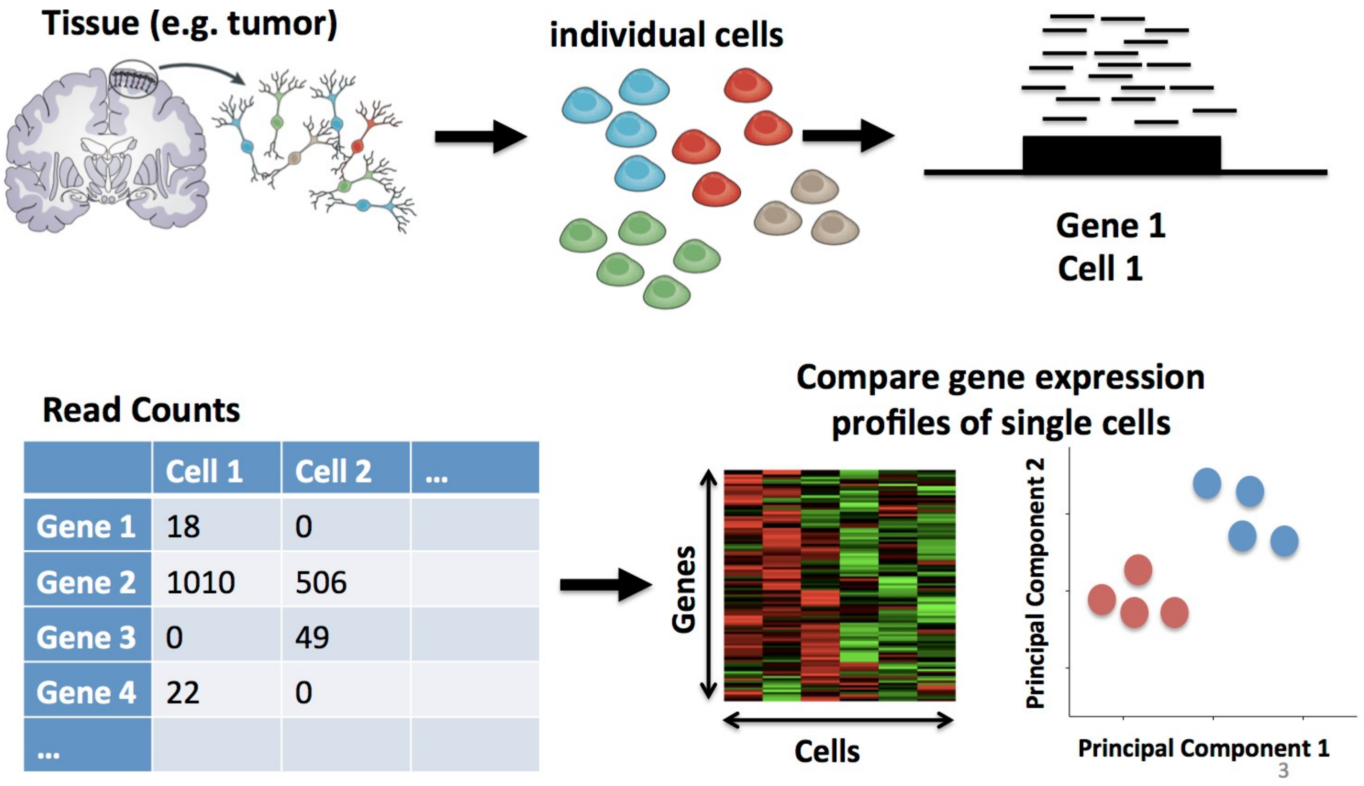
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | Cambridge Institute

