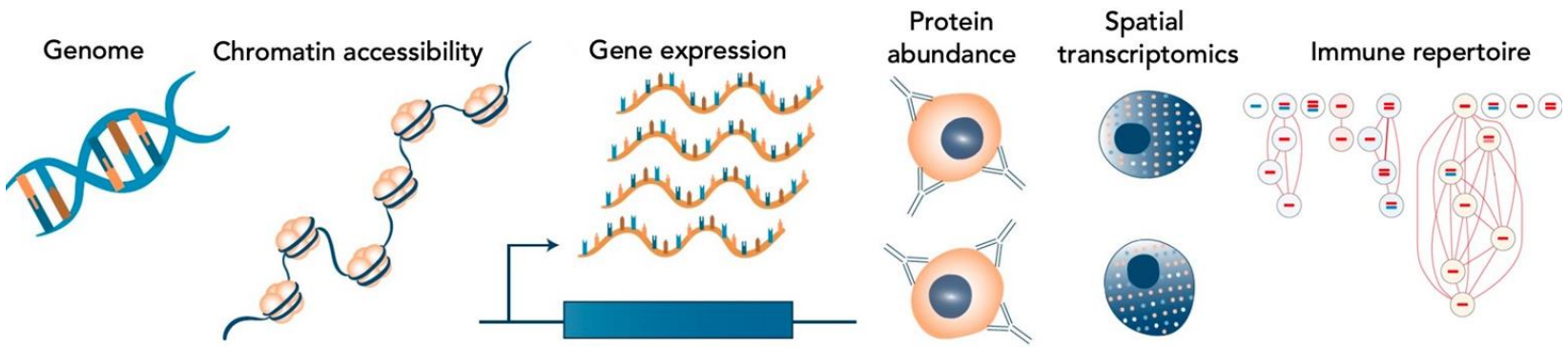
# Workshop Day 1

Bioinformatic tools (overview)

# Background Information

# Evolving software for single-cell levels




**DNA and epigenome**
- Single cell genomes (WES, WGS)
- Single cell epigenomics (HiC, ChIP, ATAC, mC )

**RNA**
- Full length (mRNA, total RNA)
- 5′ and 3′ end counting

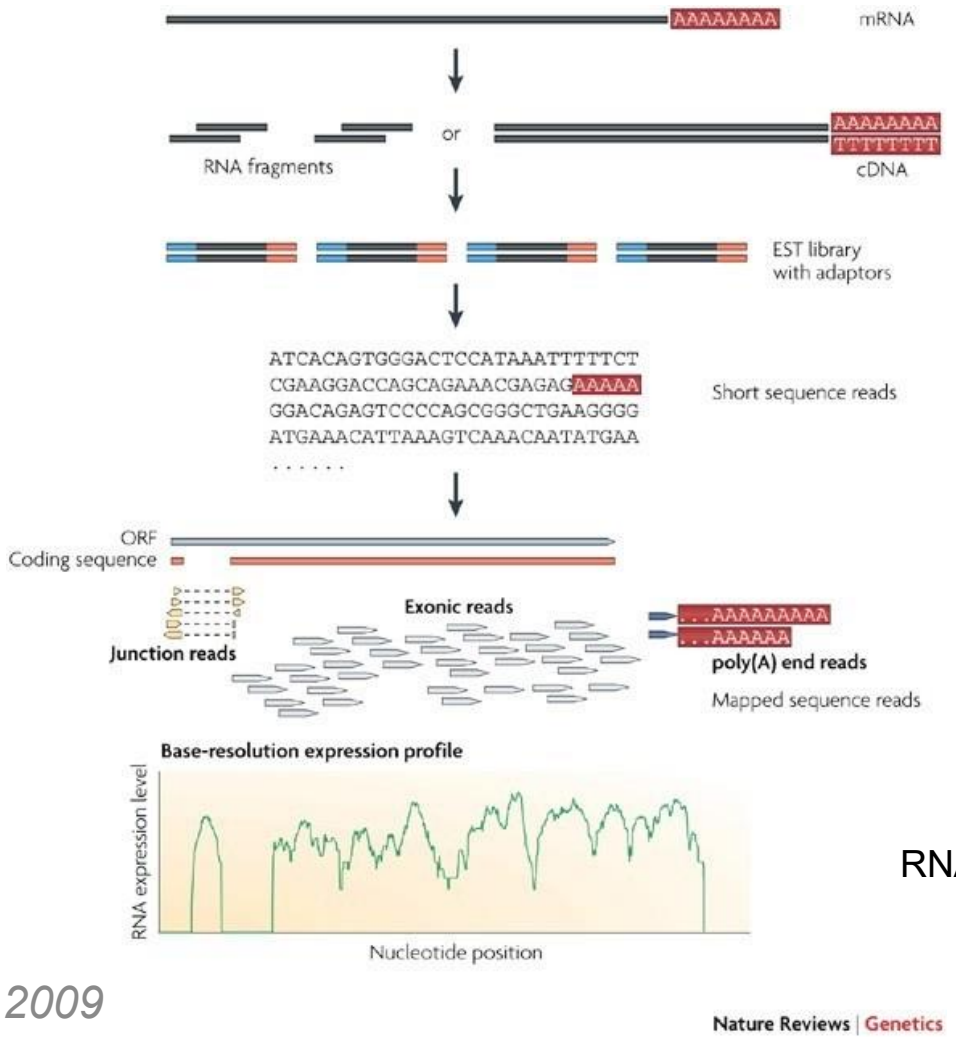
**Proteins**
- Multi-parameter flow
- Mass cytometry
- Single cell proteomics

