Preprocessing (from reads to a count matrix)

Roberta Menafra
19-10-2020

Bioinformatician LGTC (Leiden Genome Technology Center)

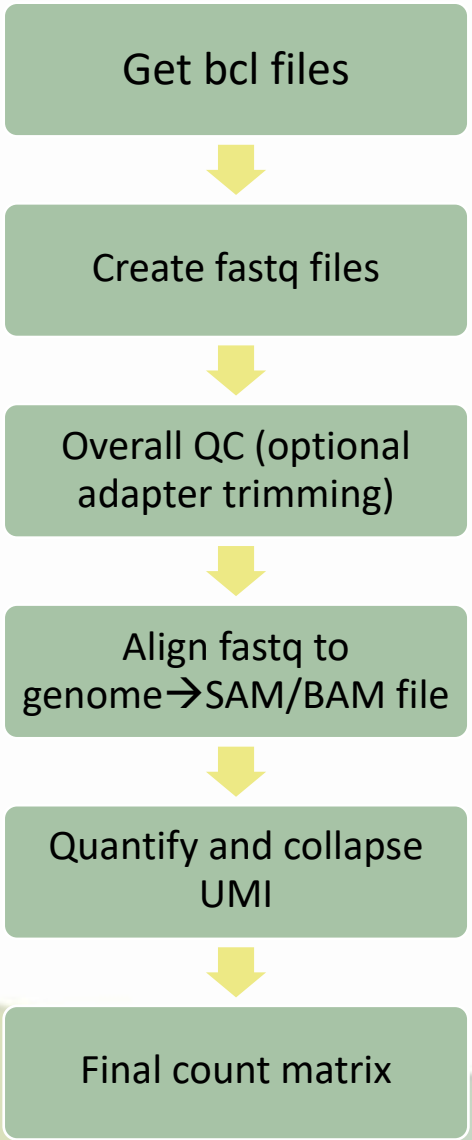
Preprocessing (from reads to a count matrix)

In DNA sequencing, a **read** is an inferred sequence of base pairs corresponding to all or part of a single DNA fragment.

In this lecture

- Sequencing data formats
- Data pre-processing
- 10X pipeline (Cell Ranger)
  - mkfastq
  - count
- Results examples

Preprocessing (from reads to a count matrix)



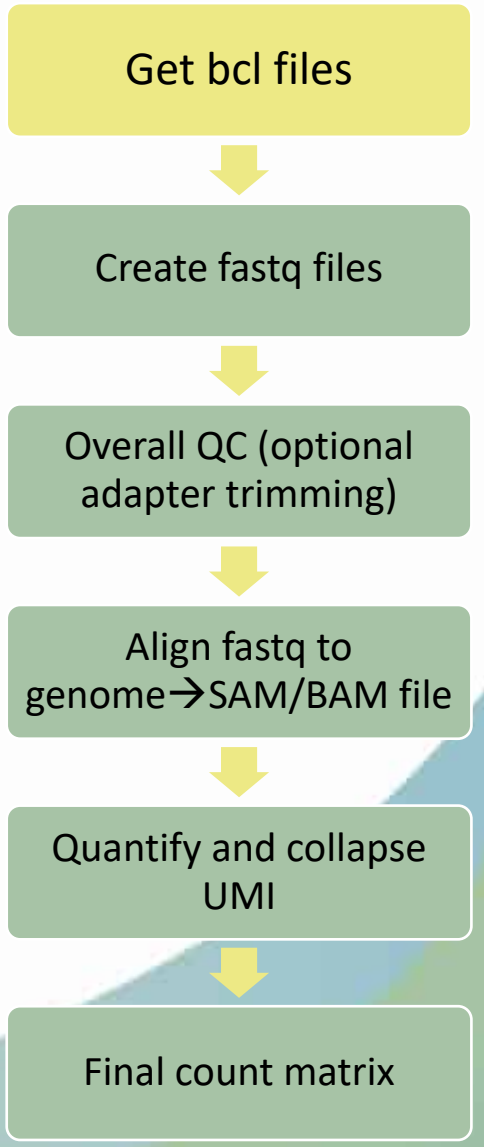
Common file formats in NGS

- **bcl**
- **fastq**
- **bam**
- mtx, tsv
- hdf5 (.h5, .h5ad)

BCL

Raw data files in binary base call format.

Illumina offers **bcl2fastq** Conversion Software to convert BCL files.
bcl2fastq is included, standalone conversion software that demultiplexes data and converts BCL files to standard FASTQ file formats for downstream analysis.



Samples demultiplexing

Sequence runs on NGS instruments are typically carried out with multiple samples pooled together. An **index tag** (also called a barcode) consisting of a unique sequence (between 6 and 12bp) is added to each sample so that the sequence reads from different samples can be identified.

The other requirement is a sample sheet

```
[Header],,,,,,,,,
IEMFileVersion,4,,,,,,,
Date,20-10-2014,,,,,,,
Workflow,GenerateFASTQ,,,,,,,
Application,FASTQ Only,,,,,,,
Assay,Nextera,,,,,,,
Description,,,,,,,,
Chemistry,Amplicon,,,,,,,
,,,,,,,,,
[Reads],,,,,,,,,
151,,,,,,,,,
151,,,,,,,,,
,,,,,,,,,
[Settings],,,,,,,,,
ReverseComplement,0,,,,,,,
Adapter,,,,,,,,,
,,,,,,,,,
[Data],,,,,,,
Lane,Sample_ID,Sample_Name,Sample_Plate,Sample_Well,I7_Index_ID,index,I5_Index_ID,index2
4,AV_1_HTO,AV_1_HTO,,,,ATTACTCG,,,
5,AV_1_HTO,AV_1_HTO,,,,ATTACTCG,,,
```

```
bcl2fastq --runfolder-dir 190826_E00603_0316_AH3GW3CCX2/ --output-dir HTO/ --sample-sheet samples_HTO.csv --barcode-mismatches 0
```

```
output-dir: HTO/Reports/html/flowcellID/all/all/all/laneBarcode.html
```

Flowcell Summary

Clusters (Raw)	Clusters(PF)	Yield (MBases)
1,242,422,208	946,733,762	285,914

Lane Summary

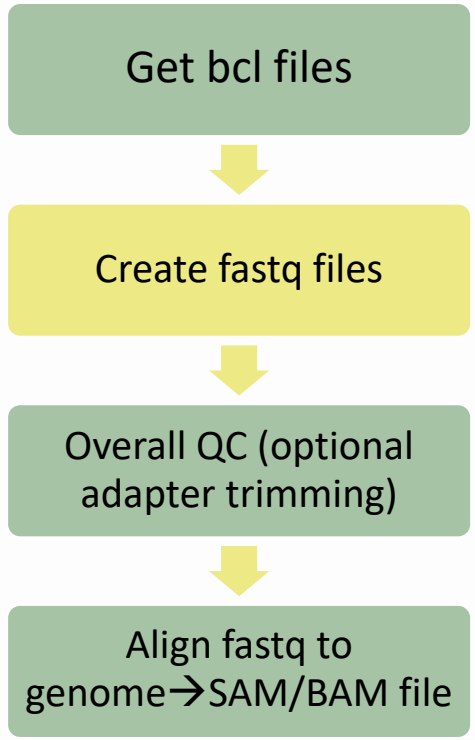
Lane	Project	Sample	Barcode sequence	PF Clusters	% of the lane	% Perfect barcode	% One mismatch barcode	Yield (Mbases)	% PF Clusters	% >= Q30 bases	Mean Quality Score
4	default	AV_1_HTO	ATTACTCG	54,321,382	11.52	100.00	NaN	16,405	100.00	46.78	26.39
4	default	Undetermined	*unknown*	417,222,614	88.48	100.00	NaN	126,001	73.60	67.87	31.88
5	default	AV_1_HTO	ATTACTCG	54,933,100	11.56	100.00	NaN	16,590	100.00	46.78	26.40
5	default	Undetermined	*unknown*	420,256,666	88.44	100.00	NaN	126,918	74.21	67.77	31.85

Top Unknown Barcodes

Lane	Count	Sequence	Lane	Count	Sequence
4	116,377,900	AGTGGAAC	5	116,758,060	AGTGGAAC
	107,277,740	GTCTCCTT		107,610,520	GTCTCCTT
	79,994,540	TCACATCA		80,406,280	TCACATCA
	74,171,580	CAGATGGG		74,495,820	CAGATGGG
	19,713,380	CAAAAGAT		19,888,220	CAAAAGAT
	705,960	GTCTCCTA		706,360	GTCTCCTA
	629,960	TCACTCAA		638,440	TCACTCAA
	600,500	CAGATGGA		606,040	AGTGAACA
	597,720	AGTGAACA		603,420	CAGATGGA
	564,060	TCCATCAA		574,720	TCCATCAA

FASTQ

- NGS data is often in FASTQ format
- FASTQ is a text-based format for storing both sequence and its corresponding quality score
- Four lines per sequence (read)
  - @ followed by the unique sequence identifier
  - The nucleotide sequence
  - + The quality line break
  - The quality scores in ASCII characters



View FASTQ Files

Viewing entire file

cat file1.fastq

Viewing first 10 lines

head file1.fastq

```
@A00379:133:HMWGLDSXX:1:1101:1163:1000 2:N:0:GAGGATCT
AAAAAAAAATAAAAAAAAATAAAAAAAATGATAAAAAAAATAAAATTAAAAAATATAAAAAAAAAAAAATTTTTTTTTTATTAAGTAAAAAAAATTAAAAATAAATTAAAAAAAAAAAAAAAAA
+
FF,:::FF,,F,:F,FF,F:FFFF:F:F,,F:F:FFFFF,,:F:,,,:FF:F,,,,F,FFF:FFFF,,,,:,F,,,,:,,,:F,,F::F,F,,,,F::F,FFF,,,,:,FFFFFF,FFFF::,
@A00379:133:HMWGLDSXX:1:1101:1253:1000 2:N:0:GAGGATCT
GATGGTGTAATGCCTGTCTTTAAGTACAAACTGCAAAGTGAAGCCACCCAGATTTATTTCCTTTCTCTGGACCAAGTGTCCCAAACTGAACACTGAGACTGAGGCAGATTTGAAATGATCCAGC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,,FFFFFFFFFFFFFFFF:FFFFF,FF:FFFFFFFFFF,FF,FFFFF
```

FASTQ

Header

```
@A00379:133:HMWGLDSXX:1:1101:1163:1000 2:N:0:GAGGATCT
+
FF,:::FF,,F,:F,FF,F:FFFF:F:F,,F:F:FFFFF,,:F:,,,:FF:F,,,,F,FFF:FFFF,,,,:,F,,,:,,,:F,,F::F,F,,,F::F,FFF,,,,:,FFFFFF,FFFF::,:
```

Instrument — RunID — FlowcellID — Read Number (Paired 2/2) — Index Sequence

Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

The quality score is associated to a probability of error, of an incorrect base call

$Q = -10\log_{10}(e)$ where **e** is the **estimated probability** of the base call being wrong.

- **Higher Q scores** indicate a smaller probability of error.
- **Lower Q scores** can result in a significant portion of the reads being unusable.

