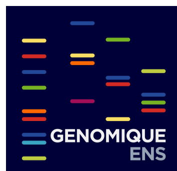
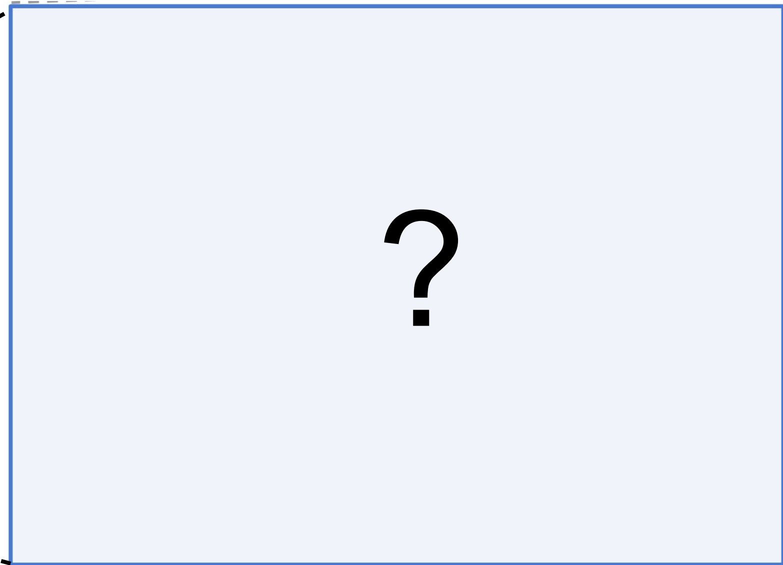
From reads mapping to count matrix

Nathalie Lehmann, Institut Pasteur, Paris



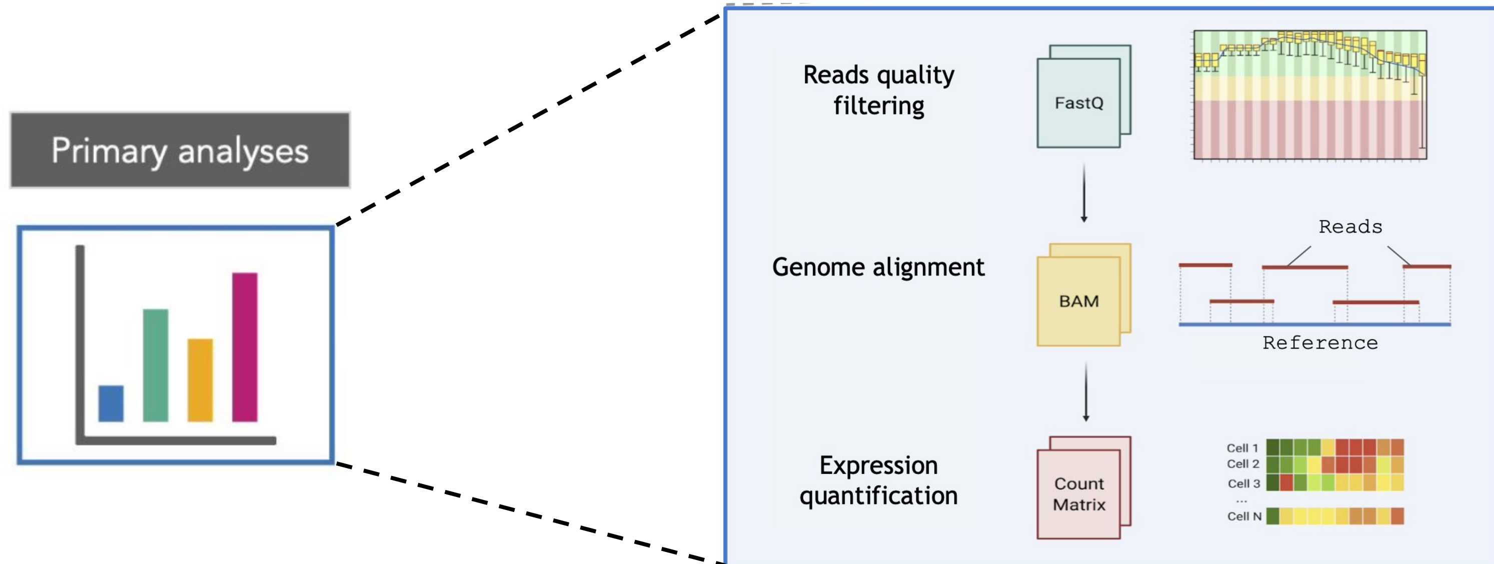
Quizz 1

What are the main steps before getting to the count matrix ?

