Alignment and feature counting

November 2025

Single Cell RNAseq Analysis Workflow

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

Single Cell RNAseq Analysis Workflow

10x single-cell isolation

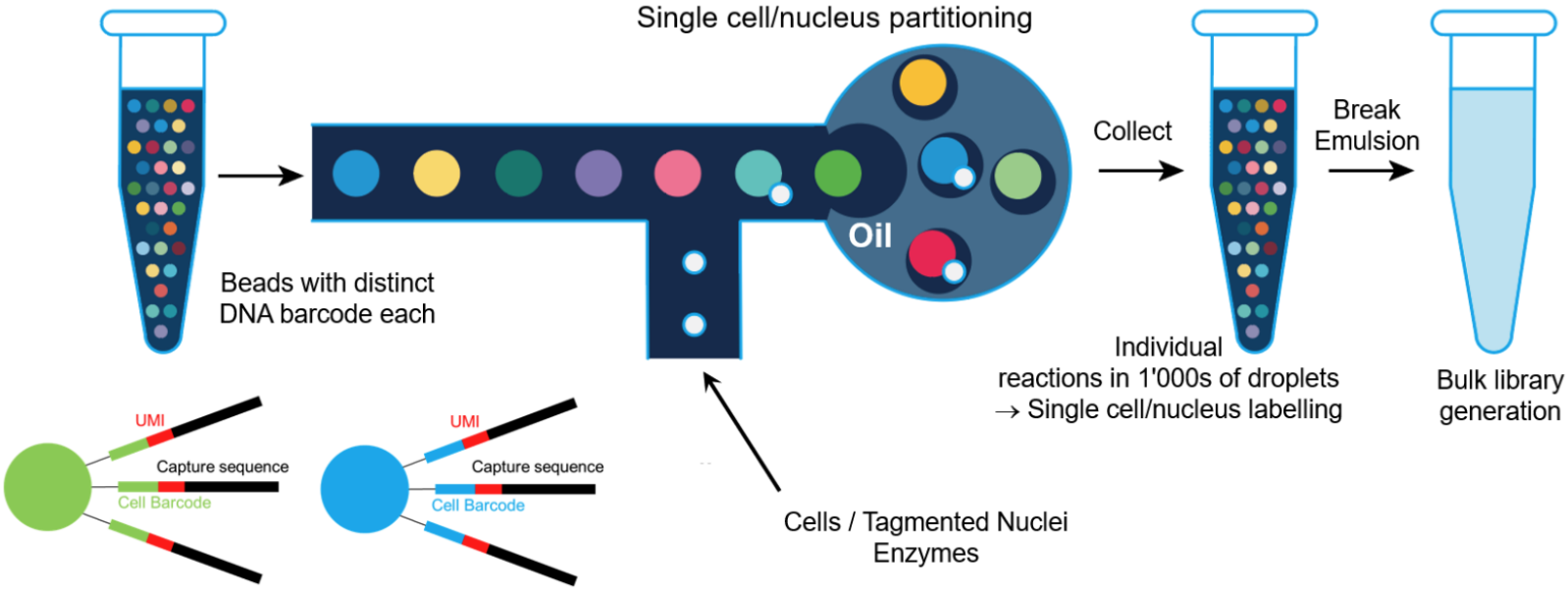
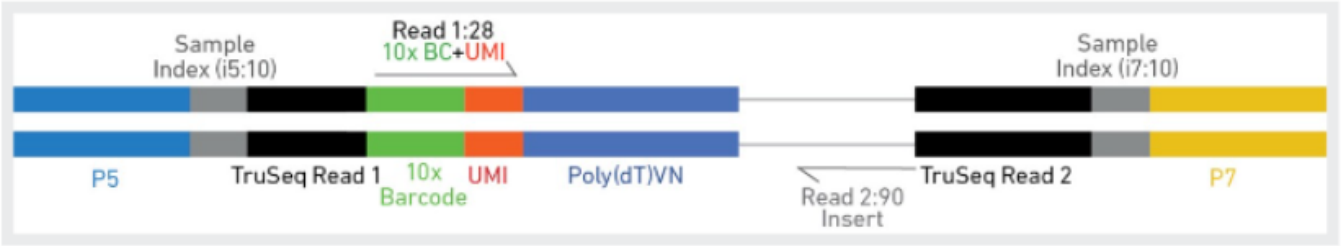


Image from https://wp.unil.ch/gtf/single-cell-experiments-10x/

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 (BC)** - identifies the droplet in the library
- **UMI (Unique Molecular Identifier)** - identifies the transcript molecule within a cell and gene
- **insert** - the transcript molecule



