ELIXIR-IIB Training Platform

Single-Cell RNA Sequencing and Data Analysis

Theory Refresher and Software Overview: Cell Ranger

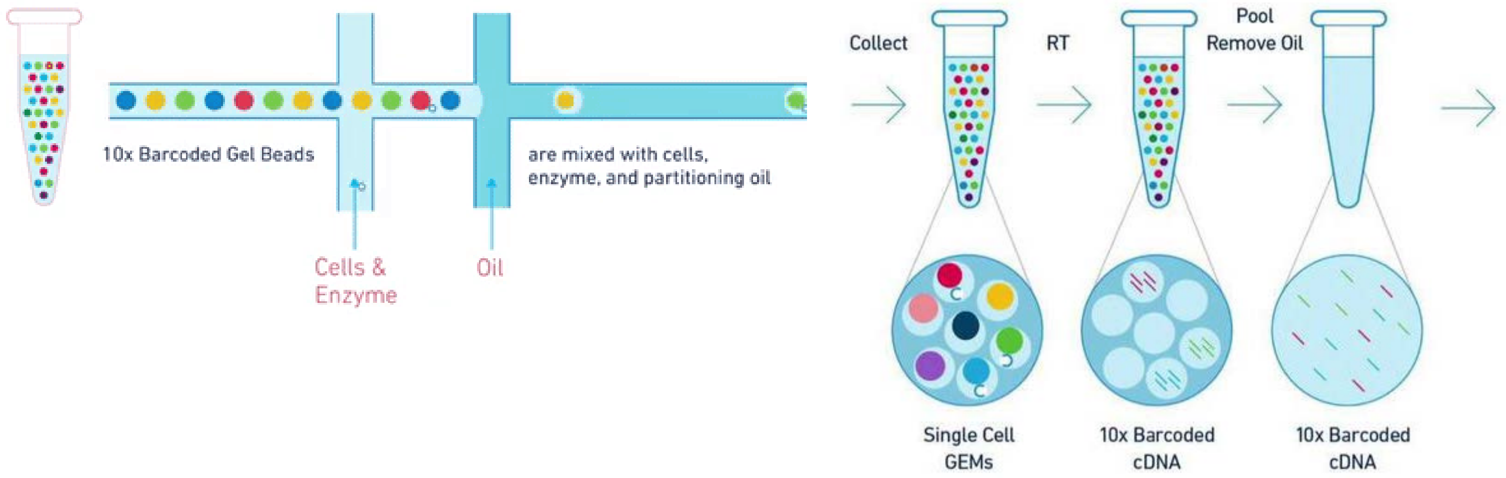
Francesco Panariello, Bioinformatician

INFRASTRUTTURA ITALIANA DI BIOINFORMATICA

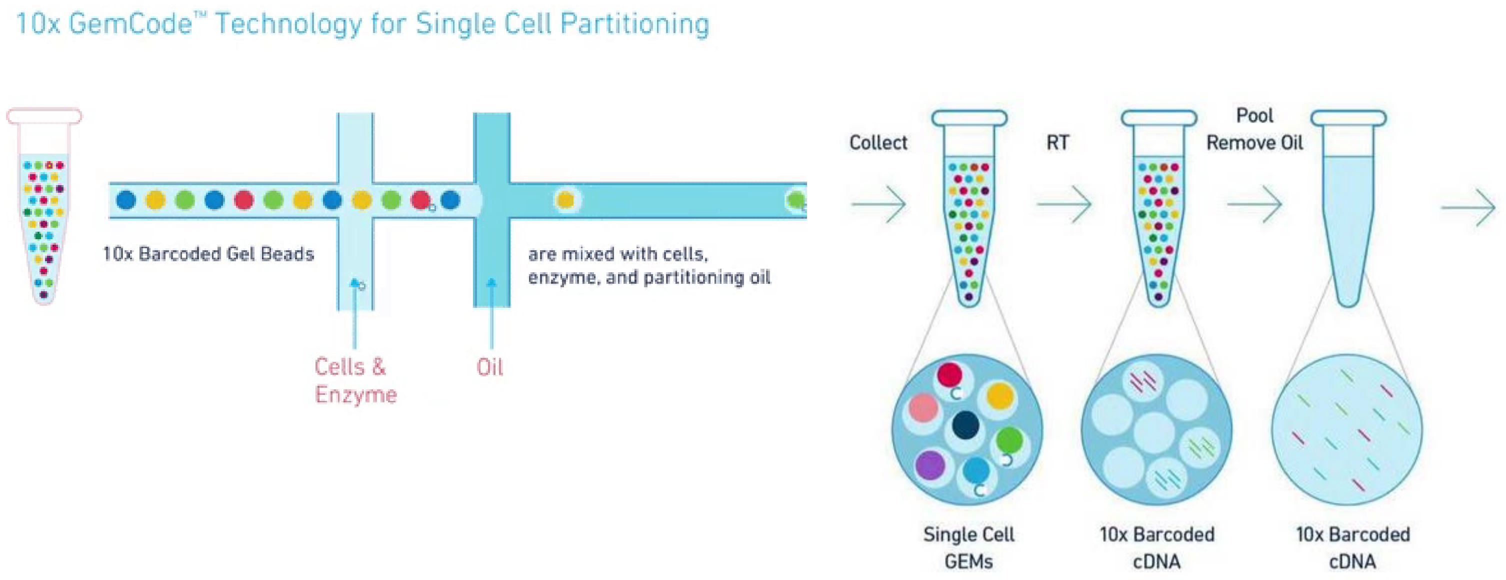
TELETHON INSTITUTE OF GENETICS AND MEDICINE

Single-cell Gene Expression Profiling: Experimental Overview

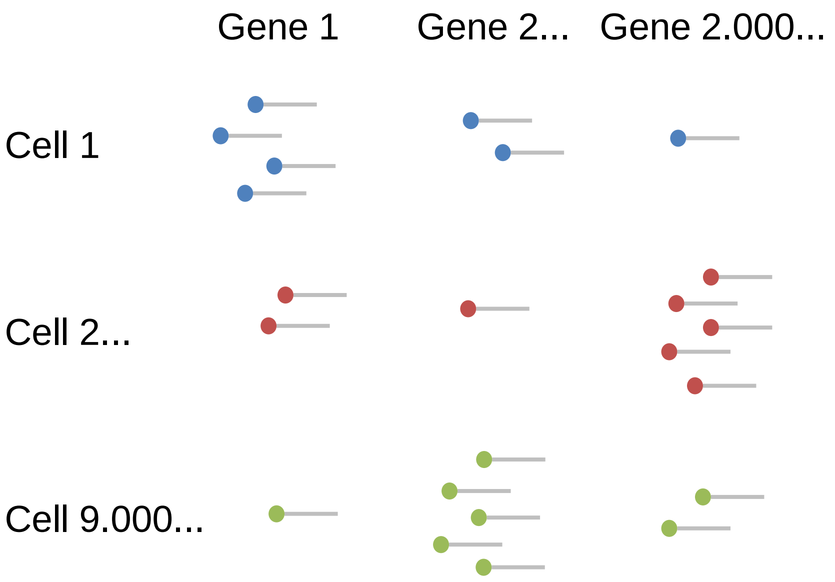
10x GemCode™ Technology for Single Cell Partitioning



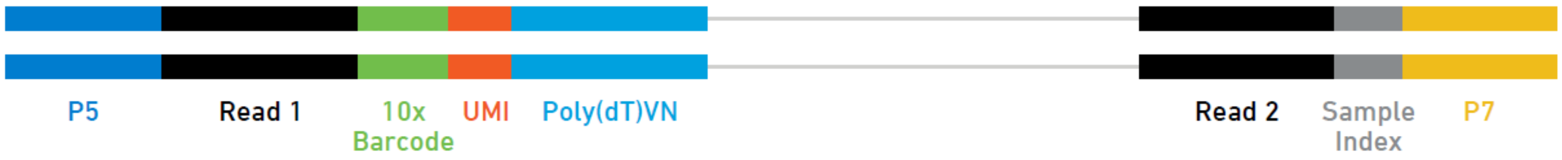
Single-cell Gene Expression Profiling: Experimental Overview



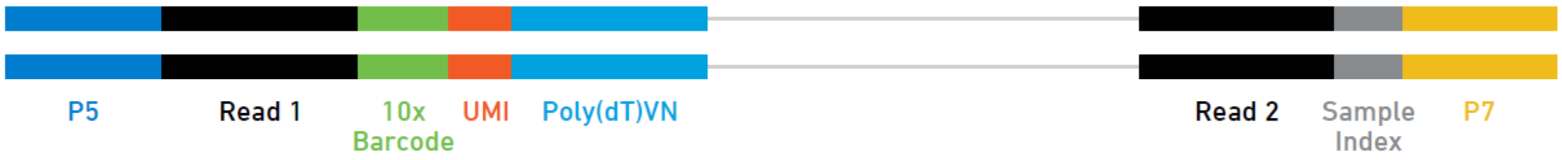
Transcriptional profiling of individual cells



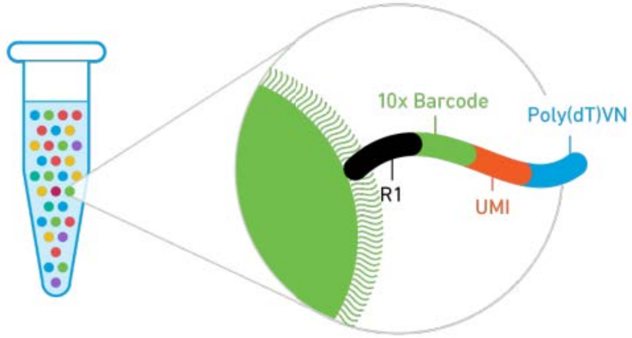
Final fragment from Chromium™ Single Cell 3' v2 library



Final fragment from Chromium™ Single Cell 3' v2 library

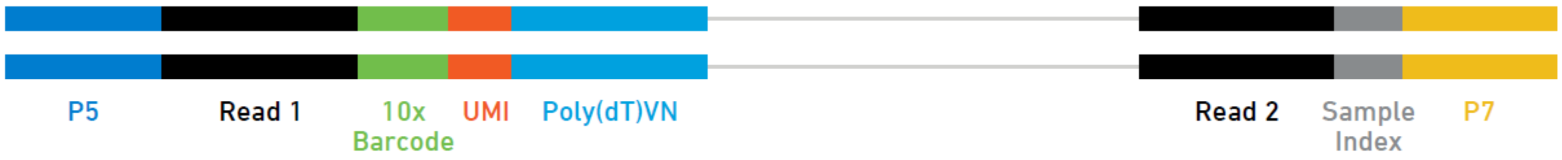


10x™ Barcode



The 16bp 10x barcode is unique to each Gel Bead and tells you which cell the transcript is from.

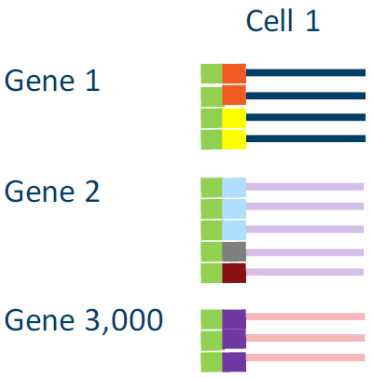
Final fragment from Chromium™ Single Cell 3' v2 library



10x™ Barcode

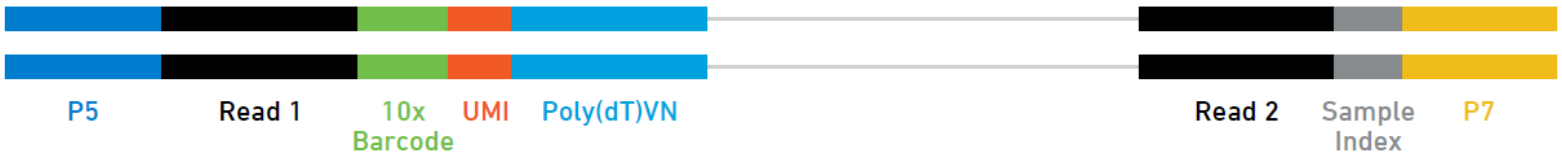
The 16bp 10x barcode is unique to each Gel Bead and tells you which cell the transcript is from.

Unique Molecular Identifier (UMI)



The 10bp UMI enables accurate quantitation of cell expression levels.

Final fragment from Chromium™ Single Cell 3' v2 library



10x™ Barcode

The 16bp 10x barcode is unique to each Gel Bead and tells you which cell the transcript is from.

Unique Molecular Identifier (UMI)

The 10bp UMI enables accurate quantitation of cell expression levels.

