Atelier scRNA-seq

Technology for scRNA-seq and data processing

Bastien Job, Gustave Roussy, Villejuif

Morgane Thomas-Chollier - GenomiqueENS, Paris + IFB-core

École de bioinformatique EB3I 2025

Organisation of this session

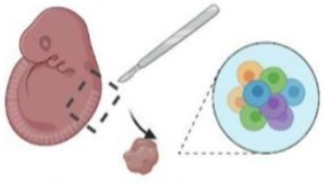
- From cells to nucleotide sequences (reads)
  - focus on the 10X genomics technology
  - how are the reads organised
- Preprocessing : from reads to raw count matrix
  - quality check (FASTQC)
  - mapping (STAR)
  - how is annotation used
  - barcode and UMI treatment

Global overview of a scRNA-seq experiment

Tissue dissection + cell dissociation

Cell partitioning + mRNA capture

Library preparation + sequencing

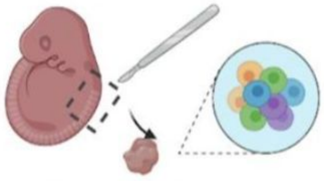


Global overview of a scRNA-seq experiment

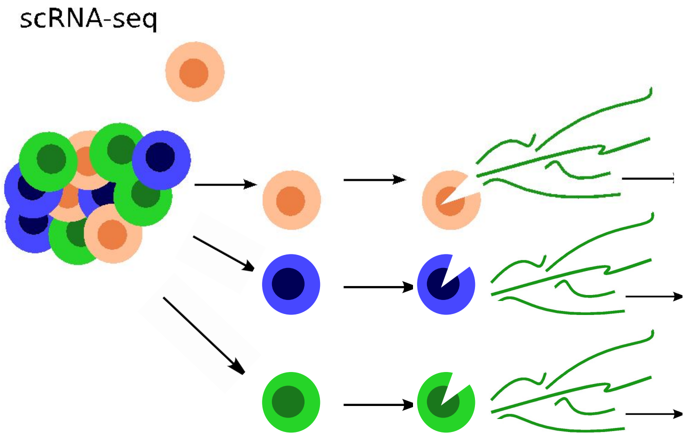
Tissue dissection + cell dissociation

Cell partitioning + mRNA capture

Library preparation + sequencing

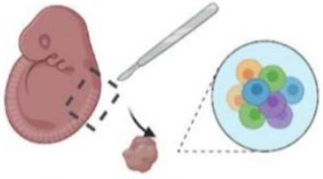


this step enables to treat each cell separately, and capture its RNA while retaining from which cell it originates



Global overview of a scRNA-seq experiment

Tissue dissection + cell dissociation

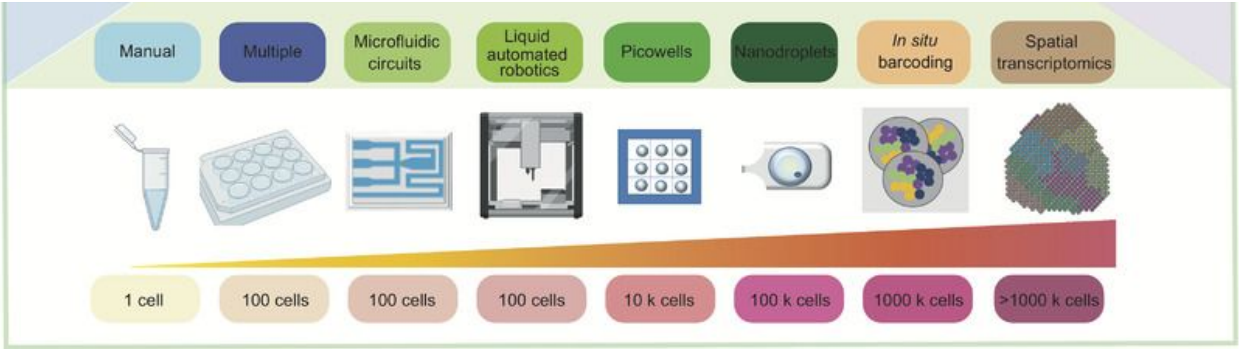


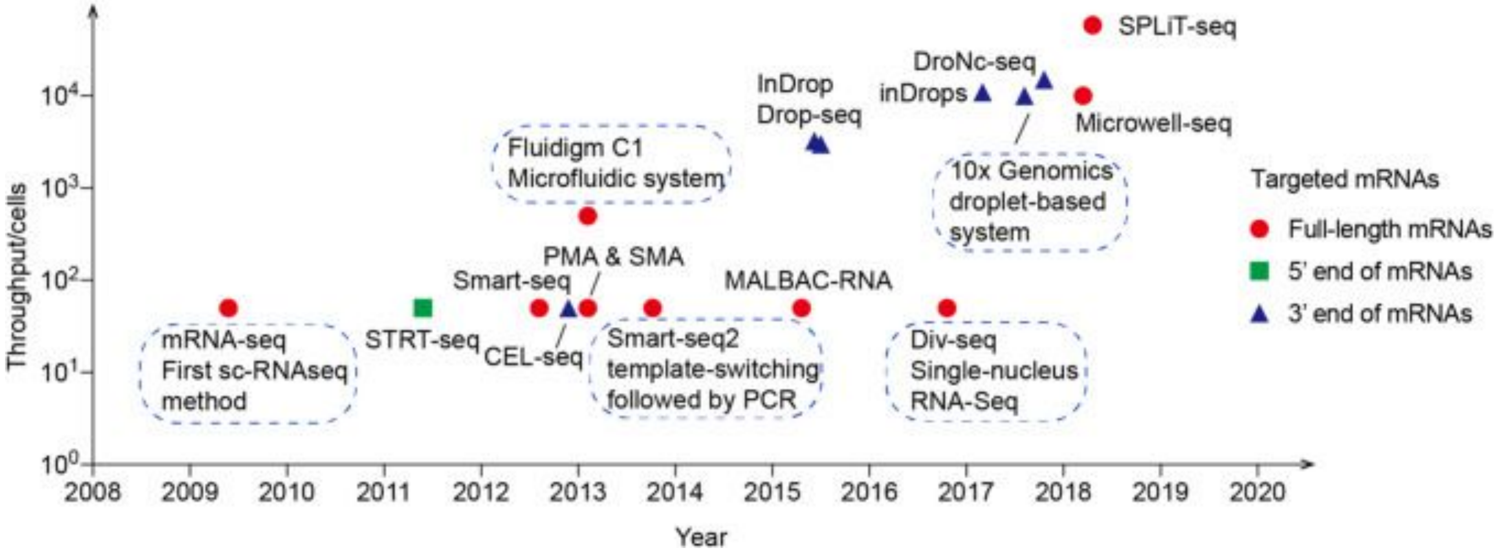
Cell partitioning + mRNA capture

various technologies developed over time for this specific step

Library preparation + sequencing

Technologies over the last decade

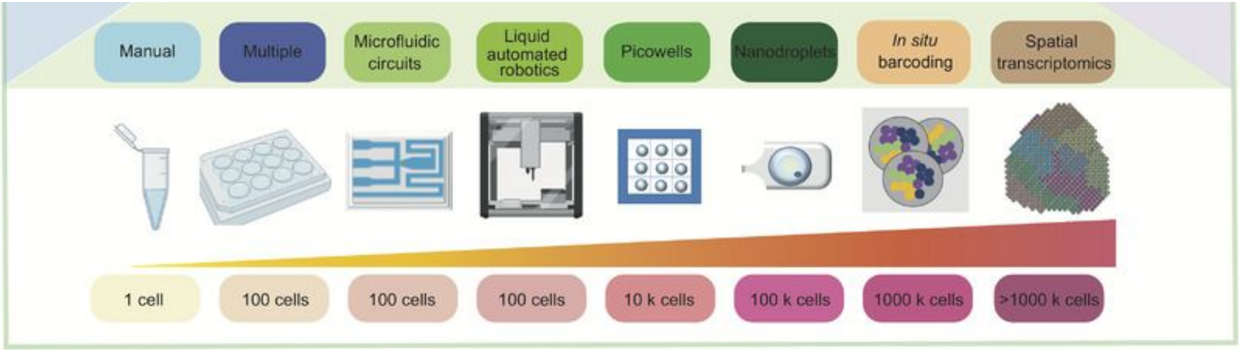




Jovic et al 2022 PMID: 35352511

Zhang et al 2022 PMID: 35318612

Technologies over the last decade

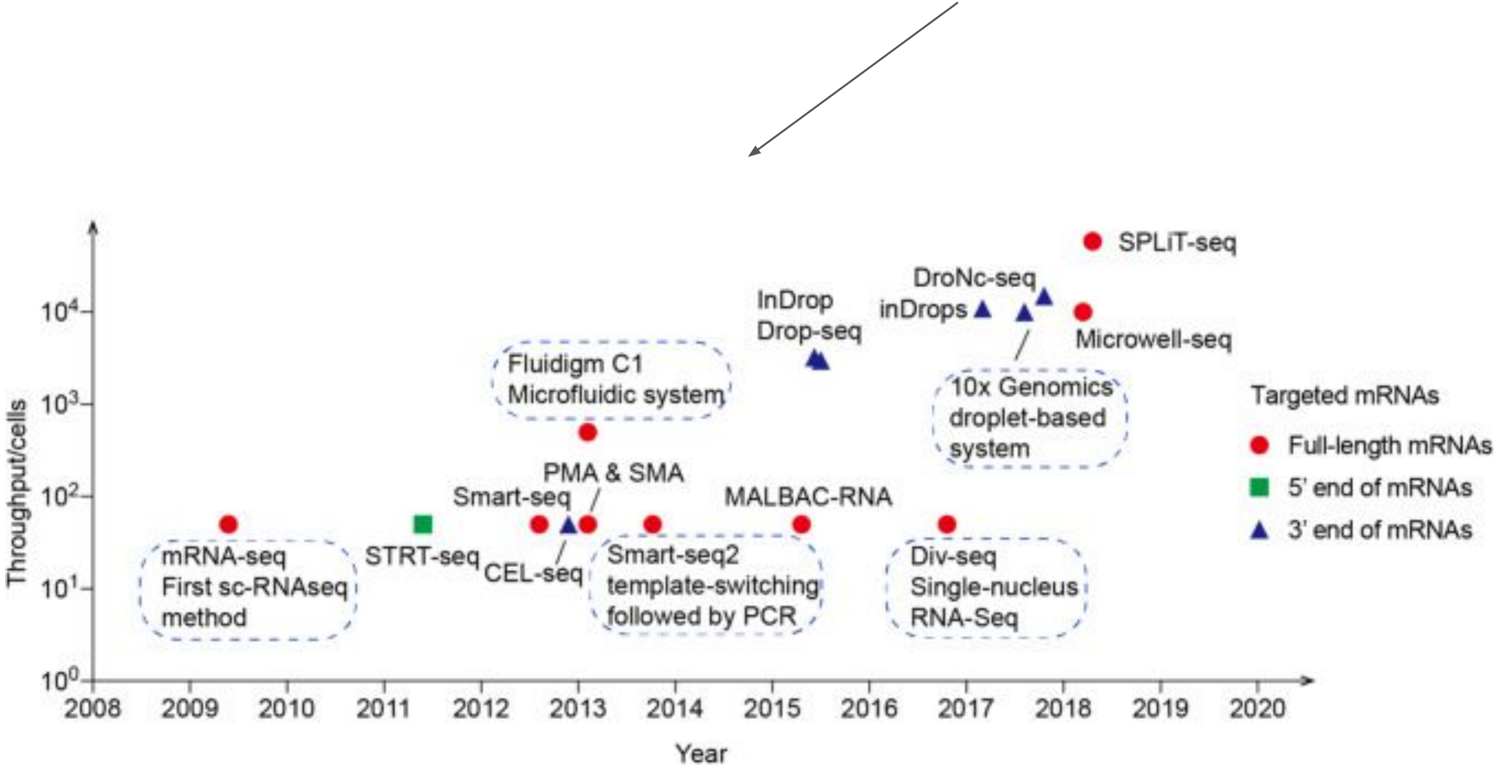


Jovic et al 2022 PMID: 35352511

Differences in cell isolation/partitioning : the device can be a simple plate to complex microfluidic machines

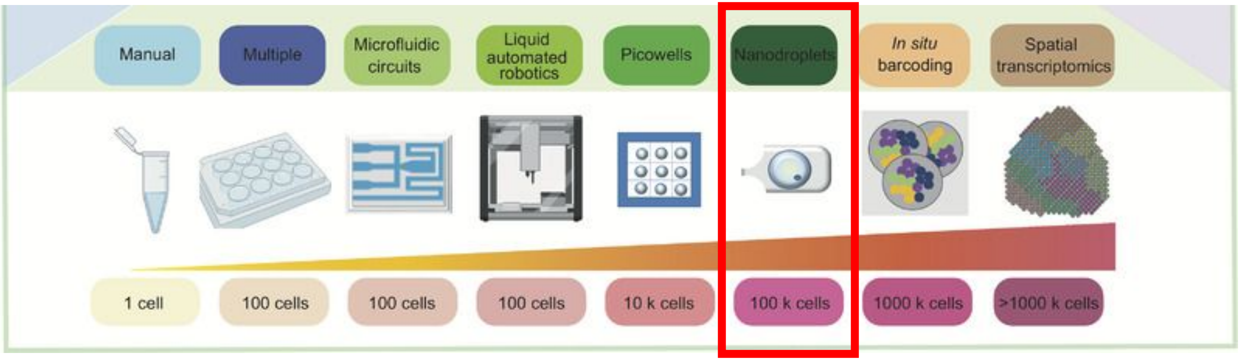
Technologies over the last decade

The number of cells that can be studied has grown from a handful to >10,000 in 10 years



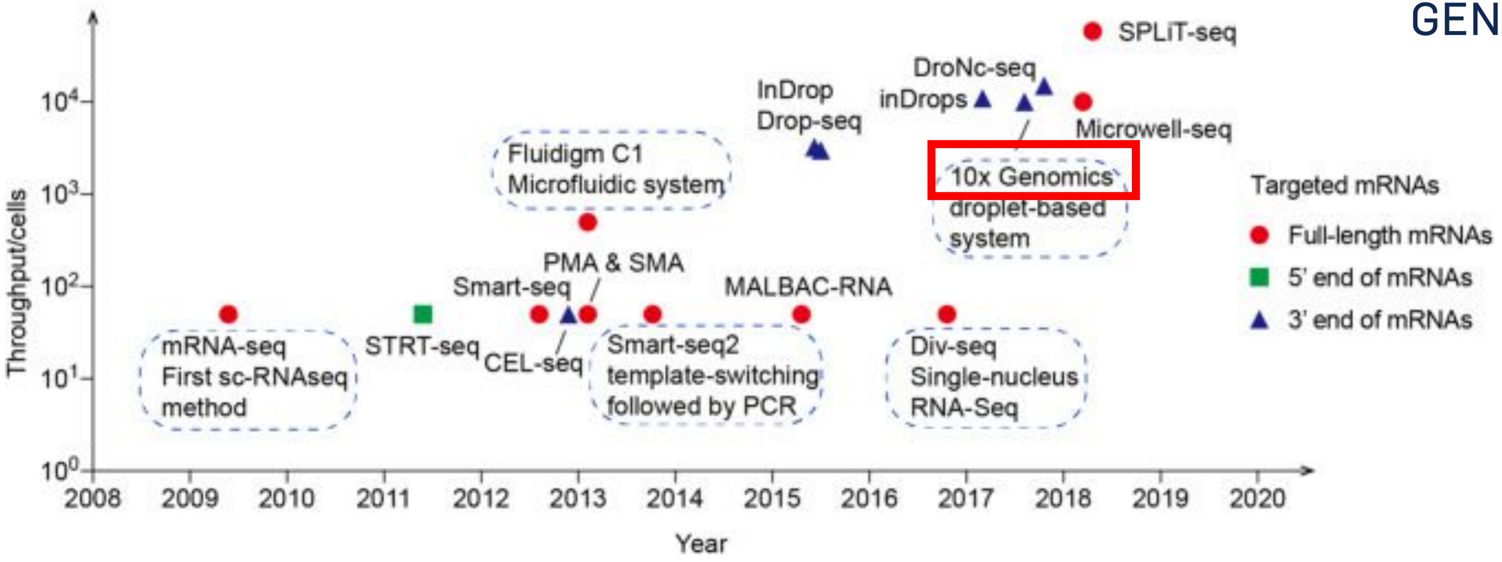
Zhang et al 2022 PMID: 35318612

Technologies over the last decade



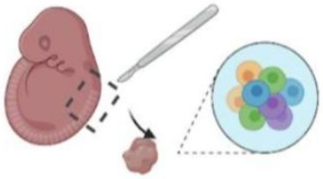
The technology that has enabled widespread usage of scRNA-seq approach is the droplet-based approach proposed by the company 10X Genomics.