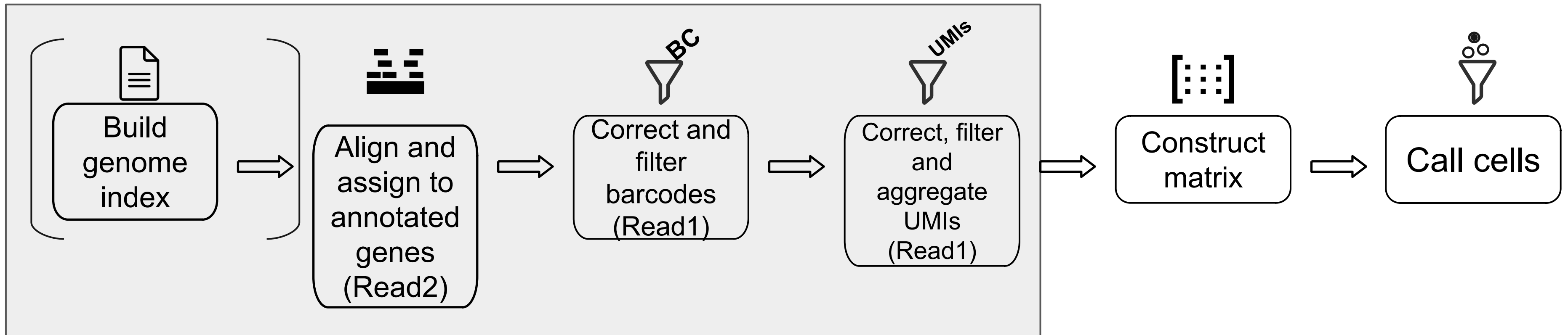


Quizz 1

What are the main steps before getting to the count matrix ?



Primary analysis : overview of the workflow



Quizz 2

How are the reads from 10x Genomics organised ?

The sequenced library

Quizz 2

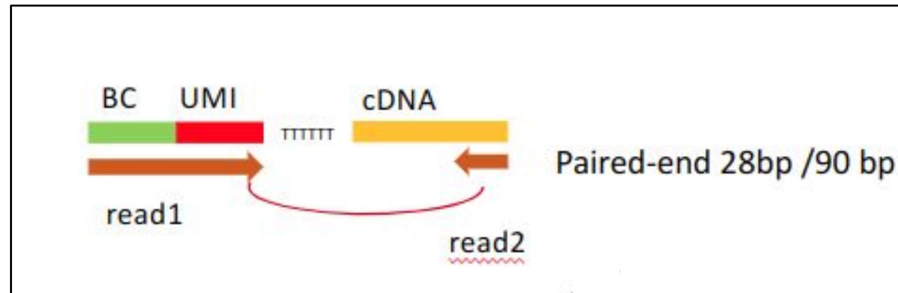
How are the reads from 10x Genomics organised ?

The sequenced library

Quizz 2

How are the reads from 10x Genomics organised ?

The sequenced library



Read 1: unique cell barcode (16 nt) + UMI (12 nt)

Read 2: RNA 3’ sequence

Organisation of the scRNA-seq course

- From cells to nucleotide sequences (reads)
  - focus on the 10X genomics technology
  - how are the reads organised
- Preprocessing : from reads to raw count matrix
  - quality check (FASTQC)
  - mapping (STAR)
  - how is annotation used
  - barcode and UMI treatment
  - visualizing the reads
  - constructing the count matrix
  - call cells / empty droplets filtering

What is the count matrix ?

Cells are in columns

Genes are in rows

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
Gene 1					
Gene 2					
Gene 3					

What is the count matrix ?

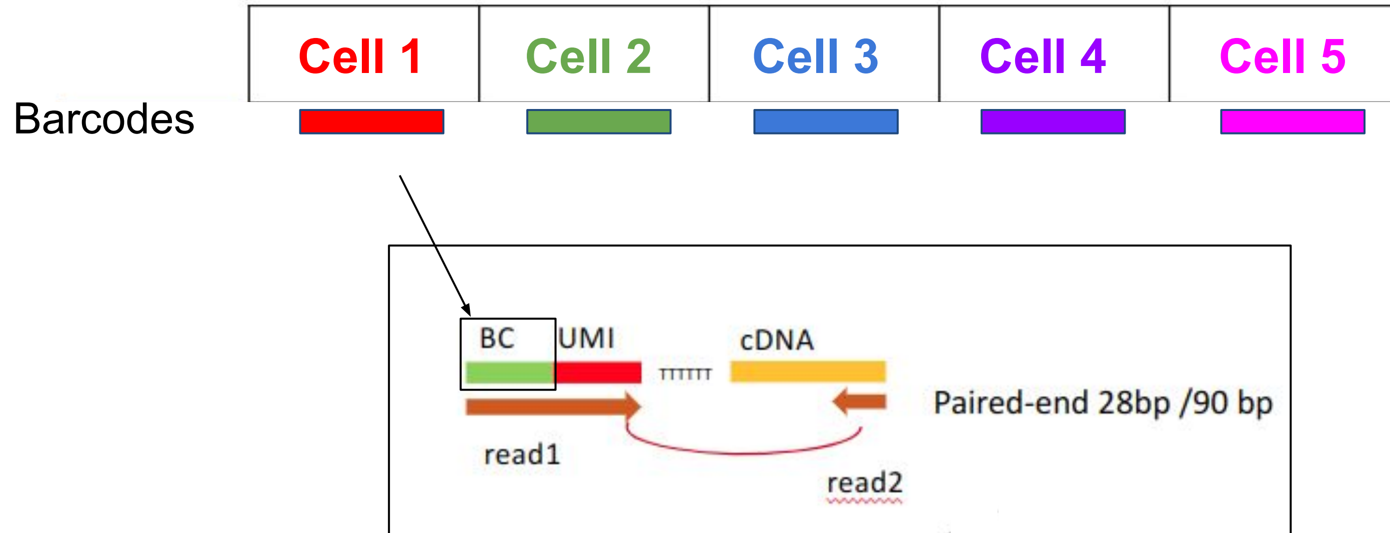
Cells are in columns

Genes are in rows

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
Gene 1	0	0	0	0	0
Gene 2	1	0	4	2	0
Gene 3	0	8	1	0	1

