# Preprocessing (from reads to a count matrix)

**Roberta Menafra**
**29-10-2024**

**Bioinformatician LGTC (Leiden Genome Technology Center)**

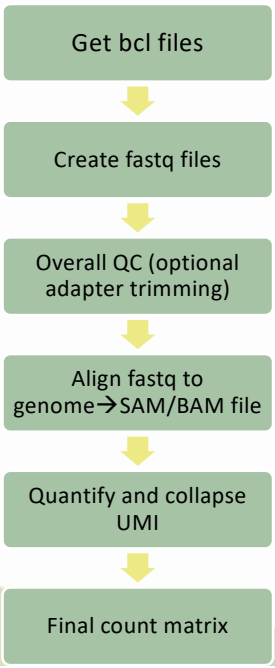
## Preprocessing (from reads to a count matrix)

In DNA sequencing, a **read** is an inferred sequence of base pairs corresponding to all or part of a single DNA fragment.

### In this lecture

- Sequencing data formats
- Data pre-processing
- 10X pipeline (Cell Ranger)
- Results examples

## Preprocessing (from reads to a count matrix)



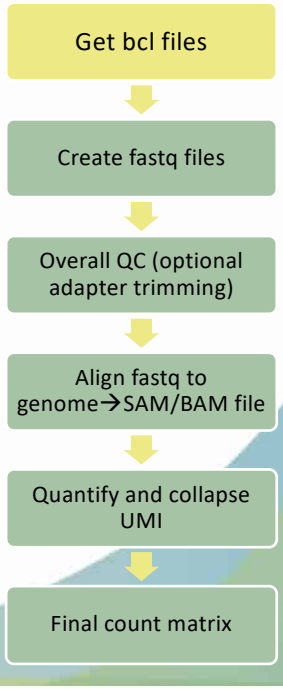
## Common file formats in NGS

- **bcl**
- **fastq**
- **bam**
- mtx, tsv
- hdf5 (.h5, .h5ad)

### BCL

Raw data files in binary base call format.

Illumina offers **bcl2fastq** Conversion Software to convert BCL files.
bcl2fastq is included, standalone conversion software that demultiplexes data and converts BCL files to standard FASTQ file formats for downstream analysis.



## Samples demultiplexing

Sequence runs on NGS instruments are typically carried out with multiple samples pooled together.
An **index tag** (also called a barcode) consisting of a unique sequence (between 6 and 12bp) is added to each sample so that the sequence reads from different samples can be identified.

The other requirement is a sample sheet

```
[Header],,,,,,,,
IEMFileVersion,4,,,,,,,
Date,20-10-2014,,,,,,,
Workflow,GenerateFASTQ,,,,,,,
Application,FASTQ Only,,,,,,,
Assay,Nextera,,,,,,,
Description,,,,,,,,
Chemistry,Amplicon,,,,,,,
,,,,,,,,
[Reads],,,,,,,,
151,,,,,,,,
151,,,,,,,,
,,,,,,,,
[Settings],,,,,,,,
ReverseComplement,0,,,,,,,
Adapter,,,,,,,,
,,,,,,,,
[Data],,,,,,,
Lane,Sample_ID,Sample_Name,Sample_Plate,Sample_Well,I7_Index_ID,index,I5_Index_ID,index2
4,AV_1_HTO,AV_1_HTO,,,,ATTACTCG,,,
5,AV_1_HTO,AV_1_HTO,,,,ATTACTCG,,,
```

```
bcl2fastq --runfolder-dir 190826_E00603_0316_AH3GW3CCX2/ --output-dir HTO/ --sample-sheet samples_HTO.csv --barcode-mismatches 0
```

```
output-dir: HTO/Reports/html/flowcellID/all/all/all/laneBarcode.html
```

**Flowcell Summary**

| Clusters (Raw) | Clusters(PF) | Yield (MBases) |
|---|---|---|
| 1,242,422,208 | 946,733,762 | 285,914 |

**Lane Summary**

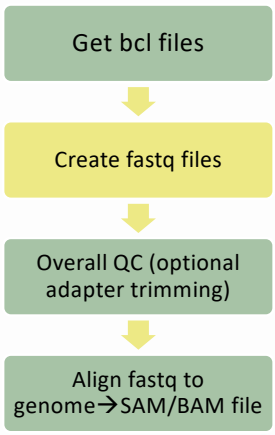
| Lane | Project | Sample | Barcode sequence | PF Clusters | % of the lane | % Perfect barcode | % One mismatch barcode | Yield (Mbases) | % PF Clusters | % >= Q30 bases | Mean Quality Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | default | AV_1_HTO | ATTACTCG | 54,321,382 | 11.52 | 100.00 | NaN | 16,405 | 100.00 | 46.78 | 26.39 |
| 4 | default | Undetermined | *unknown* | 417,222,614 | 88.48 | 100.00 | NaN | 126,001 | 73.60 | 67.87 | 31.88 |
| 5 | default | AV_1_HTO | ATTACTCG | 54,933,100 | 11.56 | 100.00 | NaN | 16,590 | 100.00 | 46.78 | 26.40 |
| 5 | default | Undetermined | *unknown* | 420,256,666 | 88.44 | 100.00 | NaN | 126,918 | 74.21 | 67.77 | 31.85 |

**Top Unknown Barcodes**

| Lane | Count | Sequence | Lane | Count | Sequence |
|---|---|---|---|---|---|
| 4 | 116,377,900 | AGTGGAAC | 5 | 116,758,060 | AGTGGAAC |
| | 107,277,740 | GTCTCCTT | | 107,610,520 | GTCTCCTT |
| | 79,994,540 | TCACATCA | | 80,406,280 | TCACATCA |
| | 74,171,580 | CAGATGGG | | 74,495,820 | CAGATGGG |
| | 19,713,380 | CAAAAGAT | | 19,888,220 | CAAAAGAT |
| | 705,960 | GTCTCCTA | | 706,360 | GTCTCCTA |
| | 629,960 | TCACTCAA | | 638,440 | TCACTCAA |
| | 600,500 | CAGATGGA | | 606,040 | AGTGAACA |
| | 597,720 | AGTGAACA | | 603,420 | CAGATGGA |
| | 564,060 | TCCATCAA | | 574,720 | TCCATCAA |

## FASTQ

- NGS data is often in FASTQ format
- FASTQ is a text-based format for storing both sequence and its corresponding quality score
- Four lines per sequence (read)
  - @ followed by the unique sequence identifier
  - The nucleotide sequence
  - + The quality line break
  - The quality scores in ASCII characters




**View FASTQ Files**

**Viewing entire file**

cat file1.fastq

**Viewing first 10 lines**

head file1.fastq

```
@A00379:133:HMWGLDSXX:1:1101:1163:1000 2:N:0:GAGGATCT
AAAAAAAAATAAAAAAAAATAAAAAAAATGATAAAAAAAATAAAATTAAAAAATATAAAAAAAAAAAAATTTTTTTTTTATTAAGTAAAAAAAATTAAAAATAAATTAAAAAAAAAAAAAAAAAA
+
FF,:::FF,,F,:F,FF,F:FFFF:F:F,,F:F:FFFFF,,:F:,,,:FF:F,,,,F,FFF:FFFF,,,,:,F,,,,:,,,:F,,F::F,F,,,,F::F,FFF,,,,:,FFFFFF,FFFF::,
@A00379:133:HMWGLDSXX:1:1101:1253:1000 2:N:0:GAGGATCT
GATGGTGTAATGCCTGTCTTTAAGTACAAACTGCAAAGTGAAGCCACCCAGATTTATTTCCTTTCTCTGGACCAAGTGTCCCAAACTGAACACTGAGACTGAGGCAGATTTGAAATGATCCAG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,,FFFFFFFFFFFFFFFF:FFFFF,FF:FFFFFFFFFFF,FF,FFFFF
```

## FASTQ

**Header**

@A00379:133:HMWGLDSXX:1:1101:1163:1000 2:N:0:GAGGATCT

Instrument | RunID | FlowcellID | Read Number (Paired 2/2) | Index Sequence

+

FF,:::FF,,F,:F,FF,F:FFFF:F:F,,F:F:FFFFF,,:F:,,,:FF:F,,,,F,FFF:FFFF,,,,:,F,,,:,,,:F,,F::F,F,,,F::F,FFF,,,,:,FFFFFF,FFFF::,:

Table 1 ASCII Characters Encoding Q-scores 0–40

| Symbol | ASCII Code | Q-Score | Symbol | ASCII Code | Q-Score | Symbol | ASCII Code | Q-Score |
|---|---|---|---|---|---|---|---|---|
| ! | 33 | 0 | / | 47 | 14 | = | 61 | 28 |
| " | 34 | 1 | 0 | 48 | 15 | > | 62 | 29 |
| # | 35 | 2 | 1 | 49 | 16 | ? | 63 | 30 |
| $ | 36 | 3 | 2 | 50 | 17 | @ | 64 | 31 |
| % | 37 | 4 | 3 | 51 | 18 | A | 65 | 32 |
| & | 38 | 5 | 4 | 52 | 19 | B | 66 | 33 |
| ' | 39 | 6 | 5 | 53 | 20 | C | 67 | 34 |
| ( | 40 | 7 | 6 | 54 | 21 | D | 68 | 35 |
| ) | 41 | 8 | 7 | 55 | 22 | E | 69 | 36 |
| * | 42 | 9 | 8 | 56 | 23 | F | 70 | 37 |
| + | 43 | 10 | 9 | 57 | 24 | G | 71 | 38 |
| , | 44 | 11 | : | 58 | 25 | H | 72 | 39 |
| - | 45 | 12 | ; | 59 | 26 | I | 73 | 40 |
| . | 46 | 13 | < | 60 | 27 | | | |

The quality score is associated to a probability of error, of an incorrect base call

**Q = -10log$_{10}$(e)** where **e** is the **estimated probability** of the base call being wrong.

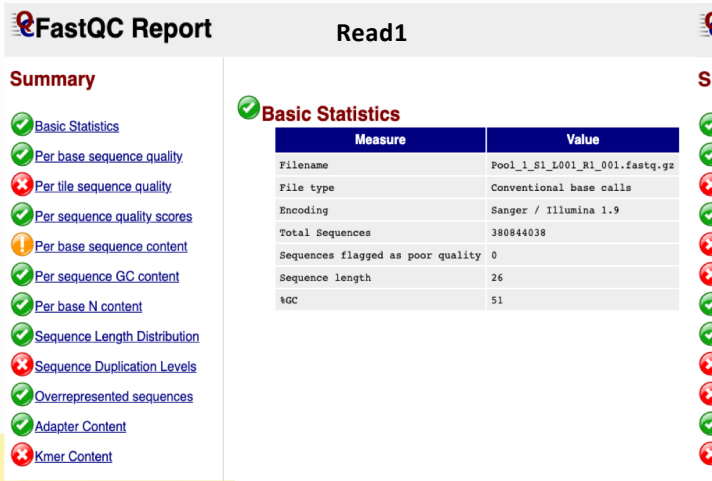
- **Higher Q scores** indicate a smaller probability of error.
- **Lower Q scores** can result in a significant portion of the reads being unusable.