Sample index

The 8bp sample index allows to assign each barcoded read to its sample of origin.

Cell Ranger^{TM} Pipeline

Transcriptomic computational workflow: a standard approach

10x GENOMICS™

RNA-seq bulk

Demultiplexing

⇩

FastQC
MultiQC

Quality control / Trimming

TrimGalore
Trimmomatic
BBDuK

⇩

STAR
BWA
Kalisto
TopHat

Reads Alignment

⇩

Reads Quantification

HT-seq
featureCounts
RSEM

⇩

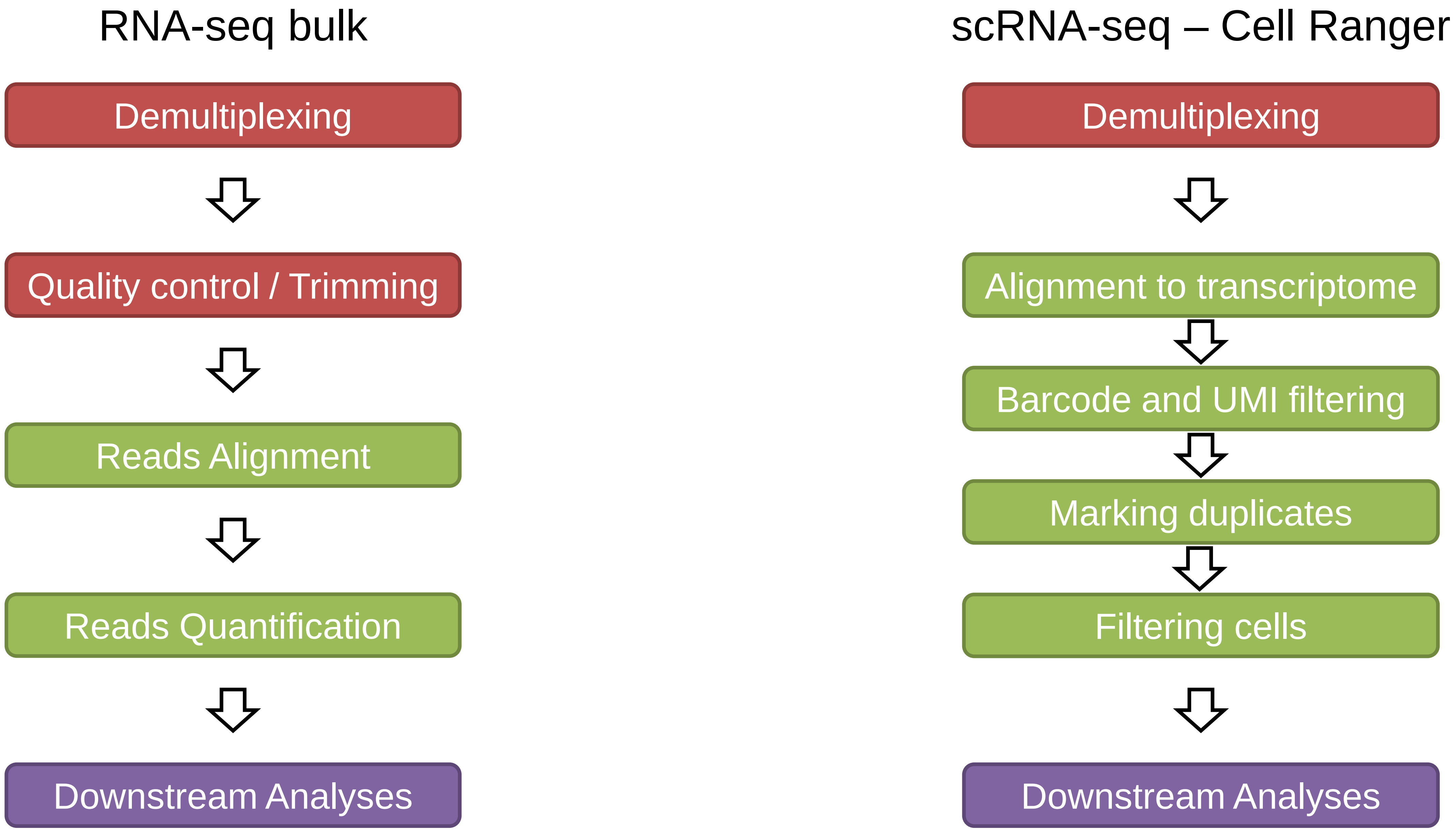
Downstream Analyses

Primary Analysis

Secondary Analysis

Tertiary Analysis

Transcriptomic computational workflow: a standard approach



What is Cell Ranger?

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

Cell Ranger includes four pipelines relevant to single-cell gene expression experiments

cellranger mkfastq

cellranger count

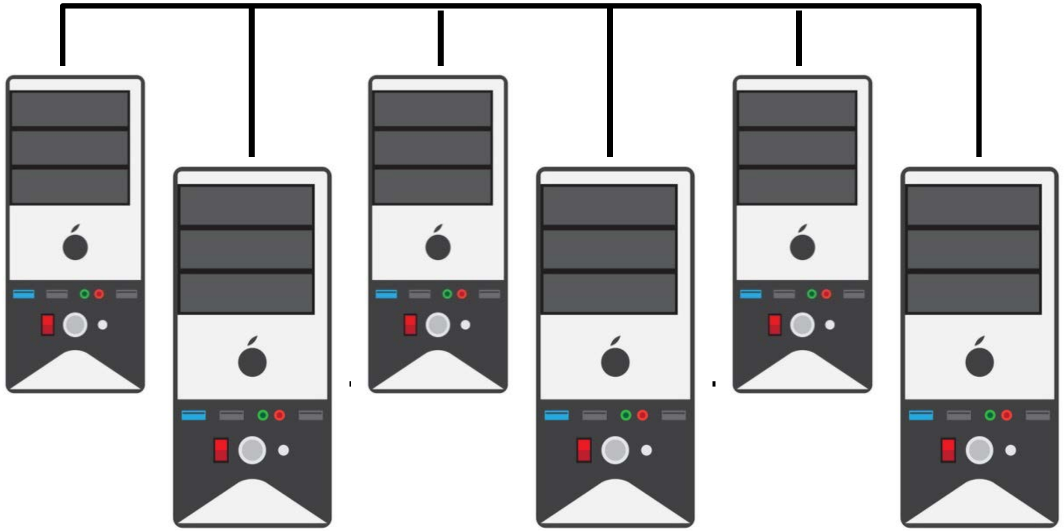
cellranger aggr

cellranger reanalyze

Cell Ranger™ Pipeline: System Requirements

Local Mode

- Run on single, standalone Linux system
- CentOS/RedHat 5.2+ or Ubuntu 8.04+
- 8+ cores, 64GB RAM



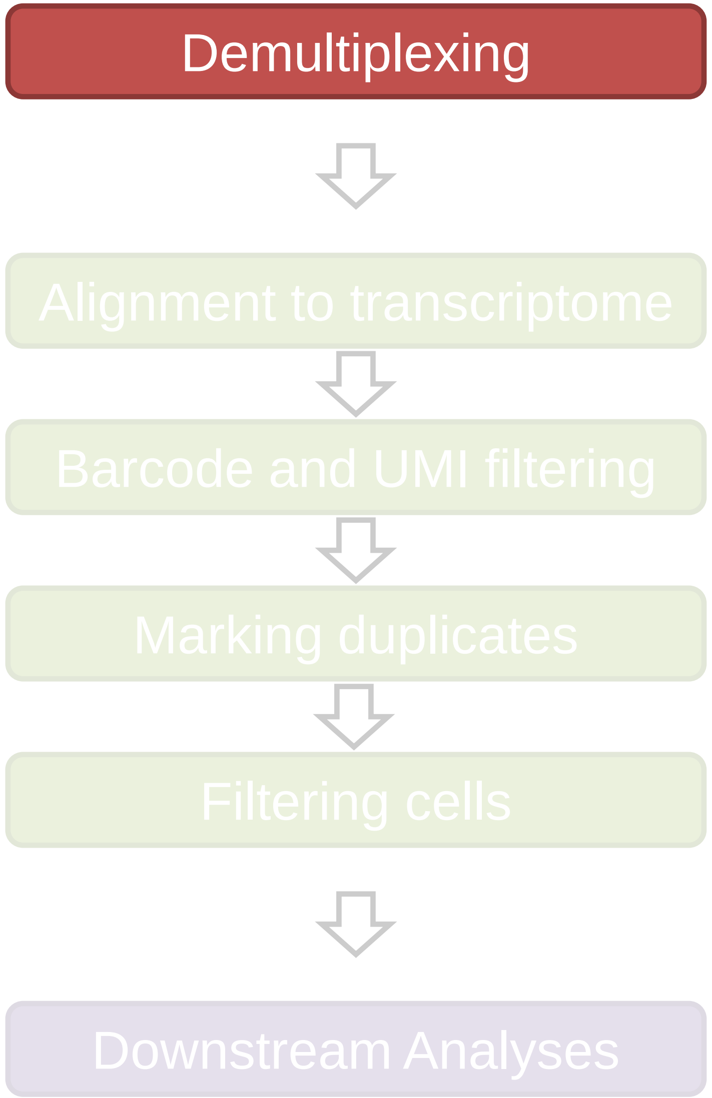
Cluster Mode

- Run on SGE and LSF
- Each node must have 8+ cores and 8GB+ RAM/core
- Shared filesystem between nodes (e.g. NFS)

Cell Ranger™ Pipeline: Demultiplexing

Demultiplexing is the step through which sequencing reads are divided into separate **fastq files** for each sample index.

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 1:N:0:12
CCTTAAAAATGTACCCATTAGGCCTAAGTAGCTAGCTGGGCCC
+
BBBBFFFFFIIIIIFIII<FFFFFIIIBBBBUUUUFIIIIDDDDIIIIFIFIFIFII
```

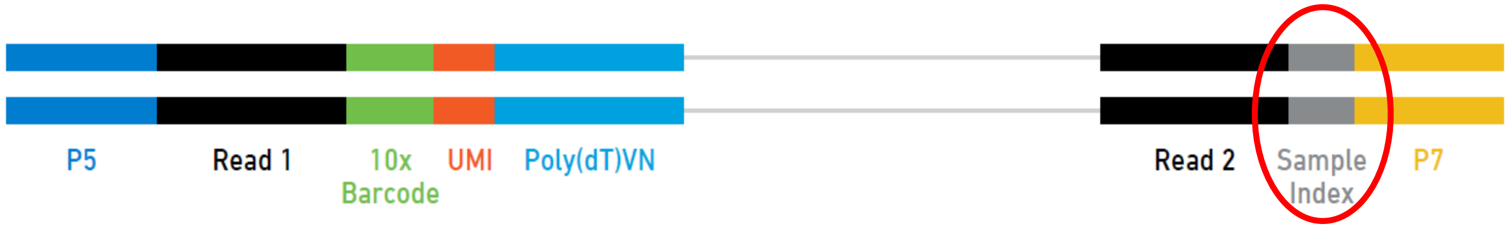


Cell Ranger^{TM} Pipeline: Demultiplexing

The *cellranger mkfastq* pipeline allows to demultiplex an Illumina sequencing run folder into FASTQ files.

Cell Ranger™ Pipeline: Demultiplexing

The *cellranger mkfastq* pipeline allows to demultiplex an Illumina sequencing run folder into FASTQ files.



- 10x samples indeces are used to assign reads to their sample of origin
- They are supplied on a 96 well plate
- Each 10x sample index is composed of 4 oligos
  - ➢ 10x index: SI-GA-A1
  - ➢ 4 Oligos: "GGTTTACT", "CTAAACGG", "TCGGCGTC", "AACCGTAA”

Cell Ranger™ Pipeline: Demultiplexing

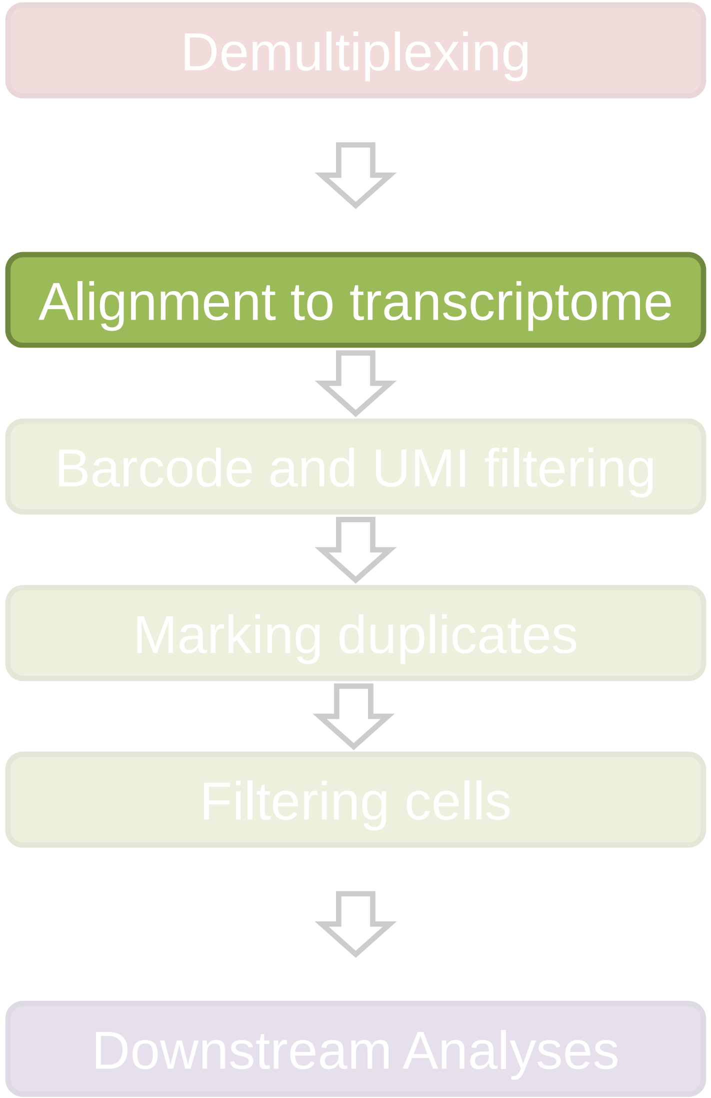
The *cellranger mkfastq* pipeline allows to demultiplex an Illumina sequencing run folder into FASTQ files.

The association between indeces and samples is provided through a samplesheet in .csv (comma-separated values) format.