**'multi-omics'**
- DNA+RNA (G+T)
- RNA+protein (T+P)
- Epigenome + RNA

*Efremova et al. Nat Methods 2020*

# Choose based on...

Throughput: # of cells/reaction can profile per run, balancing cost, workflow complexity.

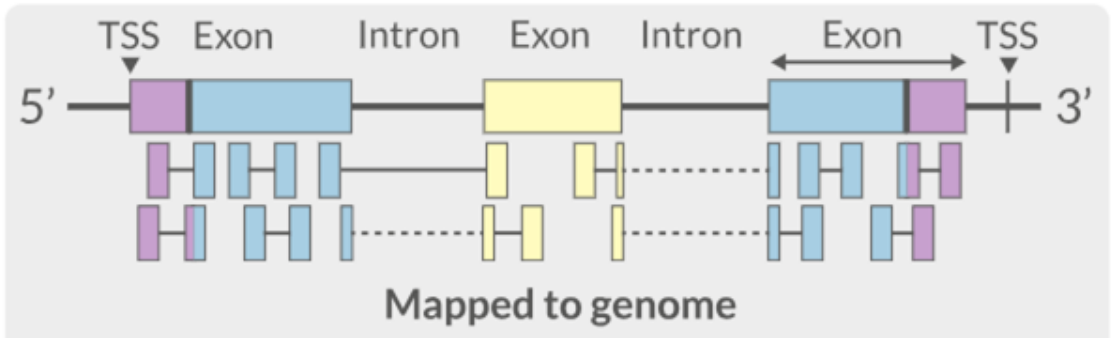


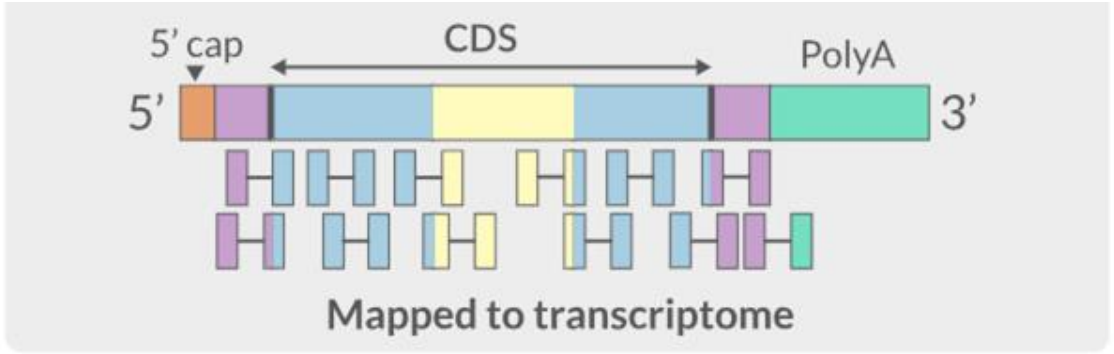

RNA-seq workflow (bulk)

