Lecture 2

Intro to Data processing:

From bcl to count matrix

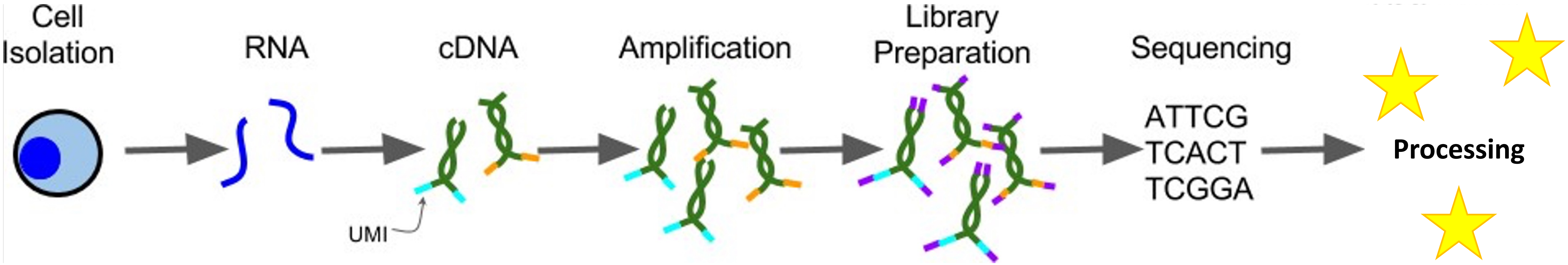
Physalia course 2025

—

Single-cell RNA-seq with R/Bioconductor

Instructors: Orr Ashenberg, Jacques Serizay, Fabrício Almeida-Silva

Experimental pipeline



Regardless of who is going to do the experimental steps, **DISCUSS WITH THEM !!!**

Experimental design is crucial for the success of a single-cell project !!!

Systematic comparison of single-cell and single-nucleus RNA-sequencing methods, Ding et al., Nat. Biotech. 2020

Flowchart: from .bcl to count matrix

- **Get bcl file** — **bcl files** – result of sequencing
- Create fastq files — **fastq files** – R1, R2 and I files showing raw reads
- QC: assess overall quality
- QC: remove low quality reads
- Align fastq to BAM — **bam file** – reads aligned to a reference genome
- QC: collapse UMI
- View alignment
- Count matrix — **count matrix** – cell by gene, showing expression of gene in cell

Flowchart: from .bcl to count matrix

- Get bcl file
- Create fastq files
- QC: assess overall quality
- QC: remove low quality reads
- Align fastq to BAM
- QC: collapse UMI
- View alignment
- Count matrix

.bcl files

.bcl:

- **Raw data** output of a sequencing run
- Binary, non-human-readable file

- Contains the **base calling** and **quality score** **per cluster per sequencing lane**

Flowchart: from .bcl to count matrix

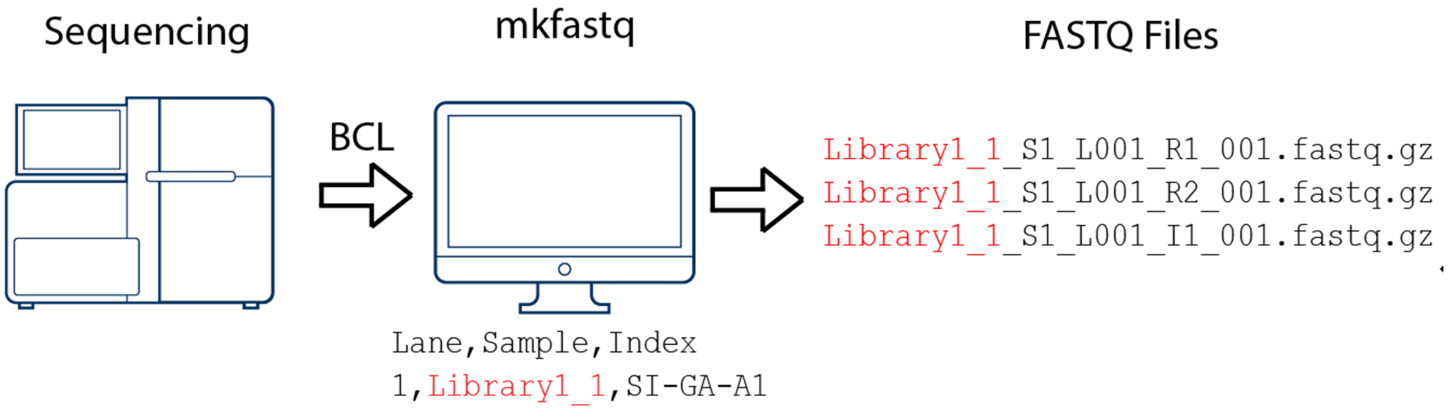
- ✓ Get bcl file
- ● Create fastq files
- ○ QC: assess overall quality
- ○ QC: remove low quality reads
- ○ Align fastq to BAM
- ○ QC: collapse UMI
- ○ View alignment
- ○ Count matrix

That is the role of the sequencing machine!

bcl2fastq

Command:

bcl2fastq --run-folder-dir <bcl_files_folder> -p 12 --output-dir <fastq_files_folder>



```
Library1_1_S1_L001_R1_001.fastq.gz
Library1_1_S1_L001_R2_001.fastq.gz
Library1_1_S1_L001_I1_001.fastq.gz
```

```
Lane,Sample,Index
1,Library1_1,SI-GA-A1
```

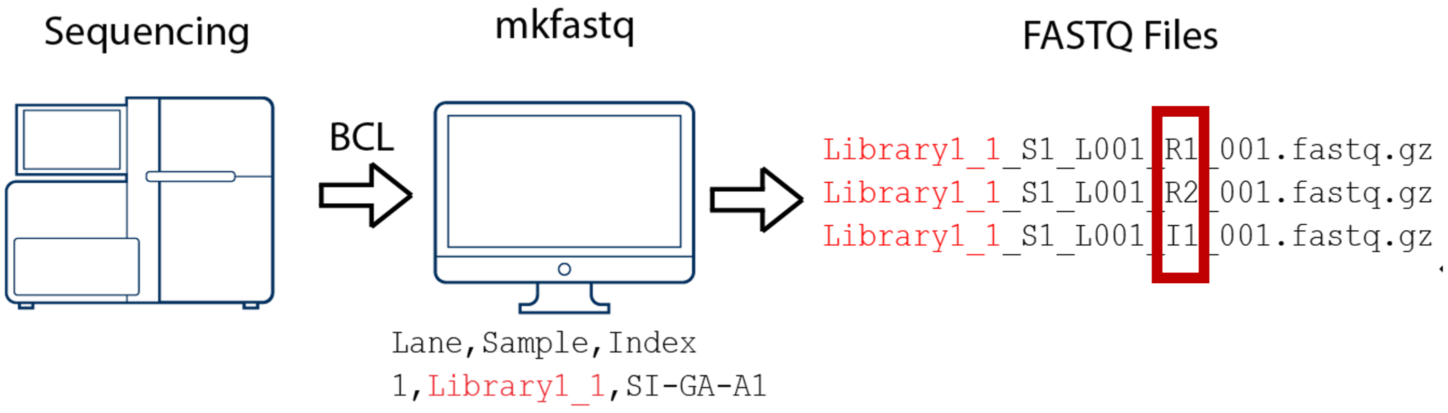
User guide:

https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/bcl2fastq/bcl2fastq_letterbooklet_15038058brpmi.pdf

bcl2fastq

Command:

```
bcl2fastq --run-folder-dir <bcl_files_folder> -p 12 --output-dir <fastq_files_folder>
```



```
Library1_1_S1_L001_R1_001.fastq.gz
Library1_1_S1_L001_R2_001.fastq.gz
Library1_1_S1_L001_I1_001.fastq.gz
```

```
Lane,Sample,Index
1,Library1_1,SI-GA-A1
```

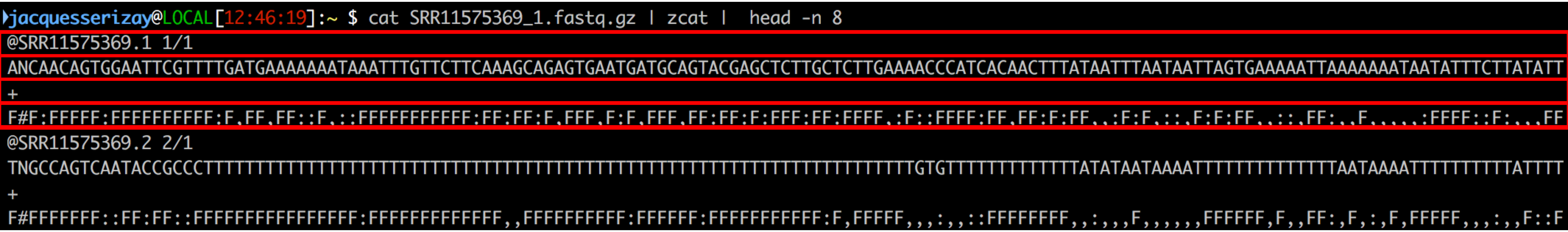
User guide:

https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/bcl2fastq/bcl2fastq_letterbooklet_15038058brpmi.pdf

How is a .fastq organized?

Each fastq file contains reads, each read is composed of 4 lines:

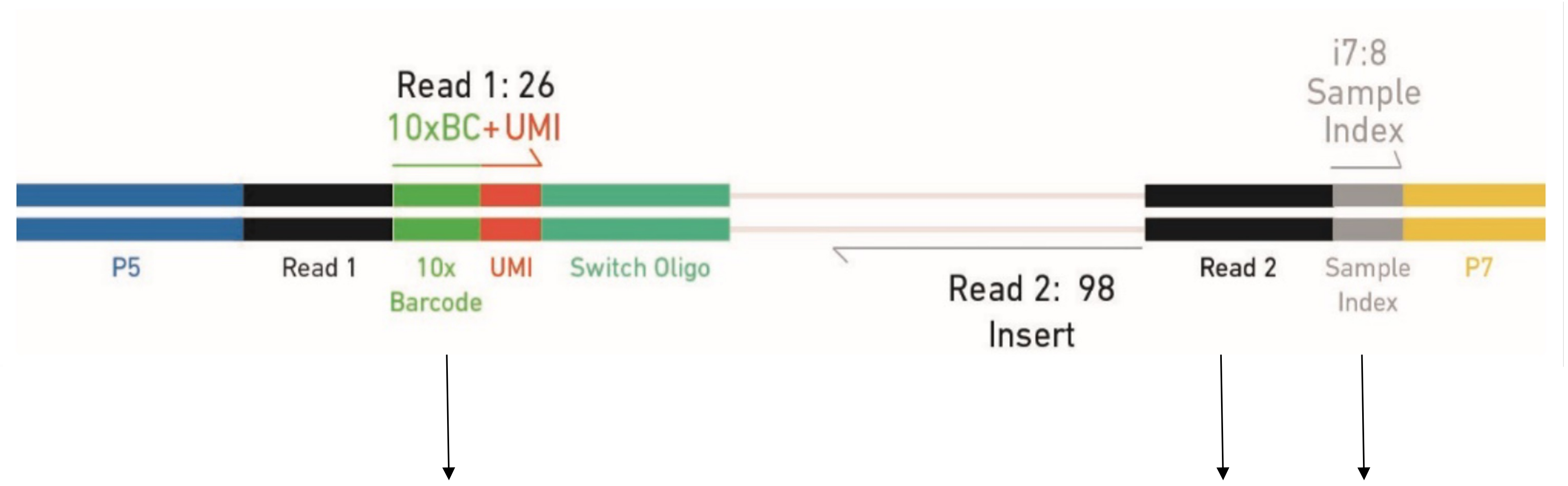
1. A sequence identifier with information about the sequencing run
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores, using ASCII characters to represent the numerical quality scores.



```
jacquesserizay@LOCAL[12:46:19]:~ $ cat SRR11575369_1.fastq.gz | zcat |  head -n 8
@SRR11575369.1 1/1
ANCAACAGTGGAATTCGTTTTGATGAAAAAAATAAATTTGTTCTTCAAAGCAGAGTGAATGATGCAGTACGAGCTCTTGCTCTTGAAAACCCATCACAACTTTATAATTTAATAATTAGTGAAAAATTAAAAAAATAATATTTCTTATATT
+
F#F:FFFFF:FFFFFFFFFF:F,FF,FF::F,::FFFFFFFFFFF:FF:FF:F,FFF,F:F,FFF,FF:FF:F:FFF:FF:FFFF,:F::FFFF:FF,FF:F:FF,,:F:F,::,F:F:FF,,::,FF:,,F,,,,,:FFFF::F:,,,FF
@SRR11575369.2 2/1
TNGCCAGTCAATACCGCCCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGTGTTTTTTTTTTTTTATATAATAAAATTTTTTTTTTTTTAATAAAATTTTTTTTTTTATTTT
+
F#FFFFFFF::FF:FF::FFFFFFFFFFFFFFFF:FFFFFFFFFFFFF,,FFFFFFFFFF:FFFFFF:FFFFFFFFFFF:F,FFFFF,,,:,,::FFFFFFFF,,:,,,F,,,,,,FFFFFF,F,,FF:,F,:,F,FFFFF,,,:,,F::F
```

Why do we end up with so many fastq files?

