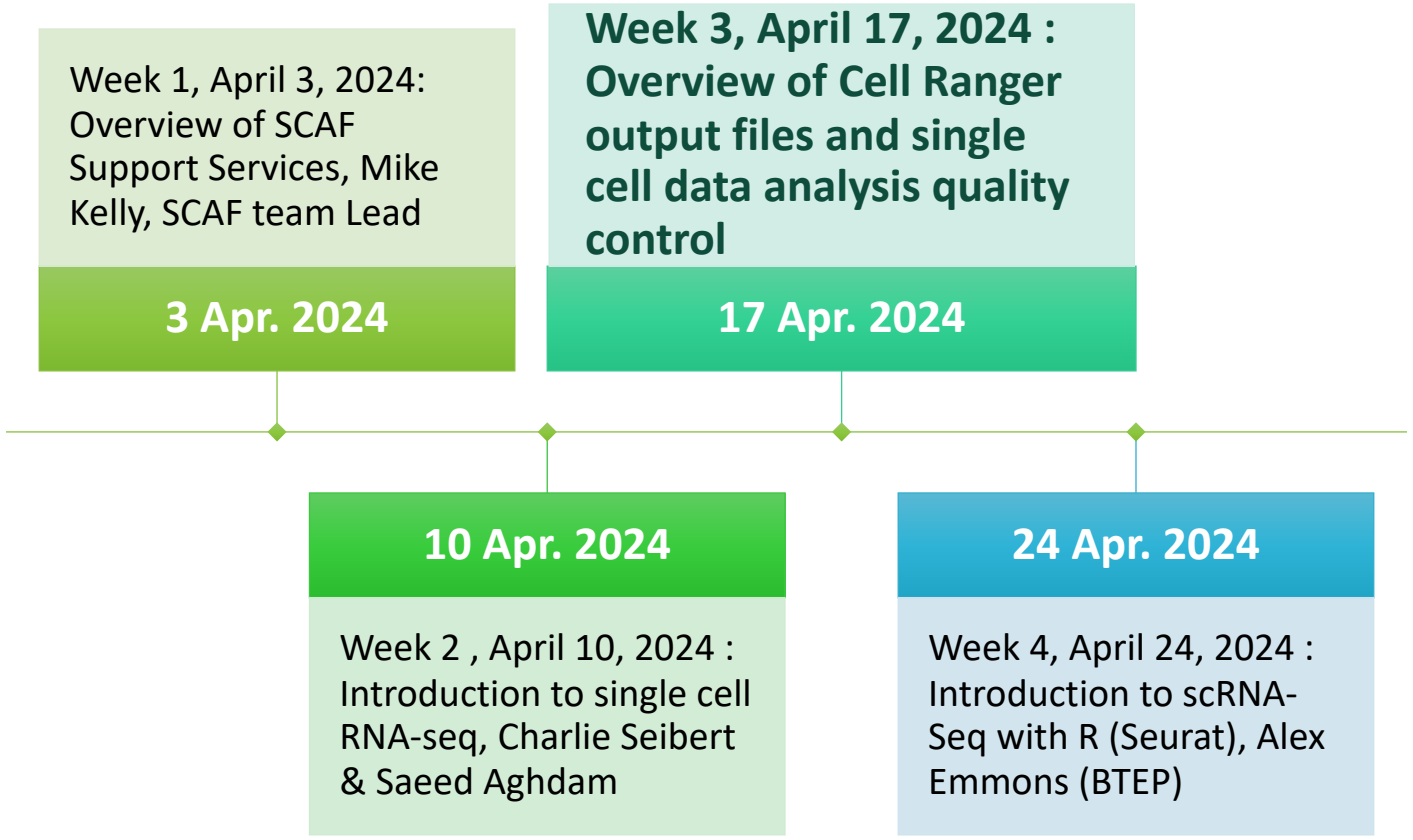
Overview of Cell Ranger output files and single cell data analysis quality control

Kimia Dadkhah
Bioinformatics Analyst
Single Cell Analysis Facility (SCAF)
April 17, 2024
BTEP Single Cell Seminar Series

BTEP Getting Started with scRNA-Seq Seminar Series



BTEP Single Cell Seminar Series

Outline

- Considerations before and after sequencing of single cell libraries
- Steps of primary analysis of single cell data
- How to use primary analysis output to check the quality and success of single cell experiment before moving on to downstream analysis

SCAF full support for 10x workflow and steps of QC in single cell experiment

Full Support

QC

Accept cell/ nuclei suspension from investigators

QC

Capture

QC

Generation of cDNA

QC

Library Prep(s)

QC

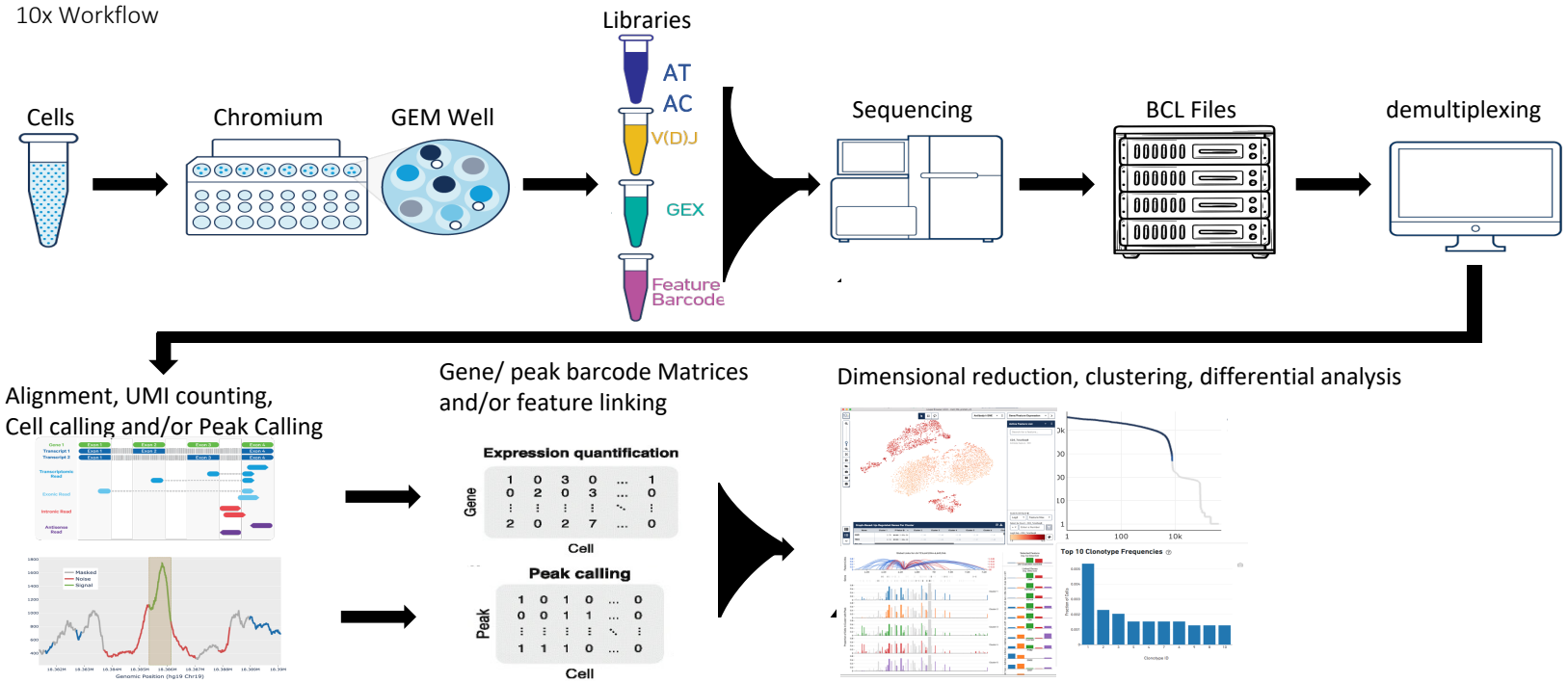
Sequence

QC

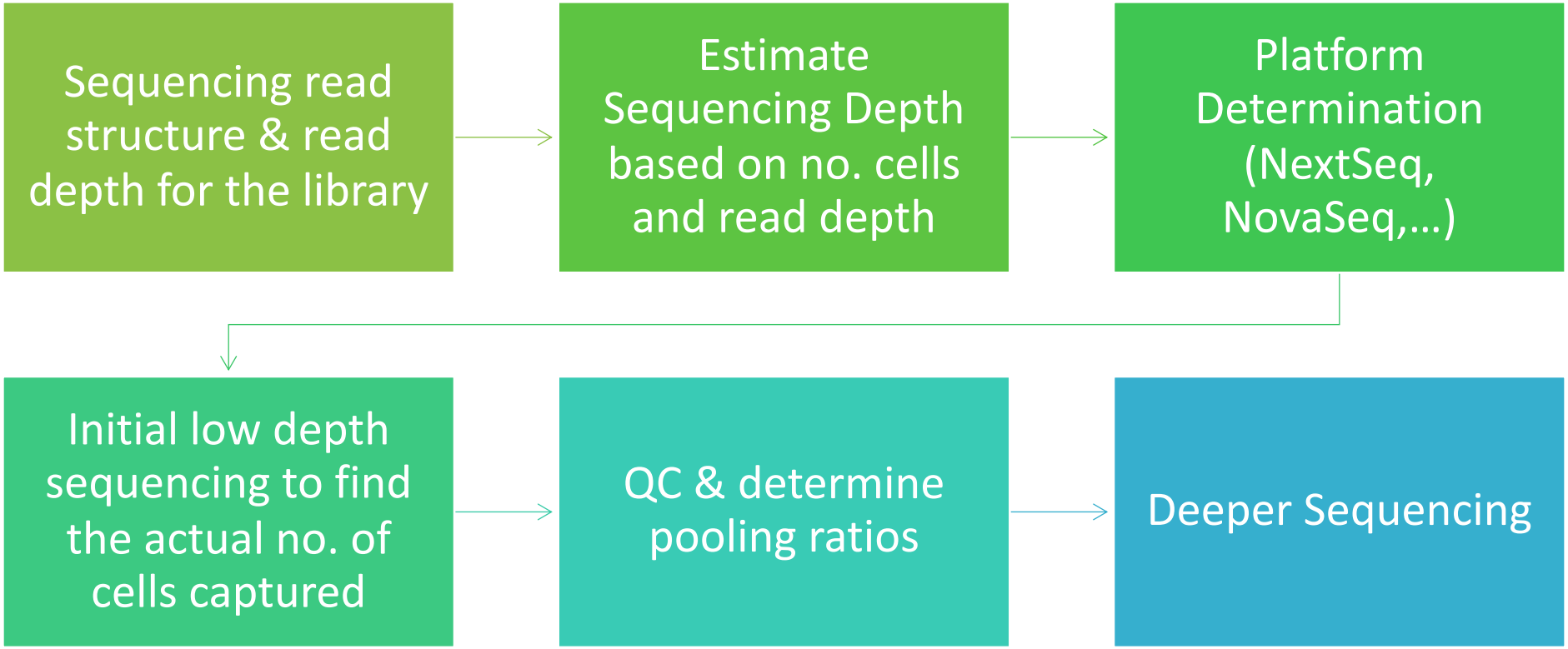
Data Processing

Data Delivery

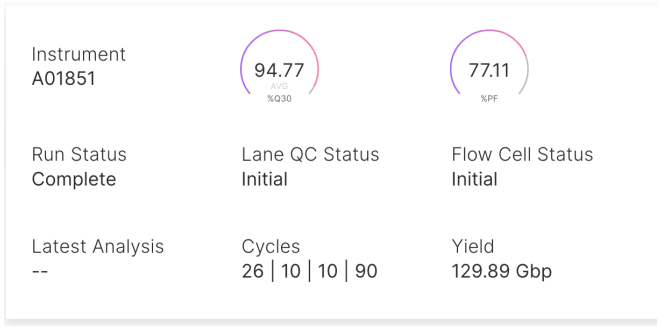
Post-Delivery Meeting

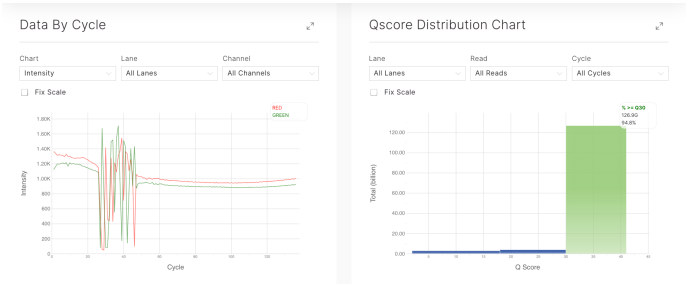


Before Sequencing considerations



QC after sequencing
Illumina BaseSpace Sequencing Hub





Per Lane Metrics

LANE	STATUS	READ	CLUSTER PF(%)	%≥ Q30	YIELD	ERROR RATE(%)	READS PF	DENSITY	TILES	LEGACY PHAS / PREPHAS(%)	COMMENTS	INTENSITY
1	Initial	Read 1	77.45±1.12	95.92	12.35 Gbp	0.06 ±0.01	494,376,384	2,961 ±0	156	0.082 / 0.048		1377±34
		Read 2 (I)		97.95	4.44 Gbp	0.00 ±0.00				0.000 / 0.000		503±56
		Read 3 (I)		96.83	4.44 Gbp	0.00 ±0.00				0.000 / 0.000		1337±65
		Read 4		94.09	43.96 Gbp	0.11 ±0.01				0.084 / 0.059		1078±25
2	Initial	Read 1	76.78±3.25	95.70	12.25 Gbp	0.08 ±0.04	490,108,416	2,961 ±0	156	0.079 / 0.045		1356±48
		Read 2 (I)		97.81	4.41 Gbp	0.00 ±0.00				0.000 / 0.000		465±26
		Read 3 (I)		96.85	4.41 Gbp	0.00 ±0.00				0.000 / 0.000		1283±70
		Read 4		93.82	43.62 Gbp	0.12 ±0.03				0.097 / 0.068		1039±50