Content of the table :
gene counts (expression levels)

Each cell is represented by a valid barcode (read1)

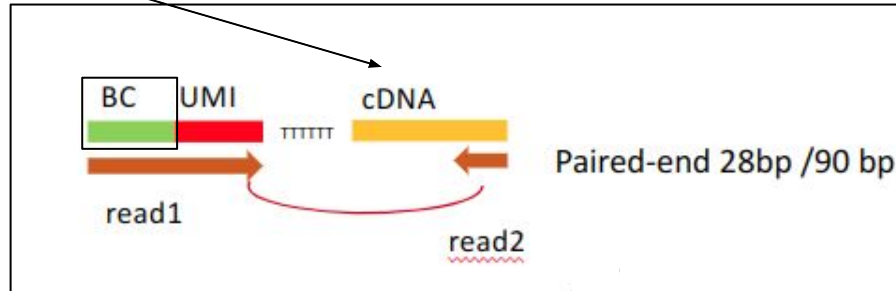


Each read2 is assigned to a gene after the mapping

Gene 1

Gene 2

Gene 3



Gene names are taken from your annotation

Structure of a GFF3 file (annotation file)

general informations

```
##gff-version 3
# gffread v0.12.1
# gffread -E --keep-genes data/raw/references/annotations/ucsc/galGal6.ensGene.gtf -o-
```

chromosome	annotation source / version	type	start	end		strand		attributes (eg. gene names, gene ID)
chr1	ensGene.v101	gene	5273	10061	.	-	.	ID=ENSGALG00000054818.1;Name=IMPDH1
chr1	ensGene.v101	transcript	5273	10061	.	-	.	ID=ENSGALT00000098984.1;Parent=ENSGALG00000054818.1
chr1	ensGene.v101	exon	5273	10061	.	-	.	Parent=ENSGALT00000098984.1
chr3	ensGene.v101	gene	1430	13328	.	+	.	ID=ENSGALG00000049712.1;Name=ENSGALG00000049712.1
chr3	ensGene.v101	transcript	1430	4395	.	+	.	ID=ENSGALT00000097407.1;Parent=ENSGALG00000049712.1
chr3	ensGene.v101	exon	1430	1820	.	+	.	Parent=ENSGALT00000097407.1
chr3	ensGene.v101	exon	4017	4395	.	+	.	Parent=ENSGALT00000097407.1
chr3	ensGene.v101	transcript	3983	13328	.	+	.	ID=ENSGALT00000093532.1;Parent=ENSGALG00000049712.1
chr3	ensGene.v101	exon	3983	4098	.	+	.	Parent=ENSGALT00000093532.1
chr3	ensGene.v101	exon	6175	6179	.	+	.	Parent=ENSGALT00000093532.1
chr3	ensGene.v101	exon	13238	13328	.	+	.	Parent=ENSGALT00000093532.1

Gene 1

Gene 2

Gene 3

chromosome

annotation source / version

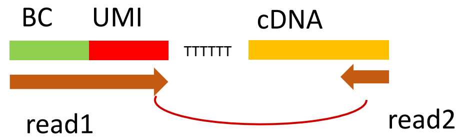
type

start / end

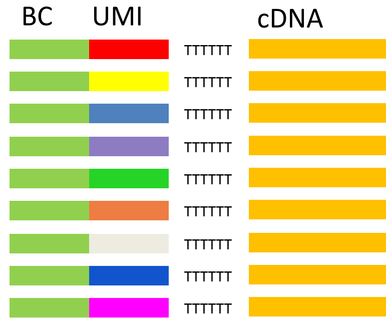
strand

attributes (eg. gene names, gene ID)

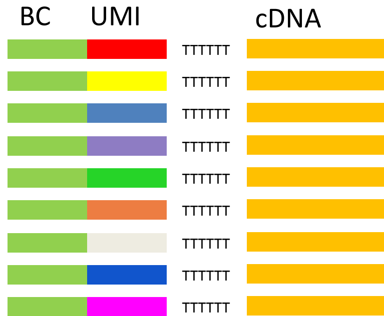
How are the reads counted ?



How are the reads counted ?



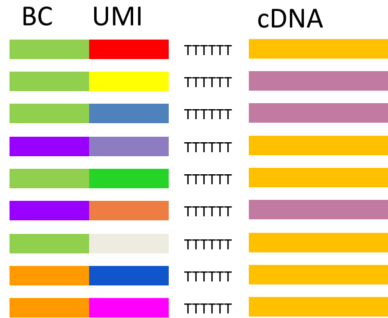
How are the reads counted ?



One cell (1 BC)

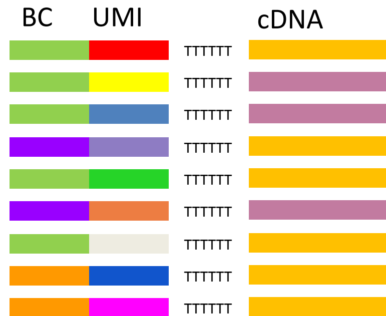
One gene, detected 9 times
(1 cDNA - 9 UMI)

How are the reads counted ?



?

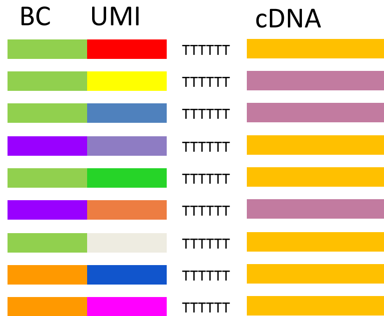
How are the reads counted ?