Alignment and feature counting

September 2022

Single Cell RNAseq Analysis Workflow

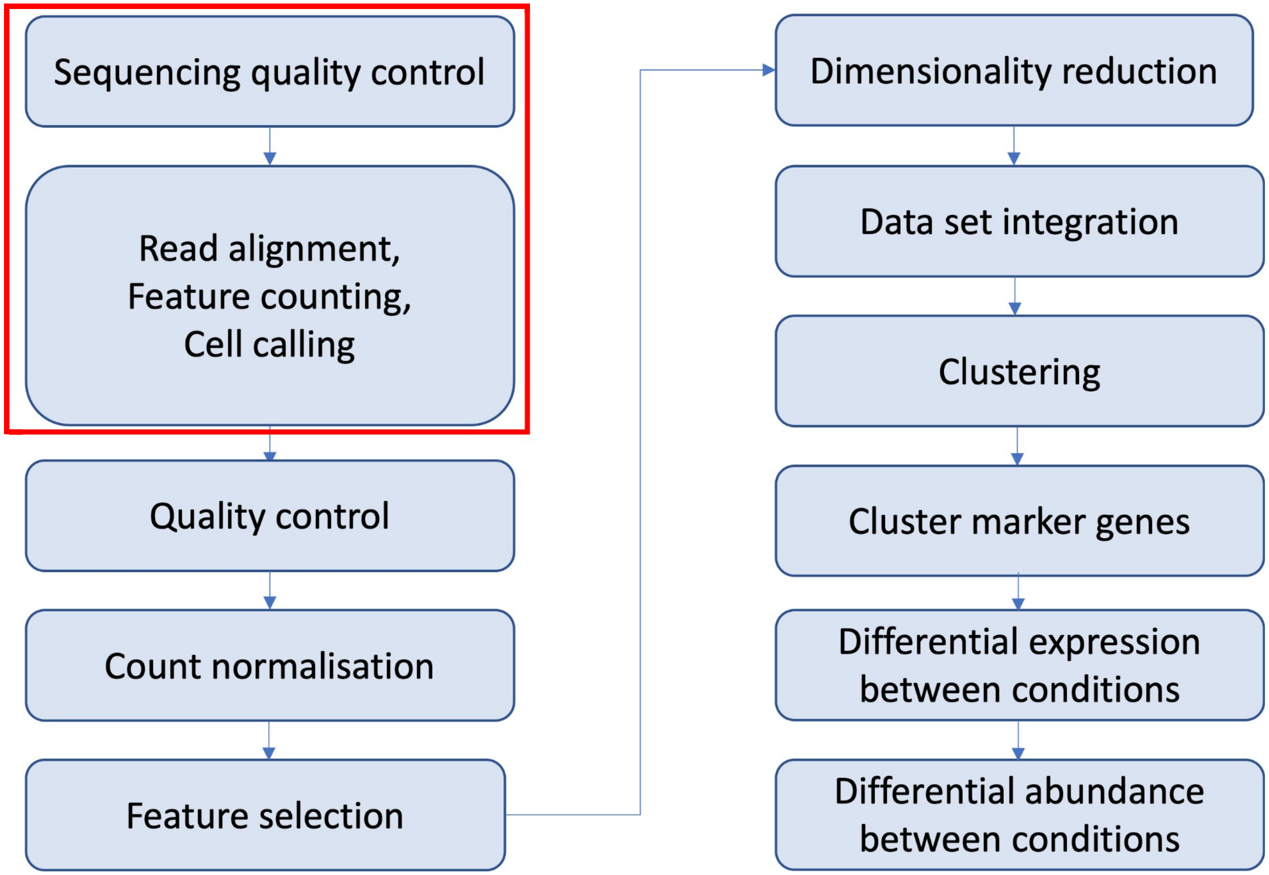


Read Counts

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

Image by Stephanie Hicks via learn.gencore.bio.nyu.edu

Single Cell RNAseq Analysis Workflow



Sequencing quality control

Read alignment,
Feature counting,
Cell calling

Quality control

Count normalisation

Feature selection

Dimensionality reduction

Data set integration

Clustering

Cluster marker genes

Differential expression between conditions

Differential abundance between conditions

10x single-cell isolation

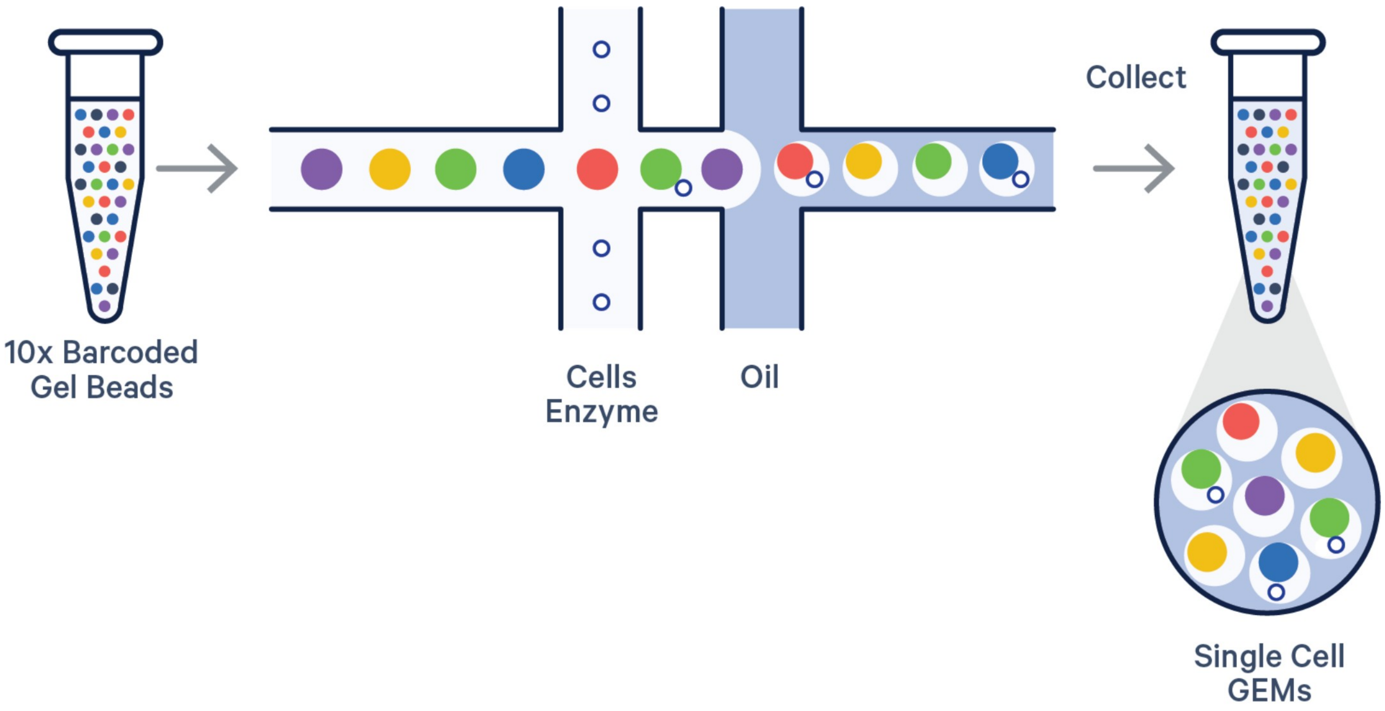
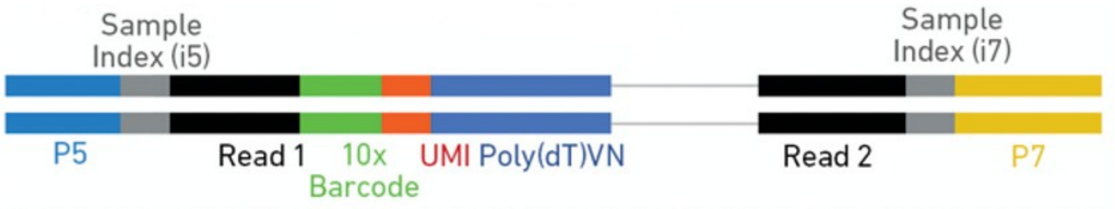


Image by 10x Genomics

10x library file structure

The 10x library contains four pieces of information, in the form of DNA sequences, for each “read”.

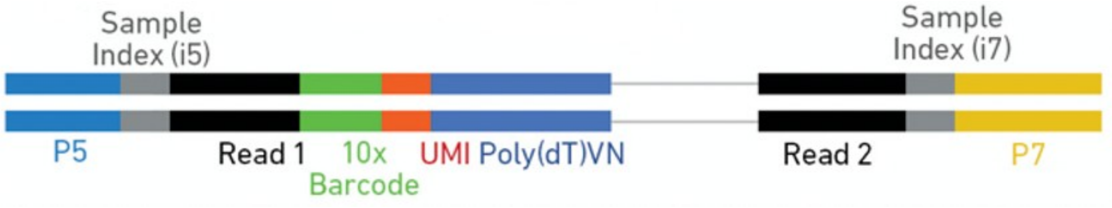
- **sample index** - identifies the library, with one or two indexes per sample
- **10x barcode** - identifies the droplet in the library
- **UMI** - identifies the transcript molecule within a cell and gene
- **insert** - the transcript molecule



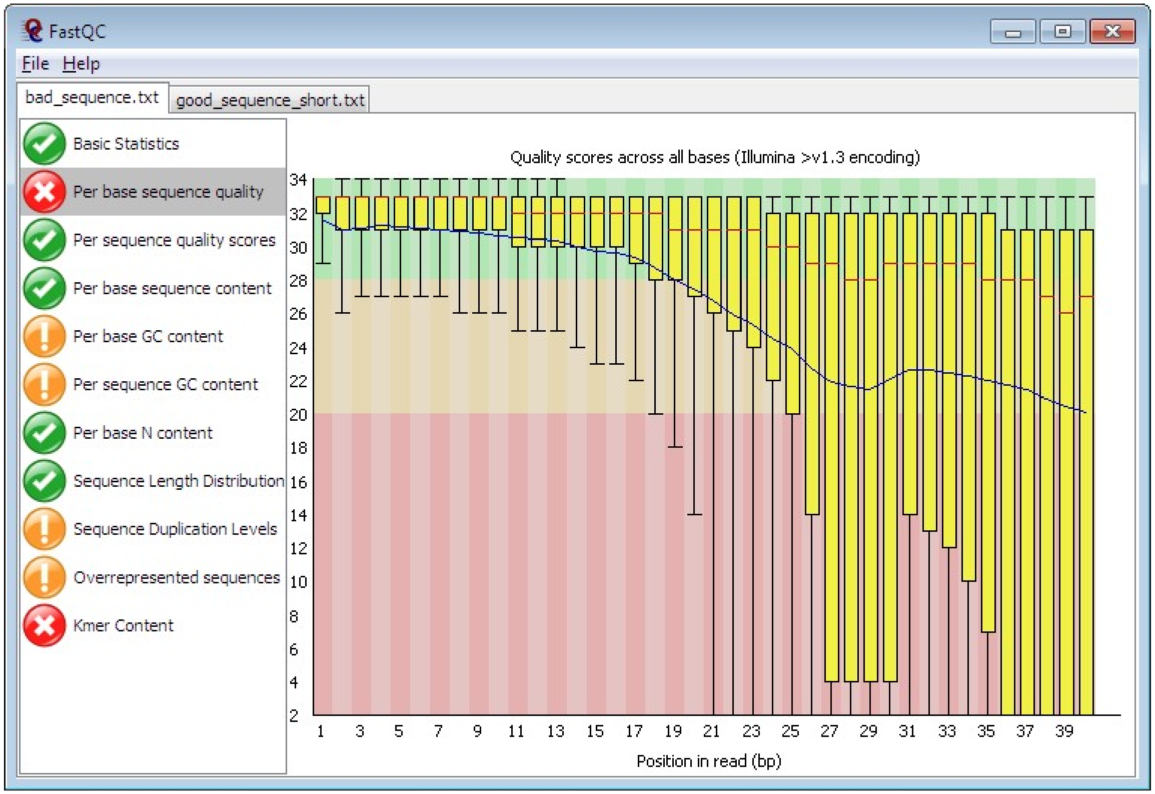
Raw fastq files

The sequences for any given fragment will generally be delivered in 3 or 4 files:

- **I1**: I7 sample index
- **I2**: I5 sample index if present (dual indexing only)
- **R1**: 10x barcode + UMI
- **R2**: insert sequence