We sequence the paired ends of each DNA fragment molecule, in 3 different sequencing “runs”.



R1.fastq

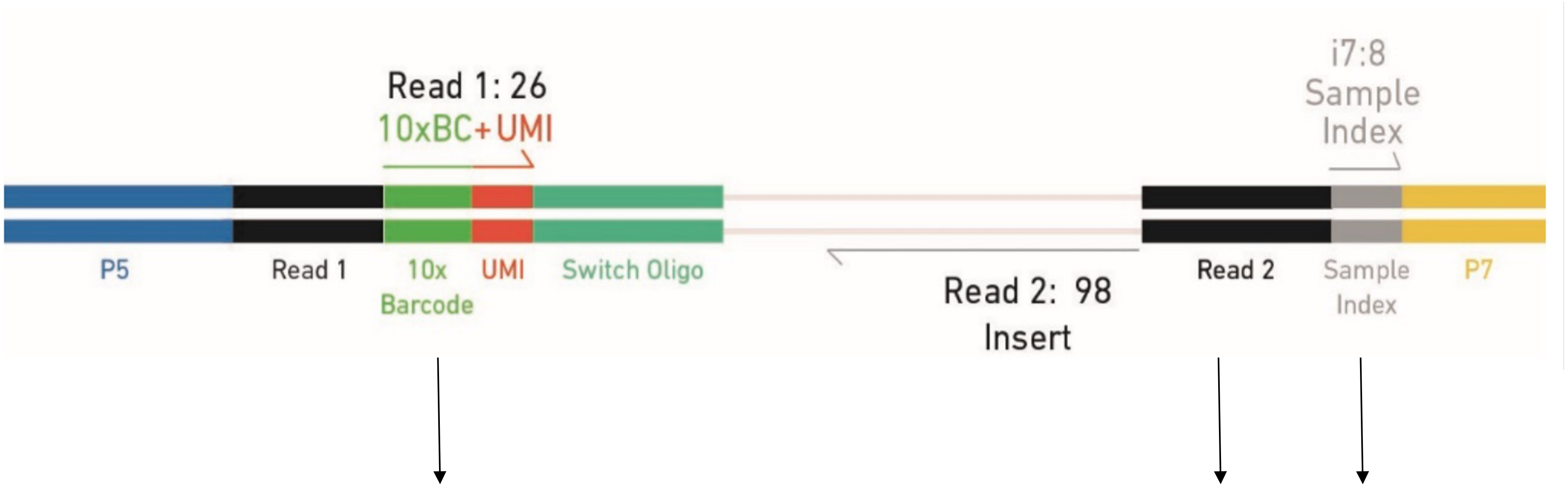
R2.fastq

I1.fastq

I1 is important for demultiplexing multiple samples simultaneously sequenced

Each fastq generated by **bcl2fastq** contains a different information:

- I1.fastq contains sample index
- R1.fastq contains cell barcode + UMI
- R2.fatsq contains transcript information



R1.fastq

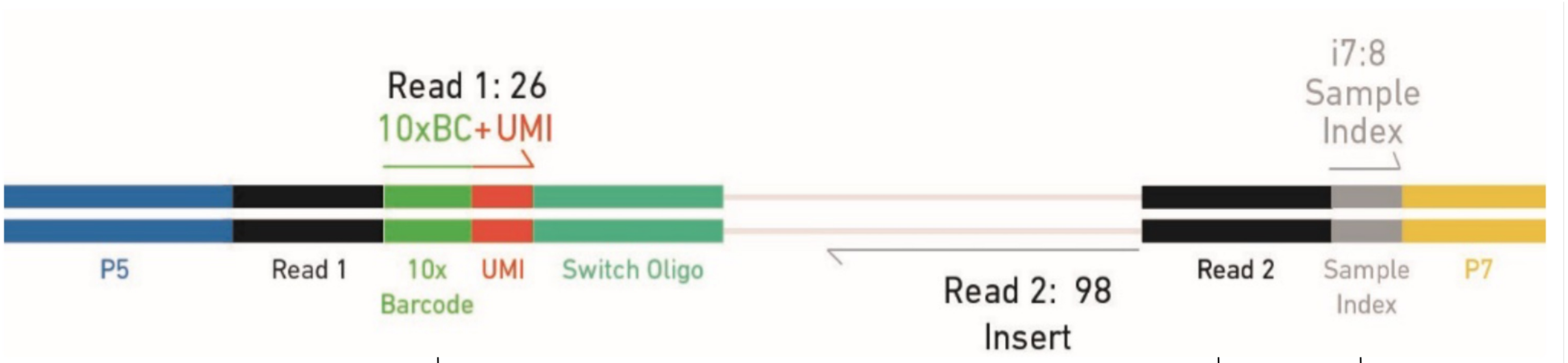
R2.fastq

I1.fastq

R1 contains information on the cell of origin (as well as a UMI)

Each fastq generated by **bcl2fastq** contains a different information:

- I1.fastq contains sample index
- R1.fastq contains cell barcode + UMI
- R2.fatsq contains transcript information



R1.fastq

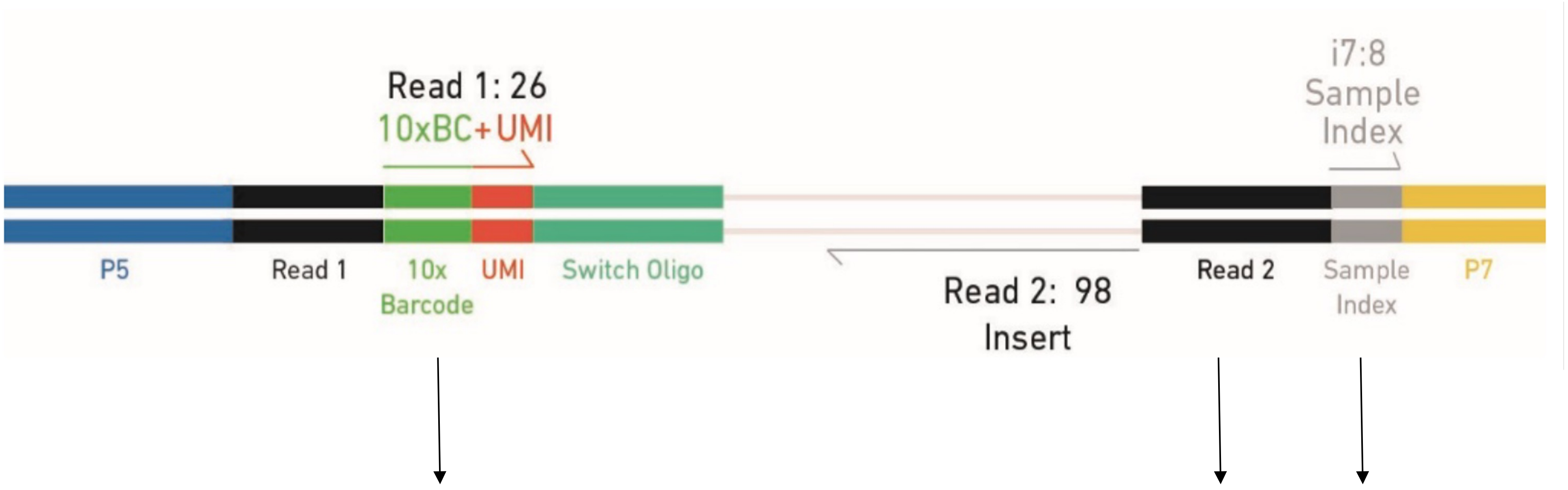
R2.fastq

I1.fastq

R2 contains sequence of the DNA fragment captured by sequencing

Each fastq generated by **bcl2fastq** contains a different information:

- I1.fastq contains sample index
- R1.fastq contains cell barcode + UMI
- R2.fatsq contains transcript information



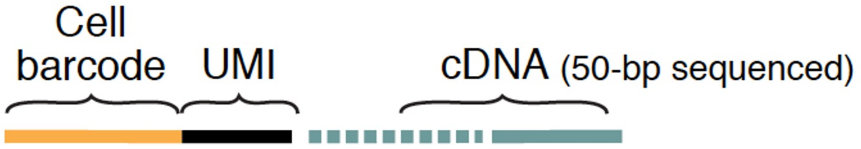
R1.fastq

R2.fastq I1.fastq

Reconstructing actual molecules loaded in the sequencer

After demultiplexing (thanks to the I1 file), here is what you have:

R1.fastq R2.fastq



```
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
GACCTACGAGTTAGTTTGTA.............GCTCATAA
GTTAAACGTACCCTAGCTGT.............GATTTTCT
ACGTCACCTTTTGTGGGGGT.............ATAAGCTC
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
AGTCCATGTGCGGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CCAAAGATGTCCTCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTTTTGACCAGTCGTGAGGG.............TTCCAAGG
ACTGTCCATGCCCCTGTGTA.............TGGTACGT
CGTAAAACAATAATCCGGTG.............TTAAACCG
...........
...........
...........
```

(Hundreds of millions of reads)