10x GENOMICS®

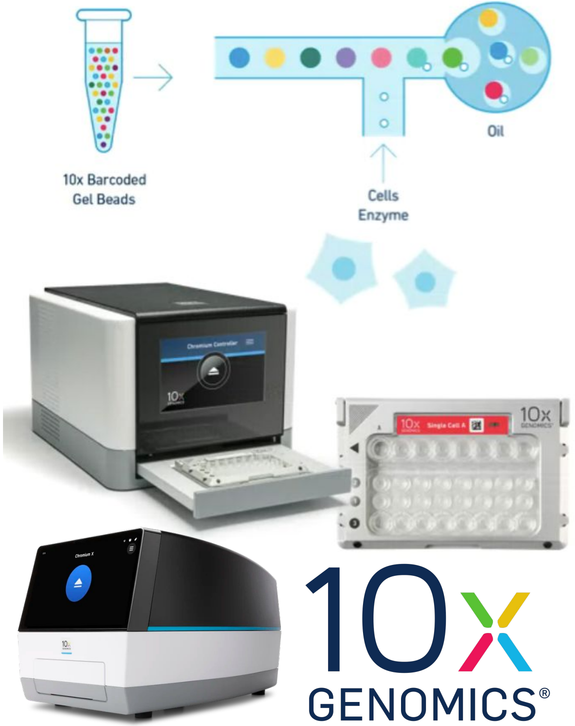


Single-cell transcriptomics with 10X genomics technology

Tissue dissection + cell dissociation



Cell partitioning + mRNA capture



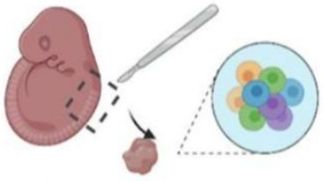
Library preparation + sequencing

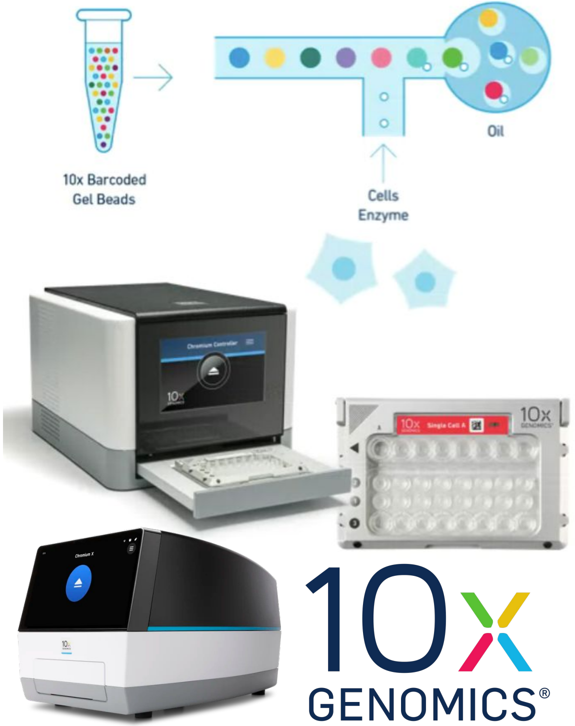
Single-cell transcriptomics with 10X genomics technology

Tissue dissection + cell dissociation

Cell partitioning + mRNA capture

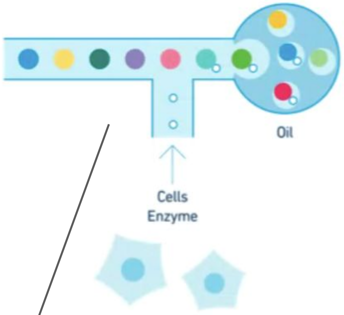
Library preparation + sequencing





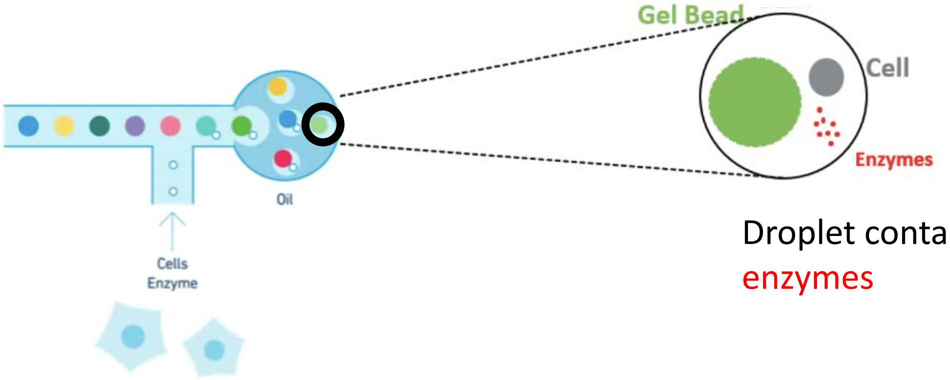
How is the 10X Genomics droplet-based system working ?

Single-cell transcriptomics with 10X genomics technology



Cells and gel beads arrive in the device from 2 separate channels

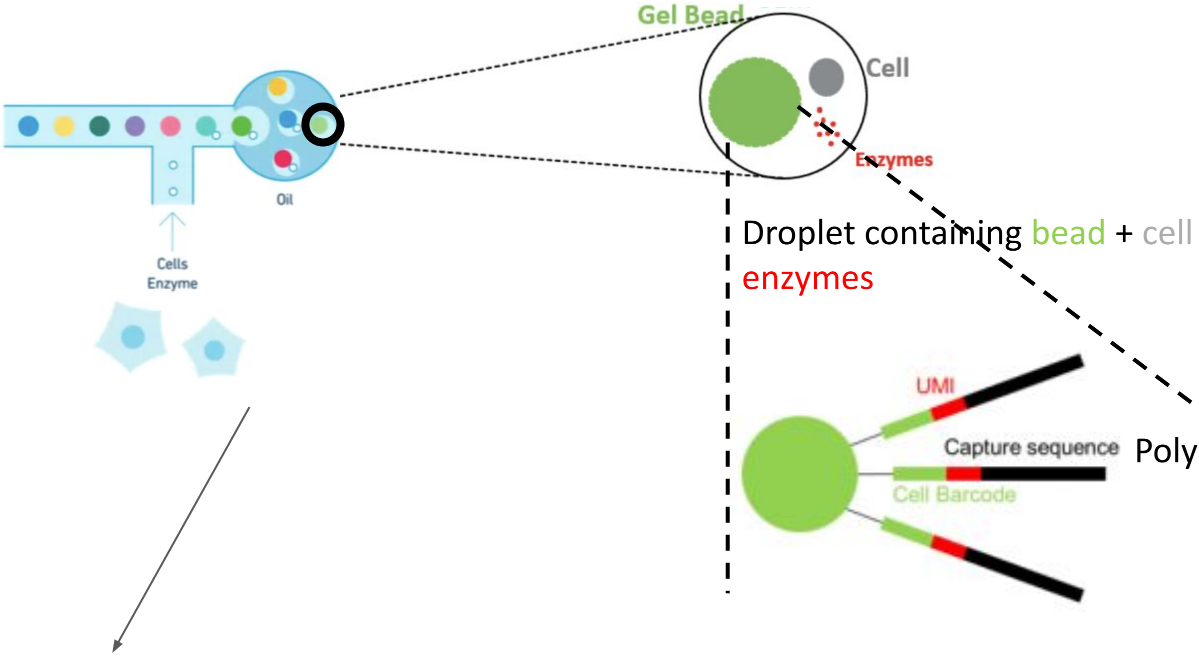
Single-cell transcriptomics with 10X genomics technology



Droplet containing bead + cell + enzymes

A single cell and a single gel bead (+ enzymes) are then encapsulated in a droplet

Single-cell transcriptomics with 10X genomics technology

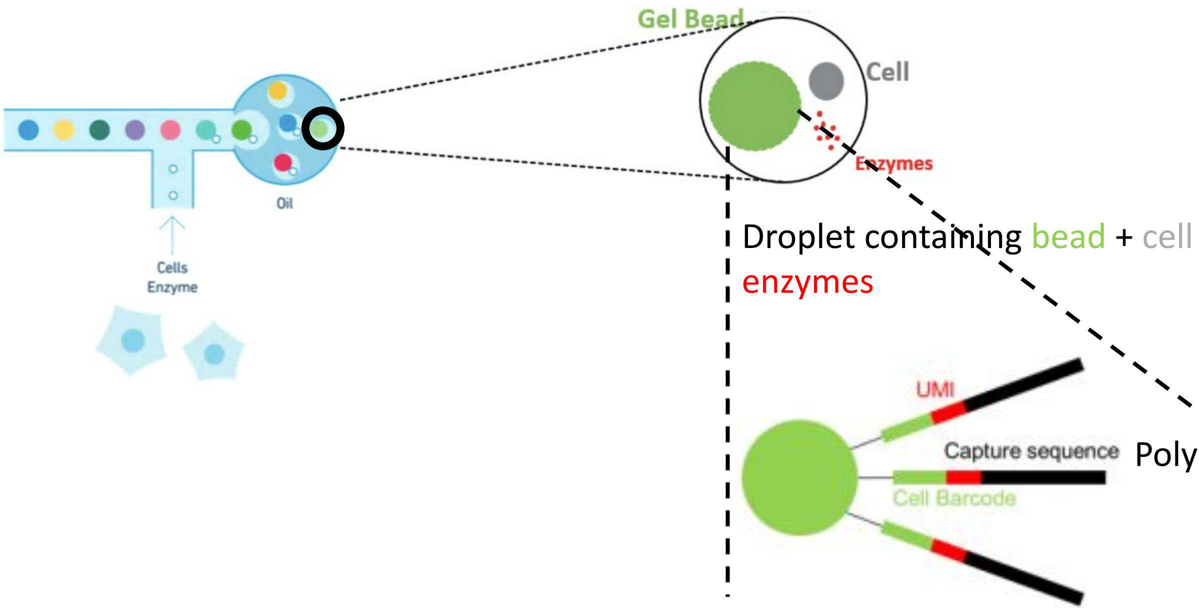


Droplet containing bead + cell + enzymes

Polyd(T) to capture polyA+ RNA

The gel bead is special : it is covered with molecules made of 3 parts

Single-cell transcriptomics with 10X genomics technology

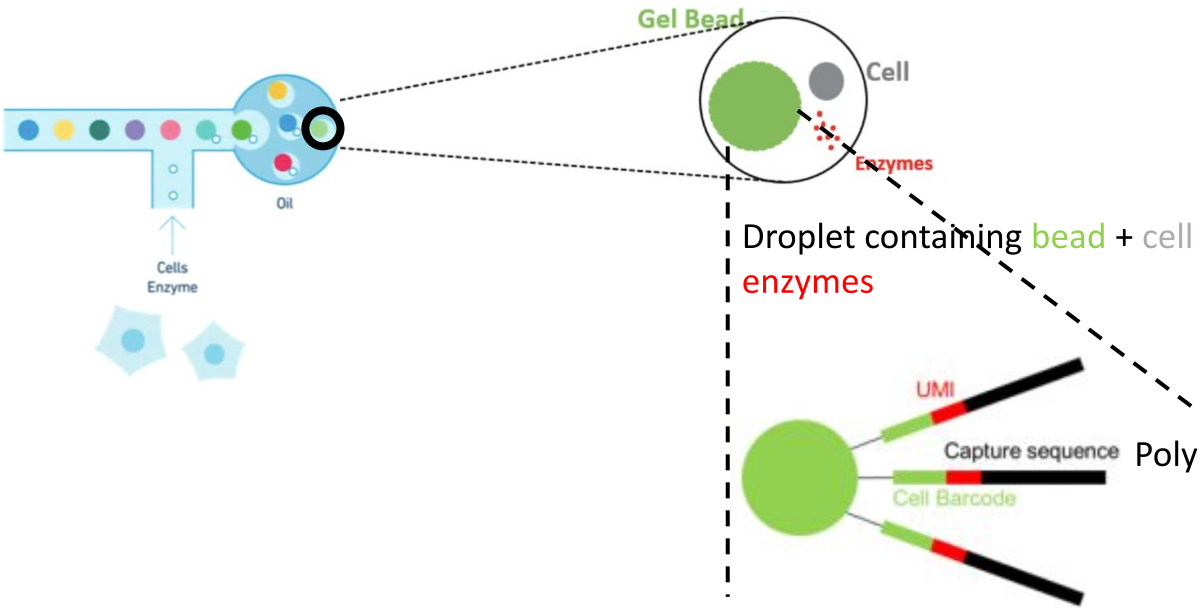


Droplet containing bead + cell + enzymes

Polyd(T) to capture polyA+ RNA

The black part is the **capture sequence** (to “catch” the RNA). 10X has various capture sequences. Here the sequence is polyd(T) to capture RNA that are polyadenylated polyA+.

Single-cell transcriptomics with 10X genomics technology



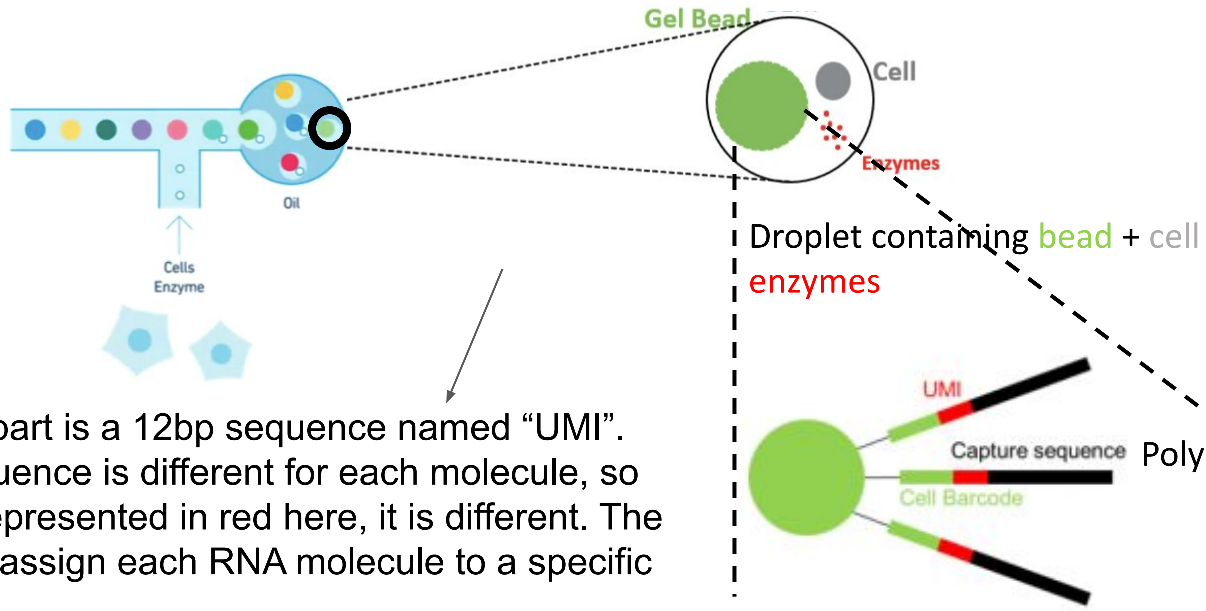
Droplet containing bead + cell + enzymes

Polyd(T) to capture polyA+ RNA

Cell Barcode (16bp) = sequence specific to each bead (so each cell)

The green part is a 16bp sequence named “barcode”. This same sequence is all over the bead. These barcodes are created by 10X and the list is available.

Single-cell transcriptomics with 10X genomics technology



Droplet containing bead + cell + enzymes

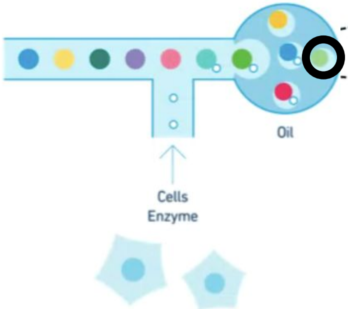
Polyd(T) to capture polyA+ RNA

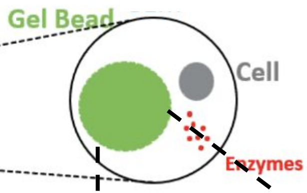
The red part is a 12bp sequence named “UMI”. This sequence is different for each molecule, so even if represented in red here, it is different. The aim is to assign each RNA molecule to a specific UMI.

Cell Barcode (16bp) = sequence specific to each bead (so each cell)

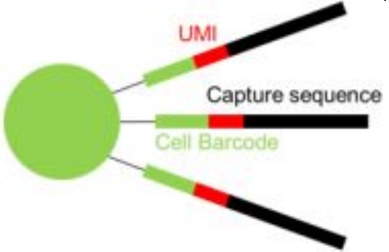
UMI (12 bp) = sequence specific to each molecule : Unique Molecular Identifier

Single-cell transcriptomics with 10X genomics technology





Droplet containing bead + cell + enzymes

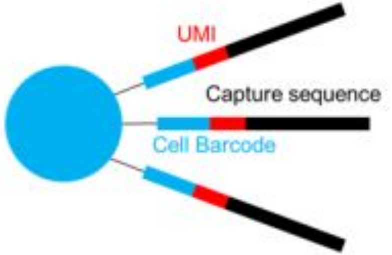


Polyd(T) to capture polyA+ RNA

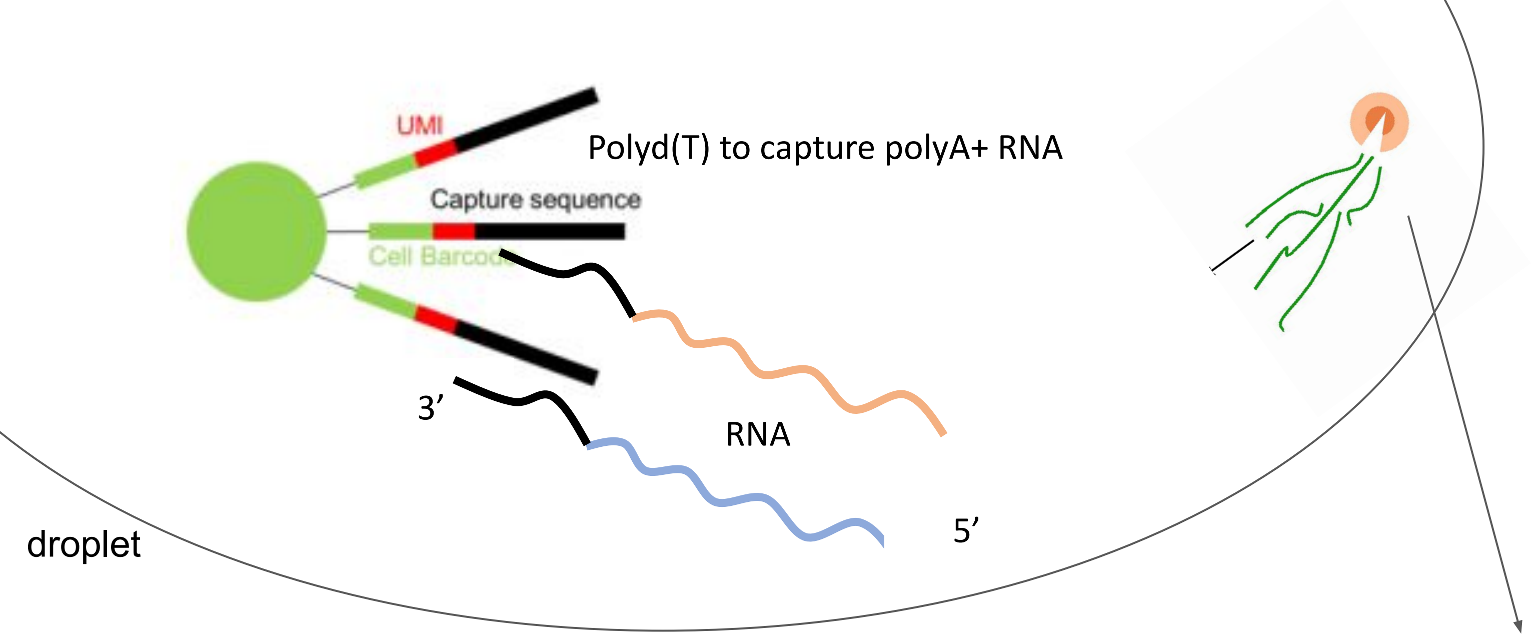
The BC+UMI = 28bp enables to distinct any RNA and assign it to its cell of origin. example : another droplet has a blue gel bead, with a different barcode from the green gel bead.