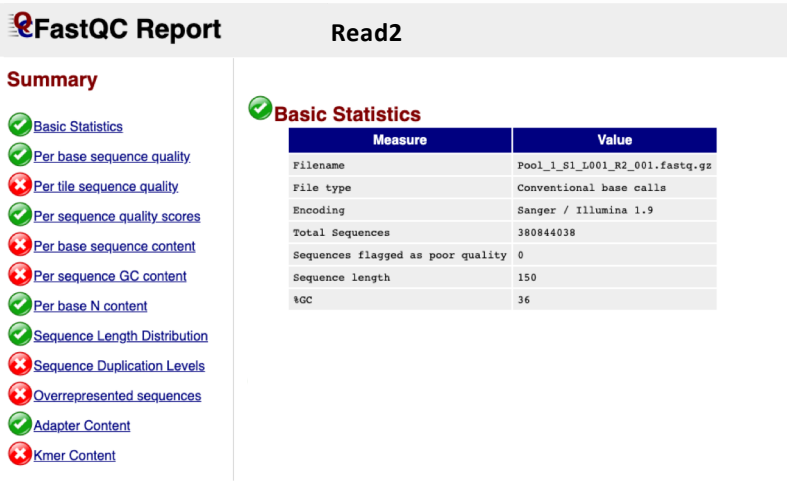
| Quality Score | Probability of Incorrect Base Call | Inferred Base Call Accuracy |
|---|---|---|
| 10 (Q10) | 1 in 10 | 90% |
| 20 (Q20) | 1 in 100 | 99% |
| 30 (Q30) | 1 in 1000 | 99.9% |

# FASTQC

Before analyzing the data to draw biological conclusions you should always perform some simple quality control

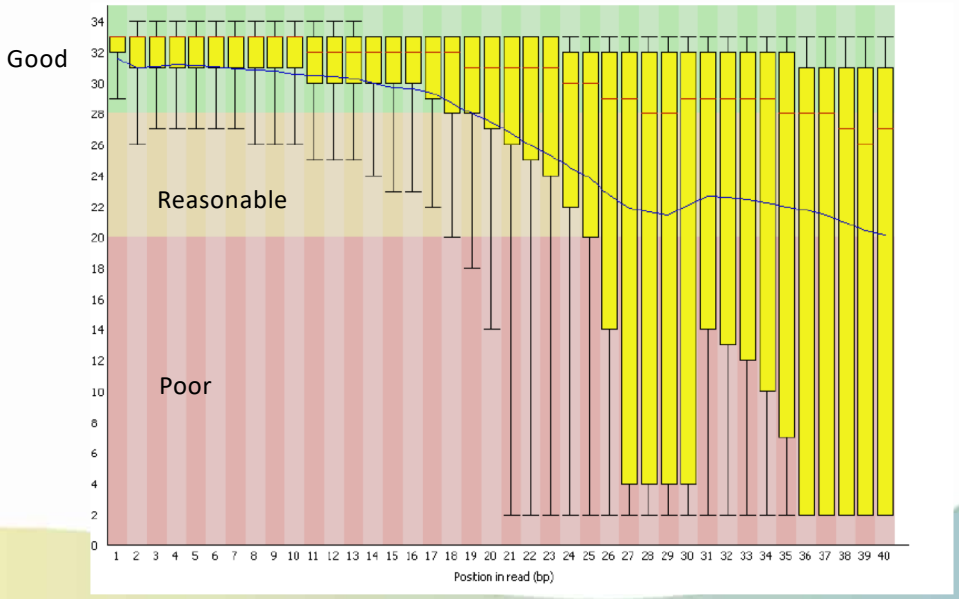
FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material



**Per Base Sequence Quality**

This view shows an overview of the range of quality values across all bases at each position in the FastQ file.

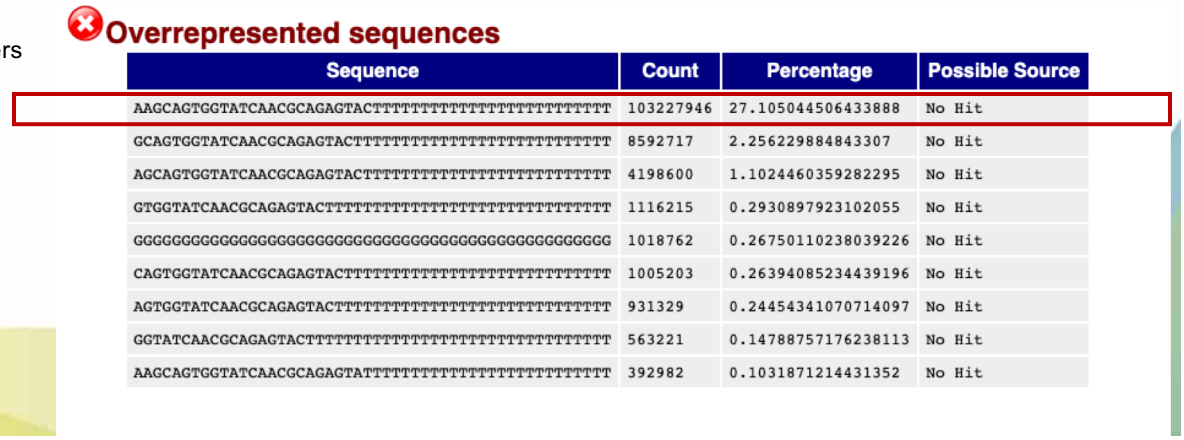


## Overrepresented Sequences

A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole.
This module lists all of the sequence which make up more than 0.1% of the total. Hits must be at least 20bp in length and have no more than 1 mismatch.
Finding a hit doesn't necessarily mean that this is the source of the contamination

**Warning**
This module will issue a warning if any sequence is found to represent more than 0.1% of the total.
**Failure**
This module will issue an error if any sequence is found to represent more than 1% of the total.

**Typical artifacts**
Primers, sequence adapters



**Overrepresented sequences**

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTT | 103227946 | 27.105044506433888 | No Hit |
| GCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTT | 8592717 | 2.256229884843307 | No Hit |
| AGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTT | 4198600 | 1.1024460359282295 | No Hit |
| GTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTT | 1116215 | 0.2930897923102055 | No Hit |
| GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG | 1018762 | 0.26750110238039226 | No Hit |
| CAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTT | 1005203 | 0.26394085234439196 | No Hit |
| AGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTT | 931329 | 0.24454341070714097 | No Hit |
| GGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT | 563221 | 0.14788757176238113 | No Hit |
| AAGCAGTGGTATCAACGCAGAGTATTTTTTTTTTTTTTTTTTTTTTTT | 392982 | 0.1031871214431352 | No Hit |


Current version: v1.9
Home
Docs
Plugins
Logo
Example Reports
MultiQC
Citations
858
Aggregate results from bioinformatics analyses across many samples into a single report
MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools.
Introduction to MultiQC
Phil Ewels
phil.ewels@scilifelab.se
Introduction to MultiQC (1:19)
Installing MultiQC (4:33)
Running MultiQC (5:21)
Using MultiQC Reports (6:06)
GitHub
Python Package Index
Documentation
92 supported tools
Publication / Citation
Get help on Gitter
Quick Install
pip install multiqc # Install
multiqc . # Run
pip
conda
manual
Need a little more help? See the full installation instructions.

MultiQC
v1.0.dev0



# General Statistics

Copy table  Showing $^{29}/_{29}$ rows and $^{3}/_{5}$ columns.

| Sample Name | % Dups | % GC | M Seqs |
|---|---|---|---|
| ARH10_S10_L001_R1_001 | 69.0% | 54% | 53.5 |
| ARH10_S10_L001_R2_001 | 73.0% | 53% | 53.5 |
| ARH11_S11_L001_R1_001 | 72.2% | 50% | 4.8 |
| ARH11_S11_L001_R2_001 | 73.0% | 50% | 4.8 |
| ARH12_S12_L001_R1_001 | 65.9% | 57% | 26.9 |
| ARH12_S12_L001_R2_001 | 69.7% | 56% | 26.9 |
| ARH13_S13_L001_R1_001 | 50.7% | 51% | 27.1 |
| ARH13_S13_L001_R2_001 | 56.6% | 51% | 27.1 |
| ARH14_S14_L001_R1_001 | 65.5% | 51% | 33.2 |
| ARH14_S14_L001_R2_001 | 68.8% | 51% | 33.2 |
| ARH15_S15_L001_R1_001 | 60.4% | 47% | 22.7 |
| ARH15_S15_L001_R2_001 | 64.2% | 47% | 22.7 |
| ARH16_S16_L001_R1_001 | 65.5% | 47% | 26.0 |
| ARH16_S16_L001_R2_001 | 68.9% | 47% | 26.0 |
| ARH17_S17_L001_R1_001 | 58.8% | 47% | 10.0 |
| ARH17_S17_L001_R2_001 | 61.3% | 46% | 10.0 |
| ARH18_S18_L001_R1_001 | 48.2% | 50% | 28.1 |

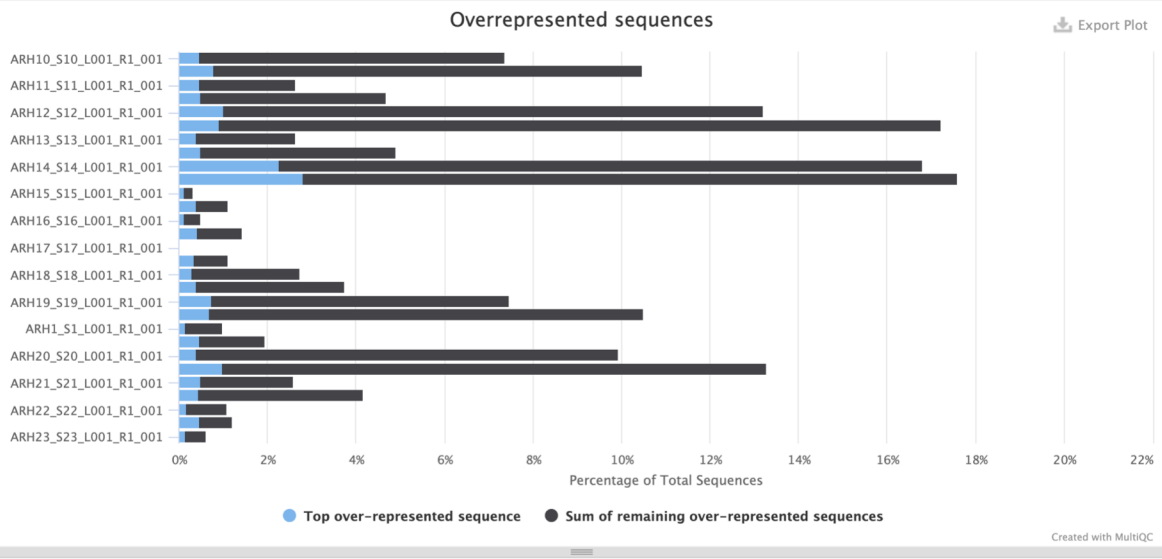
MultiQC
v1.0.dev0



## Overrepresented sequences

1 25 3

The total amount of overrepresented sequences found in each library. See the FastQC help for further information.



