

CANCER  
RESEARCH  
UK

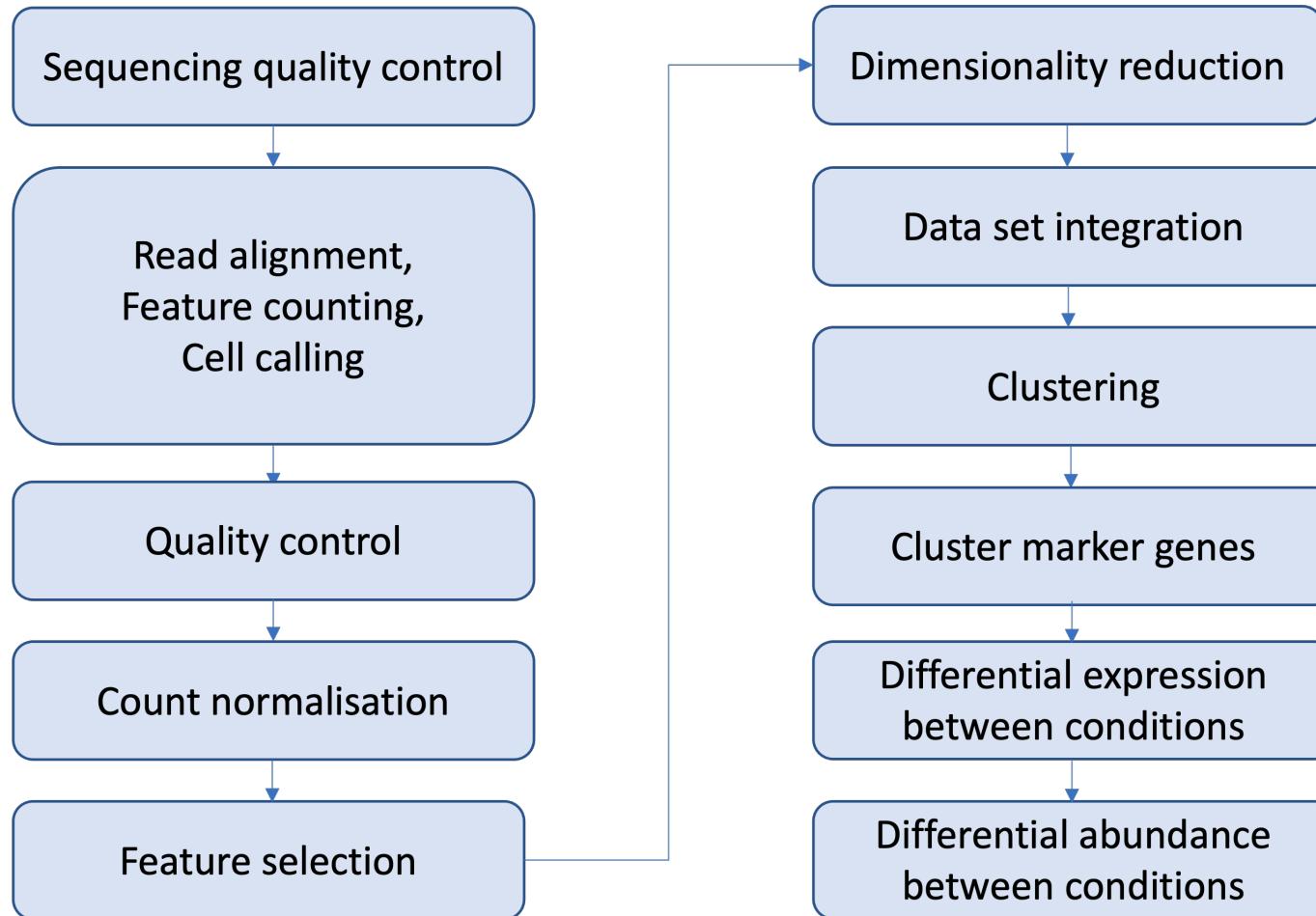
CAMBRIDGE  
INSTITUTE

# Alignment and feature counting

Ashley Sawle

April 2022

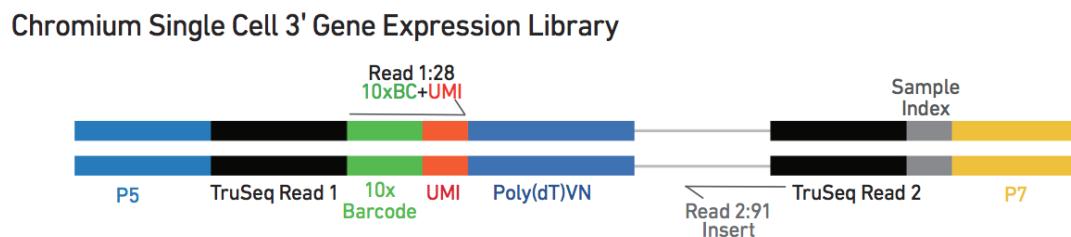
# Single Cell RNAseq Analysis Workflow



# 10x library file structure

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

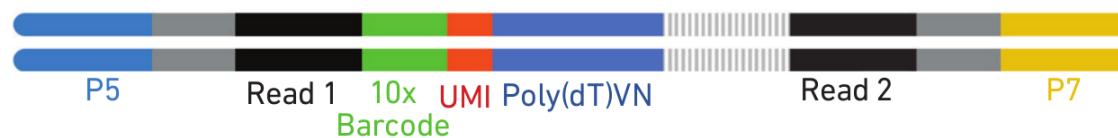
- **sample index** - identifies the library, with one or two indexes per sample
- **10x barcode** - identifies the droplet in the library
- **UMI** - identifies the transcript molecule within a cell and gene
- **insert** - the transcript molecule



