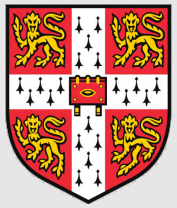


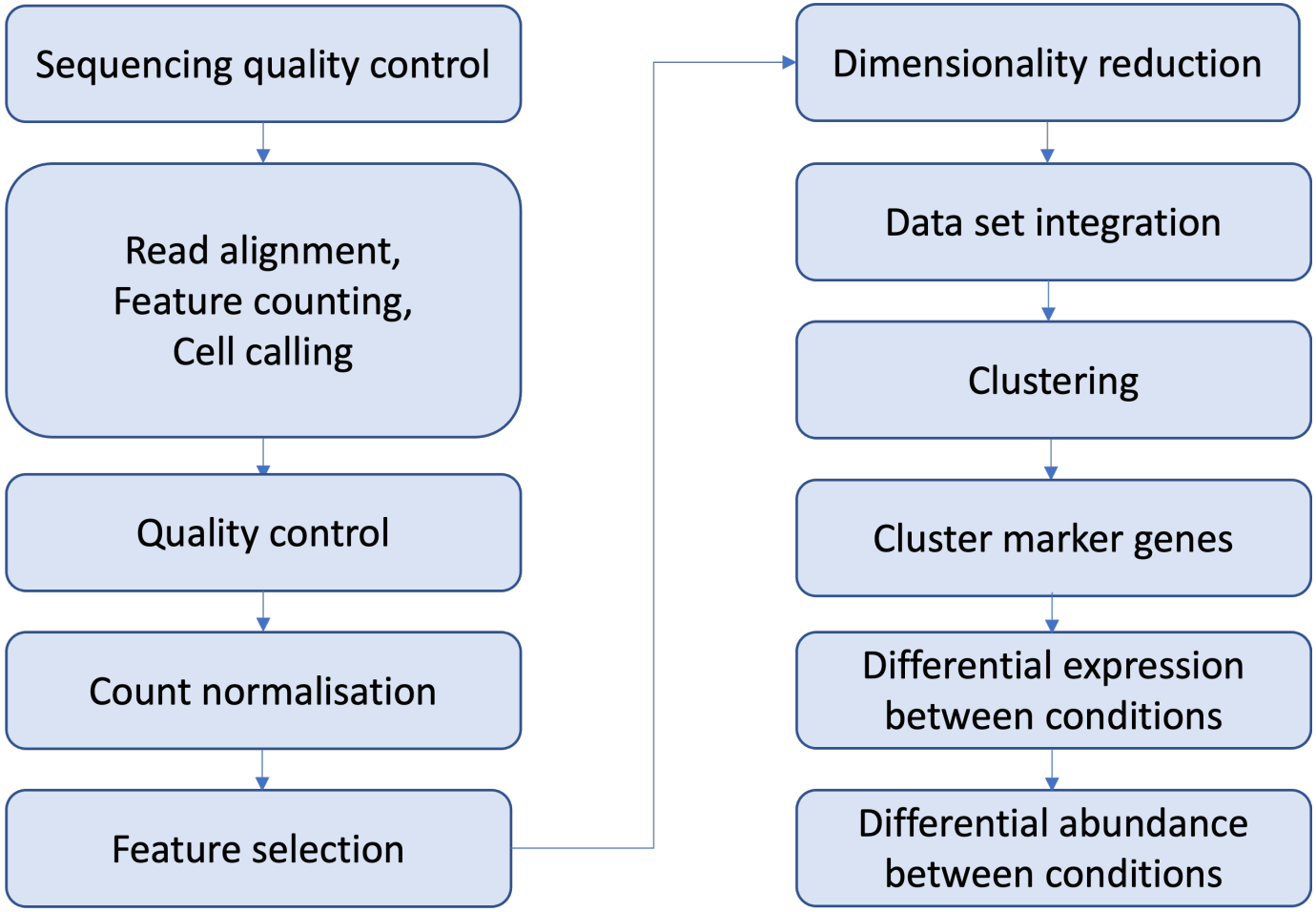
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Alignment and feature counting

September 2022

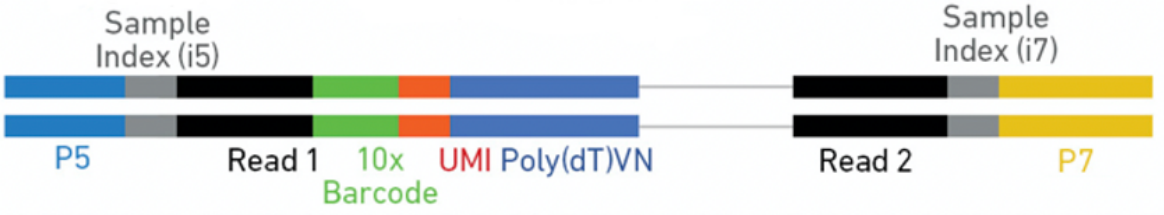
Single Cell RNAseq Analysis Workflow



10x library file structure

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

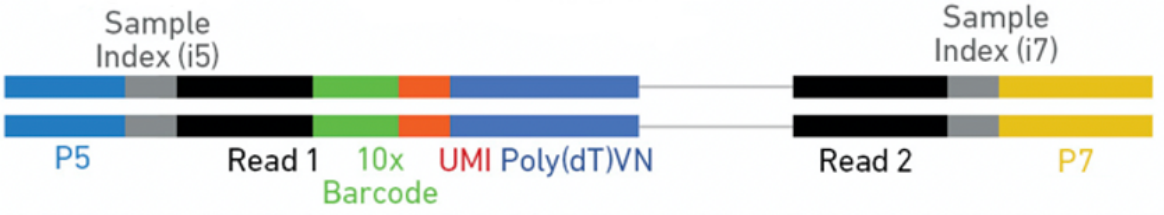
- **sample index** - identifies the library, with one or two indexes per sample
- **10x barcode** - identifies the droplet in the library
- **UMI** - identifies the transcript molecule within a cell and gene
- **insert** - the transcript molecule



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 Report

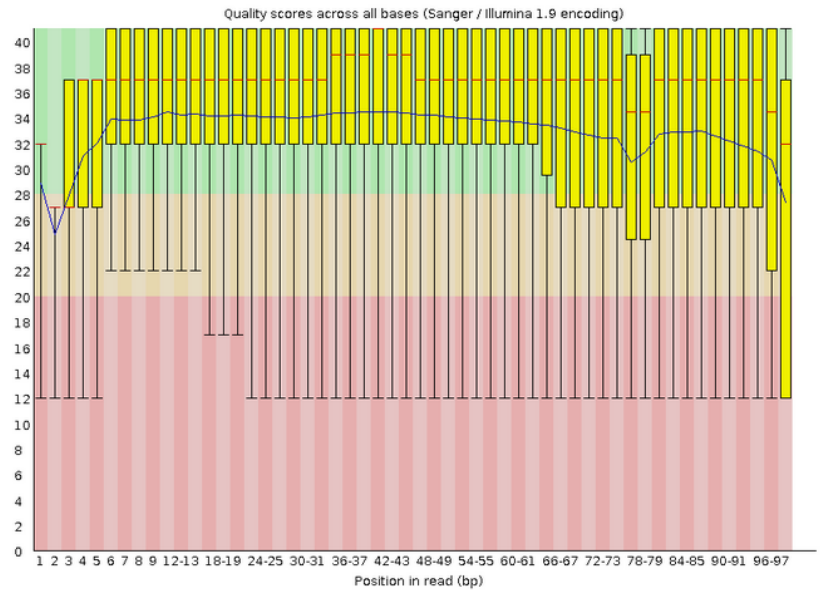
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