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

FASTQC

Before analyzing the data to draw biological conclusions you should always perform some simple quality control

FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material

FastQC Report **Read1**

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
- Kmer Content

Basic Statistics

Measure	Value
Filename	Pool_1_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	380844038
Sequences flagged as poor quality	0
Sequence length	26
%GC	51

FastQC Report **Read2**

Summary

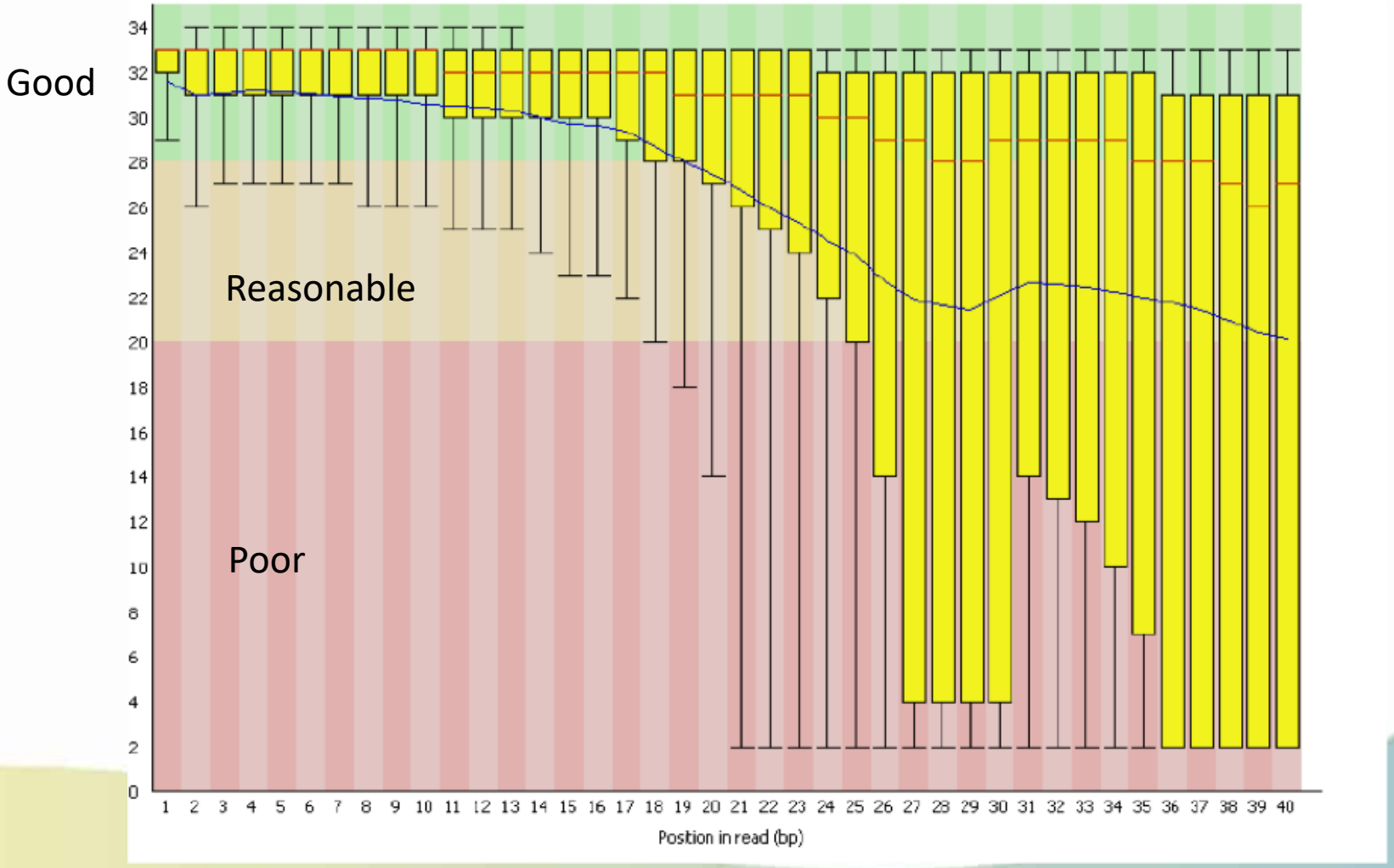
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	Pool_1_S1_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	380844038
Sequences flagged as poor quality	0
Sequence length	150
%GC	36

Per Base Sequence Quality

This view shows an overview of the range of quality values across all bases at each position in the FastQ file.



Overrepresented Sequences

A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole.
This module lists all of the sequence which make up more than 0.1% of the total. Hits must be at least 20bp in length and have no more than 1 mismatch.
Finding a hit doesn't necessarily mean that this is the source of the contamination

Warning
This module will issue a warning if any sequence is found to represent more than 0.1% of the total.
Failure
This module will issue an error if any sequence is found to represent more than 1% of the total.

Typical artifacts
Primers, sequence adapters

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTT	103227946	27.105044506433888	No Hit
GCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTT	8592717	2.256229884843307	No Hit
AGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTT	4198600	1.1024460359282295	No Hit
GTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTT	1116215	0.2930897923102055	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	1018762	0.26750110238039226	No Hit
CAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTT	1005203	0.26394085234439196	No Hit
AGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTT	931329	0.24454341070714097	No Hit
GGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	563221	0.14788757176238113	No Hit
AAGCAGTGGTATCAACGCAGAGTATTTTTTTTTTTTTTTTTTTTTTTT	392982	0.1031871214431352	No Hit


Current version: v1.9
Home
Docs
Plugins
Logo
Example Reports
MultiQC
Citations
858
Aggregate results from bioinformatics analyses across many samples into a single report
MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools.
Introduction to MultiQC
Phil Ewels
phil.ewels@scilifelab.se
Introduction to MultiQC (1:19)
Installing MultiQC (4:33)
Running MultiQC (5:21)
Using MultiQC Reports (6:06)
GitHub
Python Package Index
Documentation
92 supported tools
Publication / Citation
Get help on Gitter
Quick Install
pip install multiqc # Install
multiqc . # Run
pip
conda
manual
Need a little more help? See the full installation instructions.

MultiQC
v1.0.dev0

General Statistics

Copy table Showing $^{29}/_{29}$ rows and $^{3}/_{5}$ columns.

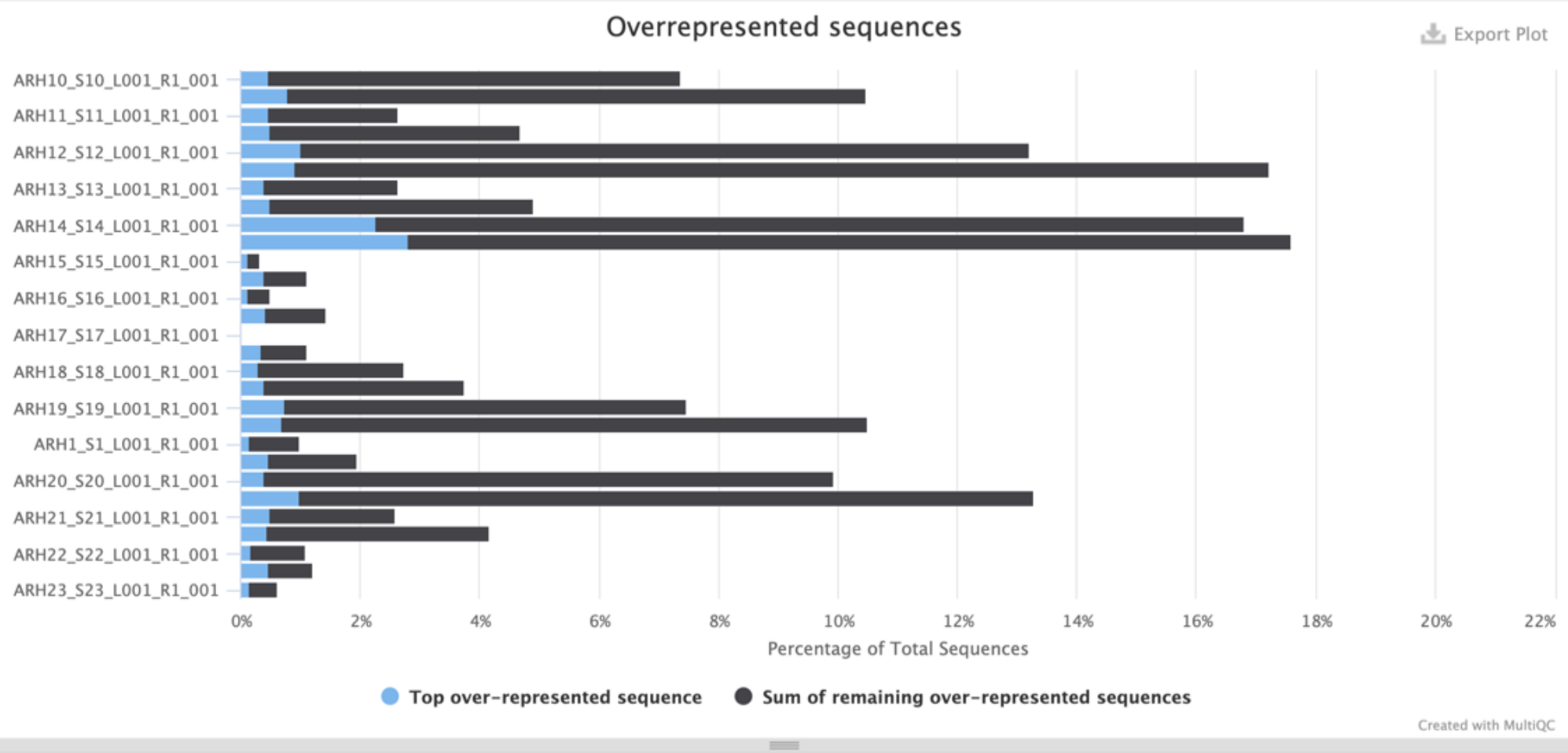
Sample Name	% Dups	% GC	M Seqs
ARH10_S10_L001_R1_001	69.0%	54%	53.5
ARH10_S10_L001_R2_001	73.0%	53%	53.5
ARH11_S11_L001_R1_001	72.2%	50%	4.8
ARH11_S11_L001_R2_001	73.0%	50%	4.8
ARH12_S12_L001_R1_001	65.9%	57%	26.9
ARH12_S12_L001_R2_001	69.7%	56%	26.9
ARH13_S13_L001_R1_001	50.7%	51%	27.1
ARH13_S13_L001_R2_001	56.6%	51%	27.1
ARH14_S14_L001_R1_001	65.5%	51%	33.2
ARH14_S14_L001_R2_001	68.8%	51%	33.2
ARH15_S15_L001_R1_001	60.4%	47%	22.7
ARH15_S15_L001_R2_001	64.2%	47%	22.7
ARH16_S16_L001_R1_001	65.5%	47%	26.0
ARH16_S16_L001_R2_001	68.9%	47%	26.0
ARH17_S17_L001_R1_001	58.8%	47%	10.0
ARH17_S17_L001_R2_001	61.3%	46%	10.0
ARH18_S18_L001_R1_001	48.2%	50%	28.1

MultiQC
v1.0.dev0

Overrepresented sequences

25 3

The total amount of overrepresented sequences found in each library. See the FastQC help for further information.



MultiQC
v1.0.dev0

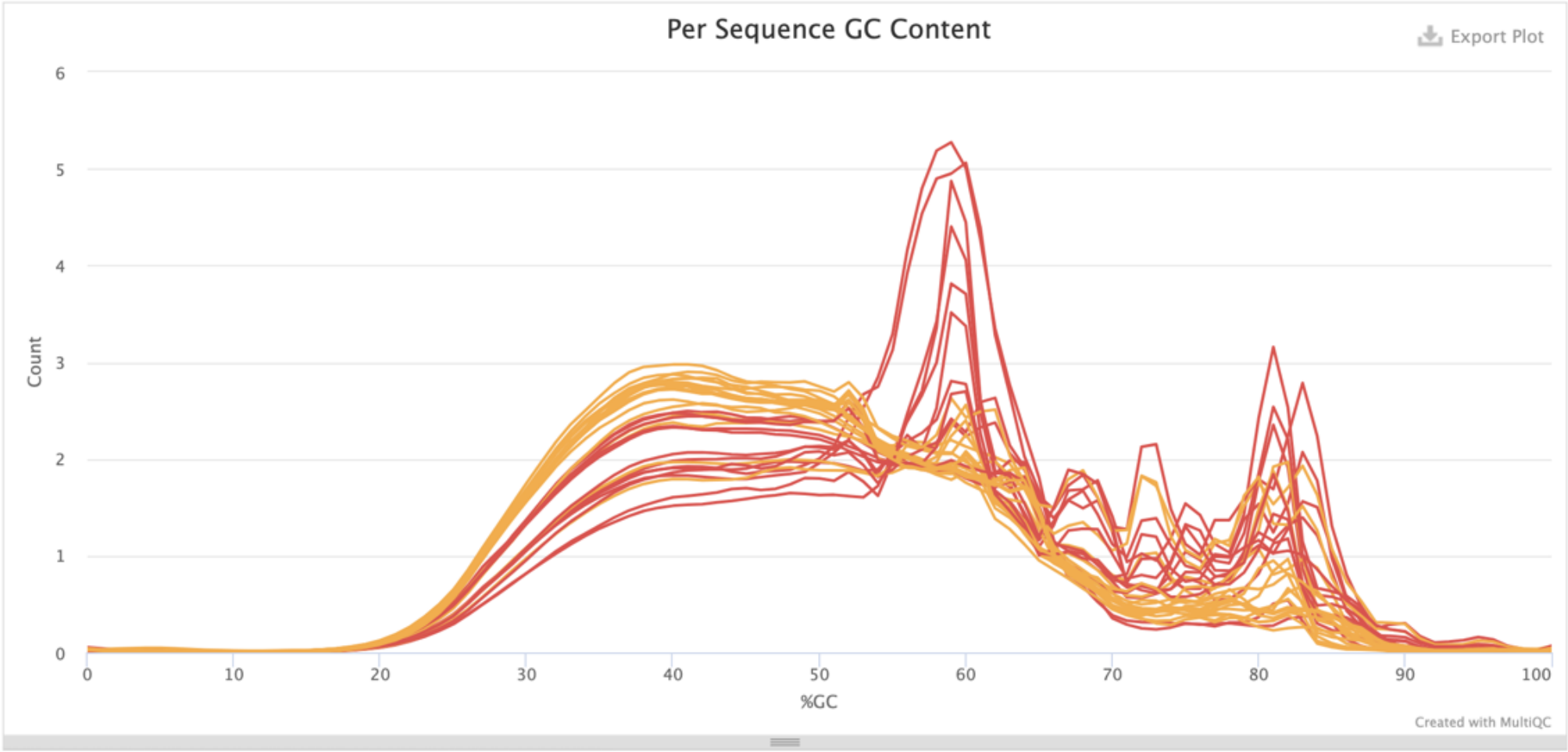
Per Sequence GC Content

16 13

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content. See the FastQC help.

