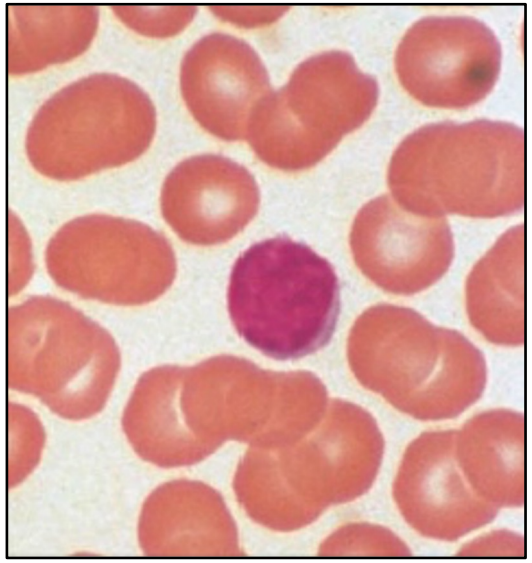
Pre-processing of scRNA-Seq data

HCA Latin America

Single-Cell RNA-Seq Computational Workshop

October 23, 2023



BROAD INSTITUTE

KLARMAN CELL OBSERVATORY AT BROAD INSTITUTE

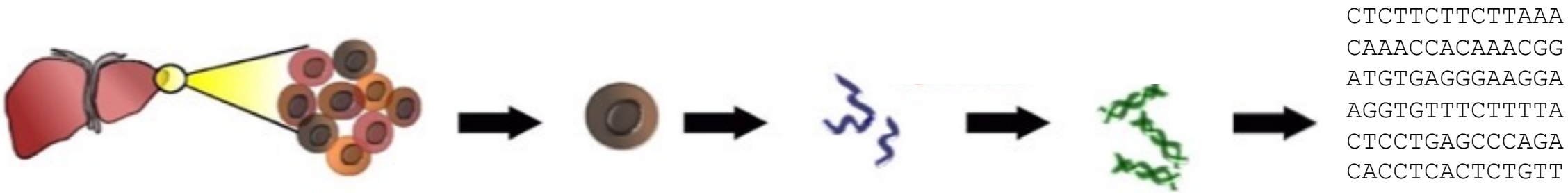
HUMAN CELL ATLAS

humancellatlas.org

Outline: Pre-processing scRNA-Seq data

- Going from raw sequencing files to gene count matrices.
  - BCL to FASTQ
  - Alignment to reference genome
  - Counting reads

Single cell RNA-Seq experimental workflow

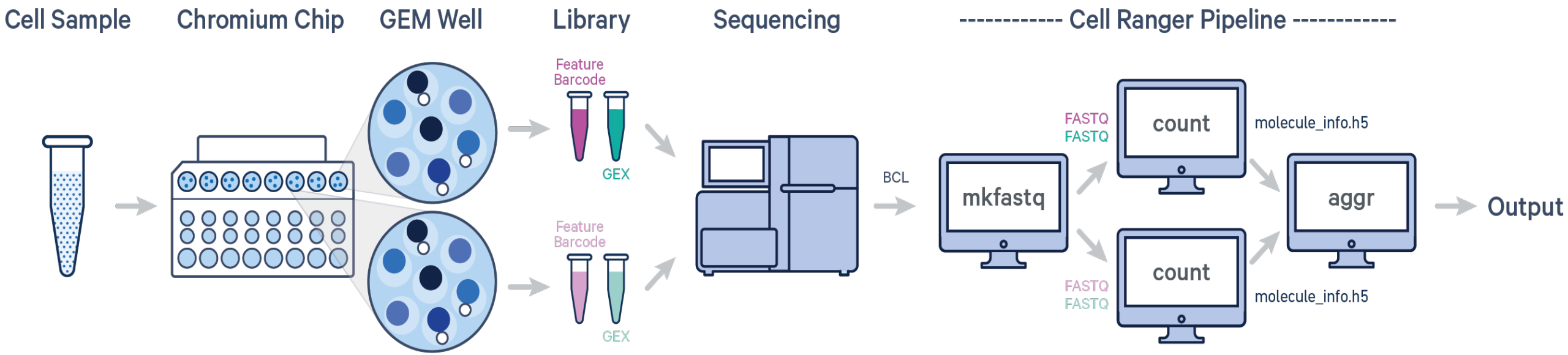


Solid Tissue Dissociation Single Cell Isolation RNA DNA Sequencing

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	…	Cell 70K
Tcf7	3	3	0	0	0	0	…	4
Bach2	2	4	1	0	0	0	…	2
Prf1	1	0	5	3	1	1	…	1
Gzma	0	0	3	1	0	0	…	0
Pdcd1	0	1	1	0	4	6	…	0
Eomes	0	0	1	0	3	3	…	1
…	…	…	…	…	…	…	…	…
Gene 20K	2	1	0	1	0	0	…	3

en.wikipedia.org/wiki/Single_cell_sequencing

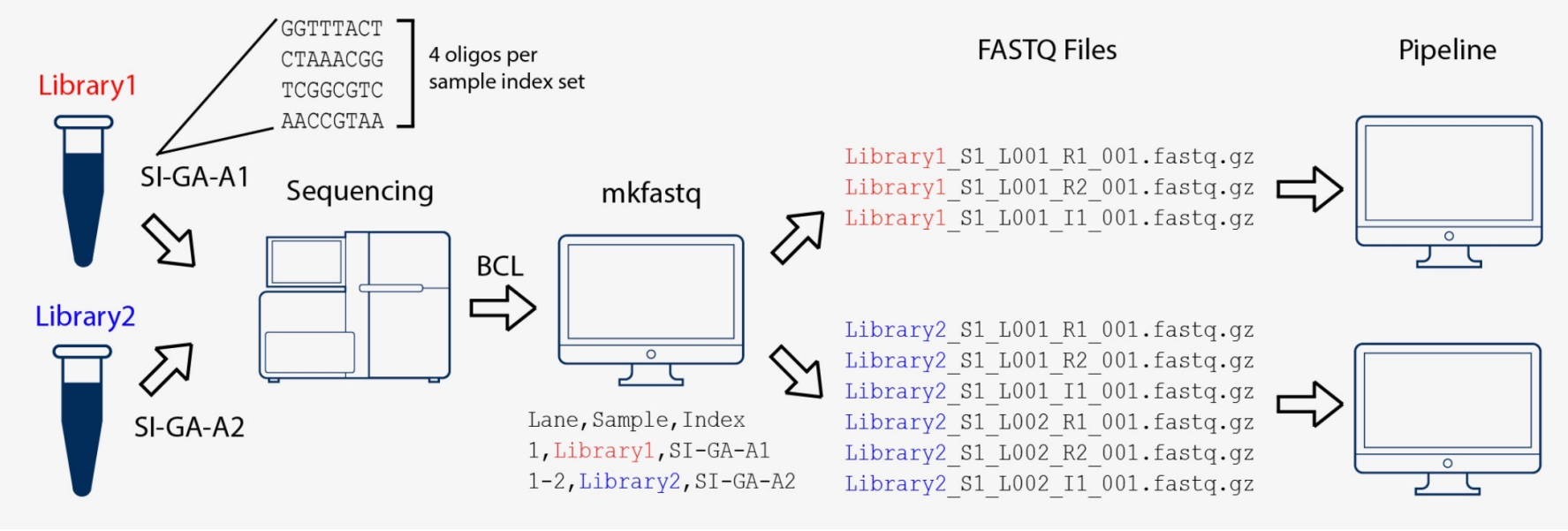
Raw sequencing files to gene count matrices (Cell Ranger)



Source: 10X Genomics

Raw sequencing files to count matrices
Step 1: BCL to FASTQ

- The primary output of Illumina sequencing instruments are per-cycle base call files in BCL format.
- The first step is to convert these BCL files to fastq files. Steps include separating reads into individual fastq files based on their barcode (demultiplexing) and moving unique molecular identifier (UMI) bases from the read to the fastq header.
- Tools: cellranger mkfastq, bcl2fastq, or BCL Convert.