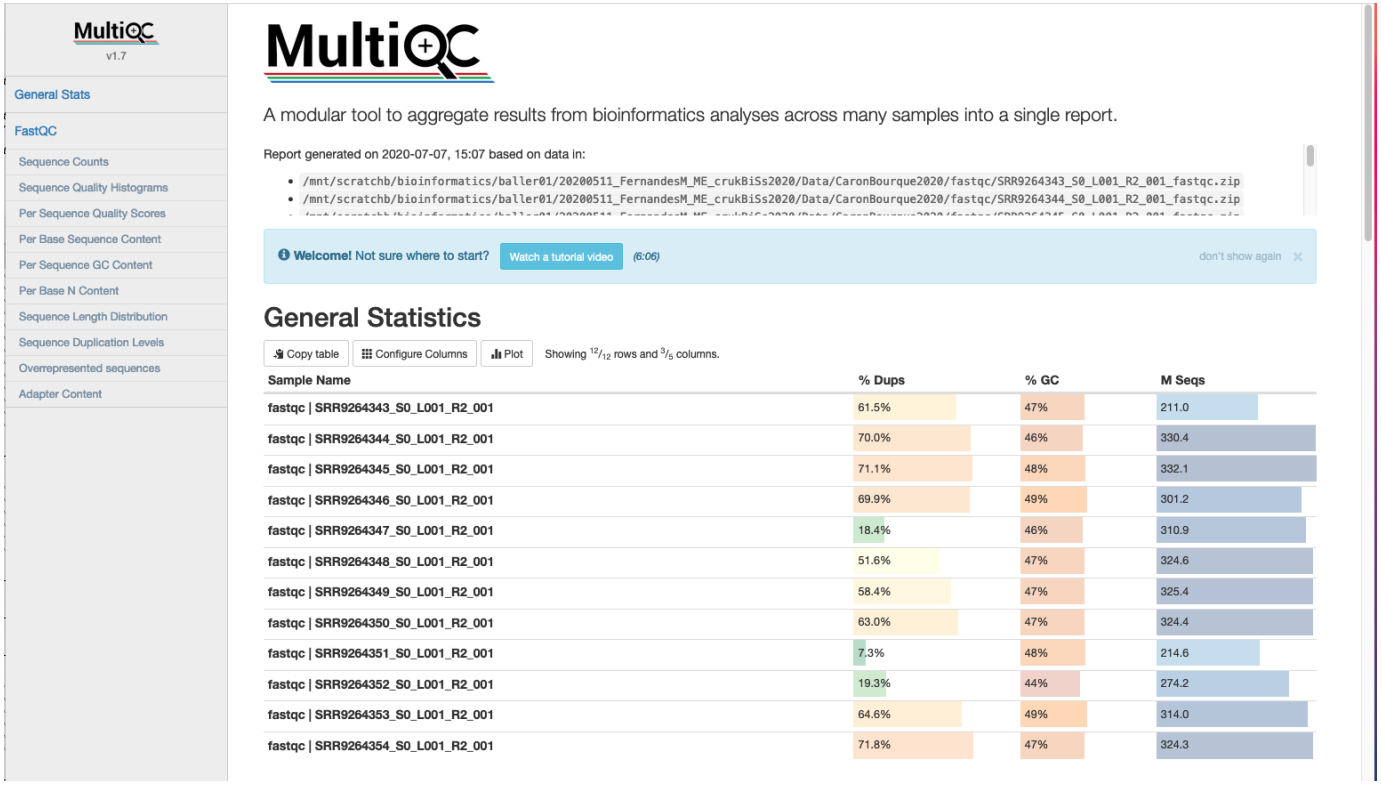
Basic Statistics

Measure	Value
Filename	SRR9264344_S0_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	330404706
Sequences flagged as poor quality	0
Sequence length	98
%GC	46