```
Lane,Sample,Index
*,Sample_1,SI-GA-A1
*,Sample_2,SI-GA-A2
*,Sample_3,SI-GA-A3
```

Cell Ranger™ Pipeline: Alignment to transcriptome

- **Read alignment** consists in the assignment of sequencing reads to the most likely locus of origin.
- Reads mapping can be performed on the genome, as well as on the transcriptome, in a **fasta** format.
- To speed up the process, fasta files are usually indexed.
- A **GTF** file is also used to provide information about gene location, biotype, etc.



Cell Ranger™ Pipeline: Alignment to transcriptome

10x GENOMICS™

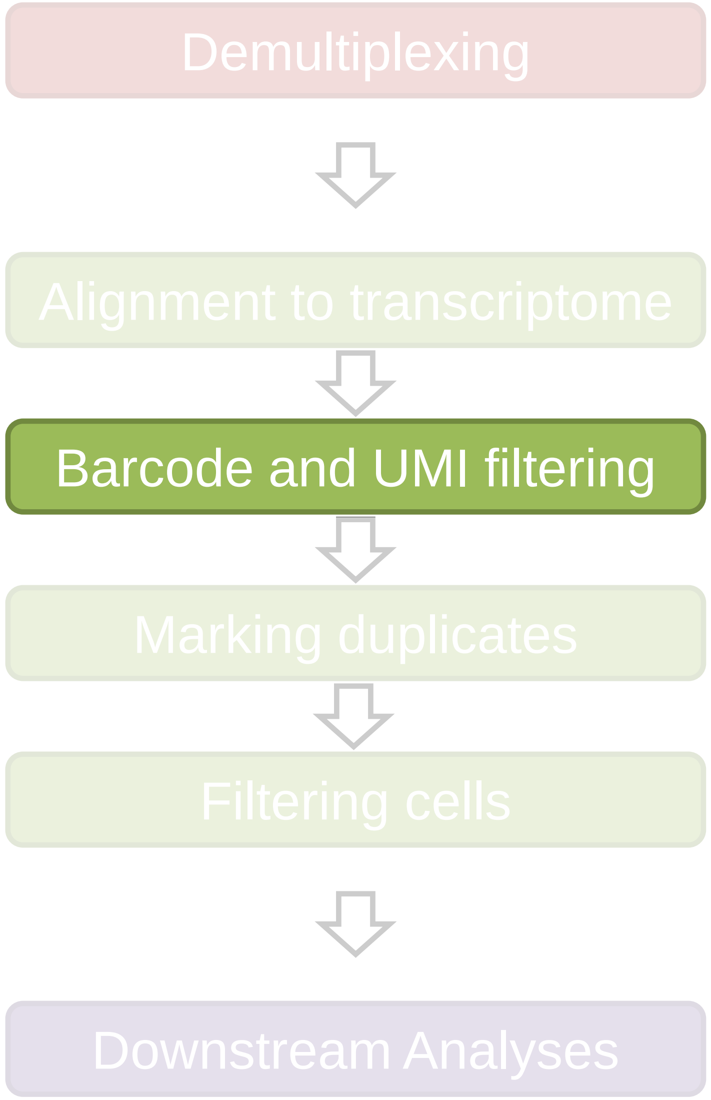
References

- 10x annotation uses ENSEMBL genomes and gene annotations.
- 10x pre-built references: Human (hg19 and GRCh38), Mouse, Human and Mouse.
- Bundled *mkgtf* utility filters a GTF file by key value pairs in the attributes column for transcript biotype (e.g. protein-coding, non-coding, linc RNA).
- Bundled *mkref* utility generates a 10x reference package from any FASTA and GTF gene file (STAR compatible).

Alignment

- Alignment done via **STAR** (Spliced Transcripts Alignment to a Reference):
  - Robust, open-source, junction-aware RNA-seq aligner.
  - Aligns reads to the genome and transcriptome simultaneously.
- STAR memory usage:
  - *mkref* script builds STAR reference such that it uses max 16 GB of memory.
- Only use confidently mapped reads aligning to transcriptome.

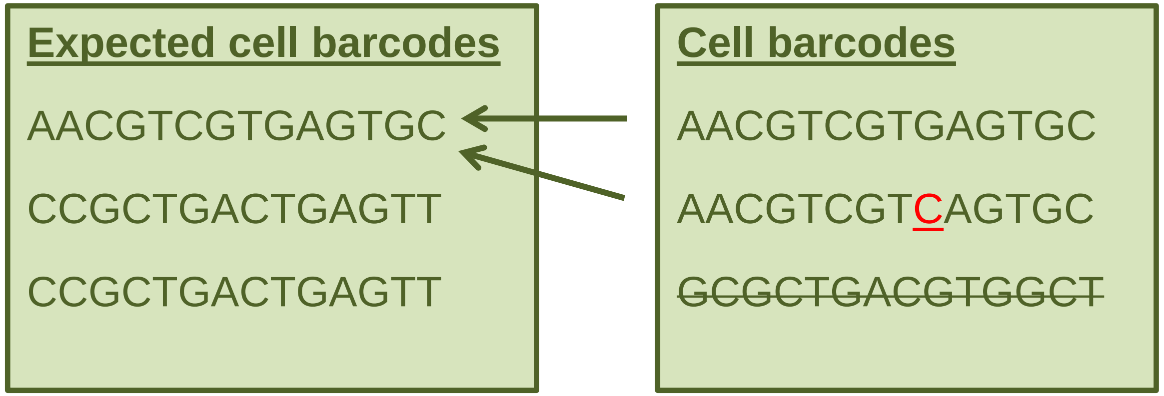
Cell Ranger™ Pipeline: Barcode and UMI filtering

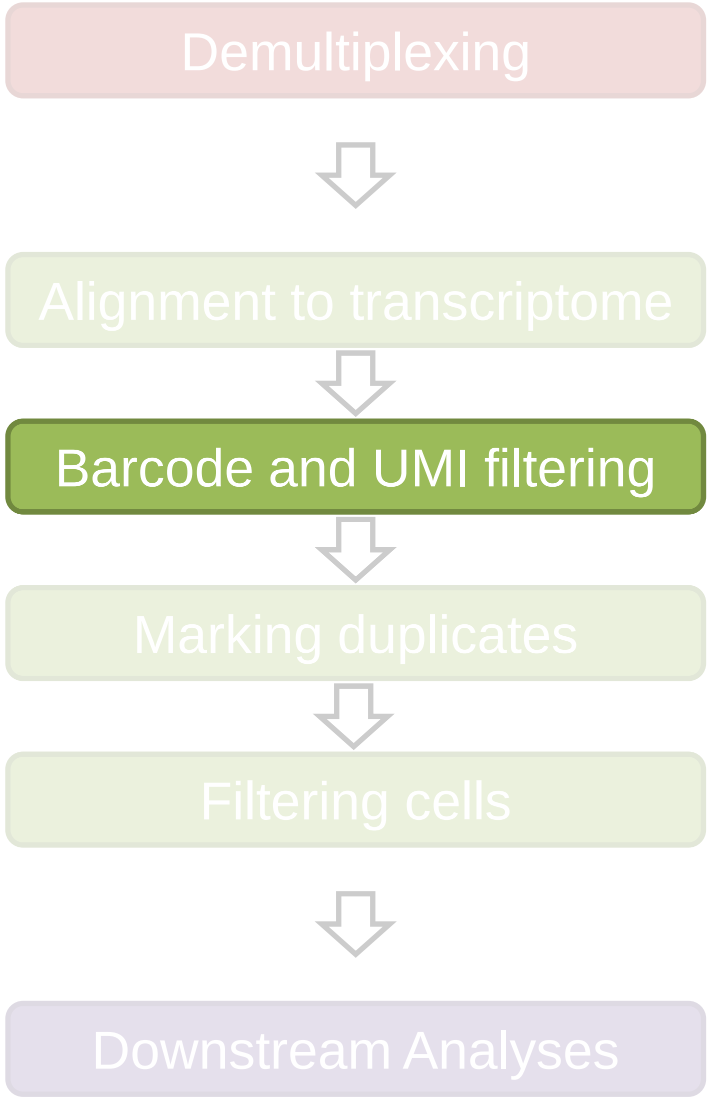


Cell Ranger[TM] Pipeline: Barcode and UMI filtering

Cell barcodes:

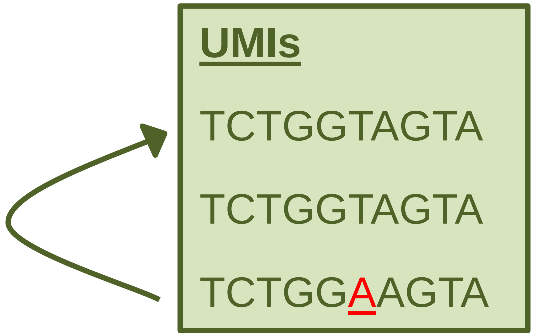
- Must be on a static list of known cell barcode sequences
- May be **one** mismatch away from the list ONLY IF the mismatch occurs at a low-quality position (the barcode is then corrected

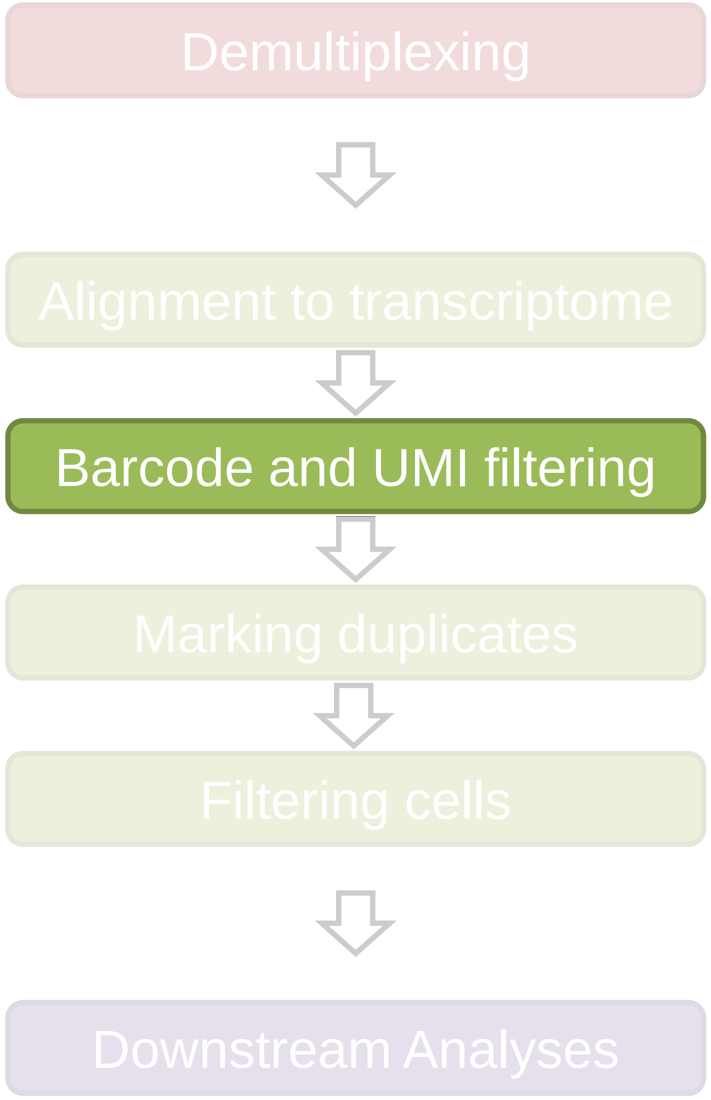




UMIs:

- Must not be a homopolymer, e.g. AAAAAAAAAA
- Must not contain N
- Must not contain bases with base quality < 10
- UMIs that are 1 nucleotide mismatch away from a higher-count UMI are corrected to that UMI if they share a cell barcode and gene.





Cell Ranger™ Pipeline: Marking duplicates and Filtering cells

Marking duplicates:

- Record which reads are duplicates of the same RNA molecule.
- Count only the unique UMIs as unique RNA molecules.

→ Unfiltered gene-barcode matrix

Filtering cells:

- Sum UMI counts for each barcode.
- Select barcodes with total UMI count ≥ 10% of the 99th percentile of the expected recovered cells.

→ Filtered gene-barcode matrix

Demultiplexing

Alignment to transcriptome

Barcode and UMI filtering

Marking duplicates

Filtering cells

Downstream Analyses

Cell Ranger^{TM} Pipeline: cellranger count

10x GENOMICS™

The *cellranger count* pipeline allows to run all the previous steps (secondary analysis), once per sample.

