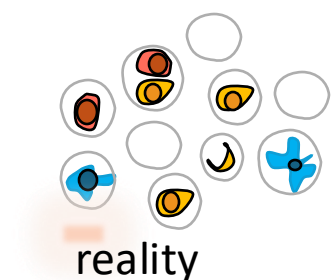Cells ⓘ

Barcode Rank Plot

Cliff — Cells
— Background

Knee

Ambient RNA

Cell Ranger

From FASTQ to expression matrices
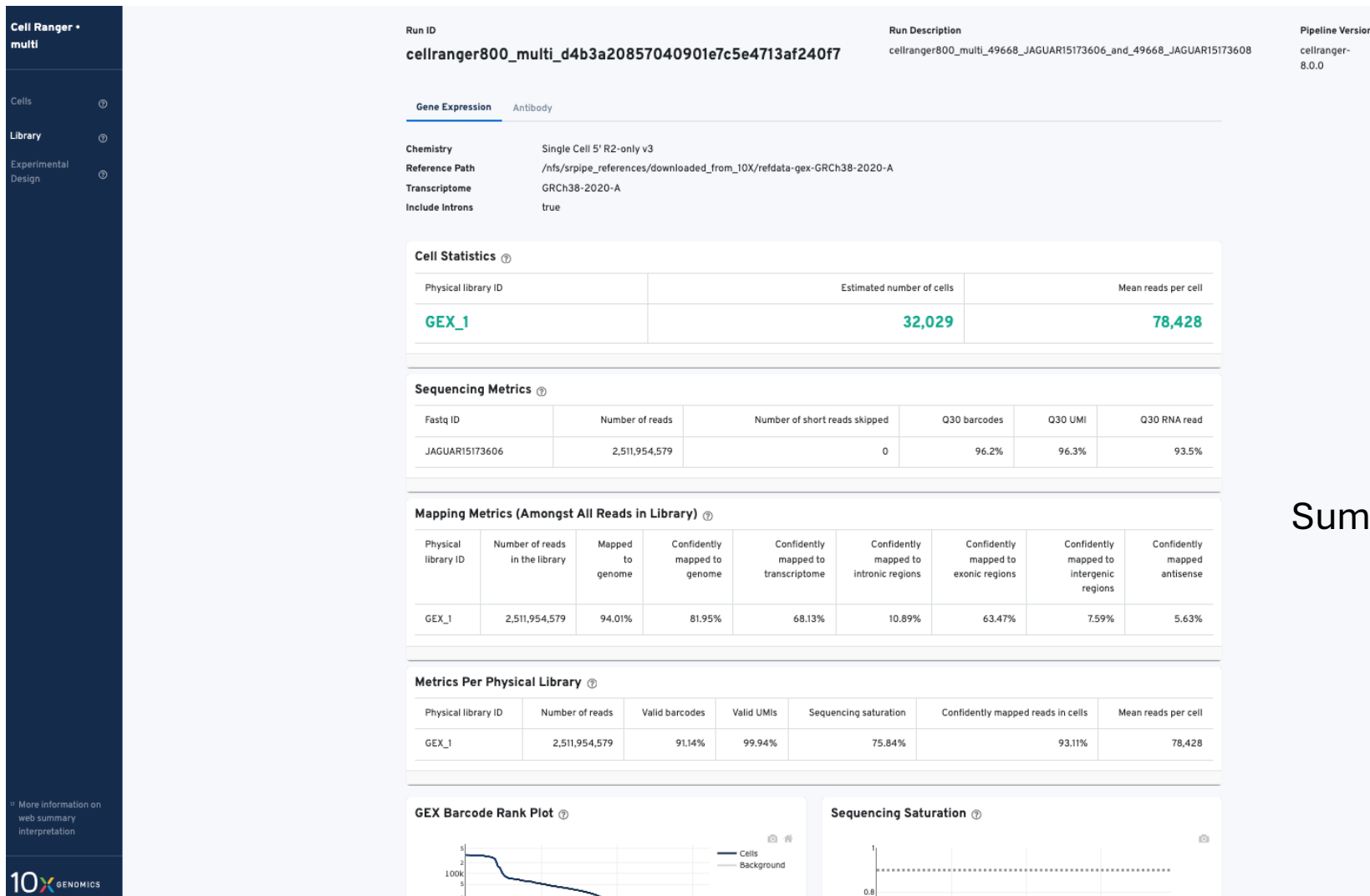
Summary of reads/transcripts/cells…