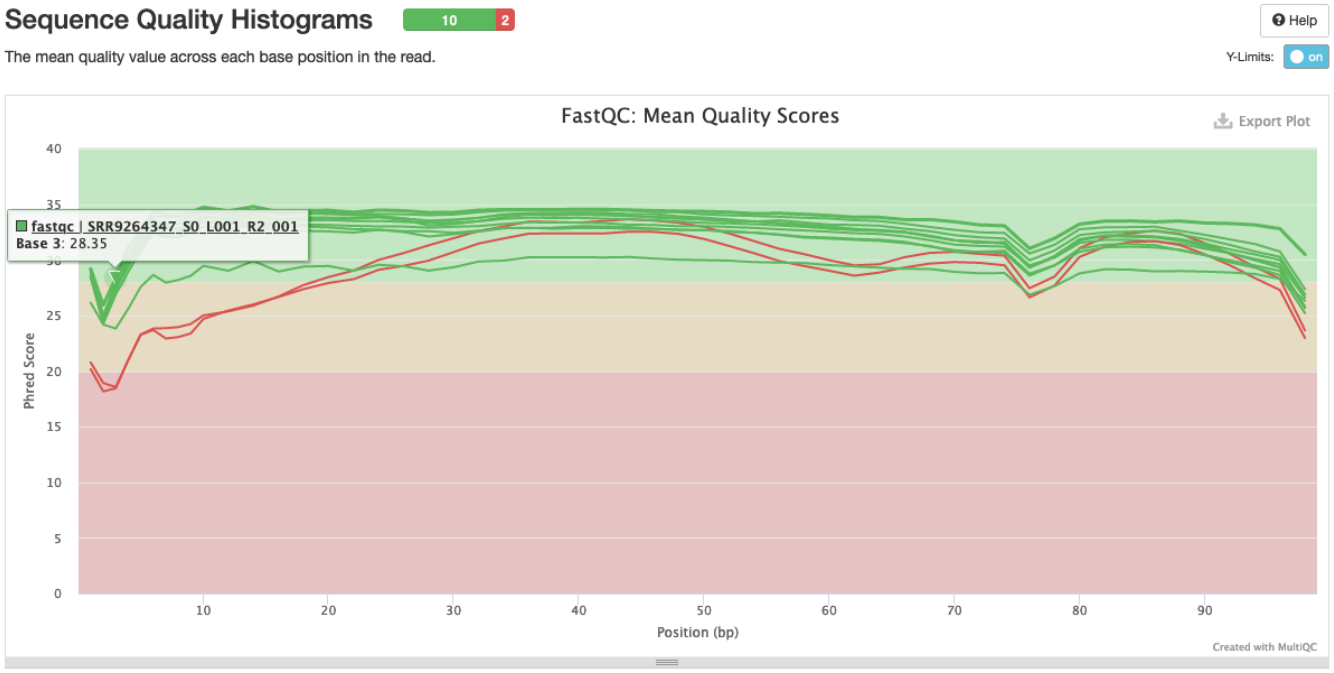
Per base sequence quality



QC of Raw Reads - MultiQC - General Statistics



QC of Raw Reads - MultiQC - Sequence Quality Histograms



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

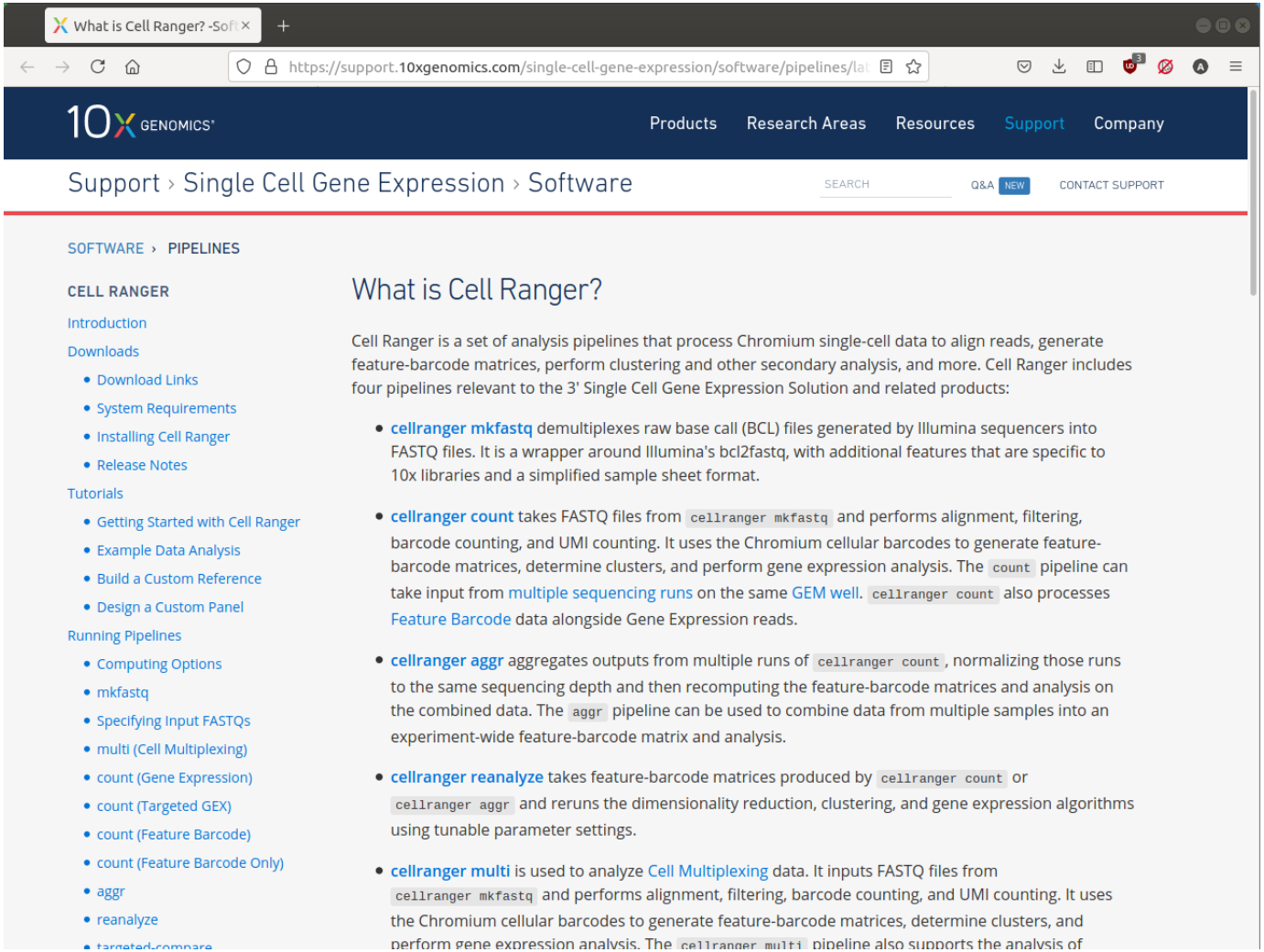
Cell Ranger

- 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



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`

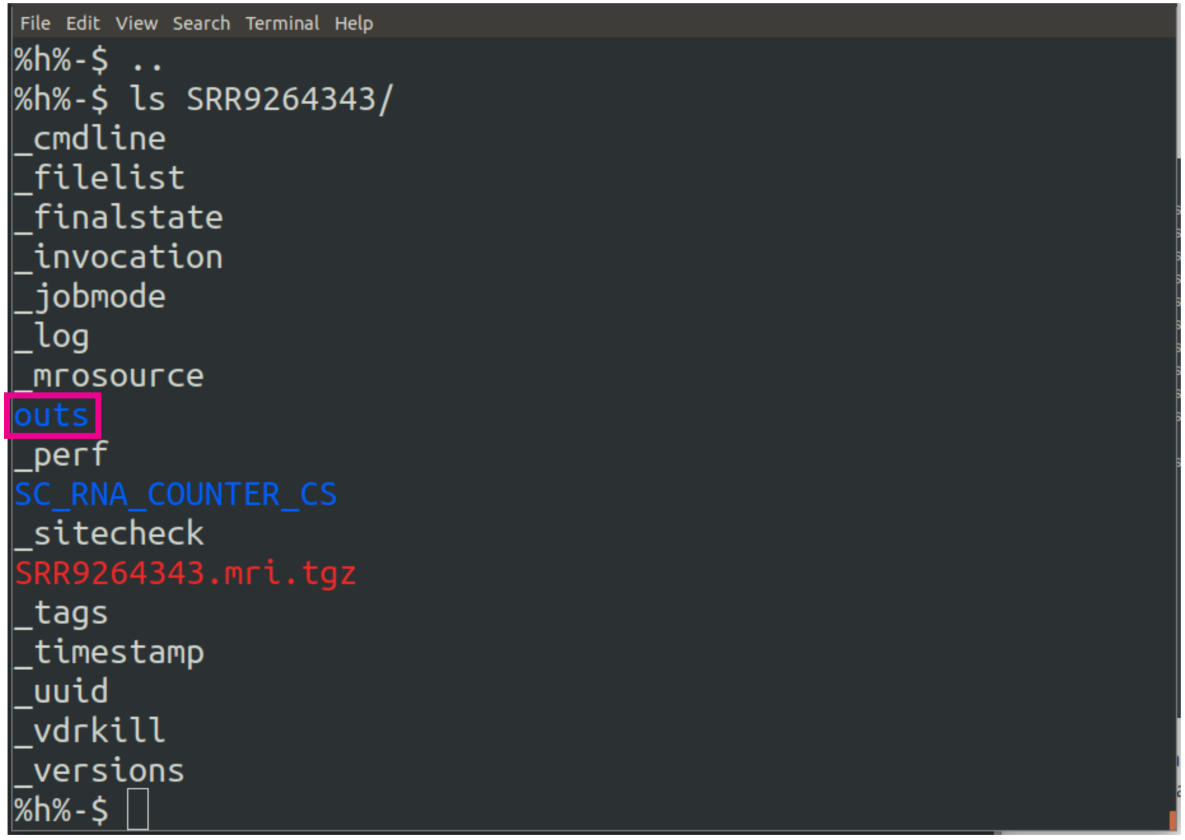