```
cellranger count --sample=Sample_1 --transcriptome=refdata-cellranger/GRCh38 --fastqs=HAD58ADXX/Sample_1
```

Cell Ranger™ Pipeline: Output files

BAM – Genome-Aligned Reads
• Indexed BAM containing position-sorted, aligned reads • Barcodes and UMIs attached as standard tags

Cell Ranger™ Pipeline: Output files

BAM – Genome-Aligned Reads

- Indexed BAM containing position-sorted, aligned reads
- Barcodes and UMIs attached as standard tags

MEX – Gene/Barcode Matrix

- "Market Exchange" format, a sparse matrix representation
- Suitable for downstream analysis in Python and R

Gene	ATCAGGGACAGA	AGGGAAAATTGA	TTGCCTTACGCG	TGGCGAAGAGAT	TACAATTAAGGC
LOXL4	0	0	0	0	0
PYROXD2	1	0	1	1	0
HPS1	23	12	9	8	3
CNNM1	0	2	1	0	0
GOT1	22	6	7	9	3

Cell Ranger™ Pipeline: Output files

10x GENOMICS™

BAM – Genome-Aligned Reads

- Indexed BAM containing position-sorted, aligned reads
- Barcodes and UMIs attached as standard tags

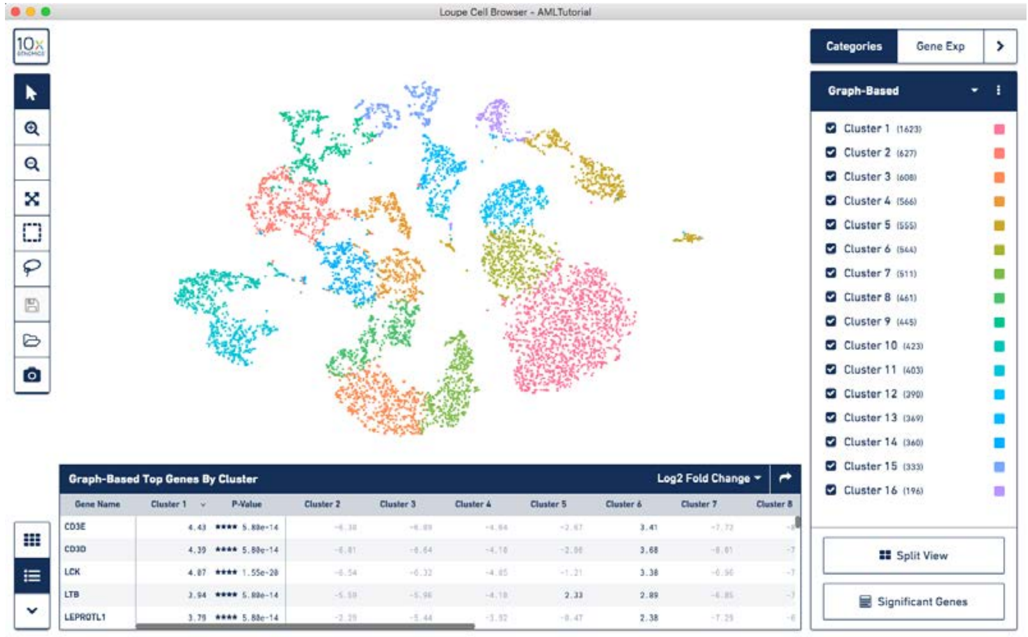
MEX – Gene/Barcode Matrix

- “Market Exchange” format, a sparse matrix representation
- Suitable for downstream analysis in Python and R

.cloupe File - Analysis

- 2D projections
- Cell clustering
- Differential expression
- Interactive exploration

Gene	ATCAGGGACAGA	AGGGAAAATTGA	TTGCCTTACGCG	TGGCGAAGAGAT	TACAATTAAGGC
LOXL4	0	0	0	0	0
PYROXD2	1	0	1	1	0
HPS1	23	12	9	8	3
CNNM1	0	2	1	0	0
GOT1	22	6	7	9	3



Cell Ranger™ Pipeline: Output files

BAM – Genome-Aligned Reads

- Indexed BAM containing position-sorted, aligned reads
- Barcodes and UMIs attached as standard tags

MEX – Gene/Barcode Matrix

- “Market Exchange” format, a sparse matrix representation
- Suitable for downstream analysis in Python and R

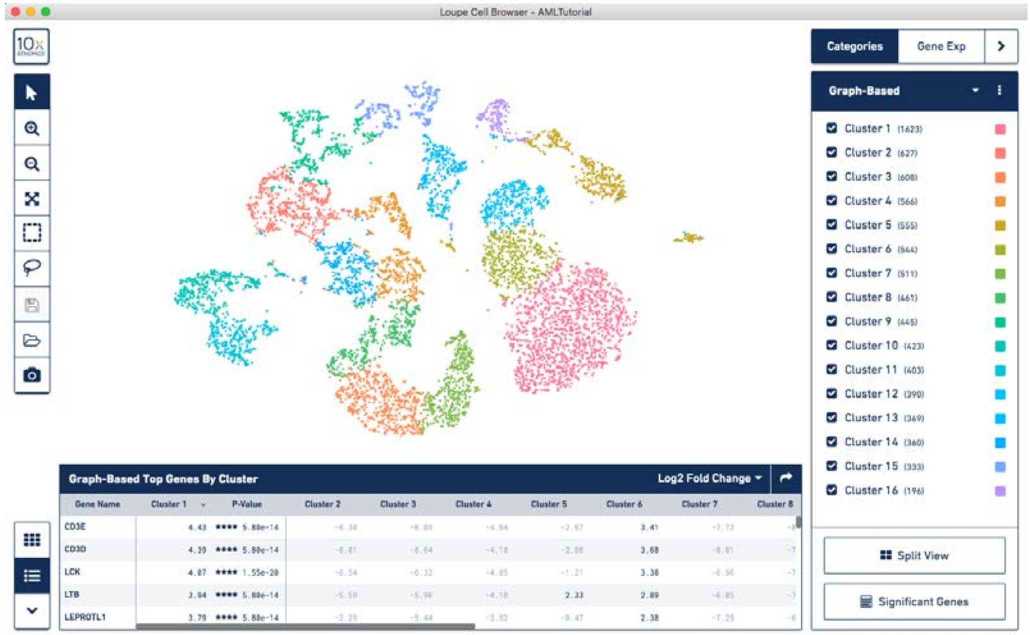
.cloupe Fie - Analysis

- 2D projections
- Cell clustering
- Differential expression
- Interactive exploration

HTML, CSV – Run Summary

- Run metrics and basic static visualizations

Gene	ATCAGGGACAGA	AGGGAAAATTGA	TTGCCTTACGCG	TGGCGAAGAGAT	TACAATTAAGGC
LOXL4	0	0	0	0	0
PYROXD2	1	0	1	1	0
HPS1	23	12	9	8	3
CNNM1	0	2	1	0	0
GOT1	22	6	7	9	3



Cell Ranger™ Pipeline: Cell Ranger Summary Tab

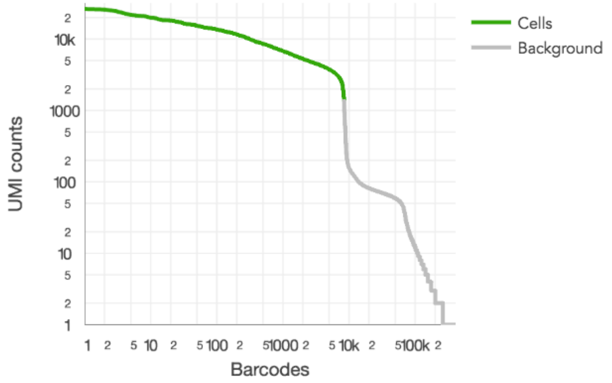
Estimated Number of Cells
8,391

Mean Reads per Cell	Median Genes per Cell
93,441	1,296

Sequencing

Number of Reads	784,064,148
Valid Barcodes	98.5%