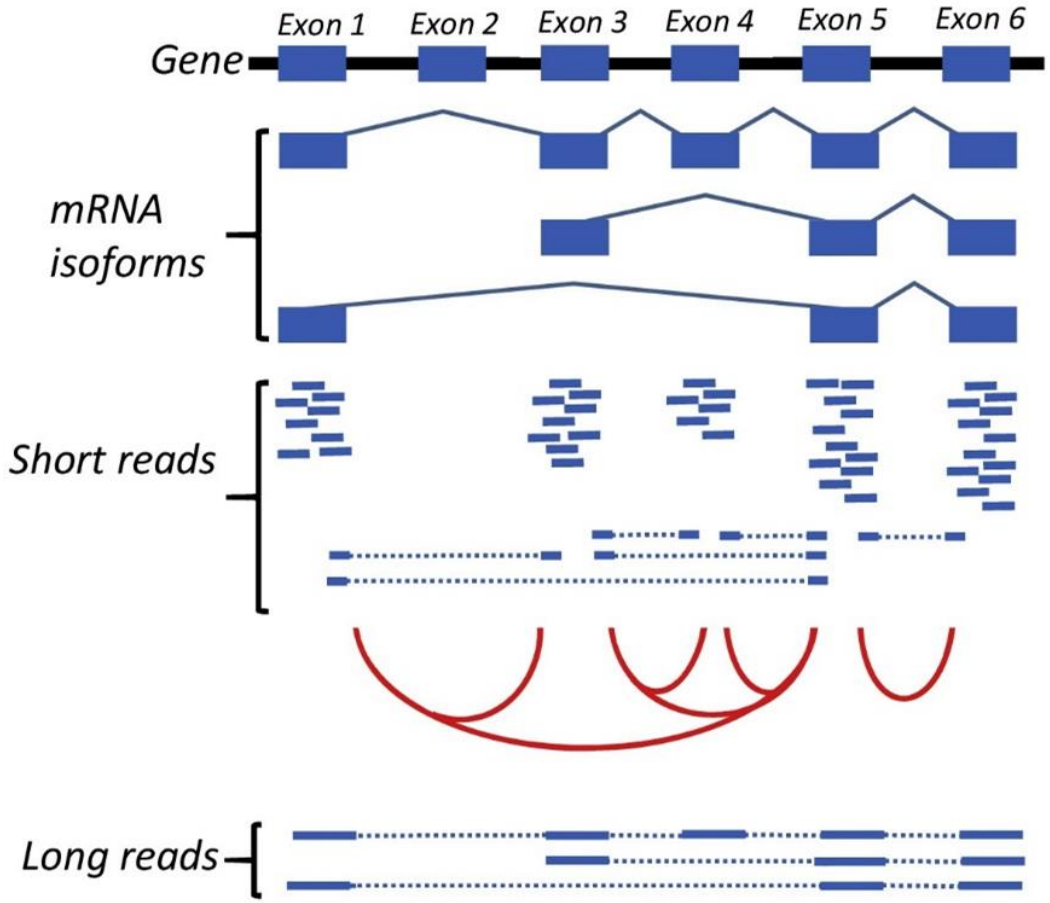
*Wang et al., Nat. Rev. Genet. 2009*

# Choose based on…

Coverage: 3’biased (Good for gene-level quantification). And, Full-length (providing better resolution for isoforms and splicing).