3 cells (3 BC)

2 genes (2 cDNA)

How are the reads counted ?

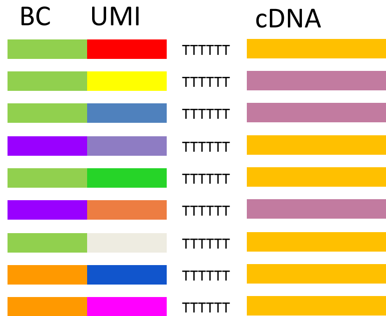


3 cells (3 BC)

2 genes (2 cDNA)

In practice, the count in the matrix corresponds to the number of **UMI** per barcode per gene

How are the reads counted ?

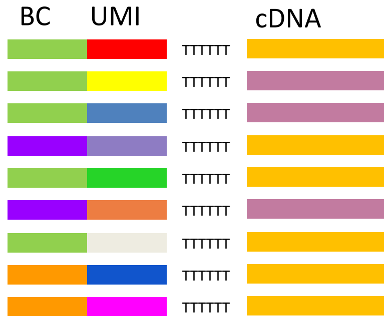


3 cells (3 BC)

2 genes (2 cDNA)

	Cell 1	Cell 2	Cell 3
Gene 1			
Gene 2			

How are the reads counted ?

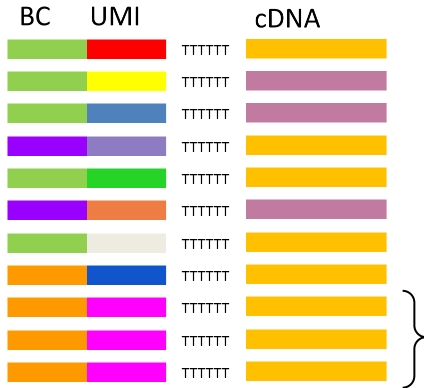


3 cells (3 BC)

2 genes (2 cDNA)

	Cell 1	Cell 2	Cell 3
Gene 1	3	1	2
Gene 2	2	1	0

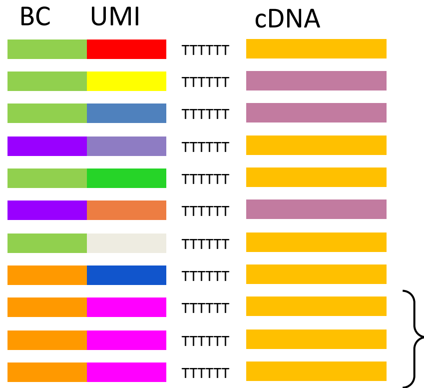
How are the reads counted ?



Reads with the same BC+UMIs are assigned to the same gene (originate from 1 unique RNA molecule) : **they count as 1**

The UMIs are used to correct for **amplification artefacts**

How are the reads counted ?

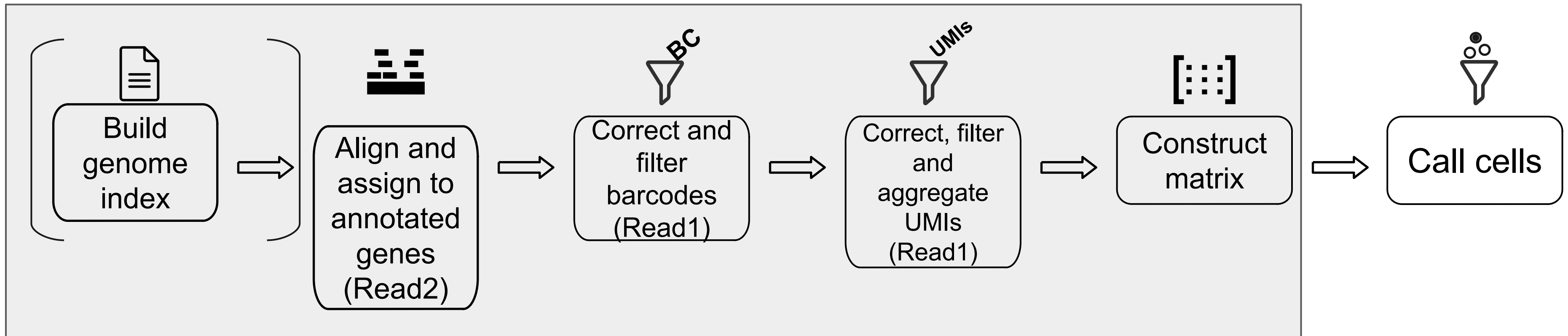


Reads with the same BC+UMIs are assigned to the same gene (originate from 1 unique RNA molecule) : **they count as 1**