Reads Mapped Confidently to Transcriptome	61.4%
Reads Mapped Confidently to Exonic Regions	65.3%
Reads Mapped Confidently to Intronic Regions	24.0%
Reads Mapped Confidently to Intergenic Regions	3.4%
Sequencing Saturation	90.5%
Q30 Bases in Barcode	98.2%
Q30 Bases in RNA Read	78.9%
Q30 Bases in Sample Index	96.4%
Q30 Bases in UMI	98.2%

Cells



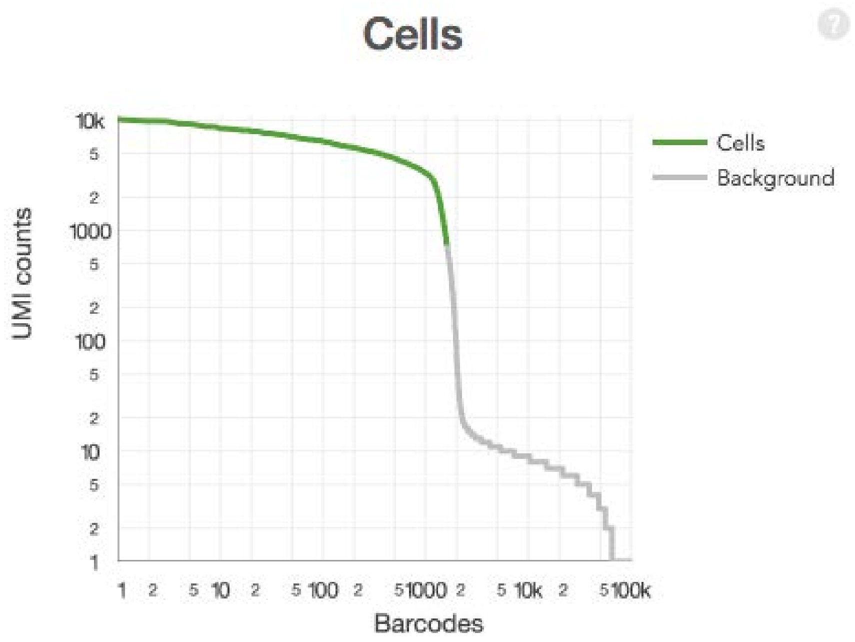
Estimated Number of Cells	8,391
Fraction Reads in Cells	93.2%
Mean Reads per Cell	93,441
Median Genes per Cell	1,296
Total Genes Detected	21,424
Median UMI Counts per Cell	4,082

Sample

Name	pbmc8k
Description	Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Transcriptome	GRCh38
Chemistry	Single Cell 3' v2
Cell Ranger Version	1.3.0

Cell Ranger™ Pipeline: QC Plot

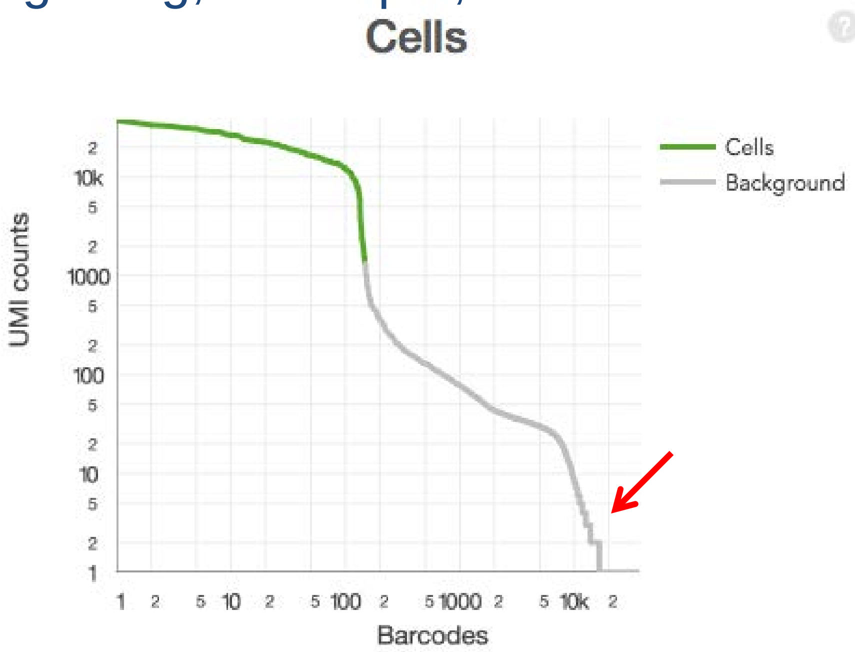
Typical Sample Profile



Defined cliff and knee

Metric	Value
Barcodes	> 90,000
Cell Barcodes	> 1,000
UMIs	> 10,000

Low Barcode Counts
(e.g. Clog, low-depth, low ambient RNA)

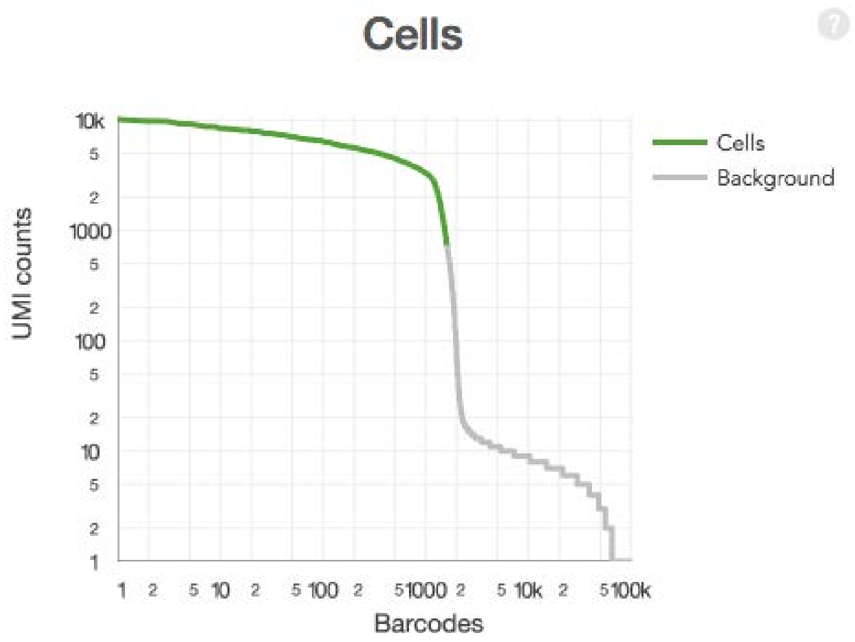


Low number of barcodes detected

Metric	Value
Barcodes	~ 15,000
Cell Barcodes	> 100
UMIs	> 10,000

Cell Ranger™ Pipeline: QC Plot

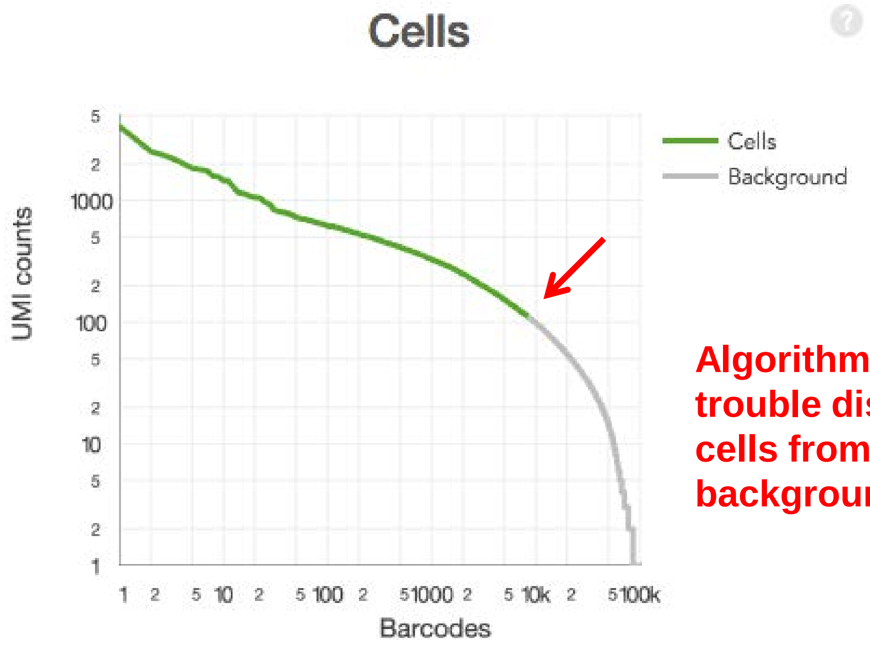
Typical Sample Profile



Defined cliff and knee

Metric	Value
Barcodes	> 90,000
Cell Barcodes	> 1,000
UMIs	> 10,000

Loss of Single Cell Behaviour (e.g. Lysis or Wetting failure)



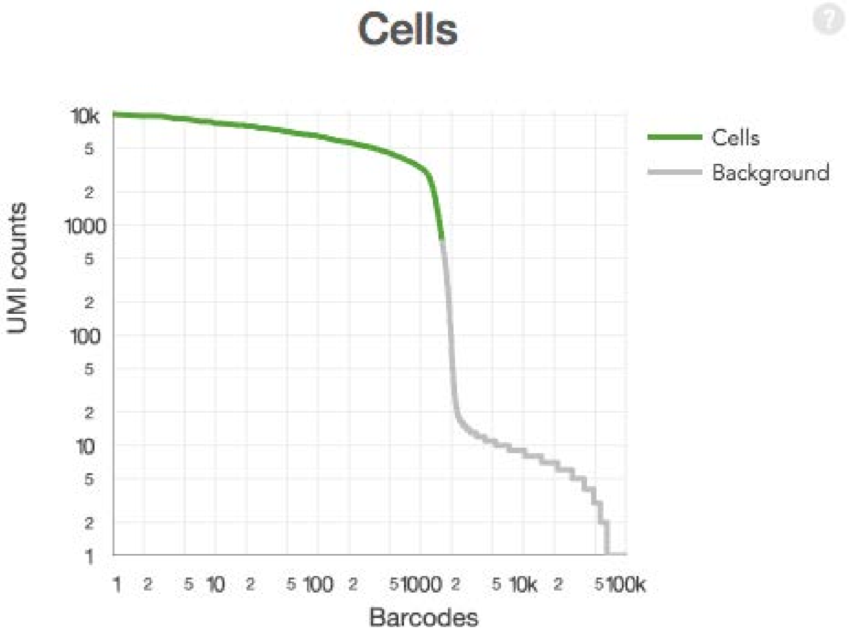
Algorithm has trouble discerning cells from the background

Lack of defined cliff and knee

Metric	Value
Barcodes	> 90,000
Cell Barcodes	~ 10,000
UMIs	> 10,000

Cell Ranger™ Pipeline: QC Plot

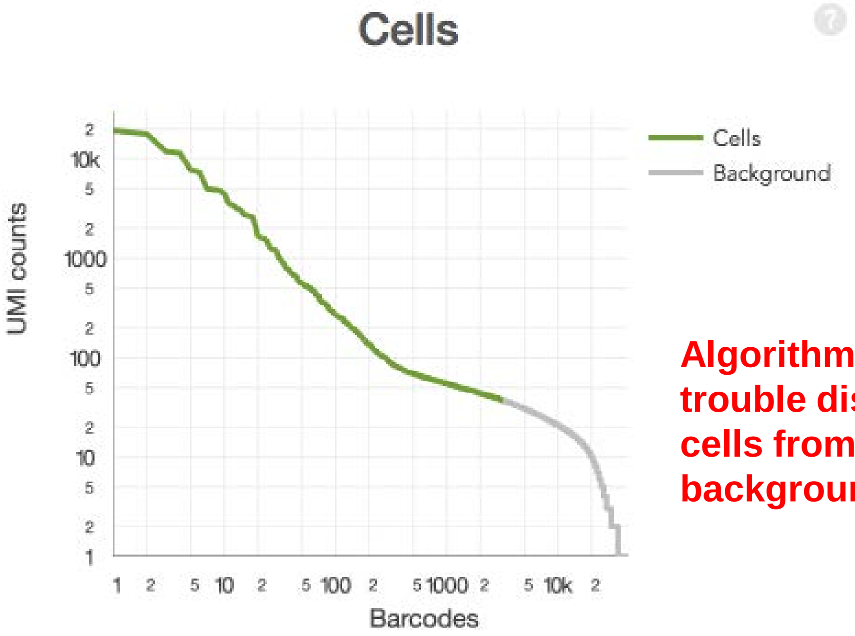
Typical Sample Profile



Defined cliff and knee

Metric	Value
Barcodes	> 90,000
Cell Barcodes	> 1,000
UMIs	> 10,000

Loss of Single Cell Behaviour (e.g. High fraction of ambient RNA)



Algorithm has trouble discerning cells from the background

Lack of defined cliff and knee

Metric	Value
Barcodes with > 1,000 UMIs	Few

Cell Ranger™ Pipeline: Additional Quality Metrics

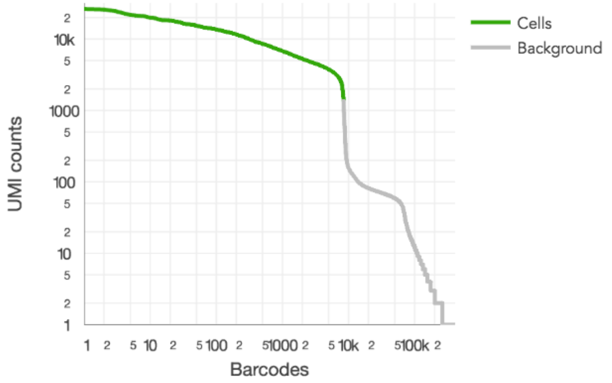
Estimated Number of Cells
8,391

Mean Reads per Cell	Median Genes per Cell
93,441	1,296

Sequencing

Number of Reads	784,064,148
Valid Barcodes	98.5%
Reads Mapped Confidently to Transcriptome ←	61.4%
Reads Mapped Confidently to Exonic Regions	65.3%