Running `cellranger count`

- Computationally very intensive
- High memory requirements

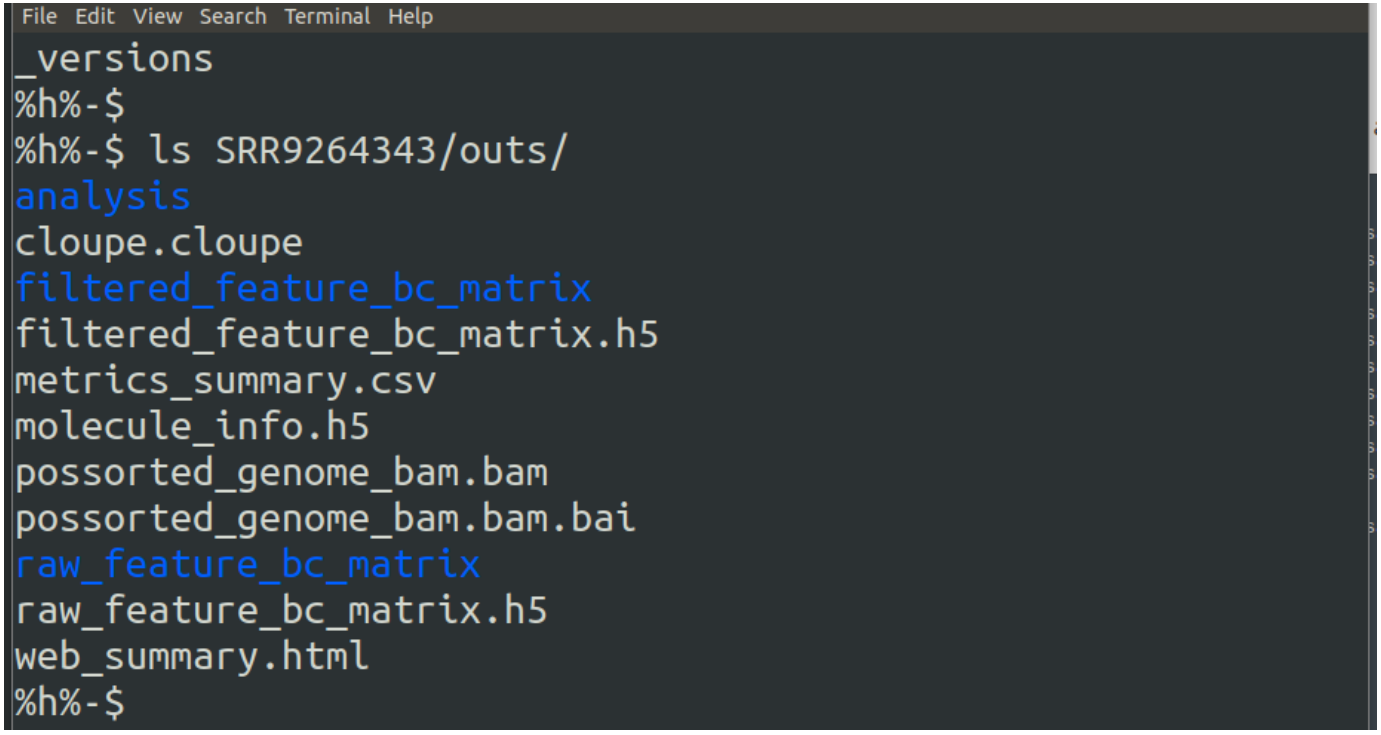
```
File  Edit  View  Search  Terminal  Help
%h%-$
%h%-$
%h%-$ cellranger count --id=SRR9264343 \
>                      --transcriptome=refdata-gex-mm10-2020-A \
>                      --fastqs=fastq \
>                      --sample=SRR9264343 \
>                      --localcores=8 \
>                      --localmem=64
```