Y-Limits: on

Percentages | Counts



Sequence filtering

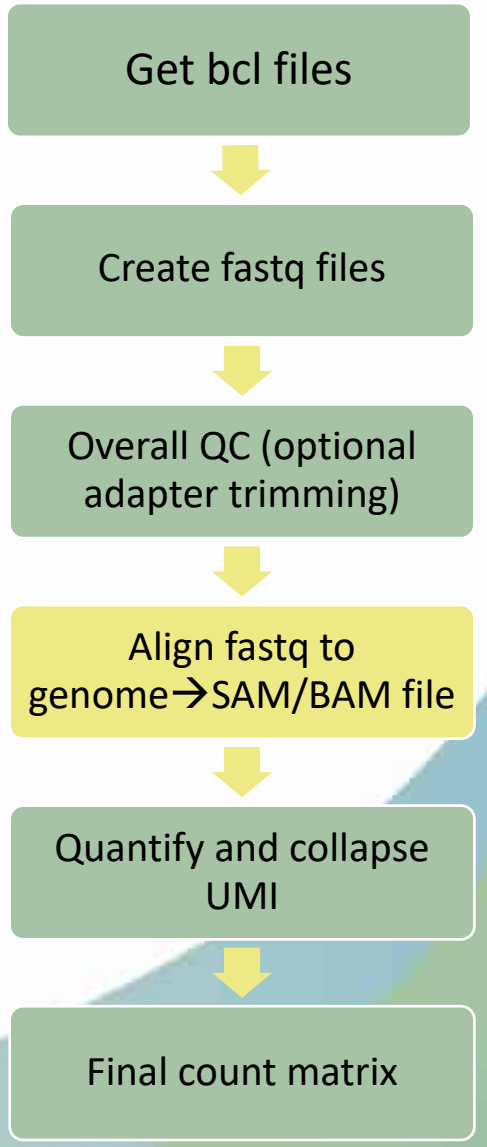
Adapter trimming
Quality trimming
Trimming fixed length

Tools

Cutadapt
SeqTK
Trimmomatic

Aligning reads to a reference

FASTQ files contain sequence information that we wish to map to genes in a genome:
1.Select your genome
2.Select your gene annotation file (generally gtf format)
3.Run the alignment program
4.Result of alignment is generally stored in a sam/bam file



Aligners

STAR, Kallisto, Bowtie

SAM Format

This is the most basic, human readable format, generated by almost every alignment algorithm that exists. It consists of a header, a row for every read in your dataset, and 11 tab-delimited fields describing that read.

SAM Header

The full list of available header fields can be found below

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

http://samtools.github.io/hts-specs/SAMv1.pdf

Bitwise Flag

The bitwise flag is a lookup code to explain certain features about the particular read.
It tells you whether the read aligned, is marked a PCR duplicate, if it's mate aligned, etc. and any combination of the available tags, seen below:

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

One important thing to note is that any combination of these flags results in one integer, which makes interpreting it a bit difficult. To make it easy you can check to either encode or decode a bitwise flag.
https://broadinstitute.github.io/picard/explain-flags.html

MapQ (Mapping Quality)

This value reports how well the read aligned to the reference. Different algorithms report it differently but nonetheless, the greater the number the better the alignment (generally).

CIGAR String

This is a shorthand way to encode an entire alignment:

Op	Description
M	Match (alignment column containing two letters). This could contain two different letters (mismatch) or two identical letters. USEARCH generates CIGAR strings containing Ms rather than X's and ='s (see below).
D	Deletion (gap in the target sequence).
I	Insertion (gap in the query sequence).
S	Segment of the query sequence that does not appear in the alignment. This is used with soft clipping, where the full-length query sequence is given (field 10 in the SAM record). In this case, S operations specify segments at the start and/or end of the query that do not appear in a local alignment.
H	Segment of the query sequence that does not appear in the alignment. This is used with hard clipping, where only the aligned segment of the query sequences is given (field 10 in the SAM record). In this case, H operations specify segments at the start and/or end of the query that do not appear in the SAM record.
=	Alignment column containing two identical letters. USEARCH can read CIGAR strings using this operation, but does not generate them.
X	Alignment column containing a mismatch, i.e. two different letters. USEARCH can read CIGAR strings using this operation, but does not generate them.

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

```
POS: 5
CIGAR: 3M1I3M1D5M
```

samtools view aligned_reads.sam | head –n 1

HS2000-940_146:5:1101:1161:63226 73 NC_000020.11 23775298 60 78M22S = 23775298 0 CTGNTAGCCCTGCTGAATCTCCCTCCTGACCCAACTCCCTCNTNNNNNNNGCTGGGTGACTGCTGNCNNCACNGGCTGTGNNNNNNNNNNNNNNCAGCTG G ?@@#4ADDDFDFFHIGGFCFHCHFGIHGCGHEHHEHD3?BH#0#######--5CEECG=?AEEHE################################## NM:i:13 MD:Z:3G37C1C0T0A0C0T0C0T15T1C0T3T5 AS:i:52 XS:i:0 RG:Z:sample_1 HS2000-940_146:5:1101:1161:63226 133 NC_000020.11 23775298 0 * = 23775298 0 NNCTCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGNNNNCNAA AGGAGCCTGGGT ## AS:i:0 XS:i:0 RG:Z:sample_1 HS2000-940_146:5:1101:1262:12434 99 NC_000020.11 23843774 60 100M = 23843977 258

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

BAM format

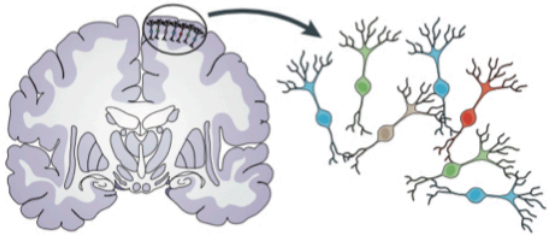
This is the same format except that it encoded in binary which means that it is significantly smaller than the SAM files and significantly faster to read, though it is not human legible and needs to be converted to another format (i.e. SAM) in order to make sense to us.
Some special tools are needed in order to make sense of BAM, such as Samtools, Picard Tools,

View BAM Files

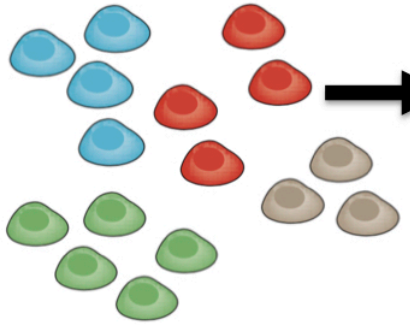
samtools view alignment.bam | head

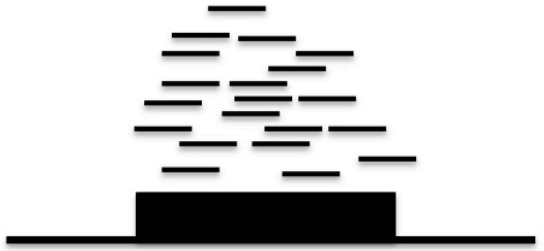
Single-cell RNA-Seq (scRNA-Seq)

Tissue (e.g. tumor)



Isolate and sequence individual cells



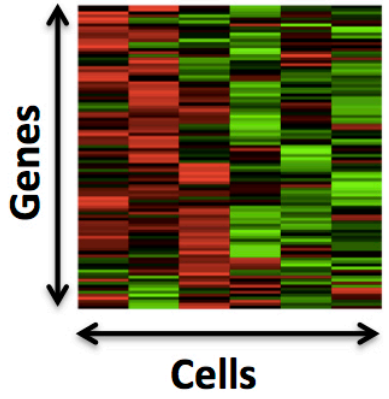


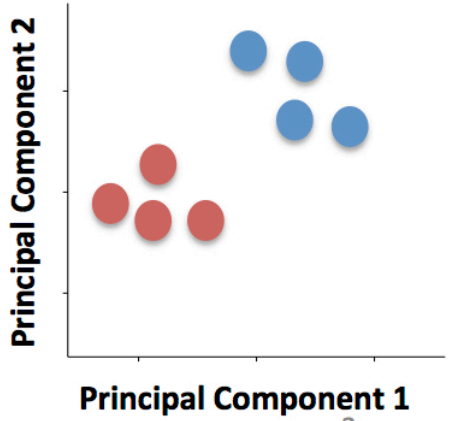
Gene 1
Cell 1

Read Counts

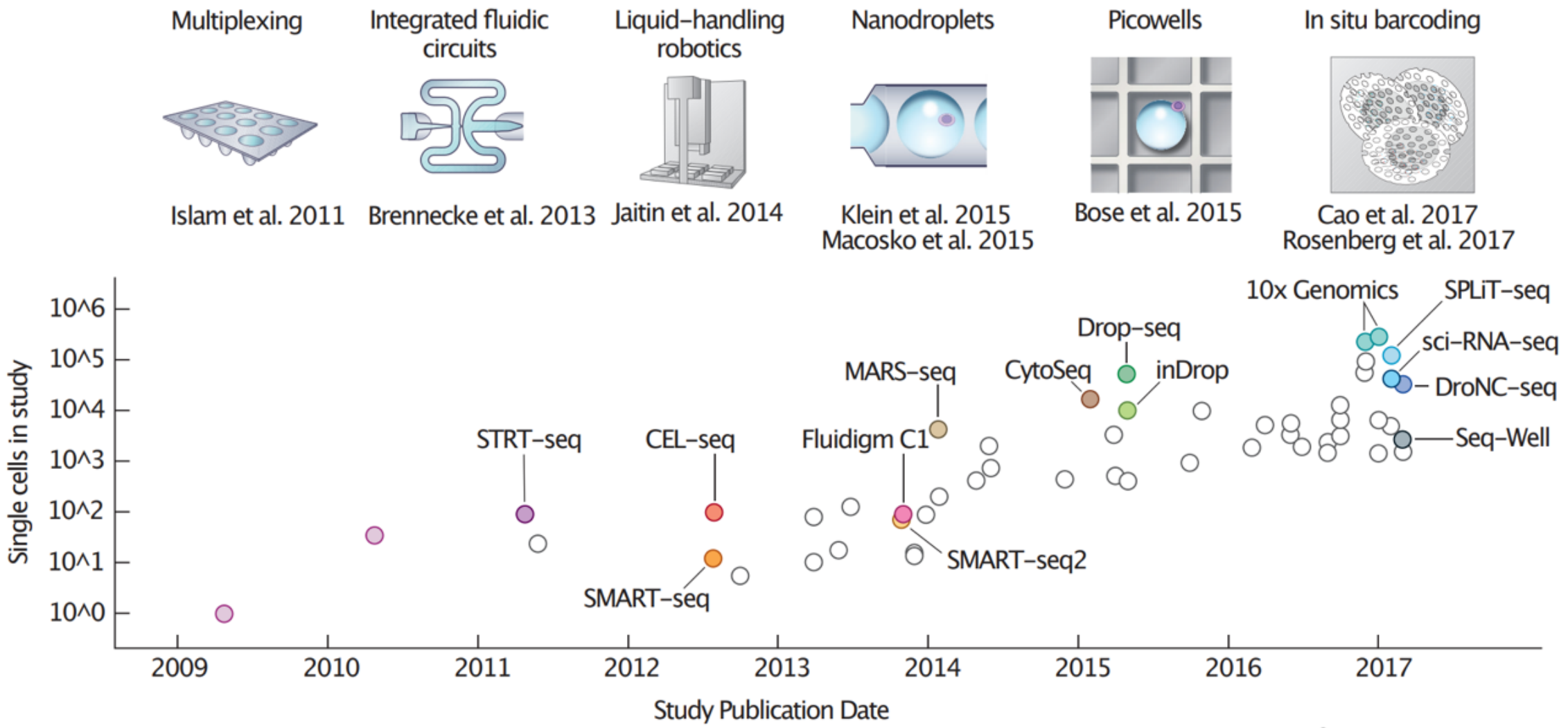
	Cell 1	Cell 2	…
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
…			