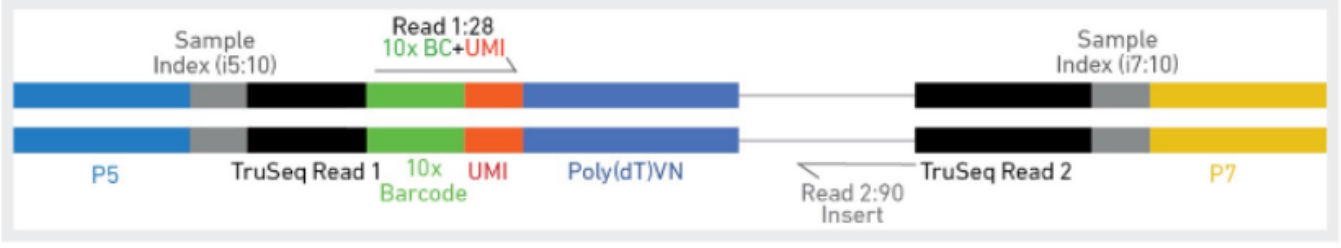
- TruSeq Read 1: Read 1 primer
- TruSeq Read 2: Read 2 primer
- P5 and P7 are short DNA adapter sequences that allow the library to bind to the Illumina flow cell

Image from 10x Genomics

Raw fastq files

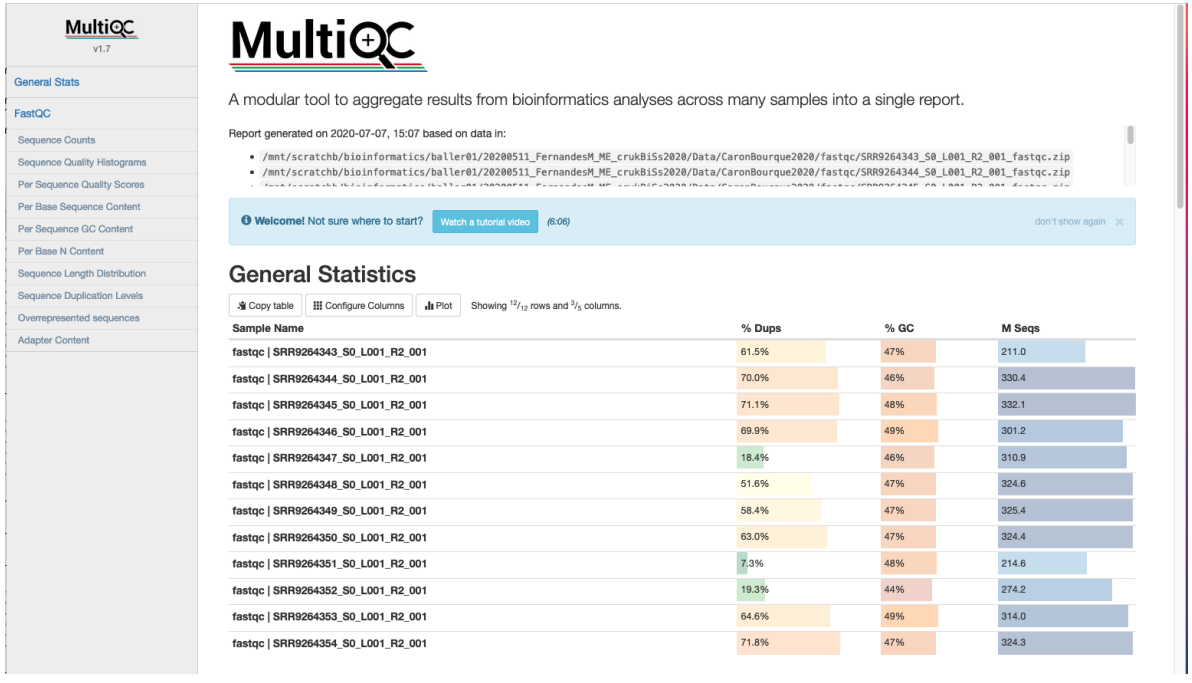
The sequences for any given fragment will generally be delivered in 3 or 4 files:

- **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

QC of Raw Reads - MultiQC - General Statistics



MultiQC
v1.7

General Stats

FastQC

- Sequence Counts
- Sequence Quality Histograms
- 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