Source: 10X Genomics

Raw sequencing files to count matrices
Step 1: BCL to FASTQ

- A FASTQ file is a text file that contains the sequence data which consists of:

  1. A sequence identifier with information about the sequencing run and the cluster.
  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 as phred scores using ASCII characters to represent the numerical quality scores

$$P = 10^{\frac{-Q}{10}}$$

Table 1: Q-Scores and Error Probabilities

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

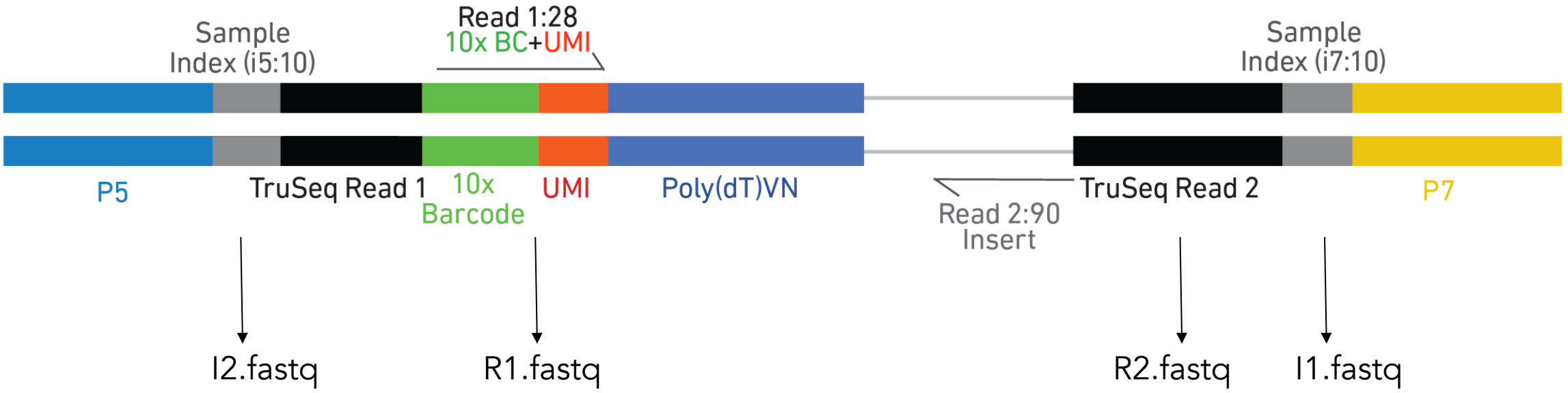
- Example sequence in a FASTQ

```
@HWI-ST808:130:H0B8YADXX:1:1101:2088:2222:CELL_GGTCCA:UMI_CCCT
AGGAAGATGGAGGAGAGAAGGCGGTGAAAGAGACCTGTAAAAAGCCACCGN
+
@@@DDBD>=AFCF+<CAFHDECII:DGGGHGIGGIIIEHGIIIGIIDHII#
```

Raw sequencing files to count matrices
Step 1: BCL to FASTQ

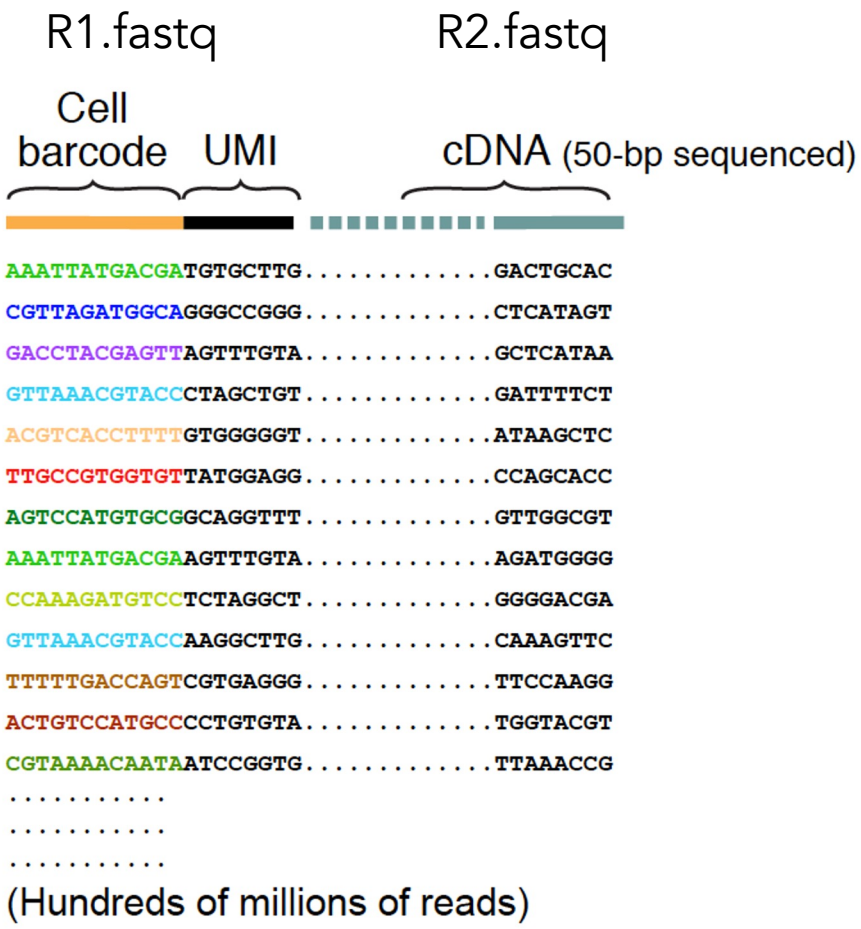
Each fastq file generated by **`bcl2fastq`** (or **`bcl-convert`**) contains different information:

- I1.fastq, I2.fastq contains sample index (for pooling together multiple samples)
- R1.fastq contains cell barcode + UMI
- R2.fatsq contains transcript information