Compare gene expression profiles of single cells





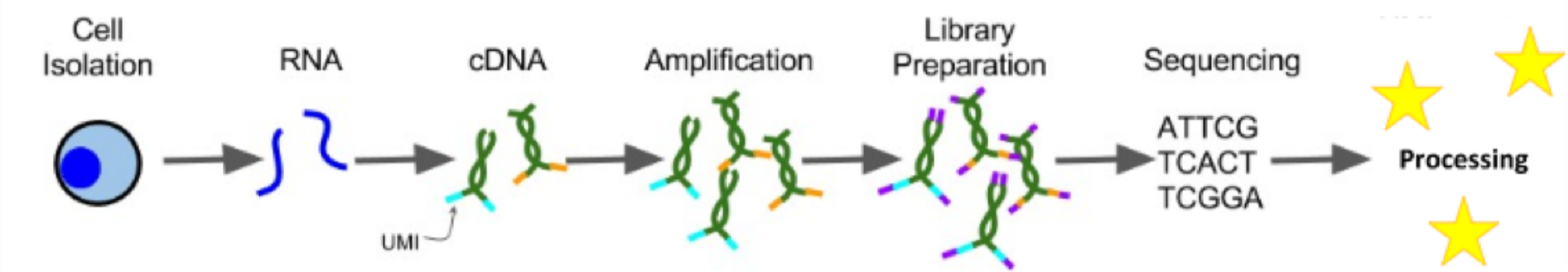
scRNA-seq output has increased significantly



Svensson et al 2018

Challenges in single cell data analysis

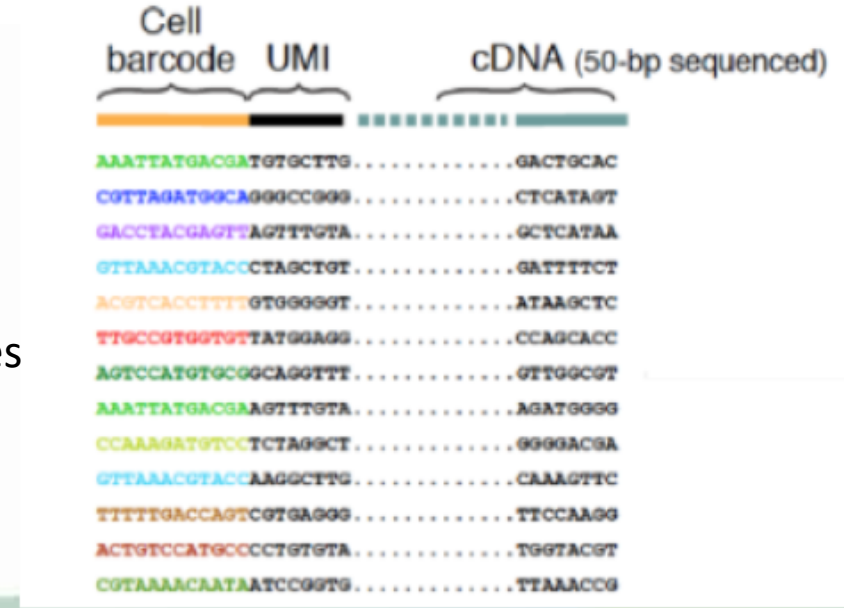
Amplification artifacts
Dropouts



With high number of PCR cycles, some transcript become over represented in the final library

UMI are unique molecule identifier added to the transcript during RT step

UMI enable sequencing reads to be assigned to individual transcript molecules and removal of amplification noise and biases



How to run the 10X cellranger software?

Pipeline download and installation

https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/tutorial_in

10X technology and Cell Ranger

Cell Ranger is a set of analysis pipelines that process Chromium single-cell RNA-seq output to align reads, generate gene-barcode matrices and perform clustering and gene expression analysis. Cell Ranger includes four pipelines relevant to single-cell gene expression experiments:

•**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 useful features that are specific to 10x libraries and a simplified sample sheet format.

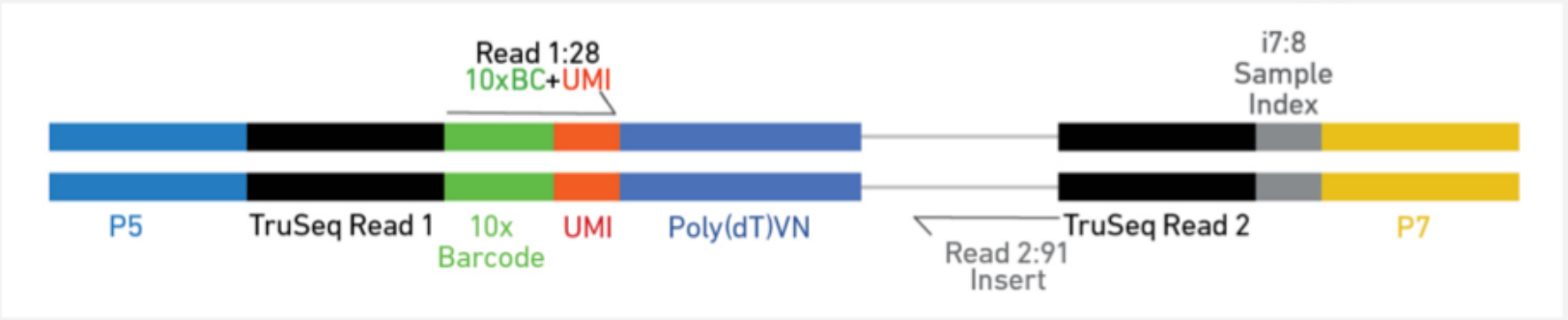
•**cellranger count** takes FASTQ files from cellranger mkfastq and performs alignment, filtering, barcode counting, and UMI counting. It uses the Chromium cellular barcodes to generate feature-barcode matrices, determine clusters, and perform gene expression analysis. The count pipeline can take input from multiple sequencing runs.

- **cellranger aggr** .
- **cellranger reanalyze**
- **cellranger mat2csv**
- **cellranger mkgtf**
- **cellranger mkref**
- **cellranger vdj**
- **cellranger mkvdjref**
- **cellranger testrun**
- **cellranger upload**
- **cellranger sitecheck**

Cell Ranger mkfastq

```
cellranger mkfastq --id=my_data --run=/path/to/bcl  --csv=samplesheet.csv
```

Cellranger mkfastq produces :
Index.fastq; R1.fastq; R2.fastq



index

```
AFCHKCCJALXX/HKCCJALXX/906_05_2776 $ zcat -f 906_05_S1_L008_I1_001.fastq.gz | head
@E00342:426:HKCCJALXX:8:1101:1661:1450 1:N:0:NAACCTCA
NAACCTCA
+
#AAAFA-F
```

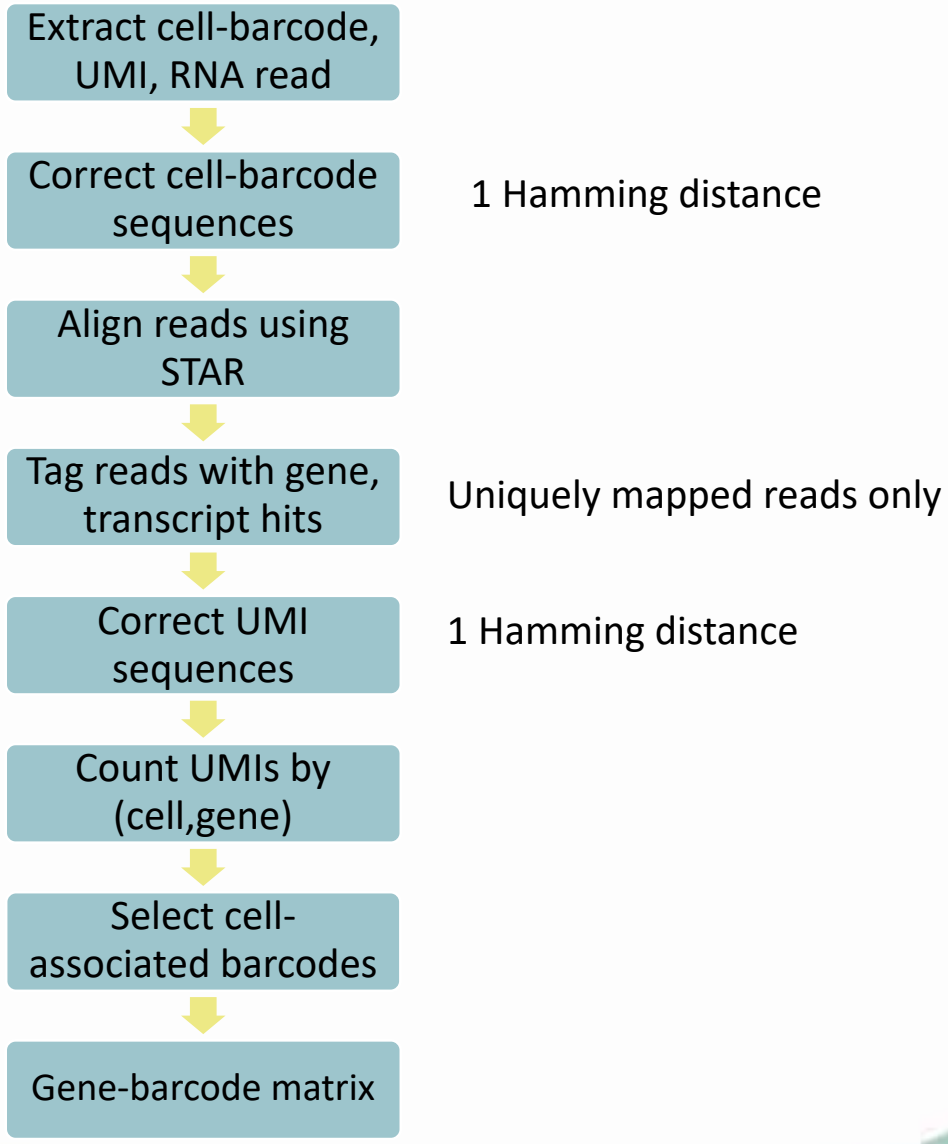
Read1

```
AFCHKCCJALXX/HKCCJALXX/906_05_2777 $ zcat -f 906_05_S1_L008_R1_001.fastq.gz | head
@E00342:426:HKCCJALXX:8:1101:1661:1450 1:N:0:NAACCTCA
GNTCAGTTCAACGGGAGCAACCCAGT
+
A#AFFFFFFF<FJJJ<AJFJJFF7<J
```

Read2

```
AFCHKCCJALXX/HKCCJALXX/906_05_2779 $ zcat -f 906_05_S1_L008_R2_001.fastq.gz | head
@E00342:426:HKCCJALXX:8:1101:1661:1450 2:N:0:NAACCTCA
NCGCGGGGCTGAATTGTGCTCTGATCTTTGGGTGAAAGATAAGTCACAATTTAAAAGCGAATTTCTTCAGGAAAAGACAGCAAAAATACACAAACTAA
+
#--A7-77FA<A7FFAA7-----<A-AF--7--77AF-AA-<---77-<<7AAFF-7A--A-7----A-7--7AA--7---<7-A7--7-------77
```

10X Cell Ranger count pipeline



Gene Expression Algorithms Overview

Genome Alignment

Cell Ranger uses an aligner called STAR, which peforms splicing-aware alignment of reads to the genome. Cell Ranger then uses the transcript annotation GTF to categorize the reads into **exonic**, **intronic**, and **intergenic**, and by whether the reads align (confidently) to the genome. A read is exonic if at least 50% of it intersects an exon, intronic if it is non-exonic and intersects an intron, and intergenic otherwise.