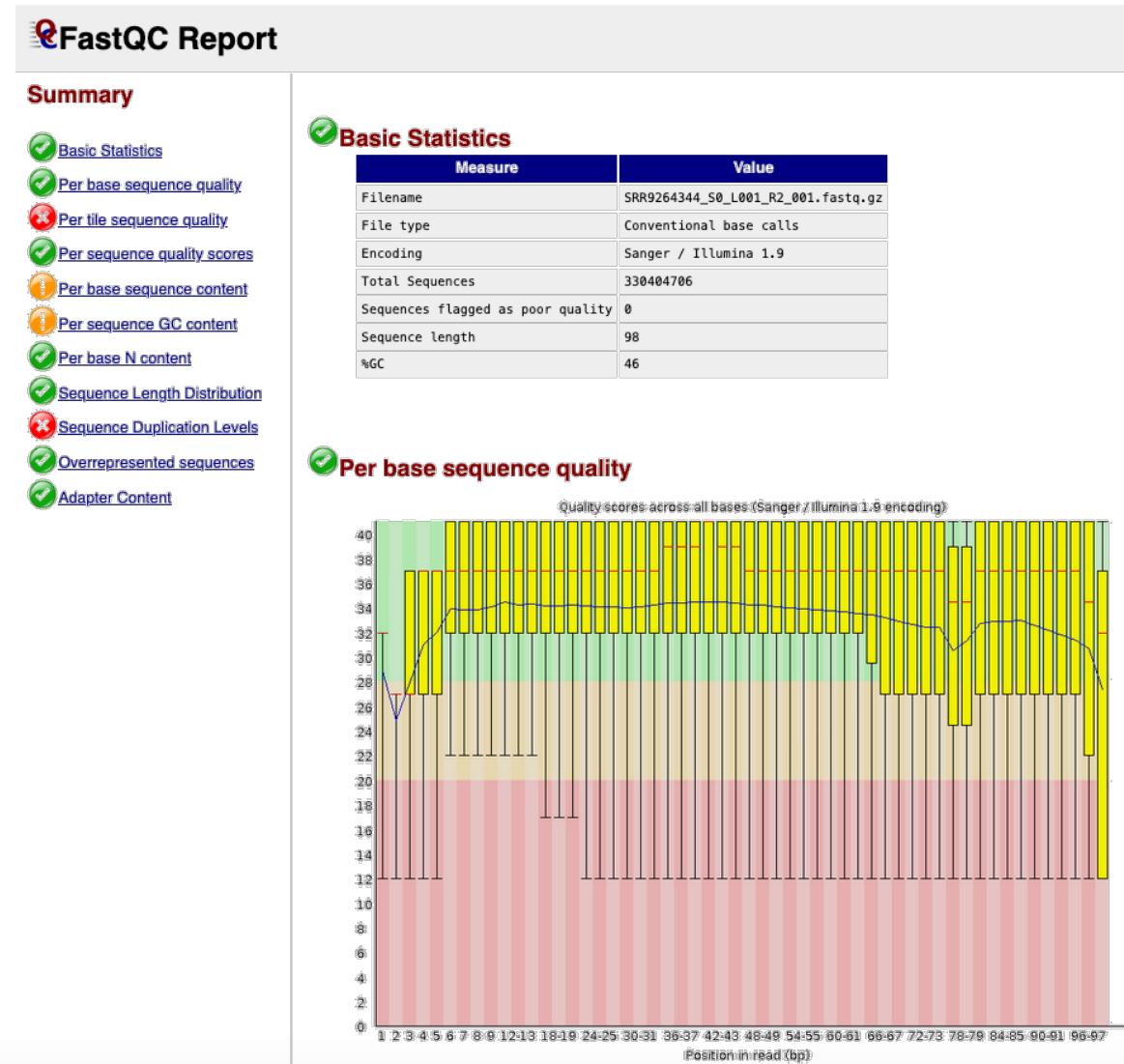
# Raw fastq files

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

**MultiQC**  
v1.11

SLX-21334

General Stats

Multi Genome Alignment

Summary

Lane 2 Statistics

Barcode Balance

Read Counts

Barcode Balance

Unknown Barcodes

Barcode Balance Summary

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

Status Checks

Single Cell

Single Cell Summary

Read Mapping

Genomic Alignment

Barcode Rank & Violin Plots - Lane 2

**MultiQC**

**SLX-21334**

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

This report is for the pool SLX-21334 as sequenced in lane 2 of NovaSeq 6000 run 211220\_A00489\_1183\_AHTLCWDRXY.

Report generated on 2021-12-21, 09:12 based on data in: /mnt/scratcha/sequencing/211220\_A00489\_1183\_AHTLCWDRXY/processing/work/2c/e1a4c6a11d9bf886f7709233aa8114

Welcome! Not sure where to start? Watch a tutorial video (6:06)

don't show again ×

**General Statistics**

Copy table Configure Columns Plot Showing 7/7 rows and 5/5 columns.

Sample Name	M Assigned	M Lost	% Dups	% GC	M Seqs
SLX-21334.HTLCWDRXY.s_2	450.5	25.9	41.6%	44%	25.9
SLX-21334.HTLCWDRXY.s_2_r_2.lostreads			59.8%	48%	76.4
SLX-21334.SITTA11.HTLCWDRXY.s_2_r_2			60.7%	47%	80.9
SLX-21334.SITTB11.HTLCWDRXY.s_2_r_2			62.2%	47%	100.4
SLX-21334.SITTG10.HTLCWDRXY.s_2_r_2			63.5%	47%	110.5
SLX-21334.SITTH10.HTLCWDRXY.s_2_r_2			59.9%	47%	82.3
SLX-21334.SITTH9.HTLCWDRXY.s_2_r_2					

**Multi Genome Alignment**

MGA (multi-genome alignment) is a quality control tool for high-throughput sequence data developed by the Bioinformatics Core at the Cancer Research UK Cambridge Institute.