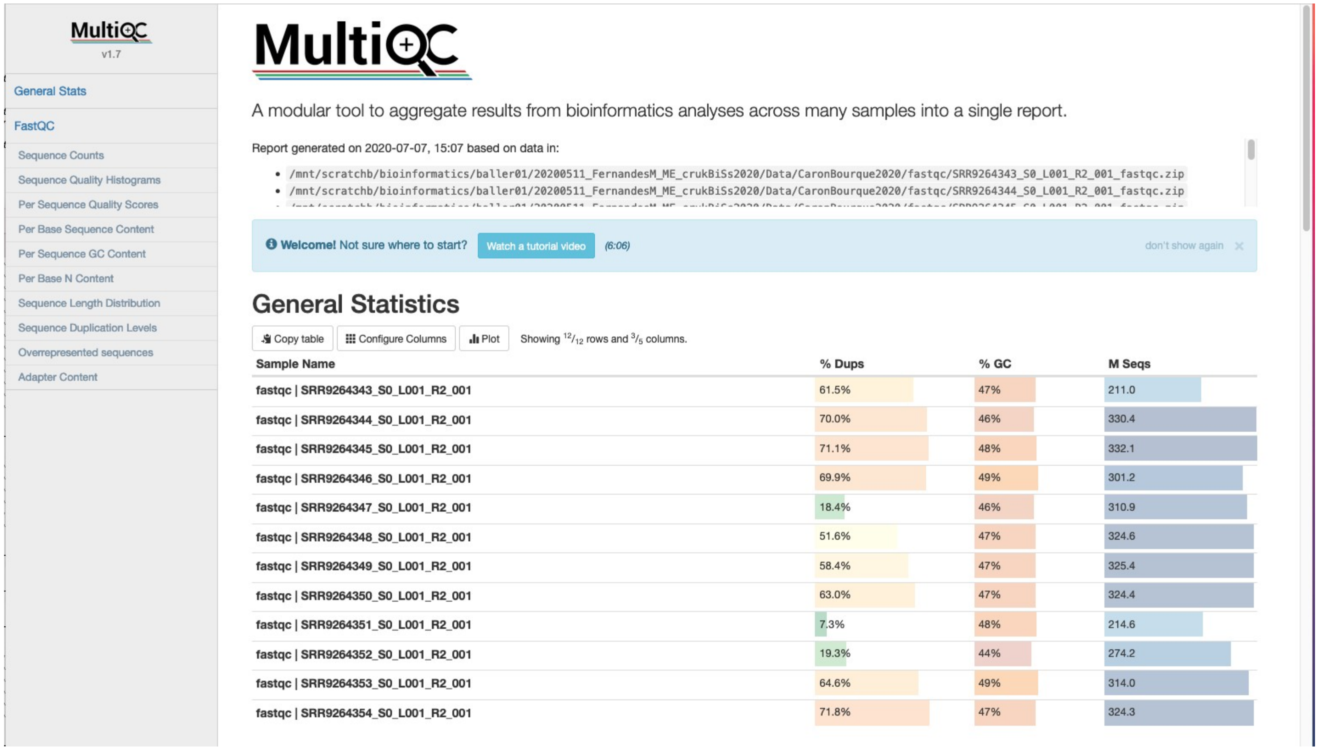


QC of Raw Reads - FASTQC



FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

QC of Raw Reads - MultiQC - General Statistics



A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2020-07-07, 15:07 based on data in:

- `/mnt/scratchb/bioinformatics/baller01/20200511_FernandesM_ME_crukBiSs2020/Data/CaronBourque2020/fastqc/SRR9264343_S0_L001_R2_001_fastqc.zip`
- `/mnt/scratchb/bioinformatics/baller01/20200511_FernandesM_ME_crukBiSs2020/Data/CaronBourque2020/fastqc/SRR9264344_S0_L001_R2_001_fastqc.zip`

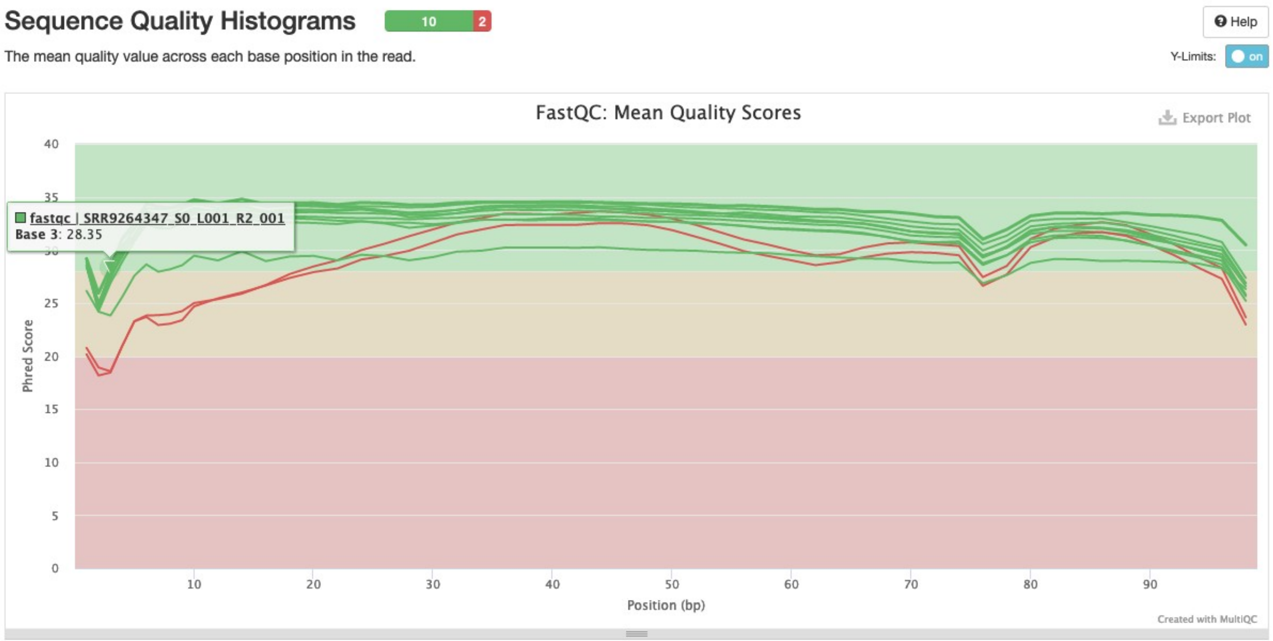
Welcome! Not sure where to start?

General Statistics

Showing $^{12}/_{12}$ rows and $^{3}/_{5}$ columns.

Sample Name	% Dups	% GC	M Seqs
fastqc | SRR9264343_S0_L001_R2_001	61.5%	47%	211.0
fastqc | SRR9264344_S0_L001_R2_001	70.0%	46%	330.4
fastqc | SRR9264345_S0_L001_R2_001	71.1%	48%	332.1
fastqc | SRR9264346_S0_L001_R2_001	69.9%	49%	301.2
fastqc | SRR9264347_S0_L001_R2_001	18.4%	46%	310.9
fastqc | SRR9264348_S0_L001_R2_001	51.6%	47%	324.6
fastqc | SRR9264349_S0_L001_R2_001	58.4%	47%	325.4
fastqc | SRR9264350_S0_L001_R2_001	63.0%	47%	324.4
fastqc | SRR9264351_S0_L001_R2_001	7.3%	48%	214.6
fastqc | SRR9264352_S0_L001_R2_001	19.3%	44%	274.2
fastqc | SRR9264353_S0_L001_R2_001	64.6%	49%	314.0
fastqc | SRR9264354_S0_L001_R2_001	71.8%	47%	324.3

QC of Raw Reads - MultiQC - Sequence Quality Histograms



Sequence Quality Histograms 10 2

The mean quality value across each base position in the read.

Alignment and counting

The first steps in the analysis of single cell RNAseq data:

- Align reads to genome
- Annotate reads with feature (gene)
- Quantify gene expression

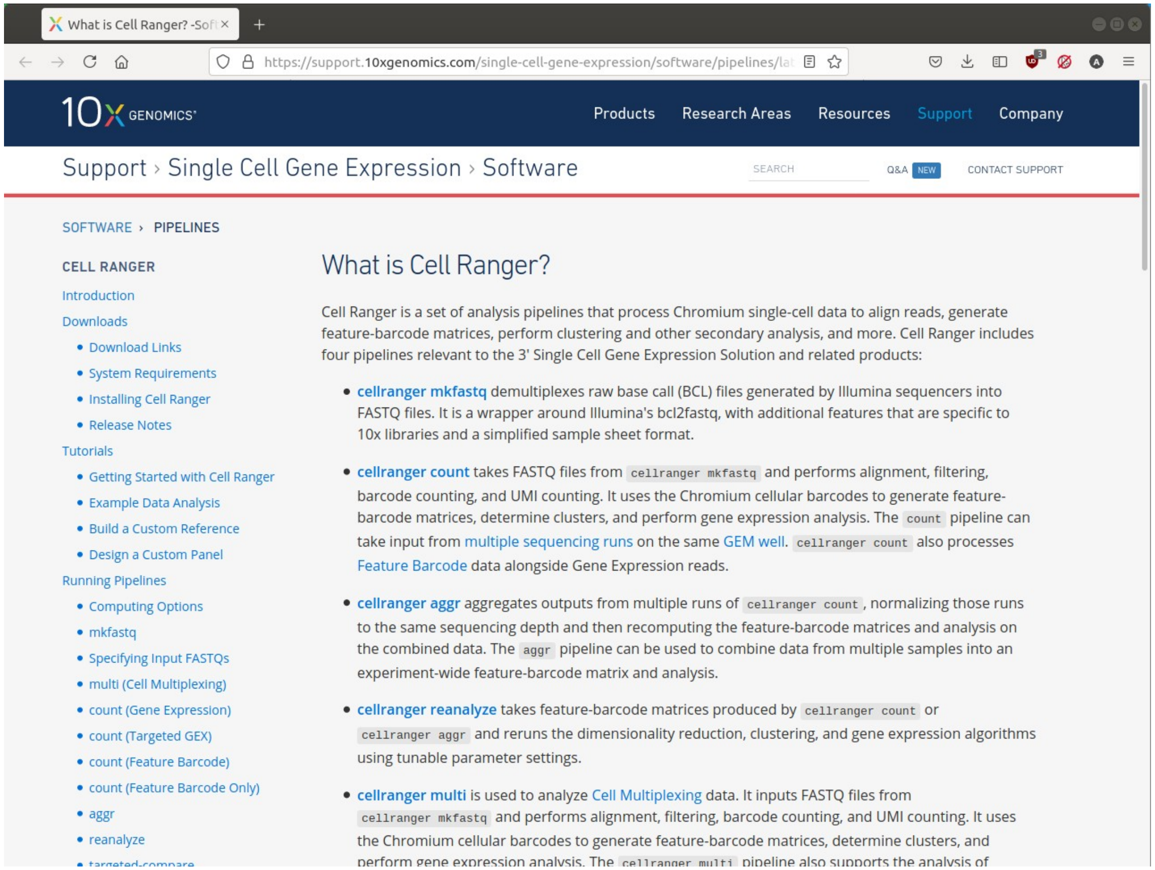
Cell Ranger

- 10x Cell Ranger - This not only carries out the alignment and feature counting, but will also:
    - Call cells
    - Generate a summary report in html format
    - Generate a “cloupe” file