**fastp**
**https://github.com/OpenGene/fastp**

# fastp

A tool designed to provide fast all-in-one preprocessing for FastQ files. This tool is developed in C++ with multithreading supported to afford high performance.

## fastp

**UMI example**
The original read:
@NS500713:64:HFKJJBGXY:1:11101:1675:1101 1:N:0:TATAGCCT+GACCCCCA
AAAAAAAAGCTACTTGGAGTACCAATAATAAAGTGAGCCCACCTTCCTGGTACCCAGACATTTCAGGAGGTCGGGA
AA + 6AAAAAEEEEE/E/EA/E/AEA6EE//AEE66/AAE//EEE/E//E/AA/EEE/A/AEE/EEA//EEEEEEEE6EEAA

After it's processed with command:
fastp -i R1.fq -o out.R1.fq -U --umi_loc=read1 --umi_len=8:

@NS500713:64:HFKJJBGXY:1:11101:1675:1101:AAAAAAAA 1:N:0:TATAGCCT+GACCCCCA
GCTACTTGGAGTACCAATAATAAAGTGAGCCCACCTTCCTGGTACCCAGACATTTCAGGAGGTCGGGAAA +
EEE/E/EA/E/AEA6EE//AEE66/AAE//EEE/E//E/AA/EEE/A/AEE/EEA//EEEEEEEE6EEAA

**Sequence filtering**

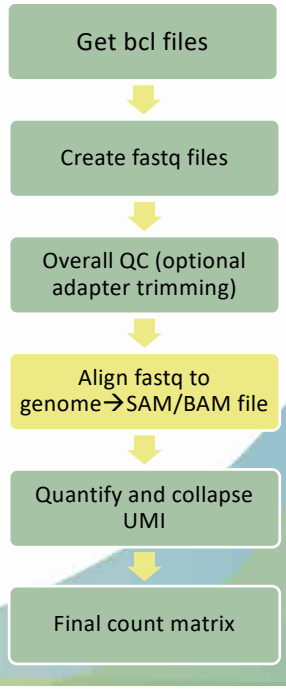
Adapter trimming
Quality trimming
Trimming fixed length

**Tools**

Cutadapt
SeqTK
Trimmomatic
fastp

## Aligning reads to a reference

FASTQ files contain sequence information that we wish to map to genes in a genome:
1.Select your genome
2.Select your gene annotation file (generally gtf format)
3.Run the alignment program
4.Result of alignment is generally stored in a sam/bam file




## Aligners

**STAR**, Kallisto, Bowtie

## SAM Format

This is the most basic, human readable format, generated by almost every alignment algorithm that exists. It consists of a header, a row for every read in your dataset, and 11 tab-delimited fields describing that read.

**SAM Header**

The full list of available header fields can be found below

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | `[!-?A-~]{1,254}` | Query template NAME |
| 2 | FLAG | Int | $[0,2^{16}-1]$ | bitwise FLAG |
| 3 | RNAME | String | `\*|[!-()+-<>-~][!-~]*` | Reference sequence NAME |
| 4 | POS | Int | $[0,2^{31}-1]$ | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | $[0,2^{8}-1]$ | MAPping Quality |
| 6 | CIGAR | String | `\*|([0-9]+[MIDNSHPX=])+` | CIGAR string |
| 7 | RNEXT | String | `\*|=|[!-()+-<>-~][!-~]*` | Ref. name of the mate/next read |
| 8 | PNEXT | Int | $[0,2^{31}-1]$ | Position of the mate/next read |
| 9 | TLEN | Int | $[-2^{31}+1,2^{31}-1]$ | observed Template LENgth |
| 10 | SEQ | String | `\*|[A-Za-z=.]+` | segment SEQuence |
| 11 | QUAL | String | `[!-~]+` | ASCII of Phred-scaled base QUALity+33 |

http://samtools.github.io/hts-specs/SAMv1.pdf

**MapQ (Mapping Quality)**

This value reports how well the read aligned to the reference. Different algorithms report it differently but nonetheless, the greater the number the better the alignment (generally).

**CIGAR String**

This is a shorthand way to encode an entire alignment:

| Op | Description |
|---|---|
| M | Match (alignment column containing two letters). This could contain two different letters (mismatch) or two identical letters. USEARCH generates CIGAR strings containing Ms rather than X's and ='s (see below). |
| D | Deletion (gap in the target sequence). |
| I | Insertion (gap in the query sequence). |
| S | Segment of the query sequence that does not appear in the alignment. This is used with soft clipping, where the full-length query sequence is given (field 10 in the SAM record). In this case, S operations specify segments at the start and/or end of the query that do not appear in a local alignment. |
| H | Segment of the query sequence that does not appear in the alignment. This is used with hard clipping, where only the aligned segment of the query sequences is given (field 10 in the SAM record). In this case, H operations specify segments at the start and/or end of the query that do not appear in the SAM record. |
| = | Alignment column containing two identical letters. USEARCH can read CIGAR strings using this operation, but does not generate them. |
| X | Alignment column containing a mismatch, i.e. two different letters. USEARCH can read CIGAR strings using this operation, but does not generate them. |

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

```
POS: 5
CIGAR: 3M1I3M1D5M
```

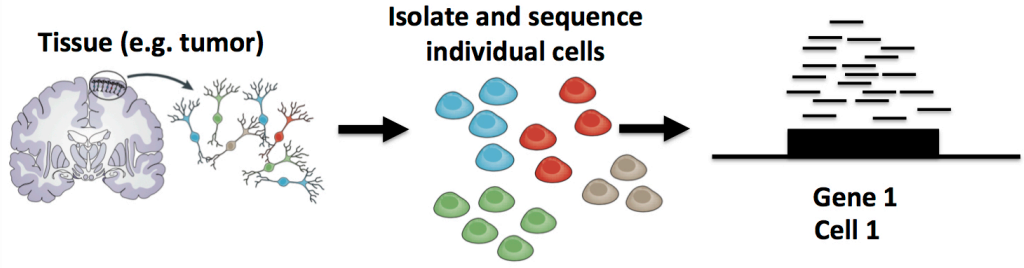
## BAM format

This is the same format except that it encoded in binary which means that it is significantly smaller than the SAM files and significantly faster to read, though it is not human legible and needs to be converted to another format (i.e. SAM) in order to make sense to us.
Some special tools are needed in order to make sense of BAM, such as Samtools, Picard Tools,

**View BAM Files**

samtools view alignment.bam | head

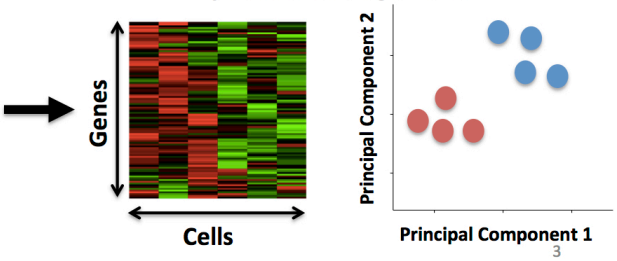
# Single-cell RNA-Seq (scRNA-Seq)




**Read Counts**

| | Cell 1 | Cell 2 | ... |
|---|---|---|---|
| Gene 1 | 18 | 0 | |
| Gene 2 | 1010 | 506 | |
| Gene 3 | 0 | 49 | |
| Gene 4 | 22 | 0 | |
| ... | | | |

**Compare gene expression profiles of single cells**



Differences between the methods are in how they capture a cell and quantify gene expression (either full-length or tag-based).

- Full-length capture tries to achieve a uniform coverage of each transcript (many reads per transcript).
- Tag-based protocols only capture either the 5'- or 3'-end of each transcript (single read per transcript). Choice in method determines what types of analyses the data can be used for.
- Full-length capture can be used to distinguish different **iso-forms**, where tag-based method is best used for only **gene abundance**.

**•Tag-based 3' counting techniques**

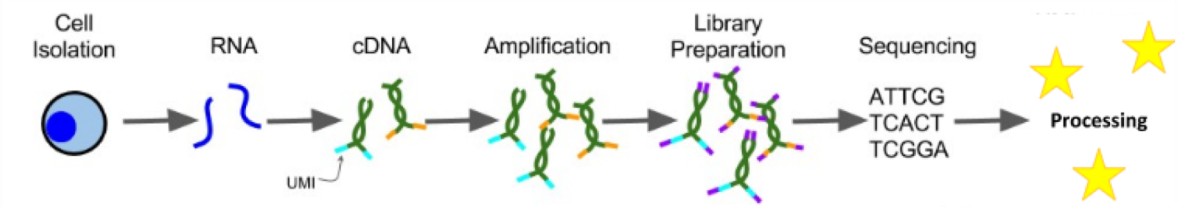
- 1 read per transcript
- Based on polyA
- Expression analysis only
- Fewer reads per cell needed (~20K reads/cell 10x V3+)
- Less noise in expression patterns

**•Full-length**

- Based on polyA
- Expression analysis
- Splicing information
- The more information desired beyond expression, the higher the reads needed per cell (~50K reads/cell to 10M reads/cell)

## Challenges in single cell data analysis

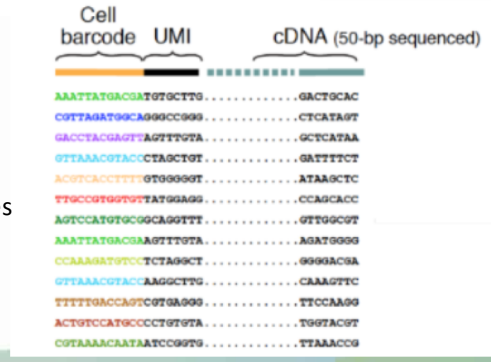
Amplification artifacts
Dropouts




With high number of PCR cycles, some transcript become over represented in the final library

UMI are unique molecule identifier added to the transcript during RT step

UMI enable sequencing reads to be assigned to individual transcript molecules and removal of amplification noise and biases



## 10X technology and Cell Ranger

**Cell Ranger is a set of analysis pipelines that process Chromium single-cell RNA-seq output** to align reads, generate gene-barcode matrices and perform clustering and gene expression analysis. Cell Ranger includes five pipelines relevant to single-cell gene expression experiments:

**cellranger mkfastq** demultiplexes raw base call (BCL) files generated by Illumina sequencers into FASTQ files. It is a wrapper around Illumina's bcl2fastq, with additional features that are specific to 10x libraries and a simplified sample sheet format.