Cell Ranger outputs

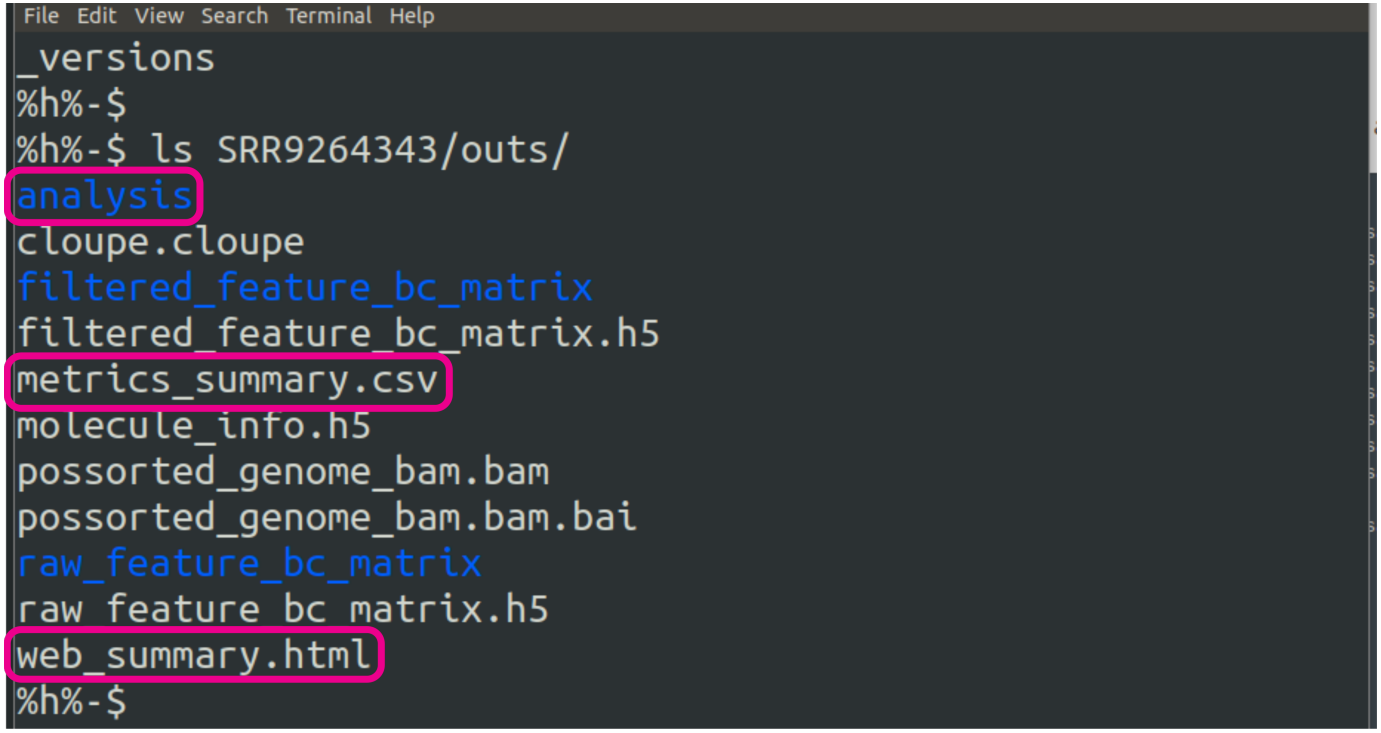
- One directory per sample



Cell Ranger outputs



Cell Ranger outputs



Cell Ranger report

10x GENOMICS Cell Ranger • count

SITTA6

Summary | Analysis

14,668
Estimated Number of Cells

20,065
Mean Reads per Cell

1,344
Median Genes per Cell

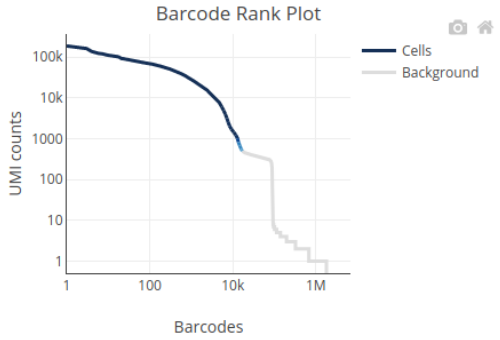
Sequencing

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

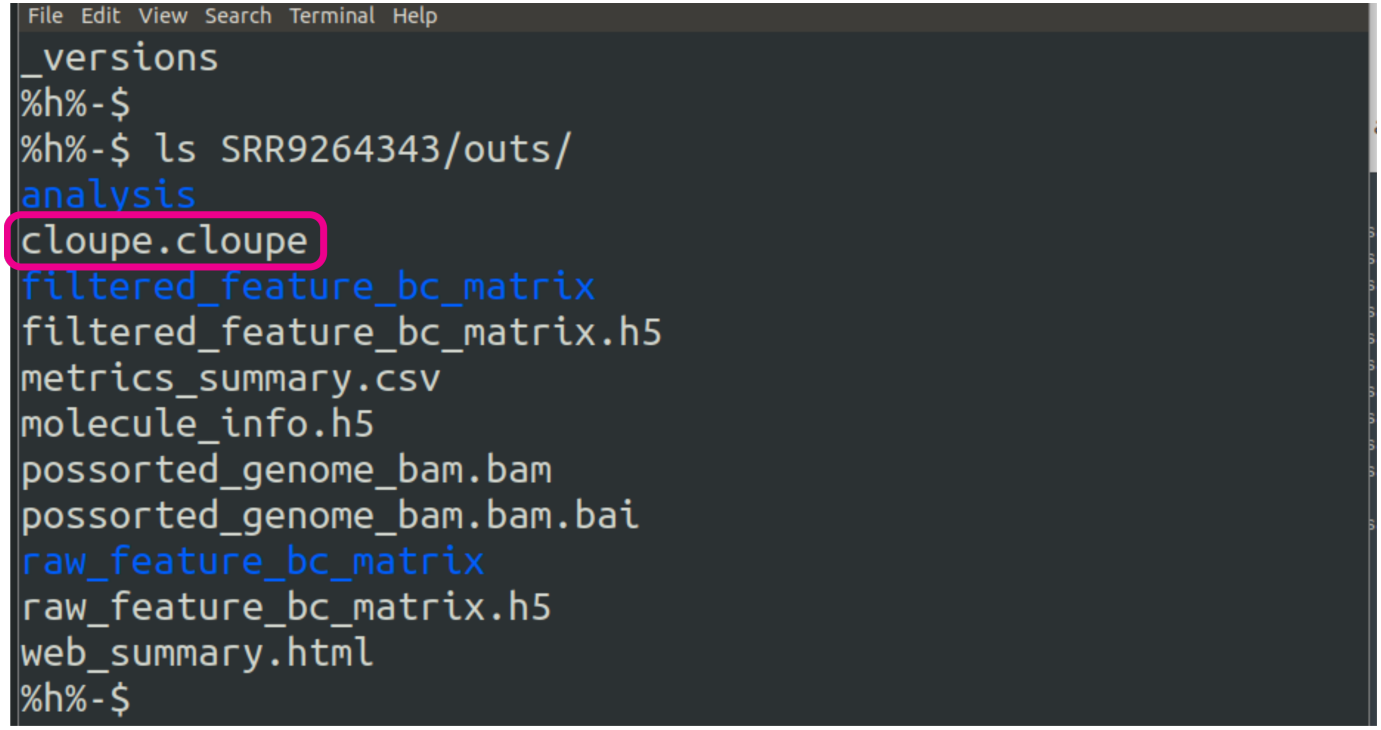


Estimated Number of Cells	14,668
Fraction Reads in Cells	80.8%
Mean Reads per Cell	20,065
Median Genes per Cell	1,344
Total Genes Detected	23,106
Median UMI Counts per Cell	2,928