Raw sequencing files to count matrices
Step 1: BCL to FASTQ

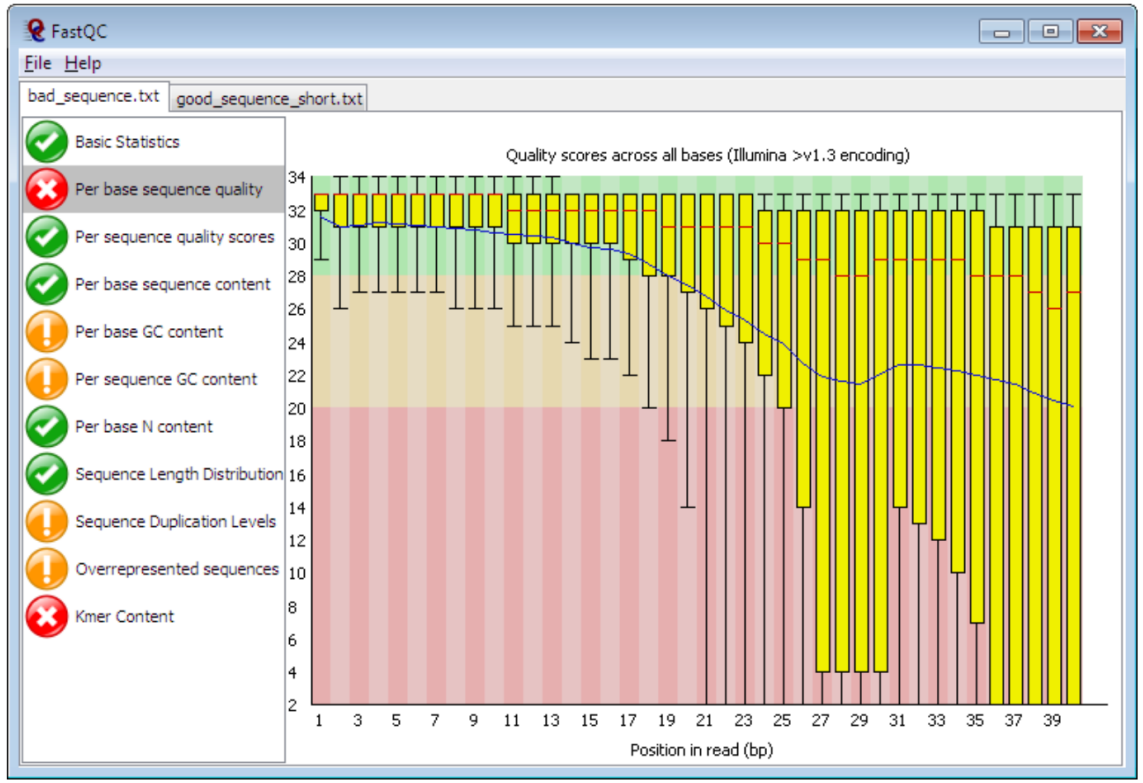
An example of what the reads look like after demultiplexing…

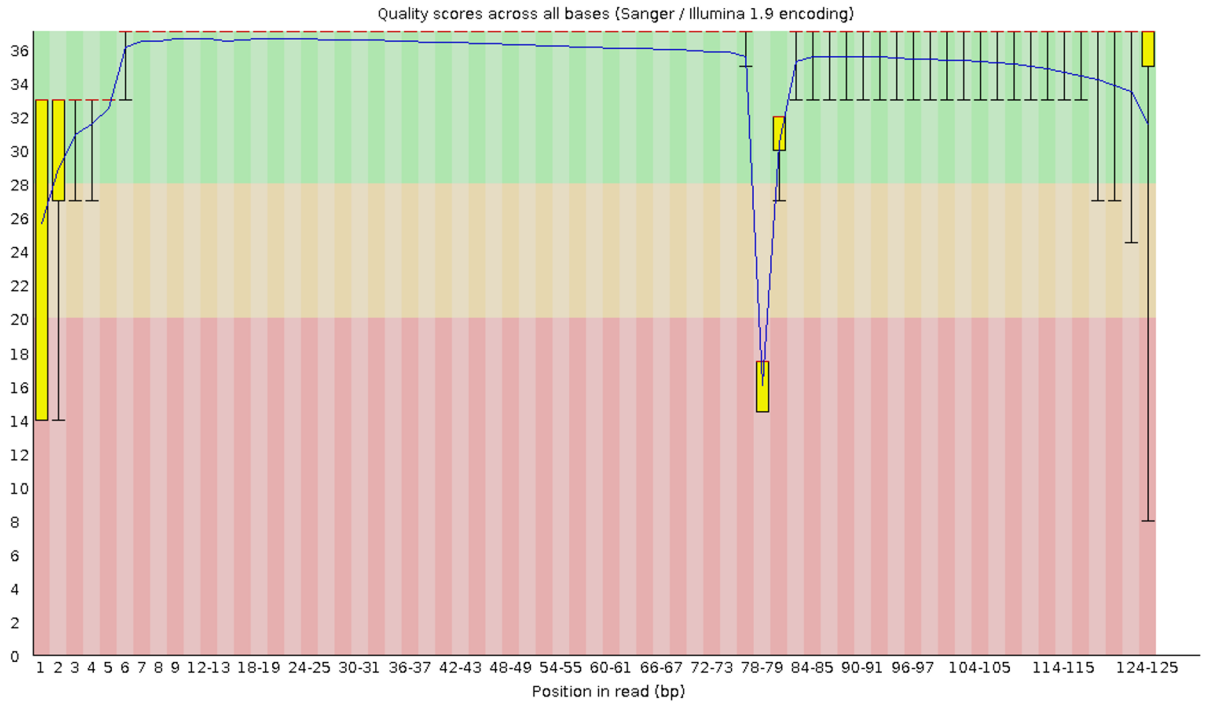


Raw sequencing files to count matrices
Sequencing read quality control

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.



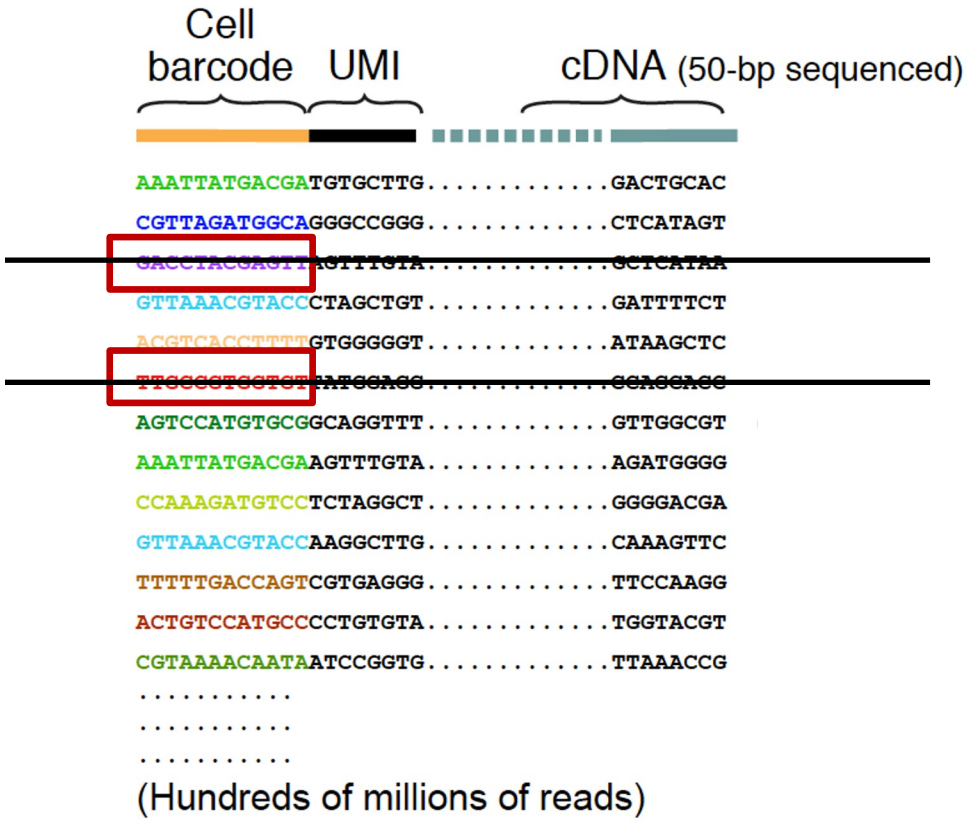


FastQC will highlight any areas where this library looks unusual and where you should take a closer look. The program is not tied to any specific type of sequencing technique and can be used to look at libraries coming from a large number of different experiment types (Genomic Sequencing, ChIP-Seq, RNA-Seq,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

Raw sequencing files to count matrices
Unrecognized cell barcodes

It is important to consider (and remove!):

