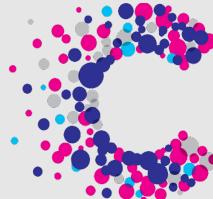




UNIVERSITY OF
CAMBRIDGE



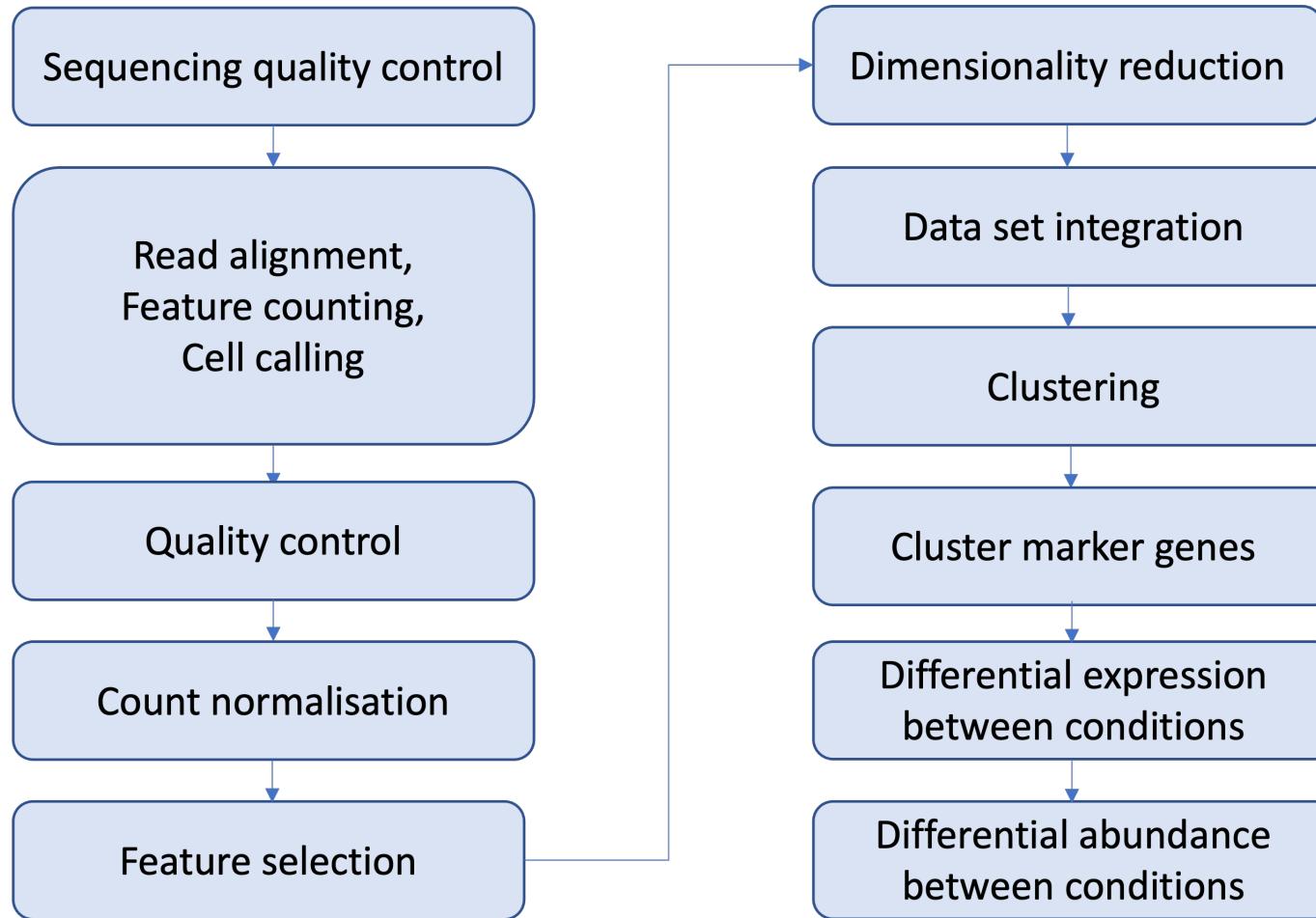
CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Alignment and feature counting