MAPQ adjustment

For reads that align to a single exonic locus but also align to 1 or more non-exonic loci, the exonic locus is prioritized and the read is considered to be confidently mapped to the exonic locus with MAPQ 255.

Transcriptome Alignment

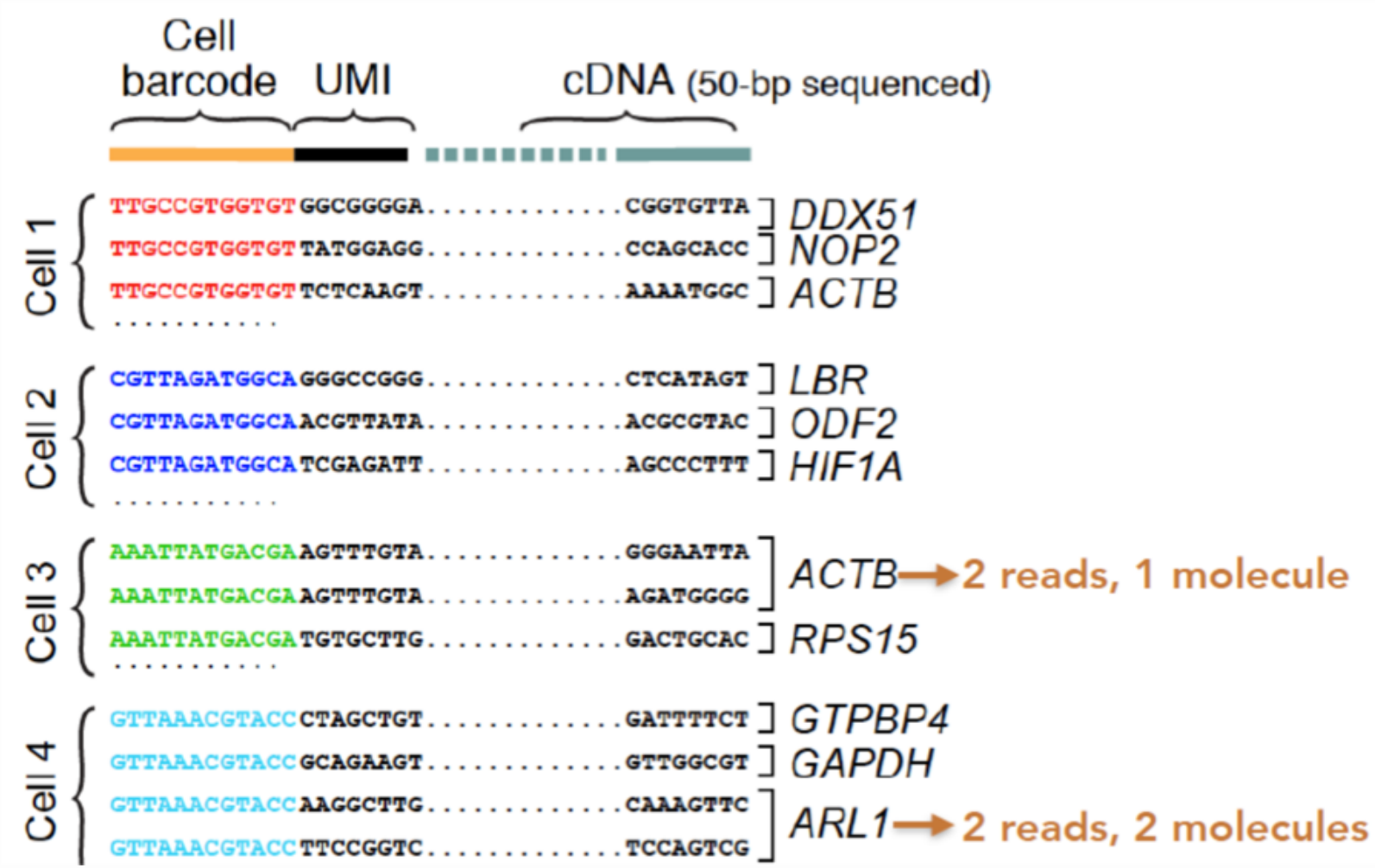
Cell Ranger further aligns exonic reads to annotated transcripts, looking for compatibility. A read that is compatible with the exons of an annotated transcript, and aligned to the same strand, is considered mapped to the transcriptome. If the read is compatible with a single gene annotation, it is considered uniquely (confidently) mapped to the transcriptome. **These confidently mapped reads are the only ones considered for UMI counting**.

UMI Counting
Before counting UMIs, Cell Ranger attempts to correct for sequencing errors in the UMI sequences.

First grouping
Reads that were confidently mapped to the transcriptome are placed into groups that share the same barcode, UMI, and gene annotation.
If two groups of reads have the **same barcode and gene, but their UMIs differ by a single base** then one of the UMIs was likely introduced by a substitution error in sequencing. The UMI of the less-supported read group is corrected to the UMI with higher support.

Second grouping
Cell Ranger again groups the reads by barcode, UMI (possibly corrected), and gene annotation. If two or more groups of reads have the same barcode and UMI, but different gene annotations, the gene annotation with the most supporting reads is kept for UMI counting.

After these two filtering steps, each observed barcode, UMI, gene combination is recorded as a UMI count in the unfiltered feature-barcode matrix. The number of reads supporting each counted UMI is also recorded in the molecule info file.


Cell barcode
UMI
cDNA (50-bp sequenced)
Cell 1
TTGCCGTGGTGTGGCGGGGA.....CGGTGTTA DDX51
TTGCCGTGGTGTTATGGAGG.....CCAGCACC NOP2
TTGCCGTGGTGTTCTCAAGT.....AAAATGGC ACTB
Cell 2
CGTTAGATGGCAGGGCCGGG.....CTCATAGT LBR
CGTTAGATGGCAACGTTATA.....ACGCGTAC ODF2
CGTTAGATGGCATCGAGATT.....AGCCCTTT HIF1A
Cell 3
AAATTATGACGAAGTTTGTA.....GGGAATTA
AAATTATGACGAAGTTTGTA.....AGATGGGG ACTB → 2 reads, 1 molecule
AAATTATGACGATGTGCTTG.....GACTGCAC RPS15
Cell 4
GTTAAACGTACCCTAGCTGT.....GATTTTCT GTPBP4
GTTAAACGTACCGCAGAAGT.....GTTGGCGT GAPDH
GTTAAACGTACCAAGGCTTG.....CAAAGTTC
GTTAAACGTACCTTCCGGTC.....TCCAGTCG ARL1 → 2 reads, 2 molecules

Filtering cells (Cell Ranger)

Cellranger 3.0 introduces and improved cell-calling algorithm to identify populations of low RNA content cells, especially when low RNA content cells are mixed into a population of high RNA content cells.
E.g tumor samples often contain large tumor cells mixed with smaller tumor infiltrating lymphocytes (TIL).
The new algorithm is based on the **EmptyDrops** method (Lun et al., 2018).

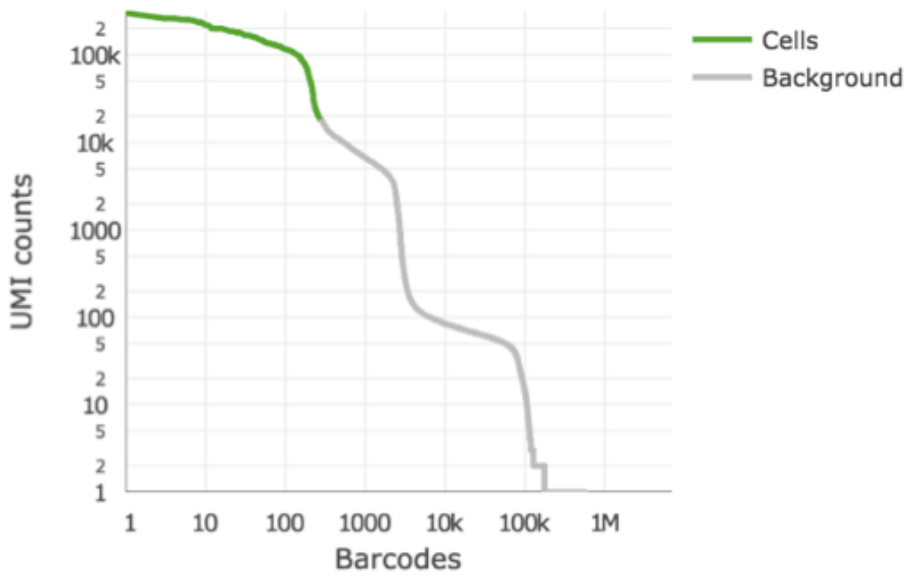
The algorithm has two key steps:
1.It uses a cutoff based on total UMI counts of each barcode to identify cells. This step identifies the primary mode of high RNA content cells.

2.Then the algorithm uses the RNA profile of each remaining barcode to determine if it is an “empty” or a cell containing partition. This second step captures low RNA content cells whose total UMI counts may be similar to empty GEMs.

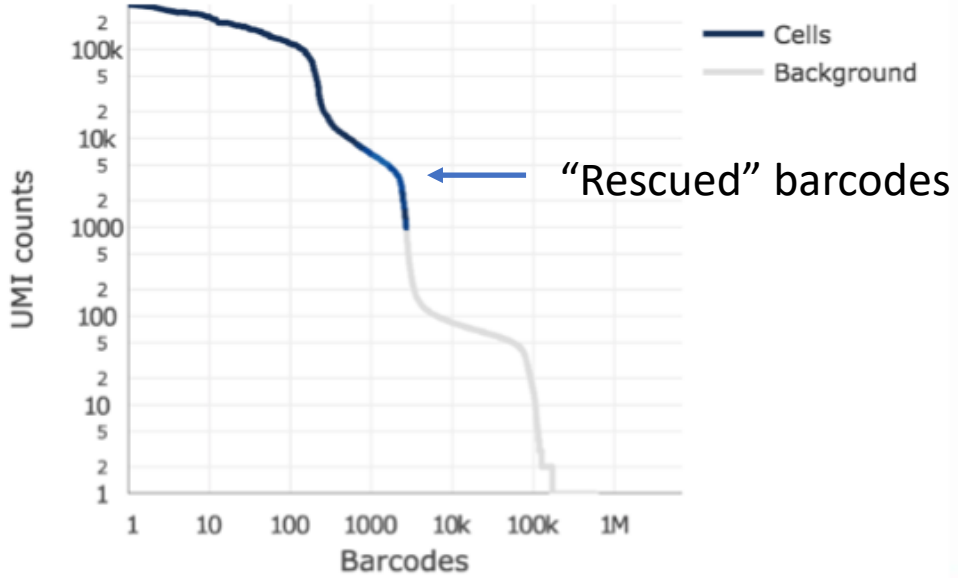
Barcodes selection steps

1. The original cellranger cell calling algorithm is used to identify high RNA content cells, using a cutoff based on the total UMI count for each barcode.
2. In the second step, a set of barcodes with low UMI counts that likely represent ‘empty’ GEM partitions is selected. A model of the RNA profile of selected barcodes is created. This second step identifies cells that are clearly distinguishable from the profile of empty GEMs, even though they may have much lower RNA content than the largest cells in the experiment.

