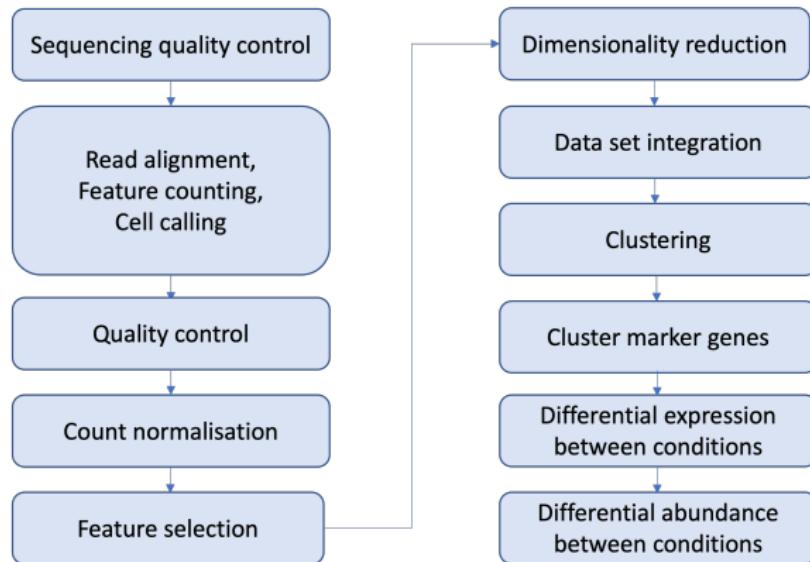


Alignment and feature counting

Ashley Sawle, Chandra Chilamakuri

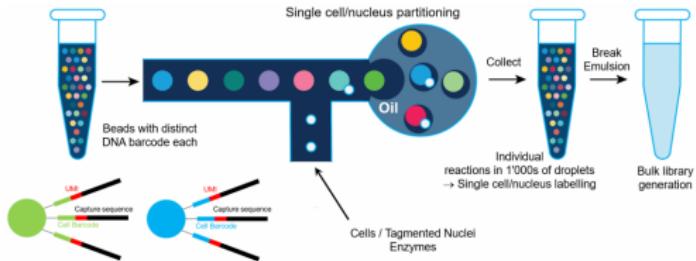
March 2024

Single Cell RNAseq Analysis Workflow



10x technology overview

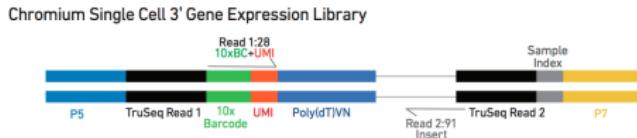
- ▶ GEM: Gel Bead-In-EMulsion
- ▶ Millions of GEMs
- ▶ Each GEM comes with thousands of oligonucleotide sequences
- ▶ Each oligo sequence has cell barcode + UMI + capture sequence



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

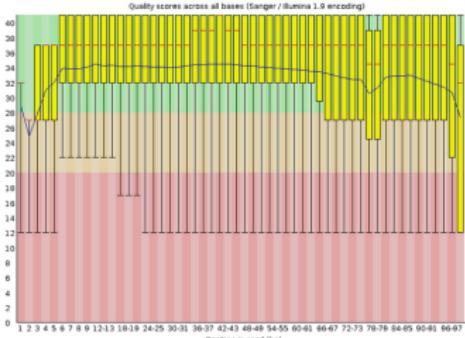
-  Basic Statistics
-  Per base sequence quality
-  Per file sequence quality
-  Per sequence quality scores
-  Per base sequence content
-  Per sequence GC content
-  Per base N content
-  Sequence Length Distribution
-  Sequence Duplication Levels
-  Overrepresented sequences
-  Adapter Content

Basic Statistics

Measure	Value
Filename	SRR3264344_S8_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	330484786
Sequences flagged as poor quality	0
Sequence length	96
NGC	46

Per base sequence quality

(Quality scores across all bases (Sanger / Illumina 1.9 encoding))



Position in read (bp)

QC of Raw Reads - MultiQC

MultiQC
v1.1.1

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
- Oversampled sequences
- Adapter Content
- Status Checks