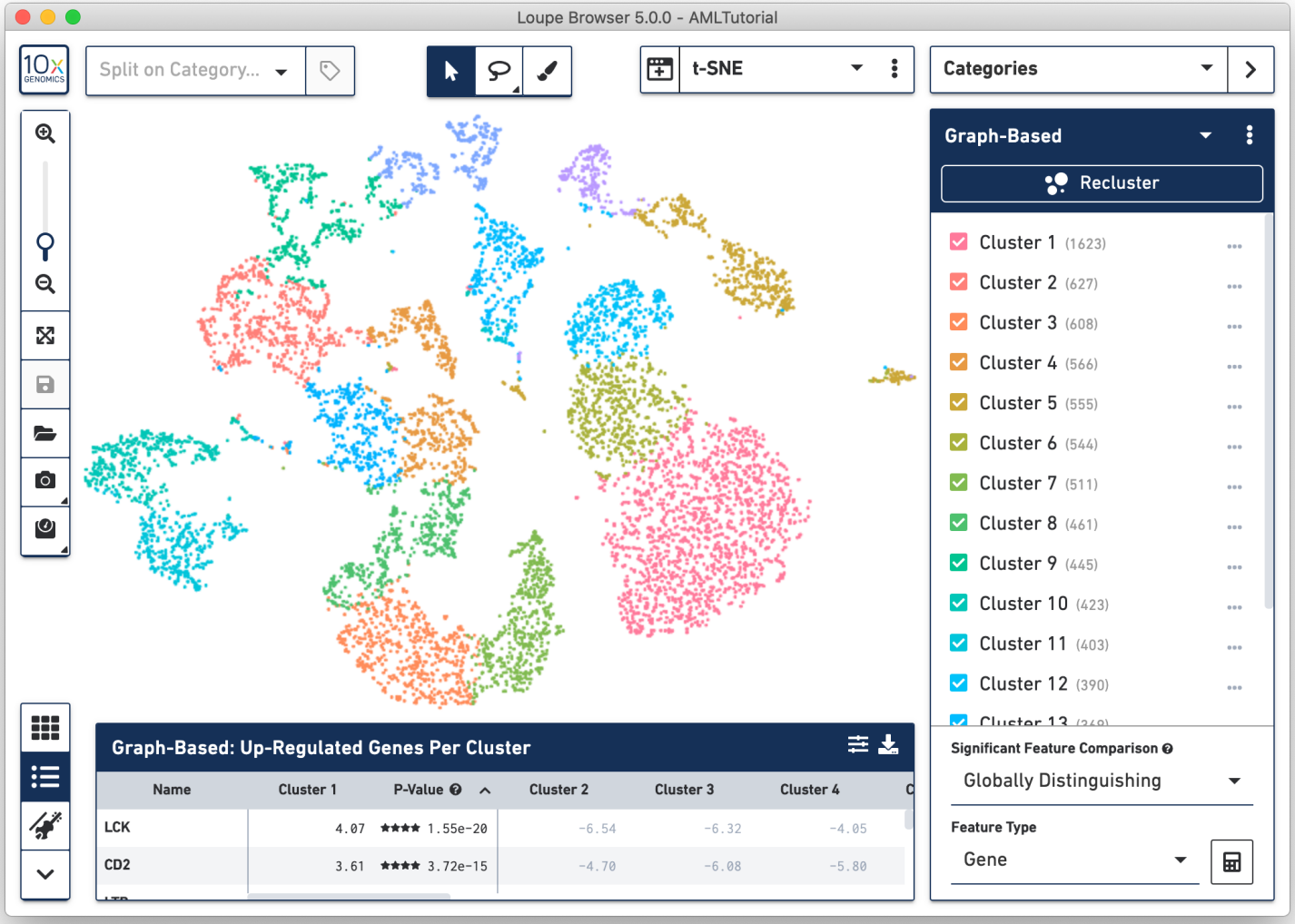
Sample

Sample ID	SITTA6
Sample Description	

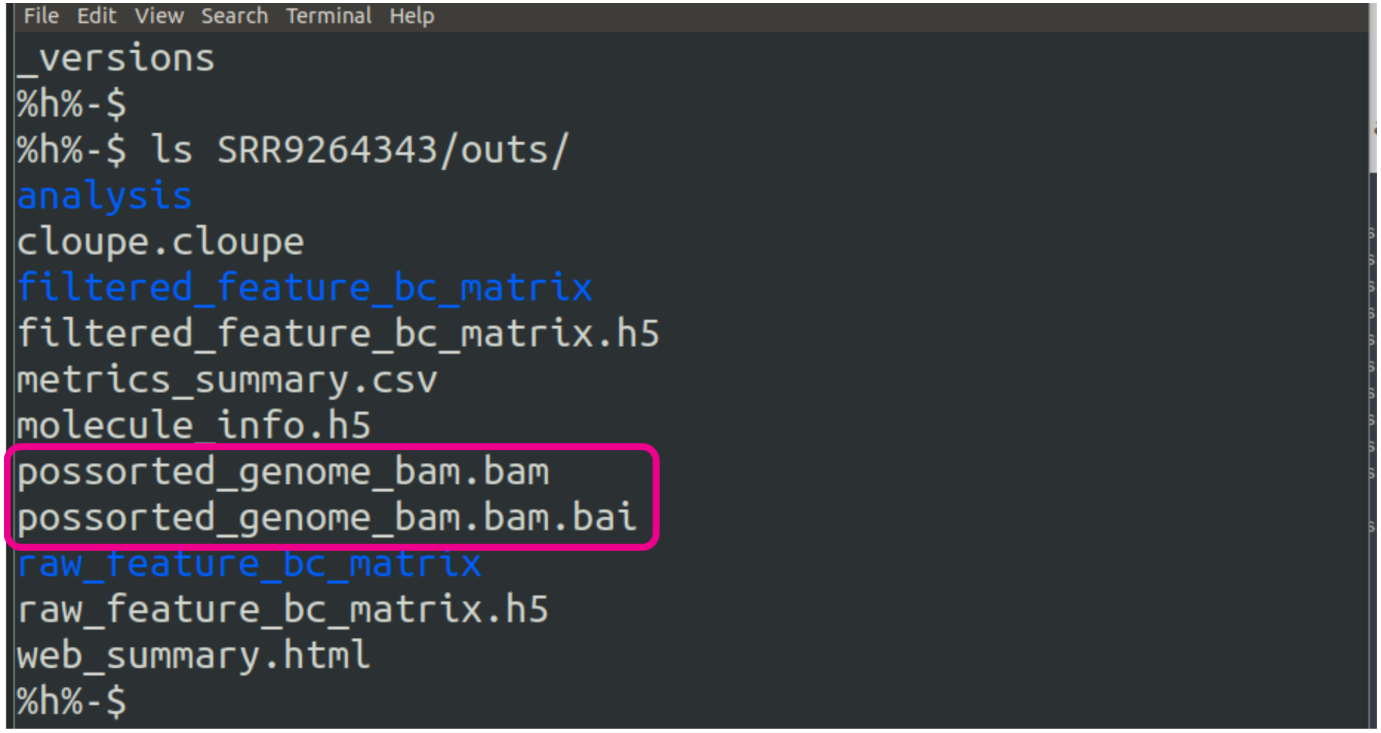
Cell Ranger outputs