Per Read Metrics

READ	CYCLES	YIELD	PROJECTED YIELD	ALIGNED (%)	ERROR RATE (%)	INTENSITY CYCLE 1	%>Q30
Read 1	26	24.60 Gbp	24.60 Gbp	0.73	0.07	1366.41	95.81
Read 2 (I)	10	8.86 Gbp	8.86 Gbp	0.00	0.00	483.88	97.88
Read 3 (I)	10	8.86 Gbp	8.86 Gbp	0.00	0.00	1309.87	96.84
Read 4	90	87.58 Gbp	87.58 Gbp	0.73	0.12	1058.70	93.96
Non-index Reads Total	116	112.18 Gbp	112.18 Gbp	0.73	0.10	1212.55	94.36
Total	**136**	**129.89 Gbp**	**129.89 Gbp**	**0.73**	**0.10**	**1054.72**	**94.77**

Primary analysis of single cell data

For the sake of this presentation, we will focus on **10x genomics Gene Expression (GEX)** droplet-based technology data analysis quality steps.

This would be just an example of performing qc by considering both the front-end of the workflow (capture) and back-end (data processing). Not all qc metrics/ all single cell platforms will be discussed here.

Steps of primary analysis

1. Demultiplex and generate FASTQs
2. Performs alignment, filtering, barcode counting, and UMI counting and generate count matrix
3. Optionally, aggregate multiple GEM wells

Cellranger

- Chromium Single Cell Software Suite for 10x Genomics experiments
- A set of analysis pipelines that process Chromium Next GEM single cell data to align reads, generate feature-barcode matrices, perform clustering and other secondary analysis.

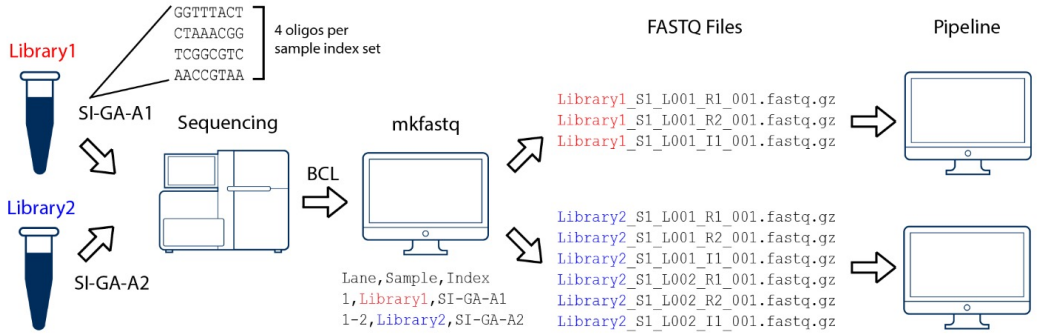
- Current version: **Cell Ranger 8.0.0 (Mar 13, 2024)**
- *New in Cell Ranger v7.0 and beyond:* ***Intronic reads*** *are counted by default for whole transcriptome gene expression data.*
- *SCAF uses the most updated version unless requested otherwise.*
- Additional information file from SCAF (any plan for combining datasets or preference on cellranger version)

Generate Fastqs

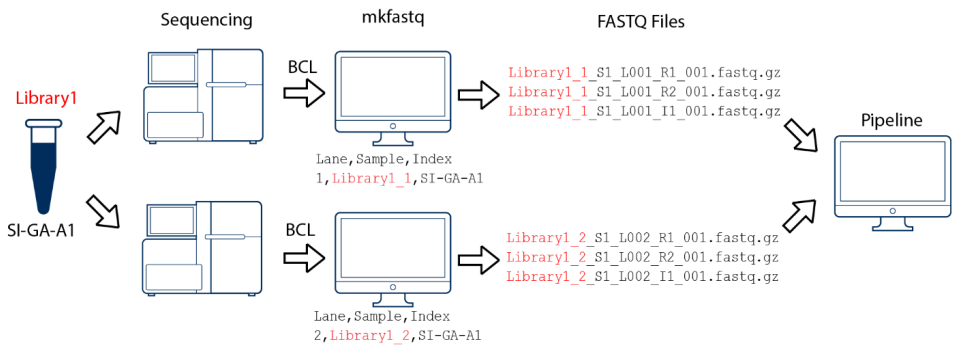
cellranger mkfastq: a wrapper of Illumina bcl2fastq, takes Illumina BCL files and demultiplex to fastqs

If you are already starting with FASTQ files, you can skip this step and proceed directly to run cellranger count.

Two libraries on the same flowcell



One library on the two different flowcells

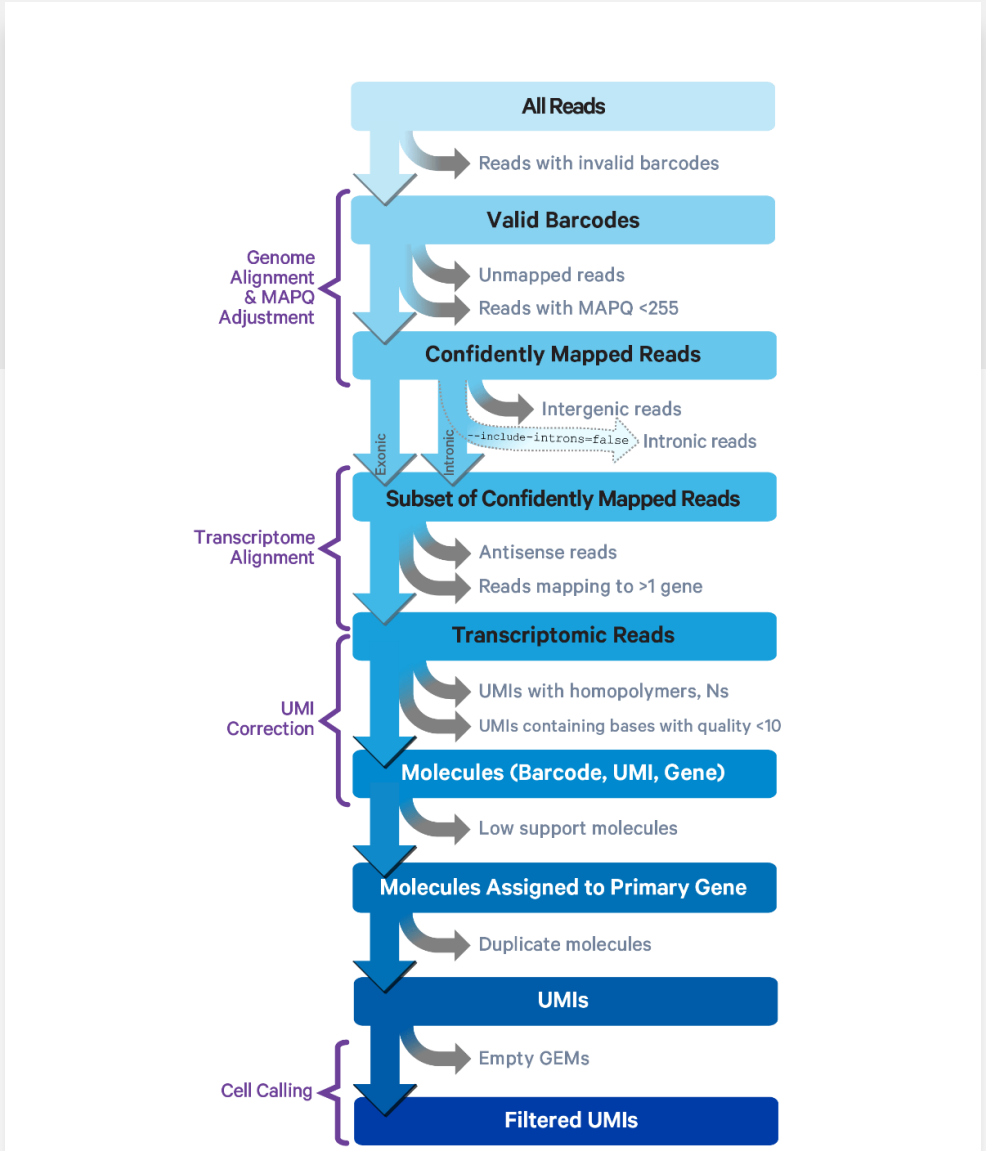


Make sure to upload fastqs from all flowcells on public databases so the results can be reproduced.

Image adapted from 10X promotional materials

Alignment and generate count matrix

- Cell Ranger count/ multi
- Input:
  - Fastqs
  - Transcriptome reference of species of interest (custom reference is supported, GFP, or CAR T sequence)
- The output will be for each GEM well that was demultiplexed separately



Cellranger aggr

- Aggregate multiple GEM wells from a single experiment that were analyzed by cellranger count and produces a single feature-barcode matrix containing all the data
- There are two modes:
  - **None:** Do not normalize at all. maximize sensitivity and plan to handle depth normalization in a downstream step
  - **Mapped (default)**: For each library type, subsample reads from higher-depth GEM wells until they all have, on average, an equal number of reads per cell that are confidently mapped to the transcriptome (Gene Expression) or assigned to known features (Feature Barcode Technology). This approach **avoids artifacts that may be introduced due to differences in sequencing depth.**

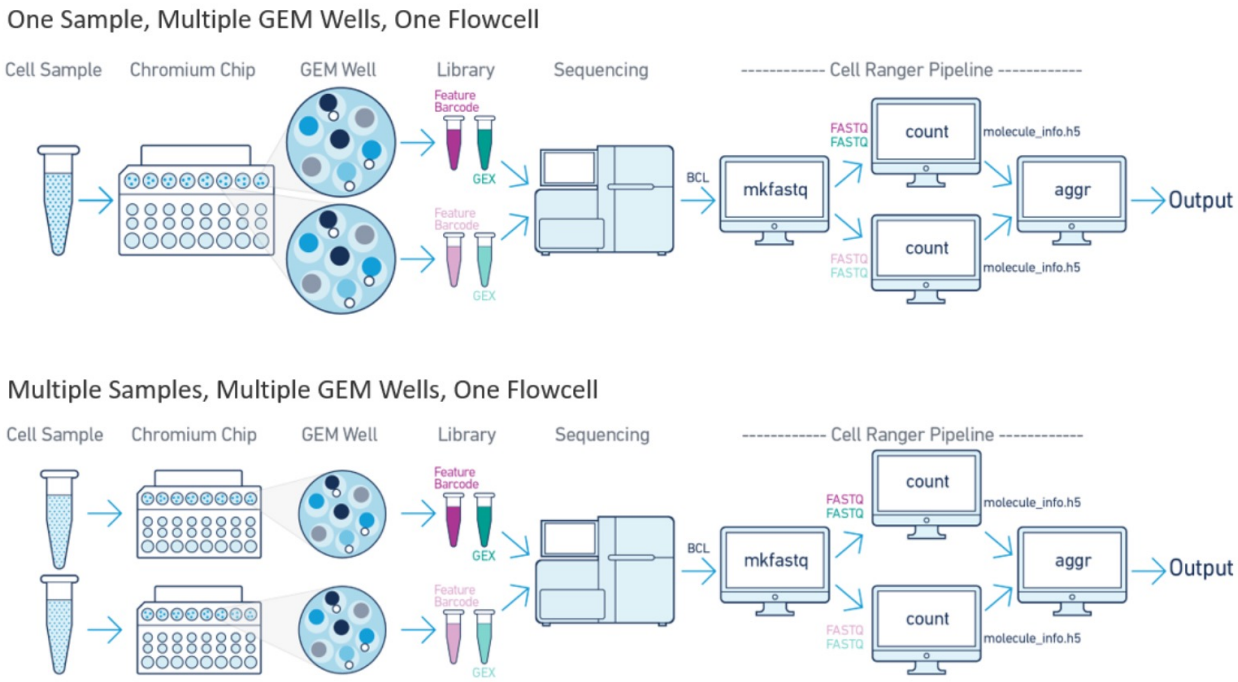
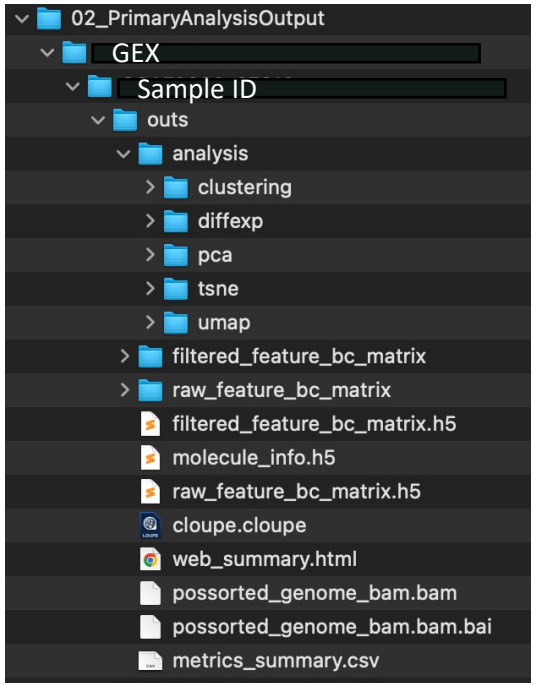