MultiQC

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

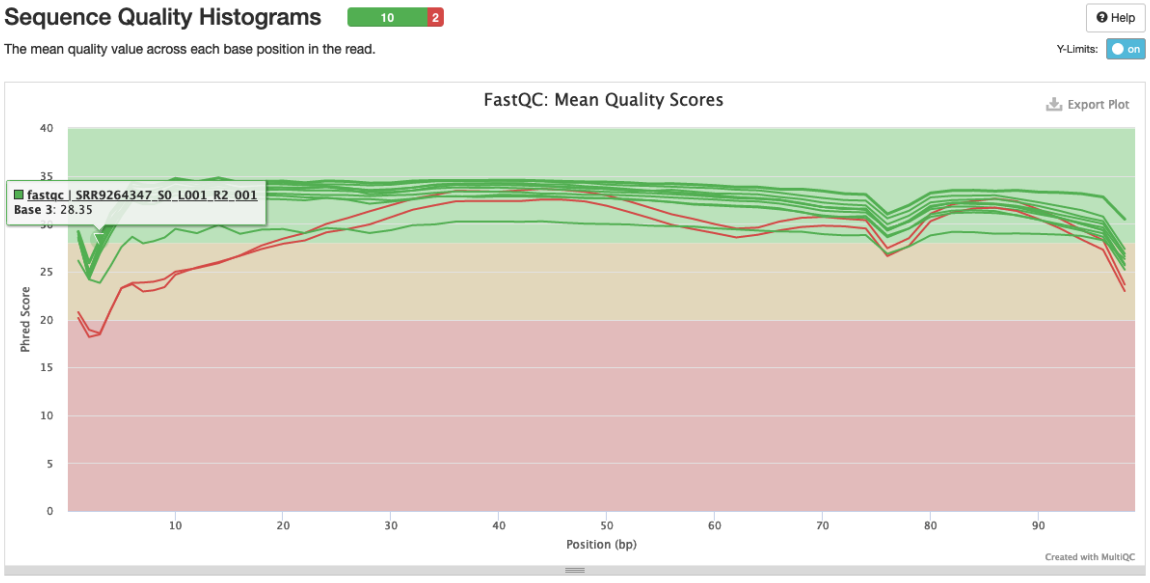
Welcome! Not sure where to start? Watch a tutorial video *(6:06)* don't show again

General Statistics

Copy table | Configure Columns | Plot | 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