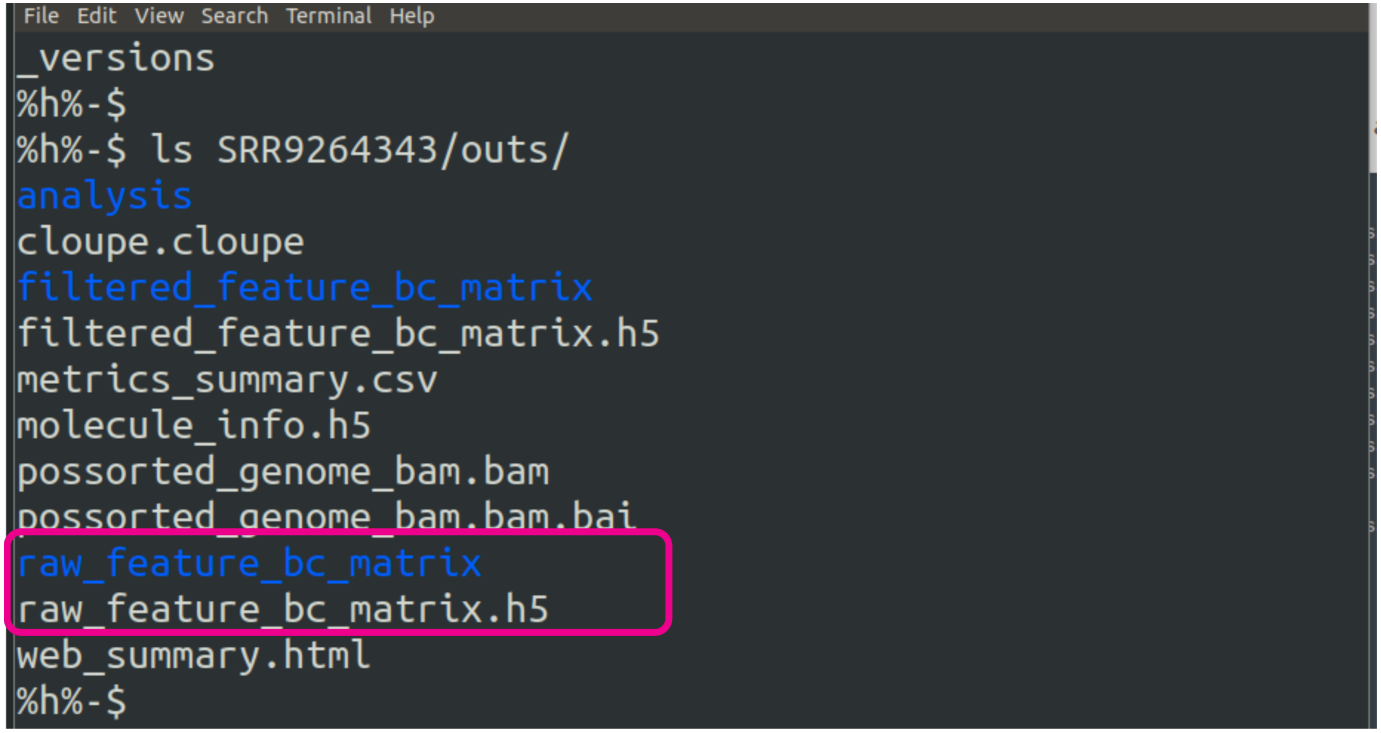
Loupe Browser



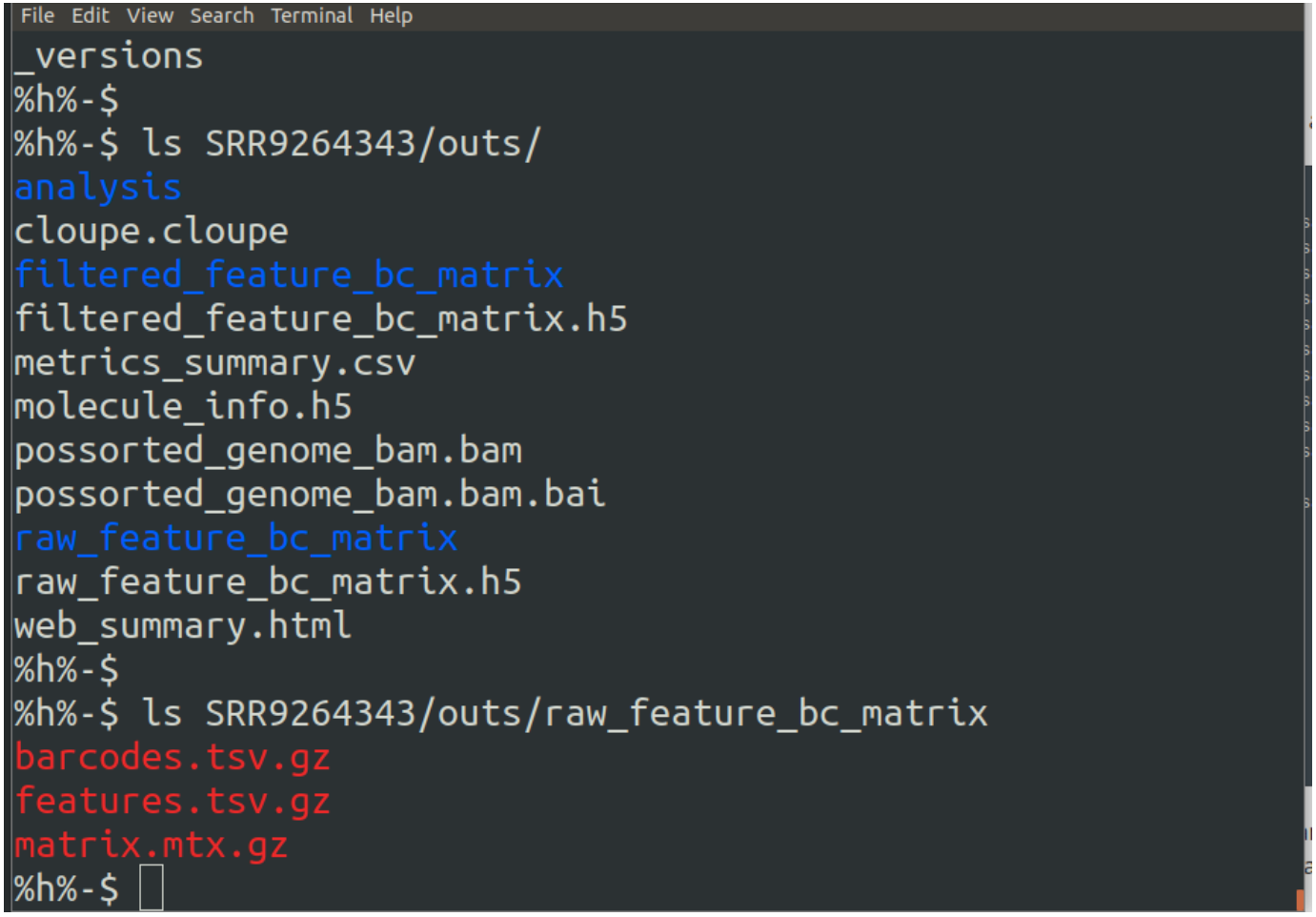
Cell Ranger outputs



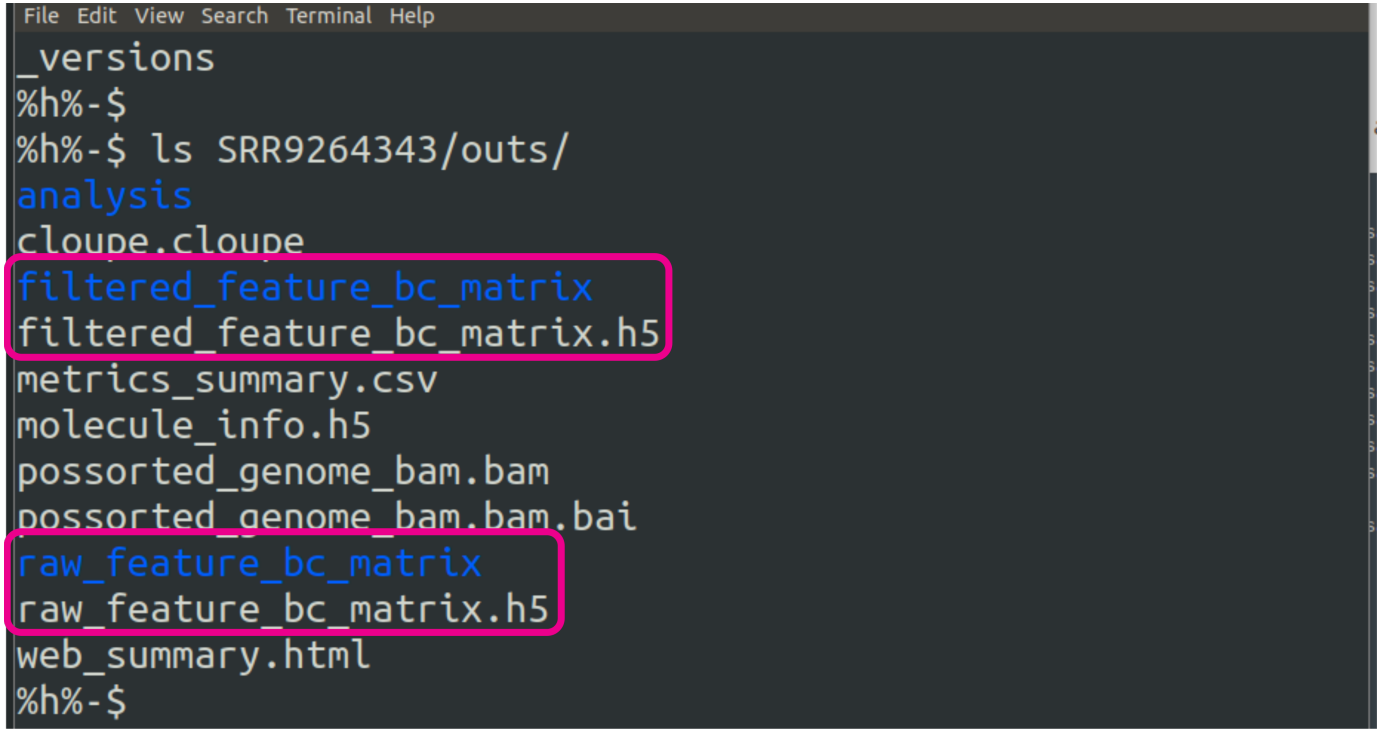