	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

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

A. Pooled cells from 8 samples / Common variants in population
B. Allelic read counts in scRNA-seq — Common variants / Single cells — Allele REF ALT NA
C. Estimated genotypes — Common variants / 8 samples — GT 0 1 2; Cell assignment probability — 8 samples / Single cells — Prob 0.0 0.2 0.4 0.6 0.8 1.0
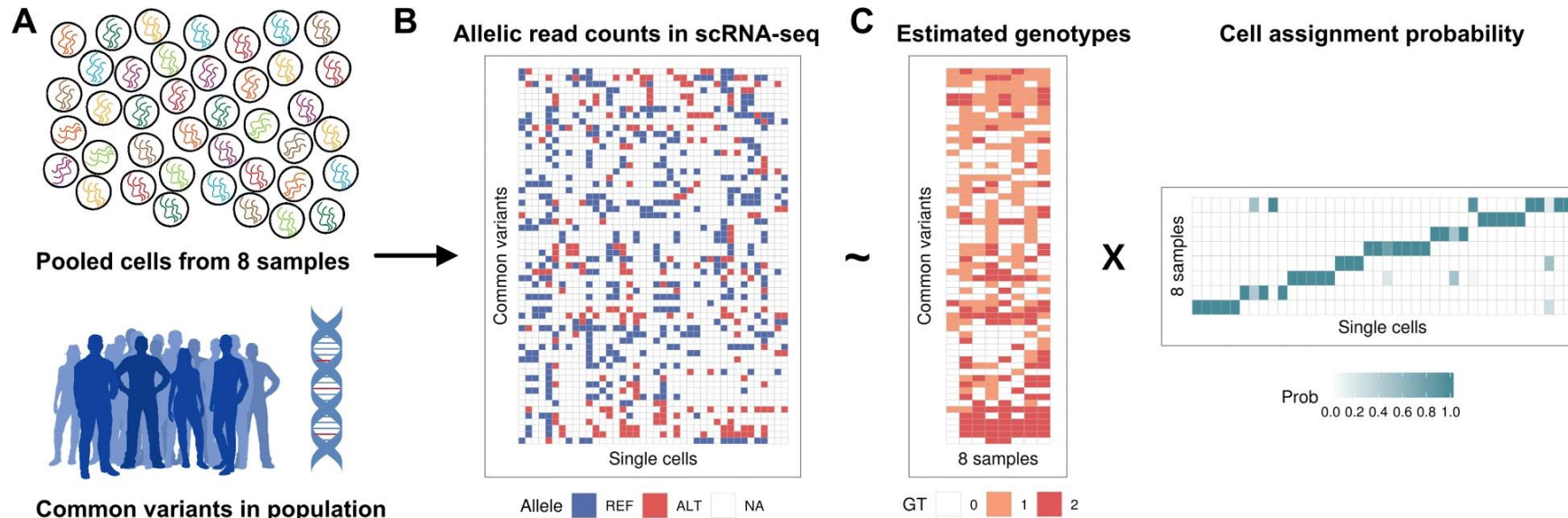
Demultiplexing by genotypes