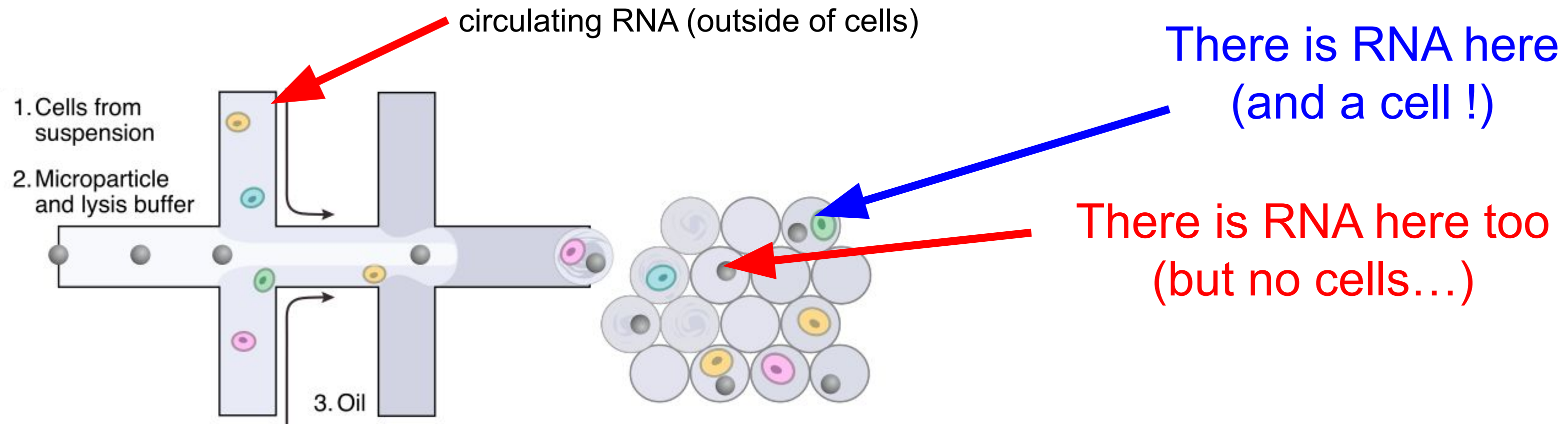
	Cell 1	Cell 2	Cell 3
Gene 1	3	1	2
Gene 2	2	1	0

Same count matrix as before

Primary analysis : overview of the workflow



Counting the cells



- A million of droplets to recover ~10k cells
- Problem : RNA from dead cells circulates and is encapsulated in droplets
- **Question : how to differentiate between “real cells” and “droplets with RNA” ?**

Empty droplets filtering

- Need to **filter** the count matrix to retain the droplets most likely containing a true cell, removing the “empty” droplets containing only ambient RNA

Empty droplets filtering

- Need to **filter** the count matrix to retain the droplets most likely containing a true cell, removing the “empty” droplets containing only ambient RNA

- **Problem**: we have no prior knowledge about whether a barcode corresponds to cell-containing or empty droplets. We need to call cells from empty droplets based on the observed expression profiles.

Empty droplets filtering

- Need to **filter** the count matrix to retain the droplets most likely containing a true cell, removing the “empty” droplets containing only ambient RNA

- **Problem**: we have no prior knowledge about whether a barcode corresponds to cell-containing or empty droplets. We need to call cells from empty droplets based on the observed expression profiles.

- **Principle** : true cells will contain many different RNA molecules, compared to empty droplets containing few ambient RNA

  => translates into : barcodes associated to many UMI are more likely to be true cells than barcodes associated to few UMIs

Identification of the "true" cells depends on UMI diversity

	Cell 1	Cell 2	Cell 3
Gene 1	3	1	2
Gene 2	2	1	0
Total	5	2	2

1 UMI = 1 single RNA molecule

Identification of the “true” cells depends on UMI diversity

Gene 1	30	1	2	1	0	6	7	0	0	2	0	9	2	0	0	1	2
Gene 2	8	1	0	4	0	2	1	4	1	3	0	3	15	1	0	1	0
Total	38	2	2	5	0	8	8	4	1	5	0	12	17	1	0	2	2

Identification of the “true” cells depends on UMI diversity

Gene 1	**30**	**1**	**2**	**1**	**0**	**6**	**7**	**0**	**0**	**2**	**0**	**9**	**2**	**0**	**0**	**1**	**2**
Gene 2	**8**	**1**	**0**	**4**	**0**	**2**	**1**	**4**	**1**	**3**	**0**	**3**	**15**	**1**	**0**	**1**	**0**
Total	**38**	**2**	**2**	**5**	**0**	**8**	**8**	**4**	**1**	**5**	**0**	**12**	**17**	**1**	**0**	**2**	**2**

“Low UMI cells” ~ “empty droplets”

Identification of the “true” cells depends on UMI diversity

Gene 1	30	1	2	1	0	6	7	0	0	2	0	9	2	0	0	1	2
Gene 2	8	1	0	4	0	2	1	4	1	3	0	3	15	1	0	1	0
Total	38	2	2	5	0	8	8	4	1	5	0	12	17	1	0	2	2

“Low UMI cells” ~ “empty droplets”