**cellranger count** takes FASTQ files from cellranger mkfastq and performs alignment, filtering, barcode counting, and UMI counting. It uses the Chromium cellular barcodes to generate feature-barcode matrices, determine clusters, and perform gene expression analysis. The count pipeline can take input from multiple sequencing runs on the same GEM well. It can also processes Feature Barcode data alongside Gene Expression reads.

**cellranger aggr** aggregates outputs from multiple runs of cellranger count, normalizing those runs to the same sequencing depth and then recomputing the feature-barcode matrices and analysis on the combined data.

**cellranger reanalyze** reruns the dimensionality reduction, clustering, and gene expression algorithms using tunable parameter settings.

**cellranger multi** is used to analyze Cell Multiplexing data. The cellranger multi pipeline also supports the analysis of Feature Barcode data.

**Primary analysis pipelines**

Cell Ranger has three analysis pipelines for primary analysis

- cellranger count
- cellranger vdj
- cellranger multi
- Cellranger arc  Chromium Single Cell Multiome ATAC + Gene Expression sequencing

| Documentation by library type | Cell Ranger pipeline |
|---|---|
| 3' Gene Expression | count |
| 3' Gene Expression + Antibody/CRISPR Guide Capture | count |
| Antibody Capture only | count |
| 3' Gene Expression + Cell Multiplexing (+ Antibody/CRISPR Guide Capture) | multi |
| Fixed RNA Profiling Gene Expression (+ Antibody/CRISPR Guide Capture) | multi |
| 5' V(D)J only | vdj |
| 5' Gene Expression only | count |
| 5' Antibody Capture (+ Gene Expression) | count |
| 5' CRISPR (FB) + Gene Expression | count |
| 5' Gene Expression + V(D)J (+ FB) | multi |
| 5' Gene Expression + V(D)J + Antigen Capture (BEAM) (+ Antibody Capture) | multi |

**How to run the 10X cellranger software?**

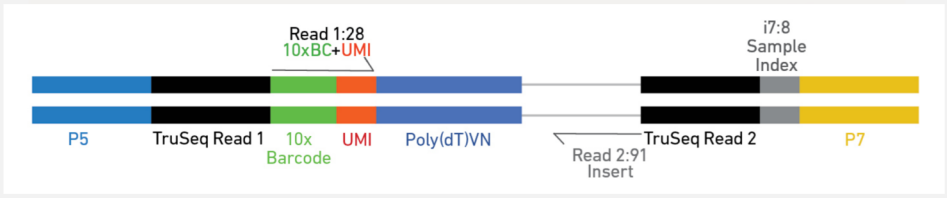
**Pipeline download and installation**

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

## Cell Ranger mkfastq

```
cellranger mkfastq --id=my_data --run=/path/to/bcl  --csv=samplesheet.csv
```

Cellranger mkfastq produces :
Index.fastq; R1.fastq; R2.fastq




index

```
AFCHKCCJALXX/HKCCJALXX/906_05_2776_$_zcat_-f_906_05_S1_L008_I1_001.fastq.gz | head
@E00342:426:HKCCJALXX:8:1101:1661:1450 1:N:0:NAACCTCA
NAACCTCA
+
#AAAFA-F
```

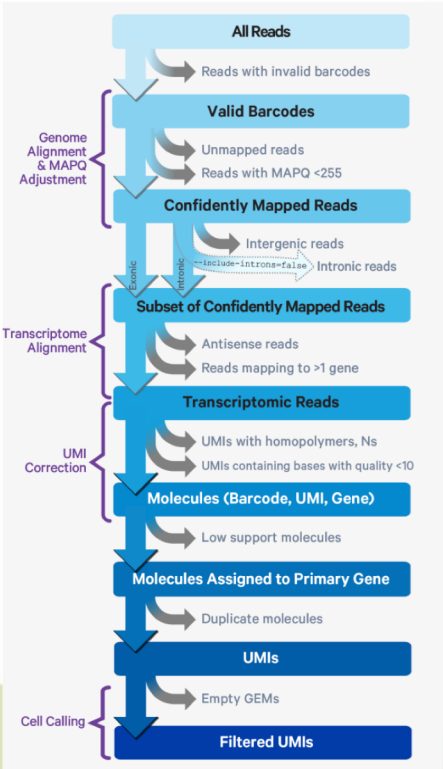
Read1

```
AFCHKCCJALXX/HKCCJALXX/906_05_2777_$_zcat_-f_906_05_S1_L008_R1_001.fastq.gz | head
@E00342:426:HKCCJALXX:8:1101:1661:1450 1:N:0:NAACCTCA
GNTCAGTTCAACGGGAGCAACCCAGT
+
A#AFFFFFFF<FJJJ<AJFJJFF7<J
```

Read2

```
AFCHKCCJALXX/HKCCJALXX/906_05_2779_$_zcat_-f_906_05_S1_L008_R2_001.fastq.gz | head
@E00342:426:HKCCJALXX:8:1101:1661:1450 2:N:0:NAACCTCA
NCGCGGGGCTGAATTGTGCTCTGATCTTTGGGTGAAAGATAAGTCACAATTTAAAAGCGAATTTCTTCAGGAAAAGACAGCAAAAATACACAAACTAA
+
#--A7-77FA<A7FFAA7-----<A-AF--7--77AF-AA-<---77-<<7AAFF-7A--A-7----A-7--7AA--7---<7-A7--7-------77
```

## 10X Cell Ranger count pipeline



## Gene Expression Algorithms Overview

**Read trimming:** Cellranger only performs read trimming to 3' gene expression assays.
A full length cDNA construct is flanked by the 30 bp template switch oligo (TSO) sequence on the 5' end and poly-A on the 3' end.

Reads derived from short RNA molecules are more likely to contain either or both TSO and poly-A sequence than longer RNA molecules.

In order to improve mapping, the TSO sequence is trimmed from the 5' end of read 2 and poly-A is trimmed from the 3' end prior to alignment.

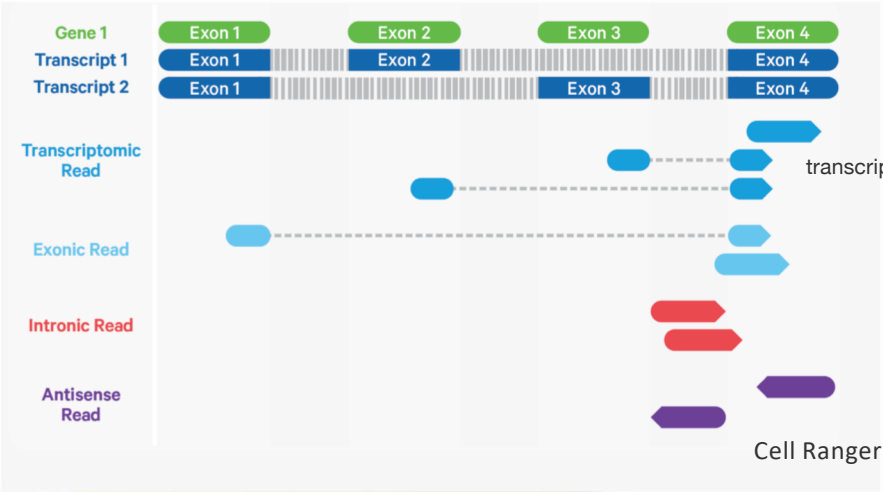
**Genome Alignment**
Cell Ranger uses an aligner called STAR, which peforms splicing-aware alignment of reads to the genome. Cell Ranger then uses the transcript annotation GTF to categorize the reads into **exonic**, **intronic**, and **intergenic**, and by whether the reads align (confidently) to the genome. A read is exonic if at least 50% of it intersects an exon, intronic if it is non-exonic and intersects an intron, and intergenic otherwise.

**MAPQ adjustment**
For reads that align to a single exonic locus but also align to 1 or more non-exonic loci, the exonic locus is prioritized and the read is considered to be confidently mapped to the exonic locus with MAPQ 255.

.

**Transcriptome Alignment**

Cell Ranger further aligns confidently mapped exonic and intronic reads to annotated transcripts by examining their compatibility with the transcriptome. As shown below, reads are classified based on whether they are sense or antisense and whether they are exonic, intronic, or have a splicing pattern compatible with transcript annotations associated with that gene. Cell Ranger prefers alignments with sense over antisense reads. A read is classified as antisense (purple) if it has any alignments to a transcript exon on the opposite strand and no sense alignments




transcriptomic reads are carried forward to UMI counting.

Cell Ranger ignores antisense reads

Starting in **Cell Ranger 7.0**, by default, the cellranger count and cellranger multi pipelines **will include intronic reads** for whole transcriptome gene expression analysis. Any reads that map in the sense orientation to a single gene - the reads labeled transcriptomic (blue) - are carried forward to UMI counting.

The default setting that includes intronic reads is recommended to maximize sensitivity, for example in cases where the input to the assay consists of nuclei, as there may be high levels of intronic reads generated by unspliced transcripts.

The fraction of gDNA was a small percentage of total UMIs (0.08%) and was a consistent 7-8% of features regardless of transcript length for the PBMC dataset (10X)

## Gene Expression Algorithms Overview

### Cell barcode and UMI filtering

- **Cell barcodes**
  - Must be on static list of known cell barcode sequences
  - May be 1 mismatch away from the list if the mismatch occurs at a low- quality position (the barcode is then corrected).
- **UMIs (Unique Molecular Index)**
  - Must not be a homopolymer, e.g. AAAAAAAAAA
  - Must not contain N
  - Must not contain bases with base quality < 10
  - UMIs that are 1 mismatch away from a higher-count UMI are corrected to that UMI if they share a cell barcode and gene.

## Filtering cells (Cell Ranger)

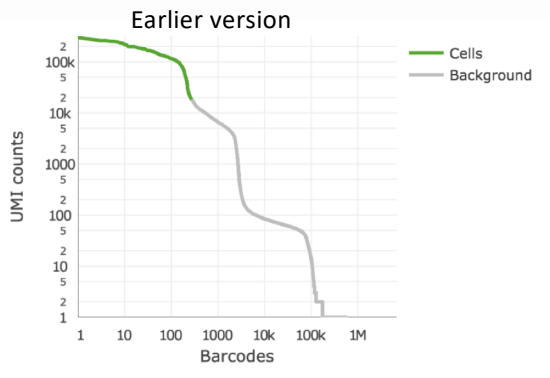
**Cellranger 3.0** introduces and improved cell-calling algorithm to identify populations of low RNA content cells, especially when low RNA content cells are mixed into a population of high RNA content cells.
E.g tumor samples often contain large tumor cells mixed with smaller tumor infiltrating lymphocytes (TIL).
The new algorithm is based on the **EmptyDrops** method (Lun et al., 2018).