If genotypes available +++

Demuxafy

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

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)
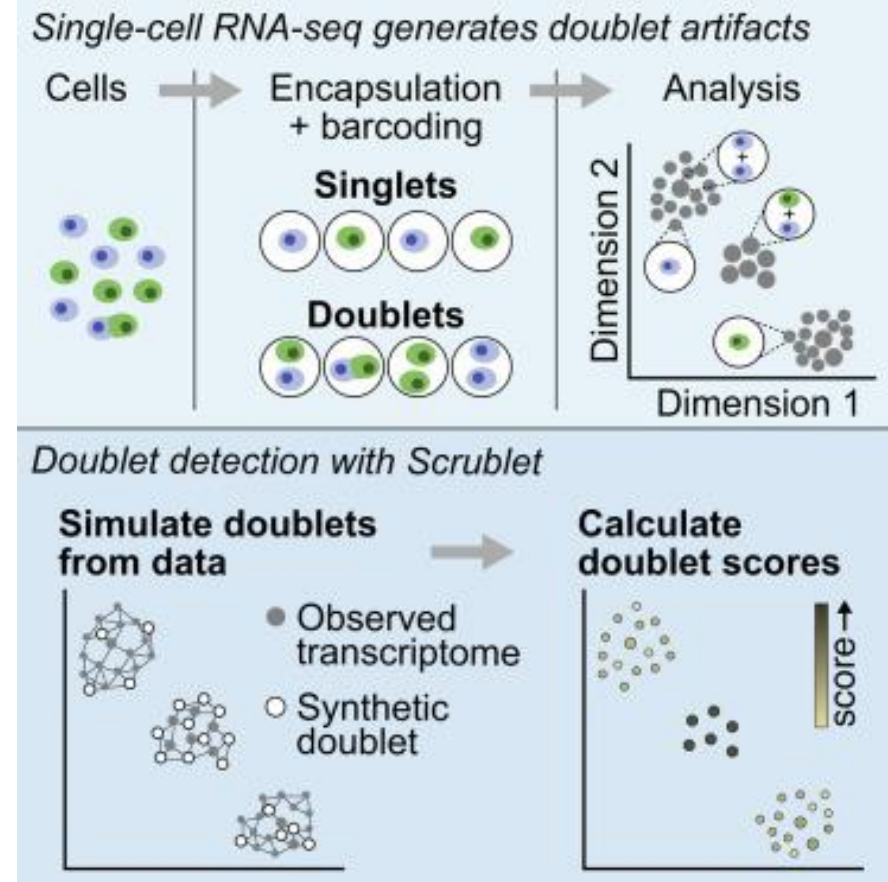
Multiplets

← Homotypic and heterotypic →
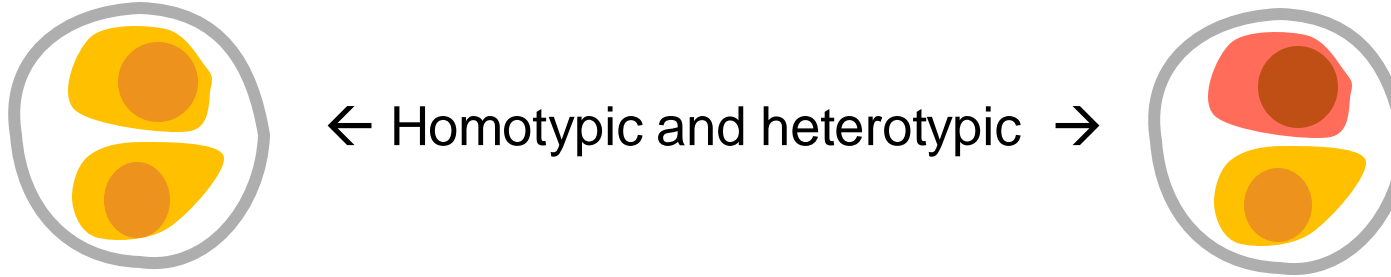
How to identify

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

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

Demuxafy

Method I recommend: scdblFinder
Germain P, Lun A, Garcia Meixide C, Macnair W, Robinson M (2022). "Doublet identification in single-cell sequencing data using scDblFinder."
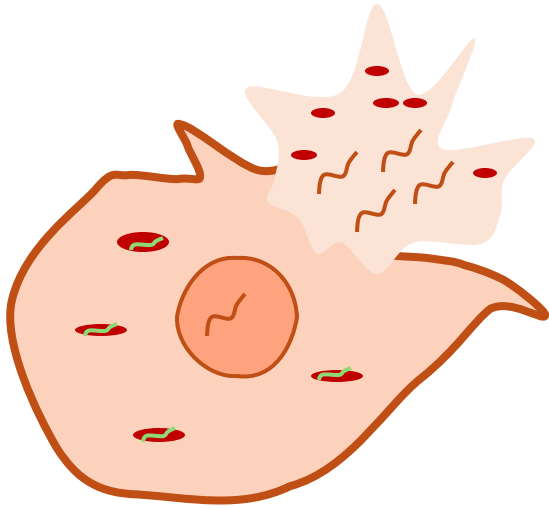*f1000research*. doi:10.12688/f1000research.73600.2.