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

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}
```

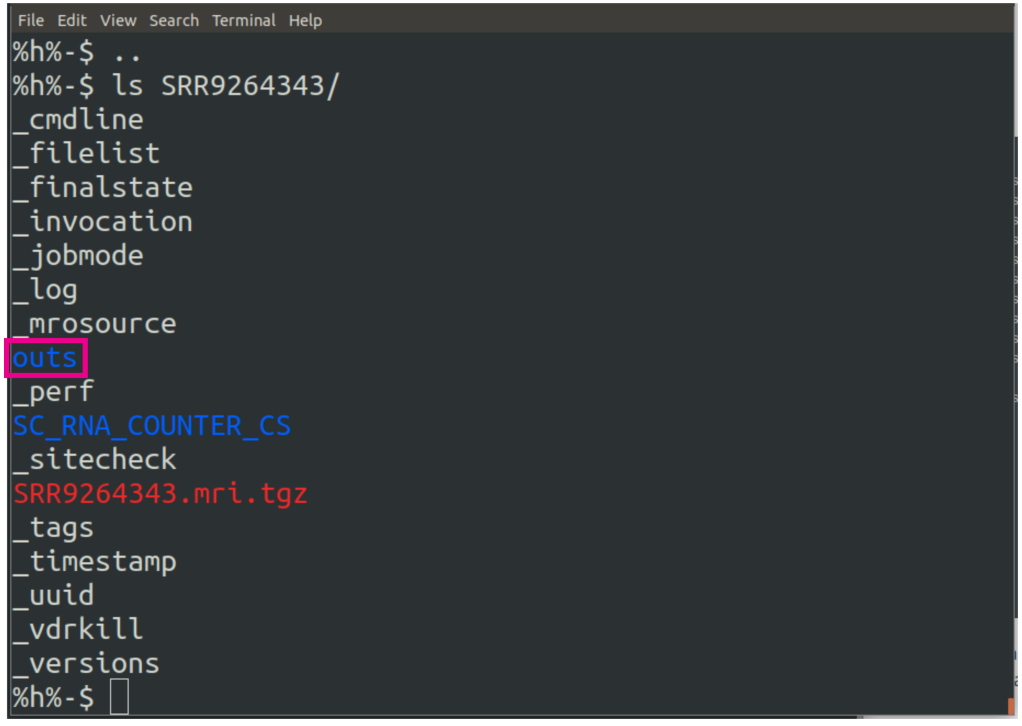
Running `cellranger count`

- Computationally very intensive
- High memory requirements

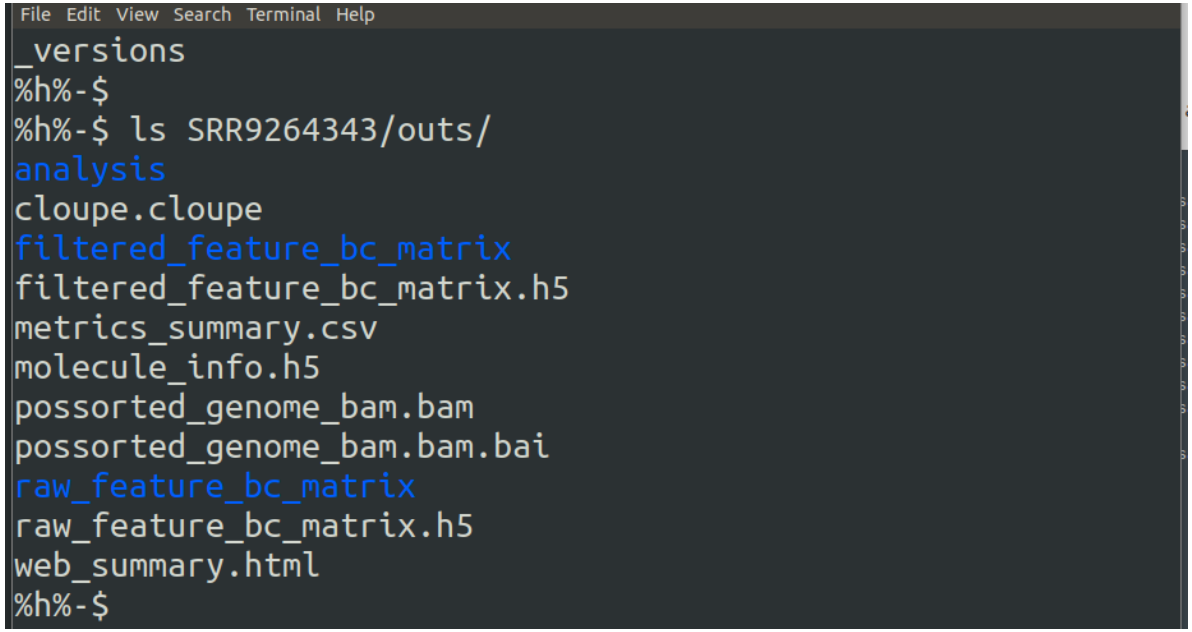
```
cellranger count \