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



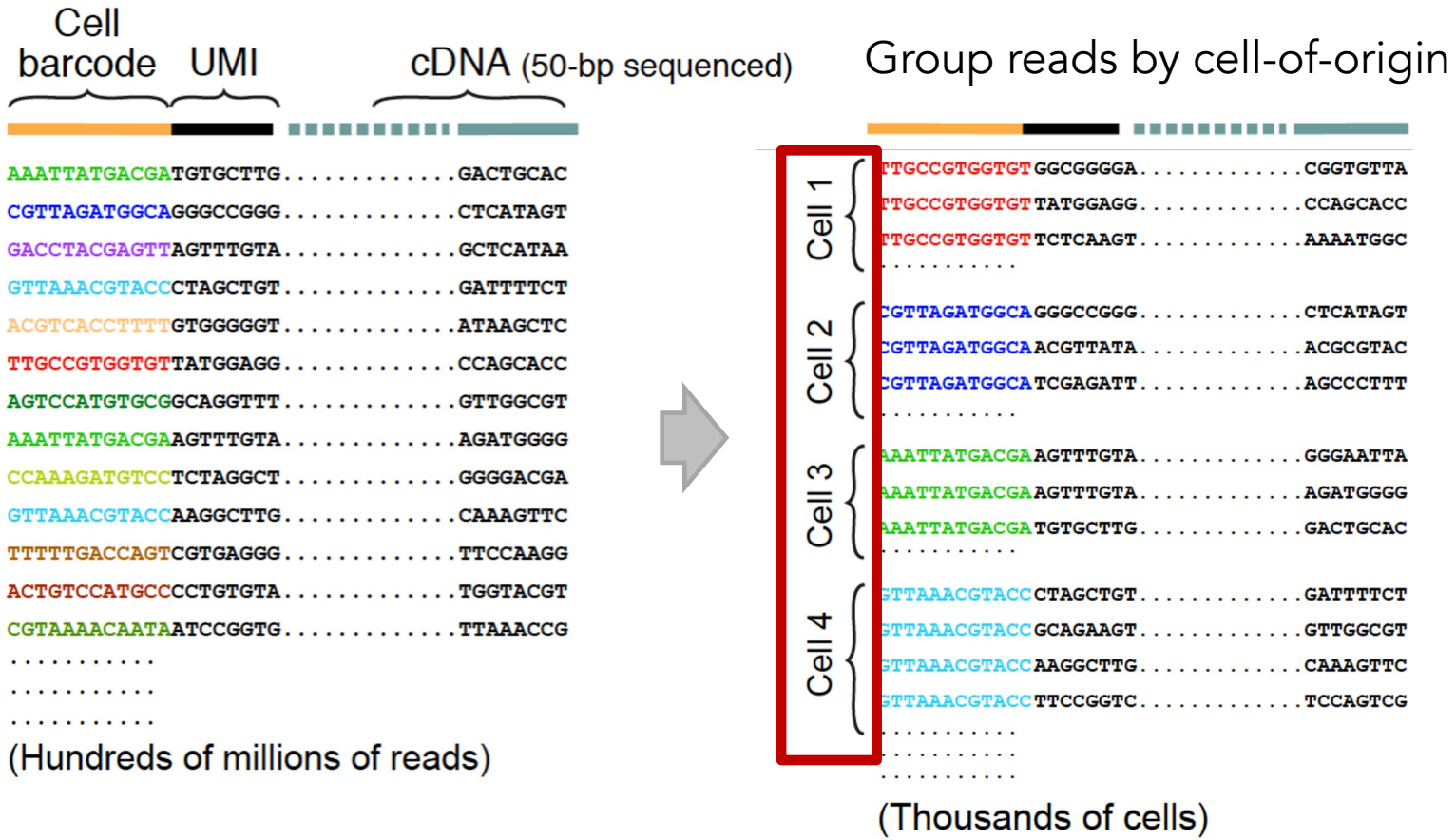
What is a barcode whitelist?
A barcode whitelist is the list of all known barcode sequences that have been included in the assay kit and are available during library preparation.

For example, there are roughly 3M cell barcodes in the whitelist for Cell Ranger's Single Cell 3'v3 applications. Here are the first 10 lines of the corresponding barcode whitelist 3M-february-2018.txt:

```
AAACCCAAGAAACACT
AAACCCAAGAAACCAT
AAACCCAAGAAACCCA
AAACCCAAGAAACCCG
AAACCCAAGAAACCTG
AAACCCAAGAAACGAA
AAACCCAAGAAACGTC
AAACCCAAGAAACTAC
AAACCCAAGAAACTCA
AAACCCAAGAAACTGC
```

Raw sequencing files to count matrices
Step 1: BCL to FASTQ

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



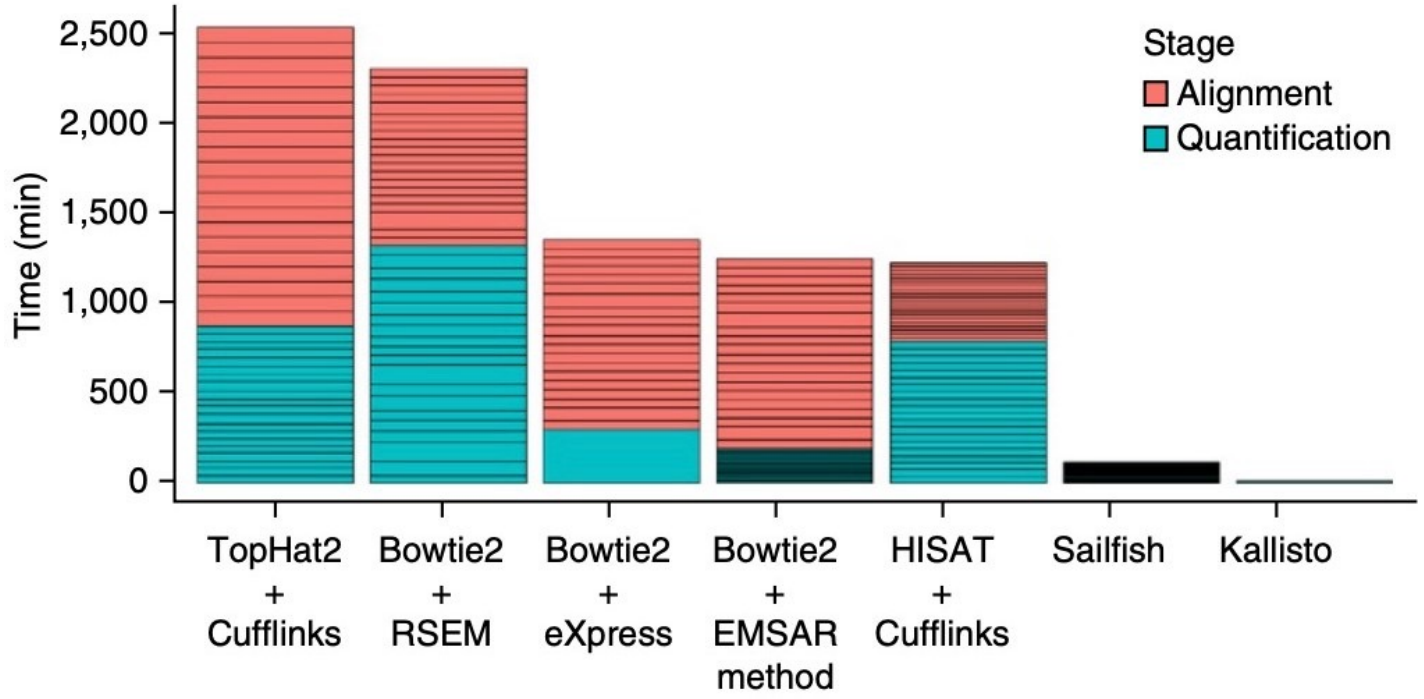
Raw sequencing files to count matrices
Step 2: Alignment to reference genome

- The next step is to determine which gene each read originated from. To do this, the read sequences are mapped to a precompiled genome reference.