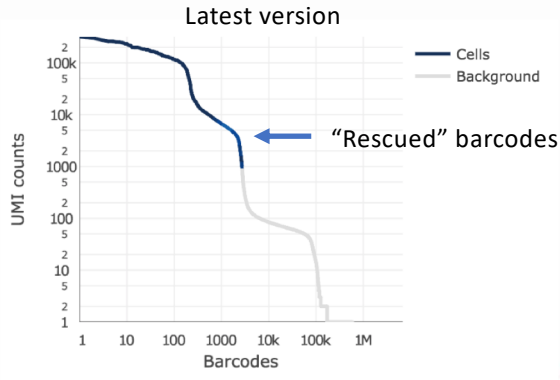
The algorithm has two key steps:
1.It uses a cutoff based on total UMI counts of each barcode to identify cells. This step identifies the primary mode of high RNA content cells.

2.Then the algorithm uses the RNA profile of each remaining barcode to determine if it is an “empty” or a cell containing partition. This second step captures low RNA content cells whose total UMI counts may be similar to empty GEMs.

**Barcodes selection steps**

1. The original cellranger cell calling algorithm is used to identify high RNA content cells, using a cutoff based on the total UMI count for each barcode.
2. In the second step, a set of barcodes with low UMI counts that likely represent ‘empty’ GEM partitions is selected. A model of the RNA profile of selected barcodes is created. This second step identifies cells that are clearly distinguishable from the profile of empty GEMs, even though they may have much lower RNA content than the largest cells in the experiment.



## Alternative pipelines

10X data or other scRNAseq technologies can also rely on independent pipelines

STARsolo

Alevin

dropEst

Salmon 1.10.2 documentation

Search




# Alevin

Alevin is a tool — integrated with the salmon software — that introduces a family of algorithms for quantification and analysis of 3′ tagged-end single-cell sequencing data. Currently alevin supports the following single-cell protocols:

1. Drop-seq
2. 10x-Chromium v1/2/3
3. inDropV2
4. CELSeq 1/2
5. Quartz-Seq2
6. sci-RNA-seq3

Alevin works under the same indexing scheme (as salmon) for the reference, and consumes the set of FASTA/Q files(s) containing the Cellular Barcode(CB) + Unique Molecule identifier (UMI) in one read file and the read sequence in the other. Given just the transcriptome and the raw read files, alevin generates a cell-by-gene count matrix (in a fraction of the time compared to other tools).

Alevin works in two phases. In the first phase it quickly parses the read file containing the CB and UMI information to generate the frequency distribution of all the observed CBs, and creates a lightweight data-structure for fast-look up and correction of the CB. In the second round, alevin utilizes the read-sequences contained in the files to map the reads to the transcriptome, identify potential PCR/sequencing errors in the UMIs, and performs hybrid de-duplication while accounting for UMI collisions. Finally, a post-abundance estimation CB whitelisting procedure is done and a cell-by-gene count matrix is generated.

ON THIS PAGE

## How to run cell ranger count
## Command-Line Argument Reference

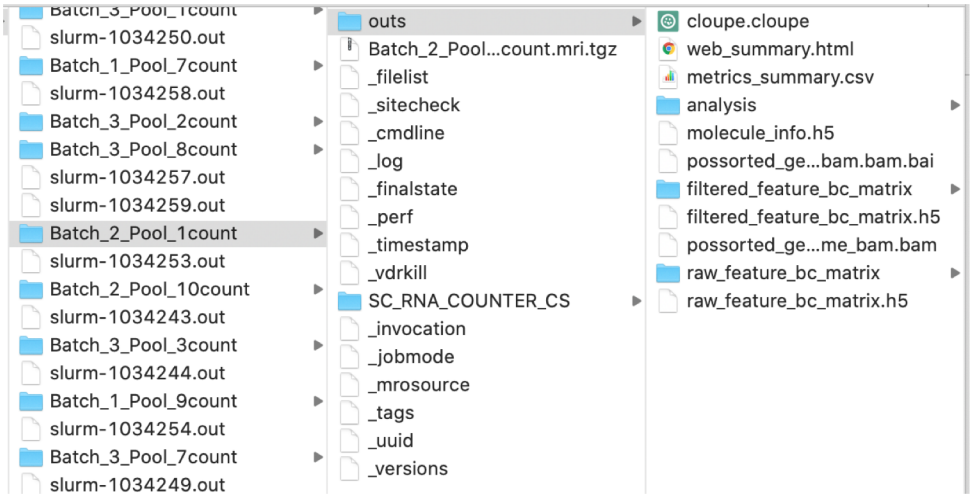
| Argument | Description |
|---|---|
| **--id** | A unique run ID string: e.g. sample345 |
| **--fastqs** | Path of the fastq_path folder generated by cellranger mkfastq |
| **--libraries** | Path to a libraries.csv file declaring FASTQ paths and library types of input libraries. Required for feature-barcoding analysis. See Feature Barcoding Analysis page for details. When using this argument, --fastqs and --sample must not be passed. |
| **--sample** | Sample name as specified in the sample sheet supplied to cellranger mkfastq. |
| **--transcriptome** | Path to the Cell Ranger compatible transcriptome reference |
| **--feature-ref** | Path to a Feature Reference CSV file declaring the Feature Barcoding reagents in use in the experiment. |
| **--expect-cells** | (optional) Expected number of recovered cells. |
| **--force-cells** | (optional) Force pipeline to use this number of cells, bypassing the cell detection algorithm. |
| **--chemistry** | (optional) |
| **--r1-length** | (optional) |
| **--r2-length** | (optional) Hard-trim the input R2 sequence to this length. |
| **--include-introns** | (optional) Add this flag to count reads mapping to intronic regions. This may improve sensitivity for samples with a significant amount of pre-mRNA molecules, such as nuclei. This flag should be used instead of the deprecated pre-mRNA reference |

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

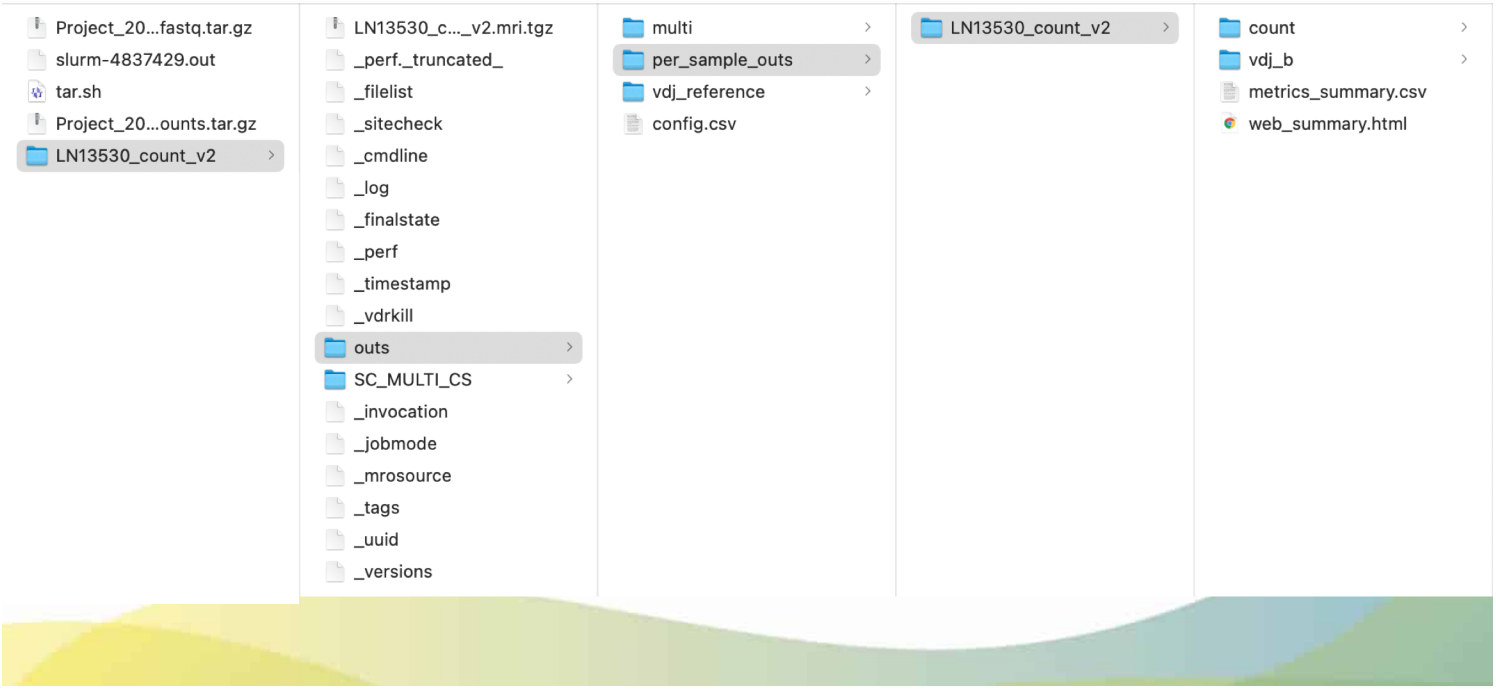
## How to run cellranger count

```
path_to_cellranger/cellranger count --id=samplename(your choice) --transcriptome=path_to/refdata-gex-GRCh38-2020-A
 --fastqs=path_to_flowcell/outs/fastq_path  --sample=sample_prefix
```

### output



path_to_cellranger/cellranger-6.0.1/cellranger multi --id=samplename(your choice)  --csv=multi-config.csv



## Matrix output

With 3 files needed to completely describe each gene x cell matrix
•matrix.mtx.gz
•features.tsv.gz
•barcode.tsv.gz

| Type | Description |
|---|---|
| Raw | gene-barcode matrices Contains every barcode from fixed list of known-good barcode sequences. This includes background and non-cellular barcodes. |
| Filtered | gene-barcode matrices Contains only detected cellular barcodes. |

## Web summary examples

### LN13530_GEX

Summary | Analysis

**11,597**
Estimated Number of Cells

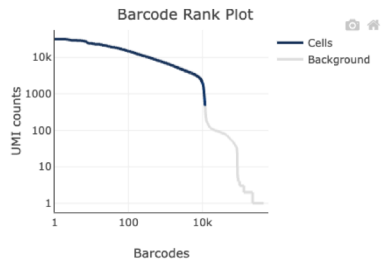
**32,179**
Mean Reads per Cell

**1,665**
Median Genes per Cell

#### Sequencing

| | |
|---|---|
| Number of Reads | 373,184,921 |
| Number of Short Reads Skipped | 0 |
| Valid Barcodes | 93.4% |
| Valid UMIs | 99.8% |
| Sequencing Saturation | 77.4% |
| Q30 Bases in Barcode | 95.5% |
| Q30 Bases in RNA Read | 86.7% |
| Q30 Bases in UMI | 94.7% |