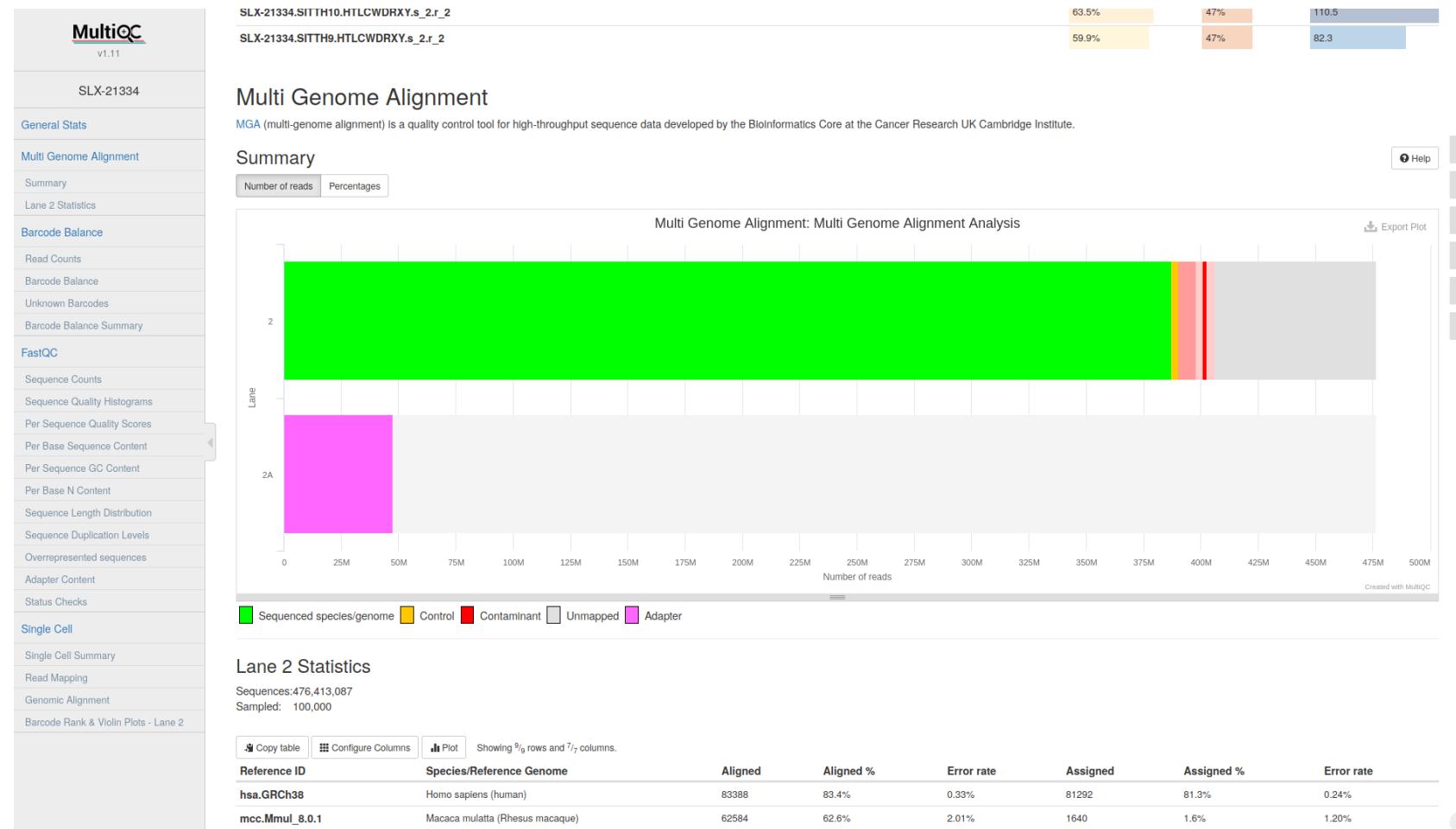
**Summary**

Number of reads Percentages

Multi Genome Alignment: Multi Genome Alignment Analysis

Export Plot

# QC of Raw Reads - MultiQC



# QC of Raw Reads - MultiQC



# QC of Raw Reads - MultiQC

MultiQC v1.11

SLX-21334

General Stats

Multi Genome Alignment

Summary

Lane 2 Statistics

Barcode Balance

Read Counts

Barcode Balance

Unknown Barcodes

Barcode Balance Summary

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

Status Checks

Single Cell

Single Cell Summary

Read Mapping

Genomic Alignment

Barcode Rank & Violin Plots - Lane 2

## Single Cell

Single Cell is a plugin to produce reports of single cell CRUK-CI sequencing.