- Be mindful of reference versions and sources. Use the latest reference when possible.
  - Human reference, GRCh38 (GENCODE v32/Ensembl 98)
  - Mouse reference, mm10 (GENCODE vM23/Ensembl 98)

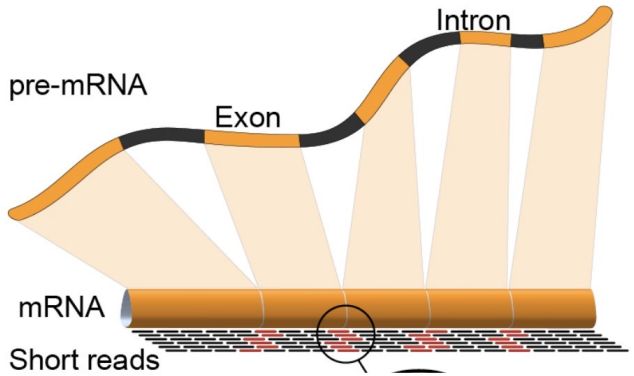
Raw sequencing files to count matrices
Step 2: Alignment to reference genome

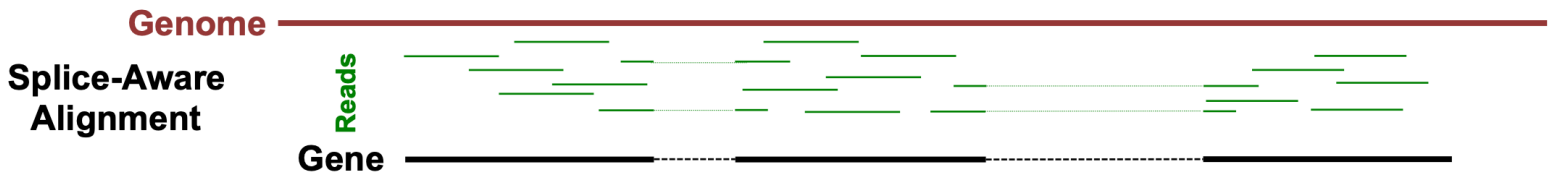
- The choice of aligner is often a personal preference and also dependent on the computational resources that are available to you.



Raw sequencing files to count matrices
Step 2: Alignment to reference genome

- 10X Cellranger uses the STAR aligner.
- Because of widespread splicing in animal genomes, read alignment against a genome should be done with a splice-aware aligner.





Raw sequencing files to count matrices
Step 2: Alignment to reference genome

Most aligners (included STAR-based cellranger) will map scRNA-Seq reads to a transcriptome index.

A transcriptome index consists of:

1. A **genome sequence** reference

FASTQ example

```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAACCCC
TAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC
CTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCC
CAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACCCTAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGACAACGCAGCTC
CGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGGCGCGCCGCGC
CGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGC
CGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGC
GCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCGTGGCGCAGGC
GCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGG
GCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGGCGCAGAGACG
GGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCACTACAGGACCC
GCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCTCTCTTGCTTA
GAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGTATAGTGGTGG
```

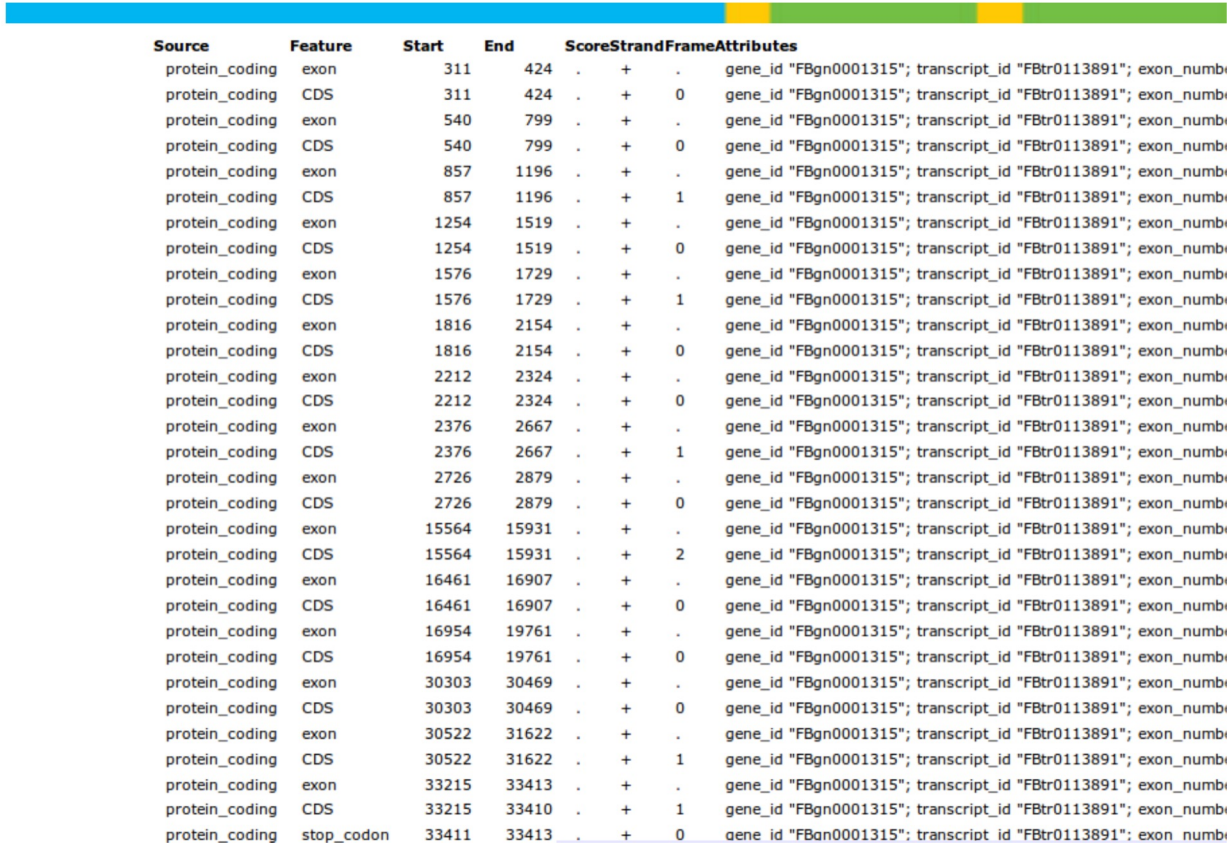
Raw sequencing files to count matrices
Step 2: Alignment to reference genome

Most aligners (included STAR-based cellranger) will map scRNA-Seq reads to a transcriptome index.