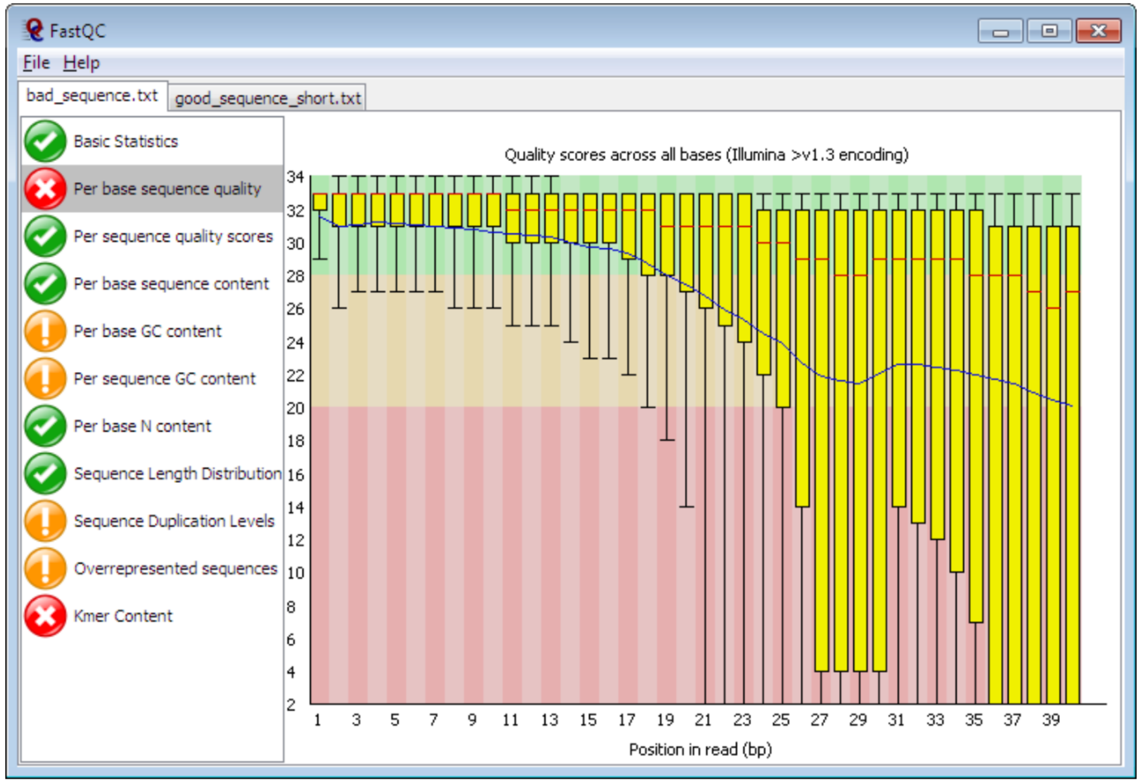
Flowchart: from .bcl to count matrix

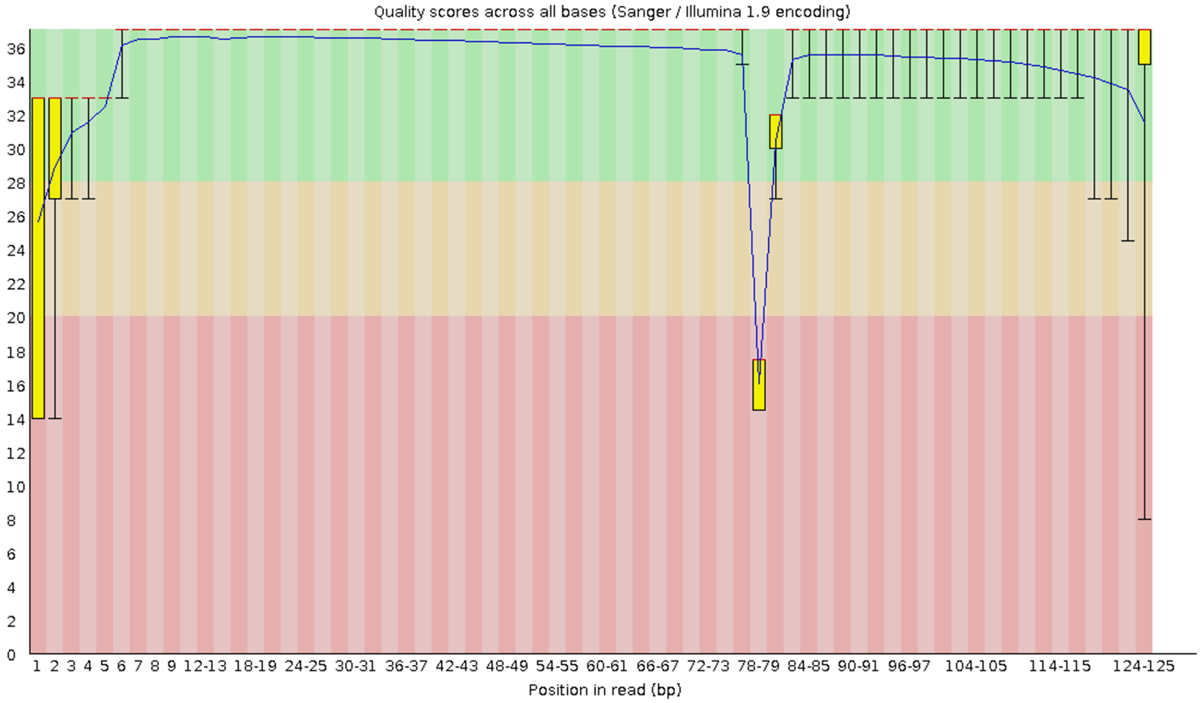
- ✓ Get bcl file — That is the role of the sequencing machine!
- ✓ Create fastq files — bcl2fastq
- ● QC: assess overall quality
- ● QC: remove low quality reads
- ● Align fastq to BAM
- ● QC: collapse UMI
- ● View alignment
- ● Count matrix

Raw sequencing QC is typically done with Fastqc

FastQC

FastQC is a program designed to spot potential problems in high througput sequencing datasets. It runs a set of analyses on one or more raw sequence files in fastq or bam format and produces a report which summarises the results.





FastQC will highlight any areas where this library looks unusual and where you should take a closer look. The program is not tied to any specific type of sequencing technique and can be used to look at libraries coming from a large number of different experiment types (Genomic Sequencing, ChIP-Seq, RNA-Seq, BS-Seq etc etc).

Flowchart: from .bcl to count matrix

	✓	Get bcl file	That is the role of the sequencing machine!
	✓	Create fastq files	bcl2fastq
	✓	QC: assess overall quality	fastqc
Cellranger	●	QC: remove low quality reads	
	○	Align fastq to BAM	
	○	QC: collapse UMI	
	○	View alignment	
	○	Count matrix	

Filtering low quality reads

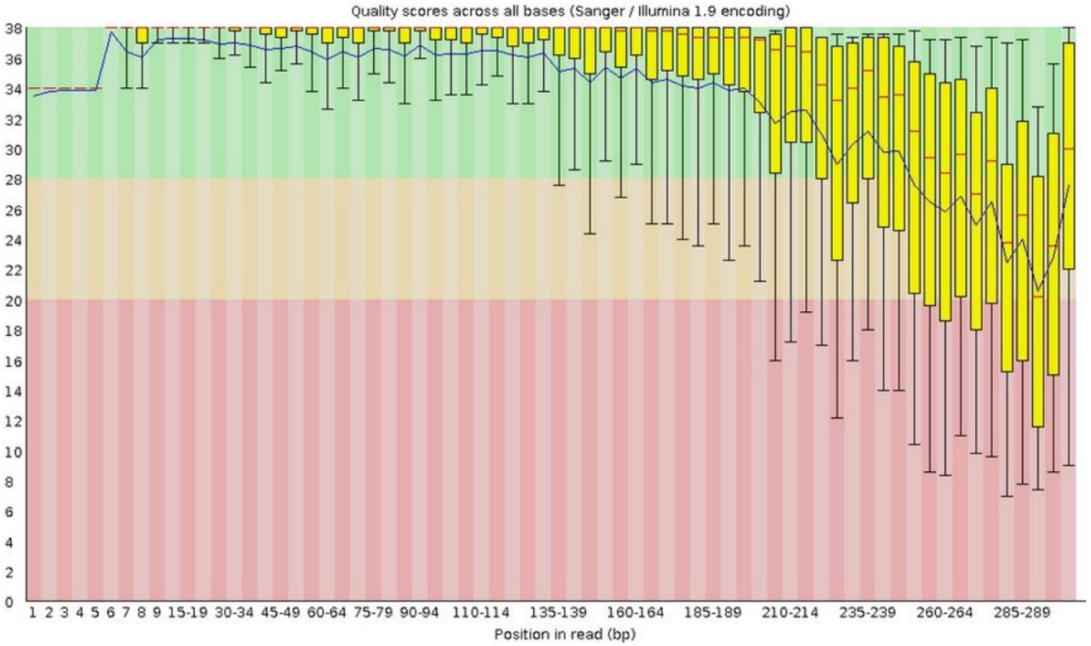
It is important to consider (and remove!) :

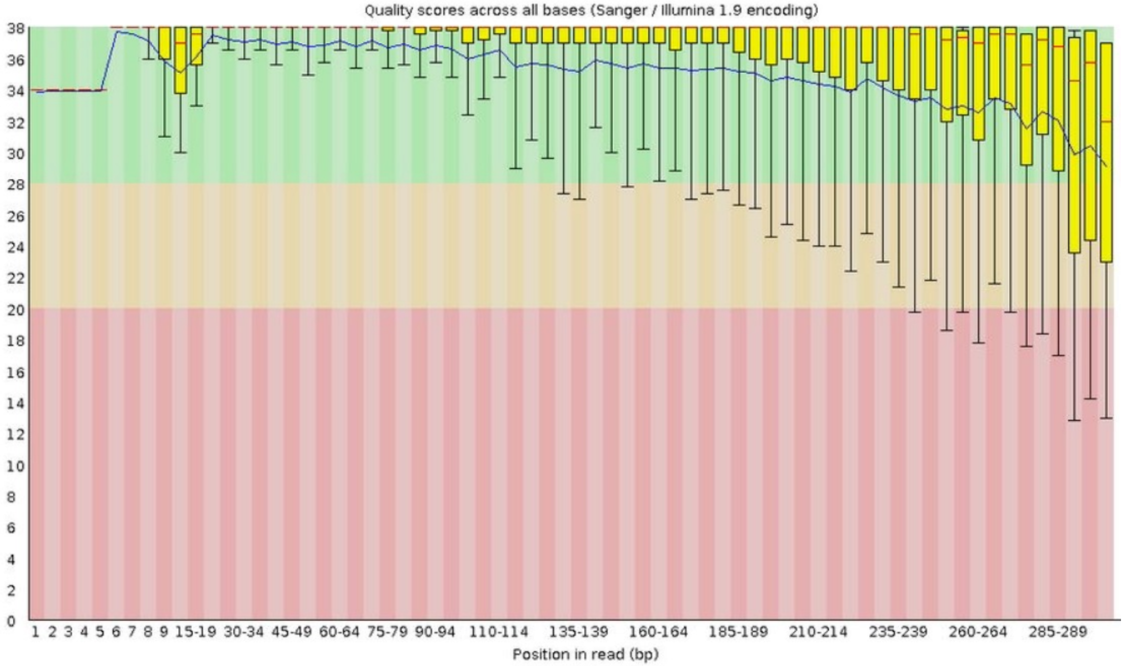
1. Reads with overall low quality
2. Unrecognized cell barcode

Filtering low quality reads

It is important to consider (and remove!) :

1. **Reads with overall low quality**
2. Unrecognized cell barcode

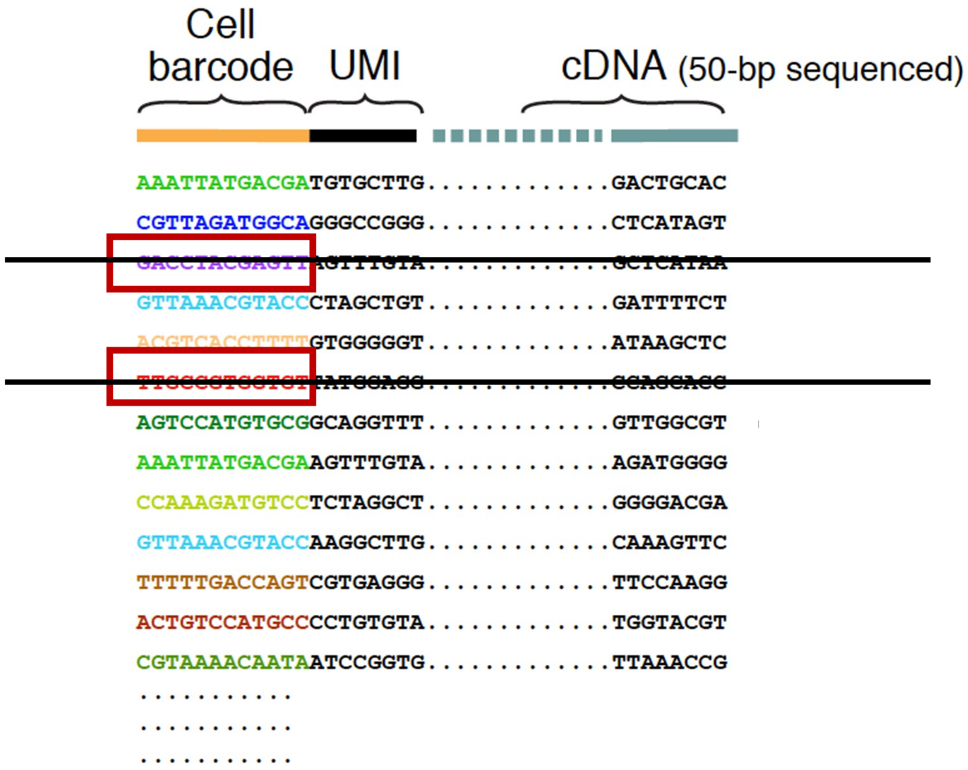




Filtering low quality reads

It is important to consider (and remove!) :

1. Reads with overall low quality
2. **Unrecognized cell barcode**


Cell barcode
UMI
cDNA (50-bp sequenced)
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
~~GACCTACGAGTTAGTTTGTA.............GCTCATAA~~
GTTAAACGTACCCTAGCTGT.............GATTTTCT
ACGTCACCTTTTGTGGGGGT.............ATAAGCTC
~~TTGCCGTGGTGTTATGGAGG.............CCAGCACC~~
AGTCCATGTGCGGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CCAAAGATGTCCTCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTTTTGACCAGTCGTGAGGG.............TTCCAAGG
ACTGTCCATGCCCCTGTGTA.............TGGTACGT
CGTAAAACAATAATCCGGTG.............TTAAACCG
....
....
....

