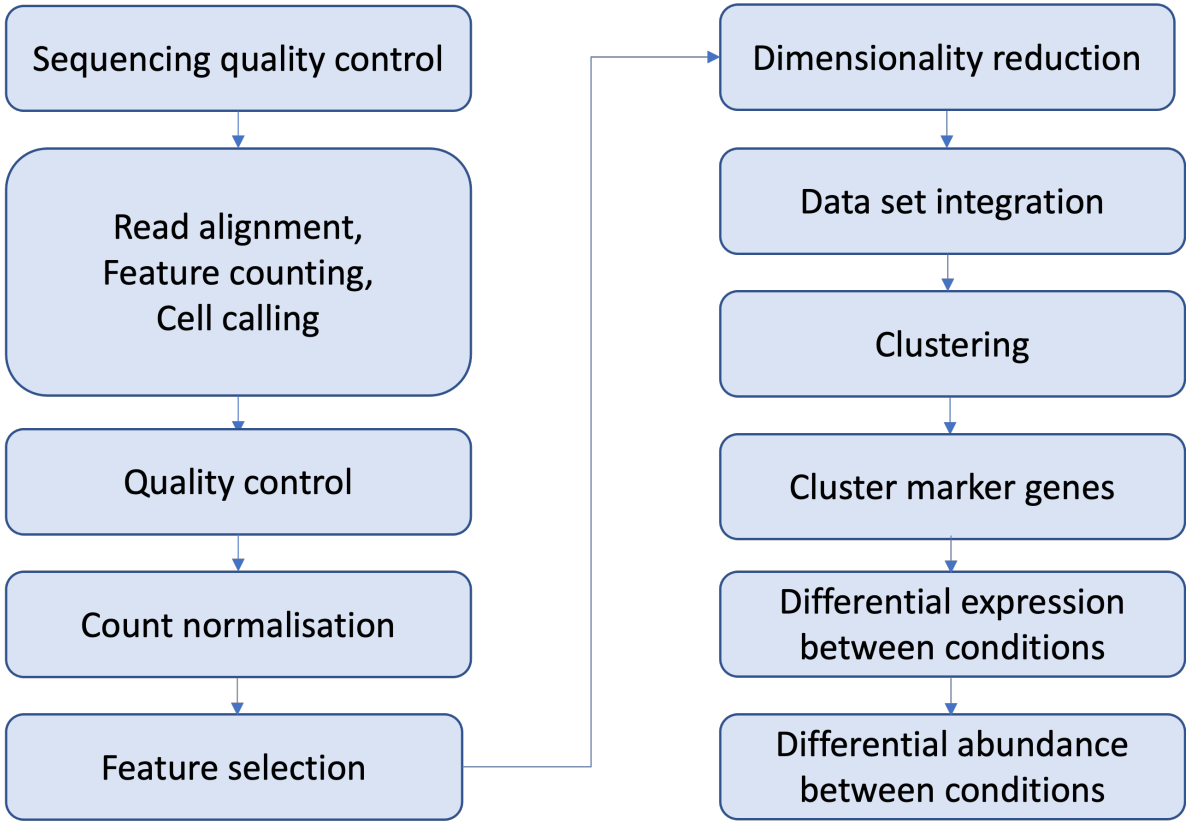
Alignment and feature counting

Ashley Sawle, Chandra Chilamakuri

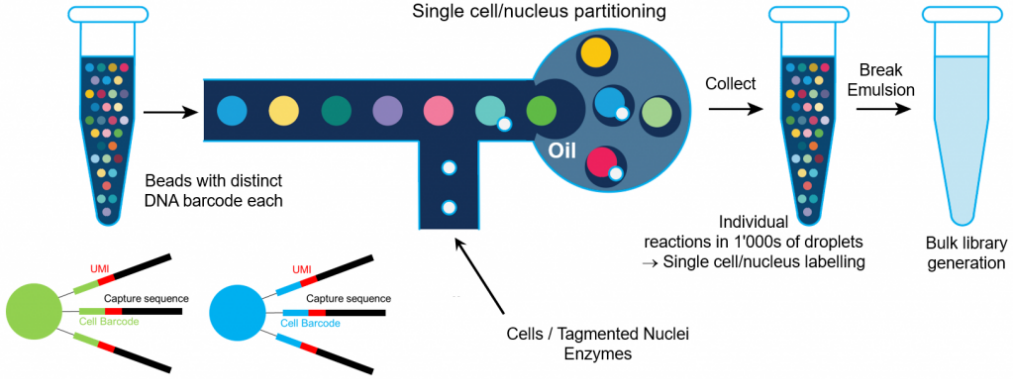
March 2024

Single Cell RNAseq Analysis Workflow



10x technology overview

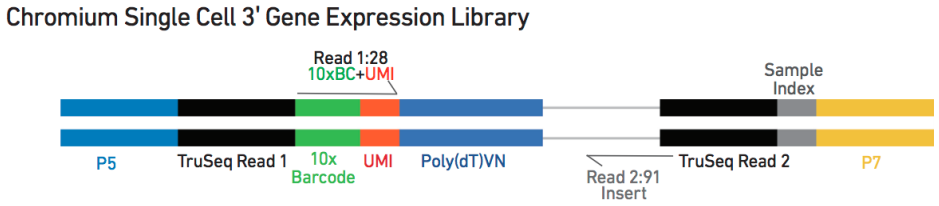
- GEM: Gel Bead-In-EMulsion
- Millions of GEMs
- Each GEM comes with thousands of oligonucleotide sequences
- Each oligo sequence has cell barcode + UMI + capture sequence