Image adapted from 10X promotional materials

Cellranger count output

- BAM
- BAM index
- Filtered feature-barcode matrices MEX/ HDF5
- Raw feature-barcode matrices MEX/ HDF5
- Per-molecule read information
- Run summary CSV
- Run summary HTML
- Loupe Browser file
- Secondary analysis output

```
├── analysis
│   ├── clustering
│   ├── diffexp
│   ├── pca
│   ├── tsne
│   └── umap
├── cloupe.cloupe
├── filtered_feature_bc_matrix
│   ├── barcodes.tsv.gz
│   ├── features.tsv.gz
│   └── matrix.mtx.gz
├── filtered_feature_bc_matrix.h5
├── metrics_summary.csv
├── molecule_info.h5
├── possorted_genome_bam.bam
├── possorted_genome_bam.bam.bai
├── raw_feature_bc_matrix
│   ├── barcodes.tsv.gz
│   ├── features.tsv.gz
│   └── matrix.mtx.gz
├── raw_feature_bc_matrix.h5
└── web_summary.html
```



```
├── possorted_genome_bam.bam
├── possorted_genome_bam.bam.bai
```

BAM and Bai file

An indexed BAM file containing **position-sorted reads** aligned to the genome and transcriptome, as well as unaligned reads

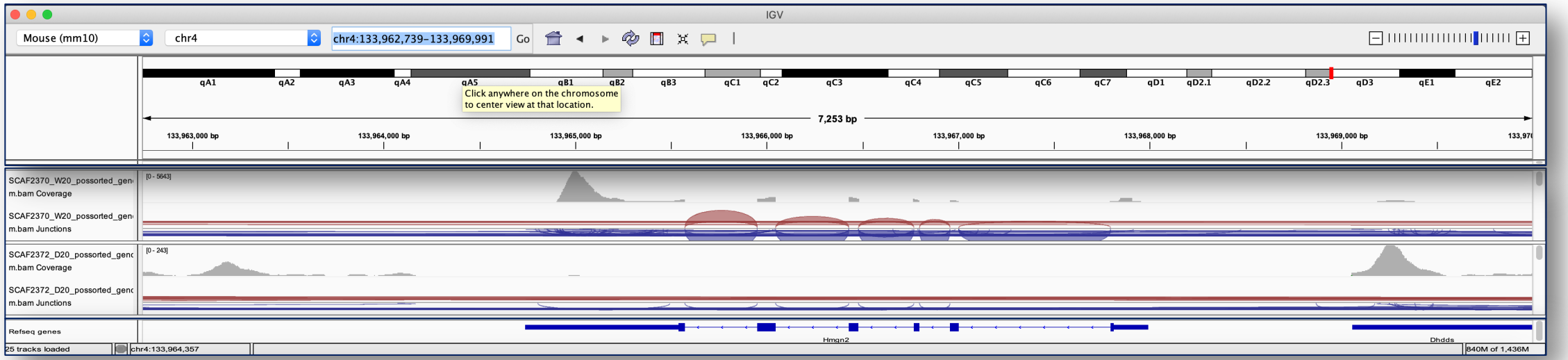
- Can be used to generate Fastq
  BAM ⟶ Fastq
  (bamtofastq available on cellranger suite)

- Can be imported to IGV (Integrative Genomics Viewer) e.g. for checking coverage of any gene of interest like knock out genes

Wild type

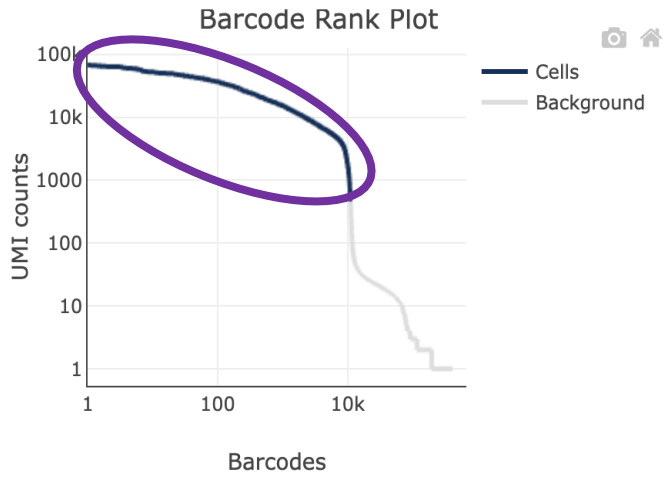
Knockout

HMGN2 gene



filtered_feature_bc_matrix

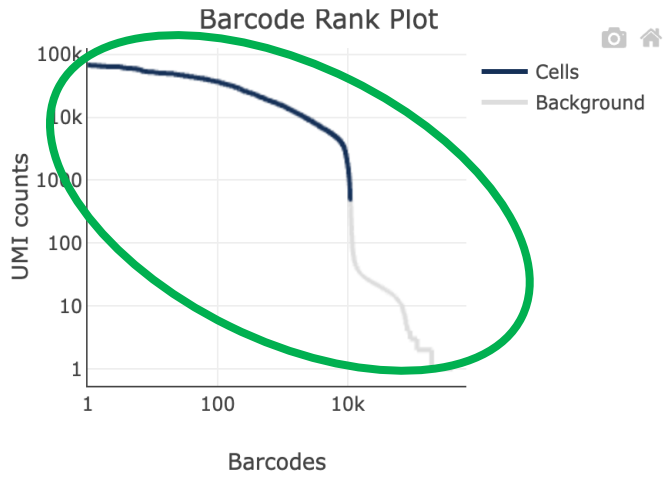
- Contains **only detected cell-associated barcodes**. Each element of the matrix is the number of UMIs associated with a feature (row) and a barcode (column.
- This file can be input into third-party packages and allows users to wrangle the barcode-feature matrix (e.g. to filter outlier cells, run dimensionality reduction, normalize gene expression).




```
├── filtered_feature_bc_matrix
│   ├── barcodes.tsv.gz
│   ├── features.tsv.gz
│   └── matrix.mtx.gz
├── filtered_feature_bc_matrix.h5
```

raw_feature_bc_matrix

- Contains **every barcode from the fixed list of known-good barcode sequences that has at least one read**. This includes background and cell-associated barcodes.
- Each element of the matrix is the number of UMIs associated with a feature (row) and a barcode (column)




```
├── raw_feature_bc_matrix
│   ├── barcodes.tsv.gz
│   ├── features.tsv.gz
│   └── matrix.mtx.gz
├── raw_feature_bc_matrix.h5
```

Two formats of feature barcode matrix

- Market Exchange Format (MEX)
- Hierarchical Data Format (HDF5)
  - H5 is a binary format that can compress and access data much more efficiently than text formats such as MEX, which is especially useful when dealing with large datasets. H5 files are supported in both R and Python.

```
(root)
    └── matrix [HDF5 group]
        ├── barcodes
        ├── data
        ├── indices
        ├── indptr
        ├── shape
        └── features [HDF5 group]
            ├─ _all_tag_keys
            ├─ target_sets [for Fixed RNA Profiling]
            │   └─ [target set name]
            ├─ feature_type
            ├─ genome
            ├─ id
            ├─ name
            ├─ pattern [Feature Barcode only]
            ├─ read [Feature Barcode only]
            └─ sequence [Feature Barcode only]
```

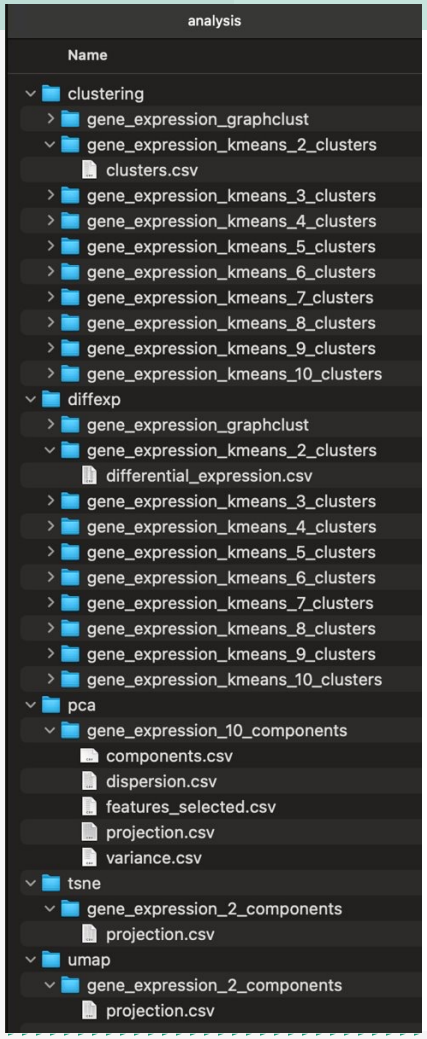
molecule_info.h5

- Contains per-molecule information for all molecules that contain a valid barcode, valid UMI, and were assigned with high confidence to a gene or Feature Barcode. This file is a required input to run cellranger aggr.