(Hundreds of millions of reads)

What is a barcode whitelist?
List of all known barcode sequences that have been included in the assay kit and are available during library preparation.

For example, there are roughly 737,000 cell barcodes in the whitelist for Cell Ranger's Single Cell 3' applications. Here are the first 10 lines of the corresponding barcode whitelist 737K-august-2016.txt:

```
AAACCTGAGAAACCAT
AAACCTGAGAAACCGC
AAACCTGAGAAACCTA
AAACCTGAGAAACGAG
AAACCTGAGAAACGCC
AAACCTGAGAAAGTGG
AAACCTGAGAACAACT
AAACCTGAGAACAATC
AAACCTGAGAACTCGG
AAACCTGAGAACTGTA
```

How to correct barcode sequencing errors?

For every observed barcode in the dataset not on the whitelist, but 1-Hamming-distance away (i.e. 1 mismatch) from a whitelist barcode:

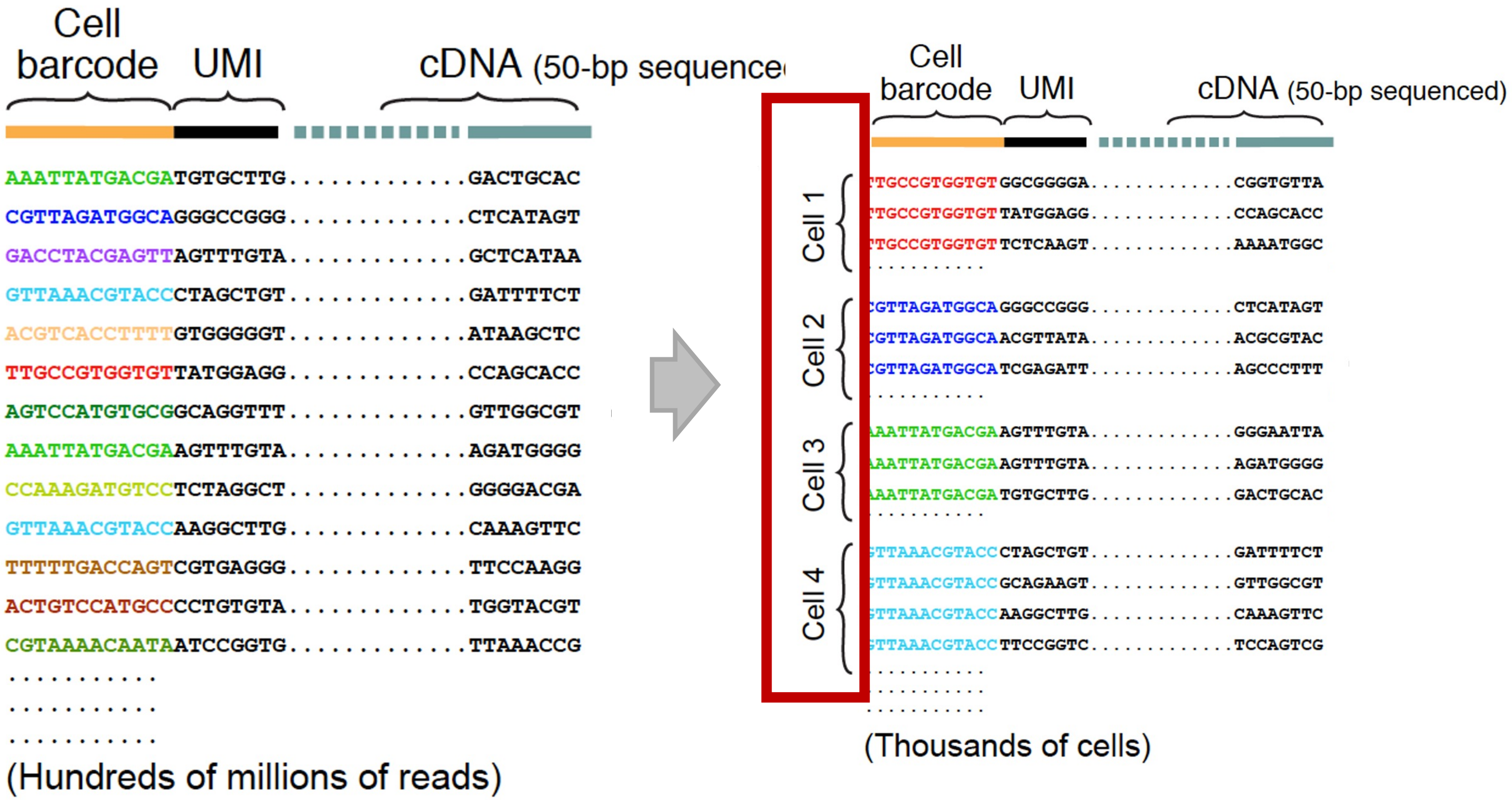
- Compute the posterior probability that the observed barcode originated from the whitelist barcode with a sequencing error at the differing base (based on the base Q score).
- Replace the observed barcode with the whitelist barcode with the highest posterior probability that exceeds 0.975.

Flowchart: from .bcl to count matrix

	✓	Get bcl file	That is the role of the sequencing machine!
	✓	Create fastq files	bcl2fastq
	✓	QC: assess overall quality	fastqc
Cellranger	✓	QC: remove low quality reads	Automatic in most aligners
	●	Align fastq to BAM	
	○	QC: collapse UMI	
	○	View alignment	
	○	Count matrix	

Aligning reads to a transcriptome reference

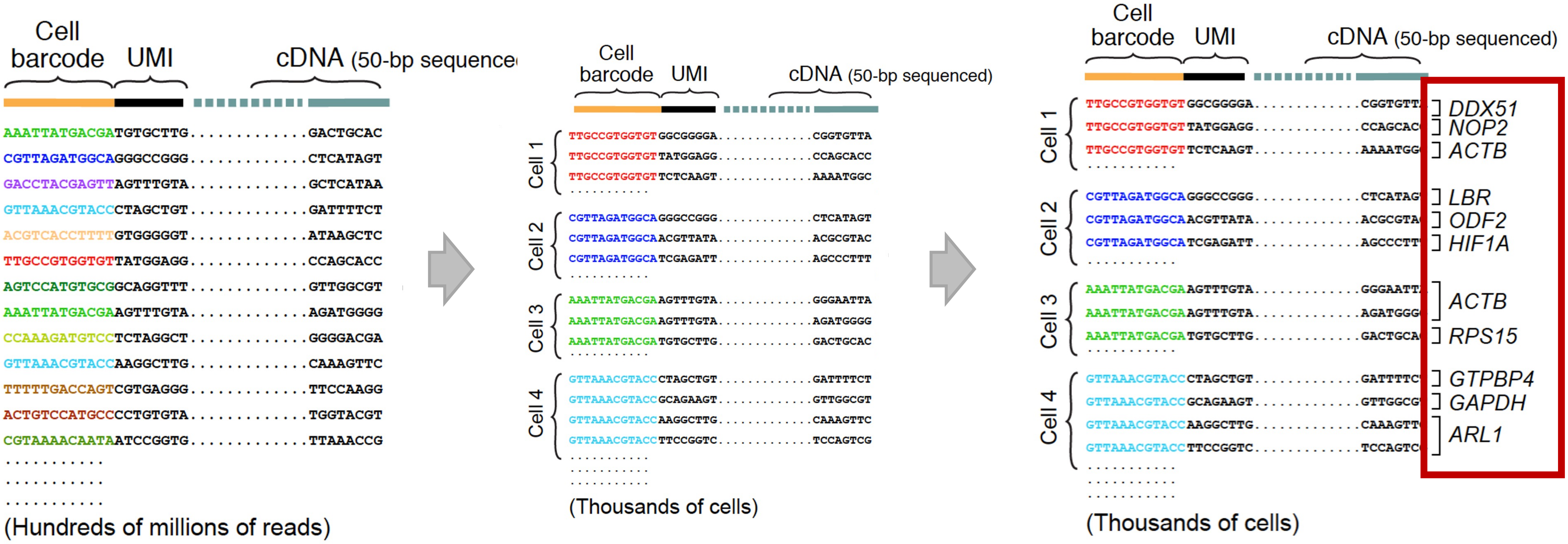
1. Group reads by cell-of-origin (using the cell barcodes)



Cell barcode
UMI
cDNA (50-bp sequence
AAATTATGACGATGTGCTTG............GACTGCAC
CGTTAGATGGCAGGGCCGGG............CTCATAGT
GACCTACGAGTTAGTTTGTA............GCTCATAA
GTTAAACGTACCCTAGCTGT............GATTTTCT
ACGTCACCTTTTGTGGGGGT............ATAAGCTC
TTGCCGTGGTGTTATGGAGG............CCAGCACC
AGTCCATGTGCGGCAGGTTT............GTTGGCGT
AAATTATGACGAAGTTTGTA............AGATGGGG
CCAAAGATGTCCTCTAGGCT............GGGGACGA
GTTAAACGTACCAAGGCTTG............CAAAGTTC
TTTTTGACCAGTCGTGAGGG............TTCCAAGG
ACTGTCCATGCCCCTGTGTA............TGGTACGT
CGTAAAACAATAATCCGGTG............TTAAACCG
....
....
....
(Hundreds of millions of reads)
Cell barcode
UMI
cDNA (50-bp sequenced)
Cell 1
TTGCCGTGGTGT GGCGGGGA............CGGTGTTA
TTGCCGTGGTGT TATGGAGG............CCAGCACC
TTGCCGTGGTGT TCTCAAGT............AAAATGGC
....
Cell 2
CGTTAGATGGCA GGGCCGGG............CTCATAGT
CGTTAGATGGCA ACGTTATA............ACGCGTAC
CGTTAGATGGCA TCGAGATT............AGCCCTTT
....
Cell 3
AAATTATGACGA AGTTTGTA............GGGAATTA
AAATTATGACGA AGTTTGTA............AGATGGGG
AAATTATGACGA TGTGCTTG............GACTGCAC
....
Cell 4
GTTAAACGTACC CTAGCTGT............GATTTTCT
GTTAAACGTACC GCAGAAGT............GTTGGCGT
GTTAAACGTACC AAGGCTTG............CAAAGTTC
GTTAAACGTACC TTCCGGTC............TCCAGTCG
....
....
....
(Thousands of cells)