A transcriptome index consists of:

1. A **genome sequence** reference
2. A **gene feature annotation** reference

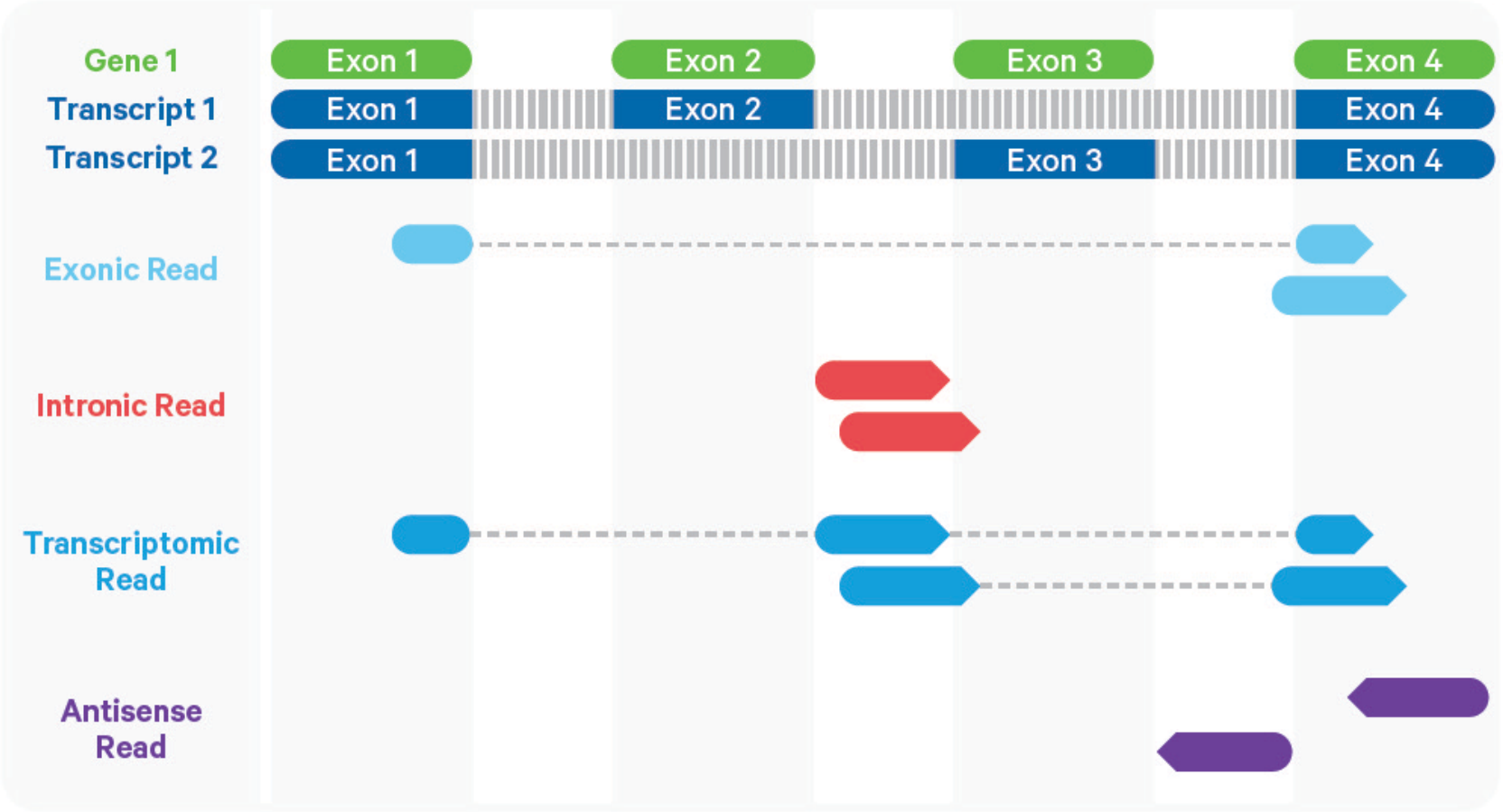
GTF example



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

Raw sequencing files to count matrices
Step 2: Alignment to reference genome

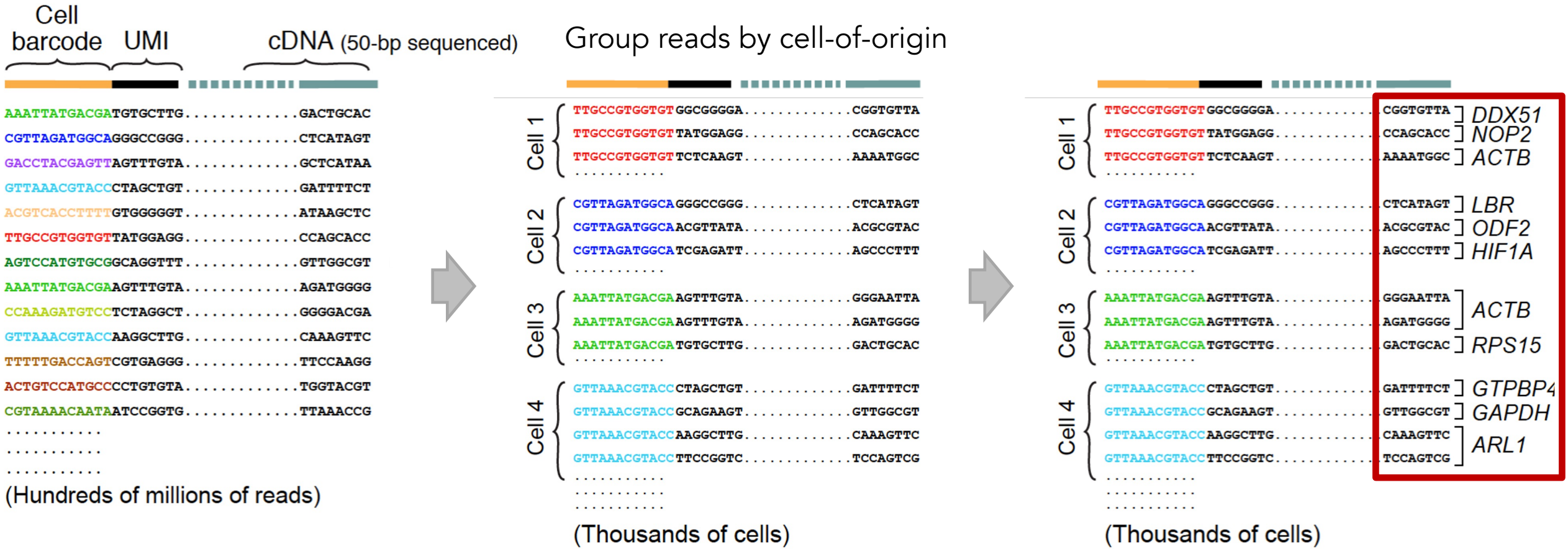
Which reads are considered for UMI counting by Cell Ranger?