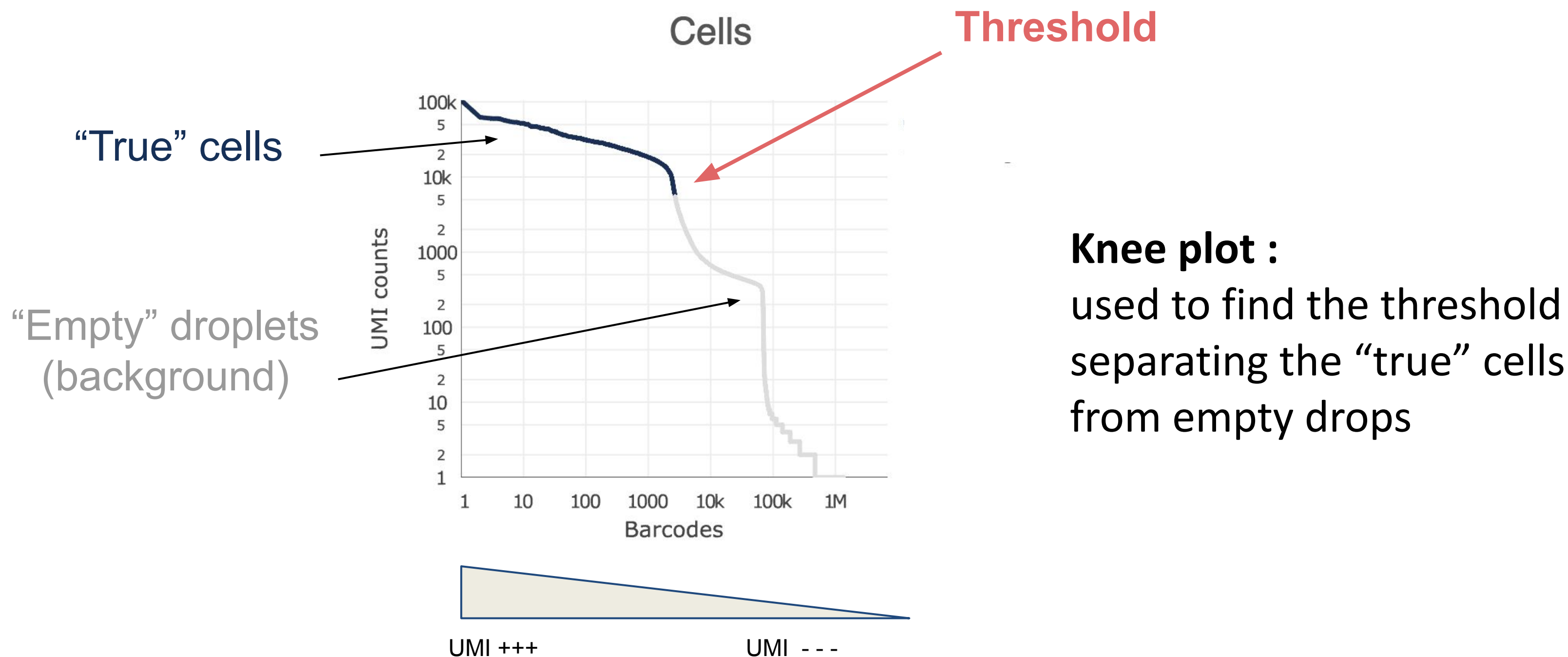
Identification of the "true" cells depends on UMI diversity

Gene 1	30	1	2	1	0	6	7	0	0	2	0	9	2	0	0	1	2
Gene 2	8	1	0	4	0	2	1	4	1	3	0	3	15	1	0	1	0
Total	38	2	2	5	0	8	8	4	1	5	0	12	17	1	0	2	2

"Low UMI cells" ~ "empty droplets"... ???

How do we set the **threshold** between "true" cells and "droplets" ?

Identification of the “true” cells depends on UMI diversity

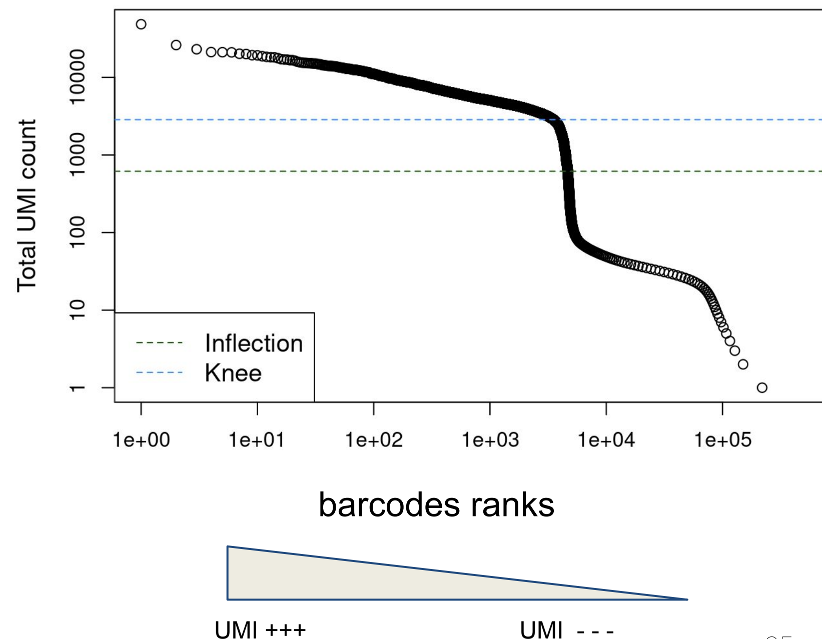


Knee plot

The knee plot (or *barcode rank plot*) is used for filtering the droplets

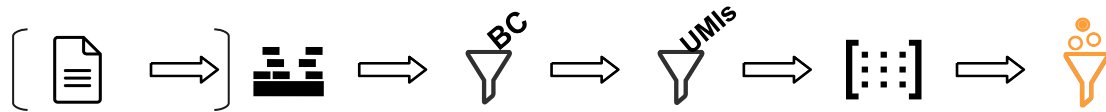
Steps :

1. Keep all barcodes over first knee point
2. Deduce background from low content droplets
3. Select droplets under knee point if the composition is very different from the background (cells with low-content RNA)