Alternative methods include:

- STAR solo:
    - Generates outputs very similar to CellRanger minus the cloupe file and the QC report
    - Will run with lower memory requirements in a shorter time than Cell Ranger
- Alevin:
    - Based on the popular Salmon tool for bulk RNAseq feature counting
    - Alevin supports both 10x-Chromium and Drop-seq derived data

Obtaining Cell Ranger


What is Cell Ranger? -Sof
https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/la
10X GENOMICS
Products
Research Areas
Resources
Support
Company
Support › Single Cell Gene Expression › Software
SEARCH
Q&A NEW
CONTACT SUPPORT
SOFTWARE › PIPELINES

CELL RANGER

- Introduction
- Downloads
  - Download Links
  - System Requirements
  - Installing Cell Ranger
  - Release Notes
- Tutorials
  - Getting Started with Cell Ranger
  - Example Data Analysis
  - Build a Custom Reference
  - Design a Custom Panel
- Running Pipelines
  - Computing Options
  - mkfastq
  - Specifying Input FASTQs
  - multi (Cell Multiplexing)
  - count (Gene Expression)
  - count (Targeted GEX)
  - count (Feature Barcode)
  - count (Feature Barcode Only)
  - aggr
  - reanalyze
  - targeted-compare

What is Cell Ranger?

Cell Ranger is a set of analysis pipelines that process Chromium single-cell data to align reads, generate feature-barcode matrices, perform clustering and other secondary analysis, and more. Cell Ranger includes four pipelines relevant to the 3' Single Cell Gene Expression Solution and related products:

- **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. `cellranger count` 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. The `aggr` pipeline can be used to combine data from multiple samples into an experiment-wide feature-barcode matrix and analysis.
- **cellranger reanalyze** takes feature-barcode matrices produced by `cellranger count` or `cellranger aggr` and reruns the dimensionality reduction, clustering, and gene expression algorithms using tunable parameter settings.
- **cellranger multi** is used to analyze Cell Multiplexing data. It inputs 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 `cellranger multi` pipeline also supports the analysis of

Setup instructions given in the course materials homepage.

Cell Ranger tools

Cell Ranger includes a number of different tools for analysing scRNAseq data, including:

- `cellranger mkref` - for making custom references
- `cellranger count` - for aligning reads and generating a count matrix
- `cellranger aggr` - for combining multiple samples and normalising the counts

Preparing the raw fastq files

Cell Ranger requires the fastq file names to follow a convention:

```
<SampleName>_S<SampleNumber>_L00<Lane>_<Read>_001.fastq.gz
```

e.g. for a single sample in the Caron data set we have:

```
    SRR9264343_S0_L001_I1_001.fastq.gz
    SRR9264343_S0_L001_R1_001.fastq.gz
    SRR9264343_S0_L001_R2_001.fastq.gz
```

Genome/Transcriptome Reference

As with other aligners Cell Ranger requires the information about the genome and transcriptome of interest to be provided in a specific format.

- Obtain from the 10x website for human or mouse (or both - PDX)
- Build a custom reference with `cellranger mkref`

```
cellranger mkref \
  --fasta={GENOME FASTA} \
  --genes={ANNOTATION GTF} \
  --genome={OUTPUT FOLDER FOR INDEX} \
  --nthreads={CPUS}
```

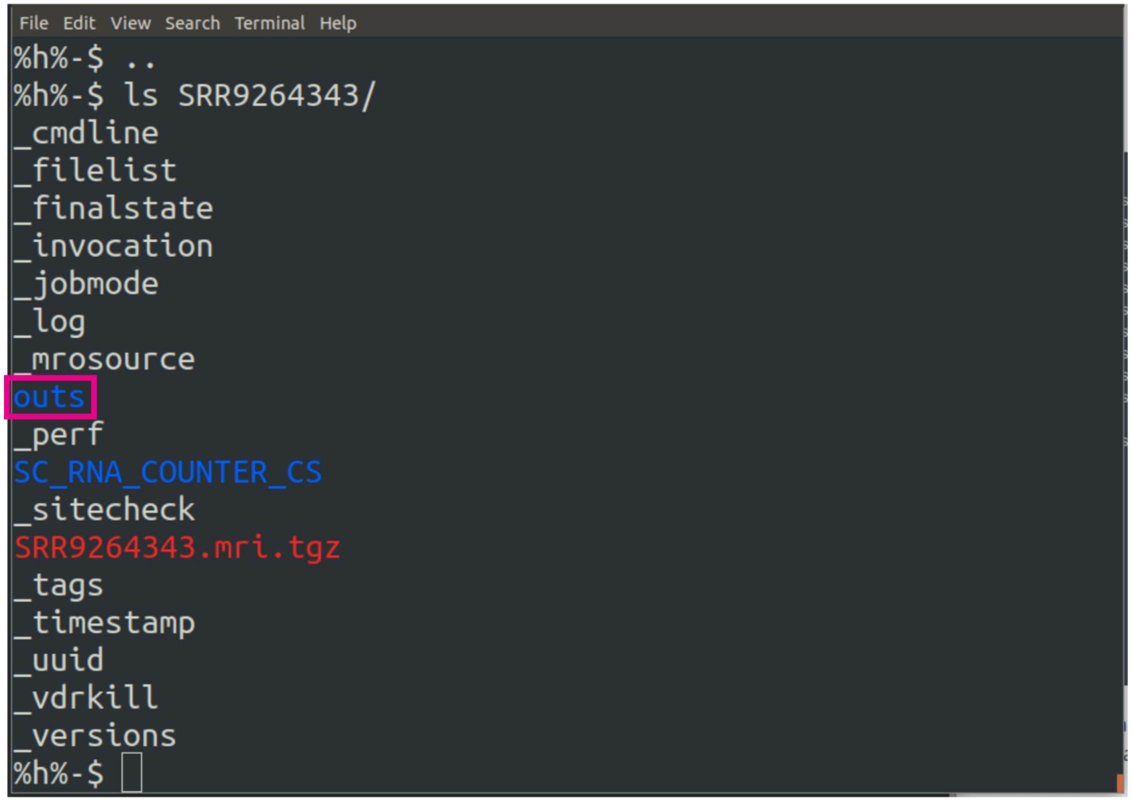
Running `cellranger count`

- Computationally very intensive
- High memory requirements

```
cellranger count \
  --id={OUTPUT_SAMPLE_NAME} \
  --transcriptome={DIRECTORY_WITH_REFERENCE} \
  --fastqs={DIRECTORY_WITH_FASTQ_FILES} \
  --sample={NAME_OF_SAMPLE_IN_FASTQ_FILES} \
  --localcores={NUMBER_OF_CPUS} \
  --localmem={RAM_MEMORY}
```