Metrics summary file in CSV format

Estimated Number of Cells	Mean Reads per Cell	Median Genes per Cell	Number of Reads	Valid Barcodes	Sequencing Saturation	Q30 Bases in Barcode	Q30 Bases in RNA Read	Q30 Bases in UMI	Reads Mapped to Genome	Reads Mapped Confidently to Genome	Reads Mapped Confidently to Intergenic Regions	Reads Mapped Confidently to Intronic Regions	Reads Mapped Confidently to Exonic Regions	Reads Mapped Confidently to Transcriptome	Reads Mapped Antisense to Gene	Fraction Reads in Cells	Total Genes Detected	Median UMI Counts per Cell
4,413	70,348	2,908	310,443,939	92.50%	74.80%	97.50%	94.40%	97.40%	93.60%	84.50%	7.50%	9.10%	67.90%	68.90%	7.50%	88.60%	45,040	7,953

These metrics are also available in html web summary



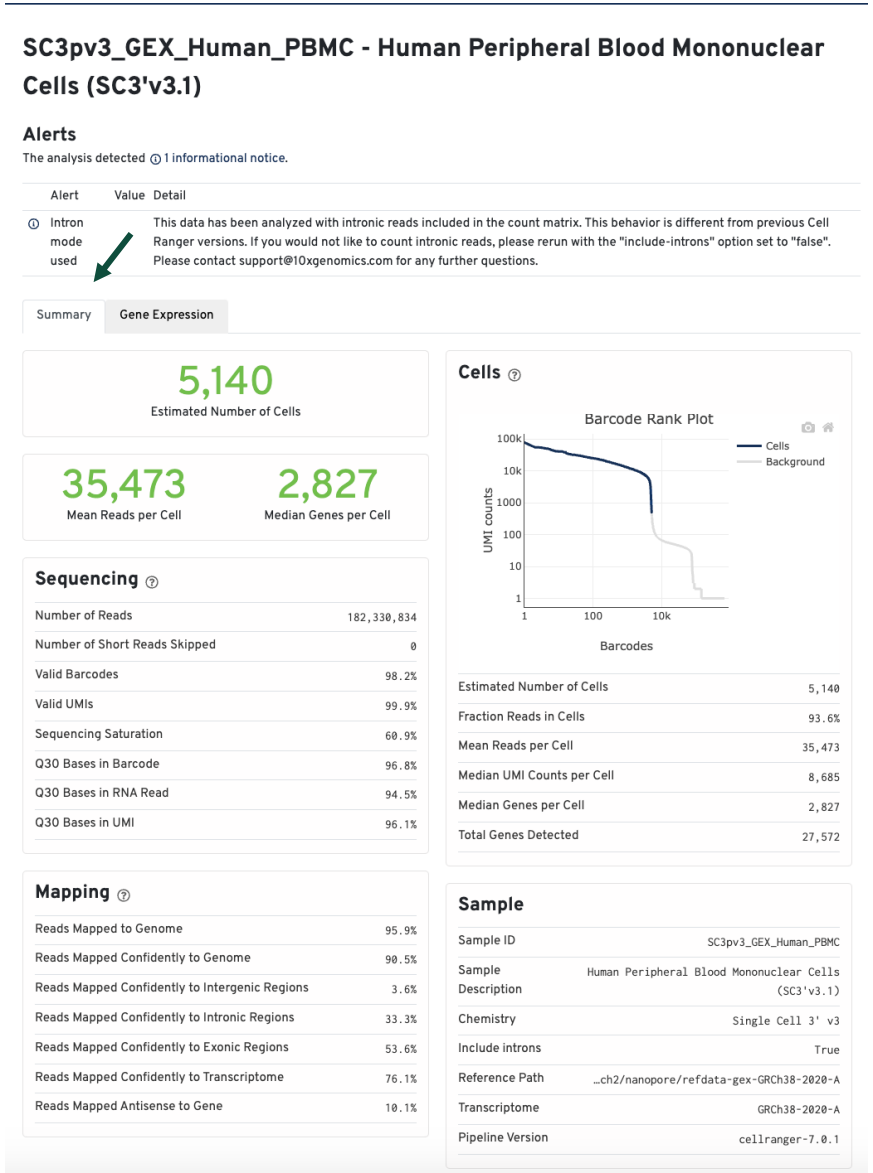

```
├── analysis
│   ├── clustering
│   ├── diffexp
│   ├── pca
│   ├── tsne
│   └── umap
```

Analysis folder

- Graph-based clusters
- K-means clustering 2-10
- Differential gene expression analysis between clusters
- PCA, t-SNE, and UMAP dimensionality reduction

web_summary.html

- Run summary metrics and charts in HTML format
- **A great place to start assessing the quality of your data**.
- Several metrics in the web summary file can be used to assess the overall success of an experiment, including sequencing, mapping, and cell metrics.



SC3pv3_GEX_Human_PBMC - Human Peripheral Blood Mononuclear Cells (SC3'v3.1)

Alerts

The analysis detected 1 informational notice.

Alert	Value	Detail
Intron mode used		This data has been analyzed with intronic reads included in the count matrix. This behavior is different from previous Cell Ranger versions. If you would not like to count intronic reads, please rerun with the "include-introns" option set to "false". Please contact support@10xgenomics.com for any further questions.

Summary | Gene Expression

5,140
Estimated Number of Cells

35,473
Mean Reads per Cell

2,827
Median Genes per Cell

Sequencing

Number of Reads	182,330,834
Number of Short Reads Skipped	0
Valid Barcodes	98.2%
Valid UMIs	99.9%
Sequencing Saturation	60.9%
Q30 Bases in Barcode	96.8%
Q30 Bases in RNA Read	94.5%
Q30 Bases in UMI	96.1%

Mapping

Reads Mapped to Genome	95.9%
Reads Mapped Confidently to Genome	90.5%
Reads Mapped Confidently to Intergenic Regions	3.6%
Reads Mapped Confidently to Intronic Regions	33.3%
Reads Mapped Confidently to Exonic Regions	53.6%
Reads Mapped Confidently to Transcriptome	76.1%
Reads Mapped Antisense to Gene	10.1%

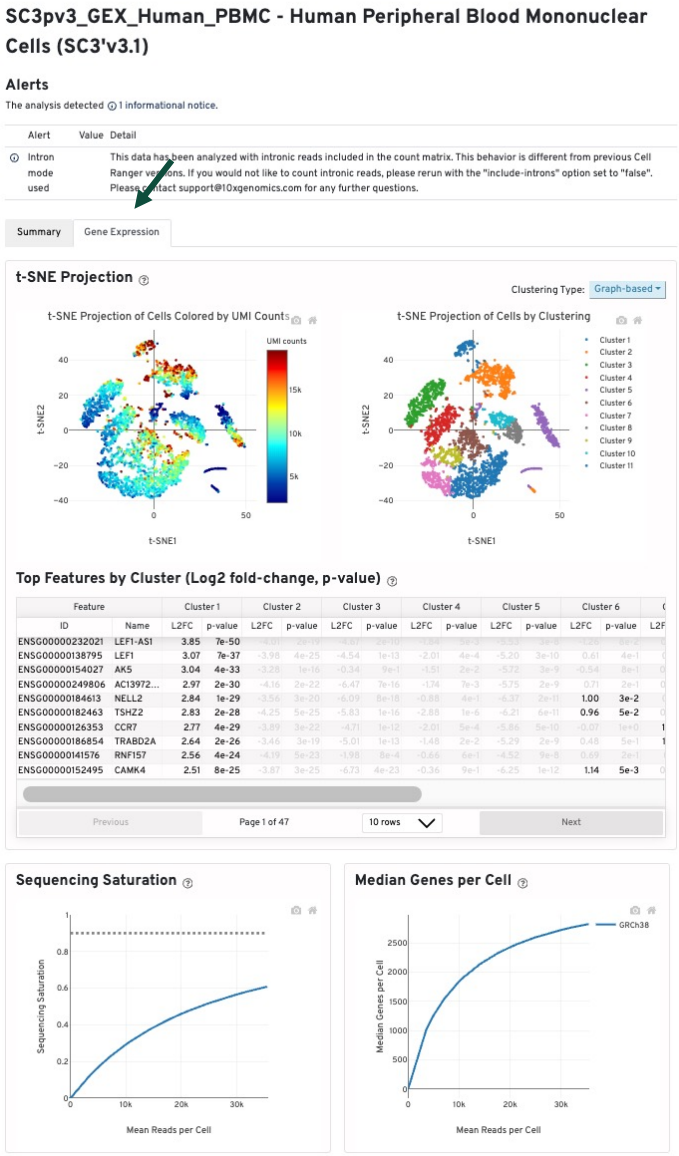
Cells

Barcode Rank Plot

Estimated Number of Cells	5,140
Fraction Reads in Cells	93.6%
Mean Reads per Cell	35,473
Median UMI Counts per Cell	8,685
Median Genes per Cell	2,827
Total Genes Detected	27,572

Sample

Sample ID	SC3pv3_GEX_Human_PBMC
Sample Description	Human Peripheral Blood Mononuclear Cells (SC3'v3.1)
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	…ch2/nanopore/refdata-gex-GRCh38-2020-A
Transcriptome	GRCh38-2020-A
Pipeline Version	cellranger-7.0.1



SC3pv3_GEX_Human_PBMC - Human Peripheral Blood Mononuclear Cells (SC3'v3.1)

Alerts

The analysis detected 1 informational notice.

Alert	Value	Detail
Intron mode used		This data has been analyzed with intronic reads included in the count matrix. This behavior is different from previous Cell Ranger versions. If you would not like to count intronic reads, please rerun with the "include-introns" option set to "false". Please contact support@10xgenomics.com for any further questions.

Summary | Gene Expression

t-SNE Projection

Clustering Type: Graph-based

Top Features by Cluster (Log2 fold-change, p-value)

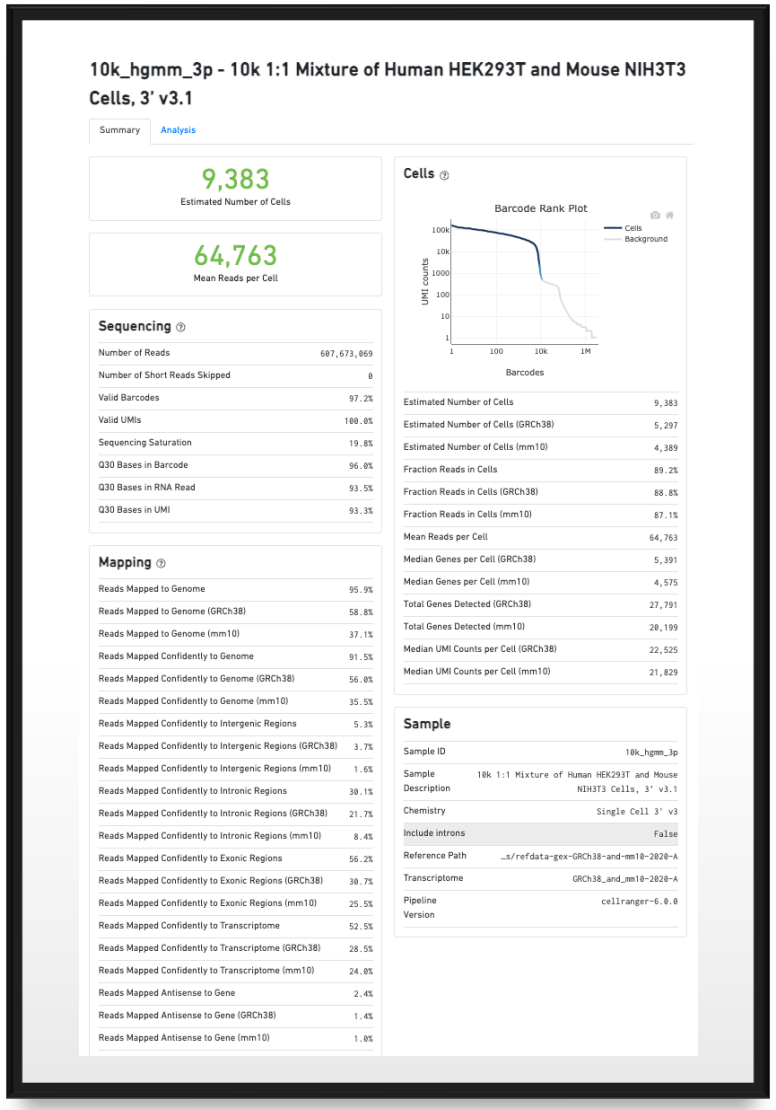
Feature		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
ID	Name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value
ENSG00000232021	LEF1-AS1	3.85	7e-50										
ENSG00000138795	LEF1	3.07	7e-37	-3.98	4e-25	-4.54	1e-13	-2.01	4e-4	-5.20	3e-10	0.61	4e-1
ENSG00000154027	AK5	3.04	4e-33	-3.28	1e-16	-0.34	9e-1	-1.51	2e-2	-5.72	3e-9	-0.54	8e-1
ENSG00000249806	AC13972…	2.97	2e-30	-4.16	2e-22	-6.47	7e-16	-1.74	7e-3	-5.75	2e-9	0.71	2e-1
ENSG00000184613	NELL2	2.84	1e-29	-3.56	3e-20	-6.09	8e-18	-0.88	4e-1	-6.37	2e-11	1.00	3e-2
ENSG00000182463	TSHZ2	2.83	2e-28	-4.25	5e-25	-5.83	1e-16	-2.88	1e-6	-6.21	6e-11	0.96	5e-2
ENSG00000126353	CCR7	2.77	4e-29	-3.89	3e-22	-4.71	1e-12	-2.01	5e-4	-5.86	5e-10	-0.07	1e+0
ENSG00000186854	TRABD2A	2.64	2e-26	-3.46	3e-19	-5.01	1e-13	-1.48	2e-2	-5.29	2e-9	0.48	5e-1
ENSG00000141576	RNF157	2.56	4e-24	-4.19	5e-23	-1.98	8e-4	-0.66	6e-1	-4.52	9e-8	0.69	2e-1
ENSG00000152495	CAMK4	2.51	8e-25	-3.87	3e-25	-6.73	4e-23	-0.36	9e-1	-6.25	1e-12	1.14	5e-3