### Single Cell Summary

Showing 5/5 rows and 13/13 columns.

Lane / Barcode	Pool	Sample	Genome	# Cells	% Mapped	Mean R	Median G	# Genes	Median UMI	Mito UMI	# Reads	% Valid	% Saturation
2 / SITTA11	SLX-21334	box12	GRCh38	4 705	78%	16 235	1 813	27 858	5 219	10	76 386 987	97.8%	17.4%
2 / SITTB11	SLX-21334	box13	GRCh38	4 595	79%	17 608	1 939	28 019	5 713	10	80 907 750	97.9%	17.9%
2 / SITTG10	SLX-21334	box15	GRCh38	4 620	82%	21 726	2 250	28 862	7 196	10	100 373 818	97.9%	20.7%
2 / SITTH10	SLX-21334	box16	GRCh38	17 532	82%	6 305	124	29 047	148	2	110 540 893	97.9%	25.3%
2 / SITTH9	SLX-21334	box14	GRCh38	3 970	80%	20 721	2 165	28 130	6 657	10	82 263 870	97.8%	20.0%

### Read Mapping

Counts Percentages

Single Cell: Reads Aligned to Reference

Number of reads

Sample

Uniquely Mapped Multimapped Unmapped

Created with MultiQC

Sample	Uniquely Mapped	Multimapped	Unmapped
SITTA11	~85	~10	~5
SITTB11	~85	~10	~5
SITTG10	~85	~10	~5
SITTH10	~85	~10	~5
SITTH9	~85	~10	~10

# 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 site at <https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest>. The page title is "What is Cell Ranger? - Software Pipelines". 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 lists four pipelines: `cellranger mkfastq`, `cellranger count`, `cellranger aggr`, and `cellranger reanalyze`. Below this, it describes the `cellranger multi` pipeline for Cell Multiplexing. The sidebar on the left contains links for "CELL RANGER", "Introduction", "Downloads", "Tutorials", "Running Pipelines", "Workflows", and "Understanding Outputs". A red banner at the bottom right of the content area reads "Cell Ranger v6.0 (latest)".

## 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 `Feature Barcode` data.

Workflows

Cell Ranger v6.0 (latest)

# 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%-$ cellranger count \
>           --id=SITTA11 \
>           --fastqs=fastq \
>           --transcriptome=references/refdata-gex-GRCh38-2020-A \
>           --expect-cells 3000 \
>           --localcores=16 \
>           --localmem=32
```

# Cell Ranger outputs

- One directory per sample

```
File Edit View Search Terminal Help
%h%-$
%h%-$ ls SITTB11
 cmdline
 filelist
 finalstate
 invocation
 jobmode
 log
 mrosource
 outs
 perf
 SC_RNA_COUNTER_CS
 sitecheck
 SITTB11.mri.tgz
 tags
 timestamp
 uuid
 vdrkill
 versions
%h%-$ 
```

# Cell Ranger outputs

```
File Edit View Search Terminal Help  
_versions  
%h%-  
%h%- ls SITTB11/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 SITTB11/outs
analysis
cloupe.clope
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