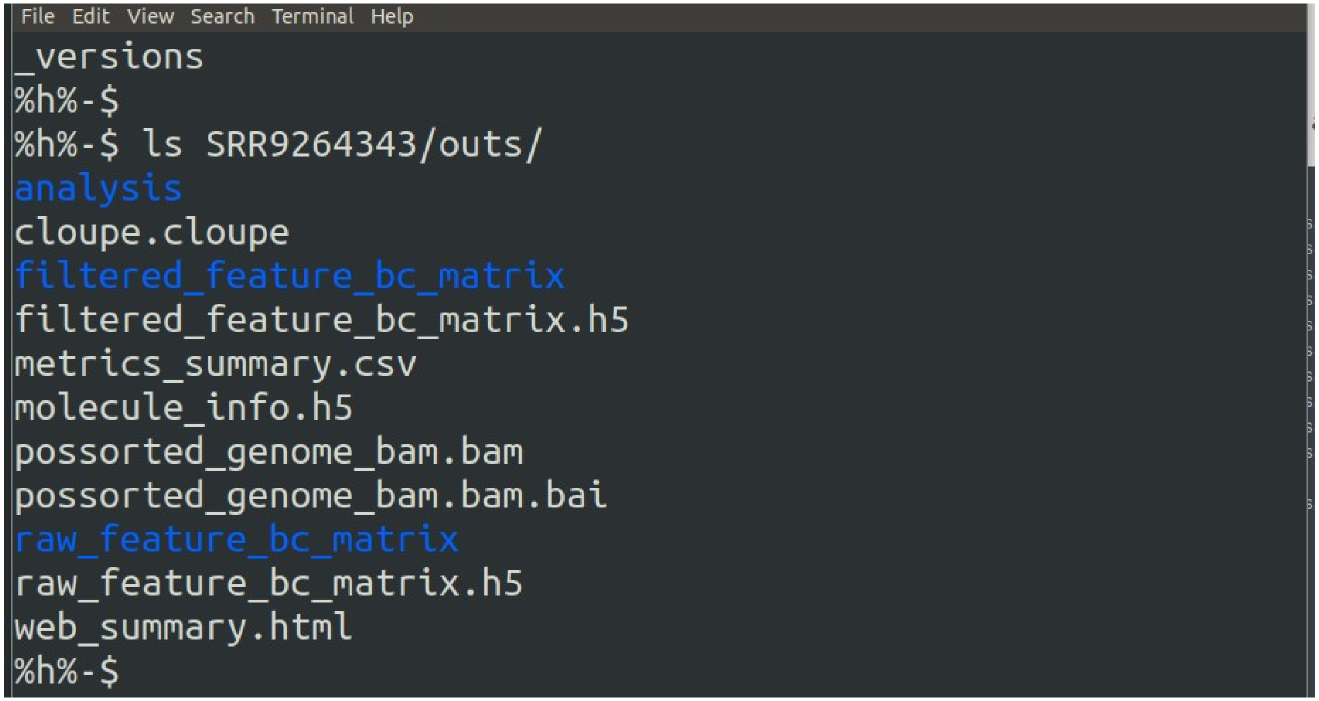
Cell Ranger outputs

- One directory per sample



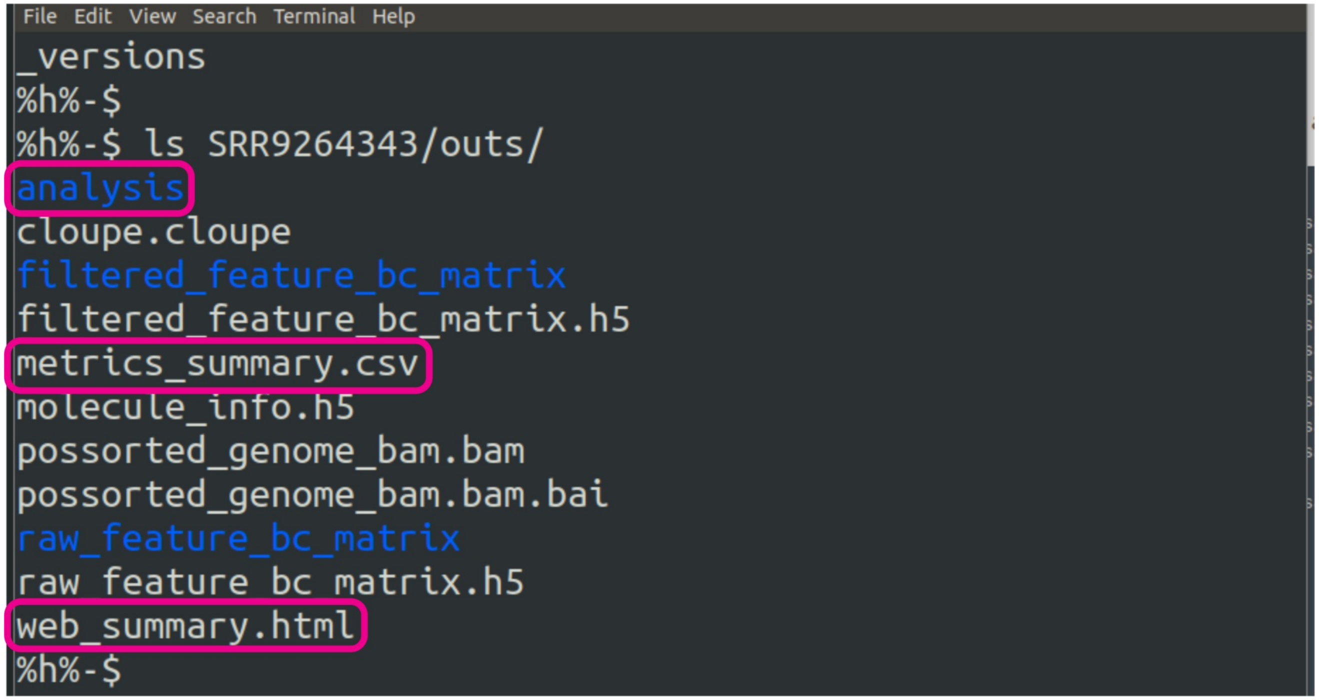
```
%h%-$ ..
%h%-$ ls SRR9264343/
_cmdline
_filelist
_finalstate
_invocation
_jobmode
_log
_mrosource
outs
_perf
SC_RNA_COUNTER_CS
_sitecheck
SRR9264343.mri.tgz
_tags
_timestamp
_uuid
_vdrkill
_versions
%h%-$
```

Cell Ranger outputs



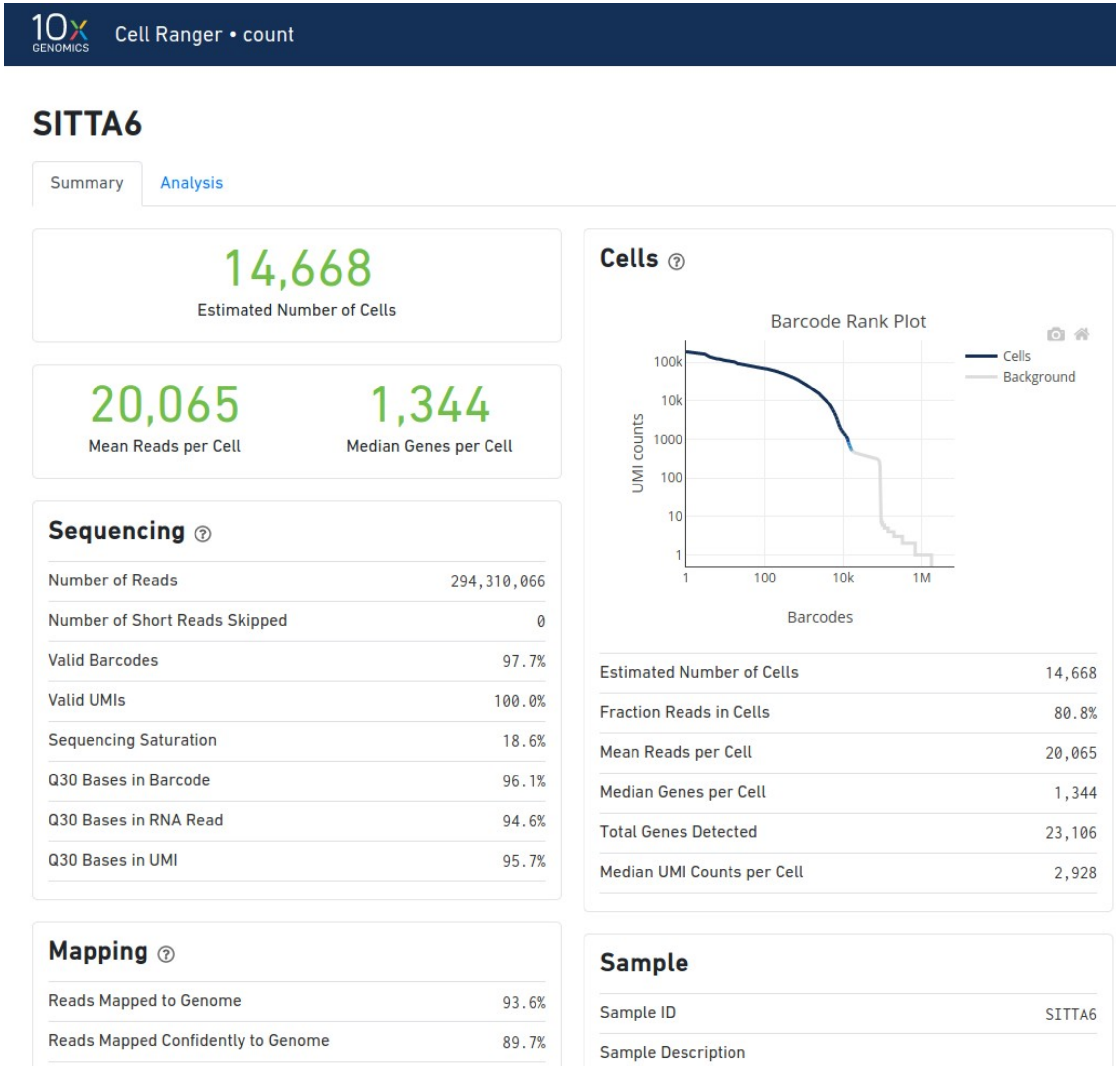
```
_versions
%h%-$
%h%-$ ls SRR9264343/outs/
analysis
cloupe.cloupe
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
metrics_summary.csv
molecule_info.h5
possorted_genome_bam.bam
possorted_genome_bam.bam.bai
raw_feature_bc_matrix
raw_feature_bc_matrix.h5
web_summary.html
%h%-$
```