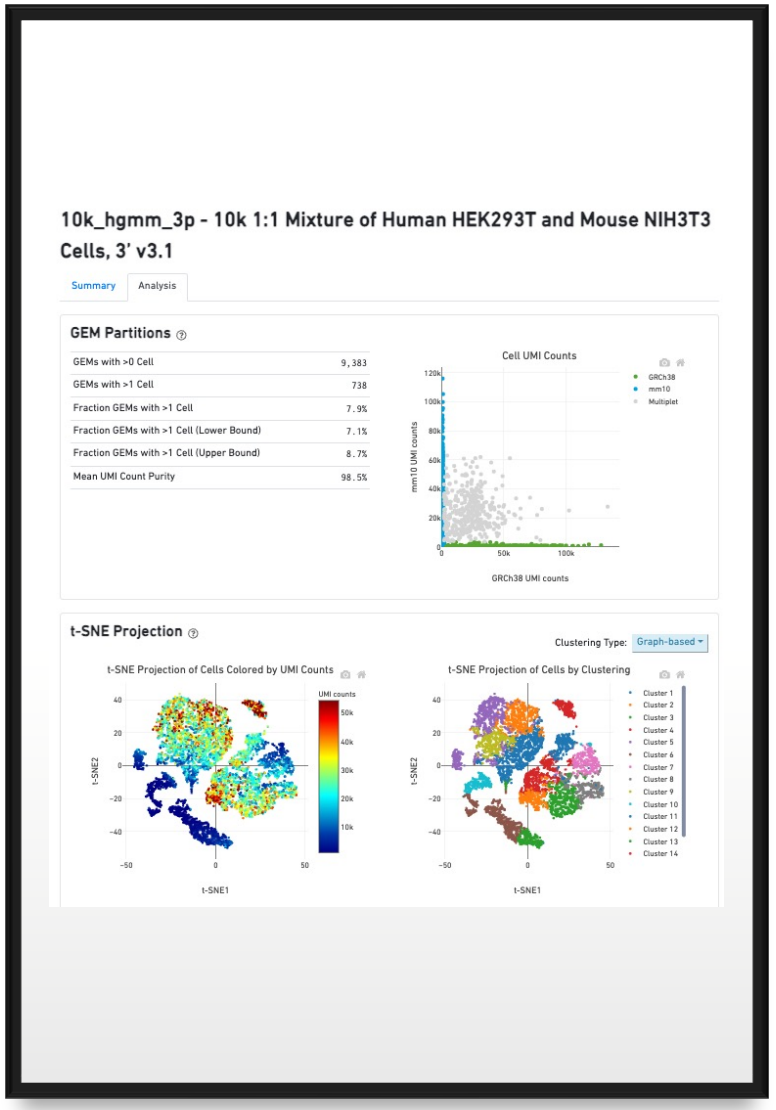
Previous | Page 1 of 47 | 10 rows | Next

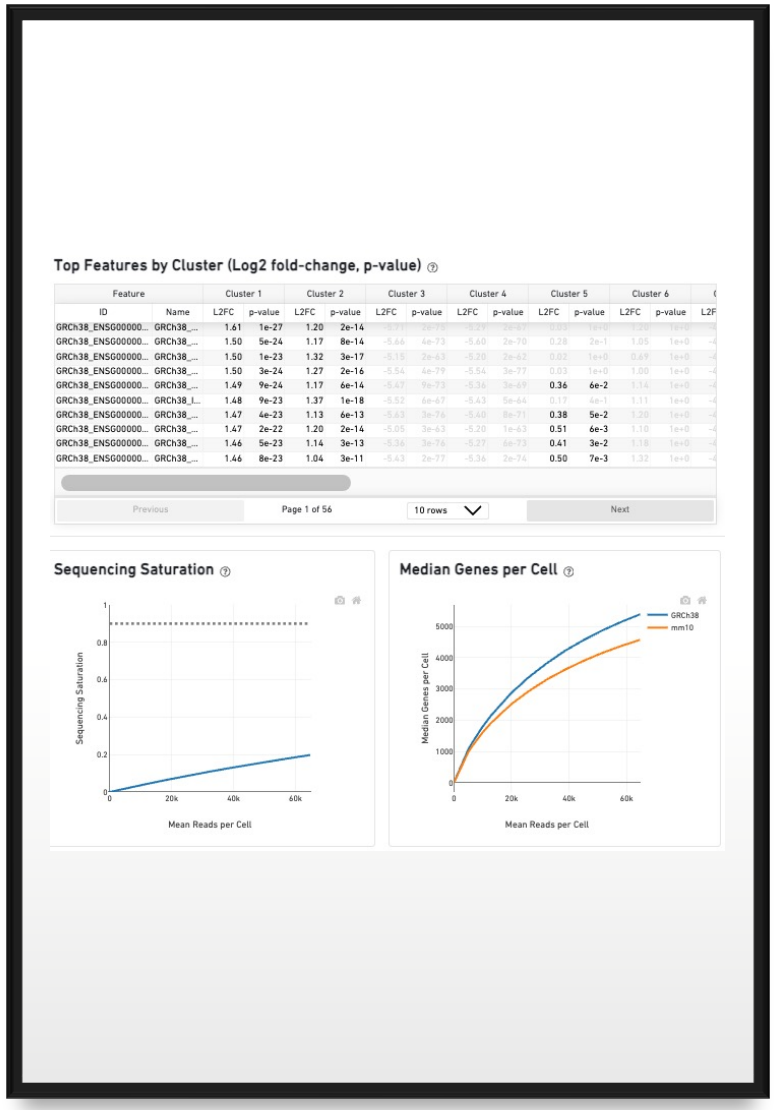
Sequencing Saturation

Median Genes per Cell

Web summary of single species



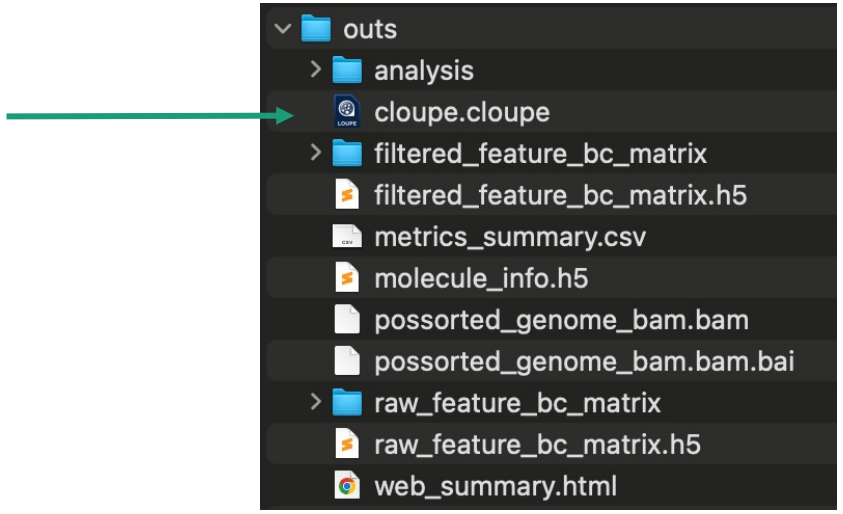




Web summary of combined human and mouse

cloupe.cloupe

- The input file for loupe browser
- Can be uploaded to Loupe Browser, a free desktop visualization software from 10x that provides the intuitive functionality to explore and analyze 10x Genomics Chromium and Visium data.
- You can also convert your Seurat objects into Loupe Browser files using the LoupeR package.



Using cellranger output for quality Control of single cell data

SC3pv3_GEX_Human_PBMC - Human Peripheral Blood Mononuclear Cells (SC3'v3.1)

Alerts

The analysis detected 1 informational notice.

Alert	Value	Detail
Intron mode used		This data has been analyzed with intronic reads included in the count matrix. This behavior is different from previous Cell Ranger versions. If you would not like to count intronic reads, please rerun with the "include-introns" option set to "false". Please contact support@10xgenomics.com for any further questions.

Summary | Gene Expression

5,140
Estimated Number of Cells

35,473
Mean Reads per Cell

2,827
Median Genes per Cell

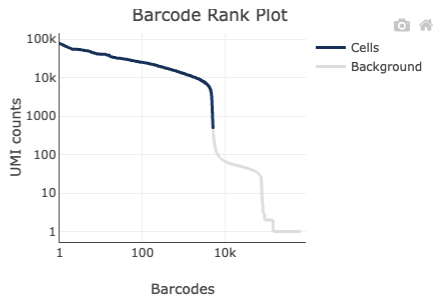
Sequencing

Number of Reads	182,330,834
Number of Short Reads Skipped	0
Valid Barcodes	98.2%
Valid UMIs	99.9%
Sequencing Saturation	60.9%
Q30 Bases in Barcode	96.8%
Q30 Bases in RNA Read	94.5%
Q30 Bases in UMI	96.1%

Mapping

Reads Mapped to Genome	95.9%
Reads Mapped Confidently to Genome	90.5%
Reads Mapped Confidently to Intergenic Regions	3.6%
Reads Mapped Confidently to Intronic Regions	33.3%
Reads Mapped Confidently to Exonic Regions	53.6%
Reads Mapped Confidently to Transcriptome	76.1%
Reads Mapped Antisense to Gene	10.1%

Cells



Estimated Number of Cells	5,140
Fraction Reads in Cells	93.6%
Mean Reads per Cell	35,473
Median UMI Counts per Cell	8,685
Median Genes per Cell	2,827
Total Genes Detected	27,572

Sample

Sample ID	SC3pv3_GEX_Human_PBMC
Sample Description	Human Peripheral Blood Mononuclear Cells (SC3'v3.1)
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	…ch2/nanopore/refdata-gex-GRCh38-2020-A
Transcriptome	GRCh38-2020-A
Pipeline Version	cellranger-7.0.1

SC3pv3_GEX_Human_PBMC - Human Peripheral Blood Mononuclear Cells (SC3'v3.1)

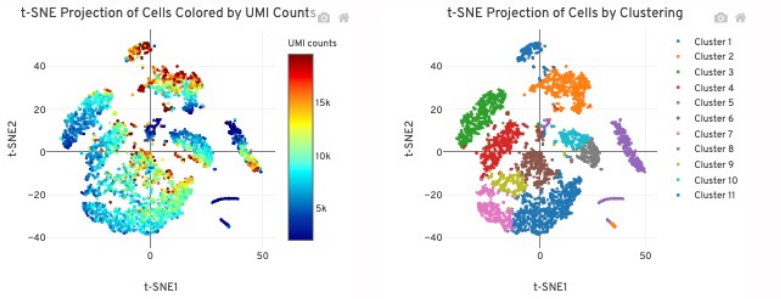
Alerts

The analysis detected 1 informational notice.

Alert	Value	Detail
Intron mode used		This data has been analyzed with intronic reads included in the count matrix. This behavior is different from previous Cell Ranger versions. If you would not like to count intronic reads, please rerun with the "include-introns" option set to "false". Please contact support@10xgenomics.com for any further questions.

Summary | Gene Expression

t-SNE Projection

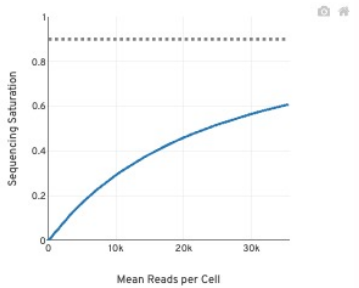


Top Features by Cluster (Log2 fold-change, p-value)

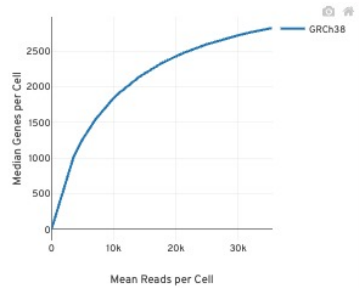
Feature		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
ID	Name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value
ENSG00000232021	LEF1-AS1	3.85	7e-50										
ENSG00000138795	LEF1	3.07	7e-37	-3.98	4e-25	-4.54	1e-13	-2.01	4e-4	-5.20	3e-10	0.61	4e-1
ENSG00000154027	AK5	3.04	4e-33	-3.28	1e-16	-0.34	9e-1	-1.51	2e-2	-5.72	3e-9	-0.54	8e-1
ENSG00000249806	AC13972…	2.97	2e-30	-4.16	2e-22	-6.47	7e-16	-1.74	7e-3	-5.75	2e-9	0.71	2e-1
ENSG00000184613	NELL2	2.84	1e-29	-3.56	3e-20	-6.09	8e-18	-0.88	4e-1	-6.37	2e-11	1.00	3e-2
ENSG00000182463	TSHZ2	2.83	2e-28	-4.25	5e-25	-5.83	1e-16	-2.88	1e-6	-6.21	6e-11	0.96	5e-2
ENSG00000126353	CCR7	2.77	4e-29	-3.89	3e-22	-4.71	1e-12	-2.01	5e-4	-5.86	5e-10	-0.07	1e+0
ENSG00000186854	TRABD2A	2.64	2e-26	-3.46	3e-19	-5.01	1e-13	-1.48	2e-2	-5.29	2e-9	0.48	5e-1
ENSG00000141576	RNF157	2.56	4e-24	-4.19	5e-23	-1.98	8e-4	-0.66	6e-1	-4.52	9e-8	0.69	2e-1
ENSG00000152495	CAMK4	2.51	8e-25	-3.87	3e-25	-6.73	4e-23	-0.36	9e-1	-6.25	1e-12	1.14	5e-3