Aligning reads to a transcriptome reference

1. Group reads by cell-of-origin (using the cell barcodes)
2. **Recover which transcript the cDNA sequence aligns to**



Cell barcode
UMI
cDNA (50-bp sequence
AAATTATGACGATGTGCTTG............GACTGCAC
CGTTAGATGGCAGGGCCGGG............CTCATAGT
GACCTACGAGTTAGTTTGTA............GCTCATAA
GTTAAACGTACCCTAGCTGT............GATTTTCT
ACGTCACCTTTTGTGGGGGT............ATAAGCTC
TTGCCGTGGTGTTATGGAGG............CCAGCACC
AGTCCATGTGCGGCAGGTTT............GTTGGCGT
AAATTATGACGAAGTTTGTA............AGATGGGG
CCAAAGATGTCCTCTAGGCT............GGGGACGA
GTTAAACGTACCAAGGCTTG............CAAAGTTC
TTTTTGACCAGTCGTGAGGG............TTCCAAGG
ACTGTCCATGCCCCTGTGTA............TGGTACGT
CGTAAAACAATAATCCGGTG............TTAAACCG
(Hundreds of millions of reads)
Cell barcode
UMI
cDNA (50-bp sequenced)
Cell 1
TTGCCGTGGTGTGGCGGGGA............CGGTGTTA
TTGCCGTGGTGTTATGGAGG............CCAGCACC
TTGCCGTGGTGTTCTCAAGT............AAAATGGC
Cell 2
CGTTAGATGGCAGGGCCGGG............CTCATAGT
CGTTAGATGGCAACGTTATA............ACGCGTAC
CGTTAGATGGCATCGAGATT............AGCCCTTT
Cell 3
AAATTATGACGAAGTTTGTA............GGGAATTA
AAATTATGACGAAGTTTGTA............AGATGGGG
AAATTATGACGATGTGCTTG............GACTGCAC
Cell 4
GTTAAACGTACCCTAGCTGT............GATTTTCT
GTTAAACGTACCGCAGAAGT............GTTGGCGT
GTTAAACGTACCAAGGCTTG............CAAAGTTC
GTTAAACGTACCTTCCGGTC............TCCAGTCG
(Thousands of cells)
Cell barcode
UMI
cDNA (50-bp sequenced)
Cell 1
TTGCCGTGGTGTGGCGGGGA............CGGTGTT
TTGCCGTGGTGTTATGGAGG............CCAGCAC
TTGCCGTGGTGTTCTCAAGT............AAAATGG
DDX51
NOP2
ACTB
Cell 2
CGTTAGATGGCAGGGCCGGG............CTCATAG
CGTTAGATGGCAACGTTATA............ACGCGTA
CGTTAGATGGCATCGAGATT............AGCCCTT
LBR
ODF2
HIF1A
Cell 3
AAATTATGACGAAGTTTGTA............GGGAATT
AAATTATGACGAAGTTTGTA............AGATGGG
AAATTATGACGATGTGCTTG............GACTGCA
ACTB
RPS15
Cell 4
GTTAAACGTACCCTAGCTGT............GATTTTC
GTTAAACGTACCGCAGAAGT............GTTGGCG
GTTAAACGTACCAAGGCTTG............CAAAGTT
GTTAAACGTACCTTCCGGTC............TCCAGTC
GTPBP4
GAPDH
ARL1
(Thousands of cells)

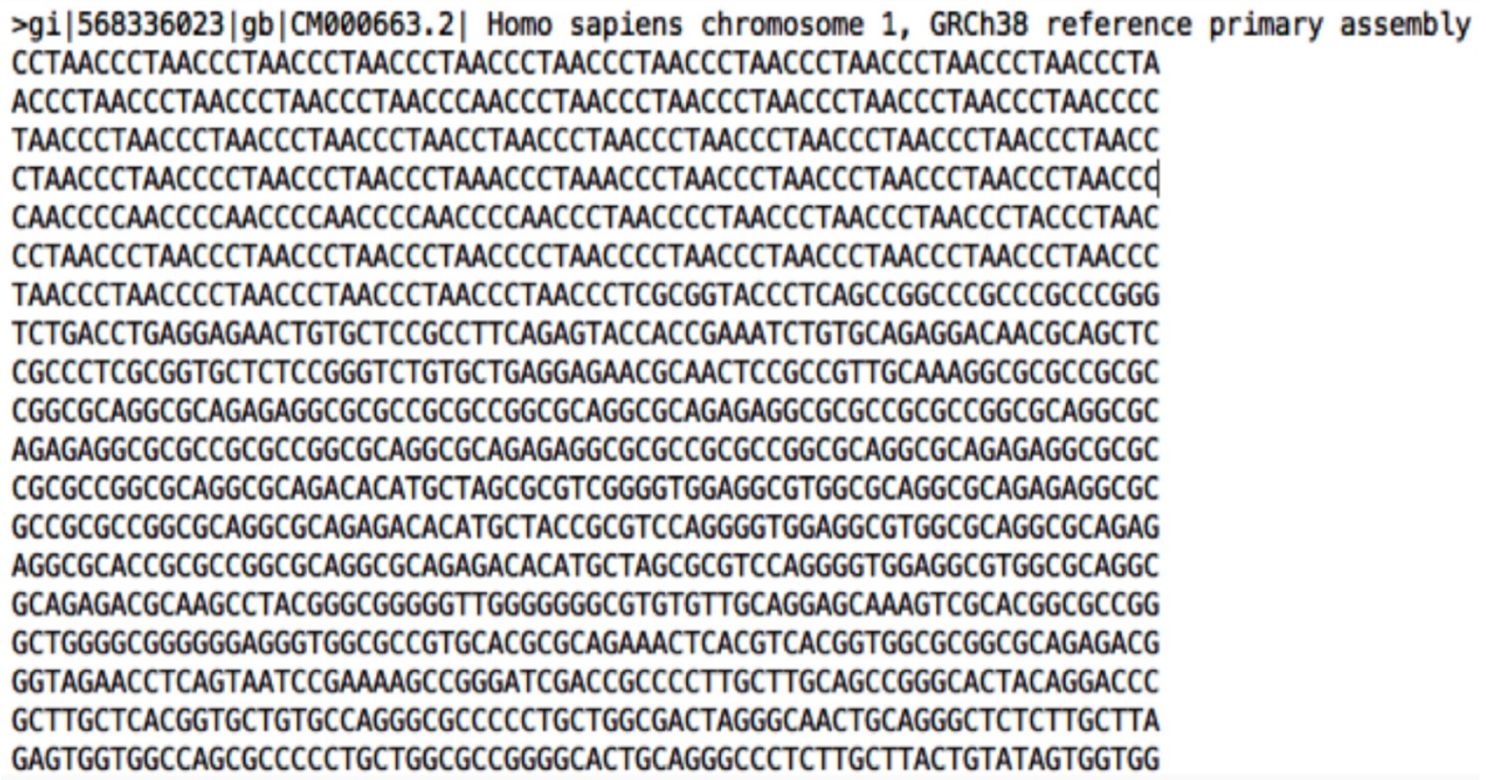
What do I need to align my reads?

Most aligners (included STAR-based cellranger) will map scRNAseq reads on a transcriptome index.

A transcriptome index consists of:

1. A <u>genome sequence </u>reference

FASTA example



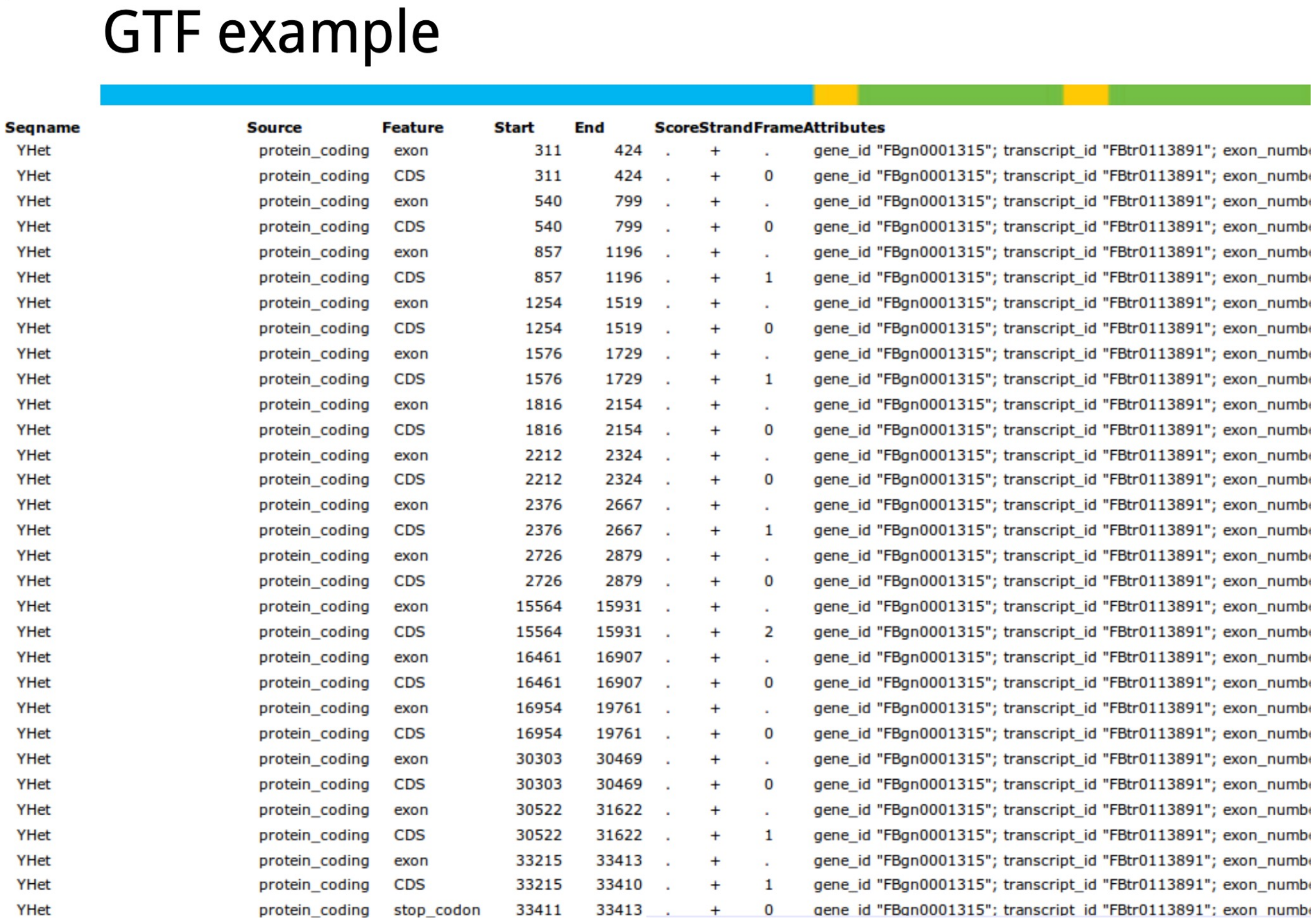
```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTC
CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGC
CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGC
GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGC
GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGG
GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGCAGAGACG
GGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACAGGACCC
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCTTGCTTA
GAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGG
```

What do I need to align my reads?

Most aligners (included STAR-based cellranger) will map scRNAseq reads on a transcriptome index.