Earlier version



Latest version



Alternative pipelines

10X data or other scRNAseq technologies can also rely on independent pipelines

STARsolo

Alevin

dropEst

How to run cell ranger count
Command-Line Argument Reference

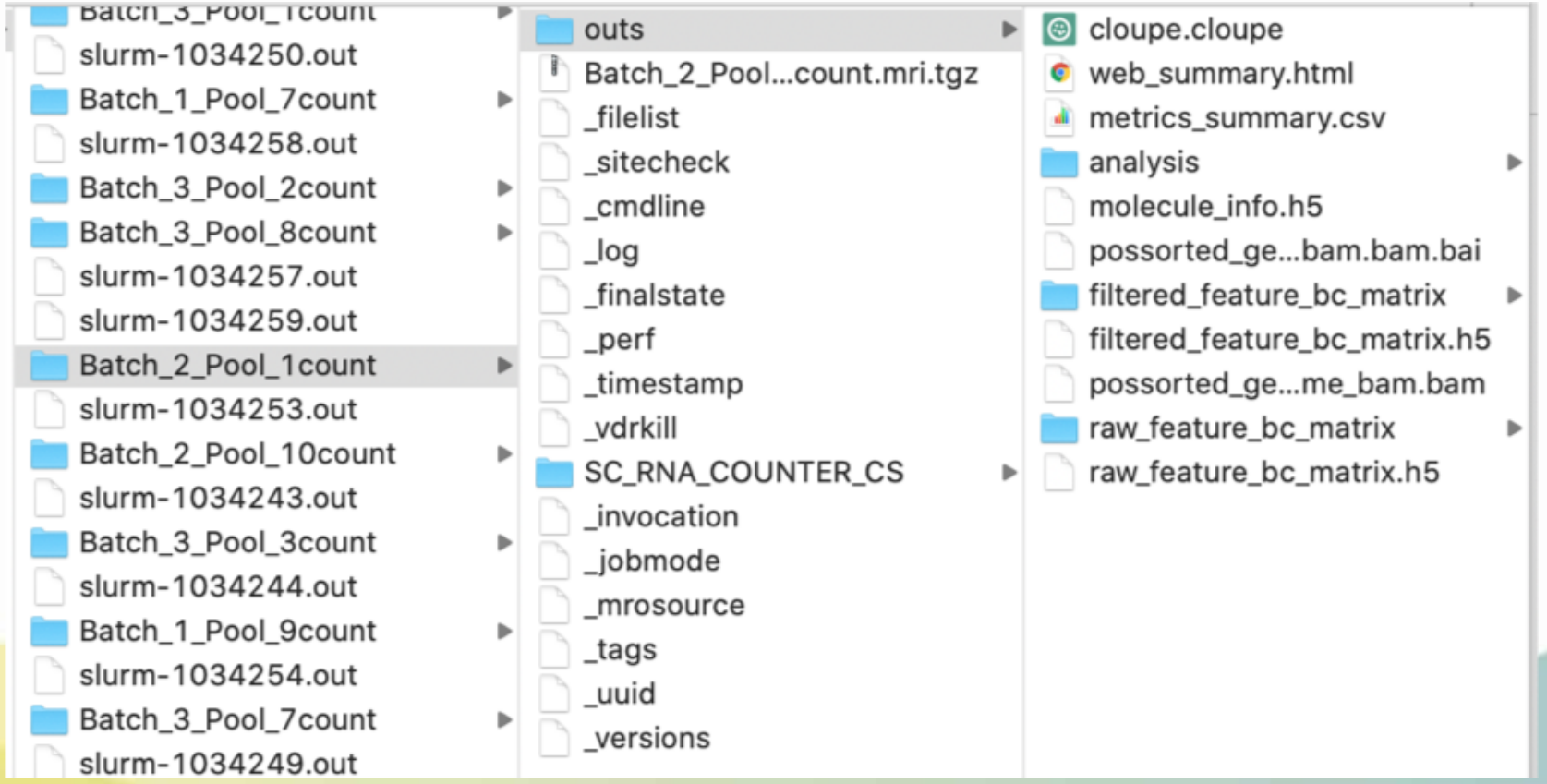
Argument	Description
--id	A unique run ID string: e.g. sample345
--fastqs	Path of the fastq_path folder generated by cellranger mkfastq
--libraries	Path to a libraries.csv file declaring FASTQ paths and library types of input libraries. Required for feature-barcoding analysis. See Feature Barcoding Analysis page for details. When using this argument, --fastqs and --sample must not be passed.
--sample	Sample name as specified in the sample sheet supplied to cellranger mkfastq.
--transcriptome	Path to the Cell Ranger compatible transcriptome reference
--feature-ref	Path to a Feature Reference CSV file declaring the Feature Barcoding reagents in use in the experiment.
--expect-cells	(optional) Expected number of recovered cells.
--force-cells	(optional) Force pipeline to use this number of cells, bypassing the cell detection algorithm.
--chemistry	(optional)
--r1-length	(optional)
--r2-length	(optional) Hard-trim the input R2 sequence to this length.

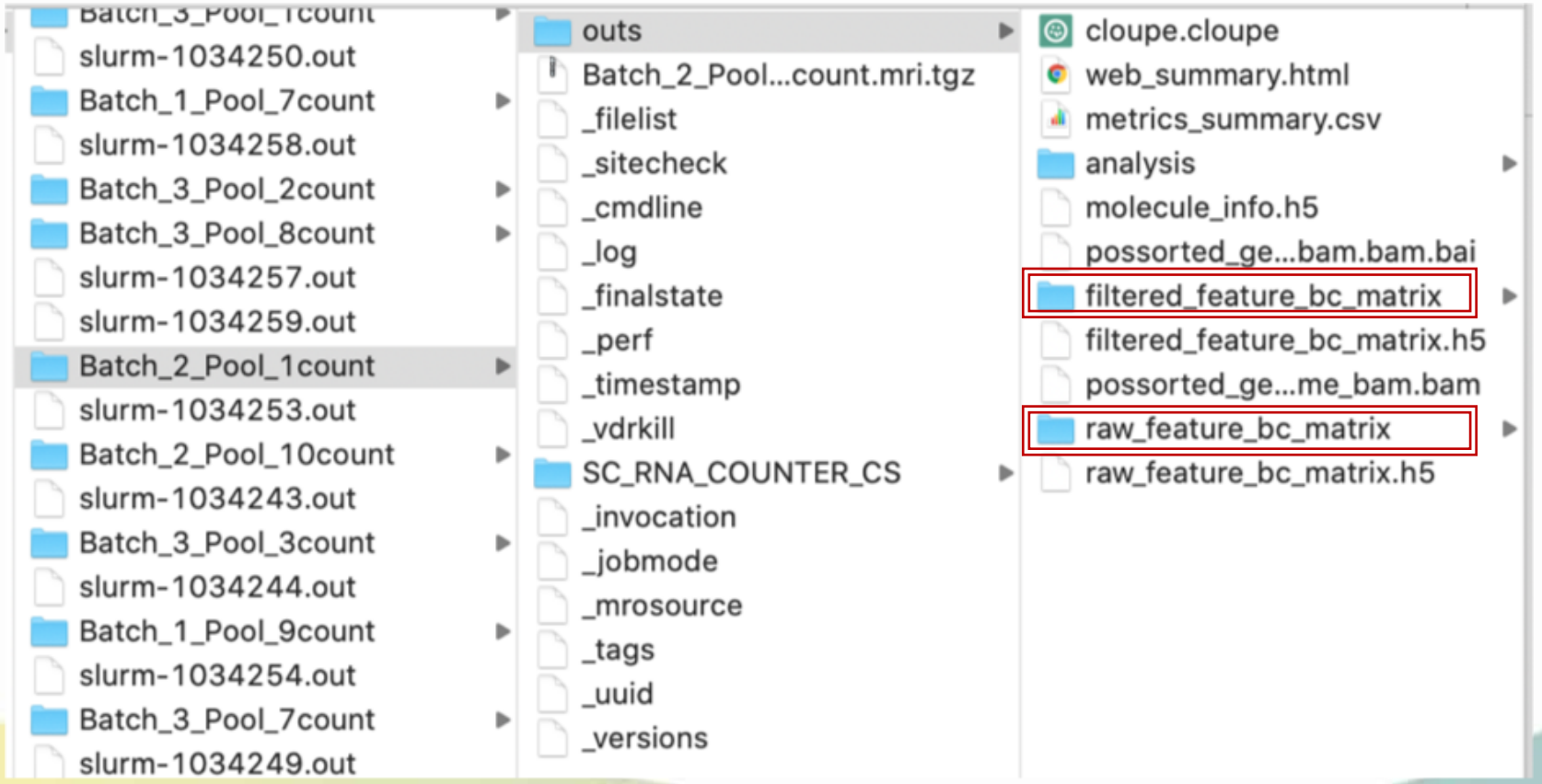
https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/count

How to run cellranger count

```
path_to_cellranger/cellranger count --id=samplename(your choice) --transcriptome=/opt/refdata-cellranger-GRCh38-3.0.0 --fastqs=/home/jdoe/runs/HAWT7ADXX/outs/fastq_path --sample=sample_prefix
```

output



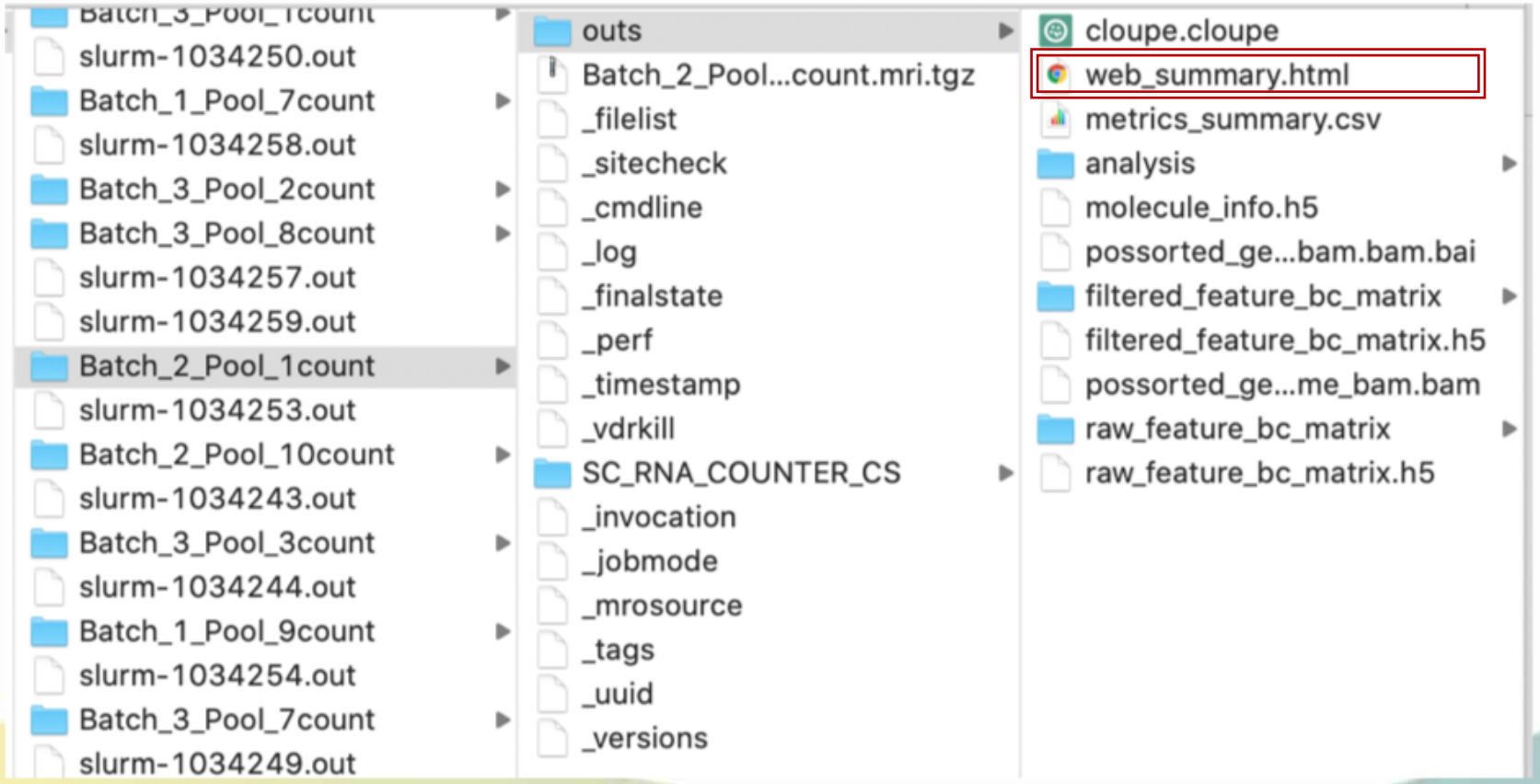

slurm-1034250.out
Batch_1_Pool_7count
slurm-1034258.out
Batch_3_Pool_2count
Batch_3_Pool_8count
slurm-1034257.out
slurm-1034259.out
Batch_2_Pool_1count
slurm-1034253.out
Batch_2_Pool_10count
slurm-1034243.out
Batch_3_Pool_3count
slurm-1034244.out
Batch_1_Pool_9count
slurm-1034254.out
Batch_3_Pool_7count
slurm-1034249.out
outs
Batch_2_Pool...count.mri.tgz
_filelist
_sitecheck
_cmdline
_log
_finalstate
_perf
_timestamp
_vdrkill
SC_RNA_COUNTER_CS
_invocation
_jobmode
_mrosource
_tags
_uuid
_versions
cloupe.cloupe
web_summary.html
metrics_summary.csv
analysis
molecule_info.h5
possorted_ge...bam.bam.bai
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
possorted_ge...me_bam.bam
raw_feature_bc_matrix
raw_feature_bc_matrix.h5

Matrix output

With 3 files needed to completely describe each gene x cell matrix
•matrix.mtx.gz
•features.tsv.gz
•barcode.tsv.gz

Type	Description
Raw	gene-barcode matrices Contains every barcode from fixed list of known-good barcode sequences. This includes background and non-cellular barcodes.
Filtered	gene-barcode matrices Contains only detected cellular barcodes.