September 2022

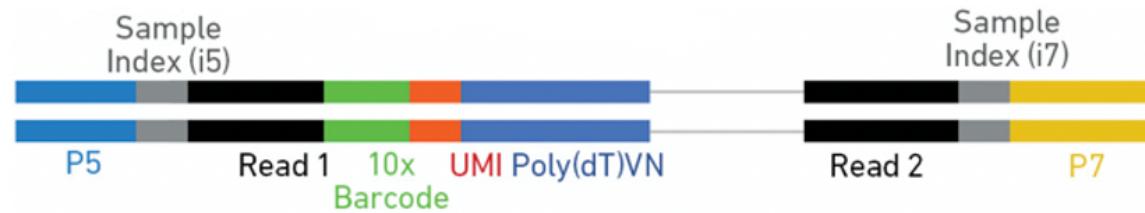
Single Cell RNAseq Analysis Workflow



10x library file structure

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

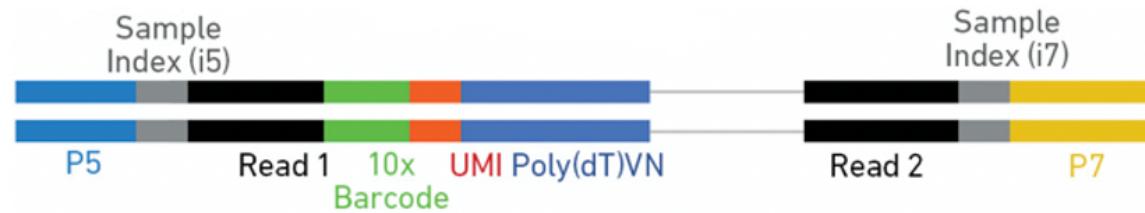
- **sample index** - identifies the library, with one or two indexes per sample
- **10x barcode** - identifies the droplet in the library
- **UMI** - identifies the transcript molecule within a cell and gene
- **insert** - the transcript molecule