  --id={OUTPUT_SAMPLE_NAME} \
  --transcriptome={DIRECTORY_WITH_REFERENCE} \
  --fastqs={DIRECTORY_WITH_FASTQ_FILES} \
  --sample={NAME_OF_SAMPLE_IN_FASTQ_FILES} \
  --localcores={NUMBER_OF_CPUS} \
  --localmem={RAM_MEMORY}
```

Cell Ranger outputs

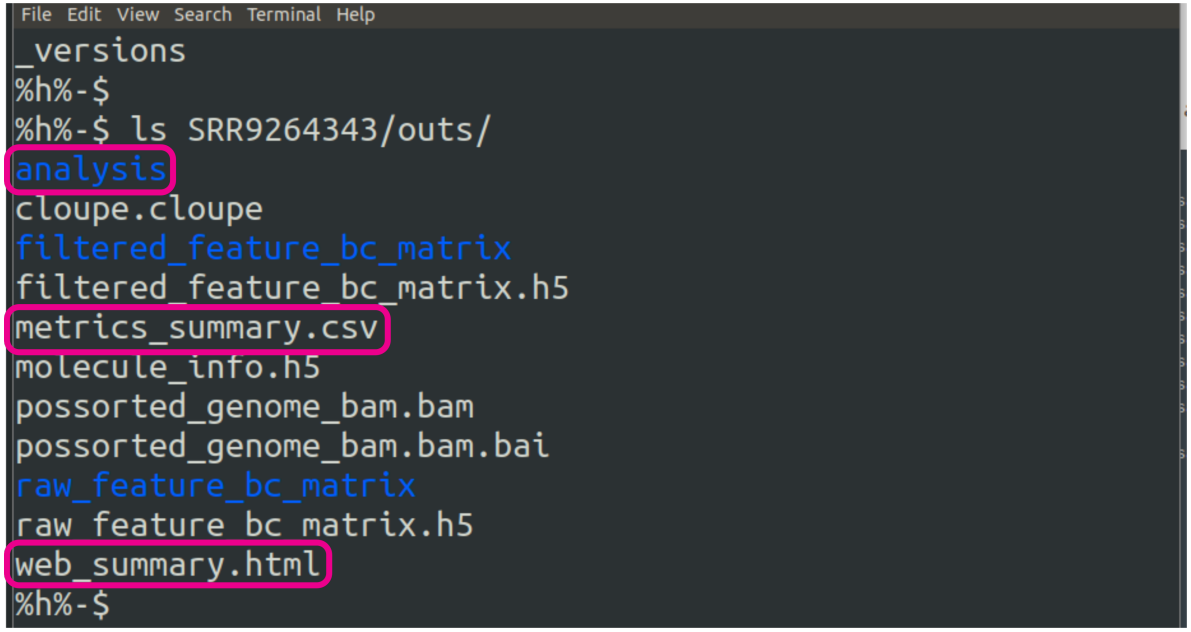
- One directory per sample