#### Cells




| | |
|---|---|
| Estimated Number of Cells | 11,597 |
| Fraction Reads in Cells | 92.1% |
| Mean Reads per Cell | 32,179 |
| Median Genes per Cell | 1,665 |
| Total Genes Detected | 20,470 |
| Median UMI Counts per Cell | 3,502 |

## Mapping

| | |
|---|---|
| Reads Mapped to Genome | 91.6% |
| Reads Mapped Confidently to Genome | 87.0% |
| Reads Mapped Confidently to Intergenic Regions | 4.8% |
| Reads Mapped Confidently to Intronic Regions | 10.1% |
| Reads Mapped Confidently to Exonic Regions | 72.1% |
| Reads Mapped Confidently to Transcriptome | 64.3% |
| Reads Mapped Antisense to Gene | 4.5% |

## Sample

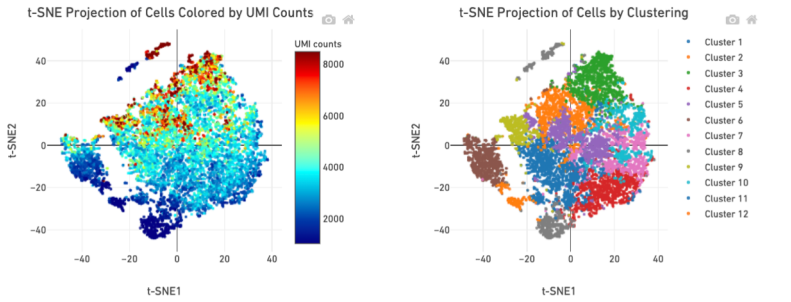
| | |
|---|---|
| Sample ID | LN13530_GEX |
| Sample Description | |
| Chemistry | Single Cell 5' R2-only |
| Include introns | False |
| Reference Path | …ts/Roberta/refdata-gex-GRCh38-2020-A |
| Transcriptome | GRCh38-2020-A |
| Pipeline Version | cellranger-6.0.1 |

# LN13530_GEX

Summary | Analysis

## t-SNE Projection

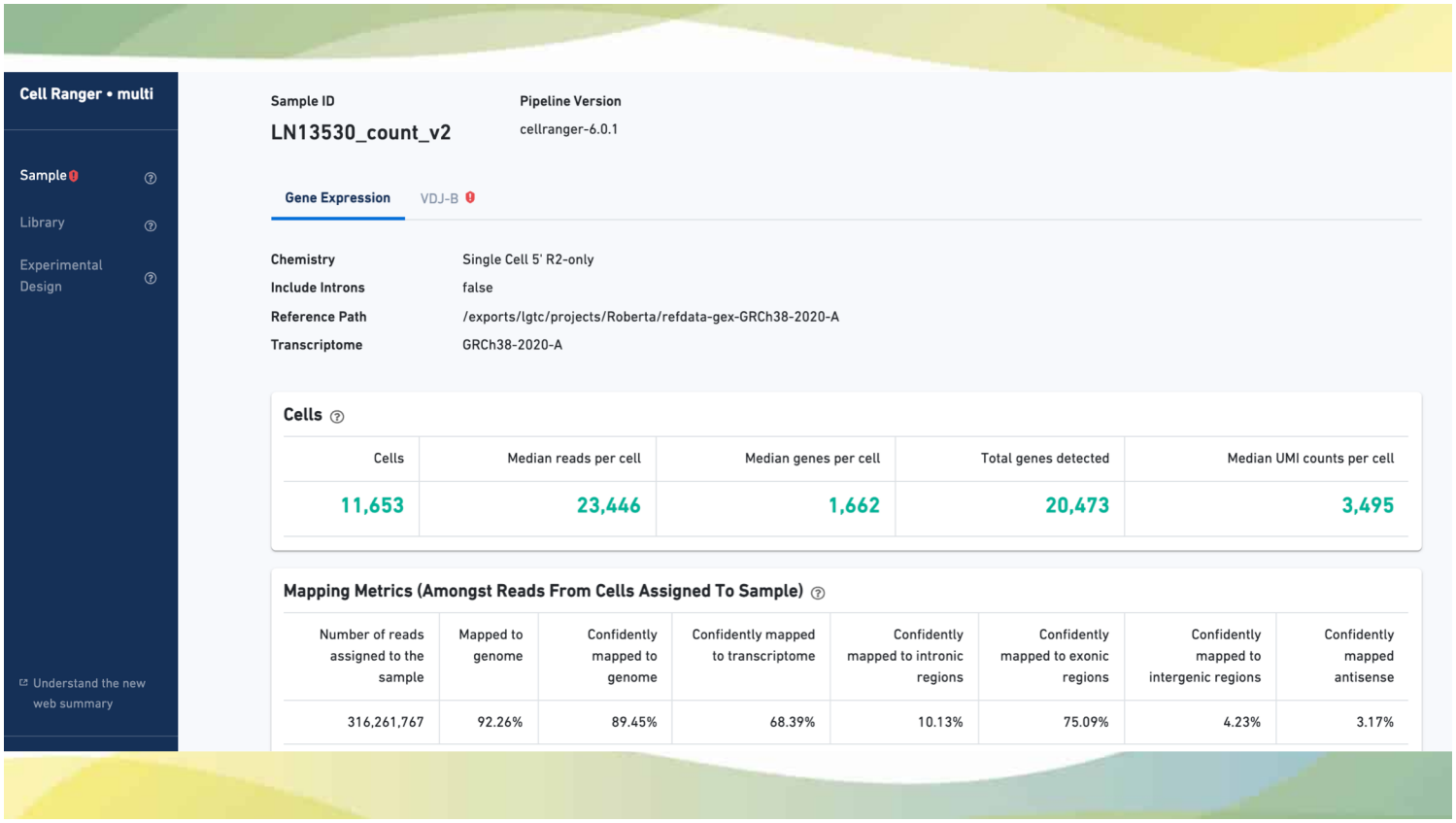
Clustering Type: Graph-based




## Top Features by Cluster (Log2 fold-change, p-value)

| Feature | | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | | Cluster 5 | | Cluster 6 | | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Name | L2FC | p-value | L2FC | p-value | L2FC | p-value | L2FC | p-value | L2FC | p-value | L2FC | p-value | L2F |
| ENSG00000111679 | PTPN6 | 1.43 | 2e-4 | 0.16 | 1e+0 | -0.93 | 9e-2 | 0.02 | 1e+0 | -0.90 | 8e-1 | 0.15 | 1e+0 | -( |
| ENSG00000158481 | CD1C | 1.19 | 7e-3 | 0.44 | 1e+0 | -1.61 | 1e-4 | -0.45 | 1e+0 | -0.03 | 1e+0 | -0.13 | 1e+0 | -( |
| ENSG00000104903 | LYL1 | 1.18 | 9e-3 | 0.20 | 1e+0 | -0.54 | 7e-1 | -0.25 | 1e+0 | -0.05 | 1e+0 | -0.14 | 1e+0 | -( |
| ENSG00000216490 | IFI30 | 1.16 | 1e-2 | -0.00 | 1e+0 | -0.44 | 9e-1 | -0.79 | 1e+0 | -0.47 | 1e+0 | -0.12 | 1e+0 | -1 |
| ENSG00000197043 | ANXA6 | 1.12 | 2e-2 | 0.18 | 1e+0 | -1.53 | 6e-4 | 0.12 | 1e+0 | -0.10 | 1e+0 | 0.53 | 5e-1 | -( |
| ENSG00000142227 | EMP3 | 0.96 | 1e-1 | 0.36 | 1e+0 | -1.22 | 1e-2 | 0.38 | 1e+0 | -0.33 | 1e+0 | -0.19 | 1e+0 | ( |
| ENSG00000137101 | CD72 | 0.90 | 2e-1 | 0.25 | 1e+0 | -0.71 | 4e-1 | -0.50 | 1e+0 | -0.11 | 1e+0 | 0.11 | 1e+0 | -( |
| ENSG00000026297 | RNASET2 | 0.87 | 2e-1 | 0.29 | 1e+0 | -0.98 | 6e-2 | 0.38 | 1e+0 | -0.19 | 1e+0 | 0.12 | 1e+0 | ( |
| ENSG00000162144 | CYB561A3 | 0.73 | 6e-1 | 0.16 | 1e+0 | -0.68 | 4e-1 | -0.22 | 1e+0 | -0.03 | 1e+0 | 0.36 | 8e-1 | -( |
| ENSG00000173457 | PPP1R14B | 0.72 | 7e-1 | 0.04 | 1e+0 | -0.18 | 1e+0 | -0.12 | 1e+0 | 0.12 | 1e+0 | -0.59 | 8e-1 | -1 |

Previous | Page 1 of 45 | 10 rows | Next


Cell Ranger • multi
Sample
Library
Experimental Design
Understand the new web summary
Sample ID
LN13530_count_v2
Pipeline Version
cellranger-6.0.1
Gene Expression
VDJ-B
Chemistry
Single Cell 5' R2-only
Include Introns
false
Reference Path
/exports/lgtc/projects/Roberta/refdata-gex-GRCh38-2020-A
Transcriptome
GRCh38-2020-A
Cells
Cells
Median reads per cell
Median genes per cell
Total genes detected
Median UMI counts per cell
11,653
23,446
1,662
20,473
3,495
Mapping Metrics (Amongst Reads From Cells Assigned To Sample)
Number of reads assigned to the sample
Mapped to genome
Confidently mapped to genome
Confidently mapped to transcriptome
Confidently mapped to intronic regions
Confidently mapped to exonic regions
Confidently mapped to intergenic regions
Confidently mapped antisense
316,261,767
92.26%
89.45%
68.39%
10.13%
75.09%
4.23%
3.17%

| | |
|---|---|
| **Chemistry** | Single Cell 5' R2-only |
| **Include Introns** | false |
| **Reference Path** | /exports/lgtc/projects/Roberta/refdata-gex-GRCh38-2020-A |
| **Transcriptome** | GRCh38-2020-A |

## Cell Statistics

| Physical library ID | Estimated number of cells | Mean reads per cell |
|---|---|---|
| GEX_1 | 11,653 | 32,025 |

## Sequencing Metrics

| Fastq ID | Number of reads | Number of short reads skipped | Q30 barcodes | Q30 UMI | Q30 RNA read |
|---|---|---|---|---|---|
| LN13530_GEX | 373,184,921 | 0 | 95.5% | 94.7% | 86.7% |

## Mapping Metrics (Amongst All Reads in Library)

| Physical library ID | Number of reads in the library | Mapped to genome | Confidently mapped to genome | Confidently mapped to transcriptome | Confidently mapped to intronic regions | Confidently mapped to exonic regions | Confidently mapped to intergenic regions | Confidently mapped antisense |
|---|---|---|---|---|---|---|---|---|
| GEX_1 | 373,184,921 | 91.61% | 87.01% | 64.28% | 10.07% | 72.11% | 4.83% | 4.47% |