Previous | Page 1 of 47 | 10 rows | Next

Sequencing Saturation

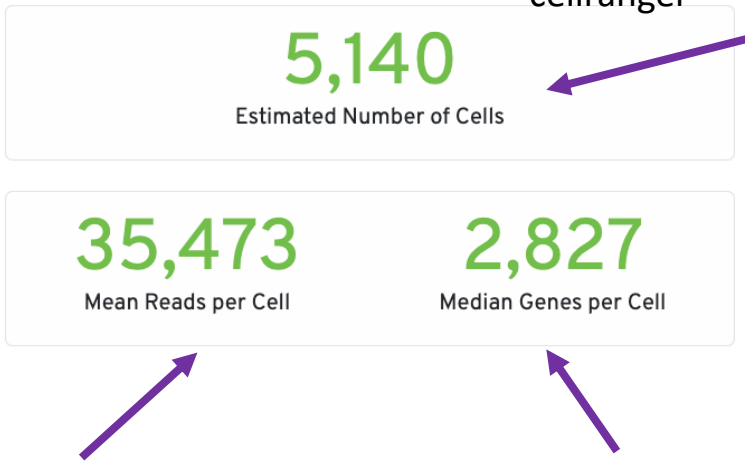


Median Genes per Cell



Web summary of single species

Number of cells picked up by cellranger



Sequencing depth for the sample

Sensitivity

Top 3 main metrics on top of web summary

Target cell number

Multiplet Rate	# of Cells Recovered	
	Next GEM	GEM-X
0.8%	1,000	2,000
1.6%	2,000	4,000
2.4%	3,000	6,000
3.2%	4,000	8,000
4.0%	5,000	10,000
4.8%	6,000	12,000
5.6%	7,000	14,000
6.4%	8,000	16,000
7.2%	9,000	18,000
8.0%	10,000	20,000

The target and cells detected by pipeline is not always the same and it depends on the quality of the cell/nuclei prep and how it affects cell counting

Sequencing and mapping metrics

Sequencing

Number of Reads	182,330,834
Number of Short Reads Skipped	0
Valid Barcodes	98.2%
Valid UMIs	99.9%
Sequencing Saturation	60.9%
Q30 Bases in Barcode	96.8%
Q30 Bases in RNA Read	94.5%
Q30 Bases in UMI	96.1%

Mapping

Reads Mapped to Genome	95.9%
Reads Mapped Confidently to Genome	90.5%
Reads Mapped Confidently to Intergenic Regions	3.6%
Reads Mapped Confidently to Intronic Regions	33.3%
Reads Mapped Confidently to Exonic Regions	53.6%
Reads Mapped Confidently to Transcriptome	76.1%
Reads Mapped Antisense to Gene	10.1%

Metrics	Definition	Expected Value	Notes
Sequencing Metrics			
Number of reads	Total number of read pairs that were assigned to this library	Sequencing output dependent	Lower than expected may indicate poor sequencing run (over-clustering, under-clustering, low % passing filter).
Valid barcodes	Fraction of reads with barcodes that match the whitelist fraction of reads originating from an already observed UMI	>75%	Low valid barcodes may indicate sequencing issues (such as low Read 1 Q30 score).
Valid UMIs	Fraction of reads with valid UMIs; i.e. UMI sequences that do not contain Ns and that are not homopolymers	>75%	Low valid UMIs may indicate issues with sequencing or library quality.
Sequencing saturation	The fraction of reads originating from an already-observed UMI. This is a function of library complexity and sequencing depth	Dependent upon sequencing depth and sample complexity	Dependent on library complexity, sequencing depth, and experiment analysis goals. Lower sequencing saturation indicates a high proportion of the library complexity has not been captured by sequencing.
Q30 bases in barcode, Sample Index, or UMI	Fraction of cell barcode, Sample Index, or UMI bases with Q-score ≥30, excluding very low quality/ no call (Q-score ≤2) bases from the denominator	Sequencing platform dependent	Low Q30 base percentages could indicate sequencing issue such as sub-optimal loading concentration.
Q30 bases in RNA read	Fraction of RNA read bases with Q-score ≥30, excluding very low quality/no-call (Q-score ≤ 2) bases from the denominator	Sequencing platform dependent, ideally >65%	Expected to be lower than Q30 Bases in Barcode or UMI (Read 1) or Sample Index (i7 read) and is sequencing platform dependent. Consult Technical Note – Chromium Single Cell 3' v2 Libraries – Sequencing Metrics for Illumina Sequencers (v2 Chemistry) Document CG000089 for more information. Low Q30 Base percentages could indicate sequencing issue such as sub-optimal loading concentration.

Metrics	Definition	Expected Value	Notes
Mapping Metrics			
Reads mapped to genome	Fraction of reads that are mapped to the genome	Variable	Dependent on the quality of genome annotation. Lower than expected values could be an indication of incorrect reference selection or library quality.
Reads mapped confidently to genome	Fraction of reads that mapped uniquely to a genome. A gene mapped to exonic loci from a single gene and also to non-exonic loci is considered uniquely mapped to one of the exonic loci	Variable	Lower than expected values could be indicative of low library quality or reference quality.
Reads mapped confidently to intergenic regions	Fraction of reads that mapped uniquely to an intergenic region of the genome	Variable	May vary based on sample type and genome annotation.
Reads mapped confidently to intronic regions	Fraction of reads that mapped uniquely to an intronic region of the genome	Variable	Sample types with low RNA content (e.g. PBMCs, nuclei) or samples with suboptimal health may have a higher fraction of reads mapping to intronic regions.
Reads mapped confidently to exonic regions	Fraction of reads that mapped uniquely to an exonic region of the genome	Variable	There is a balance between exonic and intronic reads. A sample with higher exonic reads will have lower intronic reads, and vice versa. This is highly dependent upon sample type.
Reads mapped confidently to transcriptome	Fraction of reads that mapped to a unique gene in the transcriptome. The read must be consistent with annotated splice junctions. These reads are considered for UMI counting	Variable, ideally >30%	Reference quality and sequencing configuration (shorter than recommended cycles on Read 2) can impact mapping. Lower than expected values may indicate the use of the wrong reference transcriptome.
Reads mapped antisense to gene	Fraction of reads confidently mapped to the transcriptome, but on the opposite strand of their annotated gene. A read is counted as antisense if it has any alignments that are consistent with an exon of a transcript but antisense to it, and has no sense alignments	Ideal <10%	These values may be higher if using a pre-mRNA reference or may indicate incorrect Gel Bead chemistry.

Cells metrics

Estimated Number of Cells	5,140
Fraction Reads in Cells	93.6%
Mean Reads per Cell	35,473
Median UMI Counts per Cell	8,685
Median Genes per Cell	2,827
Total Genes Detected	27,572

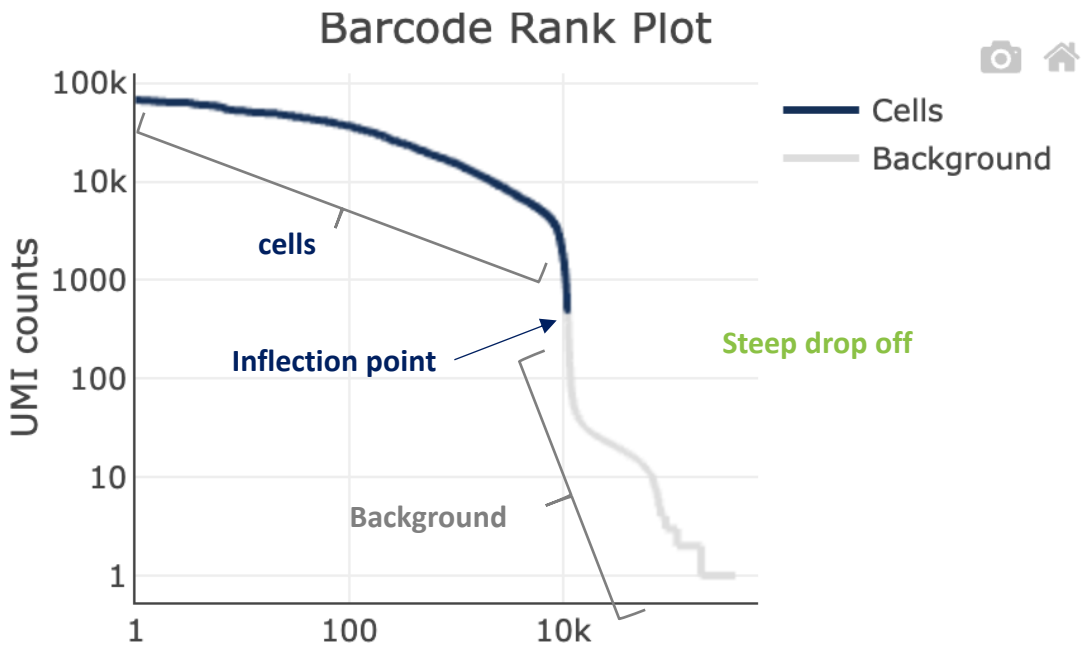
Cell Metrics			
Estimated number of cells	The number of barcodes associated with at least one cell	500-10,000	Higher or lower than expected values may indicate inaccurate cell count, cell lysis, or failures during GEM generation.
Fraction reads in cells	The fraction of reads that contain a valid barcode, are confidently mapped to the transcriptome and are associated with a barcode that is called as a cell	>70%	Lower percentages indicate that a high level of ambient RNA partitioned into all (cell-containing and non-cell-containing) GEMs.
Median reads per cell	The total number of sequenced reads divided by the number of barcodes associated with cell-containing partitions	User defined; 20,000 reads/ cell minimum recommended	The necessary sequencing depth per cell depends on the cell type (high or low RNA) and the desired analysis.
Median genes per cell	The median number of genes detected per cell-associated barcode. Detection is defined as the presence of at least 1 UMI count	Dependent on cell type and sequencing depth	Lower than expected median genes per cell may be biological (low transcriptional diversity) or may indicate low sequencing depth or library complexity.
Total genes detected	The number of genes with at least one UMI count in any cell	Dependent on cell type and sequencing depth	Lower than expected could be a result of shallower sequencing depth and/or sample/library quality.
Median UMI counts per cell	The median number of UMI dependent on cell counts per cell-associated type barcode	Dependent on cell type and sequencing depth	Lower than expected could be a result of shallower sequencing depth and/or sample/library quality.