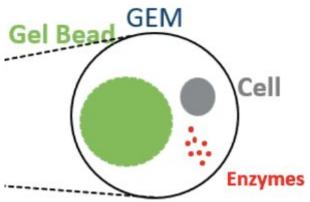
Cell Barcode (16bp) = sequence specific to each bead (so each cell)

UMI (12 bp) = sequence specific to each molecule : Unique Molecular Identifier



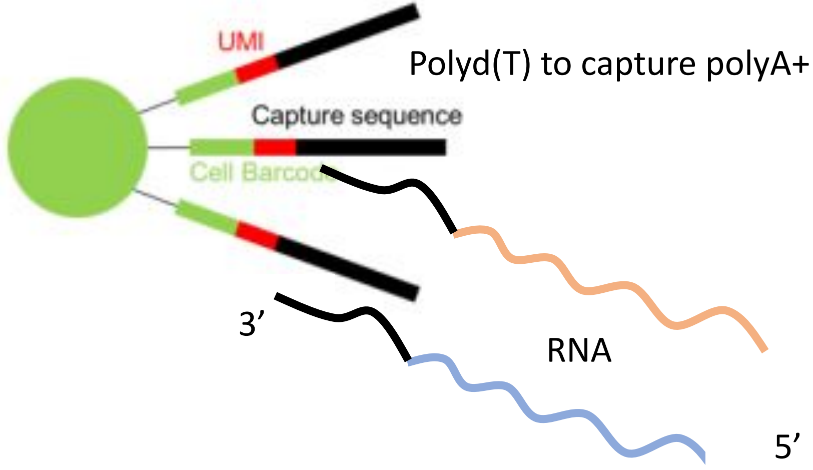
scRNA-seq with 10X targets 3' end of the transcripts





The cell is then lysed within the droplet. RNA is released (but contained in the droplet). polyA+ RNA are captured from the 3'end on the polyd(T) sequence

scRNA-seq with 10X targets 3’ end of the transcripts



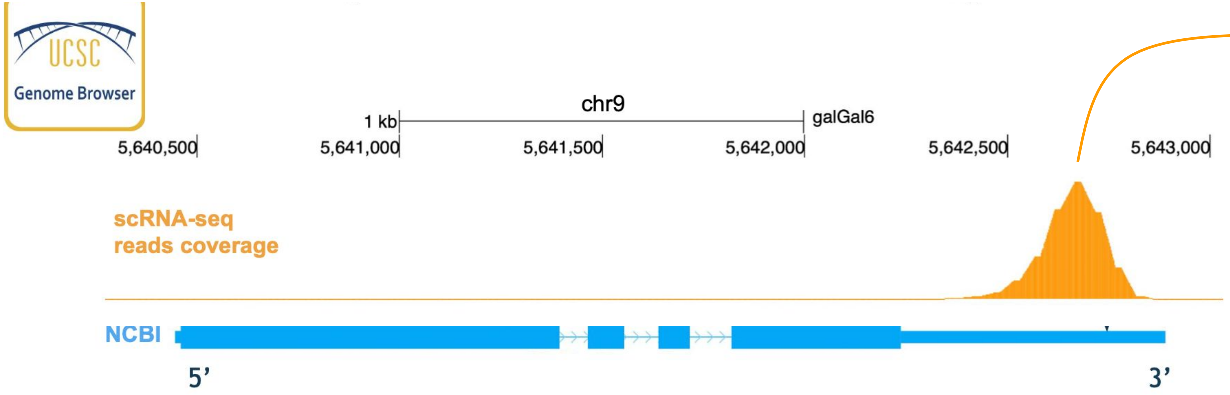
Polyd(T) to capture polyA+ RNA

3’

RNA

5’

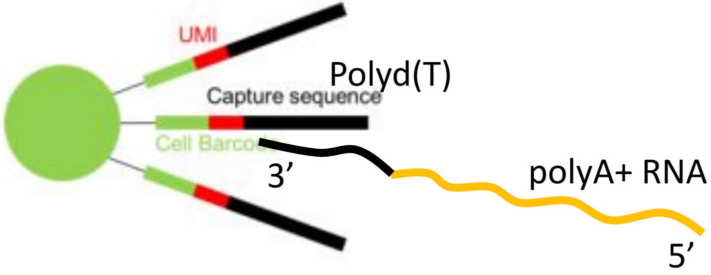
The signal is biased towards the 3’ end of the transcript (more about that tomorrow)



scRNA-seq signal

10X scRNA-seq sequenced in short reads

At the end of this step, what actually comes out of the 10X Genomics device ?

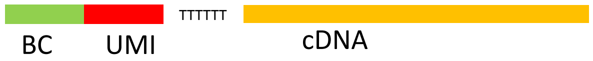


Cell Barcode = sequence specific to each cell

UMI = sequence specific to each molecule : **U**nique **M**olecular **I**dentifier

10X scRNA-seq sequenced in short reads

full length cDNA

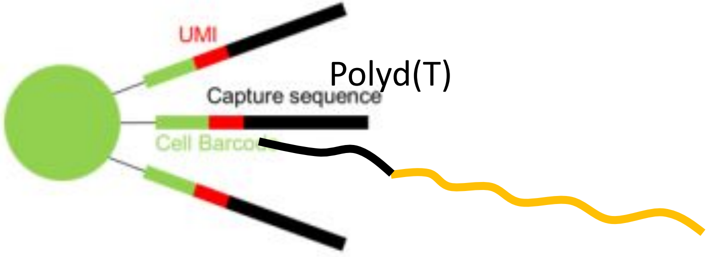


BC

UMI

TTTTTT

cDNA



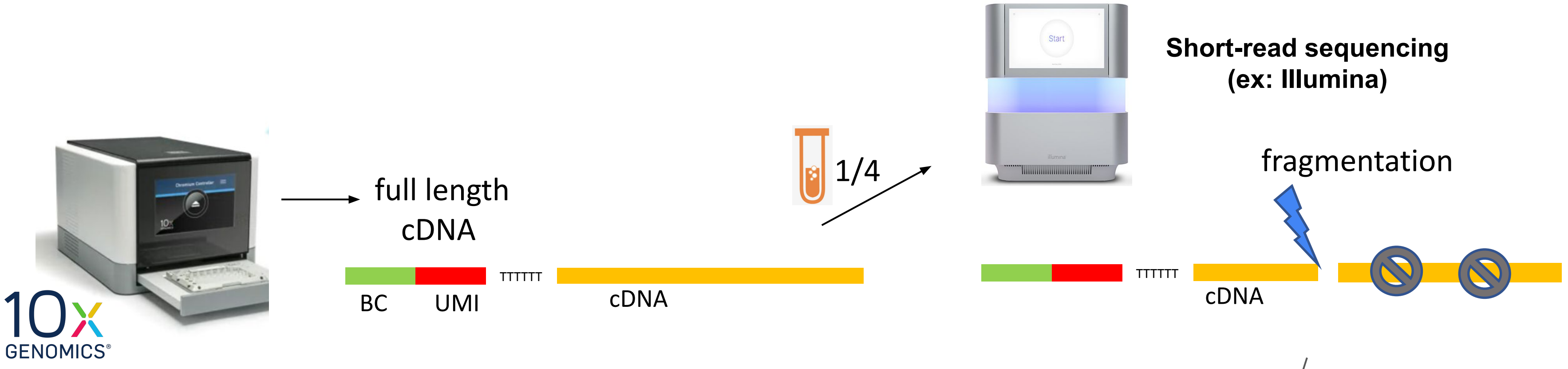
Polyd(T)

A molecule construction made of the barcode + UMI + polyd(T) + cDNA corresponding to the captured RNA

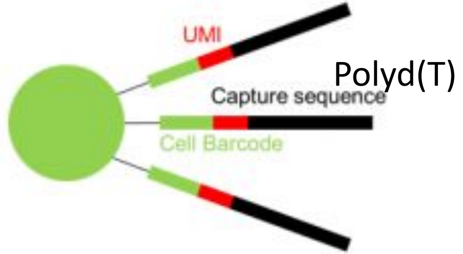
Cell Barcode = sequence specific to each cell

UMI = sequence specific to each molecule

10X scRNA-seq sequenced in short reads



Short-read sequencing (ex: Illumina)

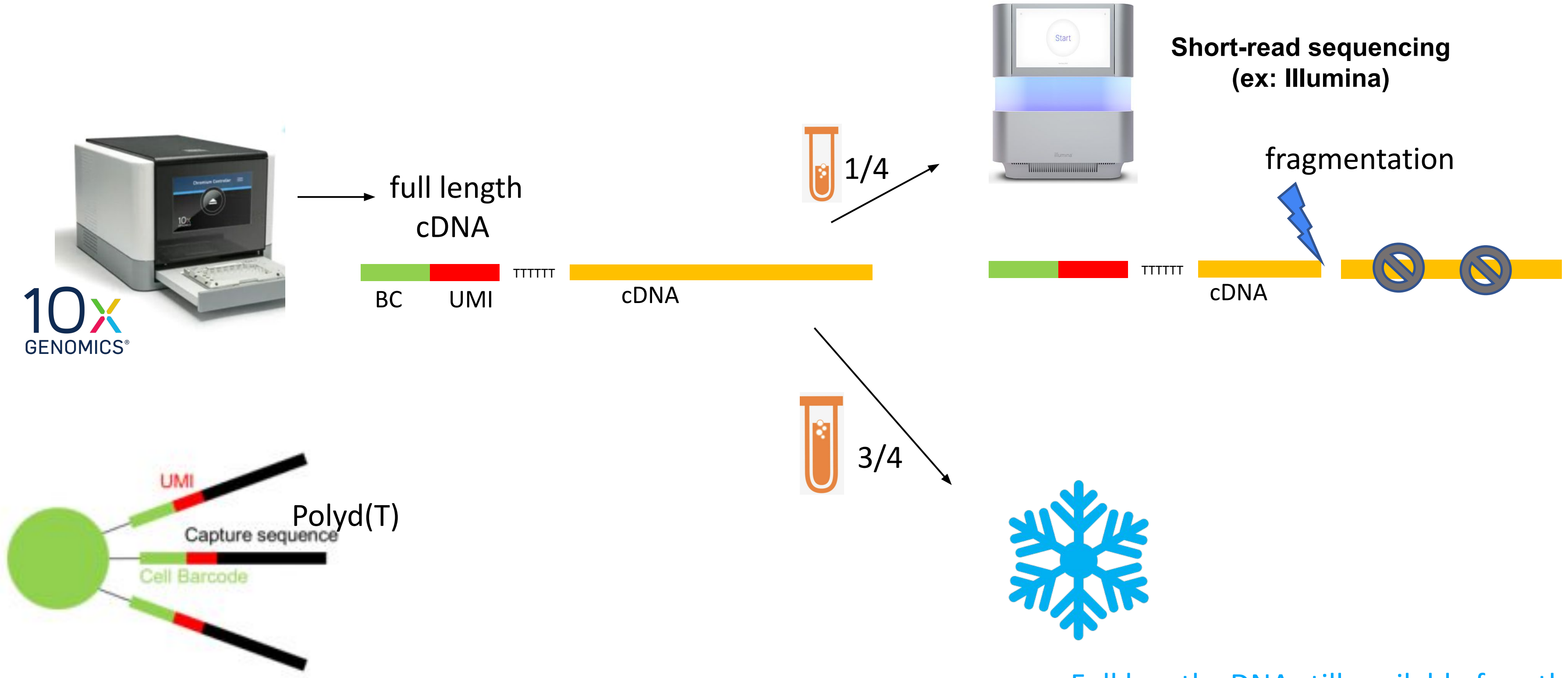


Cell Barcode = sequence specific to each cell

UMI = sequence specific to each molecule

During the library preparation for short-read sequencing, there is a fragmentation step. Only the cDNA portion linked to the BC+UMI+polyd(T) is conserved for sequencing

10X scRNA-seq sequenced in short reads

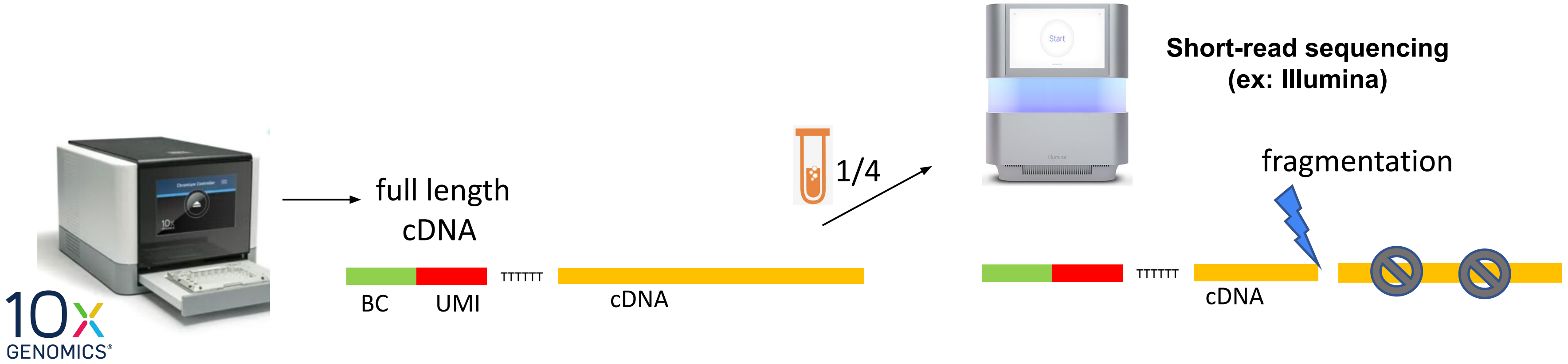


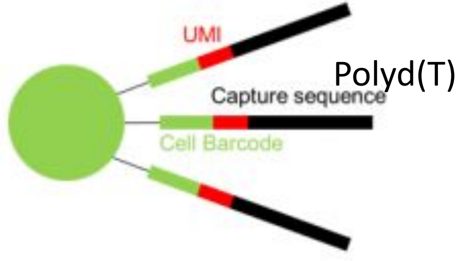
=> Full length cDNA still available for other sequencing (ex: long-read)

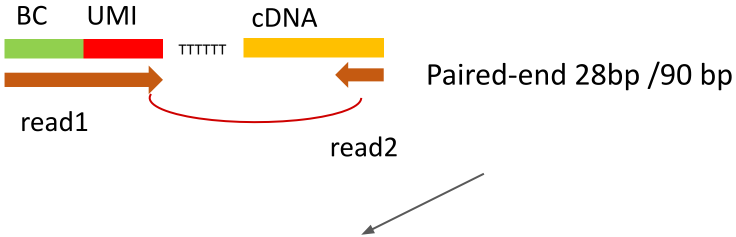
Cell Barcode = sequence specific to each cell

UMI = sequence specific to each molecule

10X scRNA-seq sequenced in short reads







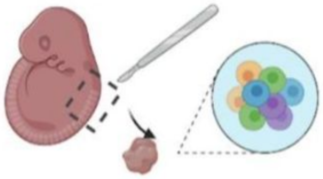
Cell Barcode = sequence specific to each cell

UMI = sequence specific to each molecule

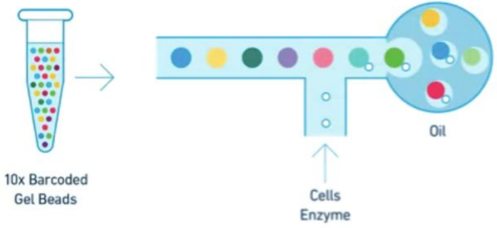
The library is sequenced in paired-end. The read 1 contains the BC+UMI (28bp). The read2 contains a 90bp portion of cDNA. Only read2 corresponds to genomic/biological DNA. Read1 stems from synthetic molecules, not the transcriptome.

10X scRNA-seq in a nutshell

Tissue dissection + cell dissociation

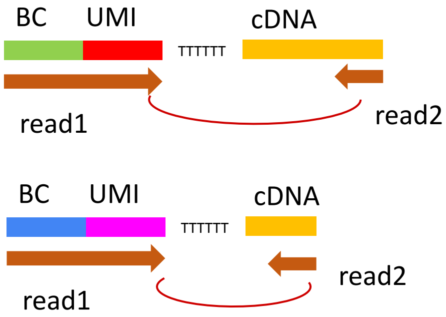


Cell partitioning + mRNA capture



Library preparation + sequencing

Paired-end 28bp /90 bp



Biases/limitations of 10X Genomics technology

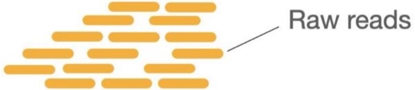
- Only the 3’ end is sequenced (with short-read protocol + 3’ kit)
- Cell size < 30um otherwise clog microfluidic channels
- 30% polyA+ transcripts captured per cell
- A droplet may contain 2 cells (= doublet)
- Some cell sub-population may be completely depleted/unfound

Considerations on experiments

- **Fresh cells** : time between dissociation and 10X experiment should be <30min, otherwise cells start to die and result in RNAs wrongly assigned to cells (RNA "soup") and many expressed genes linked to cell death
- **Frozen cells** : does not work on all cells
- **FFPE** : only in human + mouse, restricted to certain tissues
- **Dissociation + Fixation** with ACME protocol (acetic acid + methanol + glycerol): requires optimisation but successful on exotic species (GenomiqueENS)
- Charge a bit more cells (25,000)
- Many tests have been done on **PBMCs** (immune cells) that are natively dissociated. Results do not necessarily reproduce on cells dissociated from tissues
- **Q&A** section of 10X website is very informative : https://kb.10xgenomics.com/hc/en-us/categories/360000149952-Single-Cell-Gene-Expression

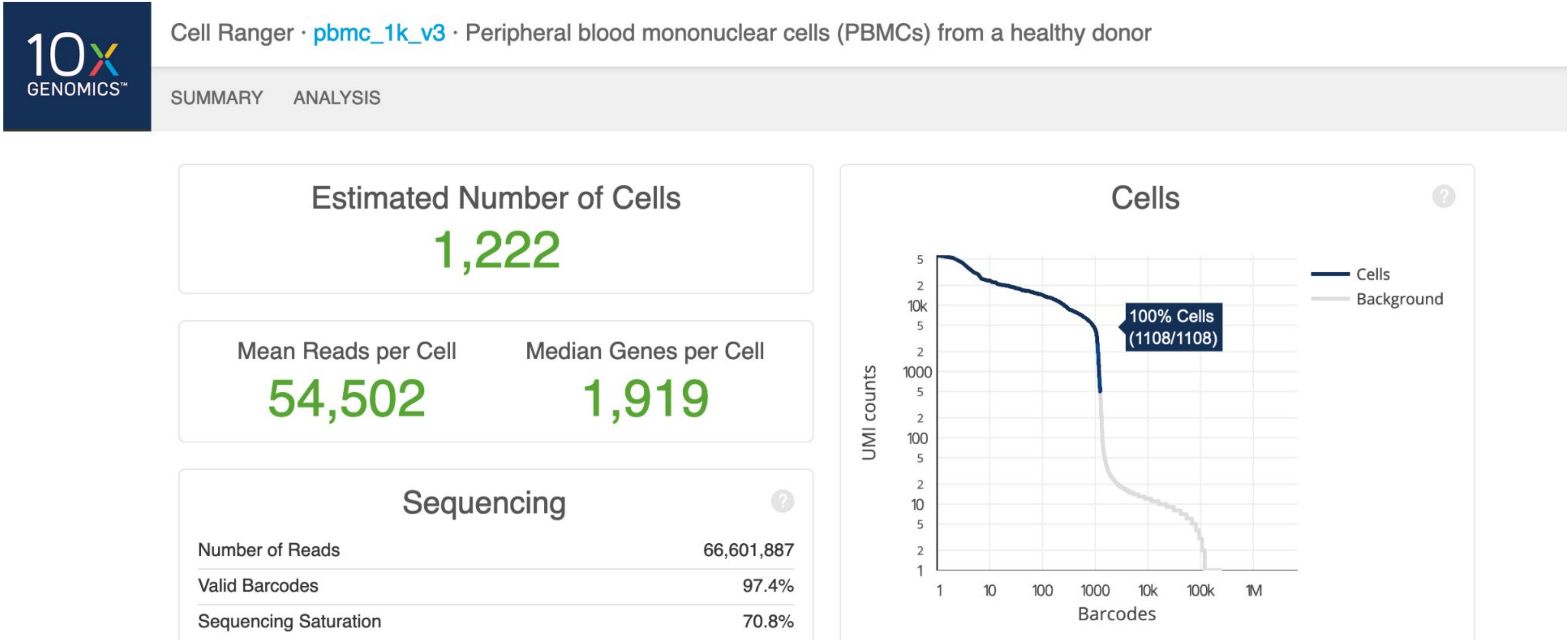
ACME : García-Castro et al 2021 : https://doi.org/10.1186/s13059-021-02302-5

Bioinformatics analysis of 10X Genomics scRNA-seq dataset



Which result file(s) did you obtain from the sequencing core facility ?