Reads Mapped Confidently to Intronic Regions	24.0%
Reads Mapped Confidently to Intergenic Regions	3.4%
Sequencing Saturation	90.5%
Q30 Bases in Barcode	98.2%
Q30 Bases in RNA Read ←	78.9%
Q30 Bases in Sample Index	96.4%
Q30 Bases in UMI	98.2%

Cells



Estimated Number of Cells	8,391
Fraction Reads in Cells	93.2%
Mean Reads per Cell	93,441
Median Genes per Cell	1,296
Total Genes Detected	21,424
Median UMI Counts per Cell	4,082

Sample

Name	pbmc8k
Description	Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Transcriptome	GRCh38
Chemistry	Single Cell 3' v2
Cell Ranger Version	1.3.0

Reads confidently mapped to transcriptome (<30%)

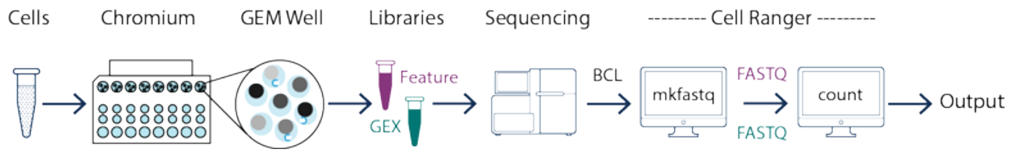
- Reads mapped to wrong genome or different strain
- Read length is too short
- Custom reference contains overlapping genes

Read2 Q30 metrics are low (<70%)

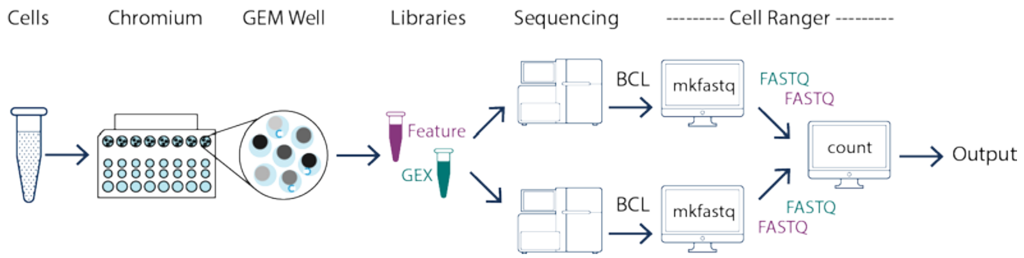
- Sequencing problems
- Suboptimal loading concentration on sequencer

Cell Ranger[TM] Pipeline: Workflows

One Sample, One GEM Well, One Flowcell

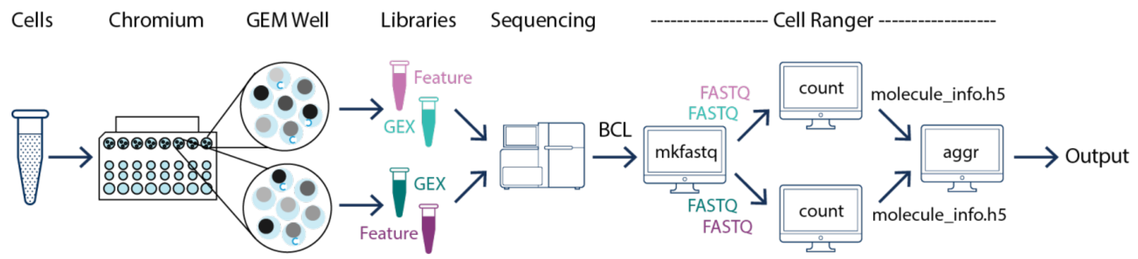


One Sample, One GEM Well, Multiple Flowcells

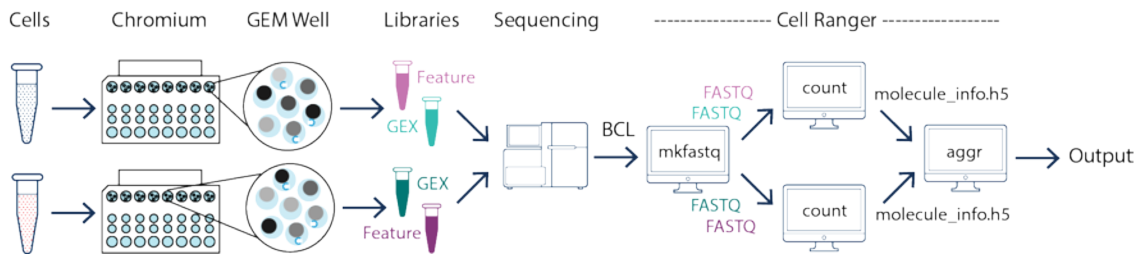


Cell Ranger^{TM} Pipeline: Workflows – cellranger aggr

One Sample, Multiple GEM Wells, One Flowcell

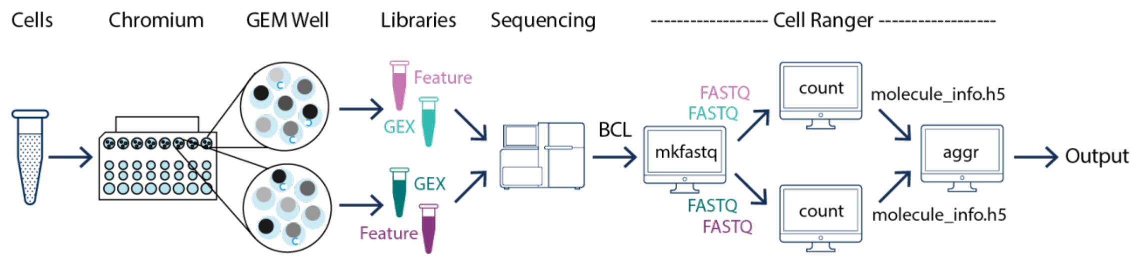


Multiple Samples, Multiple GEM Wells, One Flowcell

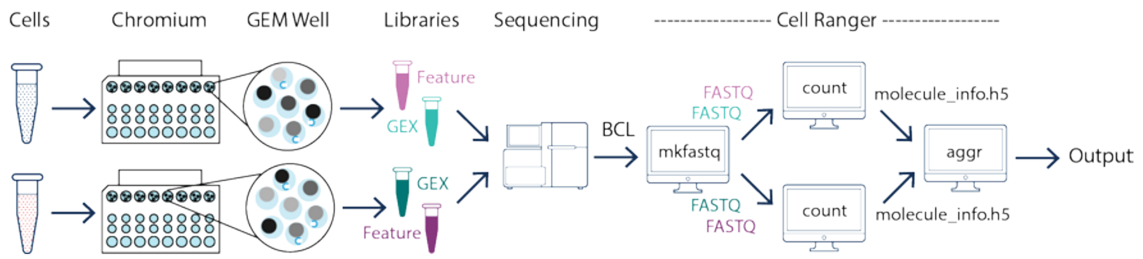


Cell Ranger™ Pipeline: Workflows – cellranger aggr

One Sample, Multiple GEM Wells, One Flowcell



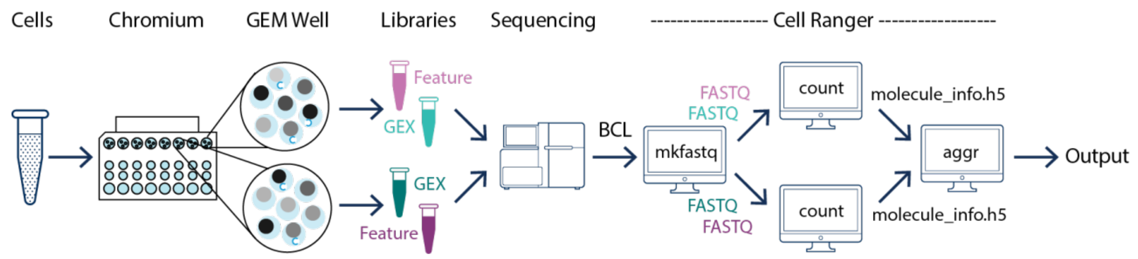
Multiple Samples, Multiple GEM Wells, One Flowcell



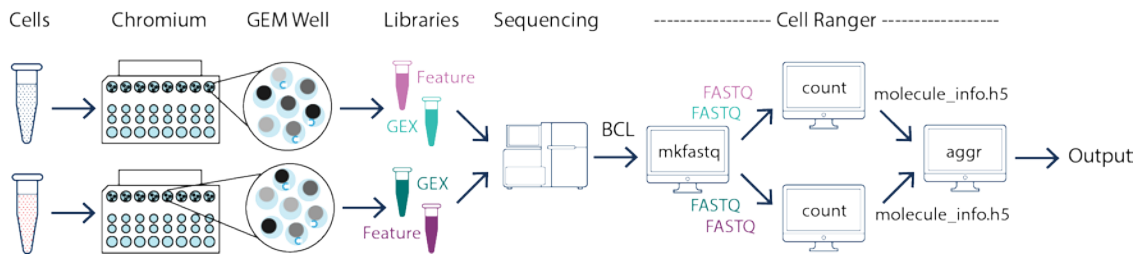
The ***cellranger aggr*** pipeline pools the results from single runs of cellranger counts, using the *molecule_info.h5* files

Cell Ranger™ Pipeline: Workflows – cellranger aggr

One Sample, Multiple GEM Wells, One Flowcell



Multiple Samples, Multiple GEM Wells, One Flowcell



The ***cellranger aggr*** pipeline pools the results from single runs of cellranger counts, using the *molecule_info.h5* files

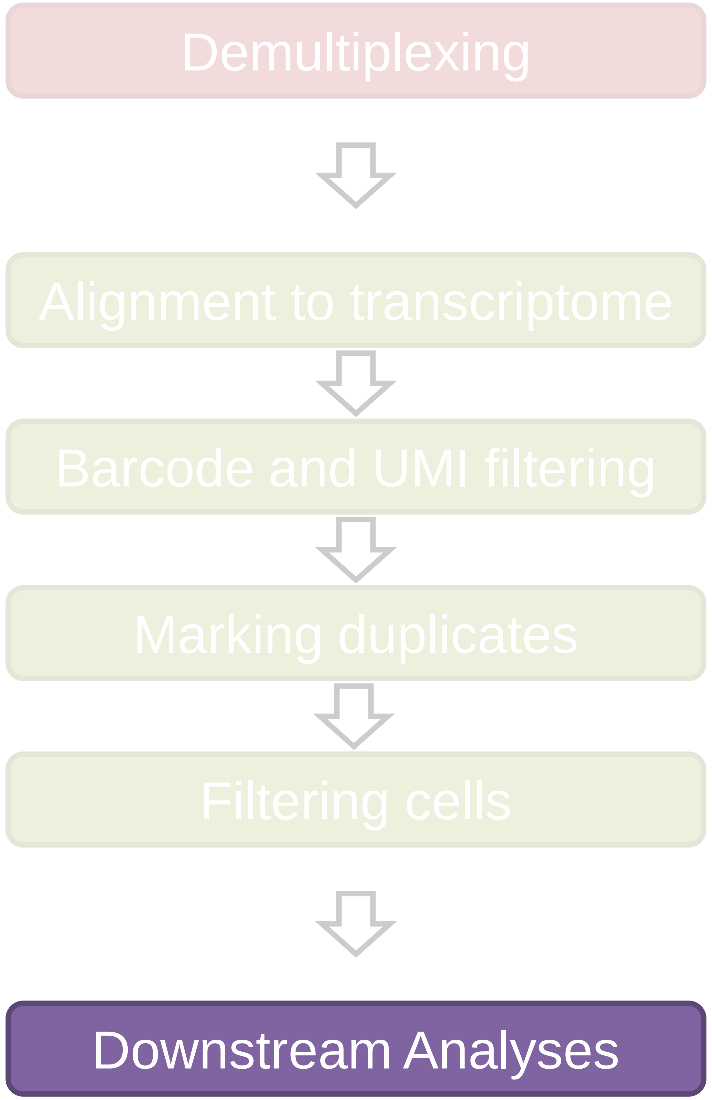
WARNING!!

By default, the reads from each GEM well are subsampled such that all GEM wells have the same effective sequencing depth, measured in terms of confidently mapped reads per cell.

Cell Ranger™ Pipeline: Downstream anaysis – cellranger reanalyze