Cell Ranger outputs



```
_versions
%h%-$
%h%-$ ls SRR9264343/outs/
analysis
cloupe.cloupe
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
metrics_summary.csv
molecule_info.h5
possorted_genome_bam.bam
possorted_genome_bam.bam.bai
raw_feature_bc_matrix
raw_feature_bc_matrix.h5
web_summary.html
%h%-$
```

Cell Ranger report



SITTA6

14,668
Estimated Number of Cells

20,065
Mean Reads per Cell

1,344
Median Genes per Cell

Sequencing

Number of Reads	294,310,066
Number of Short Reads Skipped	0
Valid Barcodes	97.7%
Valid UMIs	100.0%
Sequencing Saturation	18.6%
Q30 Bases in Barcode	96.1%
Q30 Bases in RNA Read	94.6%
Q30 Bases in UMI	95.7%

Mapping

Reads Mapped to Genome	93.6%
Reads Mapped Confidently to Genome	89.7%

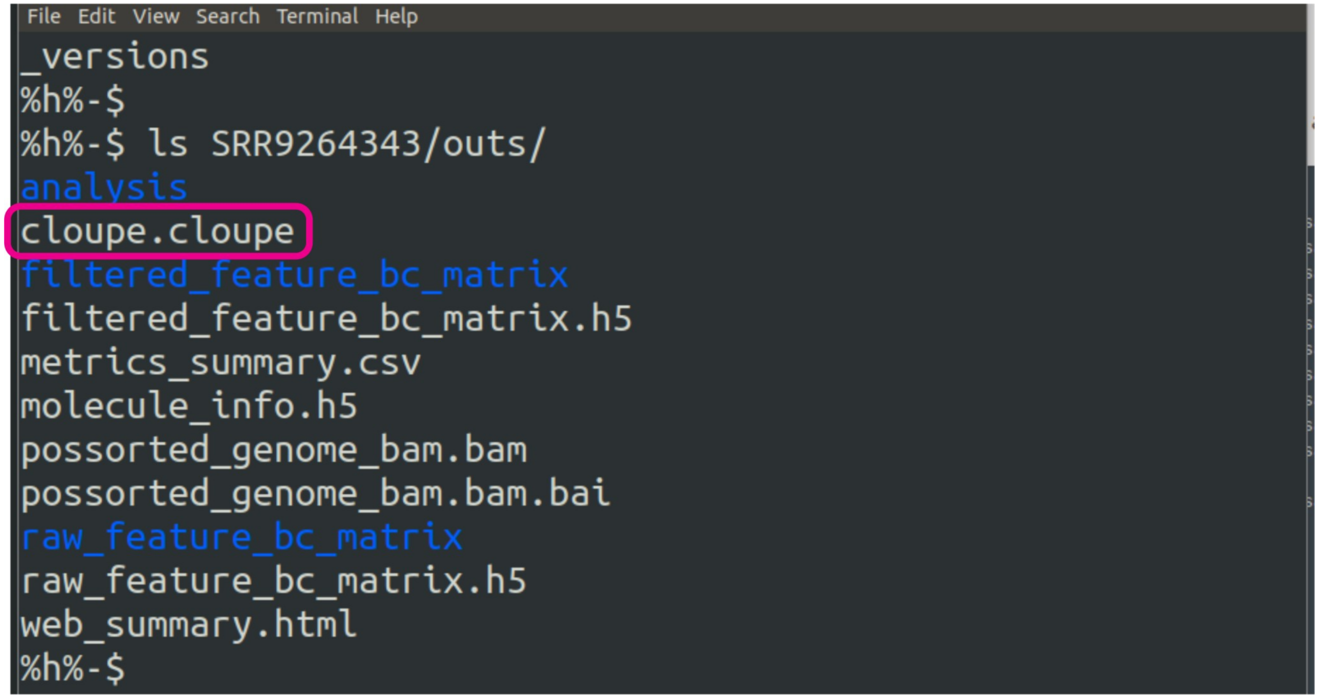
Cells

Estimated Number of Cells	14,668
Fraction Reads in Cells	80.8%
Mean Reads per Cell	20,065
Median Genes per Cell	1,344
Total Genes Detected	23,106
Median UMI Counts per Cell	2,928

Sample

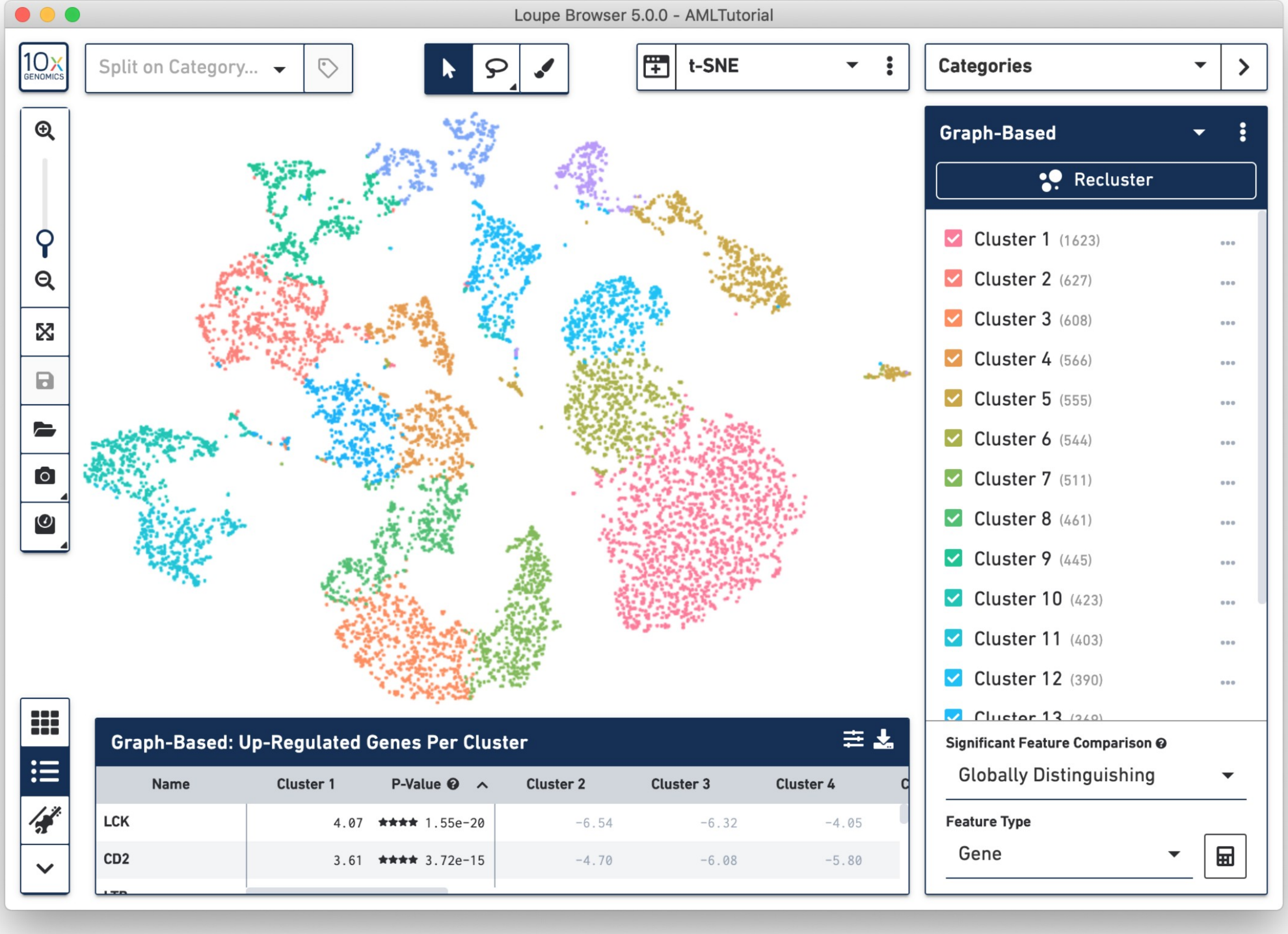
Sample ID	SITTA6
Sample Description	

Cell Ranger outputs



```
_versions
%h%-$
%h%-$ ls SRR9264343/outs/
analysis
cloupe.cloupe
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
metrics_summary.csv
molecule_info.h5
possorted_genome_bam.bam
possorted_genome_bam.bam.bai
raw_feature_bc_matrix
raw_feature_bc_matrix.h5
web_summary.html
%h%-$
```

Loupe Browser



☑ Cluster 1 (1623)
☑ Cluster 2 (627)
☑ Cluster 3 (608)
☑ Cluster 4 (566)
☑ Cluster 5 (555)
☑ Cluster 6 (544)
☑ Cluster 7 (511)
☑ Cluster 8 (461)
☑ Cluster 9 (445)
☑ Cluster 10 (423)
☑ Cluster 11 (403)
☑ Cluster 12 (390)
☑ Cluster 13 (369)