Bioinformatics analysis of 10X Genomics scRNA-seq dataset



Cell Ranger · pbmc_1k_v3 · Peripheral blood mononuclear cells (PBMCs) from a healthy donor

SUMMARY ANALYSIS

Estimated Number of Cells

1,222

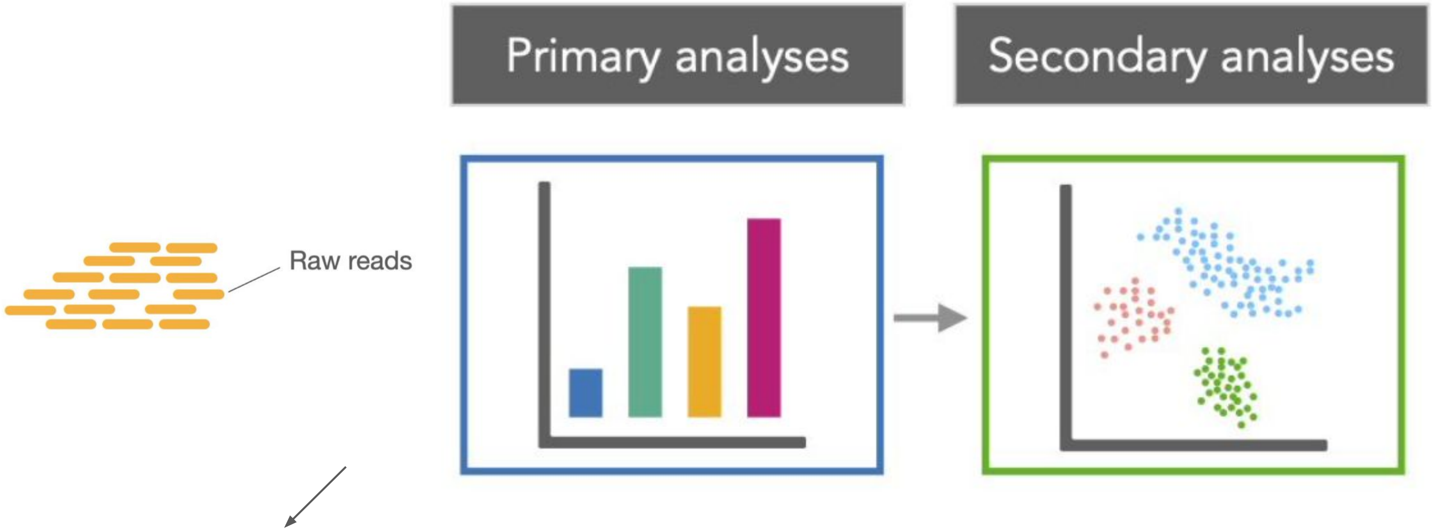
Mean Reads per Cell	Median Genes per Cell
54,502	1,919

Sequencing

Number of Reads	66,601,887
Valid Barcodes	97.4%
Sequencing Saturation	70.8%

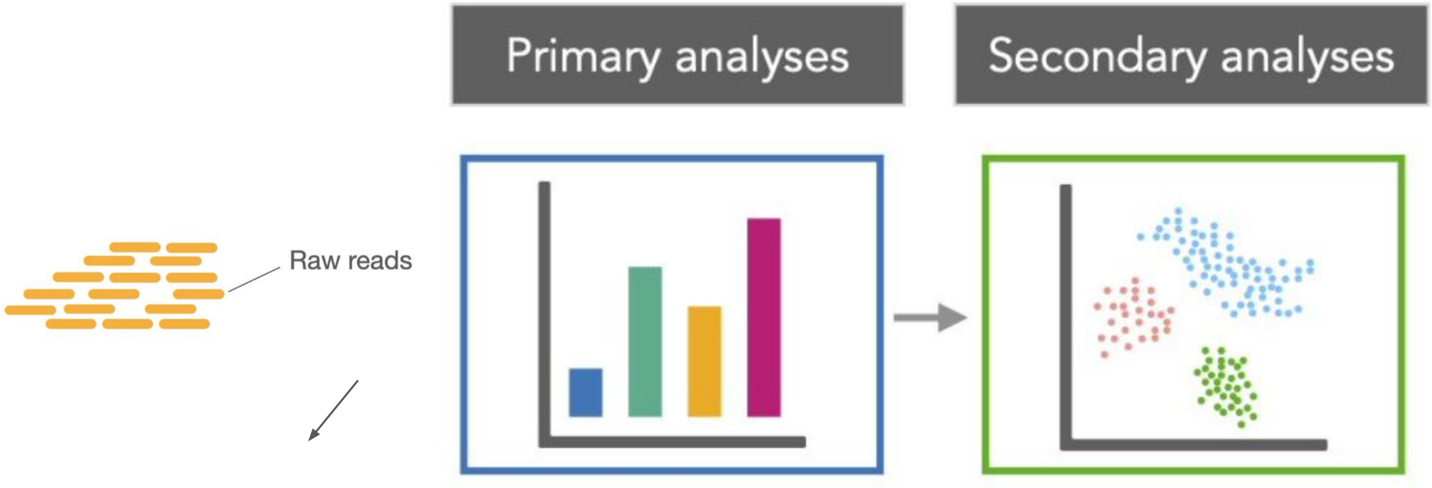
Cells

Bioinformatics analysis of 10X Genomics scRNA-seq dataset



Raw data are the sequence reads. Then the bioinformatics analysis are in 2 phases : Primary (= preprocessing) and secondary

Bioinformatics analysis of 10X Genomics scRNA-seq dataset

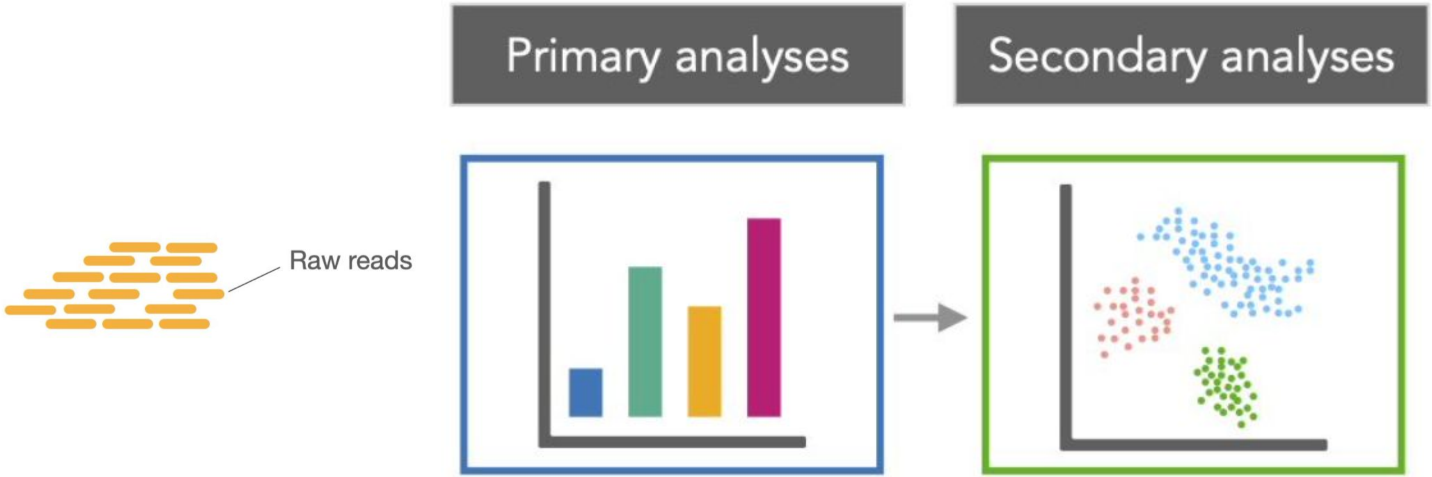


Cell Ranger is the program developed by 10X Genomics to perform the primary analysis (and a bit of secondary)

CellRanger

10x GENOMICS®

Bioinformatics analysis of 10X Genomics scRNA-seq dataset



Raw reads

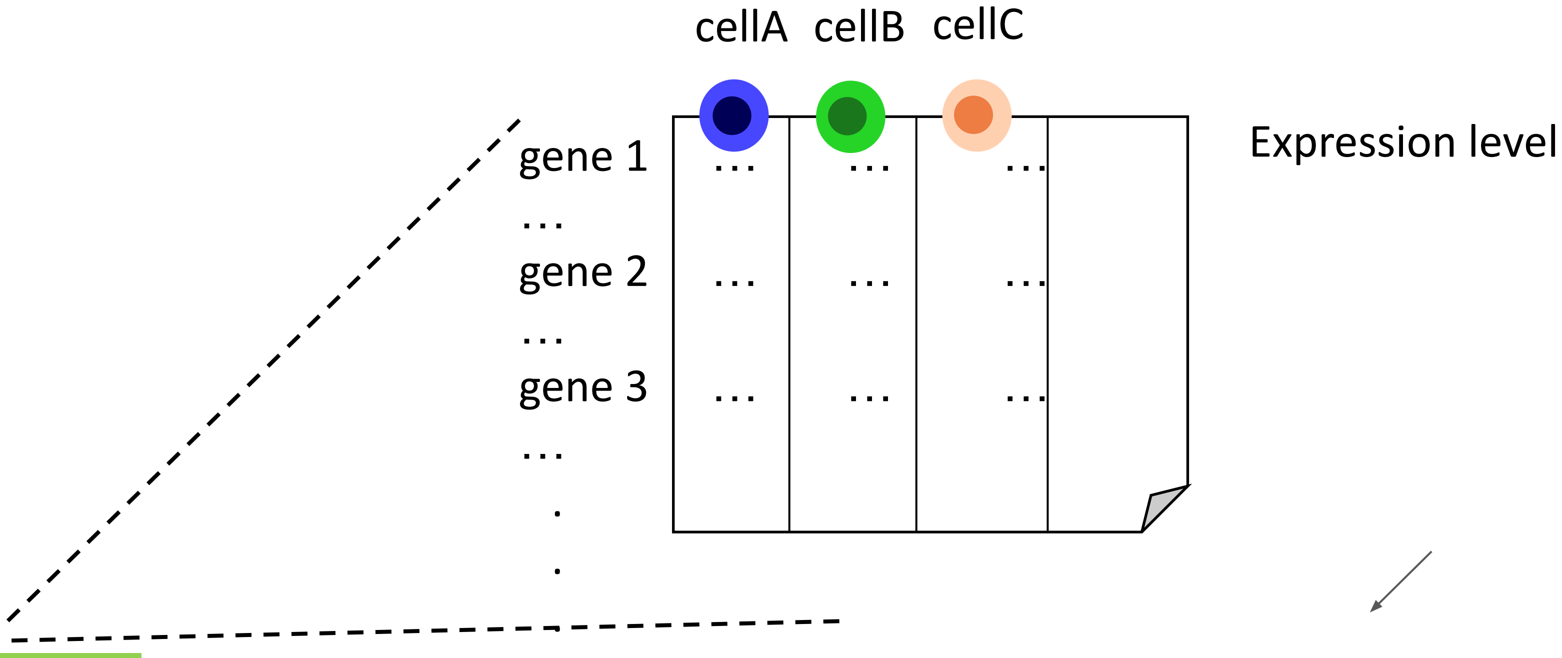
Primary analyses

Secondary analyses

Count matrix

At the end of the primary analysis, the count matrix is produced. The count matrix serves as input for all subsequent secondary analyses

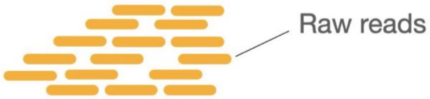
What is a count matrix ?



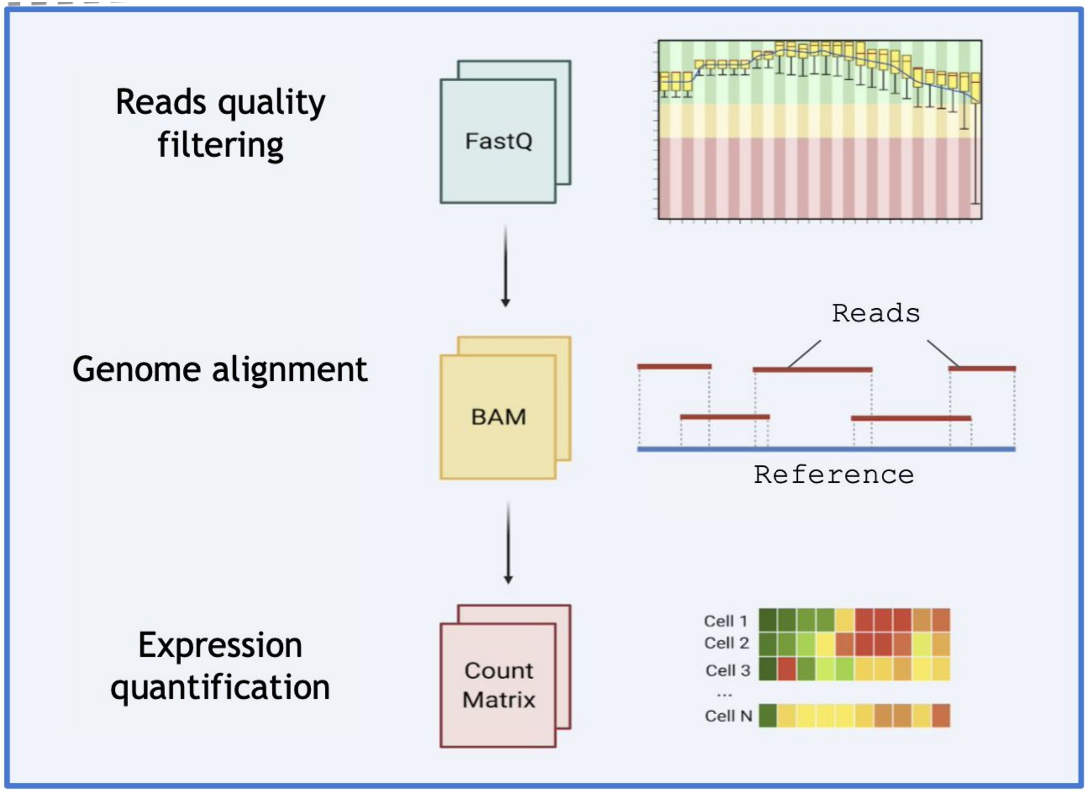
Count matrix

A table with all annotated genes as rows and all cells as columns. The content of the table are expression levels (measured as read counts)

The processing steps that are often overlooked

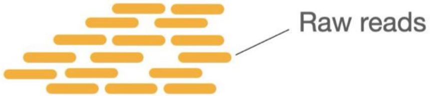


Primary analyses



These steps are often overlooked because they do not produce results linked to the biological questions, and considered as “gory details”.