## Metrics Per Physical Library

| Physical library ID | Number of reads | Valid barcodes | Valid UMIs | Sequencing saturation | Fraction reads in cells | Mean reads per cell |
|---|---|---|---|---|---|---|
| GEX_1 | 373,184,921 | 93.39% | 99.78% | 77.40% | 92.32% | 32,025 |

| Metric | Description |
|---|---|
| **Estimated Number of Cells** | The number of barcodes associated with cell-containing partitions |
| **Mean Reads per Cell** | The total number of sequenced reads divided by the estimated number of cells. |
| **Median Genes per Cell** | The median number of genes detected (with nonzero UMI counts) across all cell-associated barcodes. |
| **Number of Reads** | Total number of sequenced reads. |
| **Valid Barcodes** | Fraction of reads with cell-barcodes that match the whitelist. |
| **Reads Mapped to Genome** | Fraction of reads that mapped to the genome. |
| **Reads Mapped Confidently to Genome** | Reads Mapped Confidently to Genome. |
| **Reads Mapped Confidently to Transcriptome** | Fraction of reads that mapped to a unique gene in the transcriptome with a high mapping quality score |
| **Reads Mapped Confidently to Exonic Regions** | Fraction of reads that mapped to the exonic regions of the genome with a high mapping quality score |
| **Reads Mapped Confidently to Intronic Regions** | Fraction of reads that mapped to the intronic regions of the genome with a high mapping quality score |
| **Reads Mapped Confidently to Intergenic Regions** | Fraction of reads that mapped to the intergenic regions of the genome with a high mapping quality score |
| **Reads Mapped Antisense to Gene** | Fraction of reads confidently mapped to the transcriptome, but on the opposite strand of their annotated gene. |
| **Sequencing Saturation** | The fraction of reads originating from an already-observed UMI. This is a function of library complexity and sequencing depth. |
| **Q30 Bases in Barcode** | Fraction of bases with Q-score at least 30 in the cell barcode sequences. |
| **Q30 Bases in RNA Read** | Fraction of bases with Q-score at least 30 in the RNA read sequences. |
| **Q30 Bases in Sample Index** | Fraction of bases with Q-score at least 30 in the sample index sequences. |
| **Q30 Bases in UMI** | Fraction of bases with Q-score at least 30 in the UMI sequences. |
| **Fraction Reads in Cells** | The fraction of cell-barcoded, confidently mapped reads with cell-associated barcodes. |
| **Total Genes Detected** | The number of genes with at least one UMI count in any cell. |
| **Median UMI Counts per Cell** | The median number of total UMI counts across all cell-associated barcodes. |

## Antibody Sequencing

**Number of Reads**
Total number of Antibody library reads.

**Valid Barcodes**
Fraction of Antibody library reads with a barcode found in or corrected to one that is found in the whitelist.

**Valid UMIs**
Fraction of Antibody library reads with valid UMIs.

**Sequencing Saturation**
The fraction of Antibody library reads originating from an already-observed UMI. This is a function of library complexity and sequencing depth. More specifically, this is the fraction of confidently mapped, valid cell-barcode, valid UMI reads that had a non-unique (cell-barcode, UMI, CRISPR feature barcode).

**Q30 Bases in Barcode**
Fraction of Antibody library cell barcode bases with Q-score >= 30, excluding very low quality/no-call (Q <= 2) bases from the denominator.

**Q30 Bases in Antibody Read**
Fraction of Antibody library read bases with Q-score >= 30, excluding very low quality/no-call (Q <= 2) bases from the denominator. This is Read 2 for the Single Cell 3' v3 and Single Cell 5' chemistries.

**Q30 Bases in UMI**
Fraction of Antibody library UMI bases with Q-score >= 30, excluding very low quality/no-call (Q <= 2) bases from the denominator.

Version

## Antibody Application

**Fraction Antibody Reads**
Fraction of Antibody library reads that contain a recognized antibody barcode.

**Fraction Antibody Reads Usable**
Fraction of Antibody library reads that contain a recognized antibody barcode, a valid UMI, and a cell-associated barcode.

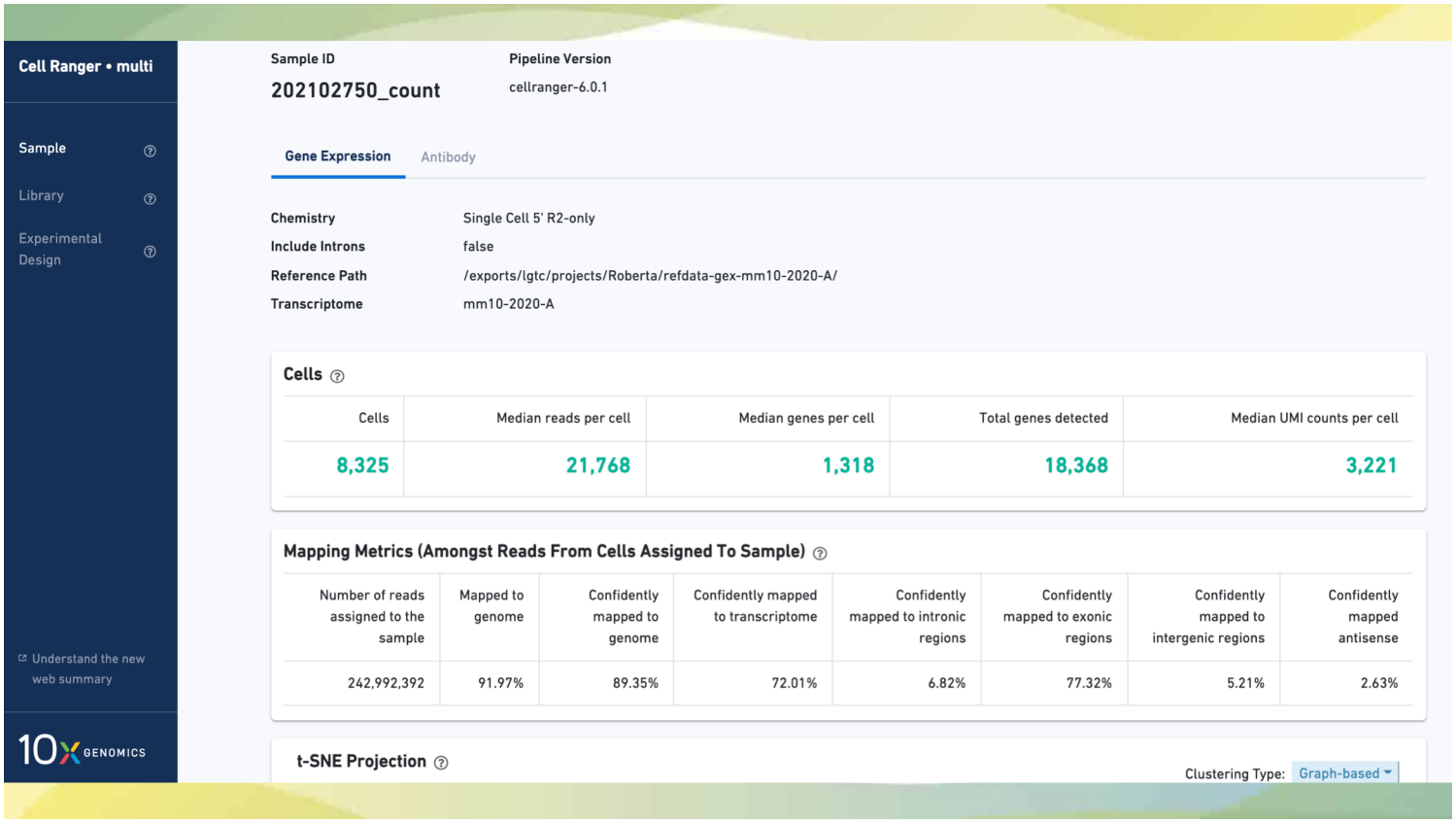
**Antibody Reads Usable per Cell**
Number of Antibody library reads usable divided by the number of cell-associated barcodes.

**Fraction Reads in Barcodes with High UMI Counts**
Fraction of Antibody library reads that was lost after removing barcodes with unusually high UMI counts (possibly aggregates).

**Fraction Unrecognized Antibody**
Among all Antibody library reads, the fraction with an unrecognizable antibody barcode.

**Antibody Reads in Cells**
Among Antibody library reads with a recognized antibody barcode, a valid UMI, and a valid barcode, the fraction associated with cell-containing partitions.

**Median UMIs per Cell (summed over all recognized antibody barcodes)**
Median UMIs per Cell (summed over all recognized antibody barcodes).


Cell Ranger • multi
Sample
Library
Experimental Design
Understand the new web summary
10X GENOMICS
Sample ID
202102750_count
Pipeline Version
cellranger-6.0.1
Gene Expression
Antibody
Chemistry
Single Cell 5' R2-only
Include Introns
false
Reference Path
/exports/lgtc/projects/Roberta/refdata-gex-mm10-2020-A/
Transcriptome
mm10-2020-A
Cells
Cells
Median reads per cell
Median genes per cell
Total genes detected
Median UMI counts per cell
8,325
21,768
1,318
18,368
3,221
Mapping Metrics (Amongst Reads From Cells Assigned To Sample)
Number of reads assigned to the sample
Mapped to genome
Confidently mapped to genome
Confidently mapped to transcriptome
Confidently mapped to intronic regions
Confidently mapped to exonic regions
Confidently mapped to intergenic regions
Confidently mapped antisense
242,992,392
91.97%
89.35%
72.01%
6.82%
77.32%
5.21%
2.63%
t-SNE Projection
Clustering Type:
Graph-based

Cell Ranger • multi

Sample

Library

Experimental Design

Understand the new web summary

10x GENOMICS

## Cell Statistics

| Physical library ID | Estimated number of cells | Mean reads per cell |
|---|---|---|
| GEX_1 | 8,325 | 35,694 |

## Sequencing Metrics

| Fastq ID | Number of reads | Number of short reads skipped | Q30 barcodes | Q30 UMI | Q30 RNA read |
|---|---|---|---|---|---|
| 202102750_GEX | 297,150,101 | 0 | 95.7% | 95.1% | 86.1% |