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

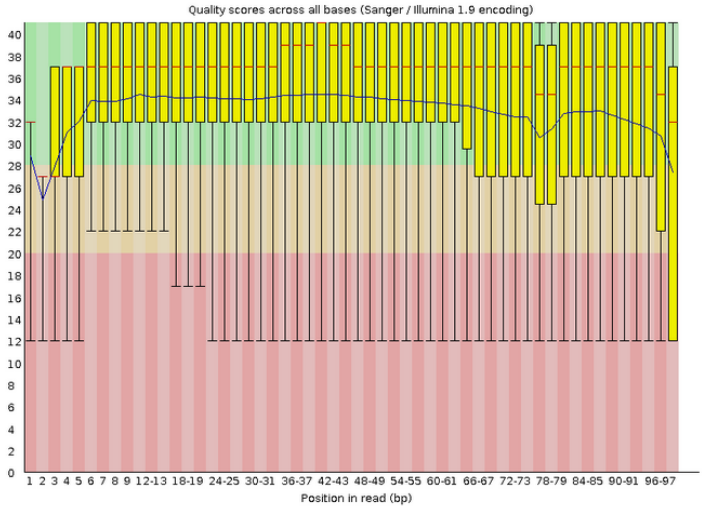
Summary

- 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

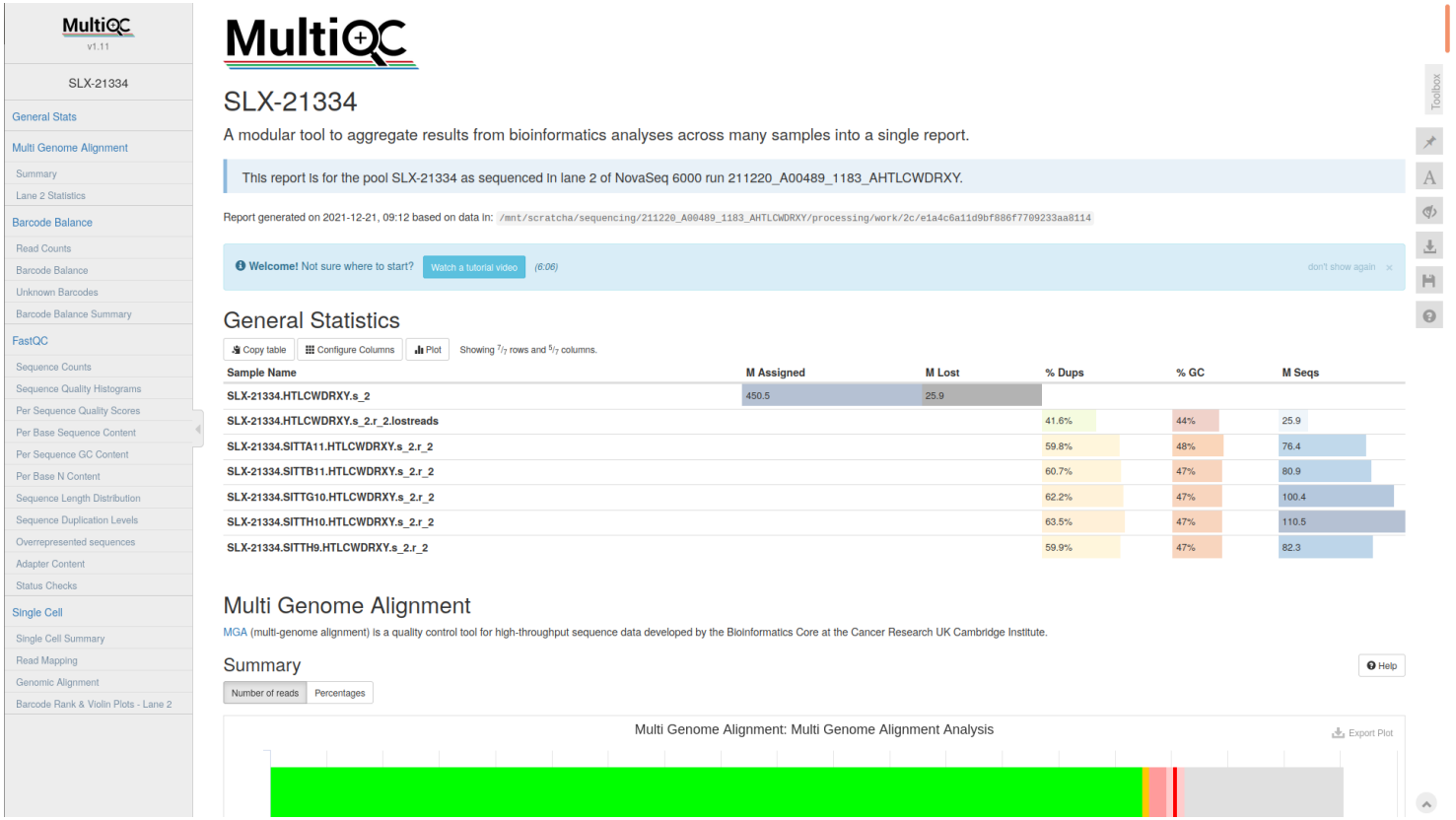
Basic Statistics

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

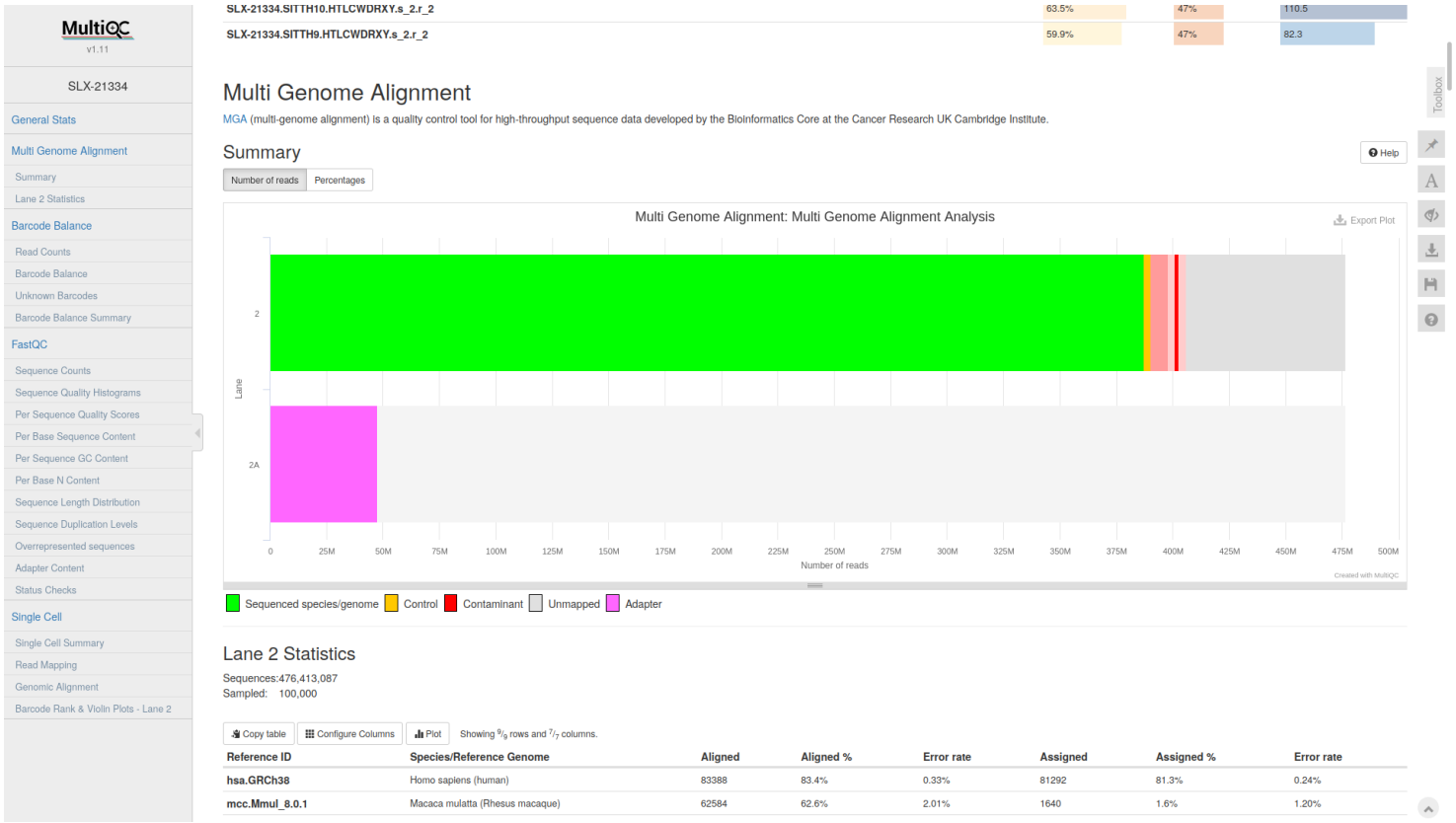
Per base sequence quality