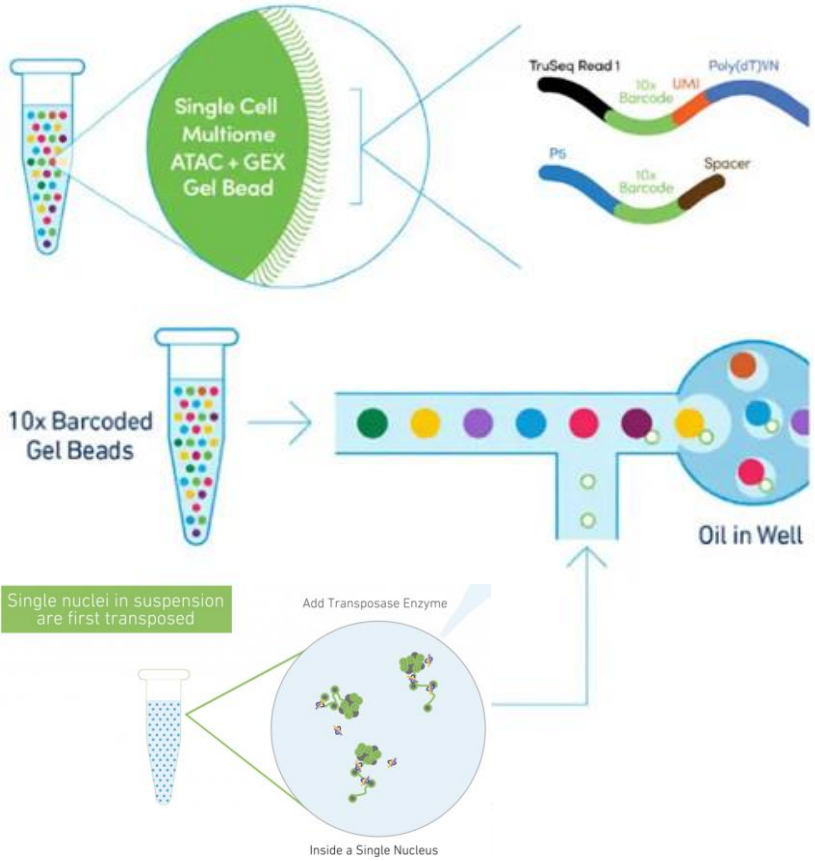


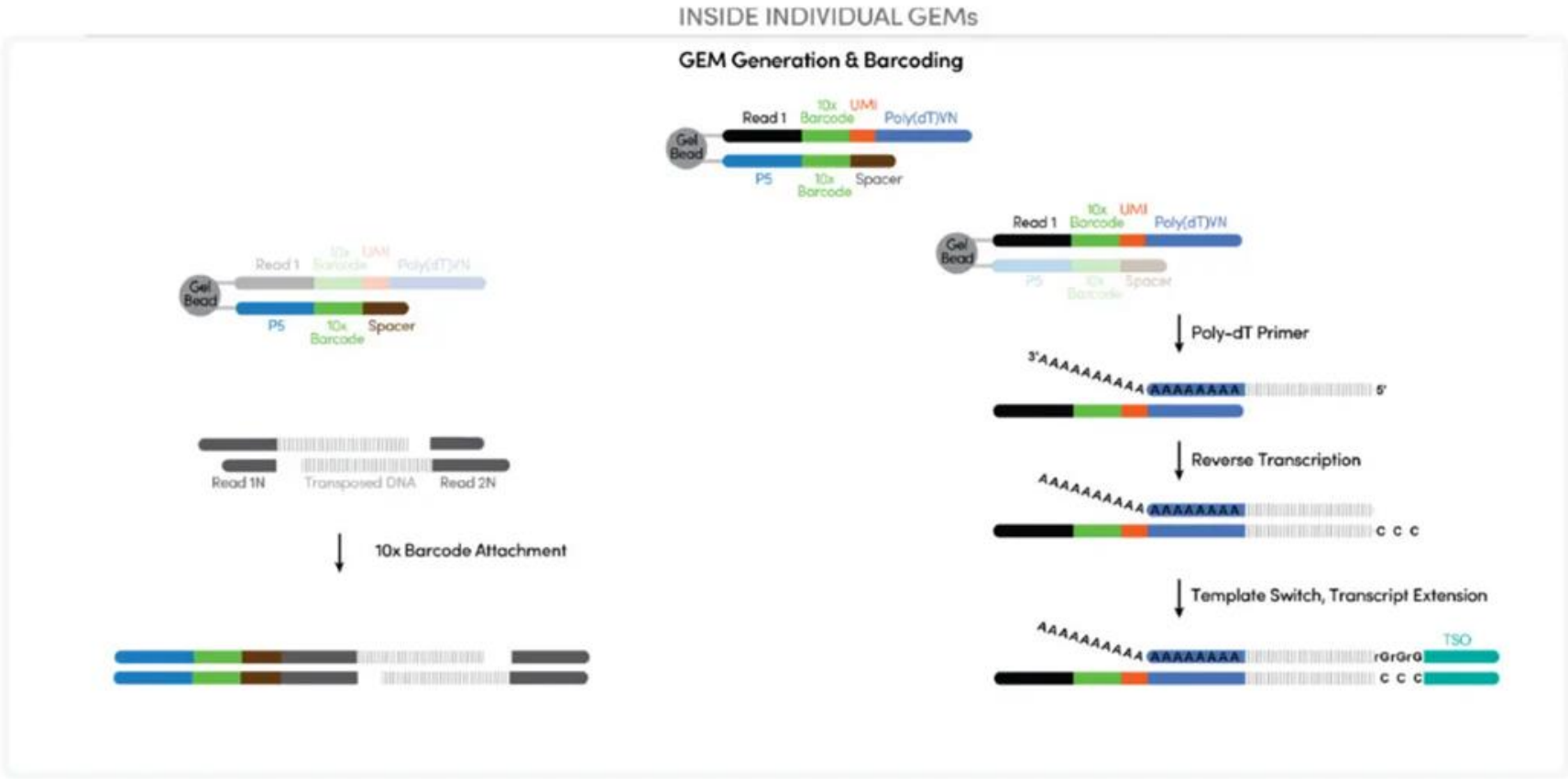