Single Cell

- Single Cell Summary
- Read Mapping
- Genome 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_AHTLCWDRX. Report generated on 2021-12-01, 09:12 based on data in: /well/kirrisona/sequencing/211220_A00489_1183_AHTLCWDRX/processing/work/2/c/ef5ddc8411090f8887779820aa8114

Welcome! Not sure where to start? [View a basic report](#) (PDF)

[don't show again](#)

General Statistics

Sample Name M Assigned M Lost % Dups % GC M Seqs

Sample Name	M Assigned	M Lost	% Dups	% GC	M Seqs
SLX-21334_AHTLCWDRX_a_2	450.5	25.0	41.0%	44%	25.9
SLX-21334_AHTLCWDRX_a_2_1_3_lowreads			58.8%	40%	76.4
SLX-21334_BTTA11_AHTLCWDRX_a_2_2			68.7%	47%	80.9
SLX-21334_BTTB11_AHTLCWDRX_a_2_2			82.2%	47%	109.4
SLX-21334_BTTT10_AHTLCWDRX_a_2_2			63.9%	47%	111.5
SLX-21334_BTTT10_AHTLCWDRX_a_2_2			58.9%	47%	82.3

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.

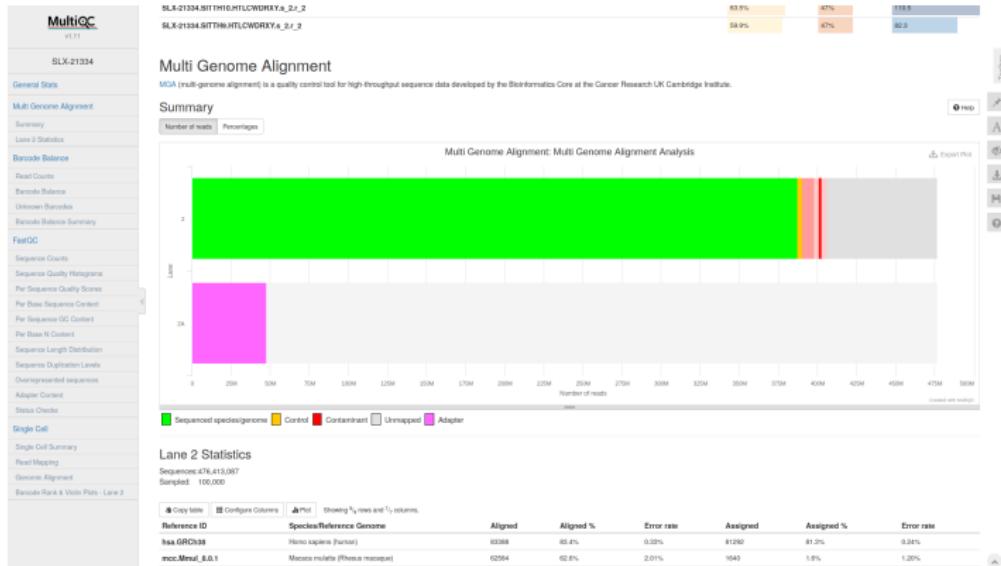
Summary

[Number of reads](#) | [Percentages](#)

Multi Genome Alignment: Multi Genome Alignment Analysis

[Export PDF](#)

QC of Raw Reads - MultiQC



QC of Raw Reads - MultiQC



QC of Raw Reads - MultiQC

MultiQC
v1.11

SLX-21354

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

Oversrepresented sequences

Adapter Content

Status Checks

Single Cell

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

Single Cell Summary

Showing 10 rows and 11 columns

Lane / Barcode	Pool	Sample	Genseq	# Cells	% Mapped	Mean R	Median G	# Gores	Median UMI	Min UMI	# Reads	% Valid	% Saturation
2 / SITTA1	SLX-21354	Isot12	GRC018	4 705	78%	16 235	1 813	27 658	5 219	10	76 386 967	97.8%	17.4%
2 / SITTB11	SLX-21354	Isot13	GRC018	4 986	79%	17 408	1 938	28 019	5 713	10	80 867 798	97.8%	17.9%
2 / SITTC16	SLX-21354	Isot15	GRC018	4 620	82%	21 706	2 256	20 982	7 196	10	100 273 616	97.9%	20.7%
2 / SITTH19	SLX-21354	Isot16	GRC018	17 532	82%	6 365	104	39 047	148	2	110 540 860	97.8%	35.3%
2 / SITTH9	SLX-21354	Isot14	GRC018	3 870	86%	20 721	2 165	28 130	6 657	10	62 263 873	97.8%	20.0%