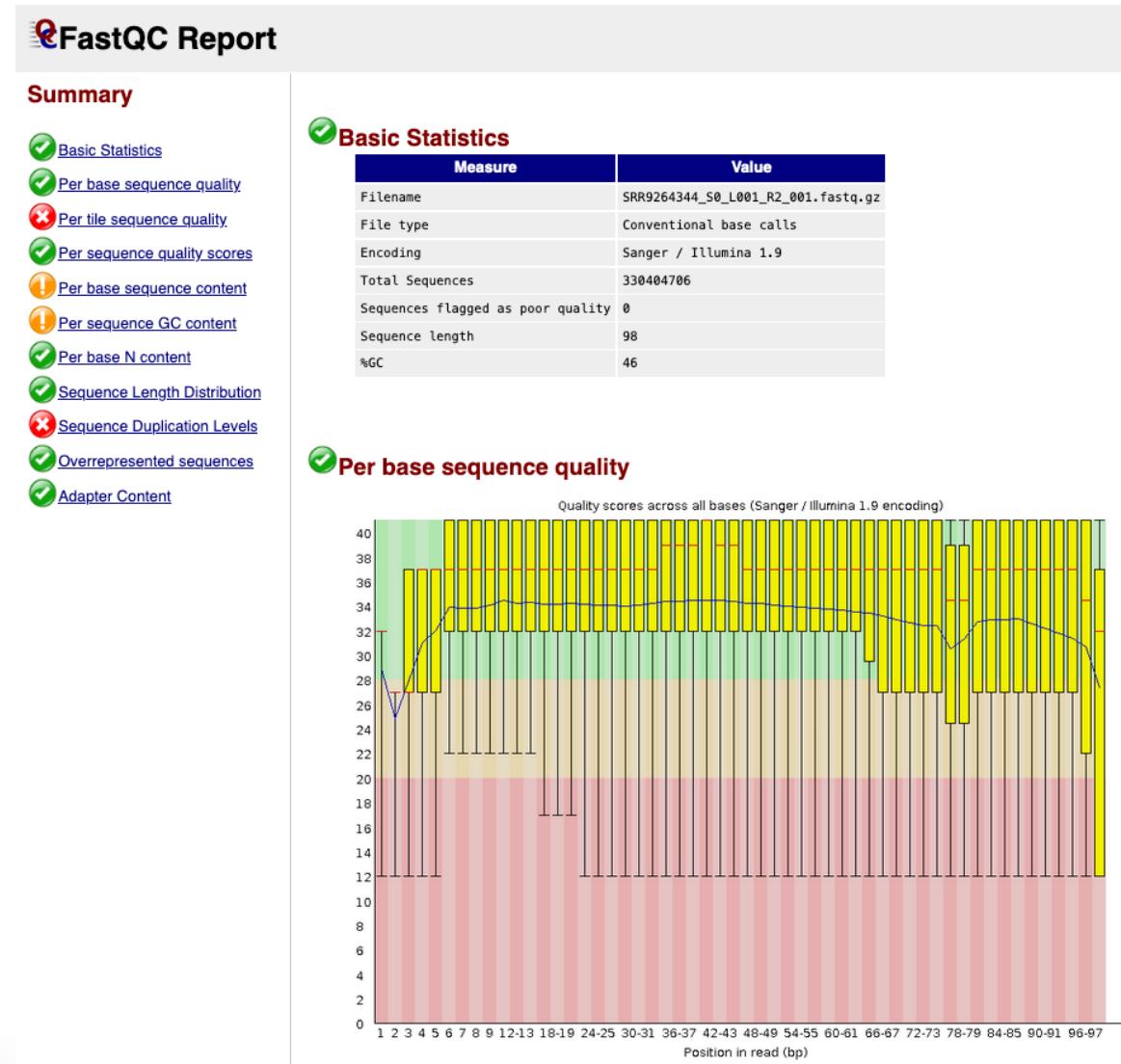
Raw fastq files

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

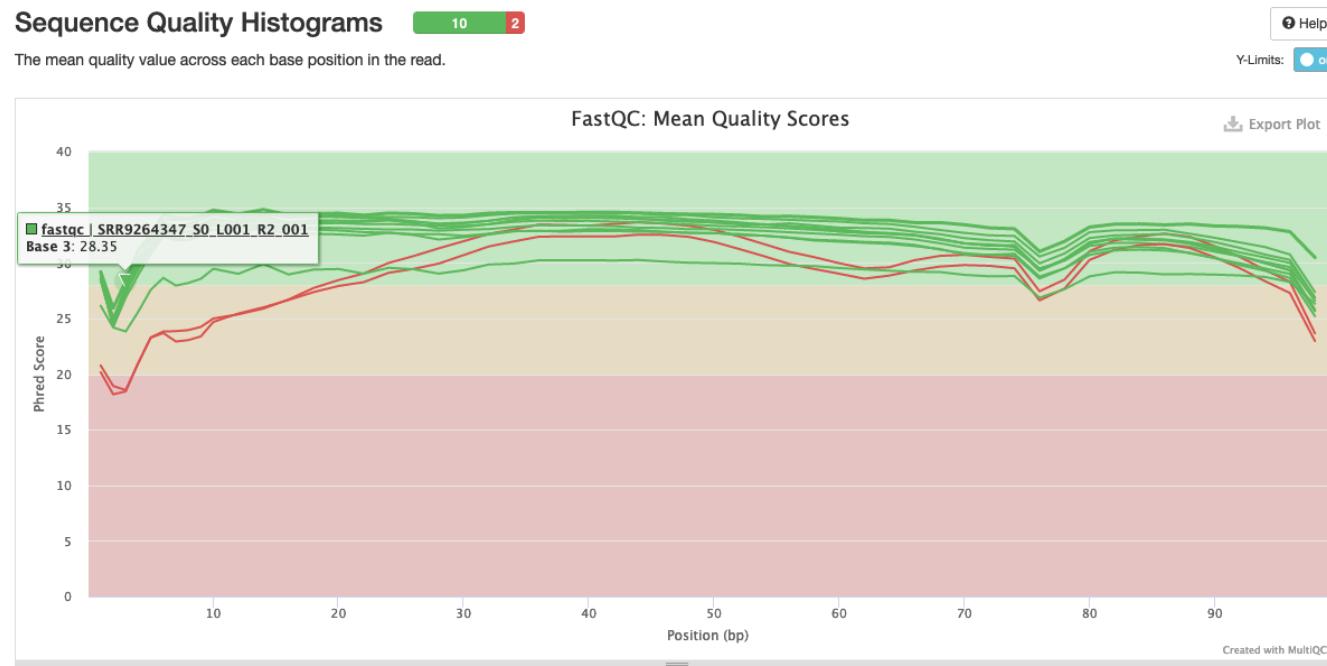


QC of Raw Reads - MultiQC - General Statistics

The screenshot shows the MultiQC interface version 1.7. The left sidebar lists analysis modules: 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, and Adapter Content. The main content area displays a "General Statistics" table with 12 rows. Each row contains a sample name, percentage of Dups, percentage of GC content, and the total number of sequences (M Seqs). The table includes buttons for "Copy table", "Configure Columns", and "Plot". A welcome message at the top right suggests watching a tutorial video.

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

The screenshot shows a web browser displaying the 10x Genomics support website at <https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latency>. The page title is "What is Cell Ranger? - Software". The main content area is titled "What is Cell Ranger?" and describes Cell Ranger as a set of analysis pipelines for Chromium single-cell data. It includes sections on the four pipelines: `cellranger mkfastq`, `cellranger count`, `cellranger aggr`, and `cellranger reanalyze`. On the left, there is a sidebar with links for "CELL RANGER", "Introduction", "Downloads", "Tutorials", and "Running Pipelines".

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 [targeted transcriptomes](#).

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

```
File Edit View Search Terminal Help
%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