## Mapping Metrics (Amongst All Reads in Library)

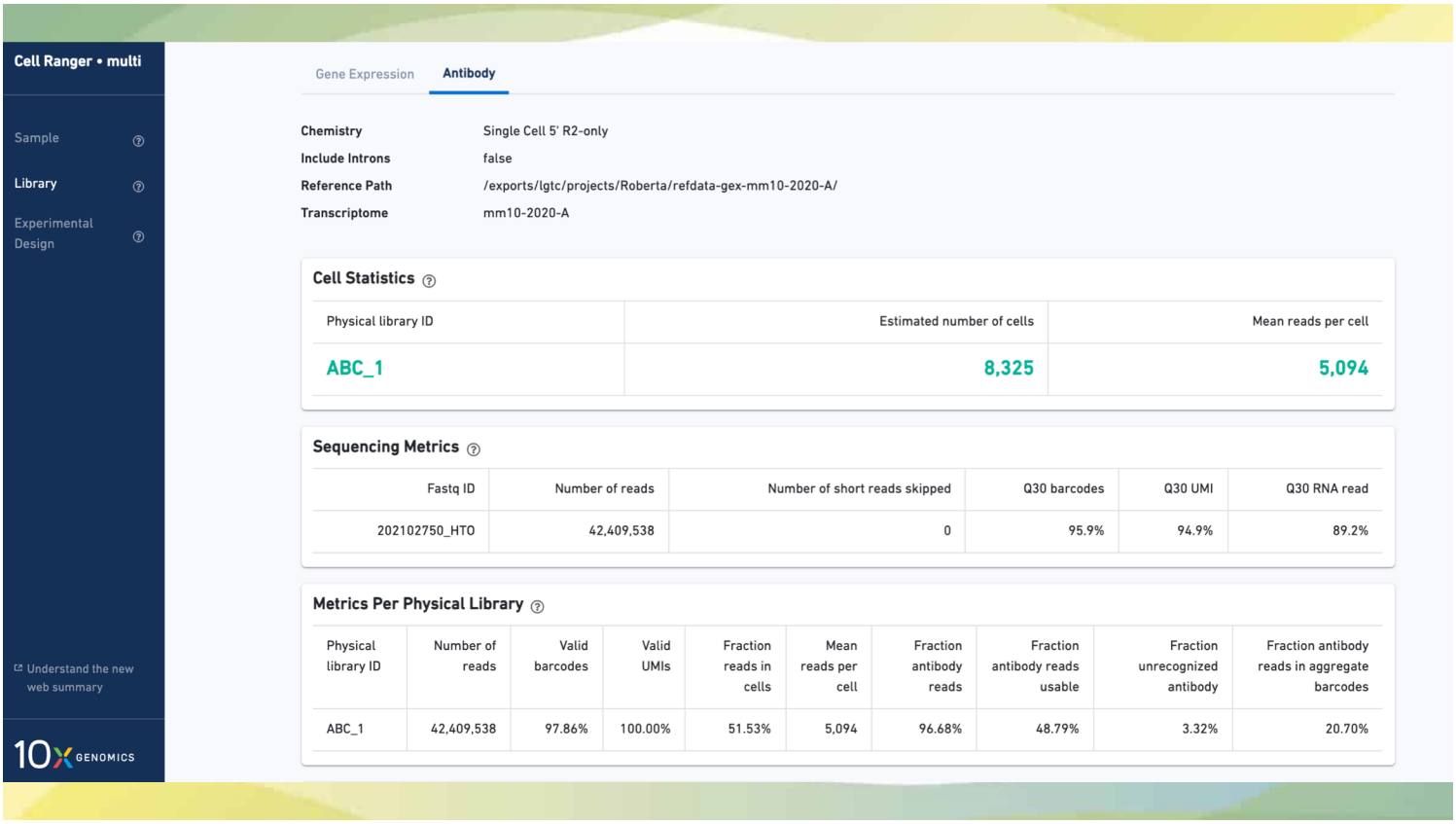
| Physical library ID | Number of reads in the library | Mapped to genome | Confidently mapped to genome | Confidently mapped to transcriptome | Confidently mapped to intronic regions | Confidently mapped to exonic regions | Confidently mapped to intergenic regions | Confidently mapped antisense |
|---|---|---|---|---|---|---|---|---|
| GEX_1 | 297,150,101 | 89.60% | 86.62% | 64.63% | 7.64% | 71.84% | 7.13% | 4.66% |

## Metrics Per Physical Library

| Physical library ID | Number of reads | Valid barcodes | Valid UMIs | Sequencing saturation | Fraction reads in cells | Mean reads per cell |
|---|---|---|---|---|---|---|
| GEX_1 | 297,150,101 | 90.63% | 99.77% | 78.84% | 94.39% | 35,694 |

| Metric | Exon-only | Intron-mode |
|---|---|---|
| Estimated Number of Cells | 11,996 | 11,984 |
| Mean Reads per Cell | 41,379 | 41,421 |
| Fraction Reads in Cells | 96.0% | 96.4% |
| Reads Mapped Confidently to Transcriptome | 52.7% | 77.7% |
| Median Genes per Cell | 2,049 | 3,268 (+59%) |
| Median UMI Counts per Cell | 6,896 | 10,186 (+48%) |
| Total Genes Detected | 25,589 | 30,351 (+19%) |

| Cell Ranger Cell Statistic Definitions | |
|---|---|
| Estimated Number of Cells | The number of barcodes identified by the cell-calling algorithm as containing a cell. |
| Mean Reads per Cell | The total number of sequenced read pairs divided by the number of cell-associated barcodes. |
| Fraction Reads in Cells | The fraction of valid-barcode, valid-UMI, confidently-mapped-to-transcriptome reads with cell-associated barcodes. |
| Reads Mapped Confidently to Transcriptome | Fraction of reads that mapped to a unique gene in the transcriptome. The read must be consistent with annotated splice junctions. These reads are considered for UMI counting. |
| Median Genes per Cell | Median number of read pairs sequenced from the cells assigned to this sample (at least one UMI count detected). |
| Median UMI Counts per Cell | Median number of UMIs obtained from the cells assigned to this sample. |
| Total Genes Detected | The number of genes with at least one UMI count in any cell. |


Cell Ranger • multi
Sample
Library
Experimental Design
Understand the new web summary
10X GENOMICS
Gene Expression
Antibody
Chemistry Single Cell 5' R2-only
Include Introns false
Reference Path /exports/lgtc/projects/Roberta/refdata-gex-mm10-2020-A/
Transcriptome mm10-2020-A
Cell Statistics
Physical library ID
Estimated number of cells
Mean reads per cell
ABC_1
8,325
5,094
Sequencing Metrics
Fastq ID
Number of reads
Number of short reads skipped
Q30 barcodes
Q30 UMI
Q30 RNA read
202102750_HTO
42,409,538
0
95.9%
94.9%
89.2%
Metrics Per Physical Library
Physical library ID
Number of reads
Valid barcodes
Valid UMIs
Fraction reads in cells
Mean reads per cell
Fraction antibody reads
Fraction antibody reads usable
Fraction unrecognized antibody
Fraction antibody reads in aggregate barcodes
ABC_1
42,409,538
97.86%
100.00%
51.53%
5,094
96.68%
48.79%
3.32%
20.70%

The analysis detected some issues. Details »

| Alert | Value | Detail |
|---|---|---|
| ⚠ **Low Fraction Reads in Cells** | 53.4% | Ideal > 70%. Application performance may be affected. Many of the reads were not assigned to cell-associated barcodes. This could be caused by high levels of ambient RNA or by a significant population of cells with a low RNA content, which the algorithm did not call as cells. The latter case can be addressed by inspecting the data to determine the appropriate cell count and using --force-cells. |

## Estimated Number of Cells

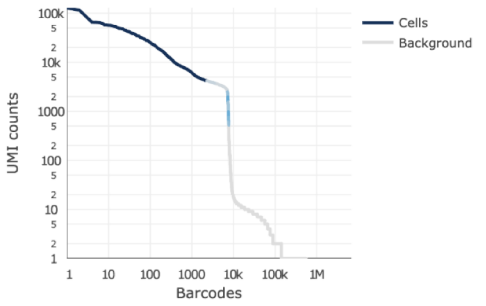
**2,553**

| Mean Reads per Cell | Median Genes per Cell |
|---|---|
| 130,565 | 2,064 |

## Sequencing

| | |
|---|---|
| Number of Reads | 333,334,060 |
| Valid Barcodes | 97.5% |
| Sequencing Saturation | 74.3% |
| Q30 Bases in Barcode | 96.1% |
| Q30 Bases in RNA Read | 90.8% |
| Q30 Bases in Sample Index | 94.8% |
| Q30 Bases in UMI | 93.3% |

## Cells




| | |
|---|---|
| Estimated Number of Cells | 2,553 |
| Fraction Reads in Cells | 53.4% |
| Mean Reads per Cell | 130,565 |
| Median Genes per Cell | 2,064 |
| Total Genes Detected | 18,966 |
| Median UMI Counts per Cell | 5,148 |

Forcing the number of cells

## Estimated Number of Cells

**8,000**

| Mean Reads per Cell | Median Genes per Cell |
|---|---|
| 41,666 | 1,586 |

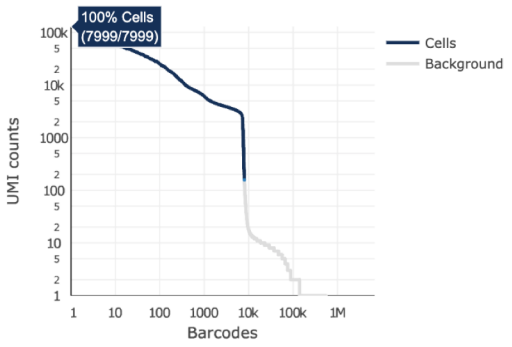
## Sequencing

| | |
|---|---|
| Number of Reads | 333,334,060 |
| Valid Barcodes | 97.5% |
| Sequencing Saturation | 74.3% |
| Q30 Bases in Barcode | 96.1% |
| Q30 Bases in RNA Read | 90.8% |
| Q30 Bases in Sample Index | 94.8% |
| Q30 Bases in UMI | 93.3% |

## Mapping

| | |
|---|---|
| Reads Mapped to Genome | 94.9% |
| Reads Mapped Confidently to Genome | 93.0% |
| Reads Mapped Confidently to Intergenic Regions | 5.0% |
| Reads Mapped Confidently to Intronic Regions | 38.4% |
| Reads Mapped Confidently to Exonic Regions | 49.6% |
| Reads Mapped Confidently to Transcriptome | 45.7% |
| Reads Mapped Antisense to Gene | 1.4% |

## Cells




| | |
|---|---|
| Estimated Number of Cells | 8,000 |
| Fraction Reads in Cells | 97.9% |
| Mean Reads per Cell | 41,666 |
| Median Genes per Cell | 1,586 |
| Total Genes Detected | 19,782 |
| Median UMI Counts per Cell | 3,614 |

## Sample

| | |
|---|---|
| Name | _GEX_count_8k |
| Description | |
| Transcriptome | GRCh38 |
| Chemistry | Single Cell 3' v3 |
| Cell Ranger Version | 3.0.0 |

## Conclusions

**Always perform QC of your libraries!**

**Be aware of library specifics → critical mindset!**