Summary

Analysis

14,668

Estimated Number of Cells

20,065

Mean Reads per Cell

1,344

Median Genes per Cell

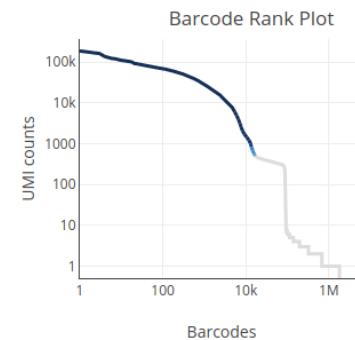
### Sequencing

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

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

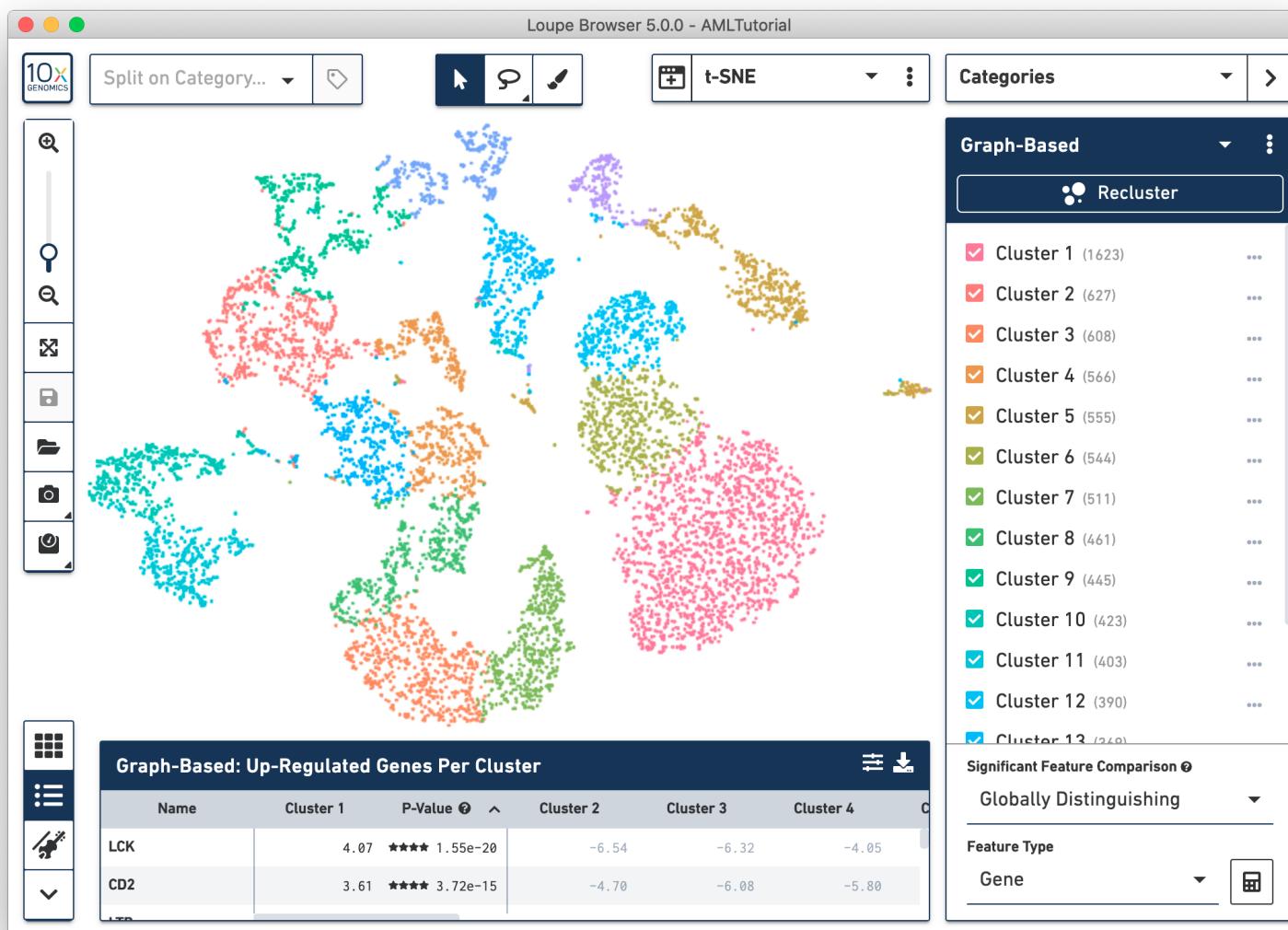
### Sample

Sample ID	SITTA6
Sample Description	

# Cell Ranger outputs

```
File Edit View Search Terminal Help
_versions
%h%-$
%h%-$ ls SITTB11/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 SITTB11/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 SITTB11/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%-$ □
```