https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview

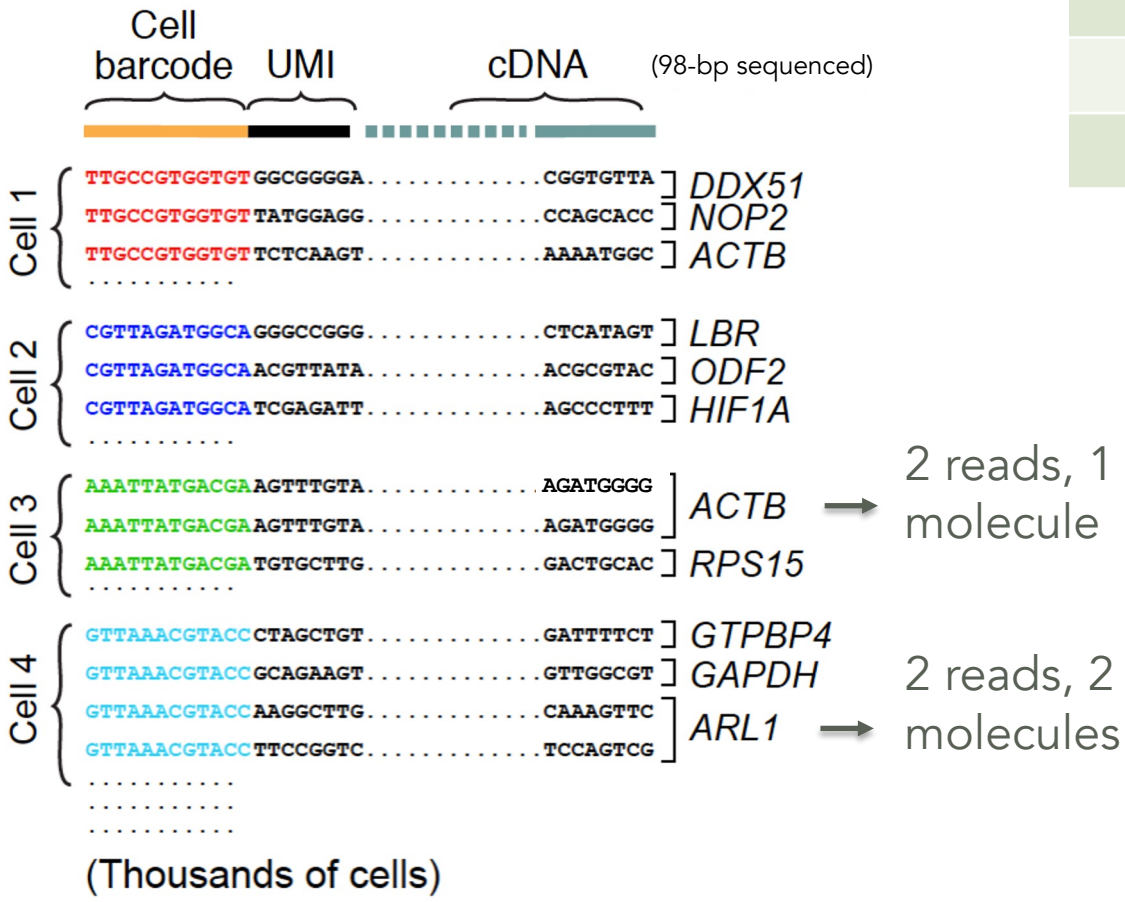
Raw sequencing files to count matrices
Step 3: Counting reads

- Reads that have been confidently mapped to the transcriptome are then assigned to cells based on their barcode (**cell barcode demultiplexing**) and the number of unique RNA molecules corresponding to each gene within each cell are counted (**UMI deduplication**).



Raw sequencing files to count matrices
Step 3: Counting reads

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



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

Columns: cells
Rows: features

Raw sequencing files to count matrices
Step 3: Counting reads

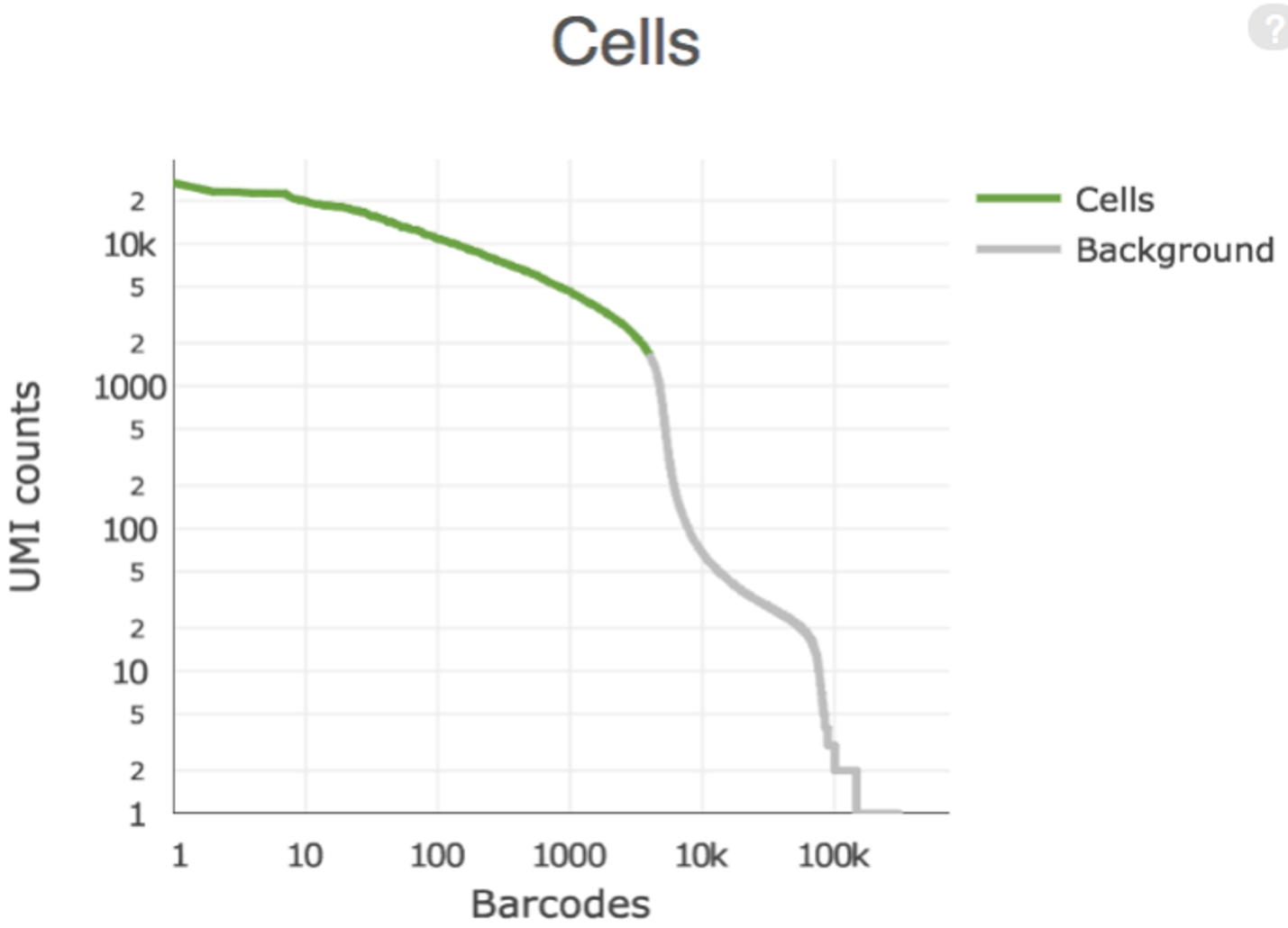
- Reads that have been confidently mapped to the transcriptome are then assigned to cells based on their barcode (**cell barcode demultiplexing**) and the the number of unique RNA molecules corresponding to each gene within each cell are counted (**UMI deduplication**).

- The result is the gene *x* cell matrix that is the starting point for downstream analysis.

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	…	Cell 70K
Tcf7	3	3	0	0	0	0	…	4
Bach2	2	4	1	0	0	0	…	2
Prf1	1	0	5	3	1	1	…	1
Gzma	0	0	3	1	0	0	…	0
Pdcd1	0	1	1	0	4	6	…	0
Eomes	0	0	1	0	3	3	…	1
…	…	…	…	…	…	…	…	…
Gene 20K	2	1	0	1	0	0	…	3

Raw sequencing files to count matrices
Cell Ranger `count` output

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



Cell Ranger `count` output

Summary | Gene Expression | Antibody

125
Estimated Number of Cells

3,200
Mean Reads per Cell

13
Median Genes per Cell

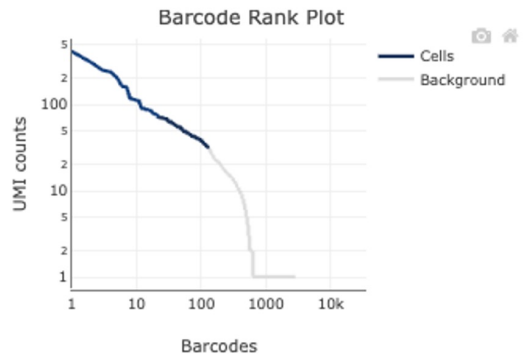
Sequencing

Number of Reads	400,000
Number of Short Reads Skipped	0
Valid Barcodes	94.0%
Valid UMIs	99.9%
Sequencing Saturation	75.0%
Q30 Bases in Barcode	96.4%
Q30 Bases in RNA Read	95.7%
Q30 Bases in UMI	96.3%