Seurat
Scanpy
Standalone software

Better expression data
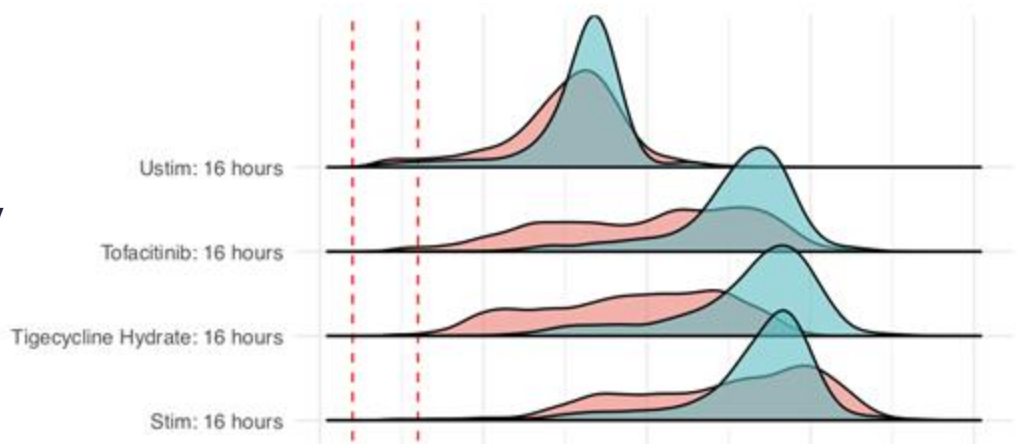
Damaged, early burst cells

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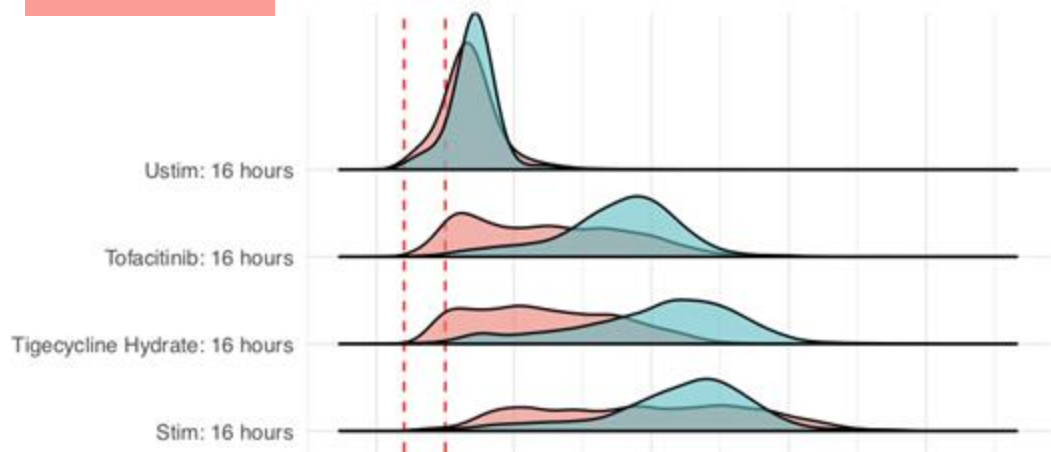- Lower total number of transcripts