QC of Raw Reads - MultiQC

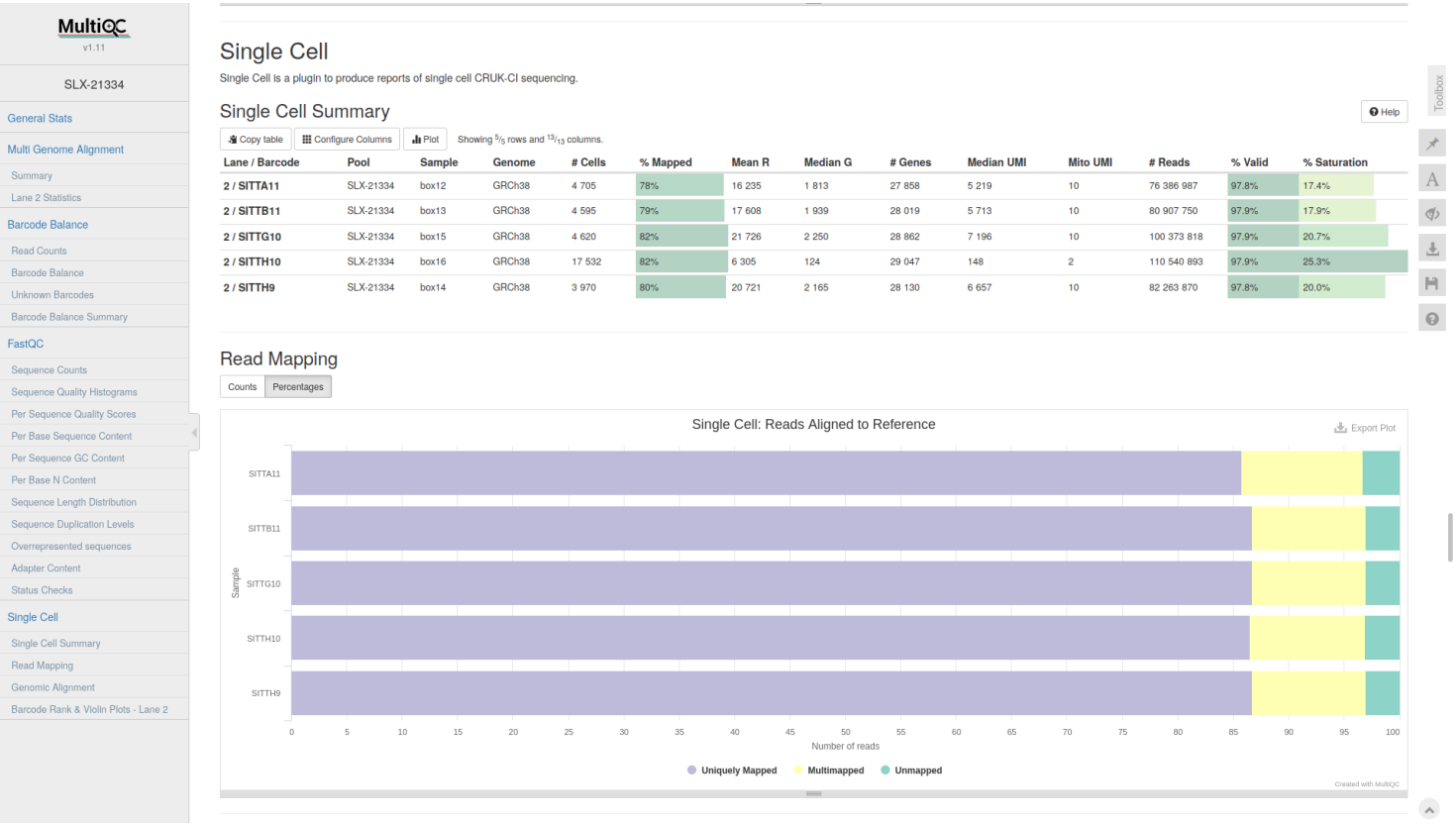


QC of Raw Reads - MultiQC



QC of Raw Reads - MultiQC

QC of Raw Reads - MultiQC



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? -Soft

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

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 we may want:

```