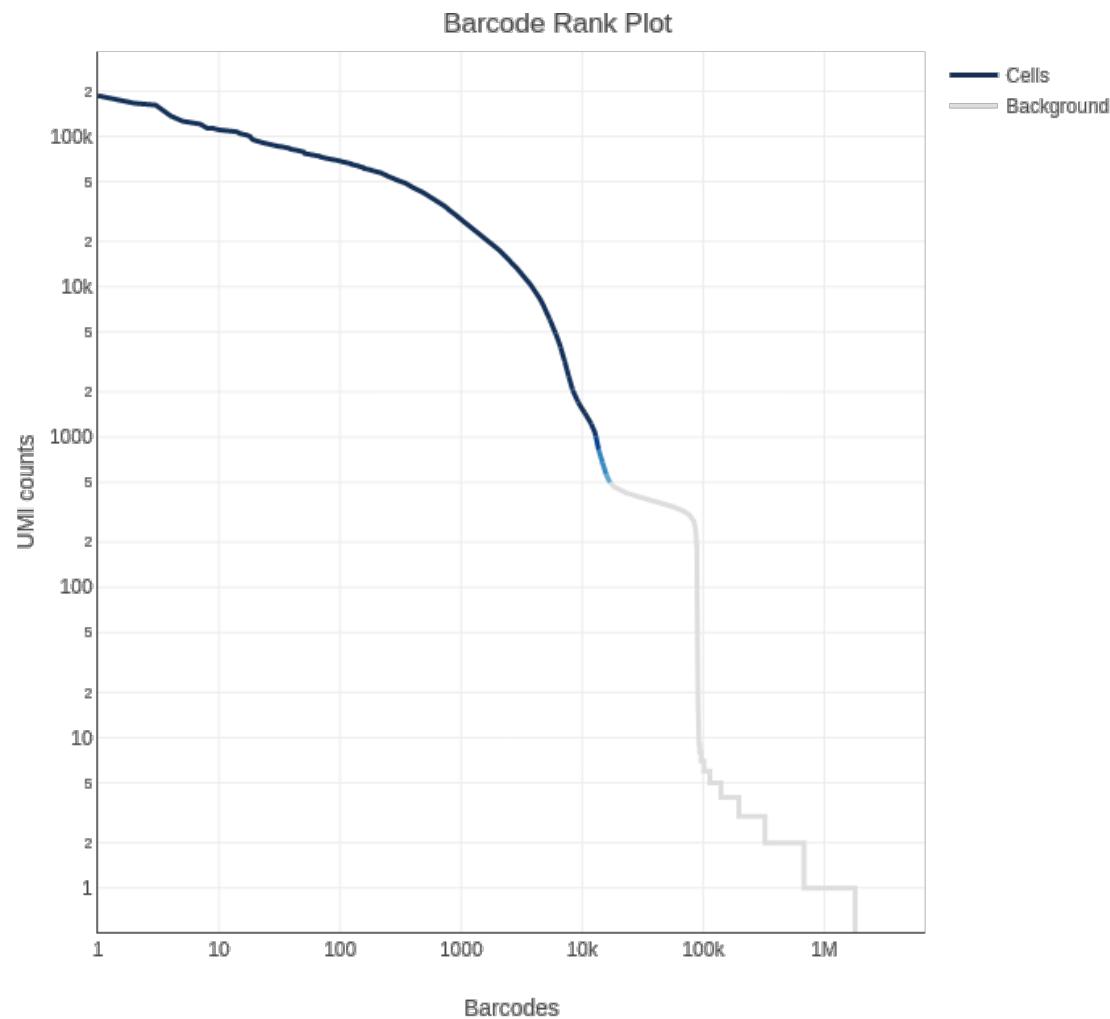
# Cell Ranger outputs

```
File Edit View Search Terminal Help
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 SITTB11/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 SITTB11/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