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

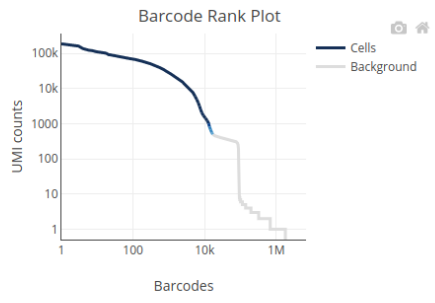
Sequencing

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

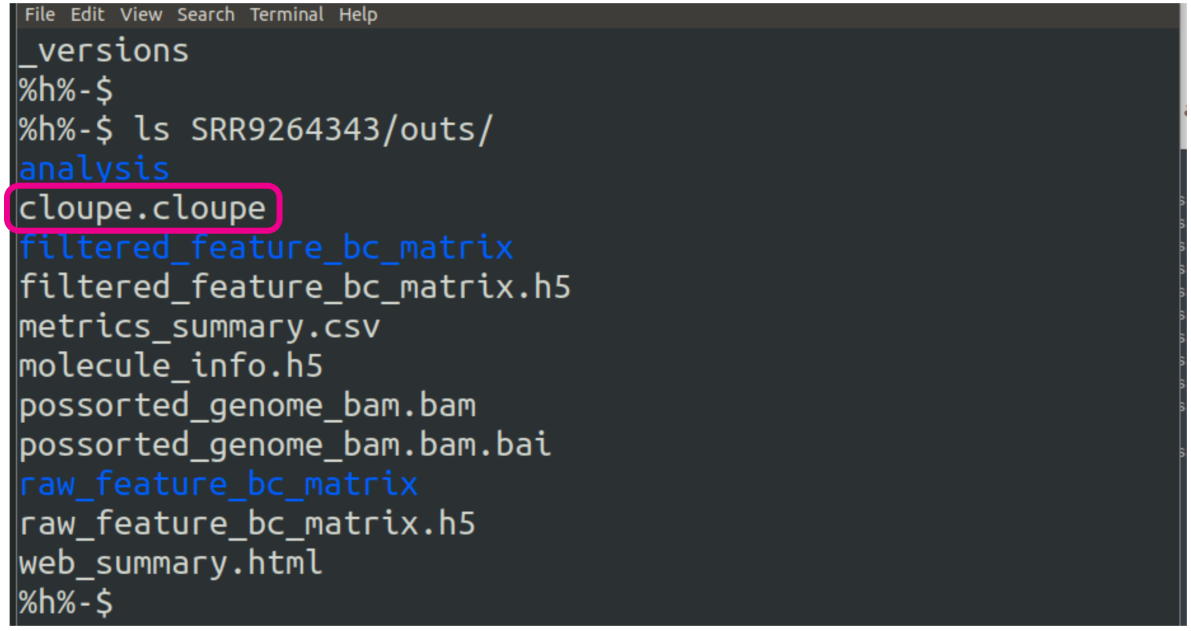


Estimated Number of Cells	14,668
Fraction Reads in Cells	80.8%
Mean Reads per Cell	20,065
Median Genes per Cell	1,344