Primary analyses : Reads

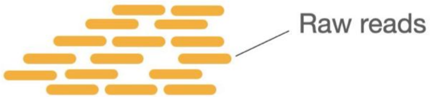


Primary analyses

- Results starts by a **BCL** file (raw base calling from the sequencer). This file needs to be treated to produce the **FASTQ** files containing the reads
- This steps is done by the program bcl2fastq from Illumina (step “mkfastq” in CellRanger)

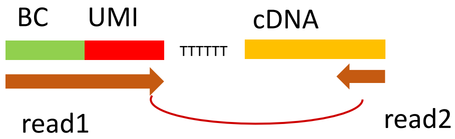
this is a detail, we just indicate it here in case you read further about Cell Ranger and step upon the notion of BCL

Primary analyses : Reads quality checking



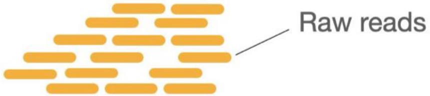
Primary analyses

Paired-end 28bp /90 bp



- 2 FASTQ files :
  - one contains all the read1
  - one contains all the read2

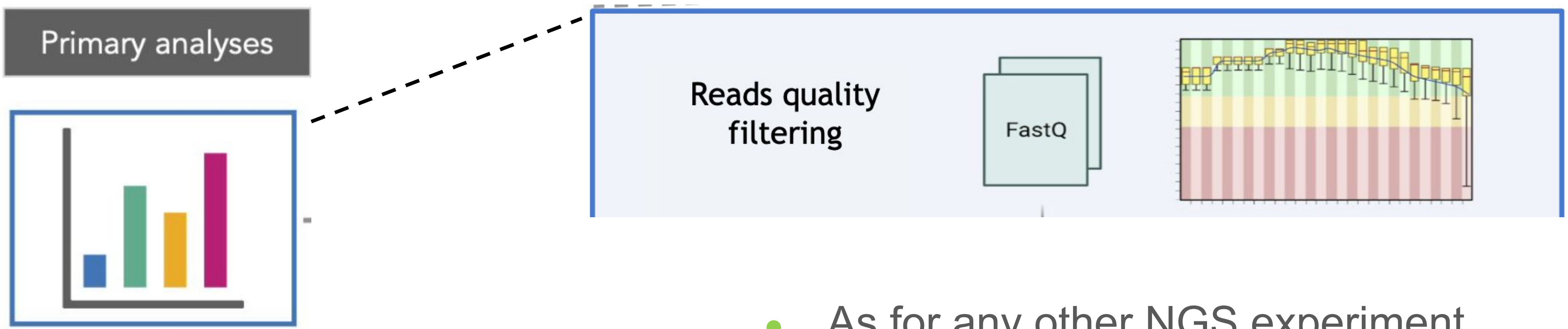
Example dataset from 10X Genomics



- **Dataset name :** pbmc_1k_v3 => 1000 human peripheral blood mononuclear cells (PBMCs) in human, freely available from 10X genomics website
- **2 files** :
    - pbmc_1k_v3_S1_L001_R1_001.fastq.gz
    - pbmc_1k_v3_S1_L001_R2_001.fastq.gz

10X provides other public datasets, for each application, chemistry…

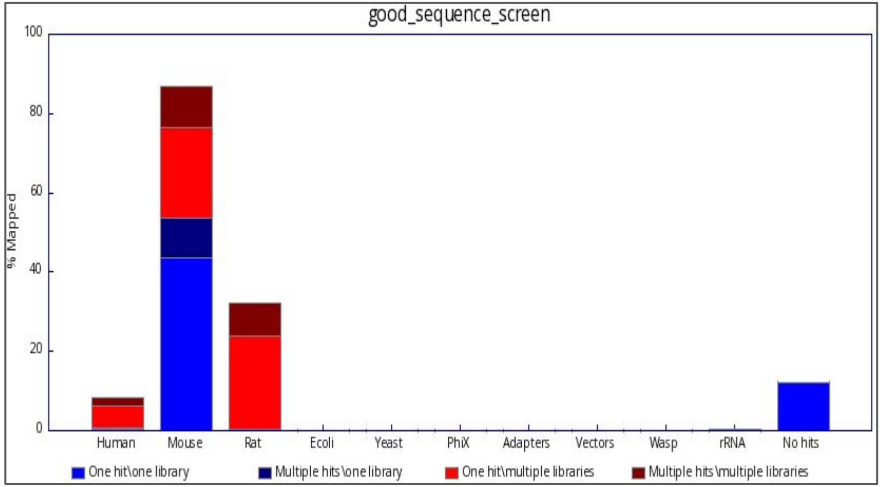
Primary analyses : Reads quality checking



Primary analyses

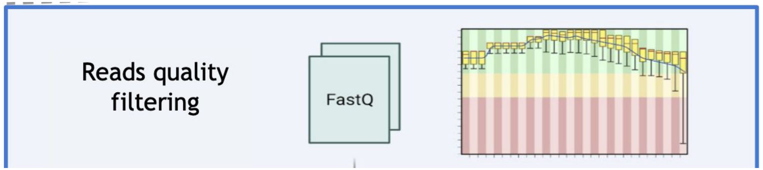
Reads quality filtering

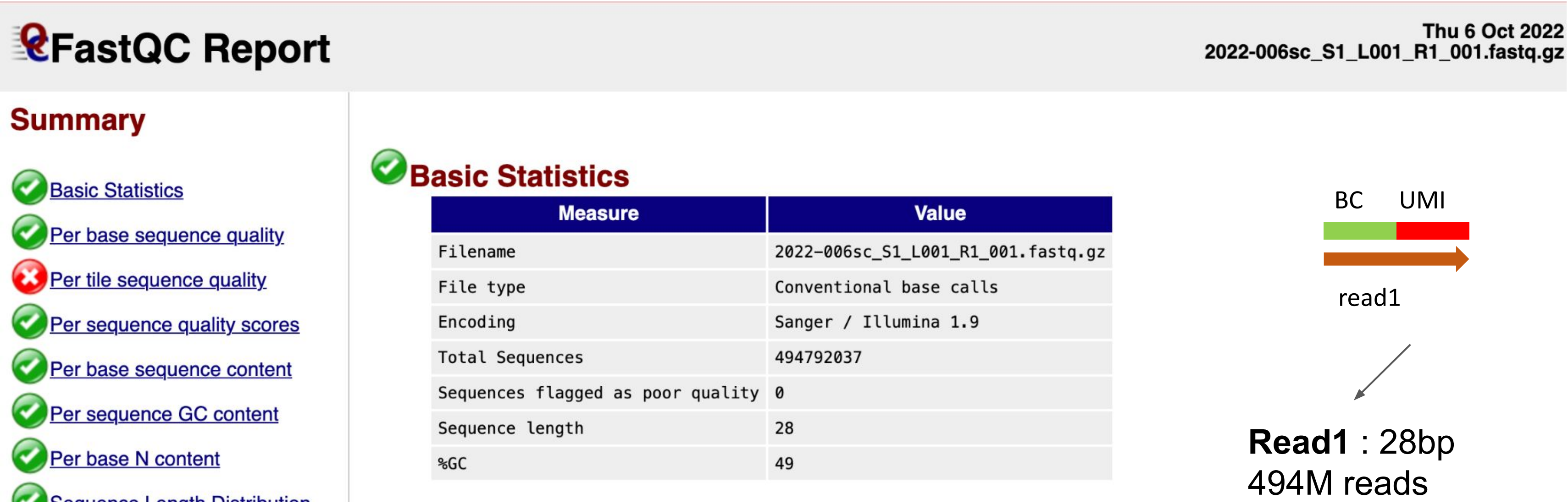
FastQ



- As for any other NGS experiment, check the quality of the reads with FASTQC.
- FastqScreen enables to check for contaminations with other organisms
- These steps are usually done by the sequencing core facility, ask for these results if not provided

Example dataset from 10X Genomics





FastQC Report

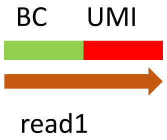
Thu 6 Oct 2022
2022-006sc_S1_L001_R1_001.fastq.gz

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution

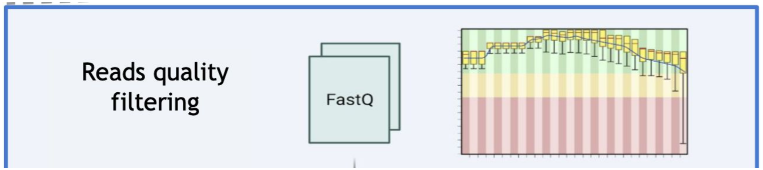
✓ Basic Statistics

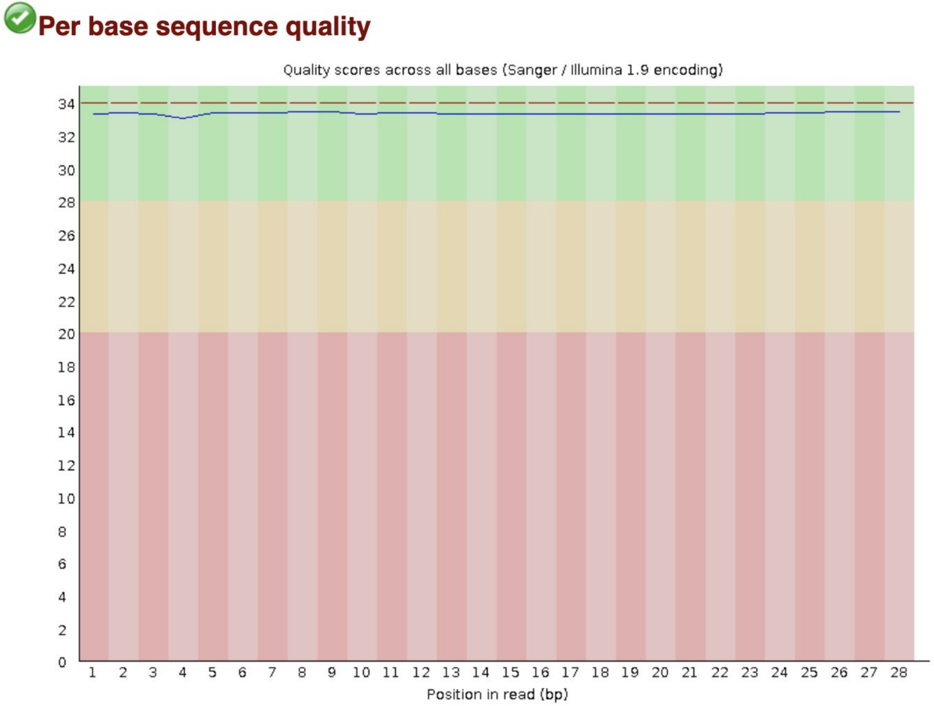
Measure	Value
Filename	2022-006sc_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	494792037
Sequences flagged as poor quality	0
Sequence length	28
%GC	49



Read1 : 28bp
494M reads

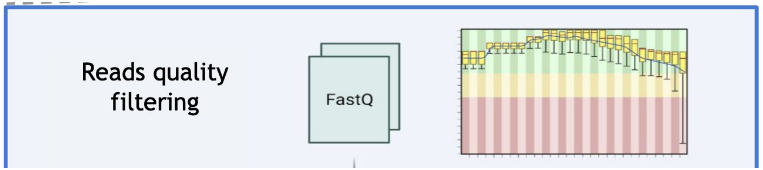
Example dataset from 10X Genomics

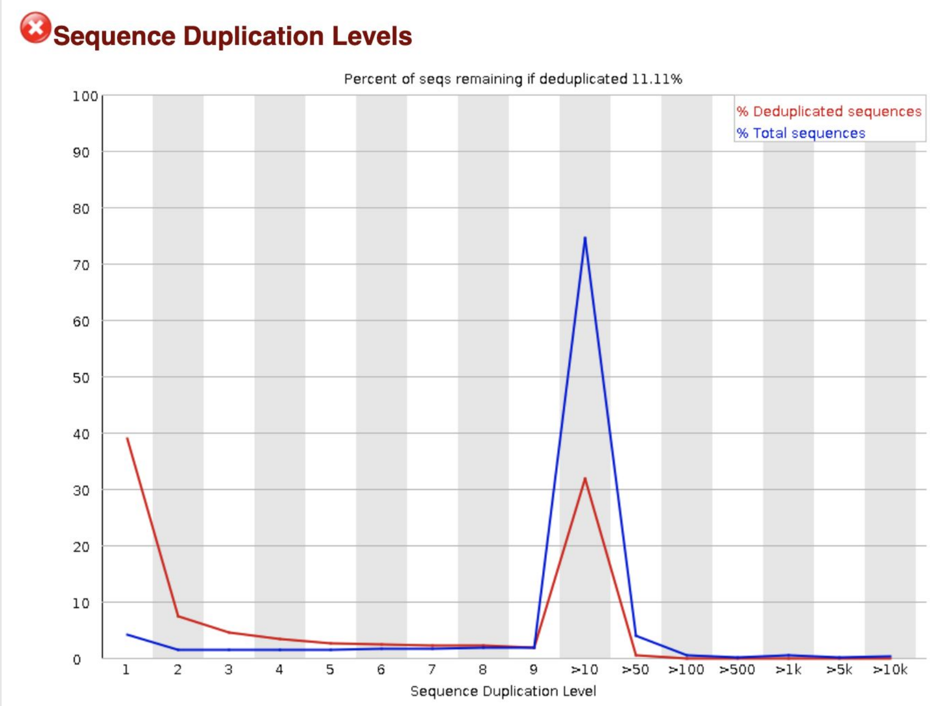



Per base sequence quality

quality is excellent

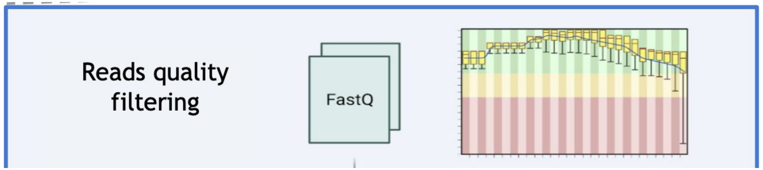
Example dataset from 10X Genomics

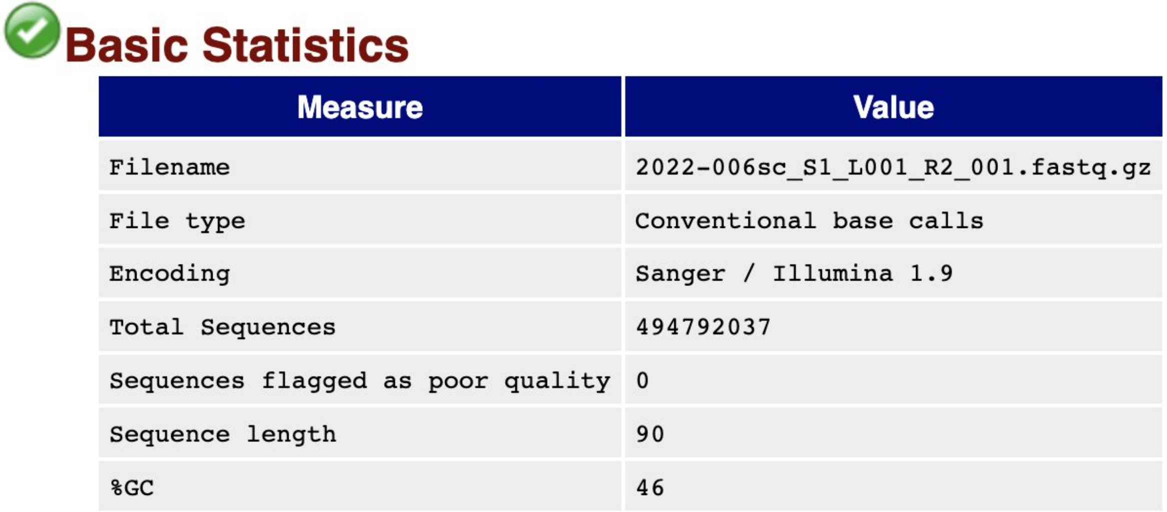




normal to have duplication level because some BC+UMI have amplification biaises

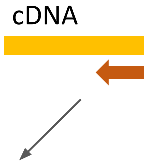
Example dataset from 10X Genomics





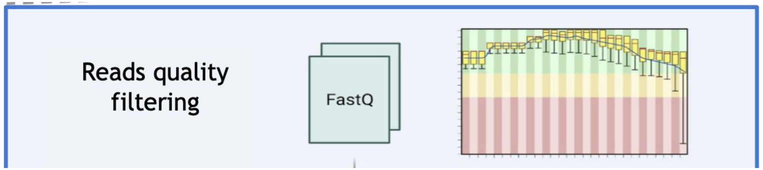
Basic Statistics

Measure	Value
Filename	2022-006sc_S1_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	494792037
Sequences flagged as poor quality	0
Sequence length	90
%GC	46