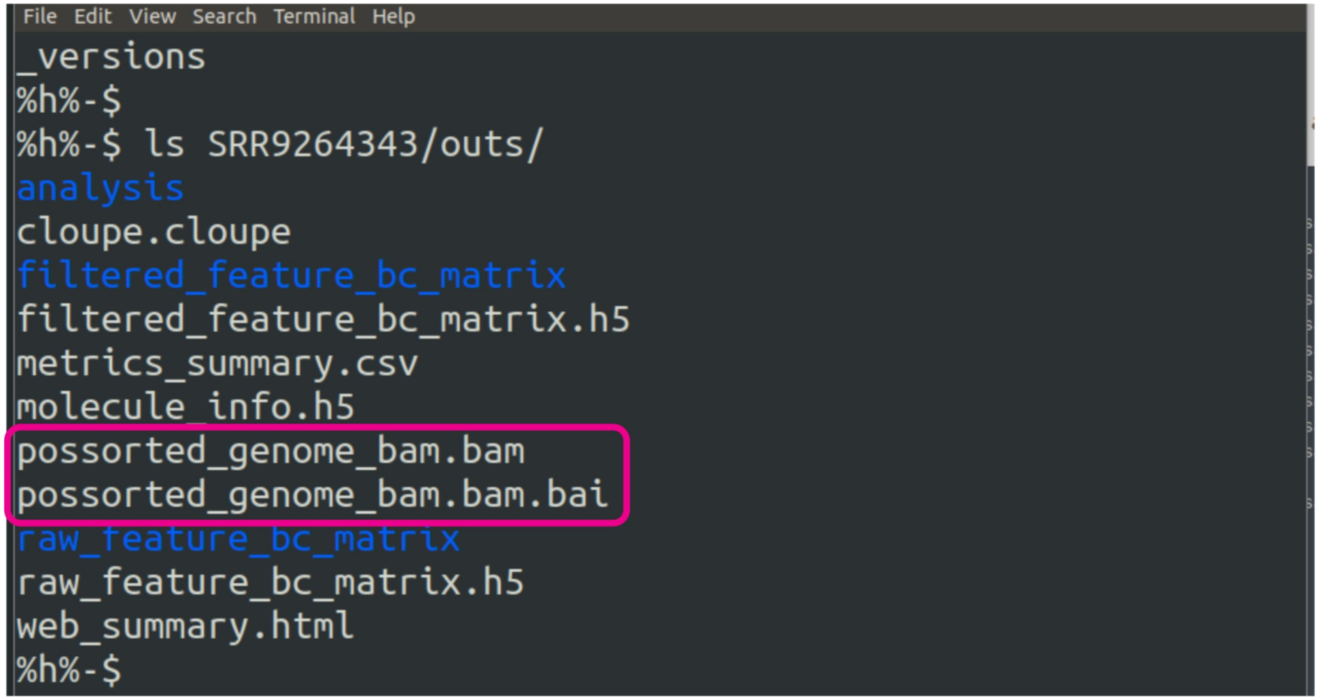
Significant Feature Comparison: Globally Distinguishing

Feature Type: Gene

Graph-Based: Up-Regulated Genes Per Cluster

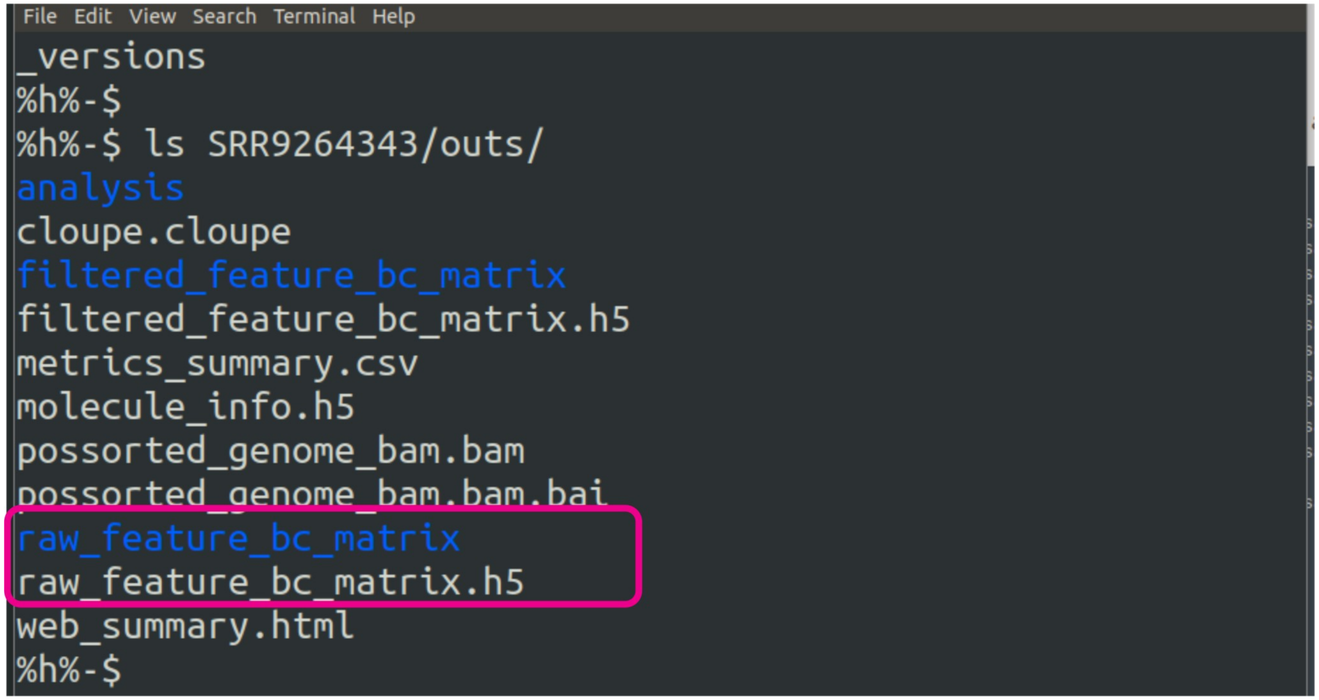
Name	Cluster 1	P-Value	Cluster 2	Cluster 3	Cluster 4
LCK	4.07	★★★★ 1.55e-20	-6.54	-6.32	-4.05
CD2	3.61	★★★★ 3.72e-15	-4.70	-6.08	-5.80

Cell Ranger outputs



```
_versions
%h%-$
%h%-$ ls SRR9264343/outs/
analysis
cloupe.cloupe
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
metrics_summary.csv
molecule_info.h5
possorted_genome_bam.bam
possorted_genome_bam.bam.bai
raw_feature_bc_matrix
raw_feature_bc_matrix.h5
web_summary.html
%h%-$
```

Cell Ranger outputs



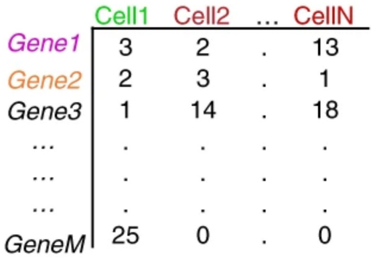
```
_versions
%h%-$
%h%-$ ls SRR9264343/outs/
analysis
cloupe.cloupe
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
metrics_summary.csv
molecule_info.h5
possorted_genome_bam.bam
possorted_genome_bam.bam.bai
raw_feature_bc_matrix
raw_feature_bc_matrix.h5
web_summary.html
%h%-$
```

Cell Ranger outputs

Two types of outputs:

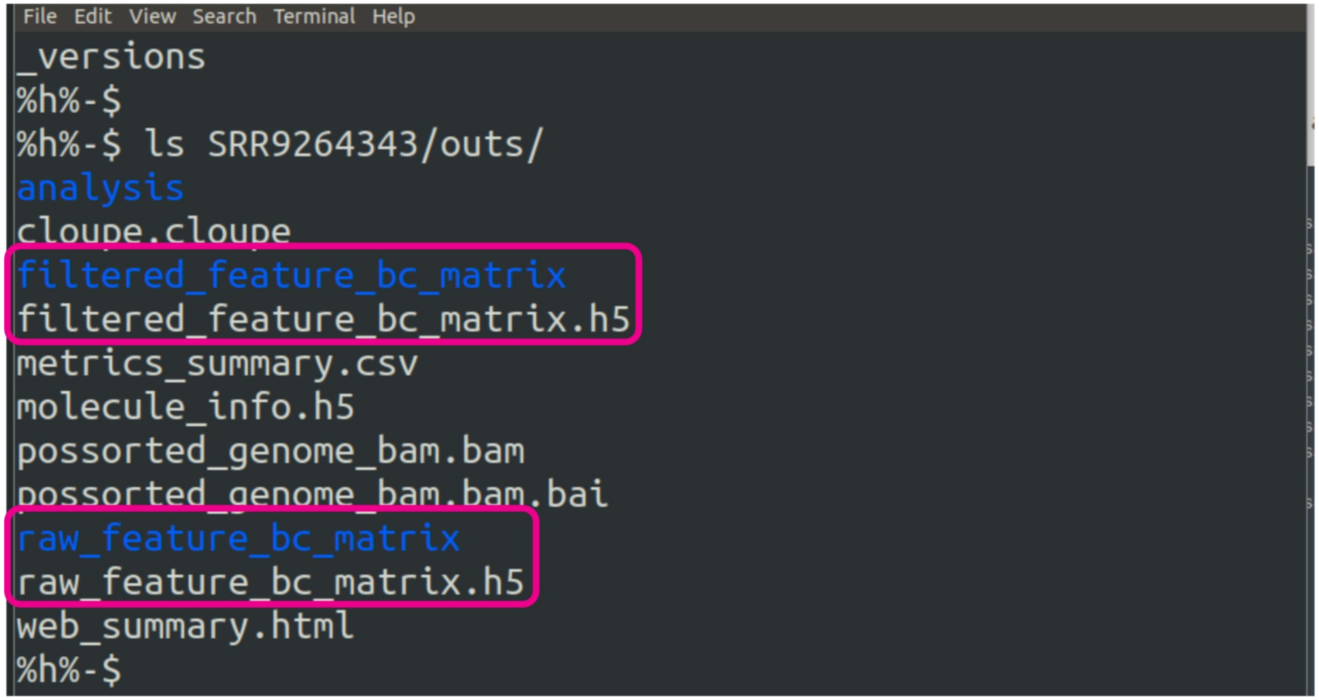
- Text-based files: `.tsv` and `.mtx`
- HDF5 files: `.h5`