```
File Edit View Search Terminal Help
_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

```
File Edit View Search Terminal Help
_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



Cell Ranger • count

SITTA6

Summary

Analysis

14,668

Estimated Number of Cells

20,065

Mean Reads per Cell

1,344

Median Genes per Cell

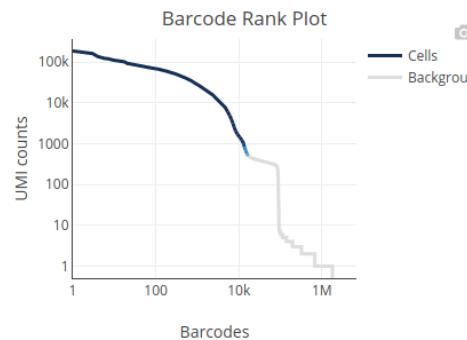
Sequencing ②

Number of Reads	294,310,066
Number of Short Reads Skipped	0
Valid Barcodes	97.7%
Valid UMI	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 ②

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

Cells ②



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

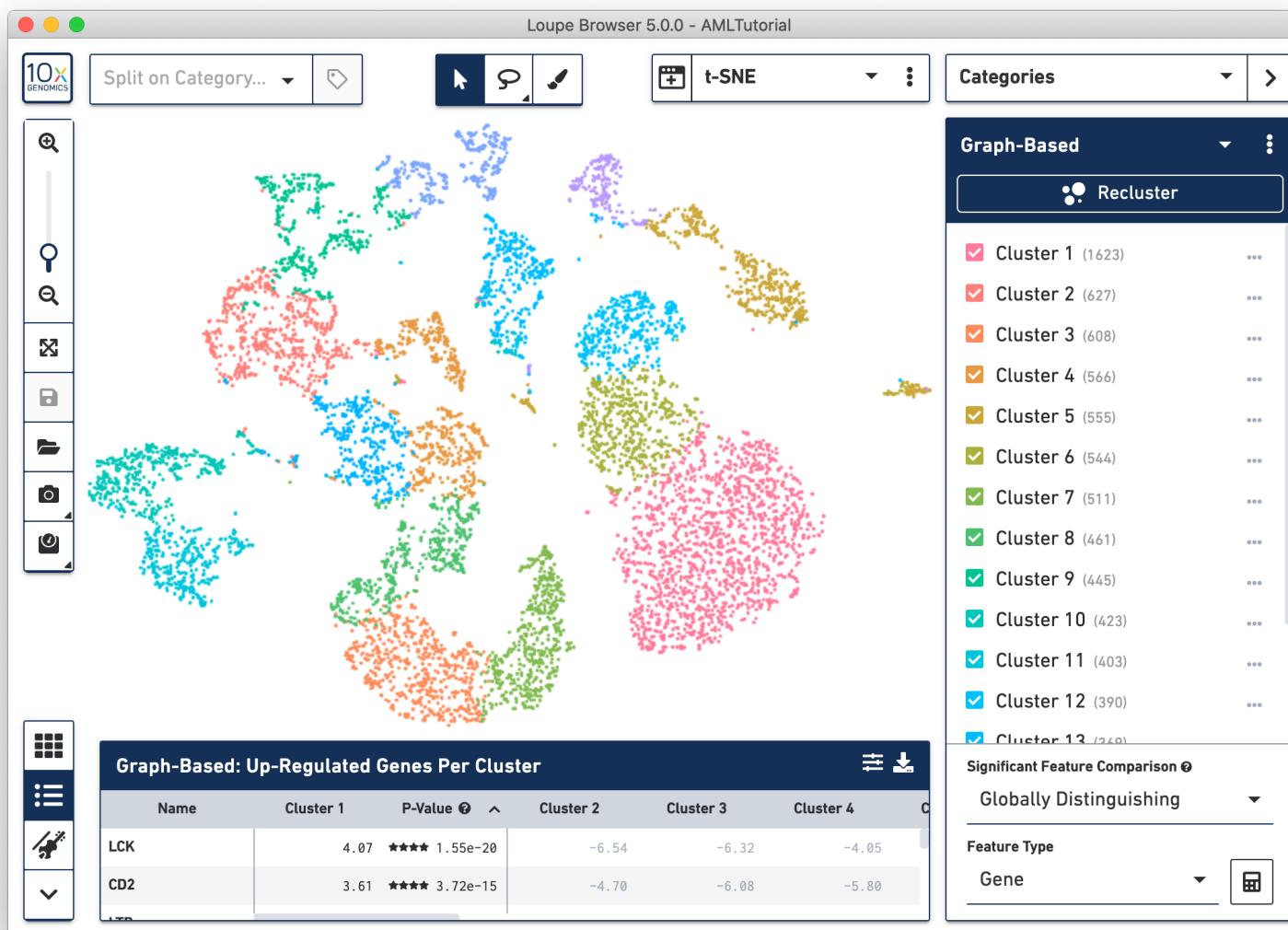
Sample

Sample ID	SITTA6
Sample Description	

Cell Ranger outputs

```
File Edit View Search Terminal Help
_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



Cell Ranger outputs

```
File Edit View Search Terminal Help
_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