Sample

Sample ID	SC3pv3_GEX_Human_PBMC
Sample Description	Human Peripheral Blood Mononuclear Cells (SC3'v3.1)
Chemistry	Single Cell 3' v3
Include introns	True
Reference Path	…ch2/nanopore/refdata-gex-GRCh38-2020-A
Transcriptome	GRCh38-2020-A
Pipeline Version	cellranger-7.0.1

Barcode Rank Plot

A distribution of cell barcodes ranked according to the number of unique molecular identifiers (UMIs) that are associated with a given barcode



UMI counts on logarithmic scale

Barcodes sorted in decreasing order from number of UMIs in Logarithmic scale

droplet-based microfluidic devices: **the majority of the droplets will not contain an actual cell.**

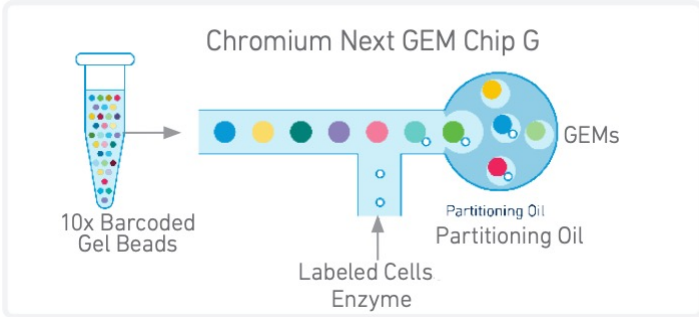


Image adapted from 10X promotional materials

Overview of cellranger cell calling algorithm

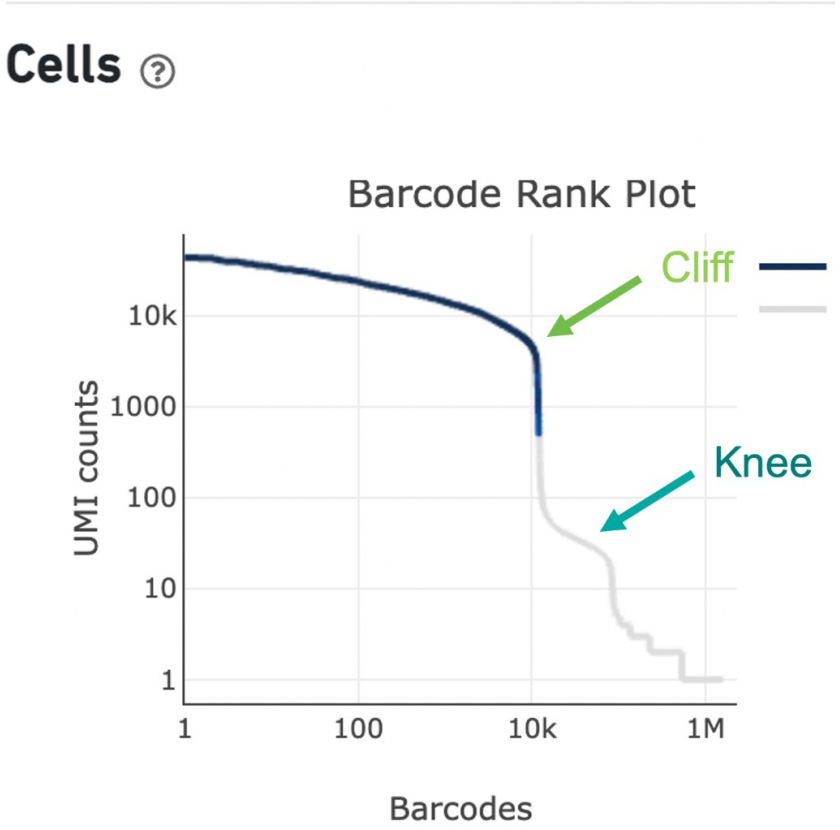
A multi-step process that determines which barcodes/GEMs are likely to contain an intact cell and uses those for downstream analysis. The cell calling algorithm can be broadly divided into two major steps:

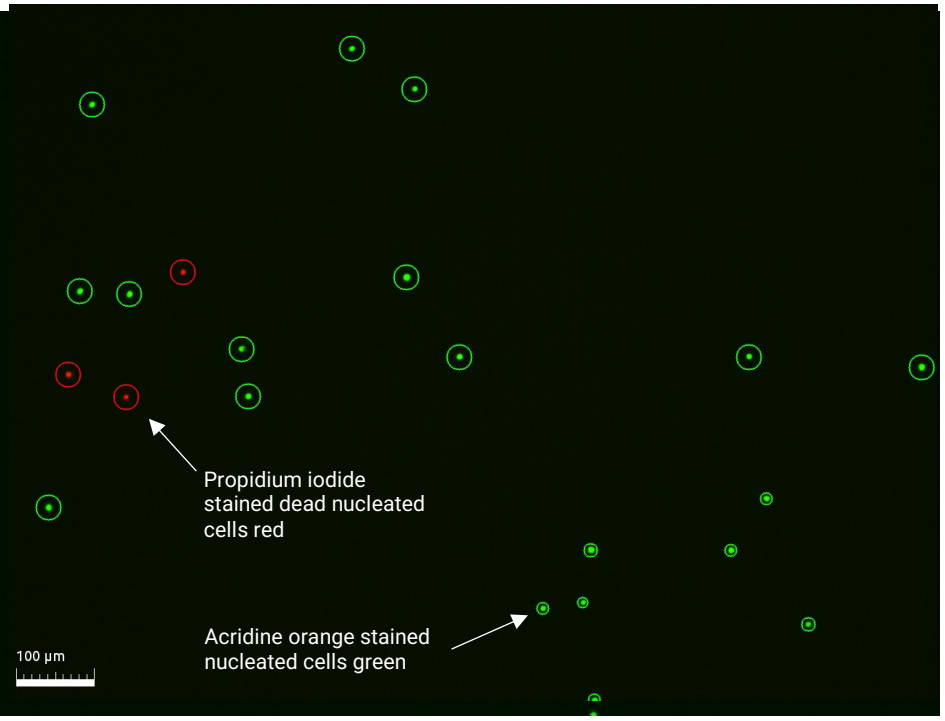
Step 1. Identify barcodes/GEMs that are likely to contain an intact cell based on the expected cell number and UMI counts.

Step 2. Distinguish low RNA content cells from empty droplets based on the **expression profiles** using the EmptyDrops method.

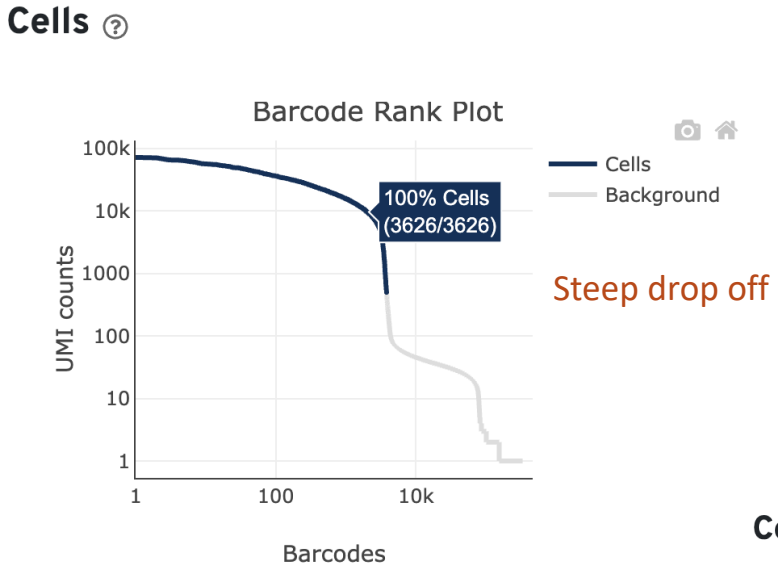
Optimal Barcode Rank Plot

- The overall shape of the Barcode Rank Plot is a useful indicator of sample quality. A “cliff-and-knee” shape in the Barcode Rank Plot is indicative of a good quality sample.
- In this case, the steep cliff, followed by the plateaued knee, demonstrates that the cell calling algorithm was able to distinguish between intact cells and background barcodes.