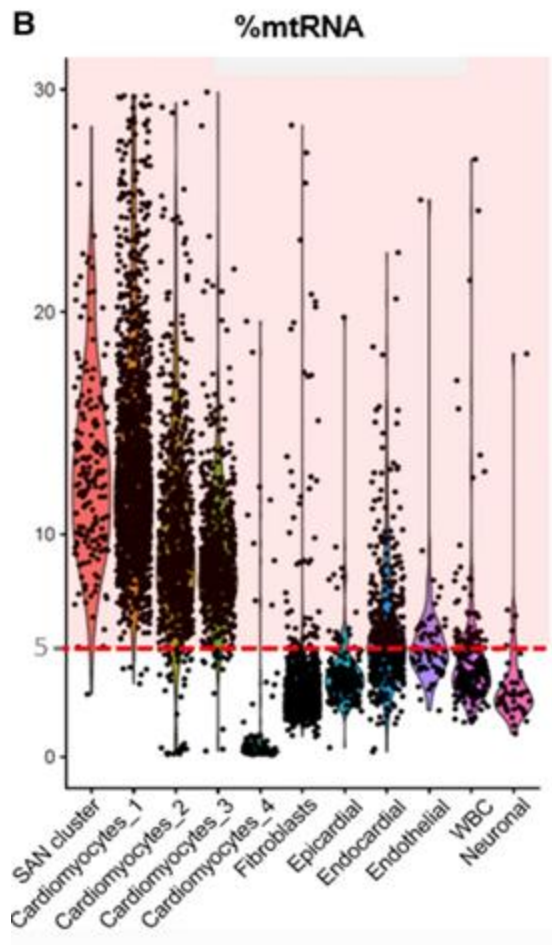Better expression data

Expression intensity
(UMI/cell)

CD4 naive

CD4 mem
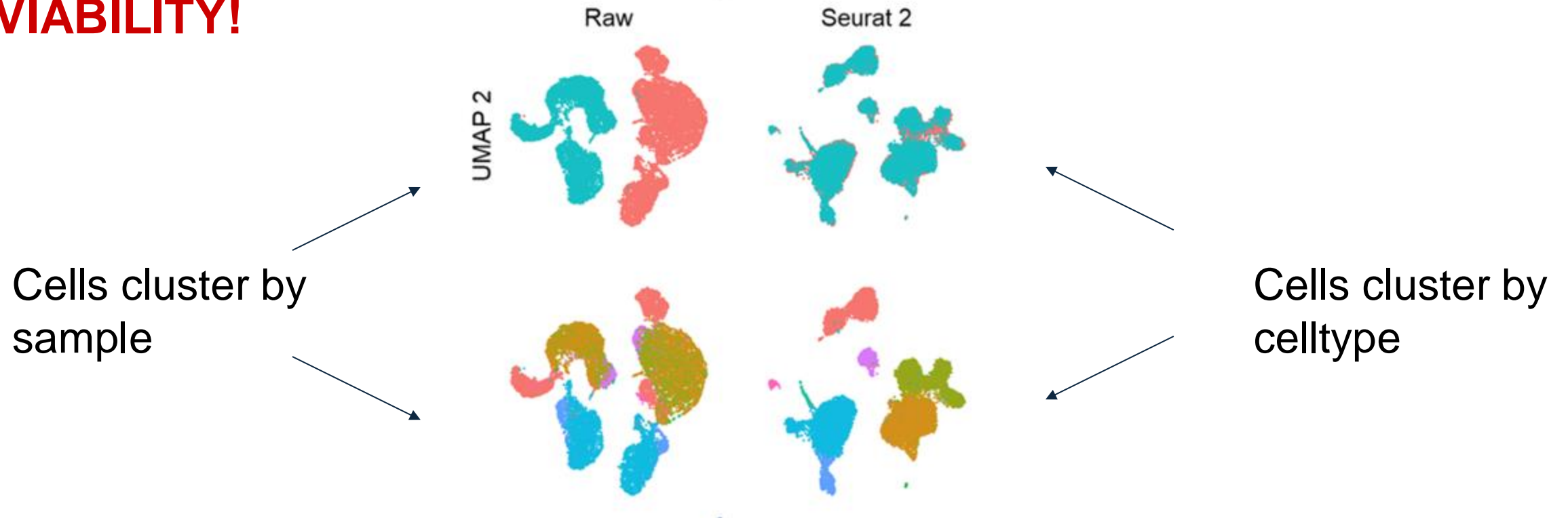
Expressed genes
(genes/cell)
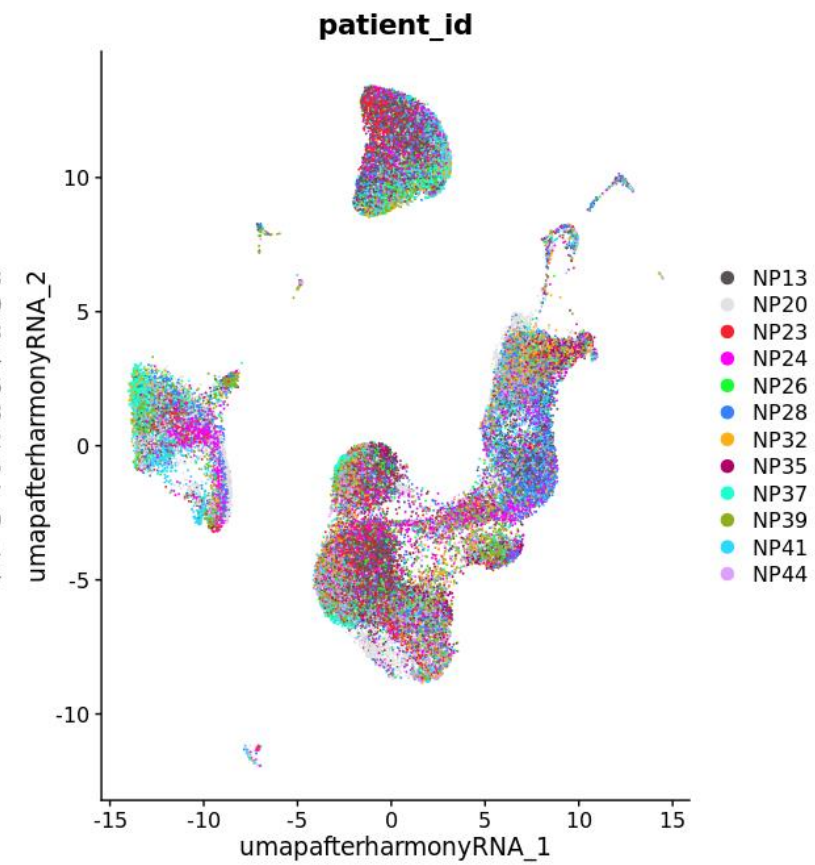
plots: Ziying Ke