SITTA11_S1_L001_I1_001.fastq.gz
SITTA11_S1_L001_I2_001.fastq.gz
SITTA11_S1_L001_R1_001.fastq.gz
SITTA11_S1_L001_R2_001.fastq.gz
```

Unfortunately, the files we receive from the Genomics server will be named like this:

```
SLX-21334.SITTA11.HTLCWDRXY.s_2.i_1.fq.gz
SLX-21334.SITTA11.HTLCWDRXY.s_2.i_2.fq.gz
SLX-21334.SITTA11.HTLCWDRXY.s_2.r_1.fq.gz
SLX-21334.SITTA11.HTLCWDRXY.s_2.r_2.fq.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

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

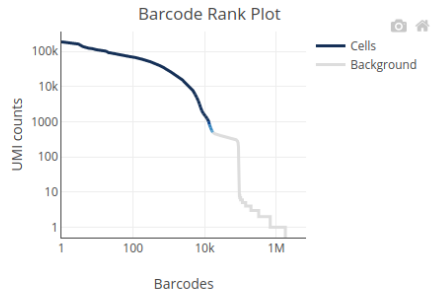
Sequencing

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

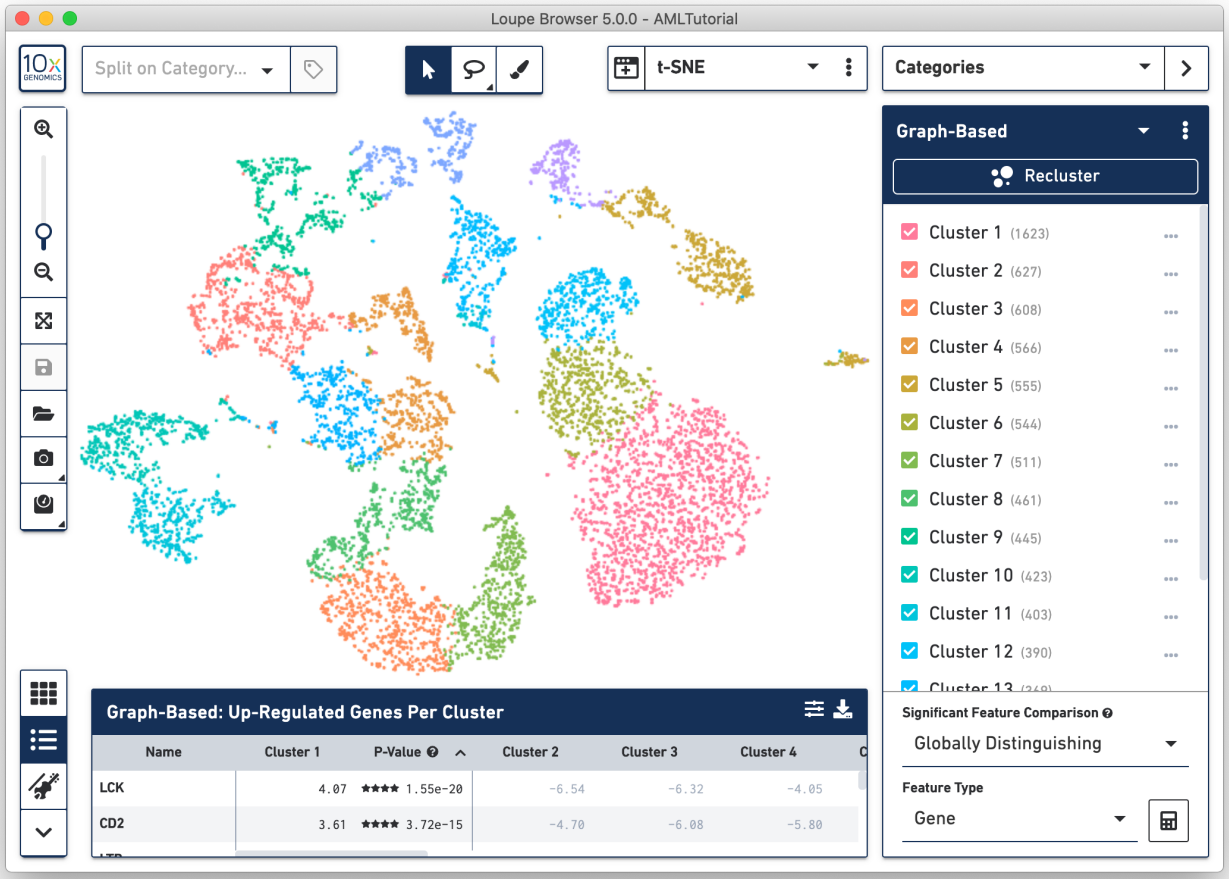
Sample

Sample ID	SITTA6
Sample Description	

Cell Ranger outputs