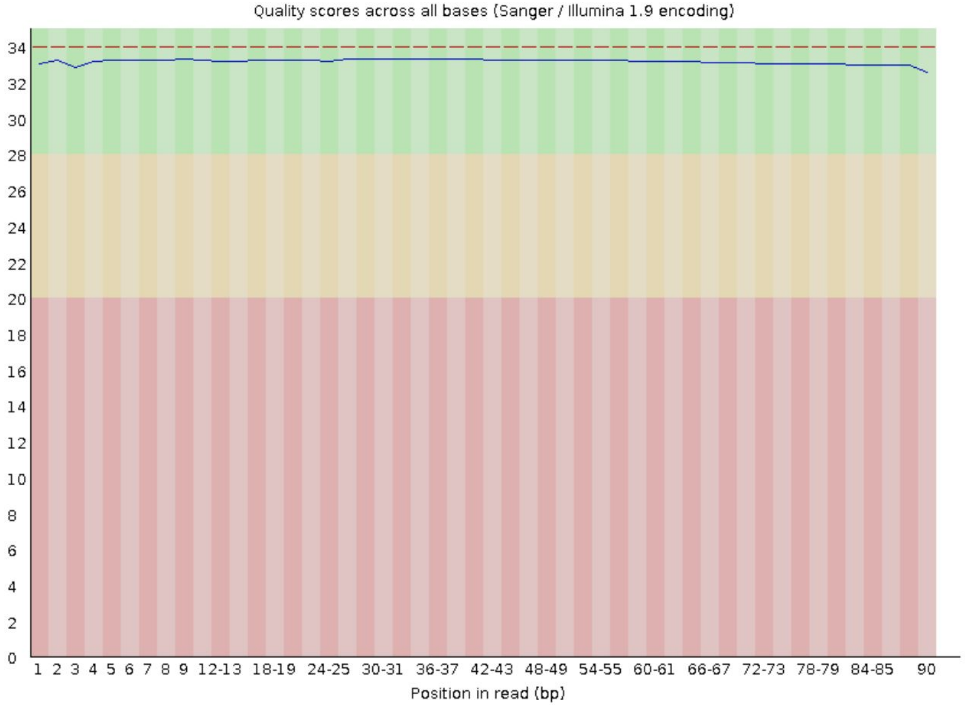


Read2 : 90bp
494M reads

Example dataset from 10X Genomics

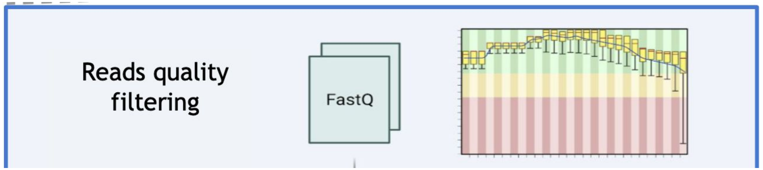


✓ **Per base sequence quality**

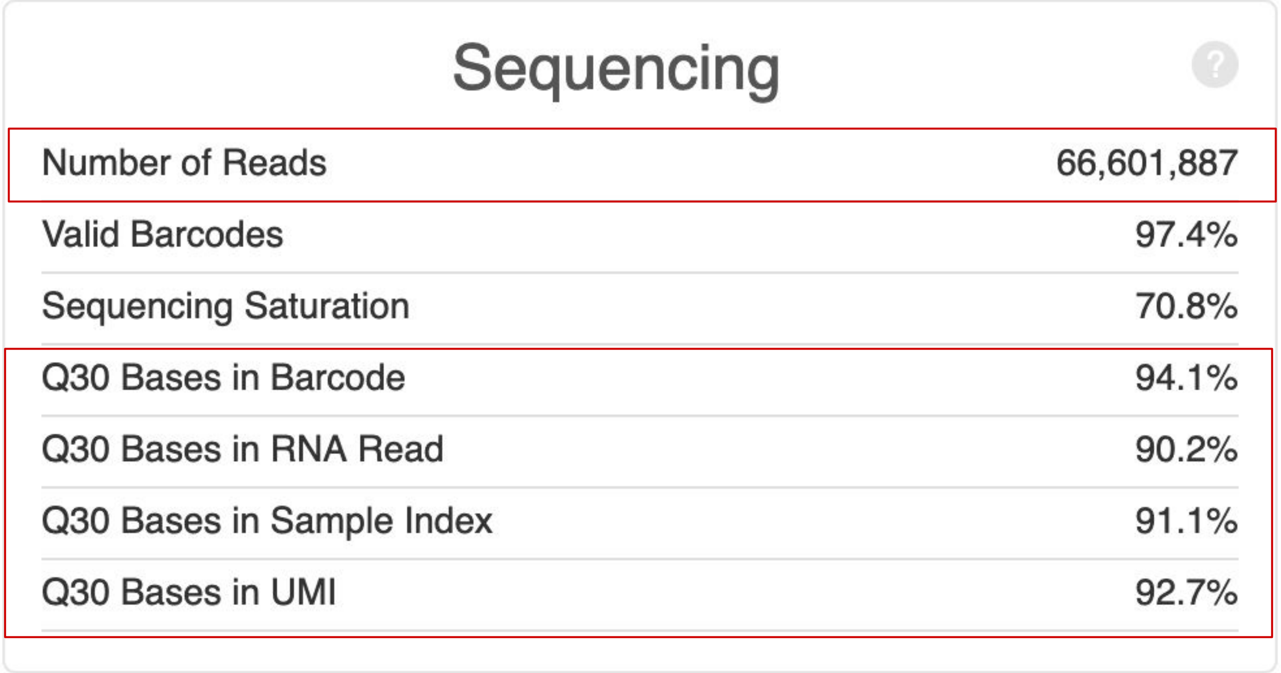


quality is excellent

Example dataset from 10X Genomics



In **CellRanger** report

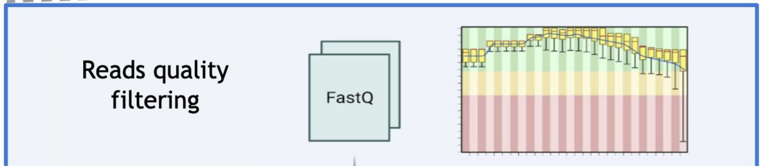


Sequencing

Number of Reads	66,601,887
Valid Barcodes	97.4%
Sequencing Saturation	70.8%
Q30 Bases in Barcode	94.1%
Q30 Bases in RNA Read	90.2%
Q30 Bases in Sample Index	91.1%
Q30 Bases in UMI	92.7%

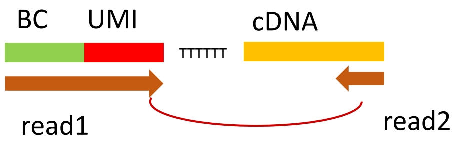
check the “sequencing” section of the report. The Q30 means “very high quality of bases”

Primary analyses : Reads quality checking



- Make sure read1 is of high quality because it contains the BC + UMI, later used to trace back the cell from which originates the RNA
- Ns and highly repeated sequences would impair read assignment.
- Any wrong base => lost read and barcode

Paired-end 28bp /90 bp



BC UMI

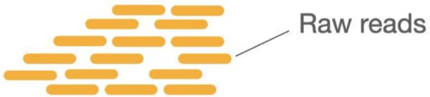
TTTTTT

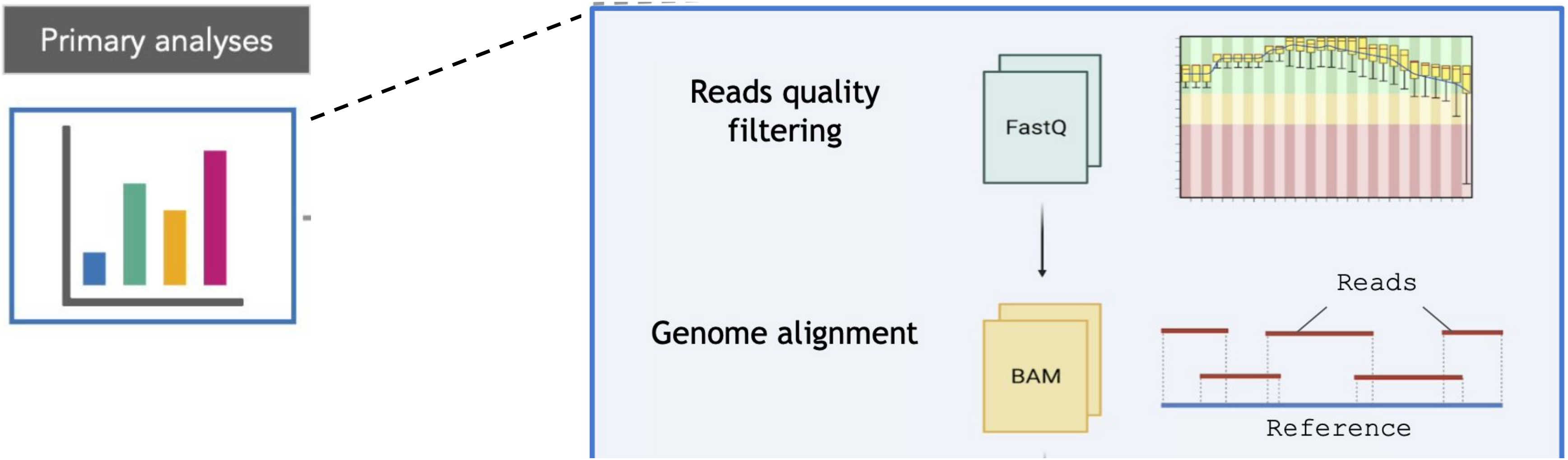
cDNA

read1

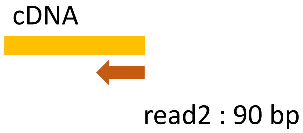
read2

Primary analyses : Mapping

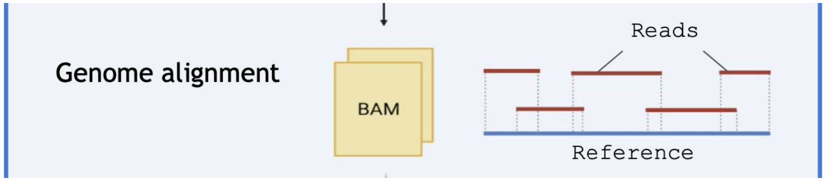


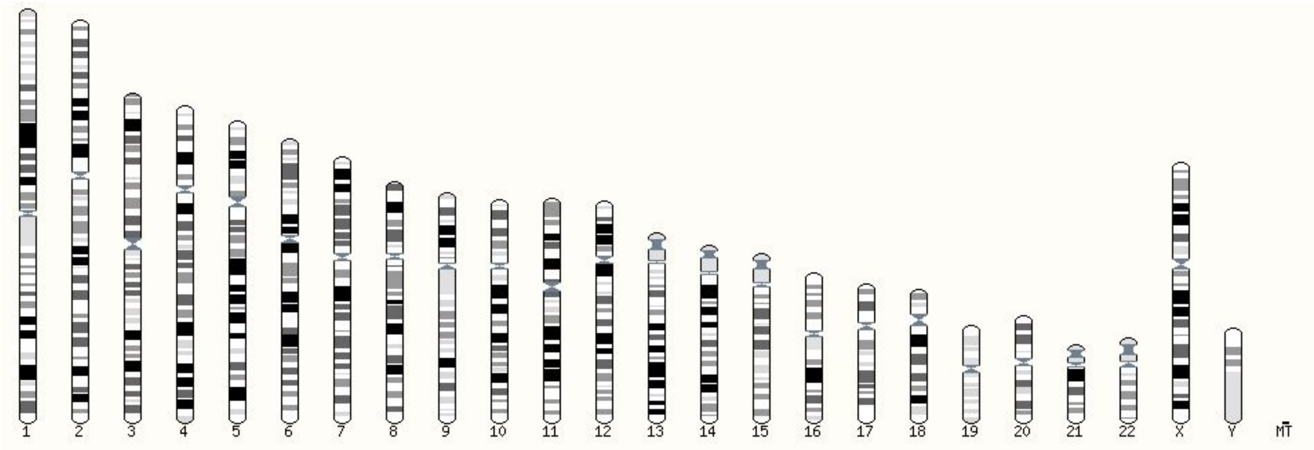


- Read1 and Read2 are then treated separately
- **Read2** corresponds to genomic sequence => **mapping step** (=infer the position on the genome from which the read originates)



Primary analyses : Mapping

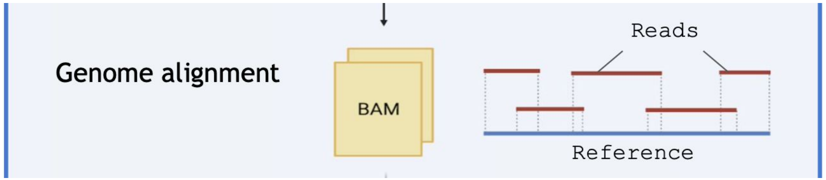


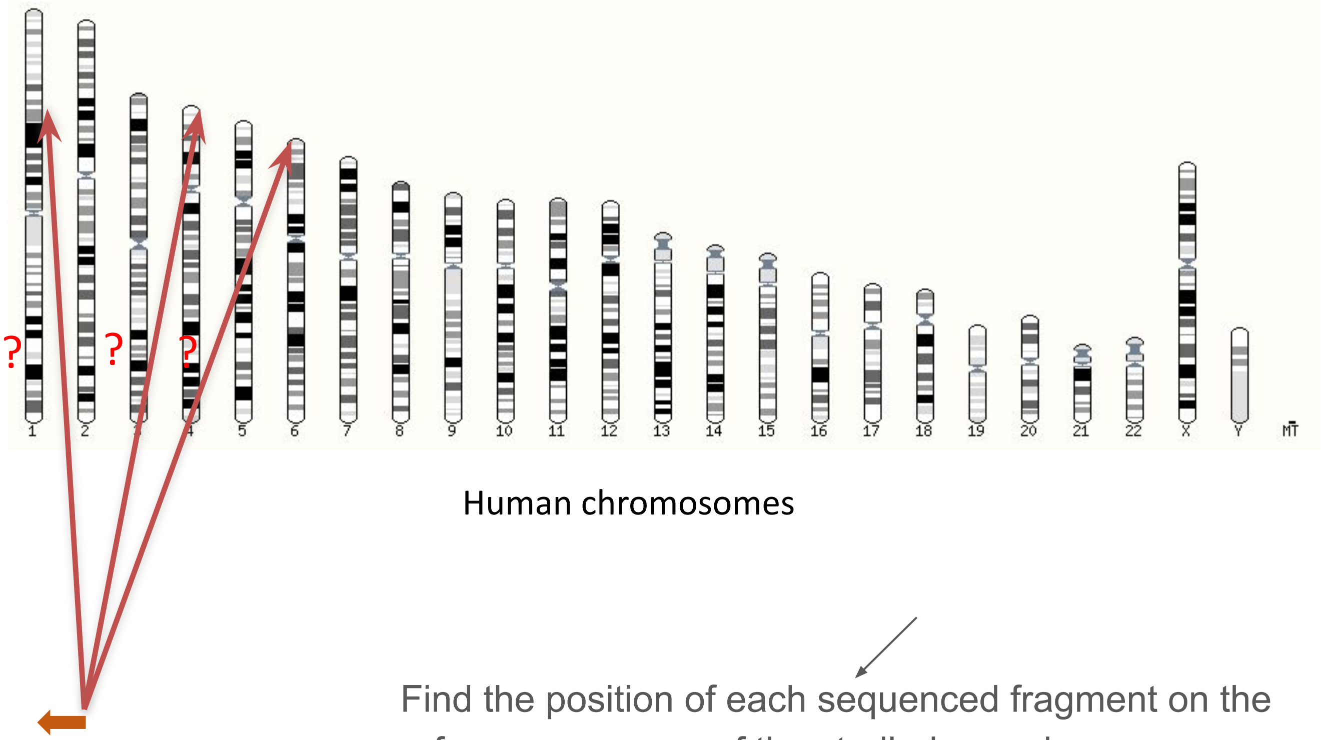


Human chromosomes

read2 : 90 bp

Primary analyses : Mapping

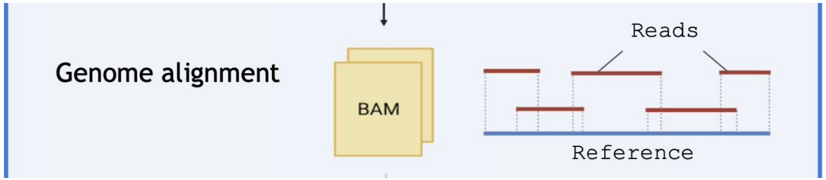


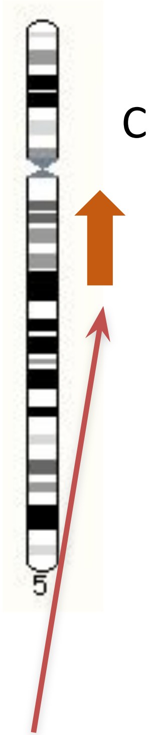


Human chromosomes

Find the position of each sequenced fragment on the reference genome of the studied organism

Primary analyses : Mapping





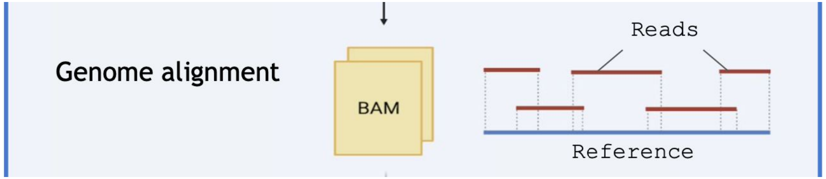
Chromosome 5

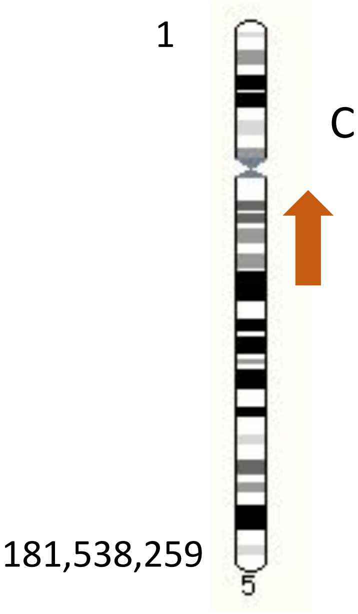
alignment

read2 : 90 bp

The best alignment is found for this read over the whole genome. Here it is on chromosome 5

Primary analyses : Mapping

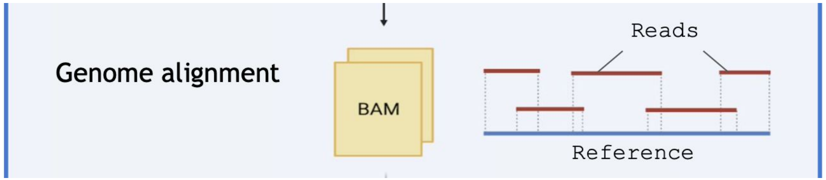




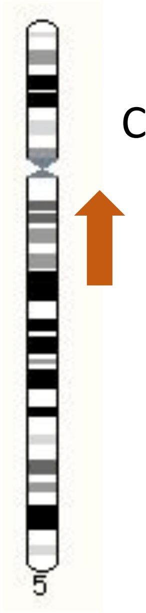
Chromosome 5

To precise the localisation of the best alignment, a coordinate system is used. First, each position of the chromosome has a particular value, corresponding to its distance from the beginning of the chromosome

Primary analyses : Mapping



1



181,538,259

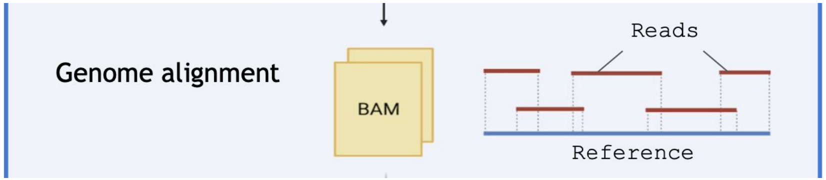
Chromosome 5

125,326,318 start

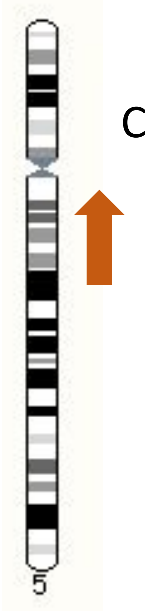
125,326,358 end

The region of alignment has a **start** and **end** position + strand orientation

Primary analyses : Mapping



1



181,538,259

Chromosome 5

125,326,318 start

125,326,358 end

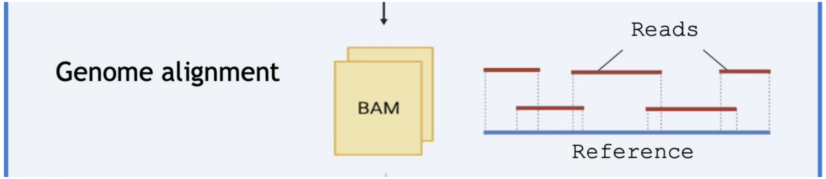
Genomic coordinates:

chr5 125326318 125326358 -

The genomic coordinates is like “GPS coordinates” to locate regions on a genome. The format is :

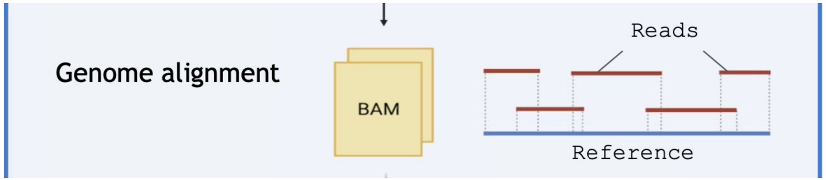
```
chromosome start end strand
```

Primary analyses : Mapping

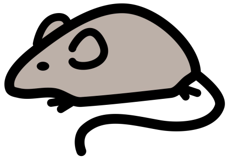


- The mapping step enables to obtain the genomic coordinates of all reads2 for which an alignment has been found.
- The output file is in **BAM** format
- Not all reads can be aligned (contaminations, differences between the sample and reference genome, …)
- Programs that perform this mapping step are often called “mappers”