A transcriptome index consists of:

1. A <u>genome sequence </u>reference
2. A <u>gene feature annotation</u> reference

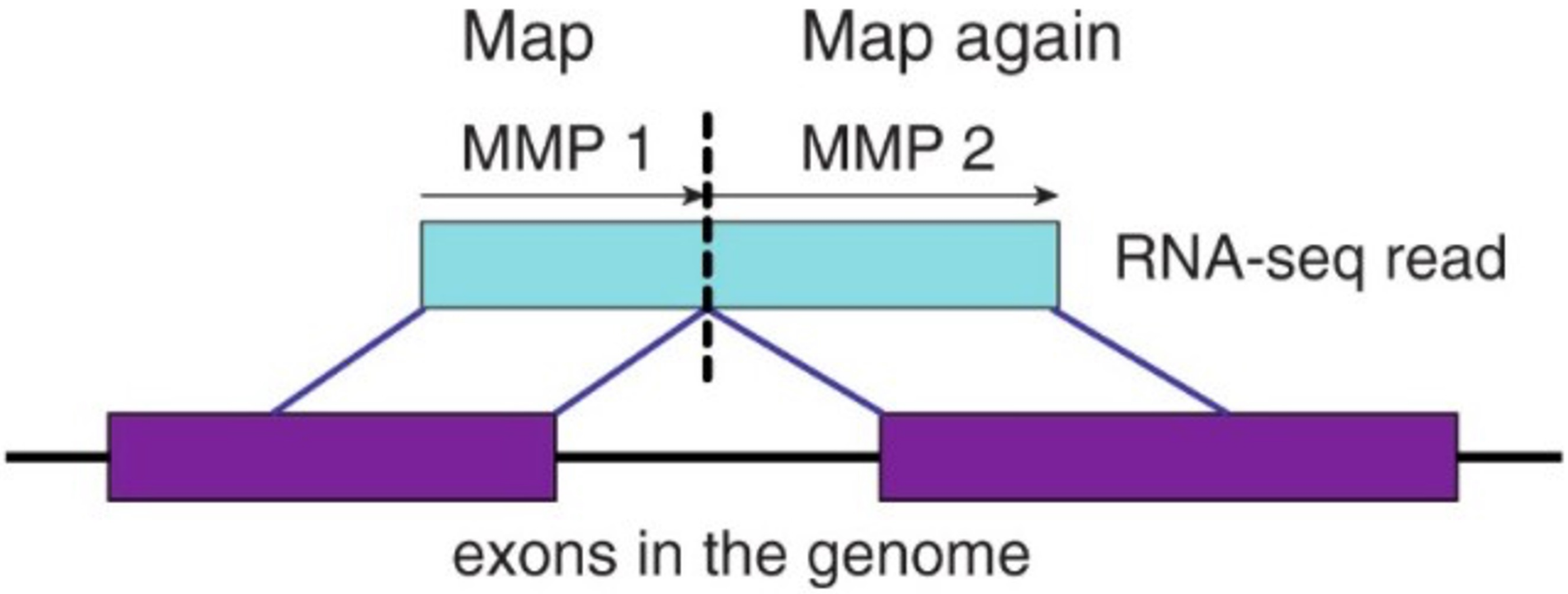


GTF example

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
YHet	protein_coding	exon	311	424	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	311	424	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	540	799	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	540	799	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	857	1196	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	857	1196	.	+	1	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	1254	1519	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	1254	1519	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	1576	1729	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	1576	1729	.	+	1	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	1816	2154	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	1816	2154	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	2212	2324	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	2212	2324	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	2376	2667	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	2376	2667	.	+	1	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	2726	2879	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	2726	2879	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	15564	15931	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	15564	15931	.	+	2	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	16461	16907	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	16461	16907	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	16954	19761	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	16954	19761	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	30303	30469	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	30303	30469	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	30522	31622	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	30522	31622	.	+	1	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	exon	33215	33413	.	+	.	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	CDS	33215	33410	.	+	1	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb
YHet	protein_coding	stop_codon	33411	33413	.	+	0	gene_id "FBgn0001315"; transcript_id "FBtr0113891"; exon_numb

STAR aligner

STAR is a traditional aligner that works by trying to find the longest possible sequence which matches one or more sequences in the reference genome.



STAR aligner

STAR is a traditional aligner that works by trying to find the longest possible sequence which matches one or more sequences in the reference genome.

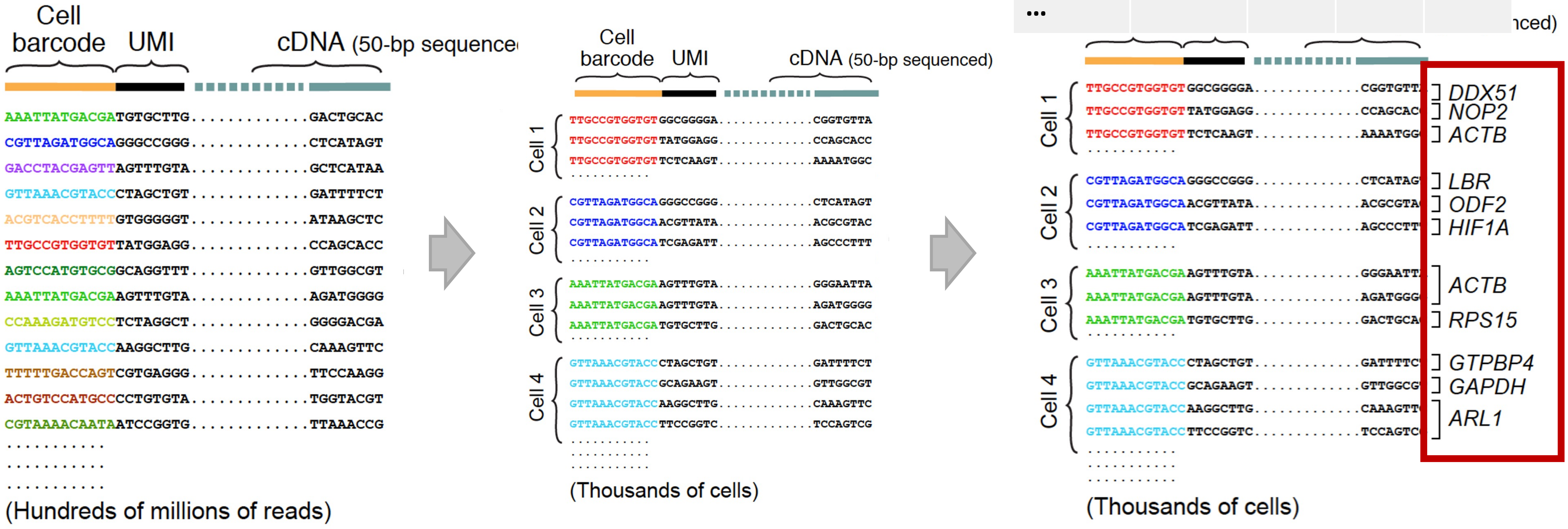
- Advantages: STAR is a splice aware aligner, making it suitable if you are interested in studying alternative splicing.
- Disadvantage: STAR requires a lot of RAM.

An alternative requiring less RAM is HISAT2.

Aligning reads to a transcriptome reference

1. Group reads by cell-of-origin (using the cell barcodes)
2. **Recover which transcript the cDNA sequence aligns to**

	Cell1	Cell2	Cell3	Cell4
ACTB	1	0	**2**	0
ARL1	0	0	0	**2**
…				



Cell barcode
UMI
cDNA (50-bp sequence
AAATTATGACGATGTGCTTG.............GACTGCAC
CGTTAGATGGCAGGGCCGGG.............CTCATAGT
GACCTACGAGTTAGTTTGTA.............GCTCATAA
GTTAAACGTACCCTAGCTGT.............GATTTTCT
ACGTCACCTTTTGTGGGGGT.............ATAAGCTC
TTGCCGTGGTGTTATGGAGG.............CCAGCACC
AGTCCATGTGCGGCAGGTTT.............GTTGGCGT
AAATTATGACGAAGTTTGTA.............AGATGGGG
CCAAAGATGTCCTCTAGGCT.............GGGGACGA
GTTAAACGTACCAAGGCTTG.............CAAAGTTC
TTTTTGACCAGTCGTGAGGG.............TTCCAAGG
ACTGTCCATGCCCCTGTGTA.............TGGTACGT
CGTAAAACAATAATCCGGTG.............TTAAACCG
....
....
....
(Hundreds of millions of reads)
Cell barcode
UMI
cDNA (50-bp sequenced)
Cell 1
TTGCCGTGGTGT GGCGGGGA.............CGGTGTTA
TTGCCGTGGTGT TATGGAGG.............CCAGCACC
TTGCCGTGGTGT TCTCAAGT.............AAAATGGC
....
Cell 2
CGTTAGATGGCA GGGCCGGG.............CTCATAGT
CGTTAGATGGCA ACGTTATA.............ACGCGTAC
CGTTAGATGGCA TCGAGATT.............AGCCCTTT
....
Cell 3
AAATTATGACGA AGTTTGTA.............GGGAATTA
AAATTATGACGA AGTTTGTA.............AGATGGGG
AAATTATGACGA TGTGCTTG.............GACTGCAC
....
Cell 4
GTTAAACGTACC CTAGCTGT.............GATTTTCT
GTTAAACGTACC GCAGAAGT.............GTTGGCGT
GTTAAACGTACC AAGGCTTG.............CAAAGTTC
GTTAAACGTACC TTCCGGTC.............TCCAGTCG
....
....
....
(Thousands of cells)
nced)
Cell 1
TTGCCGTGGTGT GGCGGGGA.............CGGTGTT
TTGCCGTGGTGT TATGGAGG.............CCAGCAC
TTGCCGTGGTGT TCTCAAGT.............AAAATGG
....
Cell 2
CGTTAGATGGCA GGGCCGGG.............CTCATAG
CGTTAGATGGCA ACGTTATA.............ACGCGTA
CGTTAGATGGCA TCGAGATT.............AGCCCTT
....
Cell 3
AAATTATGACGA AGTTTGTA.............GGGAATT
AAATTATGACGA AGTTTGTA.............AGATGGG
AAATTATGACGA TGTGCTTG.............GACTGCA
....
Cell 4
GTTAAACGTACC CTAGCTGT.............GATTTTC
GTTAAACGTACC GCAGAAGT.............GTTGGCG
GTTAAACGTACC AAGGCTTG.............CAAAGTT
GTTAAACGTACC TTCCGGTC.............TCCAGTC
....
....
....
DDX51
NOP2
ACTB
LBR
ODF2
HIF1A
ACTB
RPS15
GTPBP4
GAPDH
ARL1
(Thousands of cells)

Flowchart: from .bcl to count matrix

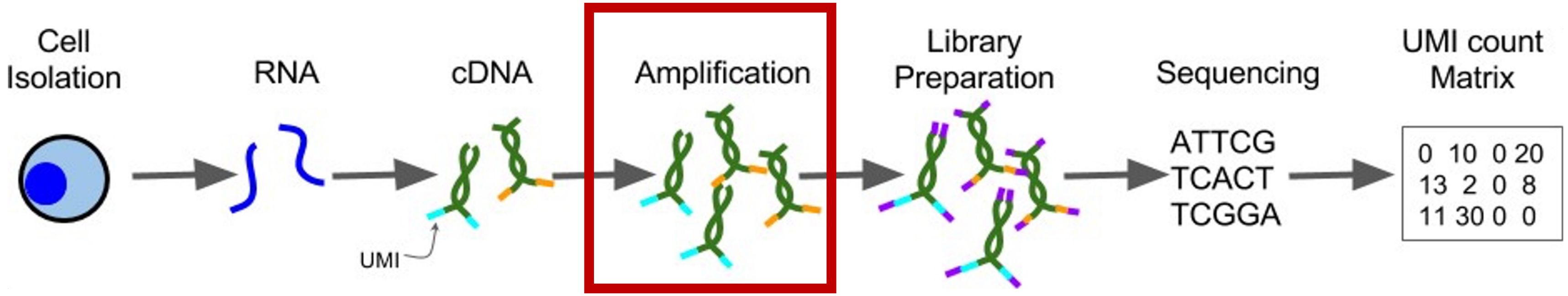
	✓	Get bcl file	That is the role of the sequencing machine!
	✓	Create fastq files	bcl2fastq
	✓	QC: assess overall quality	fastqc
Cellranger	✓	QC: remove low quality reads	Automatic in most aligners
	✓	Align fastq to BAM	STAR / Kallisto / Bowtie / …
	●	QC: collapse UMI	
	●	View alignment	
	●	Count matrix	

PCR duplicates

- PCR amplifications is an issue for both bulk and scRNAseq.

- During the PCR amplification process, some transcripts become over represented in the final library compared to their true abundance.

- The problem worsened when a high number of PCR cycles are used to generate the sequencing library, for example in scRNAseq due to the low amount of starting material.