Cell identification with CellRanger

Cellranger

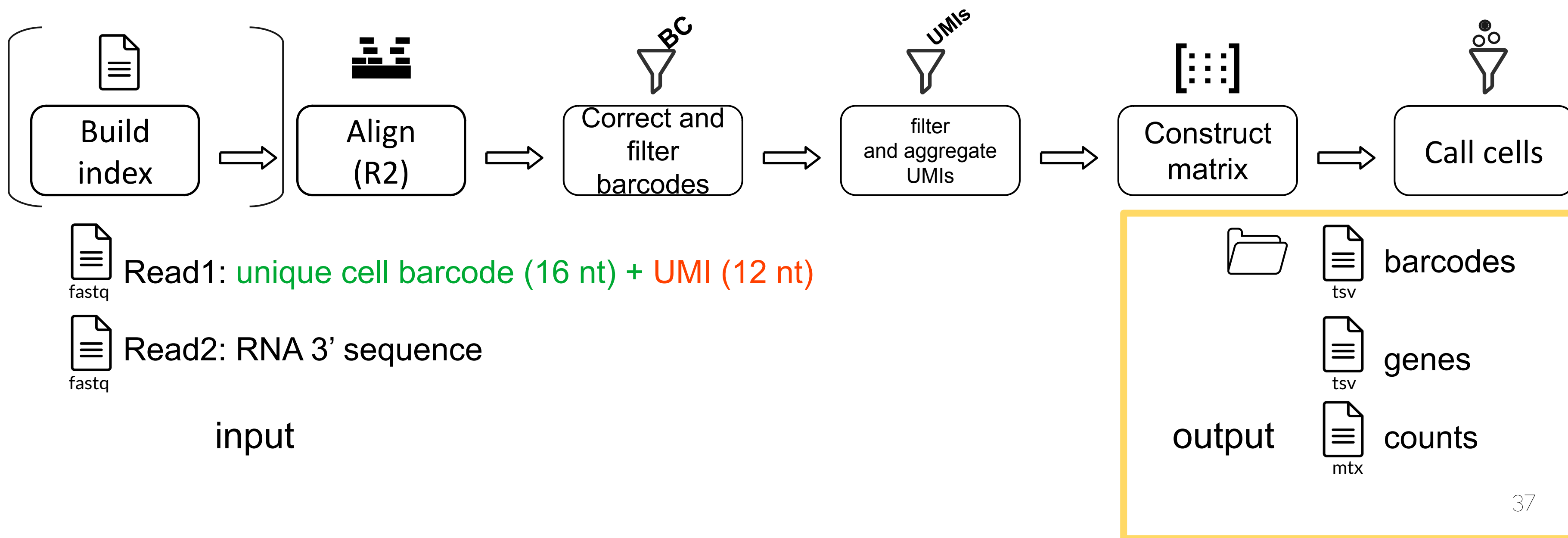


- Final number of cells can be < targeted cells
- With 10x Genomics data, cell capture is usually around 50% - 60%

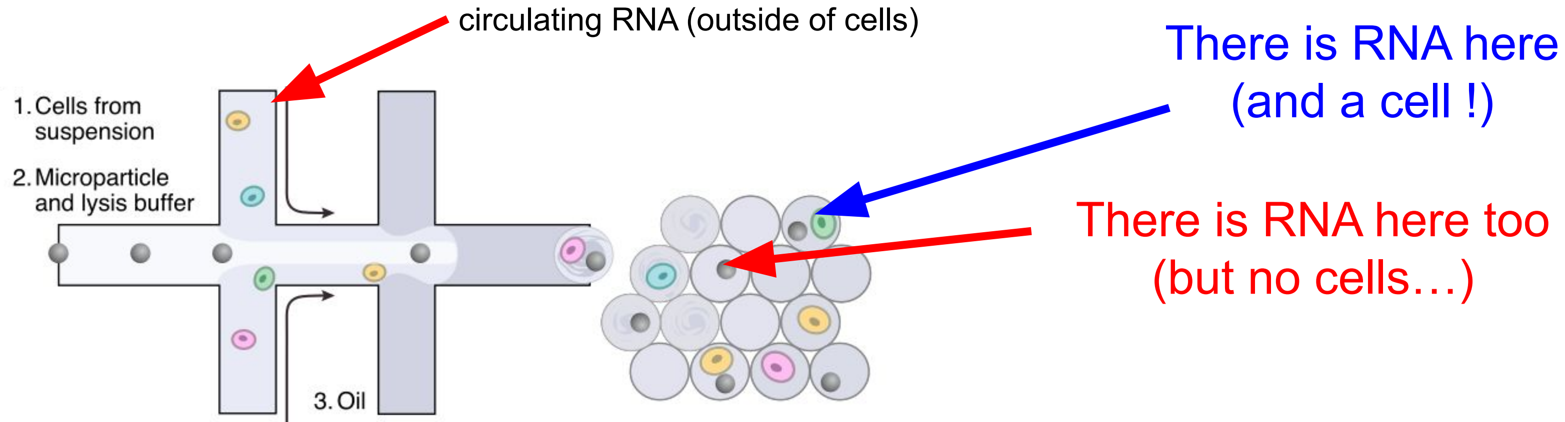
- A second round of cell filtering step is necessary. It is performed at the beginning of the data analyses (we will see that later in the course)

Output of CellRanger

Principle



Ambient RNA can be present in “true cells”



- Ambient RNA may also be encapsulated within droplets containing a cell (“true cells”)
- There are tools that allow to correct for this biais
- Ex : *SoupX* - that infers ambient RNA “soup” and removes it from the gene counting