Total Genes Detected	23,106
Median UMI Counts per Cell	2,928

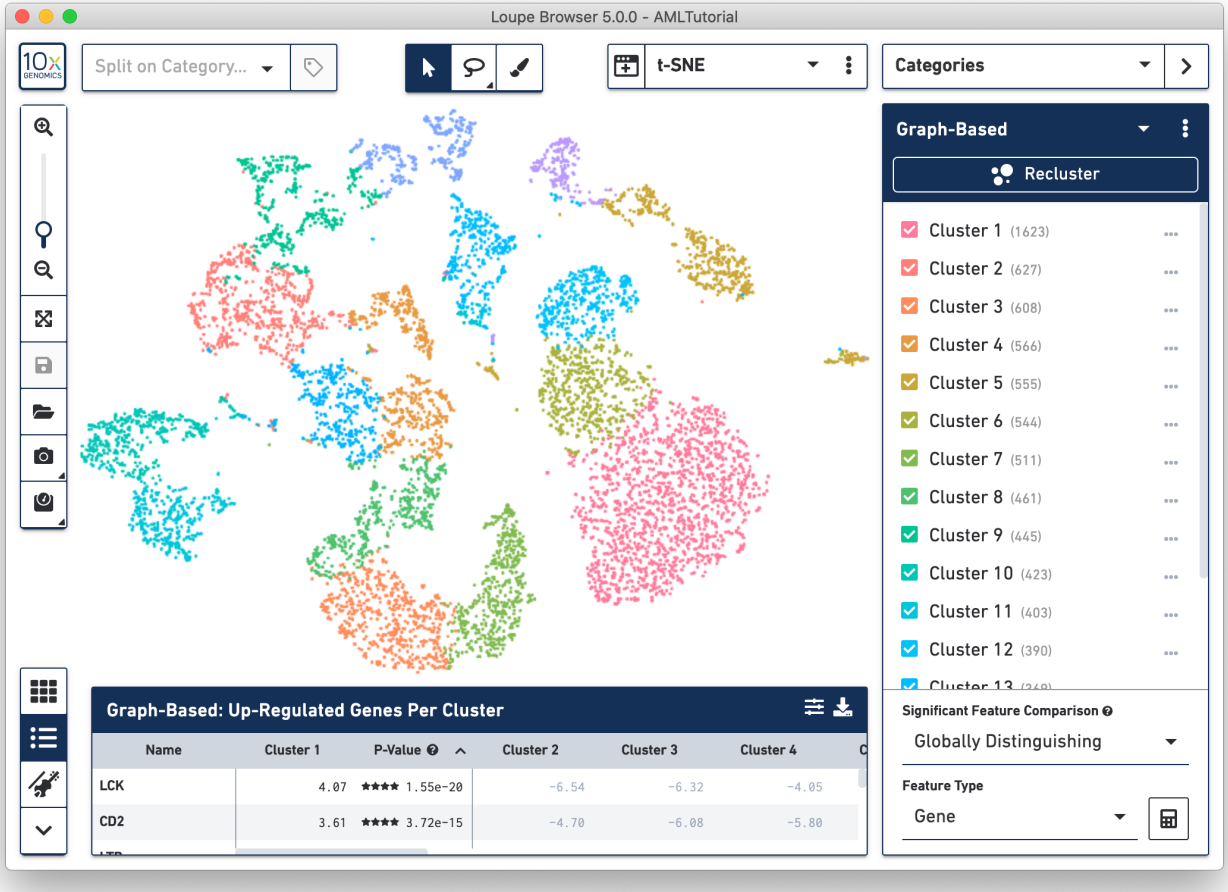
Sample

Sample ID	SITTA6
Sample Description	

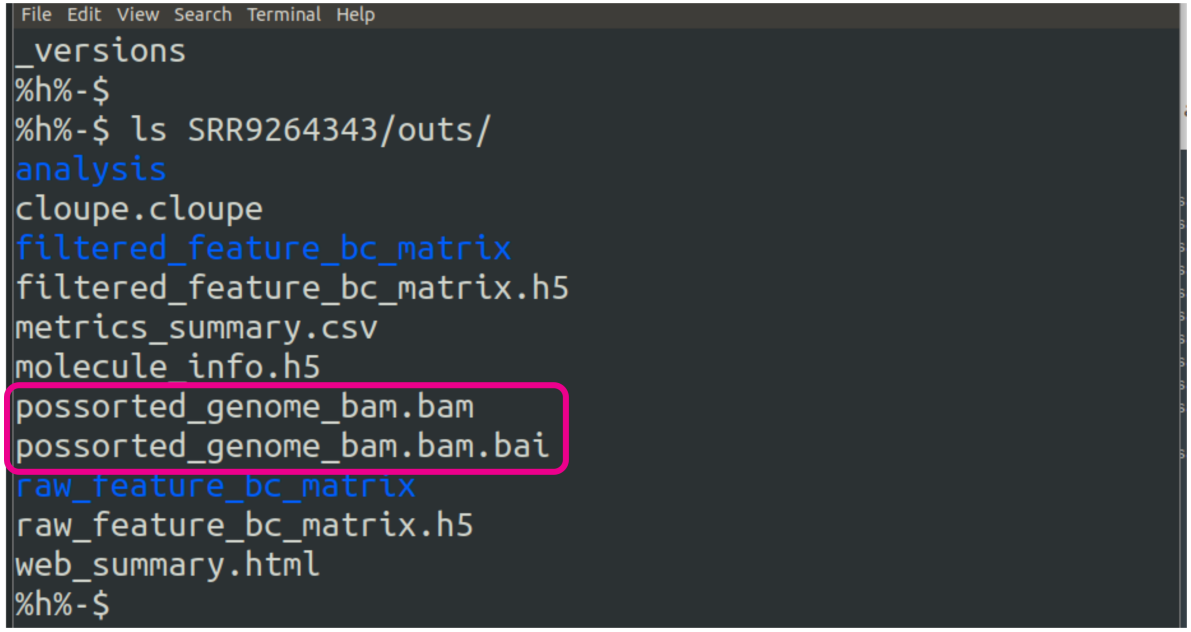
Cell Ranger outputs



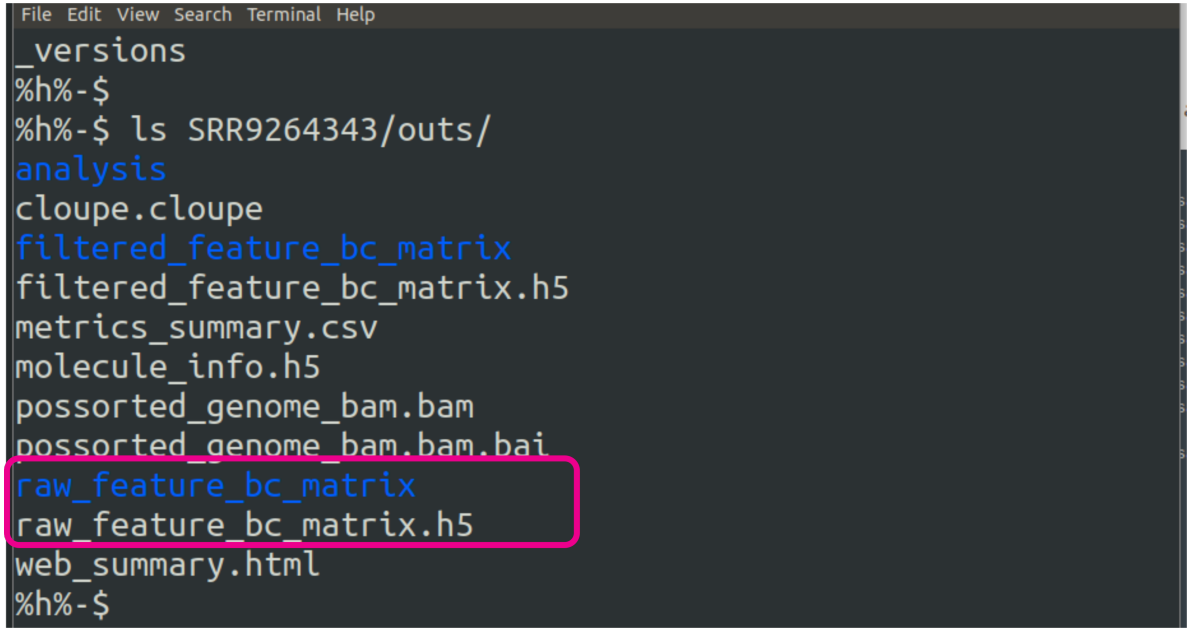
Loupe Browser



Cell Ranger outputs



Cell Ranger outputs