Both of these can be read by standard scRNA-seq analysis packages and contain data for a **unique molecular identified (UMI) count matrix**:



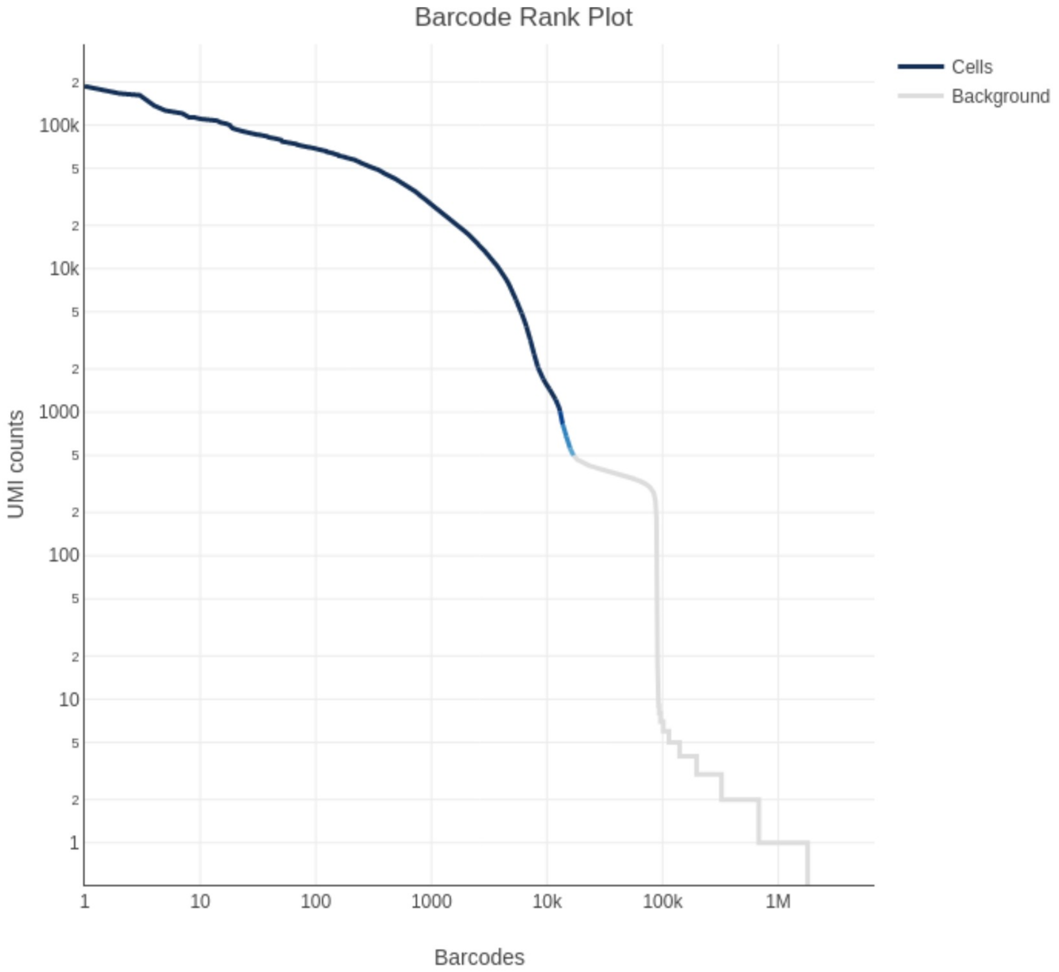
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

Cell Ranger outputs

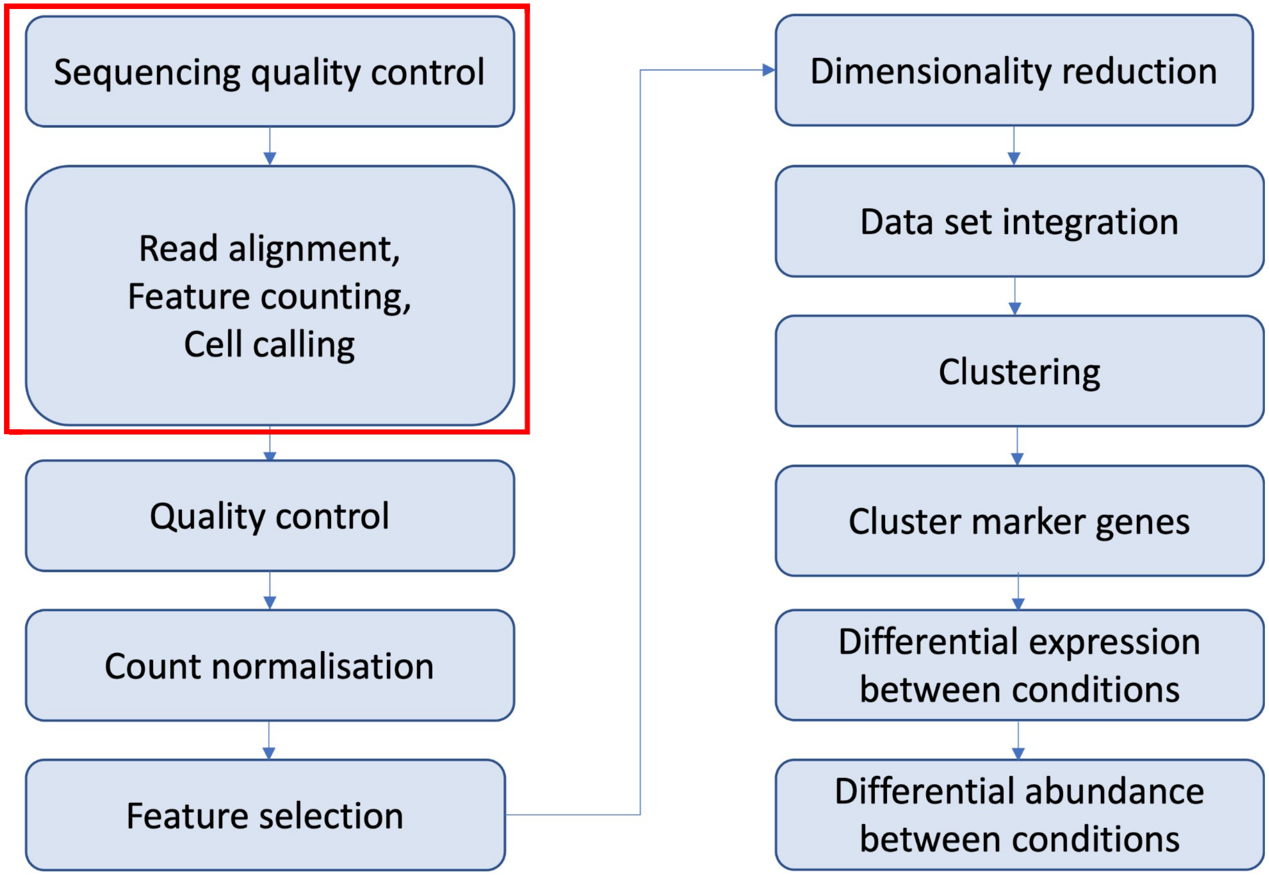


```
_versions
%h%-$
%h%-$ ls SRR9264343/outs/
analysis
cloupe.cloupe
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
metrics_summary.csv
molecule_info.h5
possorted_genome_bam.bam
possorted_genome_bam.bam.bai
raw_feature_bc_matrix
raw_feature_bc_matrix.h5
web_summary.html
%h%-$
```

Cell Ranger cell calling



Single Cell RNAseq Analysis Workflow


Sequencing quality control
Read alignment,
Feature counting,
Cell calling
Quality control
Count normalisation
Feature selection
Dimensionality reduction
Data set integration
Clustering
Cluster marker genes
Differential expression
between conditions
Differential abundance
between conditions