Web summary examples



Sample_1

Summary | Analysis

9,175
Estimated Number of Cells

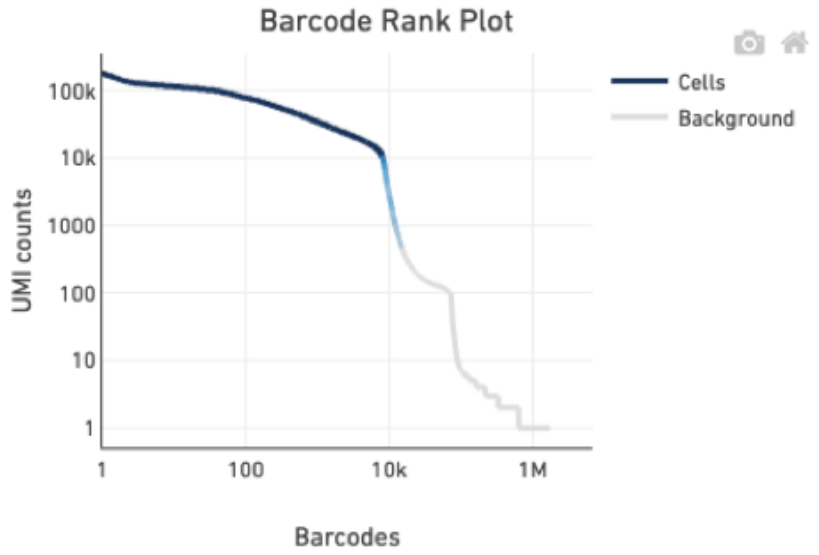
48,398
Mean Reads per Cell

3,976
Median Genes per Cell

Sequencing

Number of Reads	444,047,625
Valid Barcodes	95.0%
Valid UMIs	99.9%
Sequencing Saturation	24.4%
Q30 Bases in Barcode	95.0%
Q30 Bases in RNA Read	93.2%
Q30 Bases in Sample Index	92.8%
Q30 Bases in UMI	92.9%

Cells



Estimated Number of Cells	9,175
Fraction Reads in Cells	88.5%
Mean Reads per Cell	48,398
Median Genes per Cell	3,976
Total Genes Detected	21,633
Median UMI Counts per Cell	17,519

Mapping

Reads Mapped to Genome	90.9%
Reads Mapped Confidently to Genome	84.6%
Reads Mapped Confidently to Intergenic Regions	3.5%
Reads Mapped Confidently to Intronic Regions	12.1%
Reads Mapped Confidently to Exonic Regions	69.1%
Reads Mapped Confidently to Transcriptome	65.7%
Reads Mapped Antisense to Gene	1.6%

Sample

Sample ID	Sample_1
Sample Description	
Chemistry	Single Cell 3' v3
Transcriptome	mm10_eGFP-
Pipeline Version	3.1.0

5,663

Estimated Number of Cells

38,427

Mean Reads per Cell

1,981

Median Genes per Cell

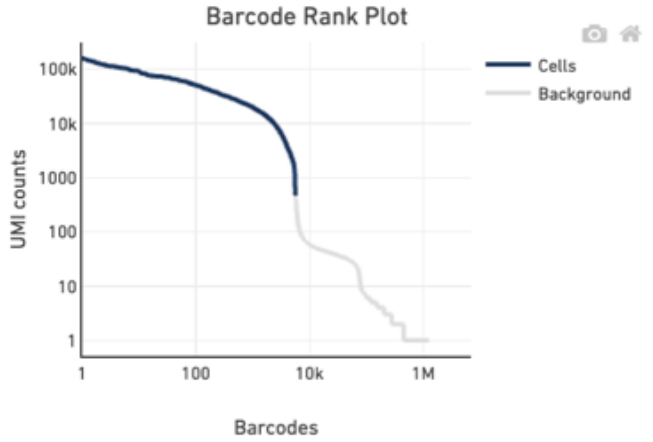
Sequencing

Number of Reads	217,612,940
Valid Barcodes	95.8%
Valid UMIs	99.9%
Sequencing Saturation	55.4%
Q30 Bases in Barcode	94.8%
Q30 Bases in RNA Read	91.1%
Q30 Bases in Sample Index	93.4%
Q30 Bases in UMI	92.2%

Mapping

Reads Mapped to Genome	94.9%
Reads Mapped Confidently to Genome	92.1%
Reads Mapped Confidently to Intergenic Regions	2.8%
Reads Mapped Confidently to Intronic Regions	9.5%
Reads Mapped Confidently to Exonic Regions	79.8%
Reads Mapped Confidently to Transcriptome	75.7%
Reads Mapped Antisense to Gene	1.1%

Cells



Estimated Number of Cells	5,663
Fraction Reads in Cells	94.1%
Mean Reads per Cell	38,427
Median Genes per Cell	1,981
Total Genes Detected	17,517
Median UMI Counts per Cell	7,938

Sample

Sample ID	Sample_1_count
Sample Description	
Chemistry	Single Cell 3' v3
Transcriptome	mm10-3.0.0
Pipeline Version	3.1.0

Metric	Description
Estimated Number of Cells	The number of barcodes associated with cell-containing partitions
Mean Reads per Cell	The total number of sequenced reads divided by the estimated number of cells.
Median Genes per Cell	The median number of genes detected (with nonzero UMI counts) across all cell-associated barcodes.
Number of Reads	Total number of sequenced reads.
Valid Barcodes	Fraction of reads with cell-barcodes that match the whitelist.
Reads Mapped to Genome	Fraction of reads that mapped to the genome.
Reads Mapped Confidently to Genome	Reads Mapped Confidently to Genome.
Reads Mapped Confidently to Transcriptome	Fraction of reads that mapped to a unique gene in the transcriptome with a high mapping quality score
Reads Mapped Confidently to Exonic Regions	Fraction of reads that mapped to the exonic regions of the genome with a high mapping quality score
Reads Mapped Confidently to Intronic Regions	Fraction of reads that mapped to the intronic regions of the genome with a high mapping quality score
Reads Mapped Confidently to Intergenic Regions	Fraction of reads that mapped to the intergenic regions of the genome with a high mapping quality score
Reads Mapped Antisense to Gene	Fraction of reads confidently mapped to the transcriptome, but on the opposite strand of their annotated gene.
Sequencing Saturation	The fraction of reads originating from an already-observed UMI. This is a function of library complexity and sequencing depth.
Q30 Bases in Barcode	Fraction of bases with Q-score at least 30 in the cell barcode sequences.
Q30 Bases in RNA Read	Fraction of bases with Q-score at least 30 in the RNA read sequences.
Q30 Bases in Sample Index	Fraction of bases with Q-score at least 30 in the sample index sequences.
Q30 Bases in UMI	Fraction of bases with Q-score at least 30 in the UMI sequences.
Fraction Reads in Cells	The fraction of cell-barcoded, confidently mapped reads with cell-associated barcodes.
Total Genes Detected	The number of genes with at least one UMI count in any cell.
Median UMI Counts per Cell	The median number of total UMI counts across all cell-associated barcodes.

SC3_v3_NextGem_DI_PBMC_CSP_1K - 1k Human PBMCs stained with a panel of TotalSeq B Antibodies, Dual Indexed

Summary | Analysis

1,200
Estimated Number of Cells

36,996
Mean Reads per Cell

1,700
Median Genes per Cell

Sequencing

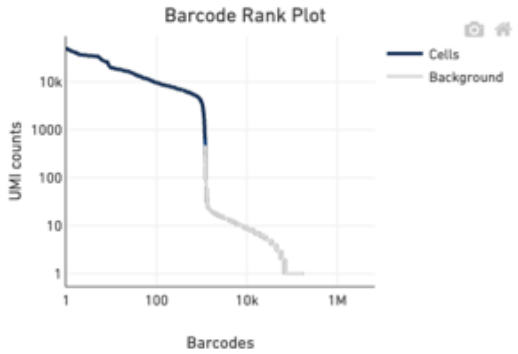
Number of Reads	44,395,031
Valid Barcodes	98.5%
Valid UMIs	99.9%
Sequencing Saturation	69.8%
Q30 Bases in Barcode	97.3%
Q30 Bases in RNA Read	95.3%
Q30 Bases in UMI	97.2%

Mapping

Reads Mapped to Genome	97.4%
Reads Mapped Confidently to Genome	94.9%
Reads Mapped Confidently to Intergenic Regions	3.9%
Reads Mapped Confidently to Intronic Regions	29.7%
Reads Mapped Confidently to Exonic Regions	61.3%
Reads Mapped Confidently to Transcriptome	58.9%
Reads Mapped Antisense to Gene	1.0%

Antibody Sequencing

Cells



Estimated Number of Cells	1,200
Fraction Reads in Cells	94.6%
Mean Reads per Cell	36,996
Median Genes per Cell	1,700
Total Genes Detected	19,169
Median UMI Counts per Cell	5,682

Sample

Sample ID	SC3_v3_NextGem_DI_PBMC_CSP_1K
Sample Description	1k Human PBMCs stained with a panel of TotalSeq B Antibodies, Dual Indexed
Chemistry	Single Cell 3' v3
Reference Path	…nger-4.0.0/refdata-gex-GRCh38-2020-A
Transcriptome	GRCh38-2020-A
Pipeline Version	cellranger-4.0.0

Antibody Sequencing

Number of Reads
Total number of Antibody library reads.

Valid Barcodes
Fraction of Antibody library reads with a barcode found in or corrected to one that is found in the whitelist.

Valid UMIs
Fraction of Antibody library reads with valid UMIs.

Sequencing Saturation
The fraction of Antibody library reads originating from an already-observed UMI. This is a function of library complexity and sequencing depth. More specifically, this is the fraction of confidently mapped, valid cell-barcode, valid UMI reads that had a non-unique (cell-barcode, UMI, CRISPR feature barcode).

Q30 Bases in Barcode
Fraction of Antibody library cell barcode bases with Q-score >= 30, excluding very low quality/no-call (Q <= 2) bases from the denominator.

Q30 Bases in Antibody Read
Fraction of Antibody library read bases with Q-score >= 30, excluding very low quality/no-call (Q <= 2) bases from the denominator. This is Read 2 for the Single Cell 3' v3 and Single Cell 5' chemistries.

Q30 Bases in UMI
Fraction of Antibody library UMI bases with Q-score >= 30, excluding very low quality/no-call (Q <= 2) bases from the denominator.

Version

Antibody Application

Fraction Antibody Reads
Fraction of Antibody library reads that contain a recognized antibody barcode.

Fraction Antibody Reads Usable
Fraction of Antibody library reads that contain a recognized antibody barcode, a valid UMI, and a cell-associated barcode.

Antibody Reads Usable per Cell
Number of Antibody library reads usable divided by the number of cell-associated barcodes.

Fraction Reads in Barcodes with High UMI Counts
Fraction of Antibody library reads that was lost after removing barcodes with unusually high UMI counts (possibly aggregates).

Fraction Unrecognized Antibody
Among all Antibody library reads, the fraction with an unrecognizable antibody barcode.

Antibody Reads in Cells
Among Antibody library reads with a recognized antibody barcode, a valid UMI, and a valid barcode, the fraction associated with cell-containing partitions.

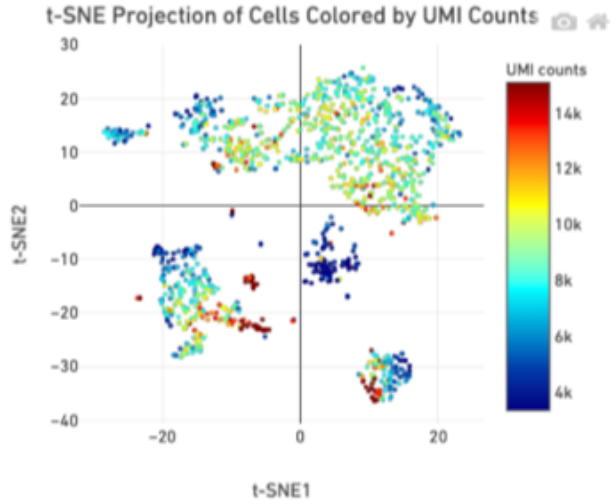
Median UMIs per Cell (summed over all recognized antibody barcodes)
Median UMIs per Cell (summed over all recognized antibody barcodes).