Read Mapping

Counts Percentages

Single Cell: Reads Aligned to Reference

Sample

SITTA1

SITTB11

SITTC16

SITTH19

0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

Number of reads

Uniquely Mapped Multimapped Unaligned

Export PDF

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 window with the URL <https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/>. The page title is "What is Cell Ranger?". 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 data. On the left sidebar, there are sections for "CELL RANGER" (Introduction, Downloads, 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 GEK), count (Feature Barcode), count (Feature Barcode Only), agg, reanalyze, integrated pipelines), and "SOFTWARE > PIPELINES". The top navigation bar includes links for Products, Research Areas, Resources, Support, and Company, along with a search bar and contact support buttons.

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%- $ []
```

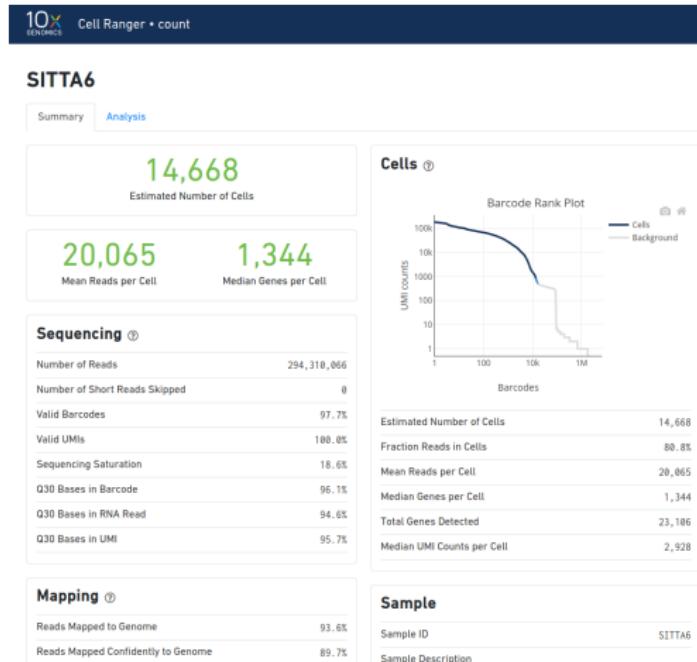
Cell Ranger outputs

```
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



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.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 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%-$
```

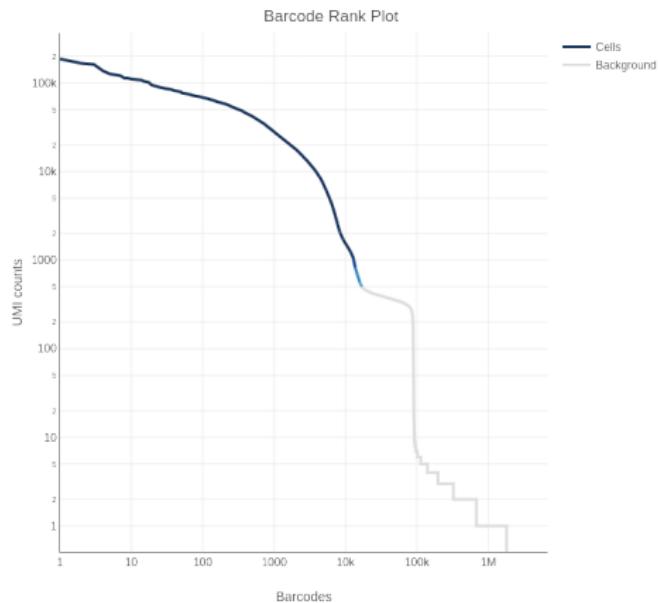
Cell Ranger outputs

```
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