Primary analyses : Mapping

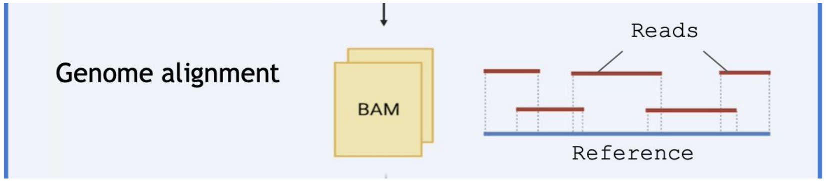


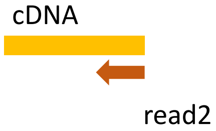
- CellRanger internally uses **STAR** as the program to align the reads on the reference genome
- The reference genome must be provided in the form of an **index**
- Ready-to-use genomes index:
  - human (hg19, GRCh38)
  - mouse (mm10)
  - both (xenografts)
- For other organisms :
  - Use the genome in FASTA format
  - convert it with *cellranger mkgtf* and *cellranger mkref*.
- If you use some specific sequences (transgenes), don’t forget to provide the sequence and rebuild the index ! (otherwise, no reads will be mapped to this region)



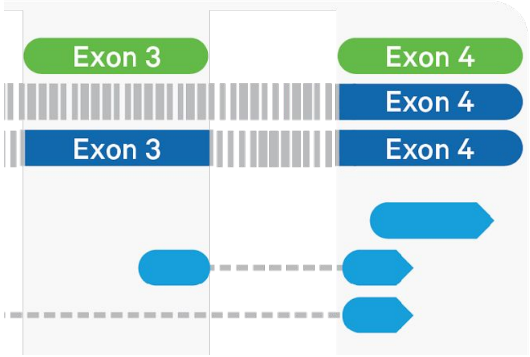
working with user-specific sequences or genomes requires more work because the genome index must be built (computer-intensive)

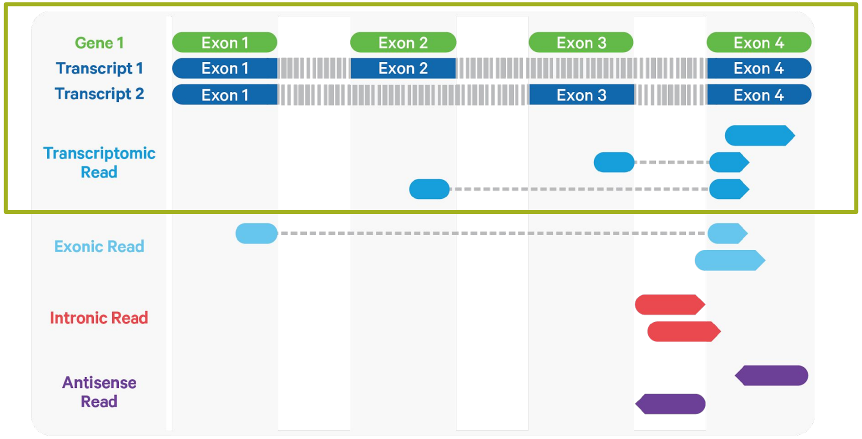
Primary analyses : Mapping





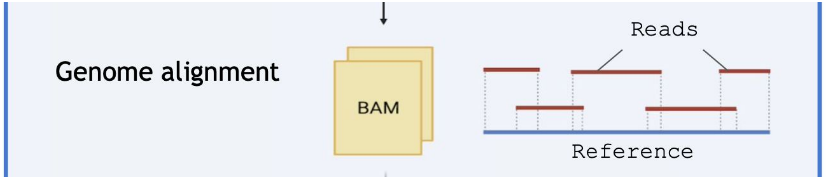
- STAR deals with RNA splicing, a read can be artificially “cut” to map to distant regions from which it originates (=different exons)

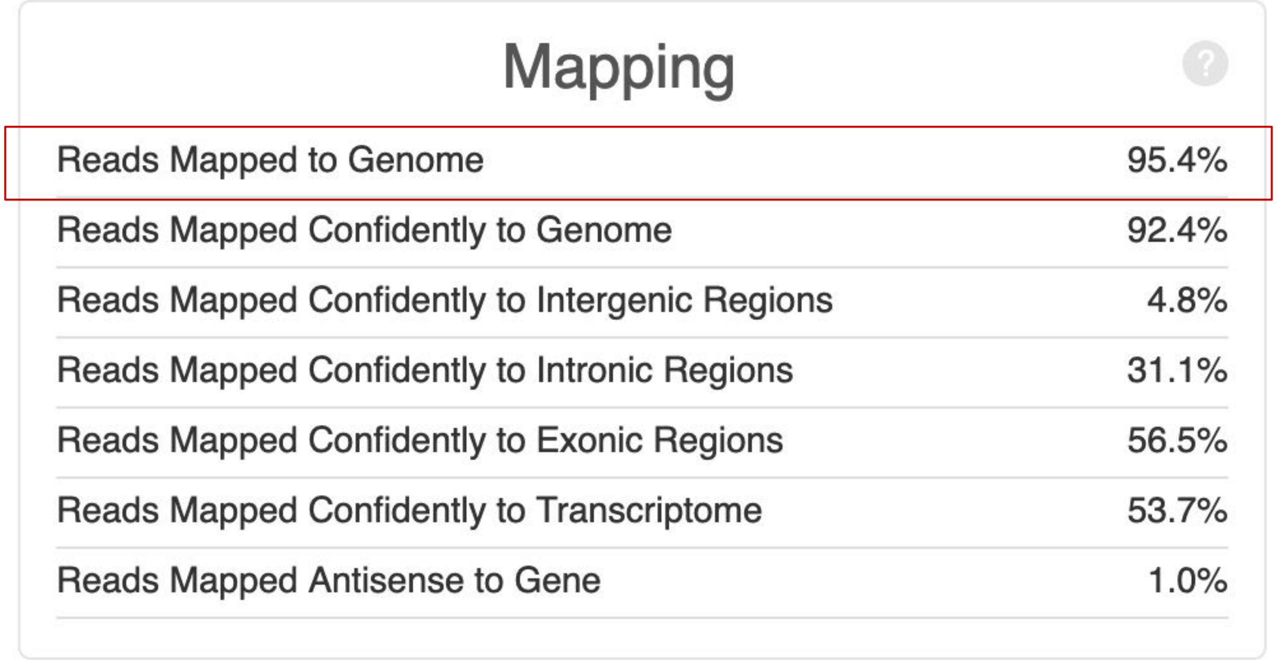




Example dataset from 10X Genomics

In CellRanger report



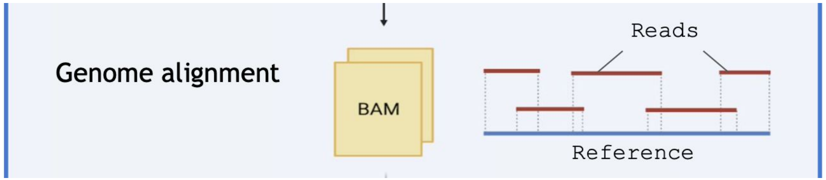


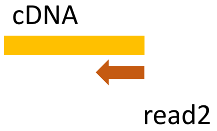
Mapping

Reads Mapped to Genome	95.4%
Reads Mapped Confidently to Genome	92.4%
Reads Mapped Confidently to Intergenic Regions	4.8%
Reads Mapped Confidently to Intronic Regions	31.1%
Reads Mapped Confidently to Exonic Regions	56.5%
Reads Mapped Confidently to Transcriptome	53.7%
Reads Mapped Antisense to Gene	1.0%

It is normal to have <100% reads aligned to the genome, because the reference genome is not exactly the genome of the studied sample. % will decrease with huge rearrangements (cancer or cell lines) or many SNPs (wild animals)

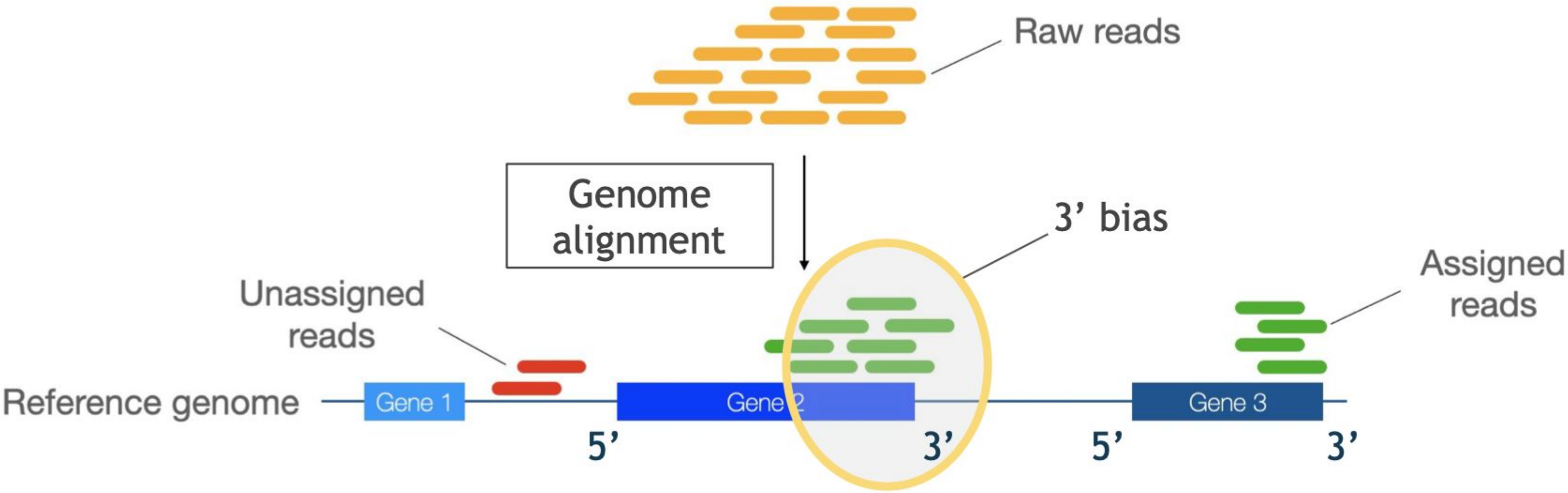
Primary analyses : Mapping



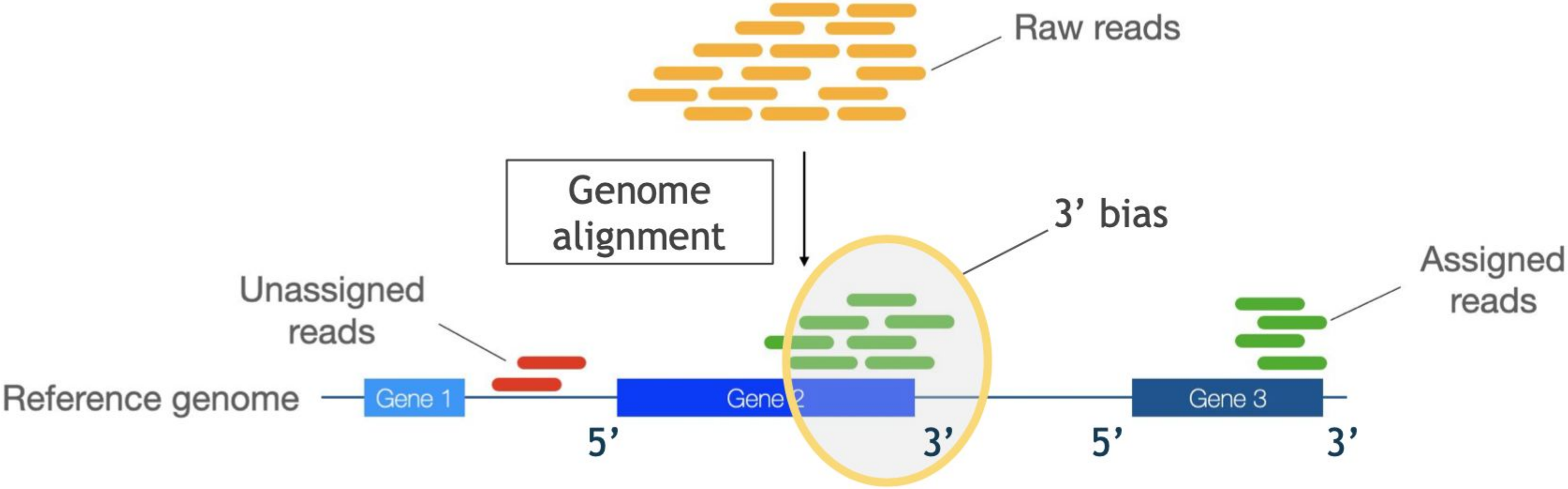


- Then, the **genome annotation** is used to assign the reads to genes
- Annotation is provided by genome portals (NCBI, Ensembl, UCSC) or consortiums of researchers working on a same organism
- genome annotation is generally provided as a file in the format **GFF** or **GTF**

How is the genome reference annotation used ?



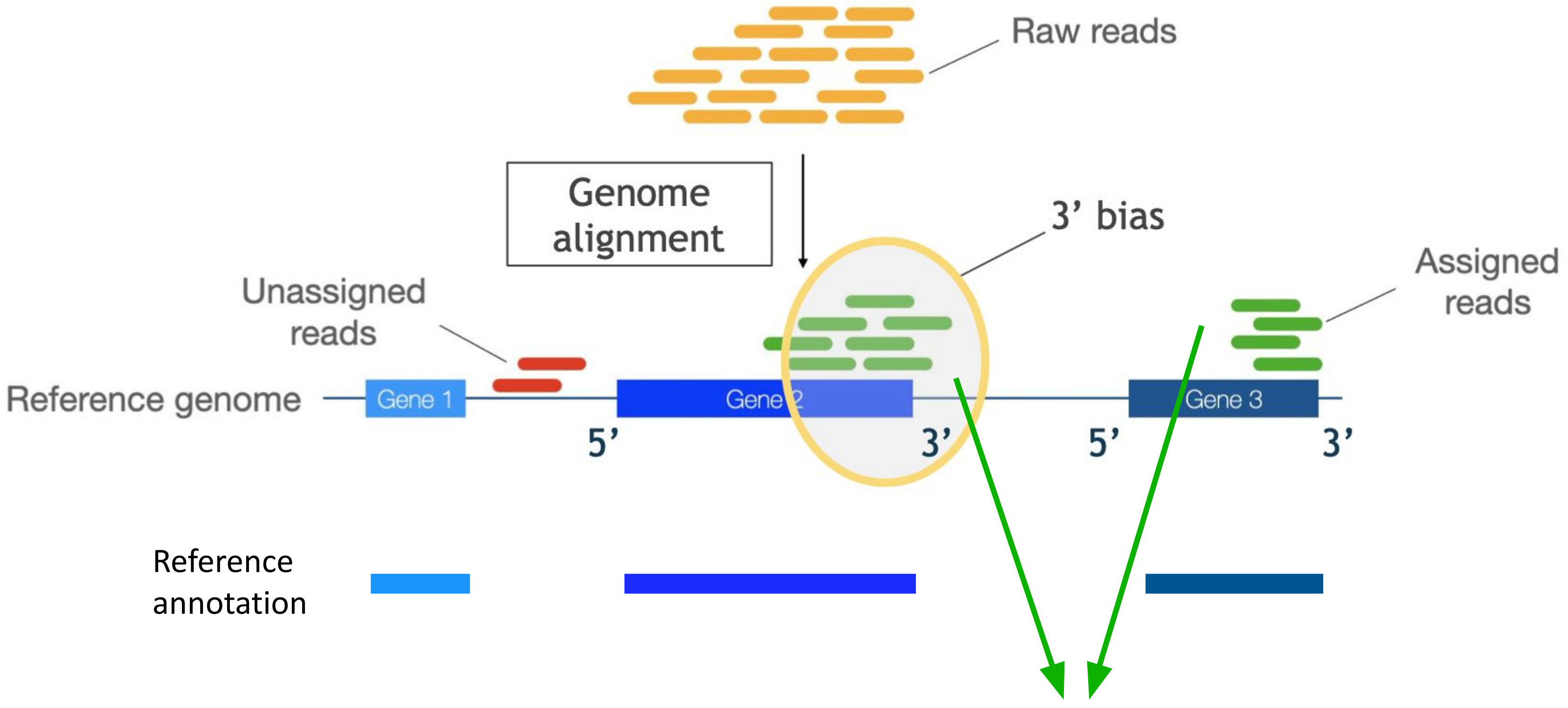
How is the genome reference annotation used ?



Reference annotation

(NCBI, UCSC, Ensembl…)

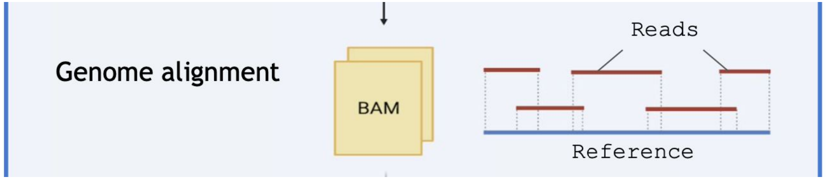
How is the genome reference annotation used ?