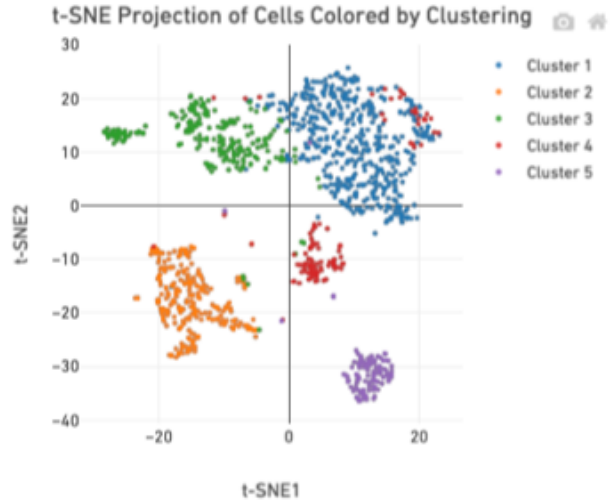
SC3_v3_NextGem_DI_PBMC_CSP_1K - 1k Human PBMCs stained with a panel of TotalSeq B Antibodies, Dual Indexed

Summary | Analysis

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	
ID	Name	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value	L2FC	p-value
ENSG00000172116	CD8B	2.88	8e-17	-3.92	2e-13	-0.92	4e-1	-0.08	1e+0	-4.54	2e-6
ENSG00000138795	LEF1	2.59	3e-17	-3.74	1e-15	-0.84	4e-1	0.69	1e+0	-4.63	4e-8
ENSG00000186854	TRABD2A	2.54	2e-16	-3.89	8e-16	-0.85	4e-1	0.64	1e+0	-2.89	2e-4
ENSG00000126353	CCR7	2.29	9e-15	-4.30	6e-20	-1.34	2e-2	0.06	1e+0	-0.27	1e+0
ENSG00000204677	FAM153CP	2.21	4e-11	-3.21	2e-10	-1.43	4e-2	1.73	1e+0	-2.08	2e-2
ENSG00000237943	PRKCQ-AS1	1.94	2e-10	-3.91	6e-17	0.01	1e+0	0.96	1e+0	-5.65	3e-10
ENSG00000081059	TCF7	1.90	2e-10	-3.54	5e-16	-0.08	1e+0	0.90	1e+0	-3.27	6e-6
ENSG00000127152	BCL11B	1.89	2e-10	-4.13	3e-20	0.00	1e+0	1.13	1e+0	-4.52	2e-9
ENSG00000104660	LEPROTL1	1.77	2e-9	-2.07	2e-7	-0.06	1e+0	-0.76	1e+0	-2.45	5e-4
ENSG00000245164	LINC00861	1.69	2e-7	-3.87	3e-14	0.14	1e+0	1.42	1e+0	-4.72	4e-7

Previous | Page 1 of 25 | 10 rows | Next

Troubleshooting examples

The analysis detected some issues. Details »

Alert	Value	Detail
⚠ **Low Fraction Reads Confidently Mapped To Transcriptome**	24.4%	Ideal > 30%. This can indicate use of the wrong reference transcriptome, a reference transcriptome with overlapping genes, poor library quality, poor sequencing quality, or reads shorter than the recommended minimum. Application performance may be affected.

Estimated Number of Cells

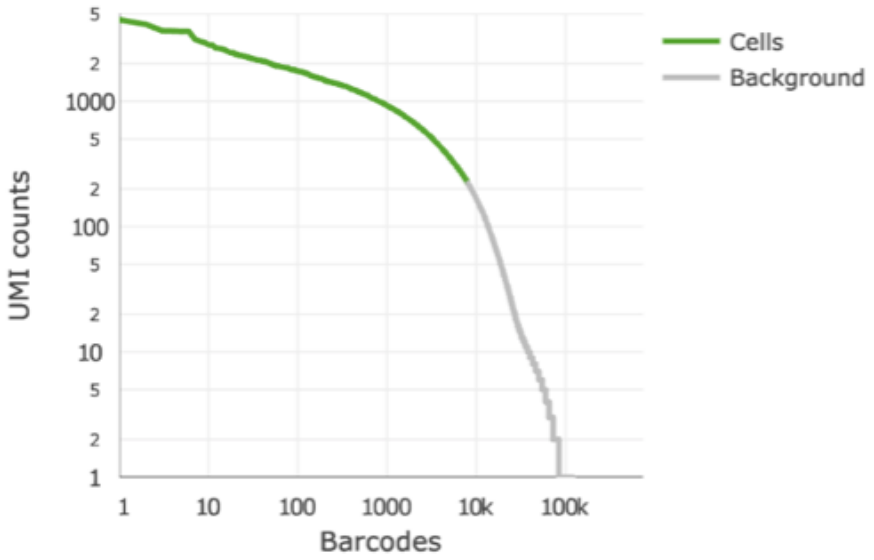
8,109

Mean Reads per Cell	Median Genes per Cell
46,965	**334**

Sequencing

Number of Reads	380,844,038
Valid Barcodes	82.0%
Sequencing Saturation	92.8%
Q30 Bases in Barcode	95.8%
Q30 Bases in RNA Read	82.9%
Q30 Bases in Sample Index	91.4%
Q30 Bases in UMI	95.2%

Cells



Estimated Number of Cells	8,109
Fraction Reads in Cells	74.8%
Mean Reads per Cell	46,965
Median Genes per Cell	334
Total Genes Detected	25,426
Median UMI Counts per Cell	426

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
- Kmer Content

Basic Statistics

Measure	Value
Filename	Pool_1_S1_L001_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	380844038
Sequences flagged as poor quality	0
Sequence length	150
%GC	36

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTT	103227946	27.105044506433888	No Hit
GCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTT	8592717	2.256229884843307	No Hit
AGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTT	4198600	1.1024460359282295	No Hit

Single Cell 5' assay after reverse transcription:

New in Cell Ranger 4.0!

When analyzing 3’ Gene Expression data, Cell Ranger 4.0 trims the template switch oligo (TSO) sequence from the 5’ end of Read-2 and the poly-A sequence from the 3’ end before aligning reads to the reference transcriptome. This behavior is different from Cell Ranger 3.1, which does not perform any trimming.

A full length cDNA molecule is normally flanked by the 30-bp TSO sequence, AAGCAGTGGTATCAACGCAGAGTACATGGG, at the 5' end and the poly-A sequence at the 3' end. Some fraction of sequencing reads are expected to contain either or both of these sequences, depending on the fragment size distribution of the library.

Reads derived from short RNA molecules are more likely to contain either or both TSO and poly-A sequence than longer RNA molecules.

Trimming results in better alignment, with the fraction of reads mapped to a gene increasing by up to 1.5%, because the presence of non-template sequence in the form of either TSO or poly-A confounds read mapping. Trimming improves the sensitivity of the assay as well as the computational efficiency of the pipeline.

The analysis detected some issues. Details »

Alert	Value	Detail
⚠ Low Fraction Reads in Cells	53.4%	Ideal > 70%. Application performance may be affected. Many of the reads were not assigned to cell-associated barcodes. This could be caused by high levels of ambient RNA or by a significant population of cells with a low RNA content, which the algorithm did not call as cells. The latter case can be addressed by inspecting the data to determine the appropriate cell count and using --force-cells.

Estimated Number of Cells

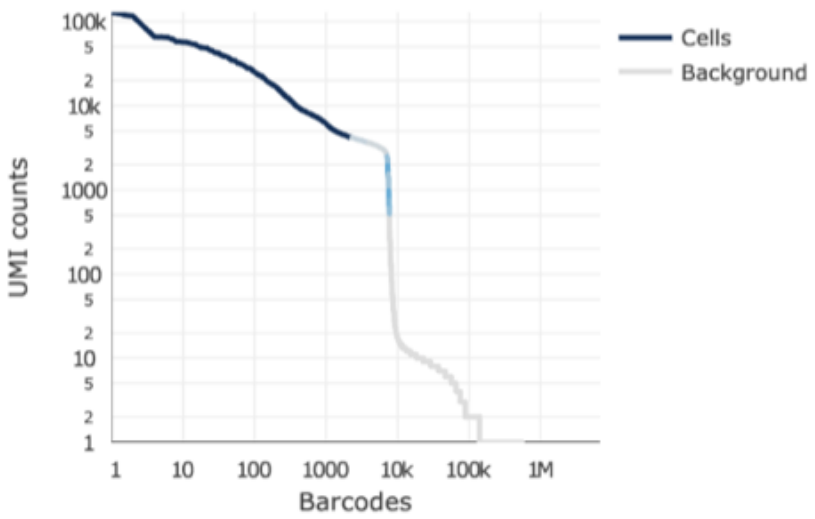
2,553

Mean Reads per Cell	Median Genes per Cell
130,565	2,064

Sequencing

Number of Reads	333,334,060
Valid Barcodes	97.5%
Sequencing Saturation	74.3%
Q30 Bases in Barcode	96.1%
Q30 Bases in RNA Read	90.8%
Q30 Bases in Sample Index	94.8%
Q30 Bases in UMI	93.3%

Cells



Estimated Number of Cells	2,553
Fraction Reads in Cells	53.4%
Mean Reads per Cell	130,565
Median Genes per Cell	2,064
Total Genes Detected	18,966
Median UMI Counts per Cell	5,148

Forcing the number of cells

Estimated Number of Cells

8,000

Mean Reads per Cell	Median Genes per Cell
41,666	1,586

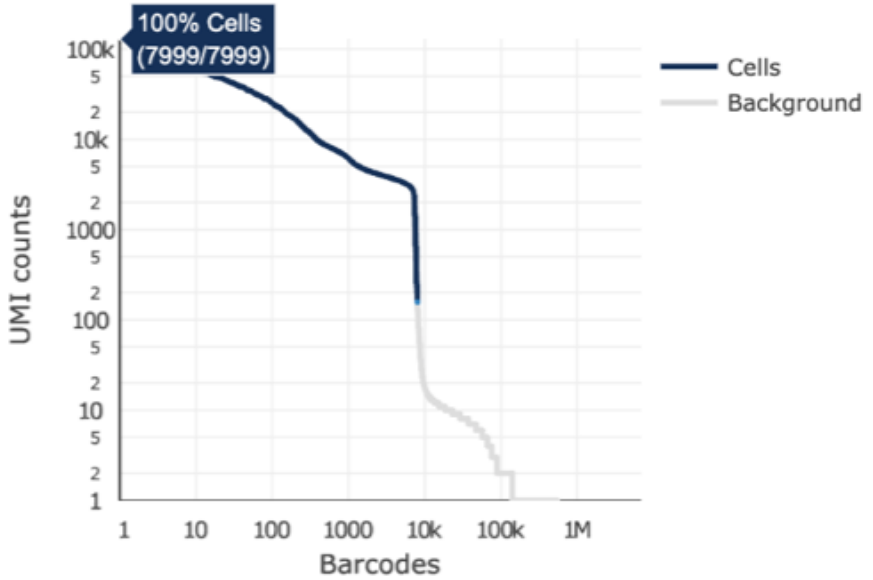
Sequencing

Number of Reads	333,334,060
Valid Barcodes	97.5%
Sequencing Saturation	74.3%
Q30 Bases in Barcode	96.1%
Q30 Bases in RNA Read	90.8%
Q30 Bases in Sample Index	94.8%
Q30 Bases in UMI	93.3%

Mapping

Reads Mapped to Genome	94.9%
Reads Mapped Confidently to Genome	93.0%
Reads Mapped Confidently to Intergenic Regions	5.0%
Reads Mapped Confidently to Intronic Regions	38.4%
Reads Mapped Confidently to Exonic Regions	49.6%
Reads Mapped Confidently to Transcriptome	45.7%
Reads Mapped Antisense to Gene	1.4%

Cells



Estimated Number of Cells	8,000
Fraction Reads in Cells	97.9%
Mean Reads per Cell	41,666
Median Genes per Cell	1,586
Total Genes Detected	19,782
Median UMI Counts per Cell	3,614

Sample

Name	_GEX_count_8k
Description	
Transcriptome	GRCh38
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.0

Conclusion

Always perform QC of your libraries!

Be aware of library specifics → critical mindset!