Cell Isolation

RNA

cDNA

UMI

Amplification

Library Preparation

Sequencing

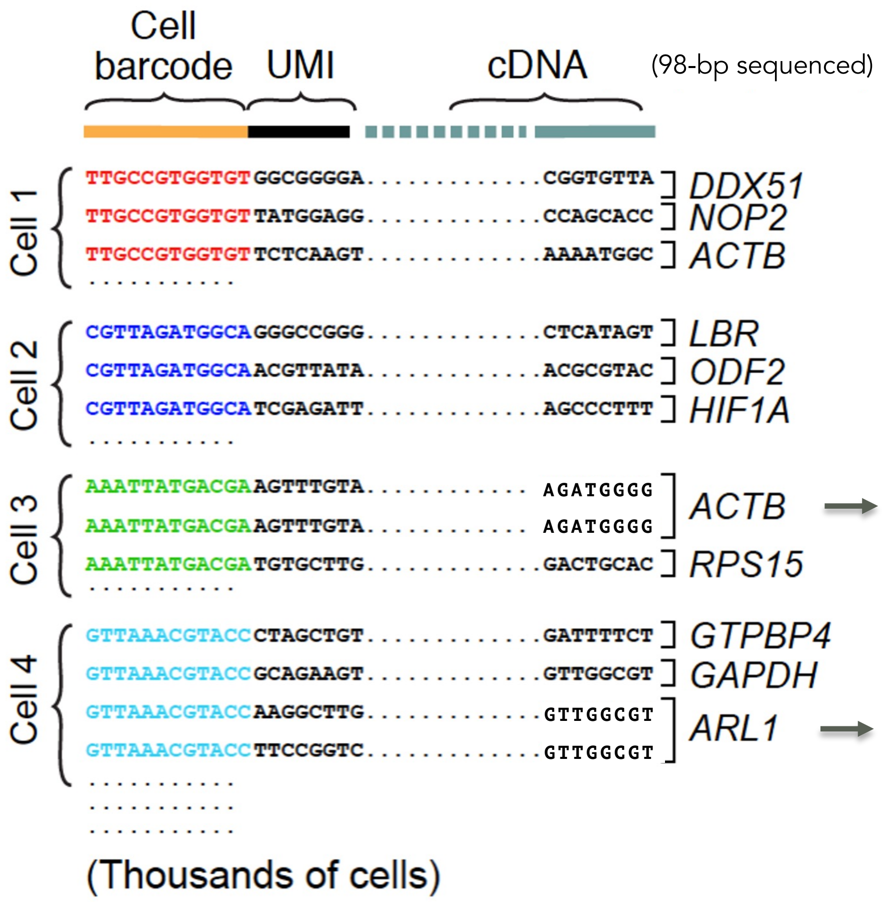
ATTCG
TCACT
TCGGA

UMI count Matrix

```
0  10  0 20
13  2  0  8
11 30  0  0
```

Solution: using UMIs

UMIs enable sequencing reads to be assigned to **individual transcript molecules** and then the removal of amplification noise and biases from scRNAseq data.



Cell barcode
UMI
cDNA
(98-bp sequenced)
Cell 1
TTGCCGTGGTGT GGCGGGGA............CGGTGTTA DDX51
TTGCCGTGGTGT TATGGAGG............CCAGCACC NOP2
TTGCCGTGGTGT TCTCAAGT............AAAATGGC ACTB
Cell 2
CGTTAGATGGCA GGGCCGGG............CTCATAGT LBR
CGTTAGATGGCA ACGTTATA............ACGCGTAC ODF2
CGTTAGATGGCA TCGAGATT............AGCCCTTT HIF1A
Cell 3
AAATTATGACGA AGTTTGTA............ AGATGGGG
AAATTATGACGA AGTTTGTA............ AGATGGGG ACTB
AAATTATGACGA TGTGCTTG............GACTGCAC RPS15
Cell 4
GTTAAACGTACC CTAGCTGT............GATTTTCT GTPBP4
GTTAAACGTACC GCAGAAGT............GTTGGCGT GAPDH
GTTAAACGTACC AAGGCTTG............GTTGGCGT
GTTAAACGTACC TTCCGGTC............GTTGGCGT ARL1
(Thousands of cells)

2 reads, 1 molecule

2 reads, 2 molecules

	Cell1	Cell2	Cell3	Cell4
ACTB	1	0	1	0
ARL1	0	0	0	**2**

Columns: cells

Rows: features

Solution: using UMIs

“Error-correcting” almost-correct UMIs:

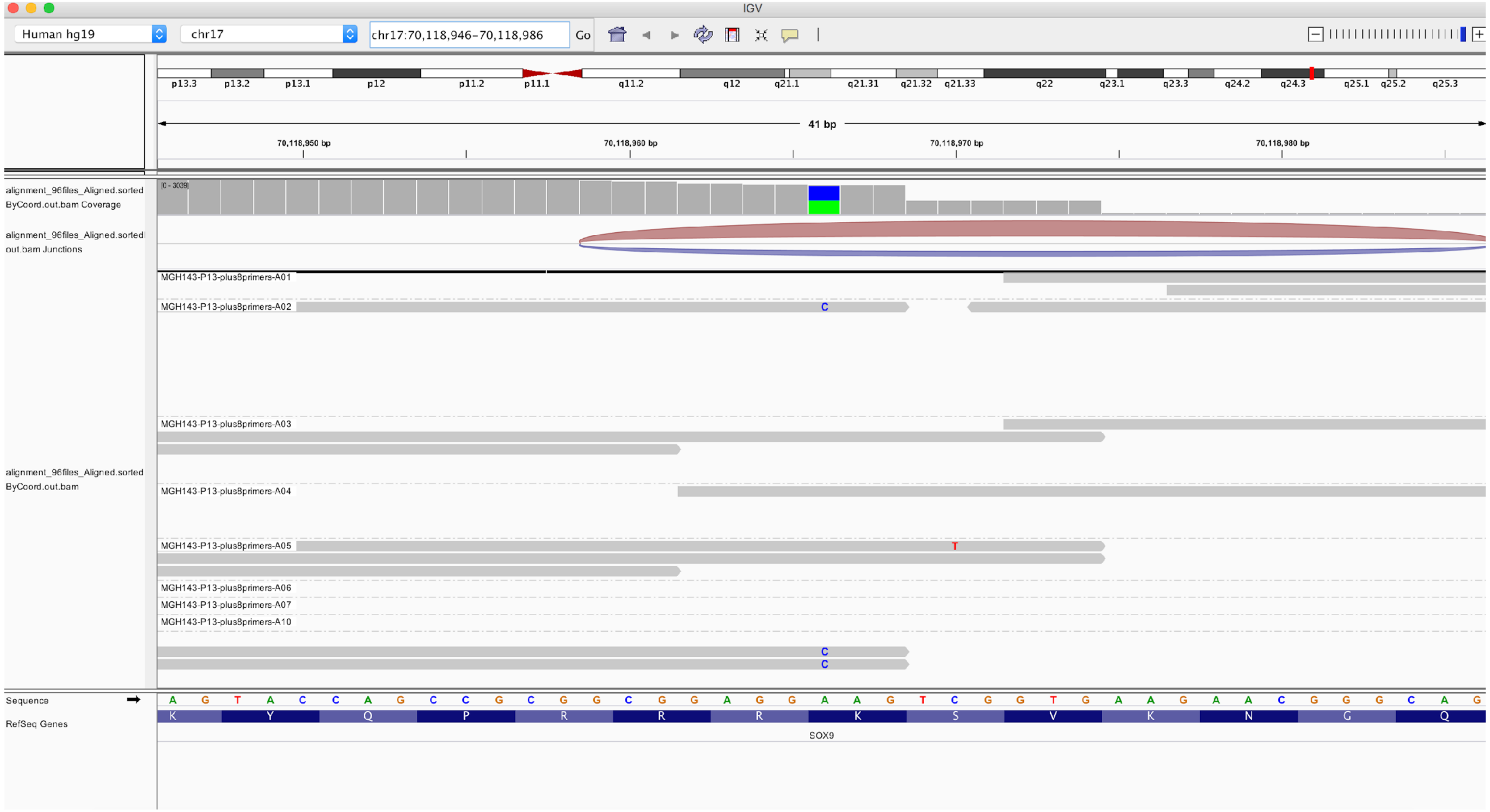
A UMI that is 1-Hamming-distance away from another UMI (with more reads) for the same cell barcode and gene is corrected to the UMI with more reads.

Flowchart: from .bcl to count matrix

	✓	Get bcl file	That is the role of the sequencing machine!
	✓	Create fastq files	bcl2fastq
	✓	QC: assess overall quality	fastqc
Cellranger	✓	QC: remove low quality reads	Automatic in most aligners
	✓	Align fastq to BAM	STAR / Kallisto / Bowtie / …
	✓	QC: collapse UMI	Automatic in most aligners
	●	View alignment	
	○	Count matrix	

Visualization with IGV

Point mutation, splice junction, …



IGV

Human hg19

chr17

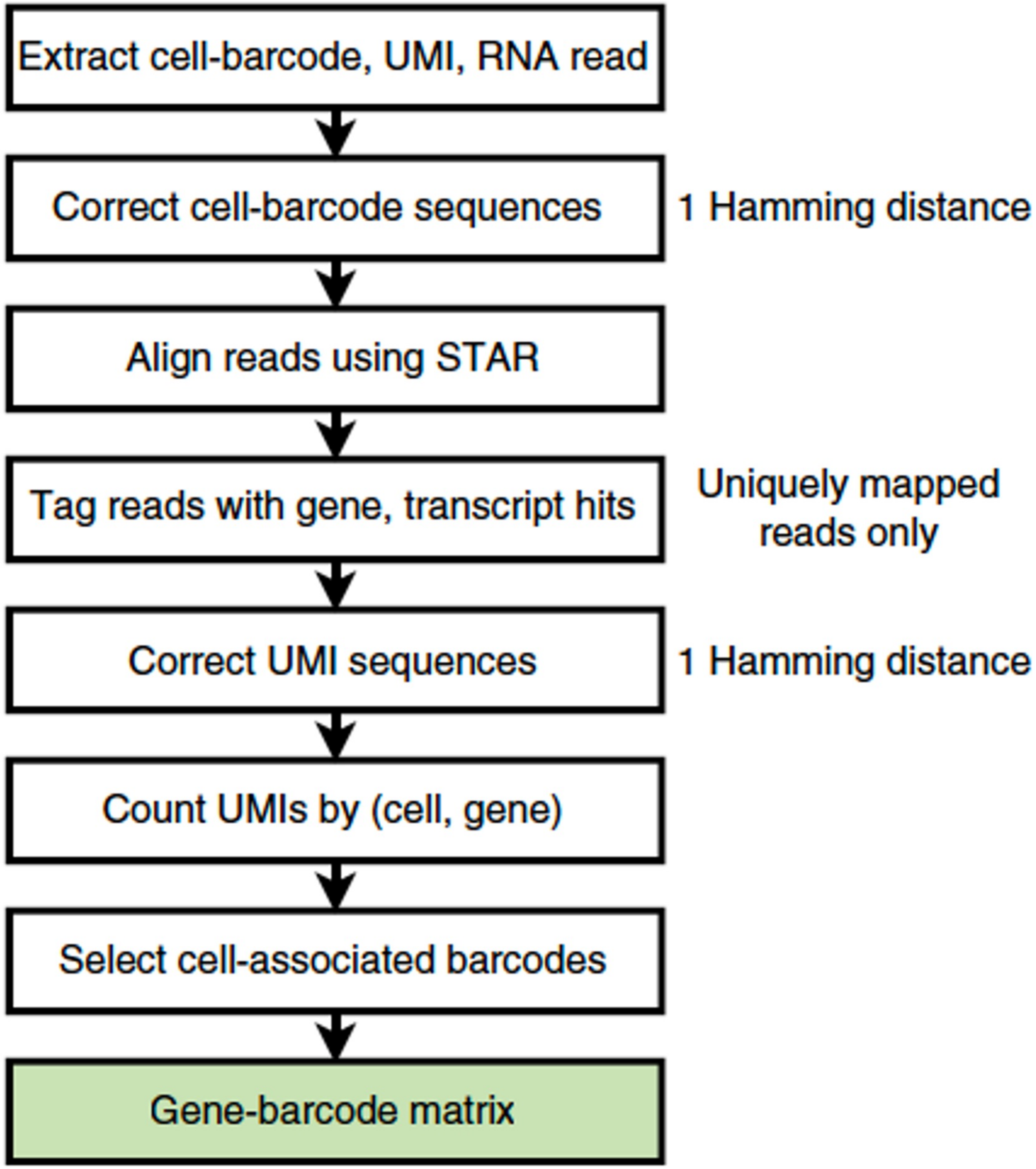
chr17:70,118,946-70,118,986

Go
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q21.33
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.2
q25.3
41 bp
70,118,950 bp
70,118,960 bp
70,118,970 bp
70,118,980 bp
alignment_96files_Aligned.sorted ByCoord.out.bam Coverage
[0 - 3039]
alignment_96files_Aligned.sorted out.bam Junctions
alignment_96files_Aligned.sorted ByCoord.out.bam
MGH143-P13-plus8primers-A01
MGH143-P13-plus8primers-A02
C
MGH143-P13-plus8primers-A03
MGH143-P13-plus8primers-A04
MGH143-P13-plus8primers-A05
T
MGH143-P13-plus8primers-A06
MGH143-P13-plus8primers-A07
MGH143-P13-plus8primers-A10
C
C
Sequence
A G T A C C A G C C G C G G C G G A G G A A G T C G G T G A A G A A C G G G C A G
RefSeq Genes
K Y Q P R R R K S V K N G Q
SOX9