CellRanger output report

Estimated Number of Cells
2,700

Mean Reads per Cell	Median Genes per Cell
197,634	3,704

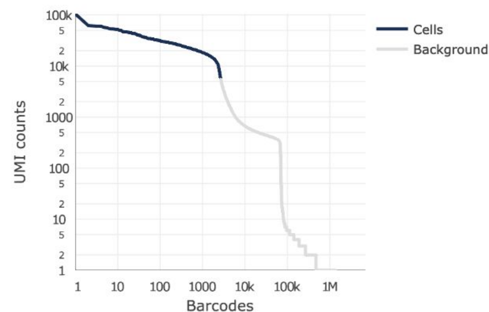
Sequencing

Number of Reads	533,613,214
Valid Barcodes	96.0%
Sequencing Saturation	67.7%
Q30 Bases in Barcode	96.1%
Q30 Bases in RNA Read	90.8%
Q30 Bases in UMI	95.2%

Mapping

Reads Mapped to Genome	74.6%
Reads Mapped Confidently to Genome	70.5%
Reads Mapped Confidently to Intergenic Regions	8.7%
Reads Mapped Confidently to Intronic Regions	8.4%
Reads Mapped Confidently to Exonic Regions	53.4%
Reads Mapped Confidently to Transcriptome	50.4%
Reads Mapped Antisense to Gene	1.0%

Cells



Estimated Number of Cells	2,700
Fraction Reads in Cells	55.1%
Mean Reads per Cell	197,634
Median Genes per Cell	3,704
Total Genes Detected	17,998
Median UMI Counts per Cell	16,440

Sample

Name	CellRanger_Report_1
Description	
Transcriptome	cellranger_mkref_output_v3_191003
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.1

CellRanger output report

- Turnkey solution
- Many QC-metrics in 1 html summary
- Some secondary analysis
- More complex experiences : VDJ analysis, feature-barcoding
- Versions for ATAC-Seq, multi-omics

- Proprietary
- Only 10X product (cannot customize BC and UMI patterns)
- Not highly configurable
- (A lot of resource and time)
  but less true for recent versions

There are other alternatives than CellRanger

Technical Overview mapper

	Cell Ranger	STARsolo	Alevin	Kallisto
Mapping scheme	Exact alignment	Exact alignment	Pseudo mapping	Pseudo mapping
Internal Mapper	Star	Star	Salmon	Kallisto
Reference	Genome	Genome	Transcriptome + Genome	Transcriptome
Supported sequence technology	10X Chromium v1 – v3	10X Chromium v2;v3, Smart-seq, Drop-seq, inDrop	10x Chromium v2;v3, Drop-seq, Cel-seq, Cel-seq2, Quartz-seq2	10x Chromium v1 – v3, Cel-seq, Cel-seq2, Drop-seq, inDrops v1-v3, SCRB-Seq, SureCell
Barcode correction	1-Hamming distance based	1-Hamming distance based	Edit distance calculation	1-Hamming distance based
Whitelisting	Whitelist based	Whitelist based	Frequency based, no whitelist needed	Whitelist based
Alternative Splicing detection	no	yes	no	no
UMI correction	Two round correction by barcode, read count and annotation	Two round correction by barcode, read count and annotation	graph based correction	NA
Index	Suffix array	Suffix array	Colored De-Bruijn Graph	Colored De-Bruijn Graph
Handling of multimapped reads	discarded	discarded	Distributing read count between genes by EM-algorithm	discarded
Output	Matrix + Bam-File and summary file as html-file with primary results as well as clustering and DEG analysis	Gene count matrix and primary results summary	Gene count matrix ready for analysis	External software required to create gene count matrix

Summary

	Cell Ranger	STARsolo	Alevin	Kallisto
Mapping performance	Lowest runtime	Similar results with Cell Ranger that are accomplished in a shorter time	Whitelisting causes loss or gain of barcodes depending on the data	Fastest runtime with highest mapping rate, more cells are detected with a small gene content
Barcode correction and filtering			Final barcode set included barcodes that are not present in the whitelist	Reports more cells with a low gene content
Gene discovery				Detection of more genes than all other tools. Highest UMI count for genes not expressed in studied tissue
MT-content	Highly affected by complete annotation including pseudogenes	See Cell Ranger	Smaller difference of MT-content between the mapping with filtered and unfiltered annotation	See Cell Ranger
Clustering	Highest Overlap with SCINA classification	Very similar to Cell Ranger with minor differences	Cell types contain lower amount of cells with SCINA classification	Cell types contain the lowest amount of cells with SCINA classification
DEG	No difference detected	No difference detected	No difference detected	No difference detected

R. S. Brüning *et al.*, Gigascience (2022)

Take-home messages

- Take time to **visualize the scRNA-seq signal** in a genome browser (IGV)
- Results can be hugely affected by the annotation
- Exotic /poorly-annotated / non-model organisms : generate a new annotation from bulk data with long-read sequencing (or reconstruct with short-reads)

- Count matrix = nb of UMI per barcode (columns) per gene (row)

- Call the cells : remove the empty droplets containing ambient RNA => use of the knee plot to decide on the threshold and obtain the number of “true” cells
- Sometimes, need to lower this threshold for small cells/low-RNA content
- More filters will be applied in the downstream analysis

Acknowledgements

- Slides taken or inspired from Bastien Job

- Some illustrations were created by
  - Kevin Lebrigand
  - Morgane Thomas-Chollier