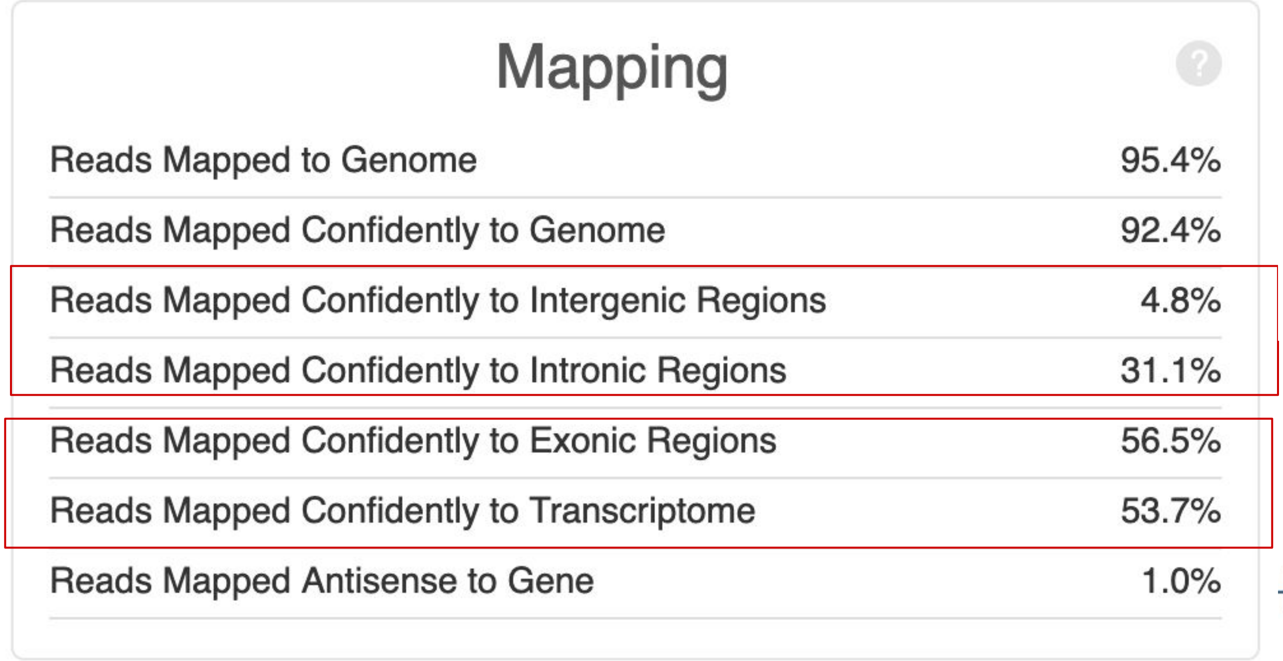


Only assigned reads are taken into account in the Count matrix

Example dataset from 10X Genomics

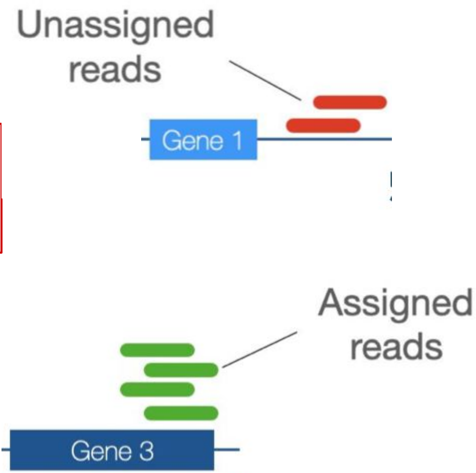
In CellRanger report





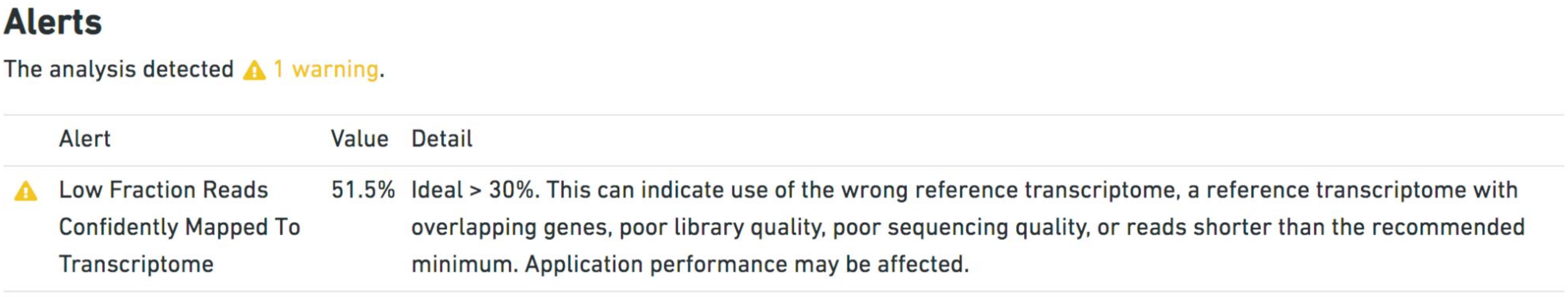
Mapping

Reads Mapped to Genome	95.4%
Reads Mapped Confidently to Genome	92.4%
Reads Mapped Confidently to Intergenic Regions	4.8%
Reads Mapped Confidently to Intronic Regions	31.1%
Reads Mapped Confidently to Exonic Regions	56.5%
Reads Mapped Confidently to Transcriptome	53.7%
Reads Mapped Antisense to Gene	1.0%



Important point on annotation

- Annotation is a crucial parameter (largely *underestimated*), as reads outside the annotated exons will not be taken into account !
- CellRanger will warn you on the report with the Alert below. In such cases, you need to visualise your signal in a genome browser (more on this tomorrow) and suspect the annotation may be problematic

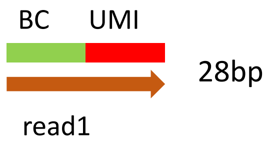


Alerts

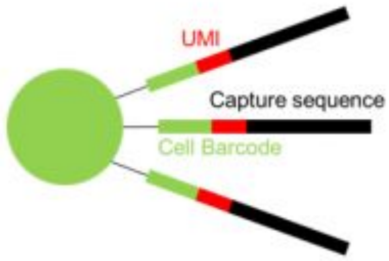
The analysis detected ⚠ 1 warning.

	Alert	Value	Detail
⚠	Low Fraction Reads Confidently Mapped To Transcriptome	51.5%	Ideal > 30%. This can indicate use of the wrong reference transcriptome, a reference transcriptome with overlapping genes, poor library quality, poor sequencing quality, or reads shorter than the recommended minimum. Application performance may be affected.

Primary analyses : barcode and UMI



- Read1 is made of BC + UMI
- Barcode and UMI are treated separately

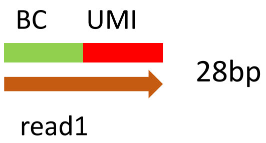


Reminder : barcode enables to trace back the read to the cell of origin ; UMI enables to distinguish each individual molecule

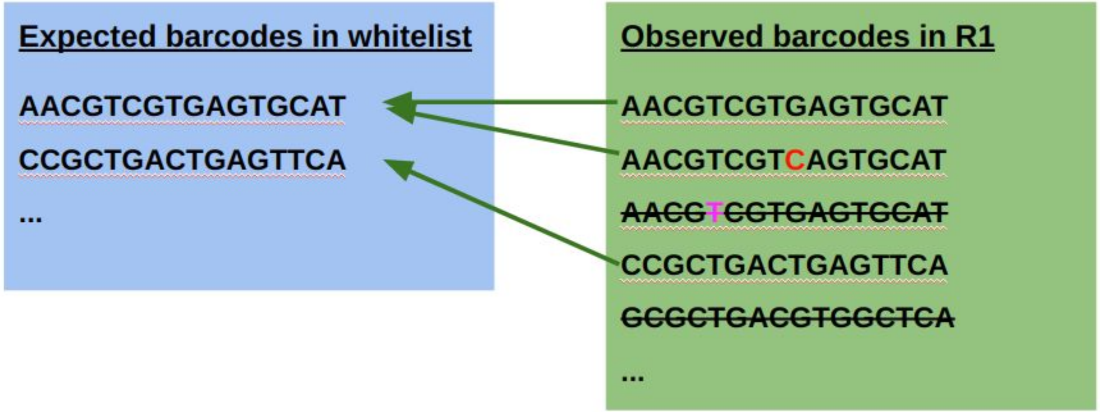
Cell Barcode (16bp) = sequence specific to each bead (so each cell)

UMI (12 bp) = sequence specific to each molecule : Unique Molecular Identifier

Primary analyses : barcode



- Barcode is extracted (16bp)
- 10X provides a **whitelist** containing all possible barcodes used on the gel beads (~3 million barcodes for the v3 chemistry)
- All barcodes are compared to this whitelist
- **Correction**: barcodes with 1 difference (1 mismatch) from the whitelist are corrected.
- **Filtering**: keep only BC in the whitelist.



Expected barcodes in whitelist

AACGTCGTGAGTGCAT

CCGCTGACTGAGTTCA

...

Observed barcodes in R1

AACGTCGTGAGTGCAT

AACGTCGTCAGTGCAT

~~AACGTCGTGAGTGCAT~~

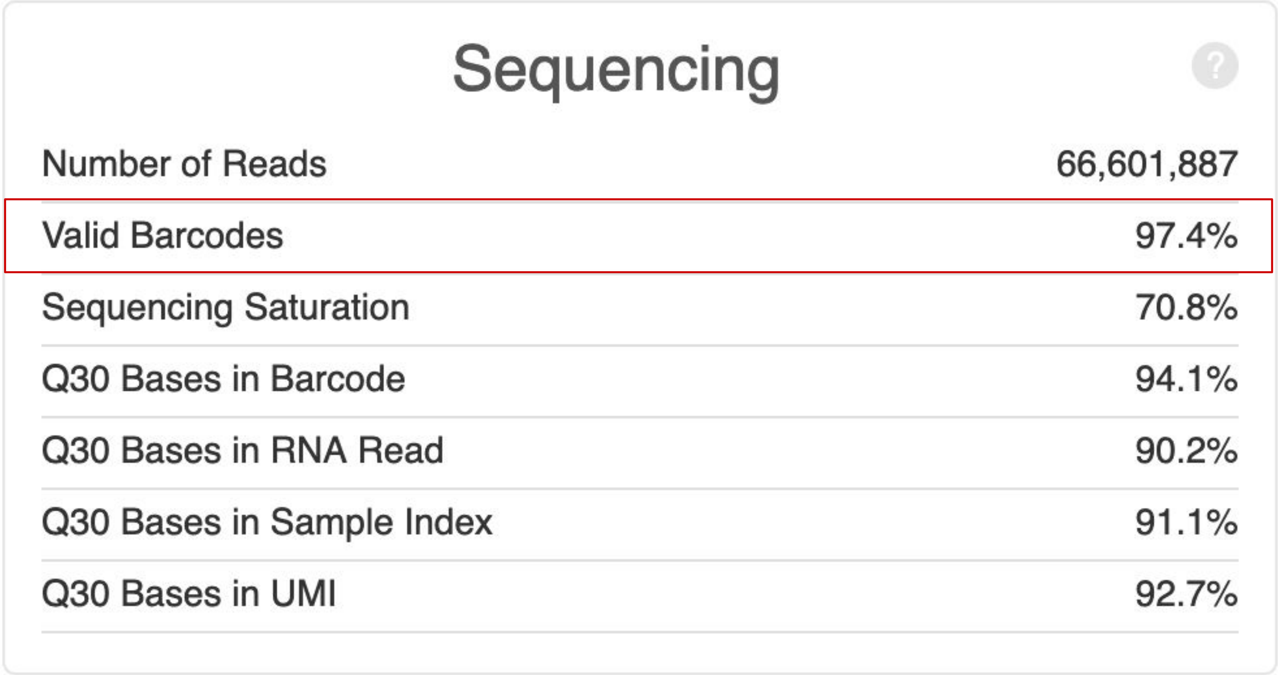
CCGCTGACTGAGTTCA

~~GCGCTGACGTGGCTCA~~

...

Example dataset from 10X Genomics

In CellRanger report

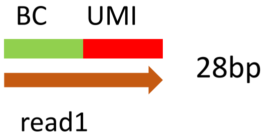


Sequencing

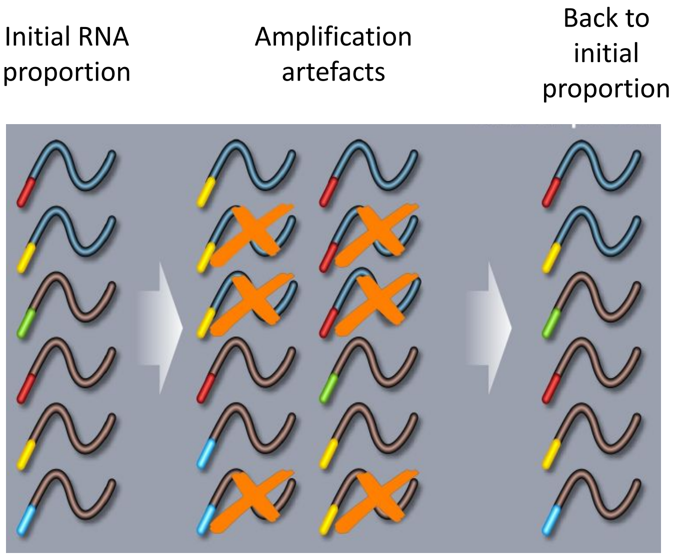
Number of Reads	66,601,887
Valid Barcodes	97.4%
Sequencing Saturation	70.8%
Q30 Bases in Barcode	94.1%
Q30 Bases in RNA Read	90.2%
Q30 Bases in Sample Index	91.1%
Q30 Bases in UMI	92.7%

% of valid barcodes is indicated in the report

Primary analyses : UMI

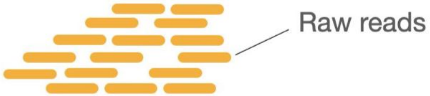


- UMI is extracted (12bp)
- UMI are randomised sequences, there is no whitelist
- **Correction**: UMI with 1 difference (1 mismatch) from a higher-count UMI are corrected to the higher count UMI if they share a cell barcode.
- **Filtering**: remove incorrect UMIs:
  - homopolymers (e.g. AAAAAAAAAA)
  - Contains 1 or several N
  - contains any base with BASEQ < 10

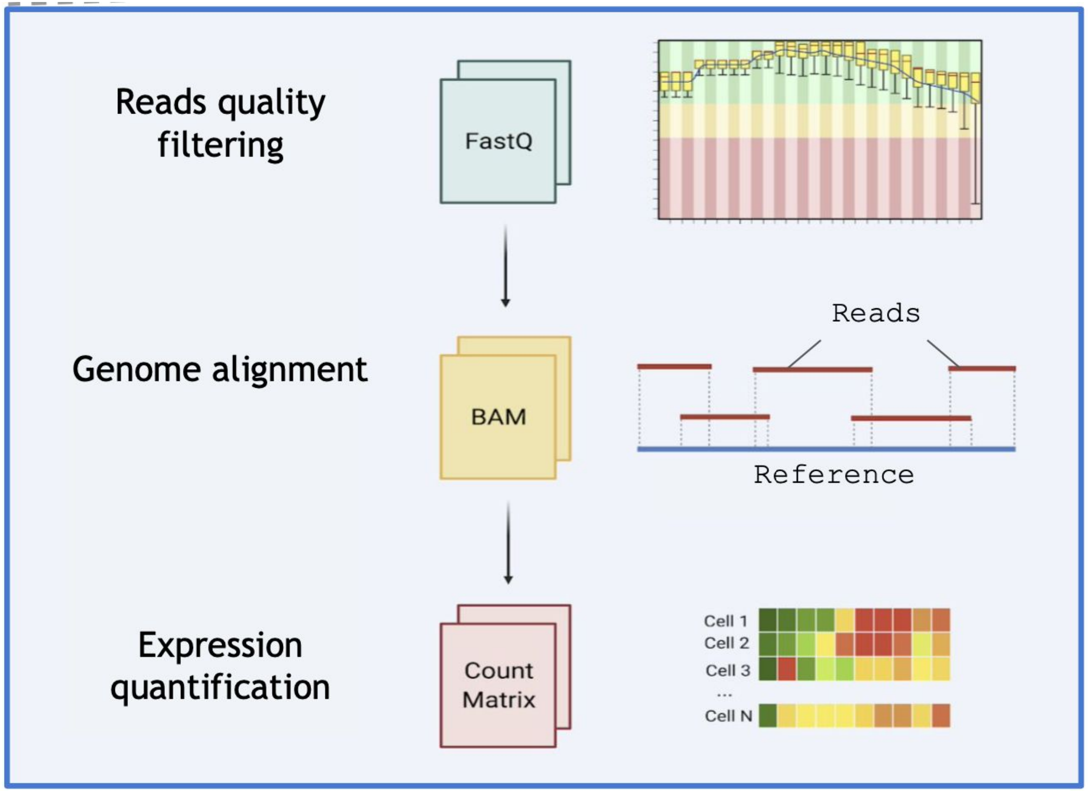


UMI aims at correcting amplification artefacts (more details tomorrow)

Last step : generation of the count matrix



Primary analyses



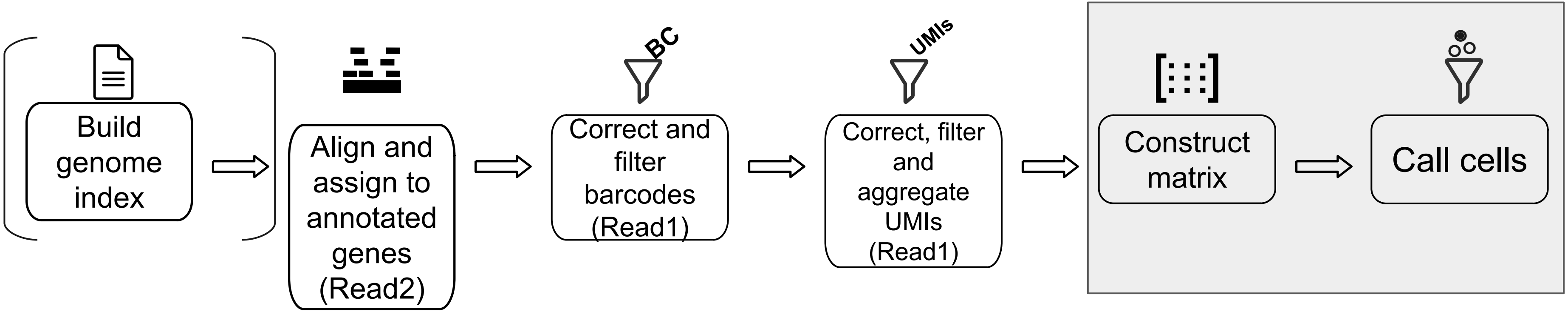
Reads quality filtering

Genome alignment

Expression quantification

will be presented in slides part 2

Overview of the workflow for primary analysis



Build genome index
Align and assign to annotated genes (Read2)
BC
Correct and filter barcodes (Read1)
UMIs
Correct, filter and aggregate UMIs (Read1)
Construct matrix
Call cells

Take-home messages

- **Primary analysis is important !** If this step has issues, the resulting count matrix will have issues that will be propagated to all downstream analyses
- These steps are often overlooked
- **Cell Ranger** : program provided by 10X Genomics that perform primary analysis (and a bit more). Cell Ranger is reliable but it is necessary to understand what it does and its limits
- You will hear that “the raw data is the count matrix” => this is wrong, remember **the raw data are the reads**
- Only **read2 is mapped** to the genome ; read1 is synthetic Barcode+UMI
- Alternative ways to perform primary analysis exist

Acknowledgements

- Some illustrations were created by Nathalie Lehmann and Rémi Montagne