AM Gallow, 2021

40

questions

end

Batch effects

VIABILITY!

Cells cluster by sample

Cells cluster by celltype

Tran et al. Genome Biology (2020)

Seurat/Scanpy/etc, many methods

43

patient_id (left panel: umapbeforeharmonyRNA_1 vs umapbeforeharmonyRNA_2)

patient_id (right panel: umapafterharmonyRNA_1 vs umapafterharmonyRNA_2)

Legend: NP13, NP20, NP23, NP24, NP26, NP28, NP32, NP35, NP37, NP39, NP41, NP44
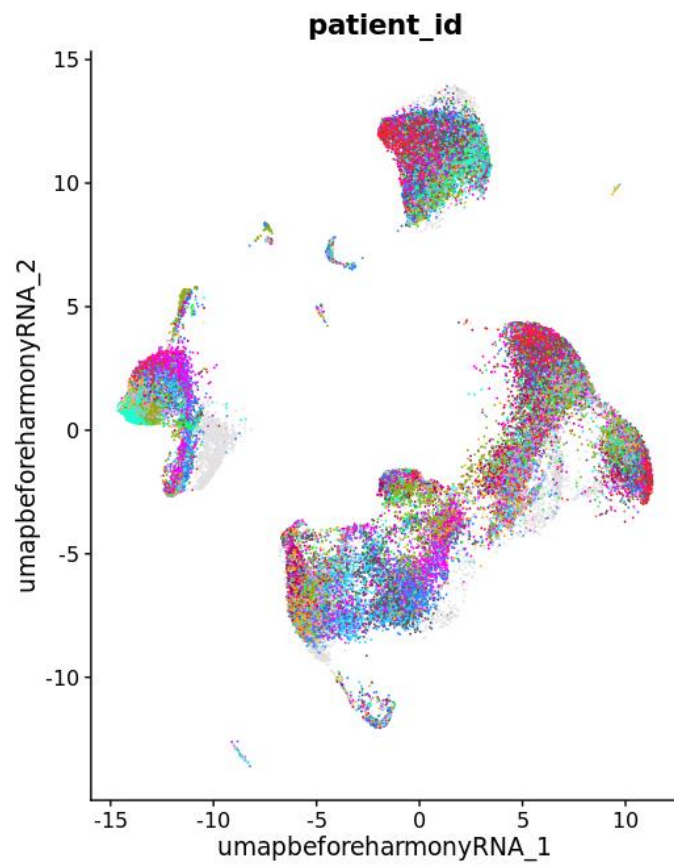
QC and filtering → post-CellRanger

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