Cellranger counts produces a .cloupe file (accessible through the Desktop app **LOUPE**) containing standard downstream analyses, run with default parameters:

- 2D projections
- Cell clustering
- Differential expression
- Interactive exploration

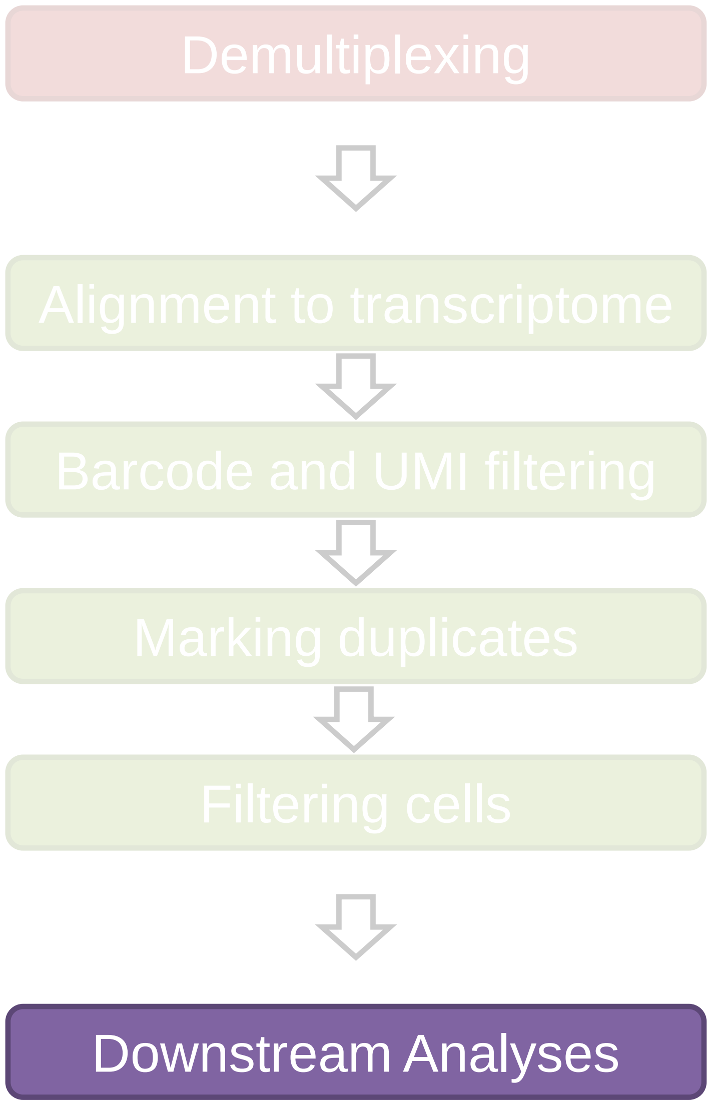


Cell Ranger™ Pipeline: Downstream anaysis – cellranger reanalyze

Cellranger counts produces a .cloupe file (accessible through the Desktop app **LOUPE**) containing standard downstream analyses, run with default parameters:

- 2D projections
- Cell clustering
- Differential expression
- Interactive exploration

The ***cellranger reanalyze*** pipeline re-runs tertiary analyses performed on the feature-barcode matrix using custom parameter settings.



10x Single Cell Softwares

https://support.10xgenomics.com/single-cell-gene-expression/software/overview/welcome

10x GENOMICS

PRODUCTS TECHNOLOGY COMPANY CAREERS SUPPORT COMMUNITY

Support › Single Cell Gene Expression › Software

SEARCH Q&A NEW CONTACT SUPPORT

SOFTWARE › OVERVIEW

- **Overview**
- System Requirements
- Downloads
- Cell Ranger: Pipelines
- Loupe Cell Browser

Overview of Single Cell Software

The Chromium Single Cell Software Suite is a set of software applications for analyzing and visualizing single cell 3' RNA-seq data produced by the 10x Chromium Platform. It is also used to analyze data generated from Feature Barcoding technology. The software suite includes:

- **Cell Ranger: Pipelines**
  A set of analysis pipelines that perform sample demultiplexing, barcode processing, and single cell 3' gene counting. Also includes the ability to process data from Feature Barcoding technology.
- **Loupe Cell Browser**
  Quickly and interactively find significant genes, cell types, and substructure within your single cell data, no programming required.

If your question is not answered here, please email us at: support@10xgenomics.com

Data pre-filtering

Data pre-filtering overview

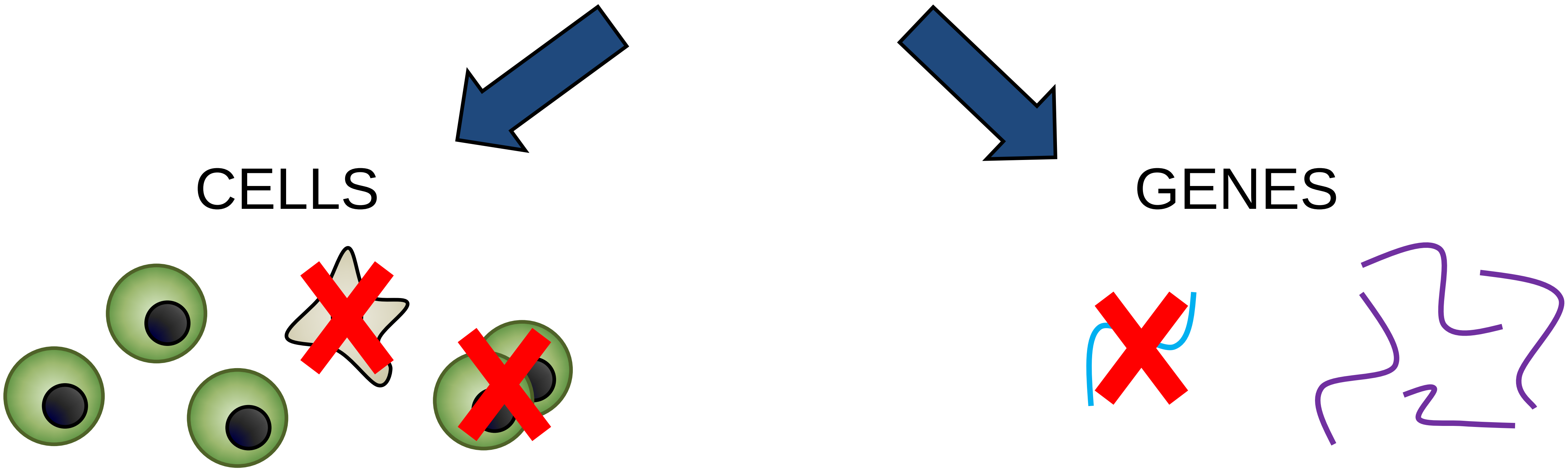
Several factors (variables) influence scRNA-seq data:

- Multiplets
- Apoptotic cells
- Drop-out effect

Data pre-filtering overview

Several factors (variables) influence scRNA-seq data:

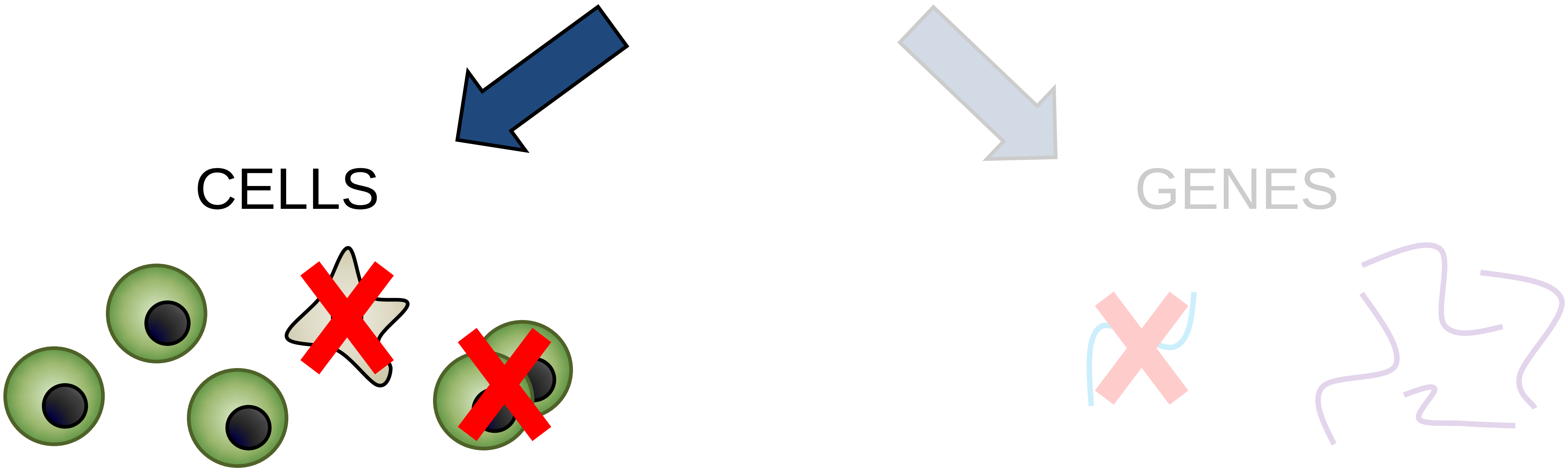
- Multiplets
- Apoptotic cells
- Drop-out effect