Mapping

Reads Mapped to Genome	100.0%
Reads Mapped Confidently to Genome	21.4%
Reads Mapped Confidently to Intergenic Regions	2.6%
Reads Mapped Confidently to Intronic Regions	12.5%
Reads Mapped Confidently to Exonic Regions	6.3%
Reads Mapped Confidently to Transcriptome	16.3%
Reads Mapped Antisense to Gene	2.0%

Cells

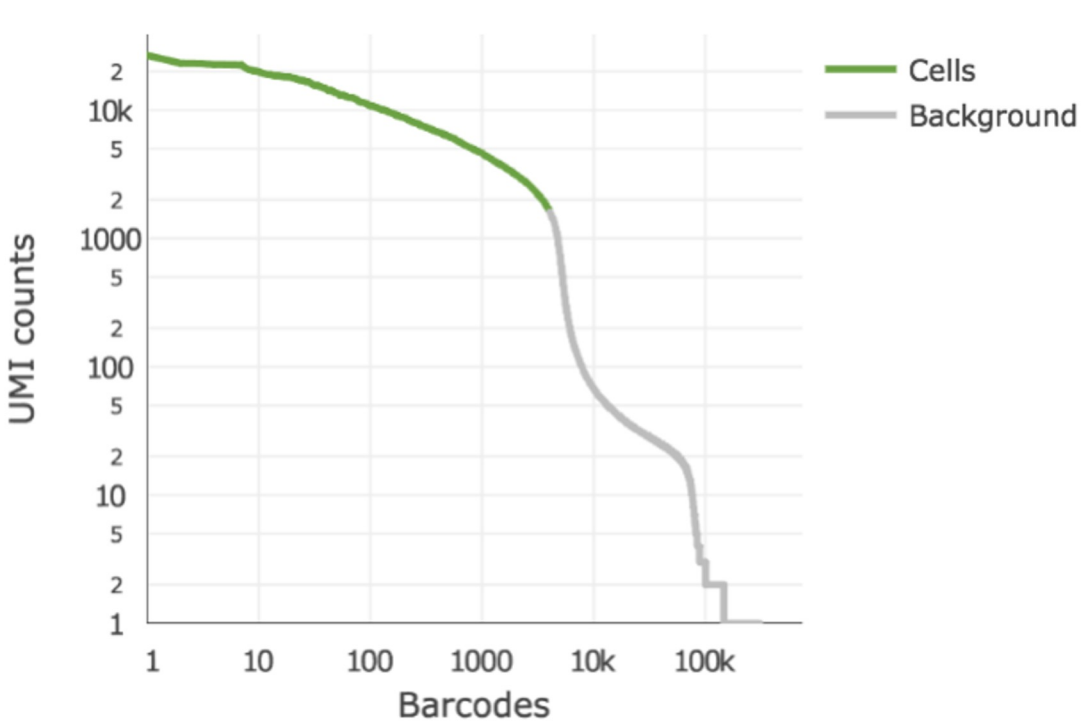


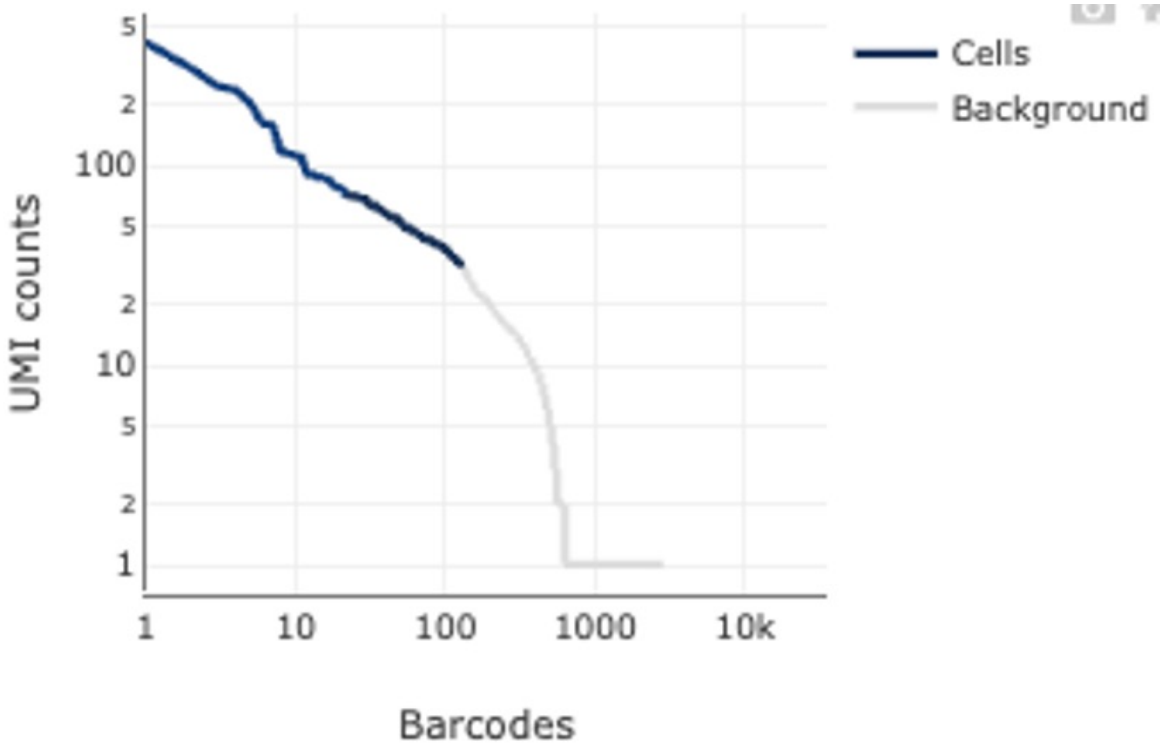
Estimated Number of Cells	125
Fraction Reads in Cells	49.1%
Mean Reads per Cell	3,200
Median UMI Counts per Cell	46
Median Genes per Cell	13
Total Genes Detected	78

Sample

Sample ID	78388_chr21_400K
Sample Description	
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	…refdata-cellranger-chr21-3.0.0
Transcriptome	GRCh38_chr21-3.0.0
Pipeline Version	7.0.0

Raw sequencing files to count matrices
Which cell-barcode rank plot looks higher quality?





There are several alternatives to Cell Ranger pipelines

- Kallisto bustools
  - *We introduced a format for scRNA-seq data that makes possible the development of efficient workflows by virtue of decoupling the computationally demanding step of associating reads to transcripts and genes (alignment) from the other steps required for scRNA-seq preprocessing10. This format, called BUS (barcode, UMI, set), can be produced by pseudoalignment, and rapidly manipulated by a suite of tools called bustools.*

- STARsolo
  - *Built directly into the RNA-seq aligner STAR, which is widely used for mapping bulk RNA-seq data. In STARsolo, read mapping, read-to-gene assignment, cell barcode demultiplexing and UMI collapsing are tightly integrated (Methods 5.1), avoiding input/output bottlenecks and boosting the processing speed. Importantly, STARsolo performs read alignment to the full genome, resulting in a higher accuracy compared to the alignment to transcriptome only.*

Kaminow B. et al. (2021) *bioRxiv*.
Melsted P. et al. (2021) *Nature Biotechnology*.