```
File Edit View Search Terminal Help
_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%-$_
```

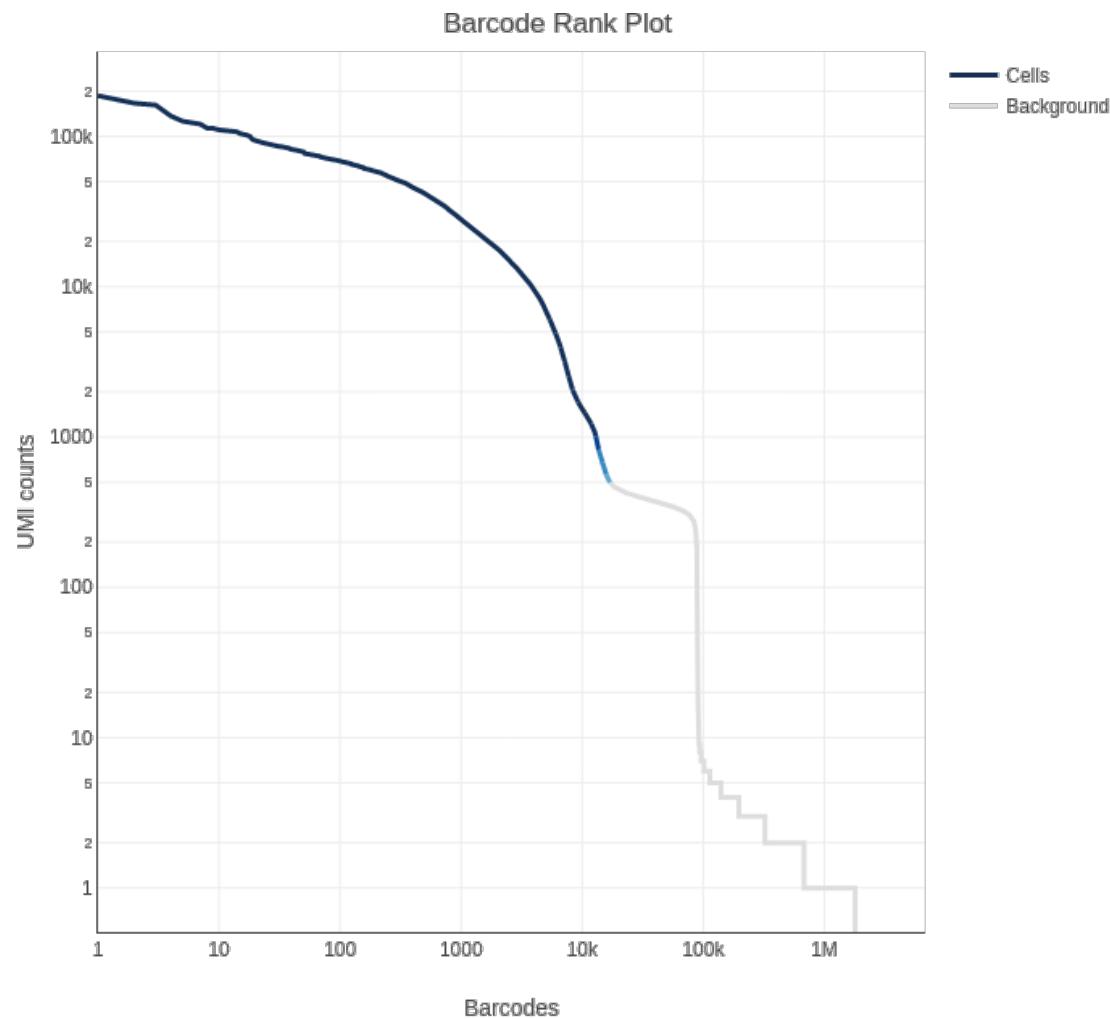
Cell Ranger outputs

```
File Edit View Search Terminal Help
_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%- $
%h%- $ ls SRR9264343/outs/raw_feature_bc_matrix
barcodes.tsv.gz
features.tsv.gz
matrix.mtx.gz
%h%- $ 
```

Cell Ranger outputs

```
File Edit View Search Terminal Help
_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