```
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%-$
```

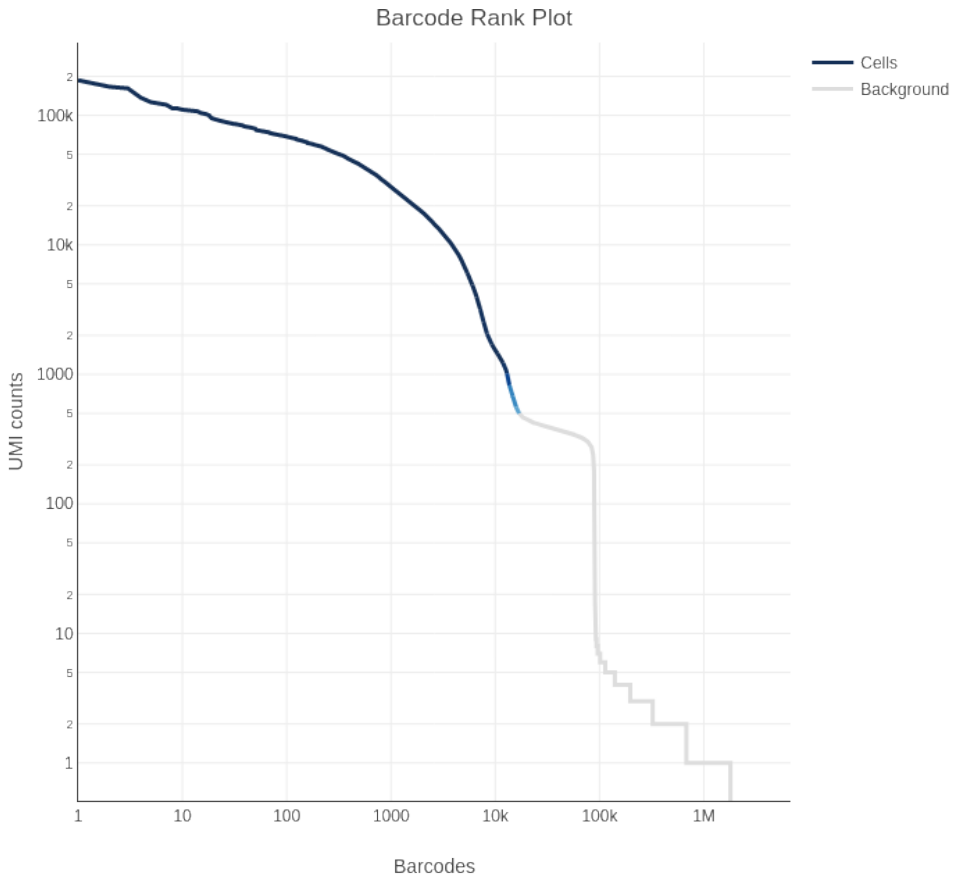
Cell Ranger outputs

```
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