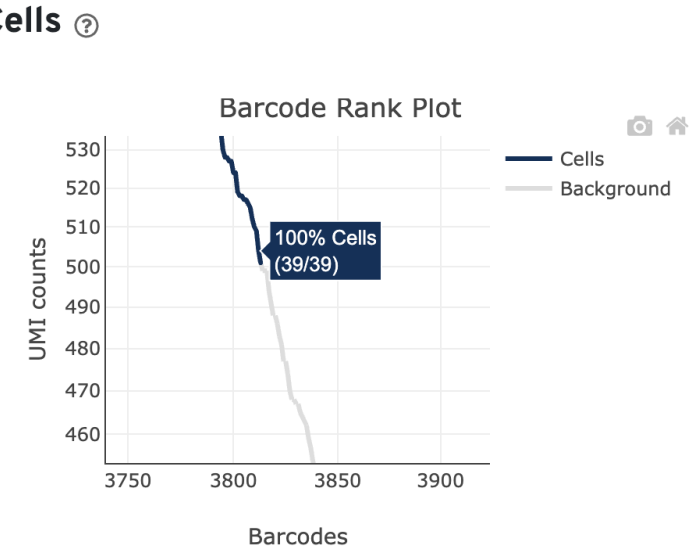


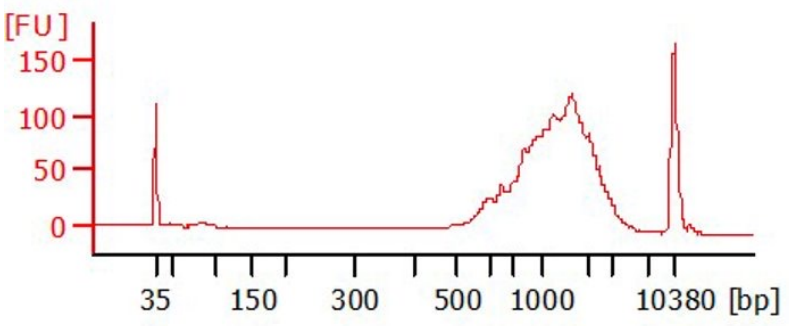


Optimal barcode rank plot

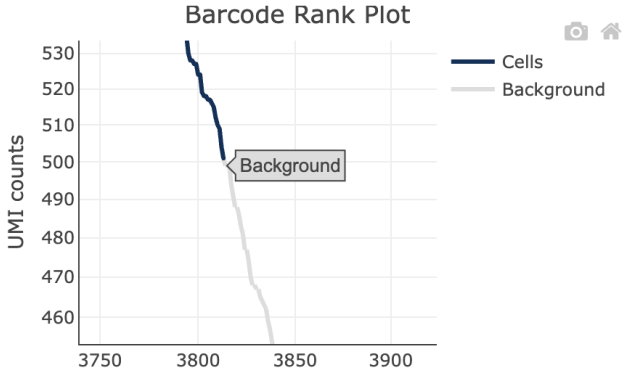


Steep drop off

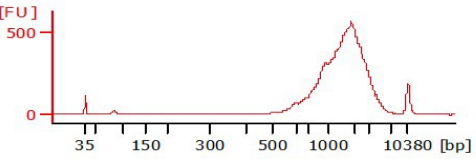


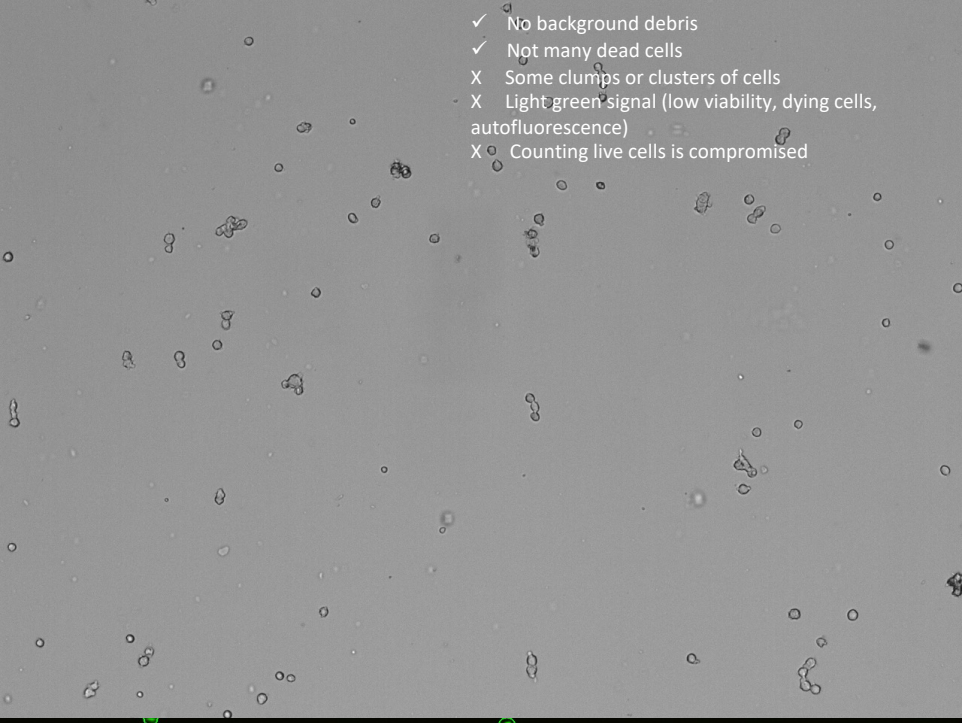


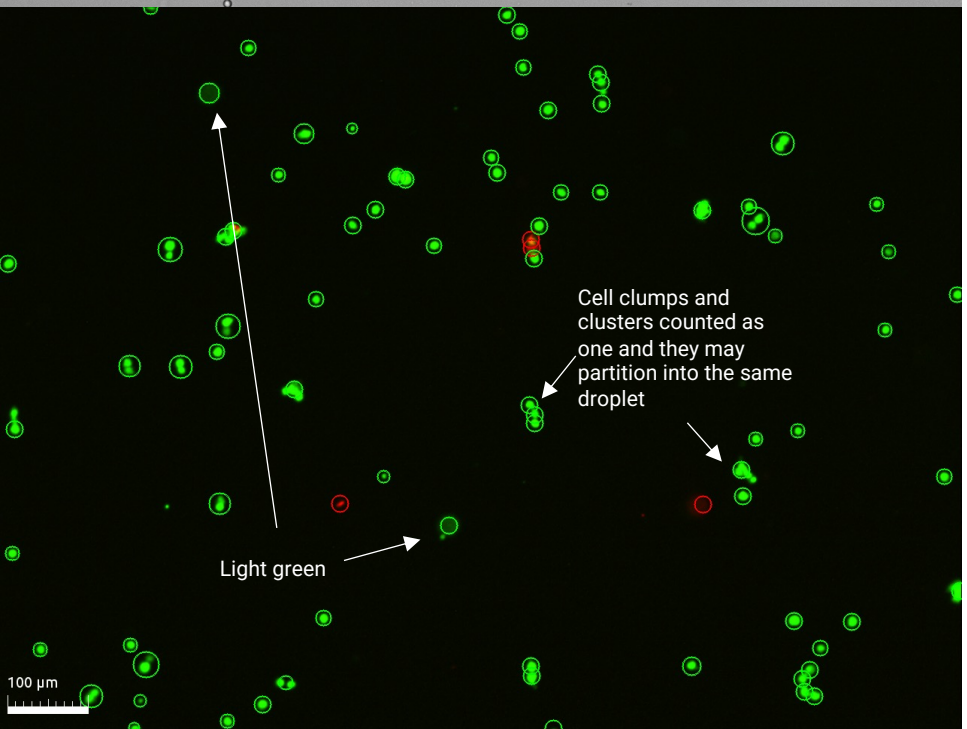
cDNA profile



Example of an optimal cDNA



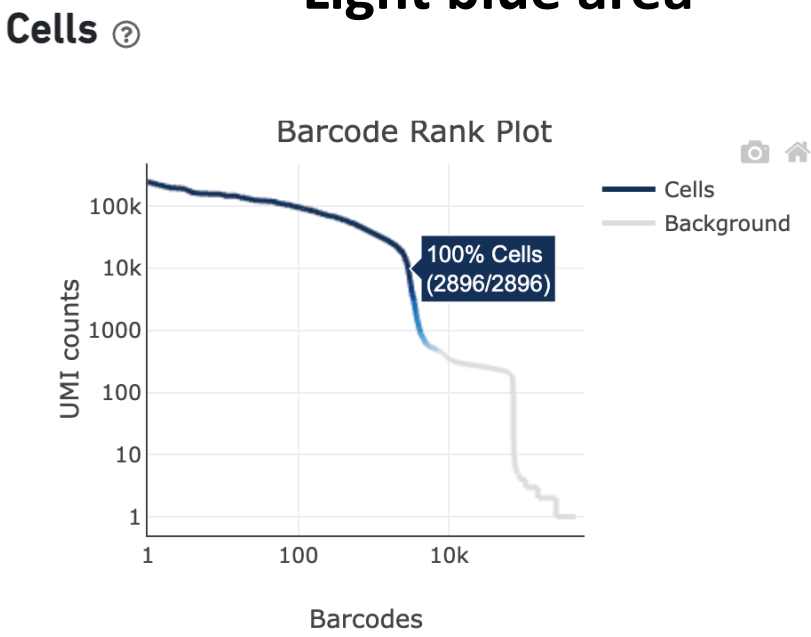




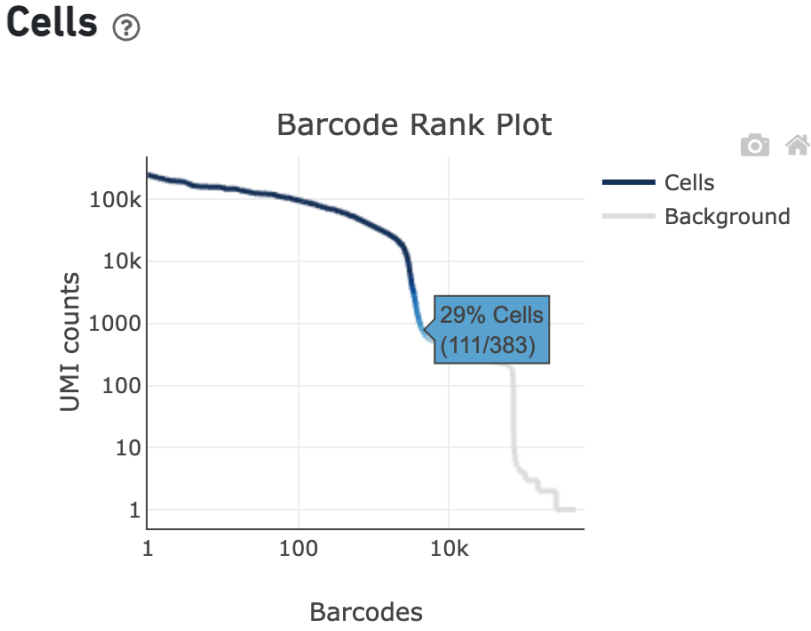
- Steep drop off, however, there is a range of UMI that some called as cell-associated GEMS and some as background.

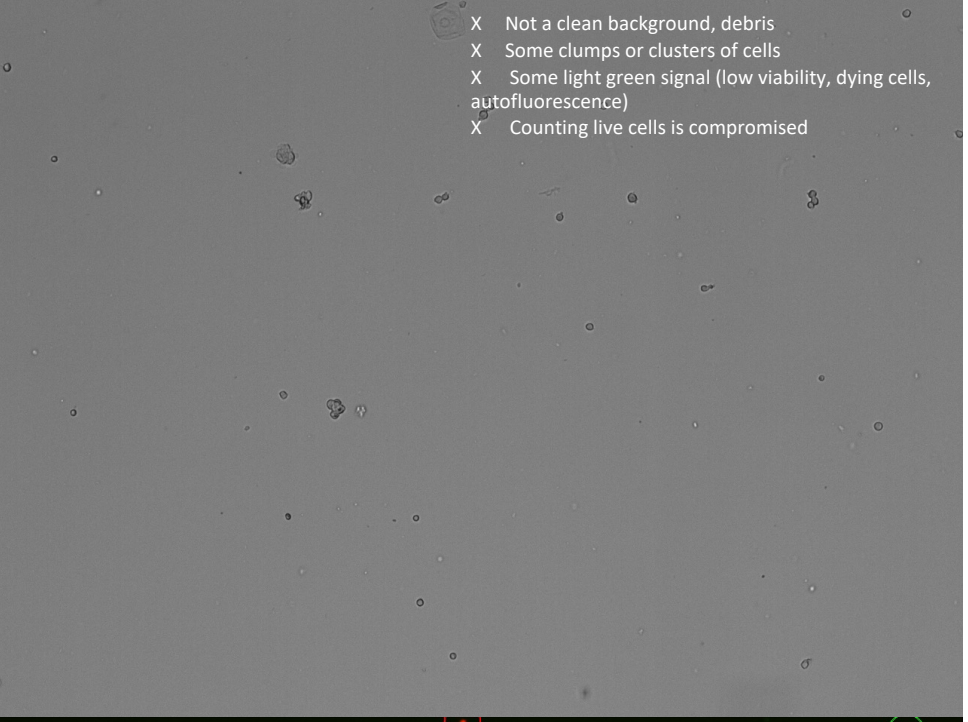
Light blue area

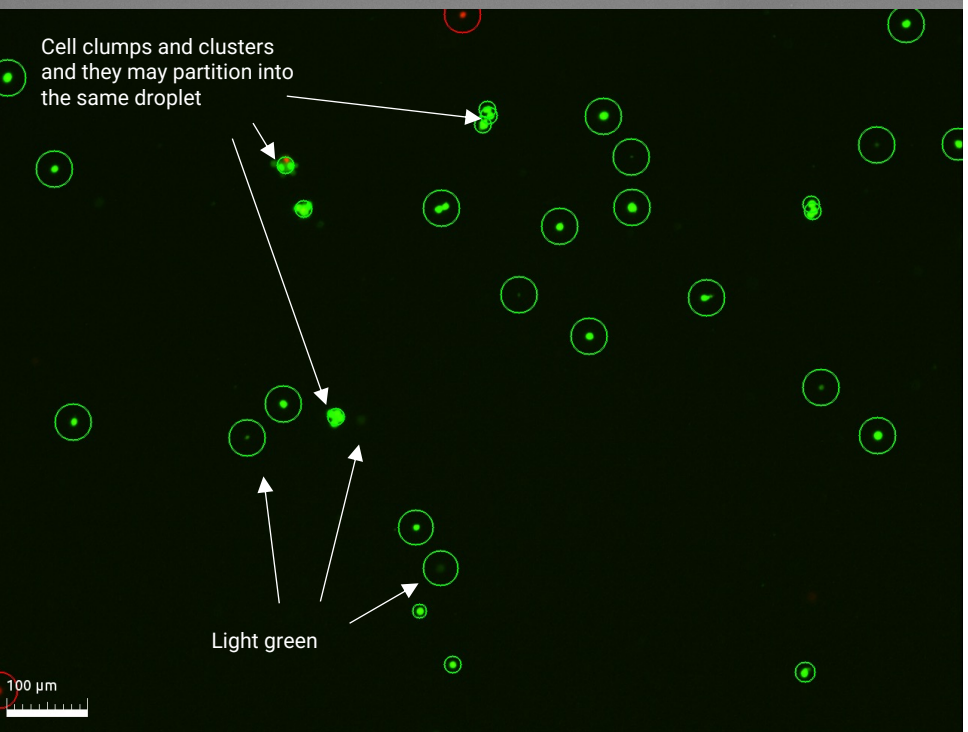
2,896 out of 2,896 barcodes in this UMI range were called as cells.



111 out of 383 barcodes in this UMI range were called as cells.

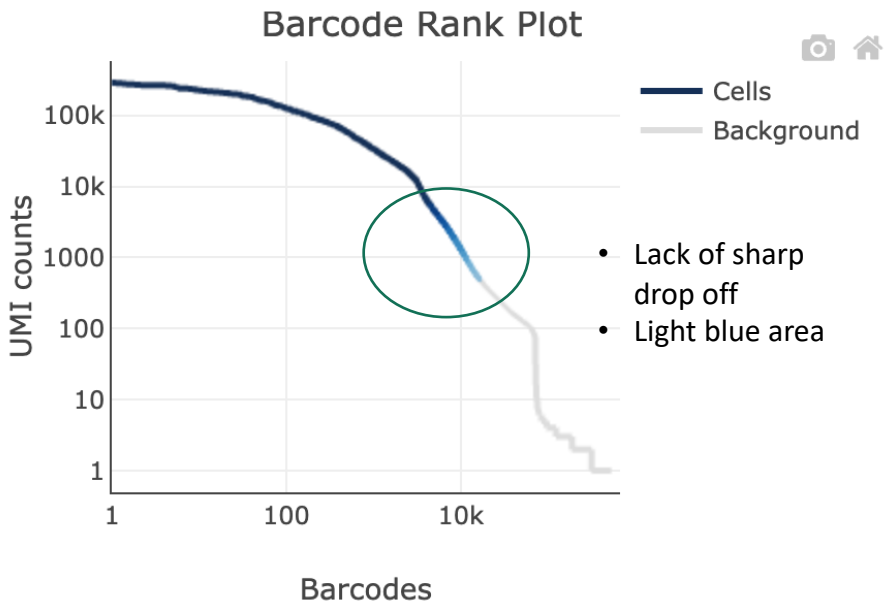






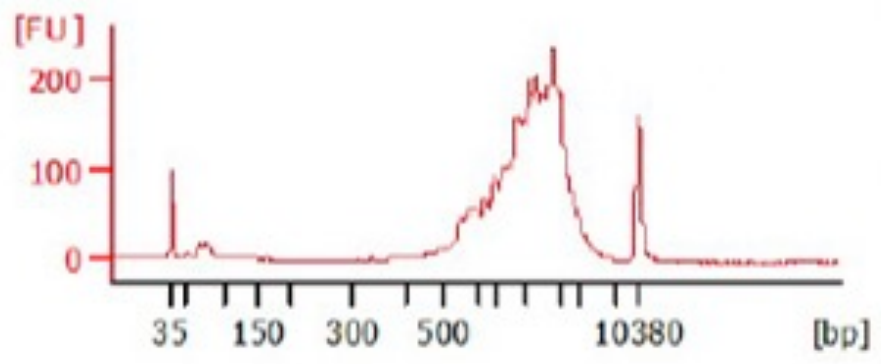
Cells

Curvy barcode rank plot



- Lack of sharp drop off
- Light blue area

Add more sequencing depth may increase the cell number as well and will end up lots of sequencing to get the recommended read count per cell



Example of an optimal cDNA