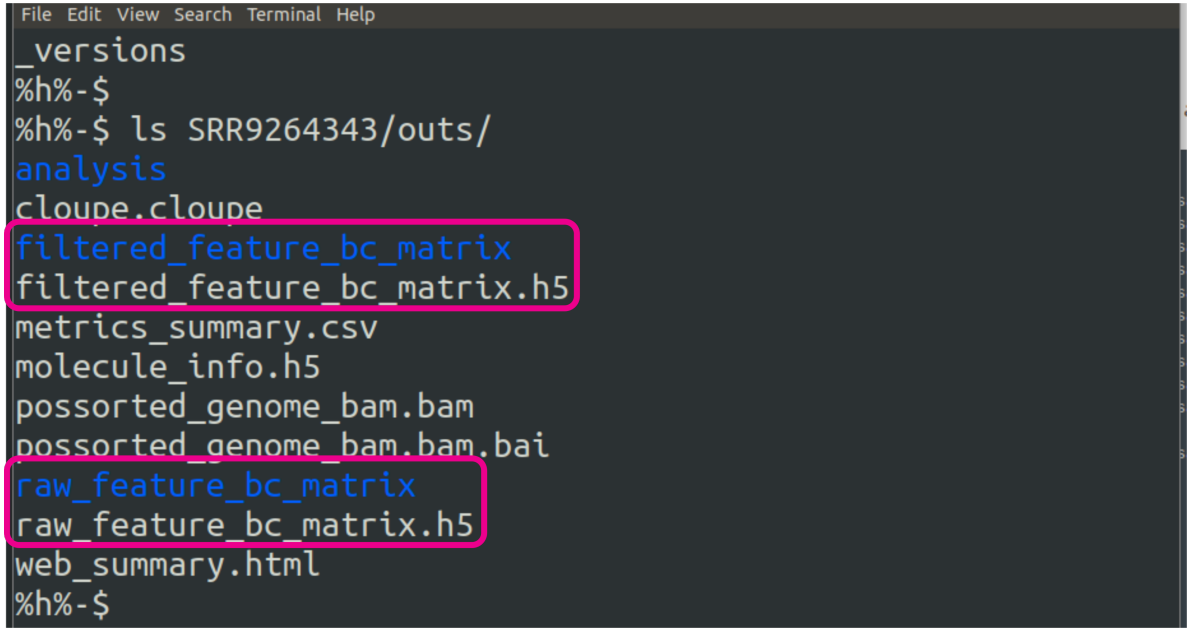
Cell Ranger outputs

Two types of outputs:

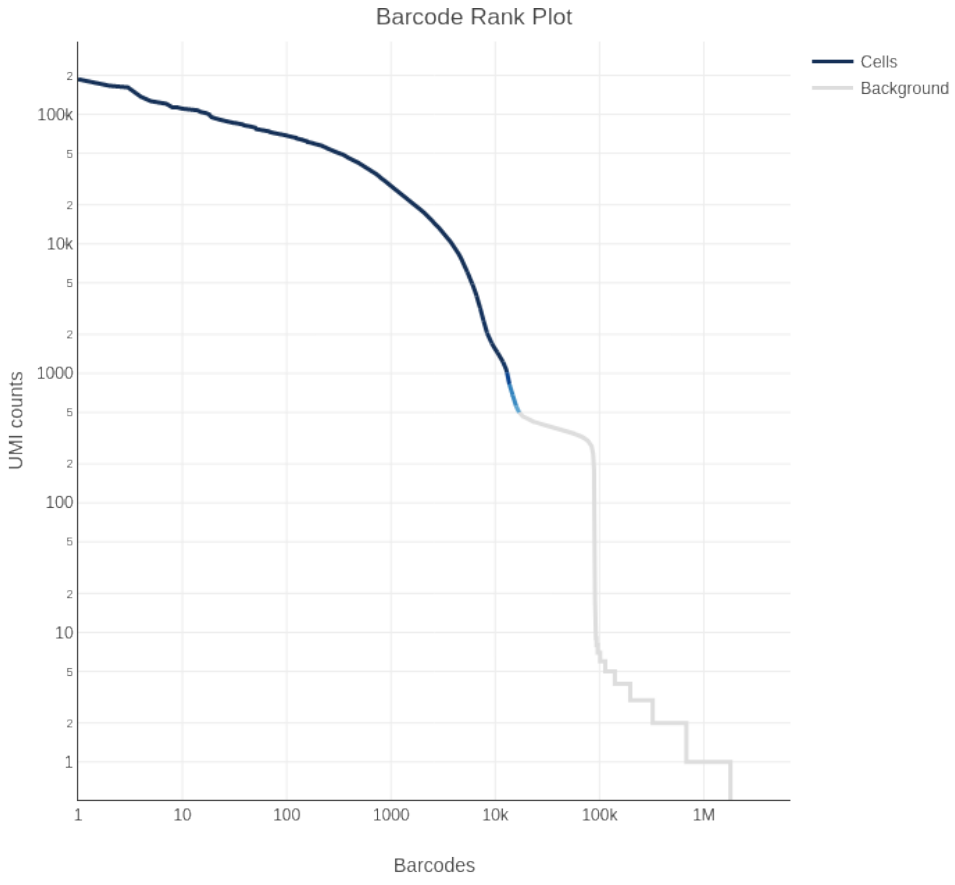
- Text-based files: `.tsv` and `.mtx`
- HDF5 files: `.h5`

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

Cell Ranger outputs



Cell Ranger cell calling



Single Cell RNAseq Analysis Workflow