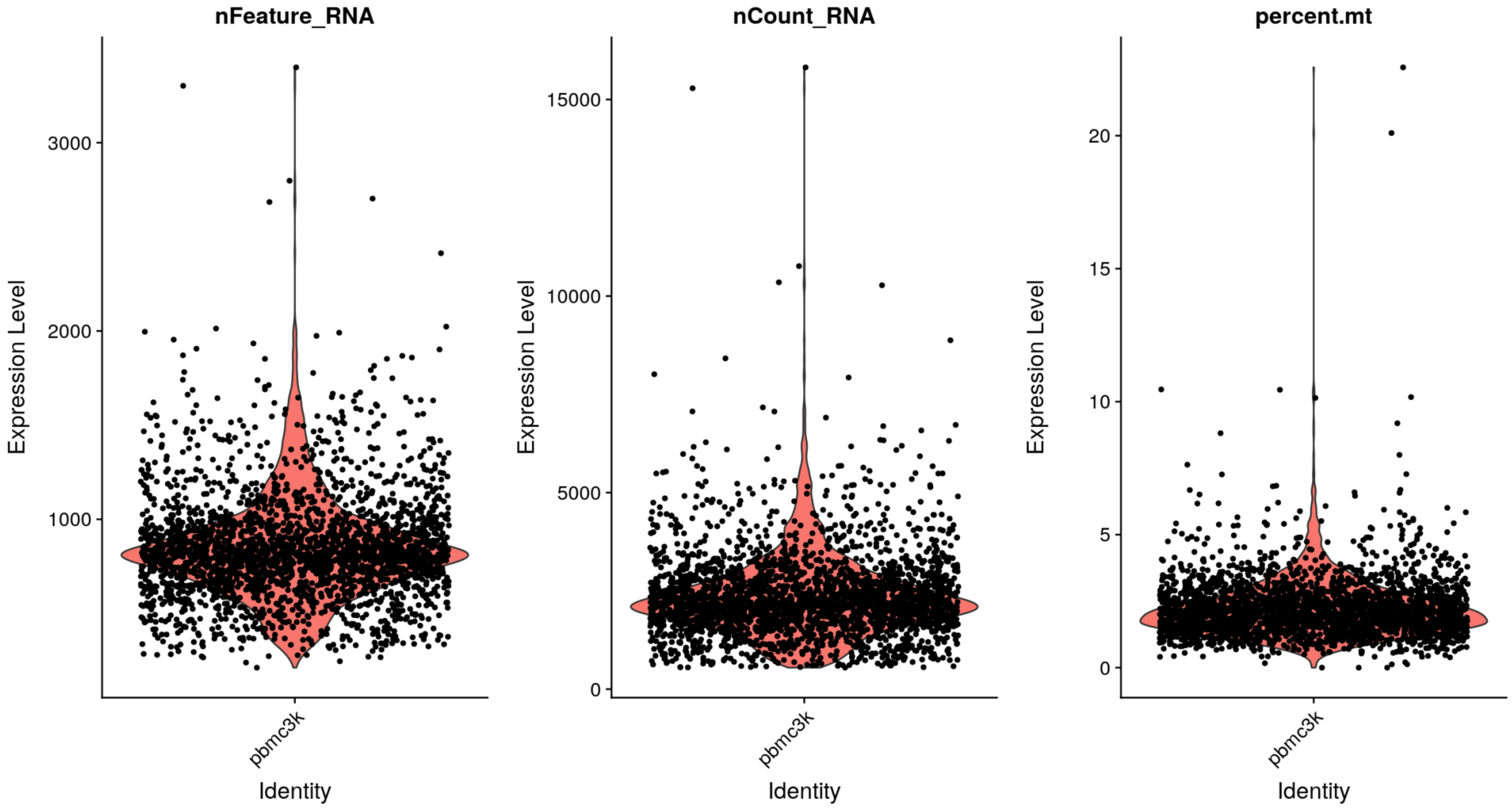
Data pre-filtering overview

Several factors (variables) influence scRNA-seq data:

- Multiplets
- Apoptotic cells
- Drop-out effect



Data pre-filtering: CELLS

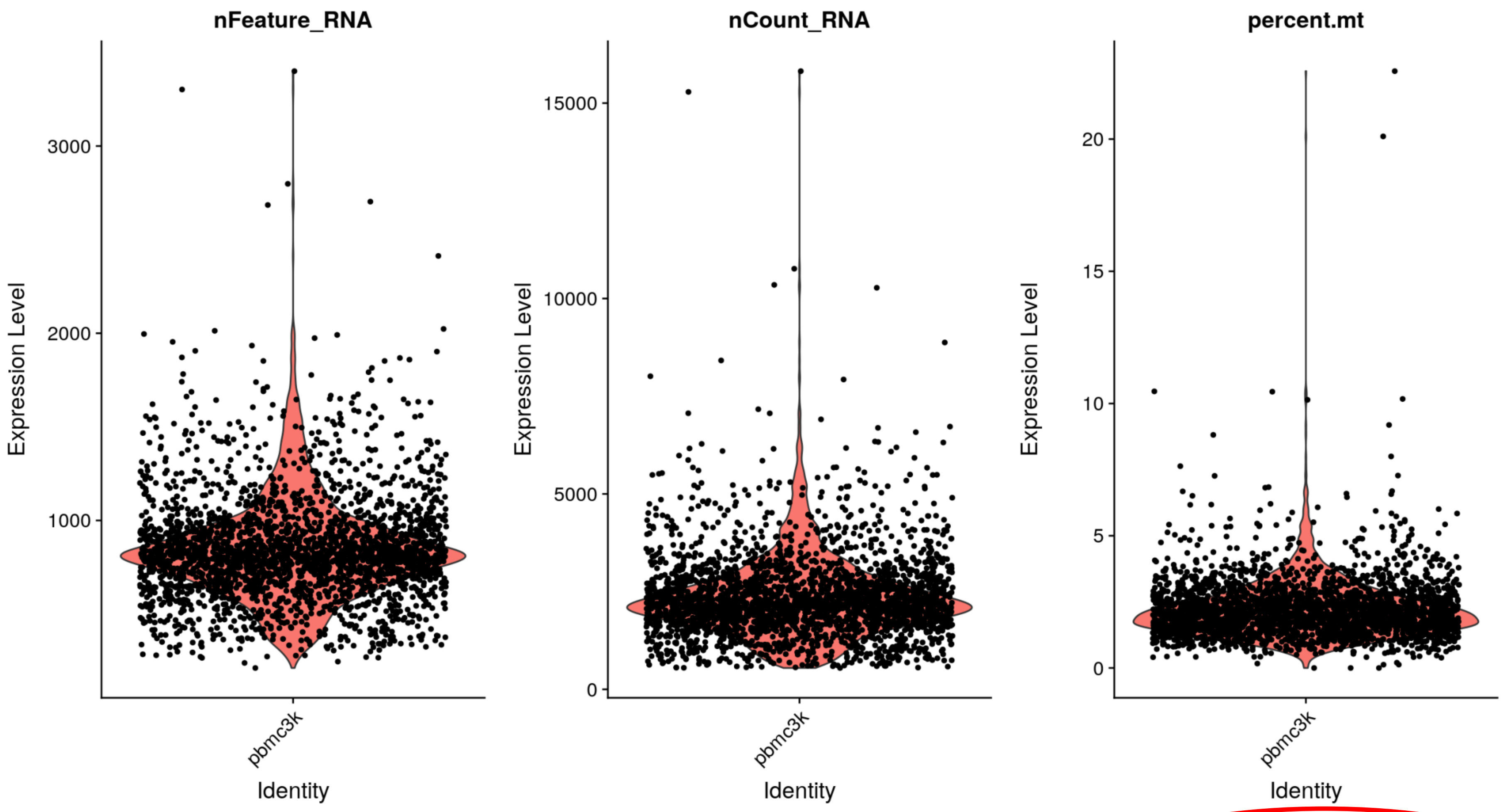


Number of detected genes

Number of total reads per cell

% of reads aligned to mitochondrial genes

Data pre-filtering: CELLS

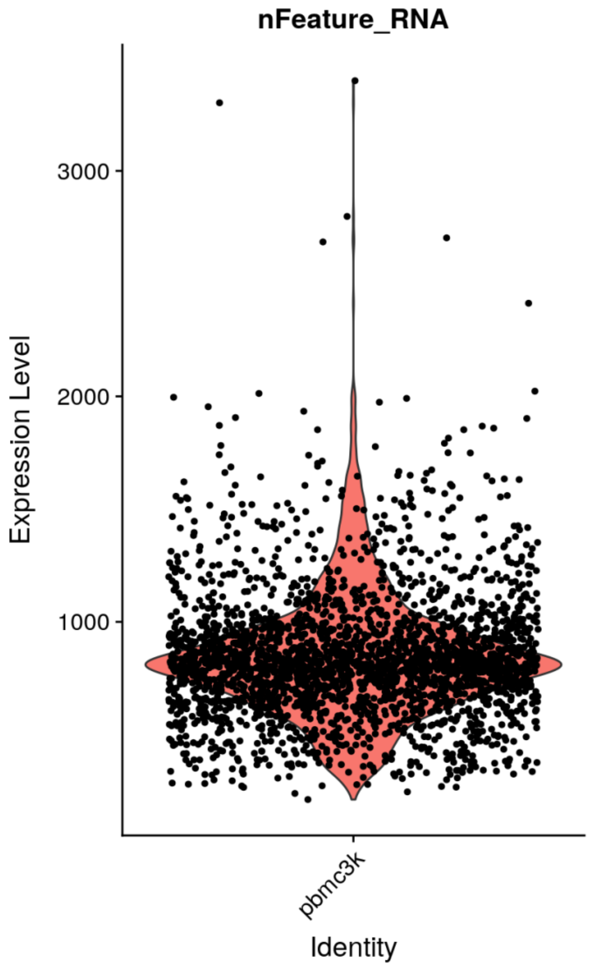


Number of detected genes

Number of total reads per cell

% of reads aligned to mitochondrial genes

Data pre-filtering: CELLS



Number of detected genes

Suggestion from several tools (Seurat, Scanpy):

- Lower limit → More than 200

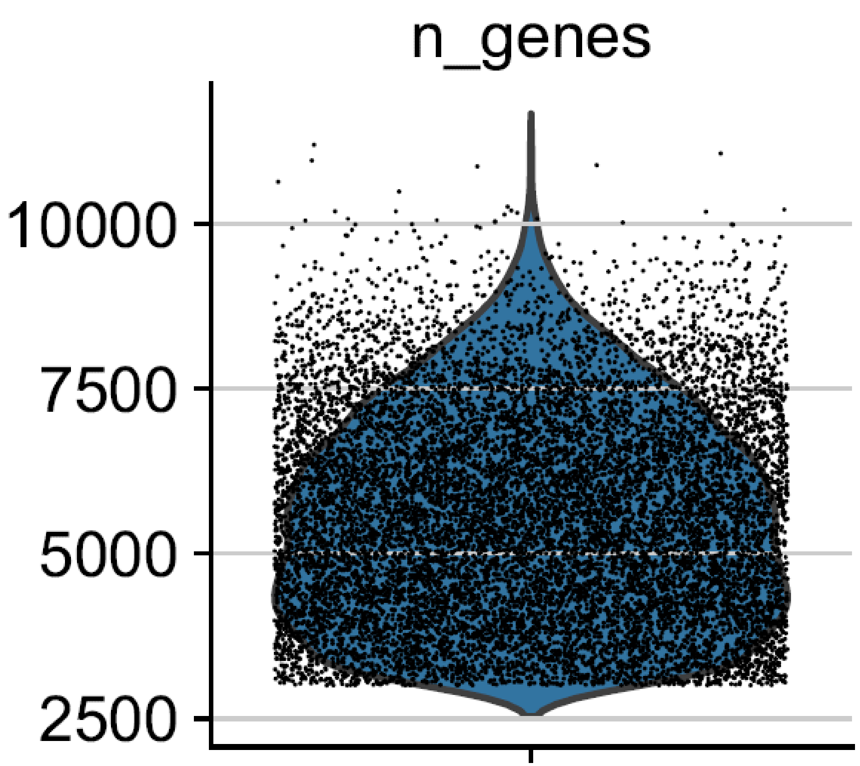
Remove cells poorly informative

- Upper limit → Less than 2,500-3,000

Remove outlier cells/multiplets (?)

- Doublet Detection
https://github.com/JonathanShor/DoubletDetection/blob/master/docs/DoubletDetection.pdf
- Mixed gene expression
- Classification of low quality cells from single-cell RNA-seq data (Ilicic et al. 2016)

Data pre-filtering: CELLS



Number of detected genes

Suggestion from several tools (Seurat, Scanpy):

- Lower limit → More than 200

Remove cells poorly informative

- Upper limit → Less than 2,500-3,000

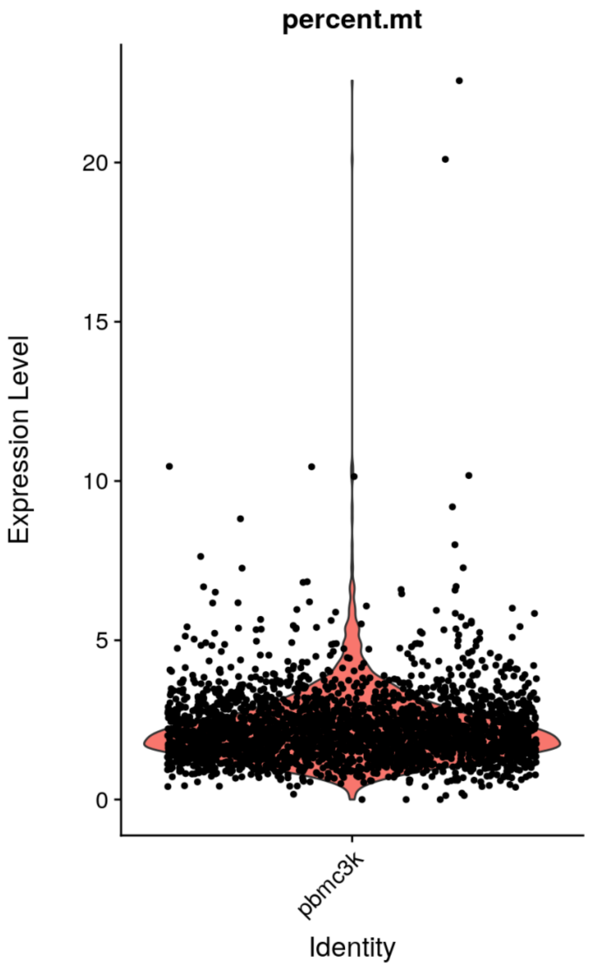
Remove outlier cells/multiplets (?)

- Doublet Detection
  https://github.com/JonathanShor/DoubletDetection/blob/master/docs/DoubletDetection.pdf
- Mixed gene expression
- Classification of low quality cells from single-cell RNA-seq data (Ilicic et al. 2016)

Data pre-filtering: CELLS

High percentage of mitochondrial gene expression may be due to:

- Apoptotic cells
- Lysed cells



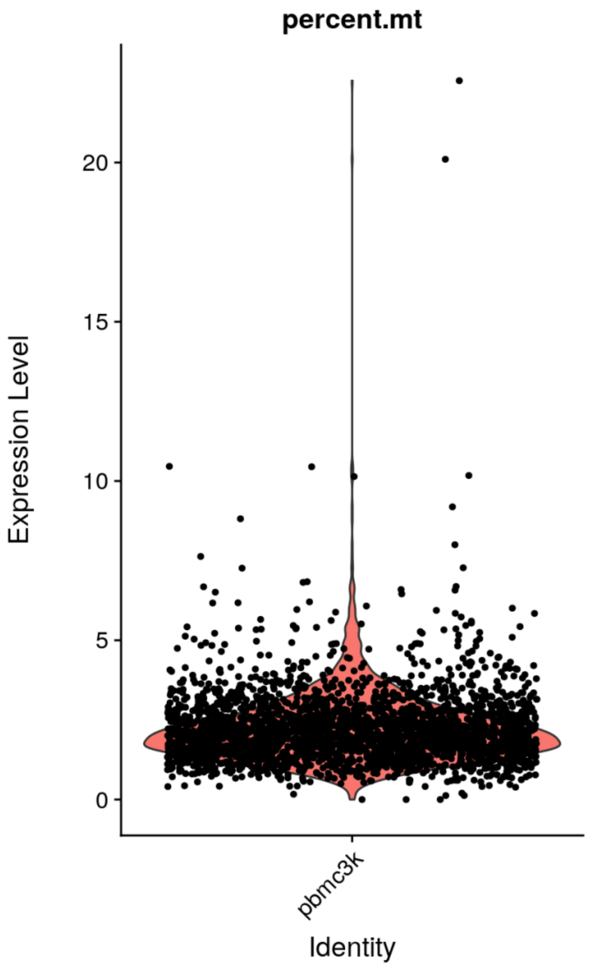
% of reads aligned to mitochondrial genes

Data pre-filtering: CELLS

High percentage of mitochondrial gene expression may be due to:

- Apoptotic cells
- Lysed cells

Cells with more than 5-7% of mtRNAs should be removed



% of reads aligned to mitochondrial genes

Aknowledgements

Organizers

Davide Cacchiarelli

Vincenza Colonna

ELIXIR-IIB Training Platform

10x Genomics

Chiara Reggio

Bashir Sadet

Carlo Erba

Stefano Tonacchera

Cacchiarelli’s Lab

Patrizia Annunziata

Valentina Bouché

Antonio Grimaldi

Anna Manfredi

Lorenzo Vaccaro

Bioinformaticians

Annamaria Carissimo

Gennaro Gambardella