Flowchart: from .bcl to count matrix

	✓	Get bcl file	That is the role of the sequencing machine!
	✓	Create fastq files	bcl2fastq
	✓	QC: assess overall quality	fastqc
Cellranger	✓	QC: remove low quality reads	Automatic in most aligners
	✓	Align fastq to BAM	STAR / Kallisto / Bowtie / …
	✓	QC: collapse UMI	Automatic in most aligners
	✓	View alignment	Automatic in most aligners
	●	Count matrix	

Cellranger count pipeline



Extract cell-barcode, UMI, RNA read
Correct cell-barcode sequences
1 Hamming distance
Align reads using STAR
Tag reads with gene, transcript hits
Uniquely mapped reads only
Correct UMI sequences
1 Hamming distance
Count UMIs by (cell, gene)
Select cell-associated barcodes
Gene-barcode matrix

Cellranger count pipeline

Which reads are considered for UMI counting by Cell Ranger?

1. Only reads with a valid UMI and a valid 10x barcode.

2. No bases with base quality < 10.

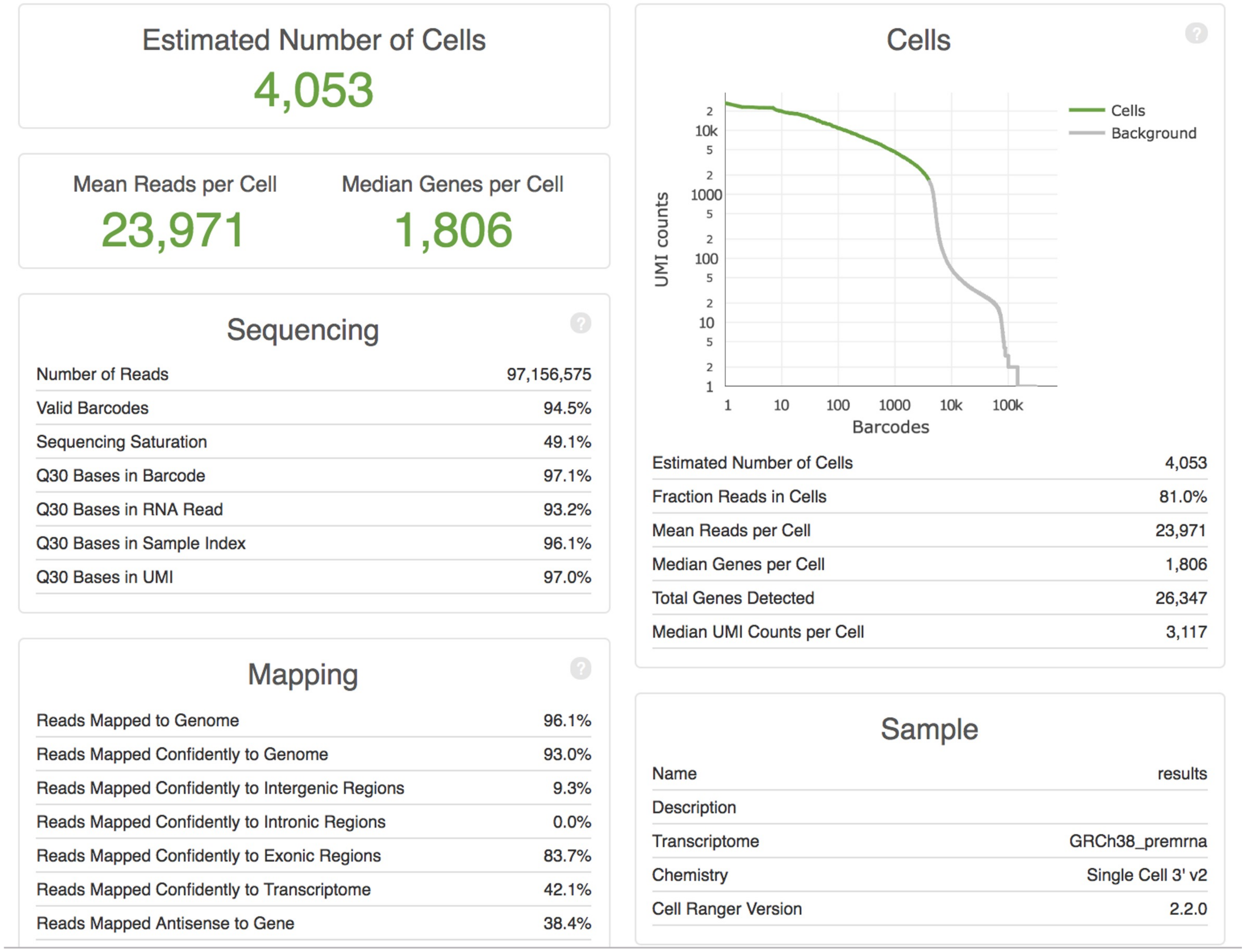
3. Read maps to exactly one gene.

4. Overlaps an exon by at least 50% in a way consistent with annotated splice junctions and strand annotation.

5. Multiple reads that map to the same UMI will only count once.

Cellranger count pipeline

Cellranger output



Estimated Number of Cells

4,053

Mean Reads per Cell	Median Genes per Cell
23,971	1,806

Sequencing

Number of Reads	97,156,575
Valid Barcodes	94.5%
Sequencing Saturation	49.1%
Q30 Bases in Barcode	97.1%
Q30 Bases in RNA Read	93.2%
Q30 Bases in Sample Index	96.1%
Q30 Bases in UMI	97.0%

Mapping

Reads Mapped to Genome	96.1%
Reads Mapped Confidently to Genome	93.0%
Reads Mapped Confidently to Intergenic Regions	9.3%
Reads Mapped Confidently to Intronic Regions	0.0%
Reads Mapped Confidently to Exonic Regions	83.7%
Reads Mapped Confidently to Transcriptome	42.1%
Reads Mapped Antisense to Gene	38.4%

Cells

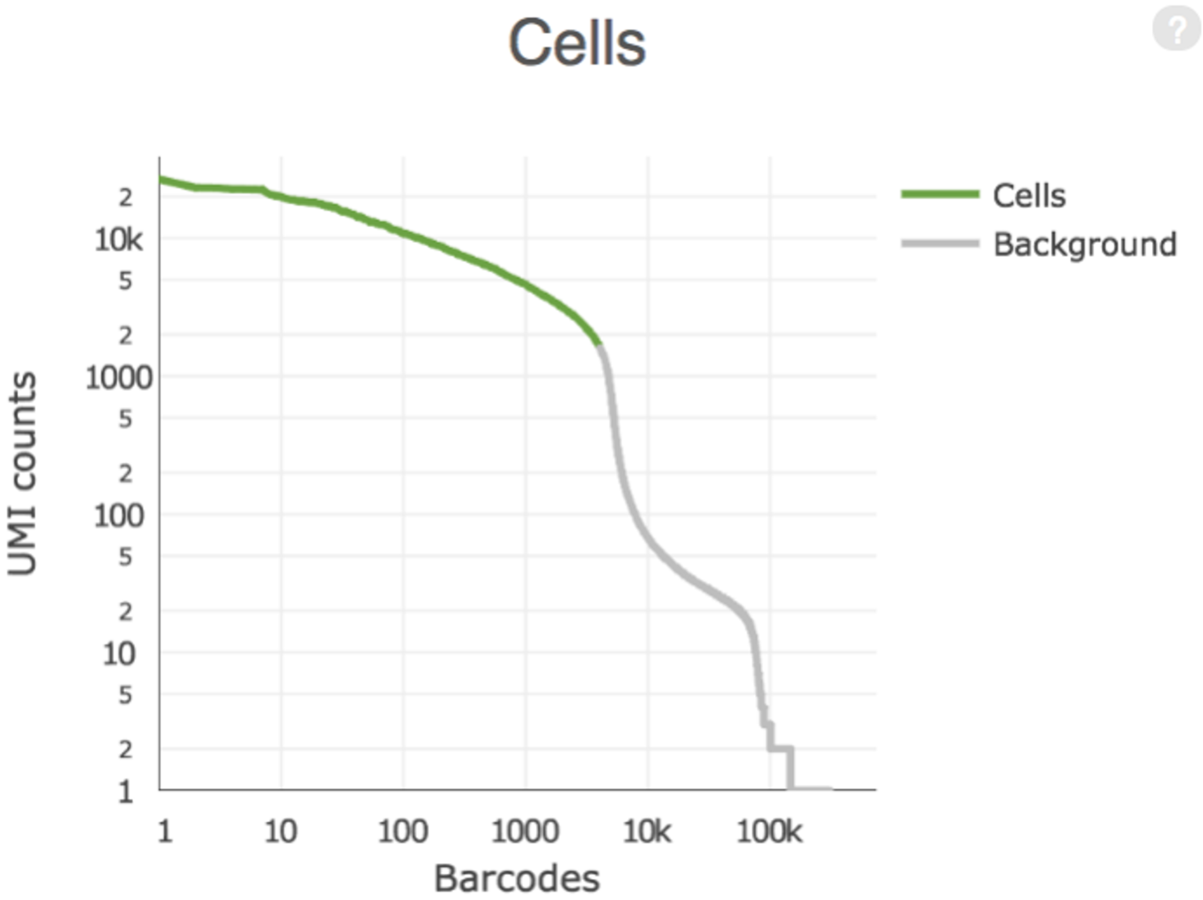
Estimated Number of Cells	4,053
Fraction Reads in Cells	81.0%
Mean Reads per Cell	23,971
Median Genes per Cell	1,806
Total Genes Detected	26,347
Median UMI Counts per Cell	3,117

Sample

Name	results
Description	
Transcriptome	GRCh38_premrna
Chemistry	Single Cell 3' v2
Cell Ranger Version	2.2.0

Cellranger output

Some cell barcodes have many UMIs, but most do not.



Sequencing depth

How deeply do I need to sequence?

- At high sequencing depth, we can detect:

  - The expression of rare splice variants.
  - Quantitative modulations in transcript abundance.

- Lower sequencing depth can still be very powerful :

  - Catalog of cell types.
  - Catalog of transcriptional programs.

Flowchart: from .bcl to count matrix

	✓	Get bcl file	That is the role of the sequencing machine!
	✓	Create fastq files	bcl2fastq
	✓	QC: assess overall quality	fastqc
Cellranger	✓	QC: remove low quality reads	Automatic in most aligners
	✓	Align fastq to BAM	STAR / Kallisto / Bowtie / …
	✓	QC: collapse UMI	Automatic in most aligners
	✓	View alignment	Automatic in most aligners
	✓	Count matrix	Cellranger

Any question?

Physalia Courses