Cell Ranger outputs



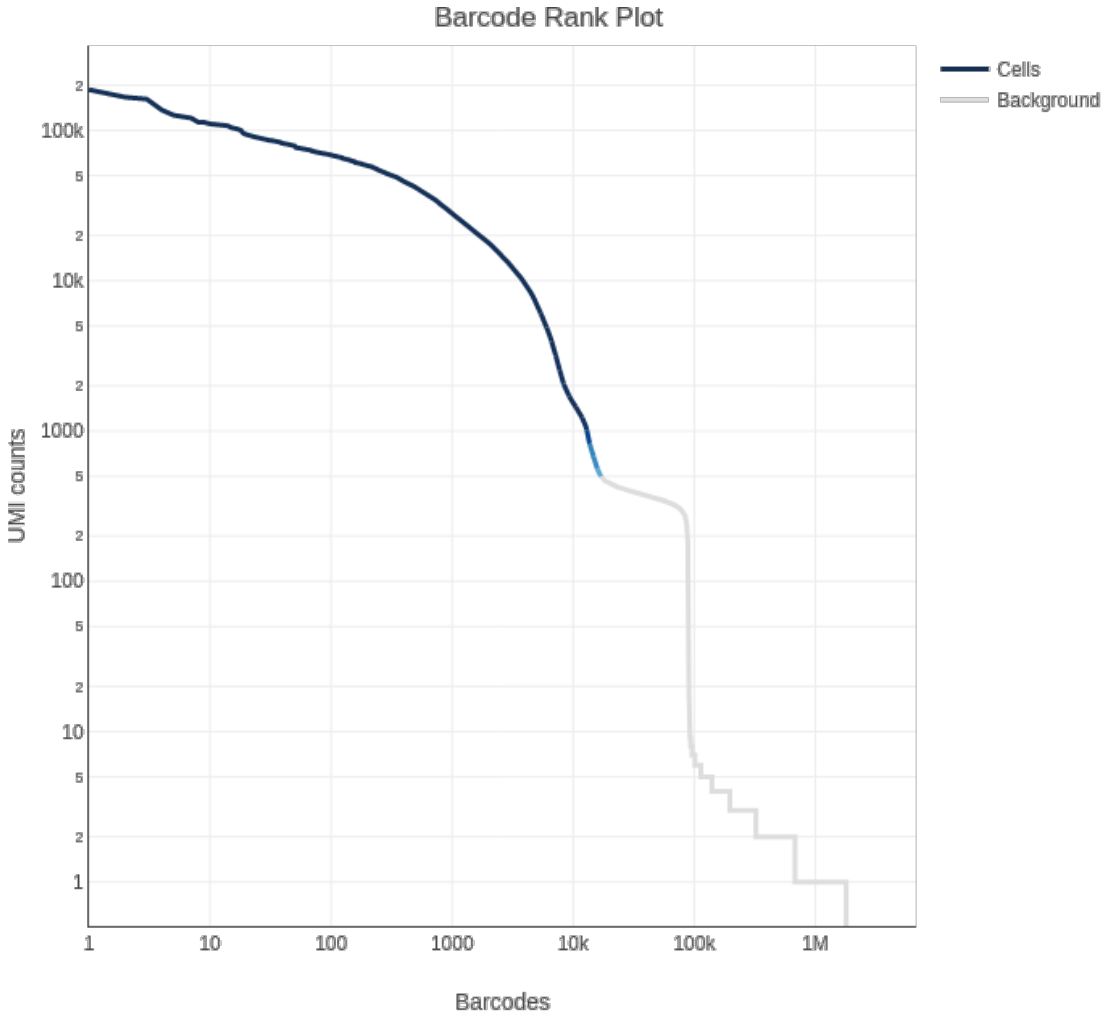
Cell Ranger outputs



Cell Ranger outputs



Cell Ranger cell calling



Single Cell RNAseq Analysis Workflow