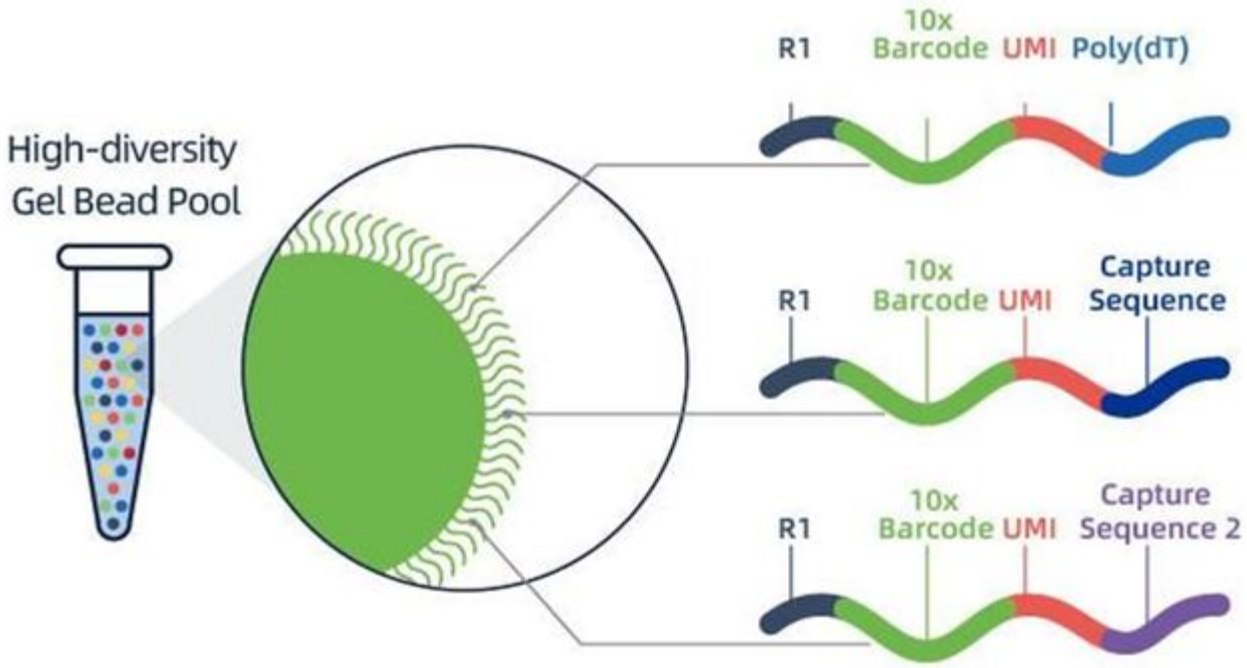
# Choose based on…

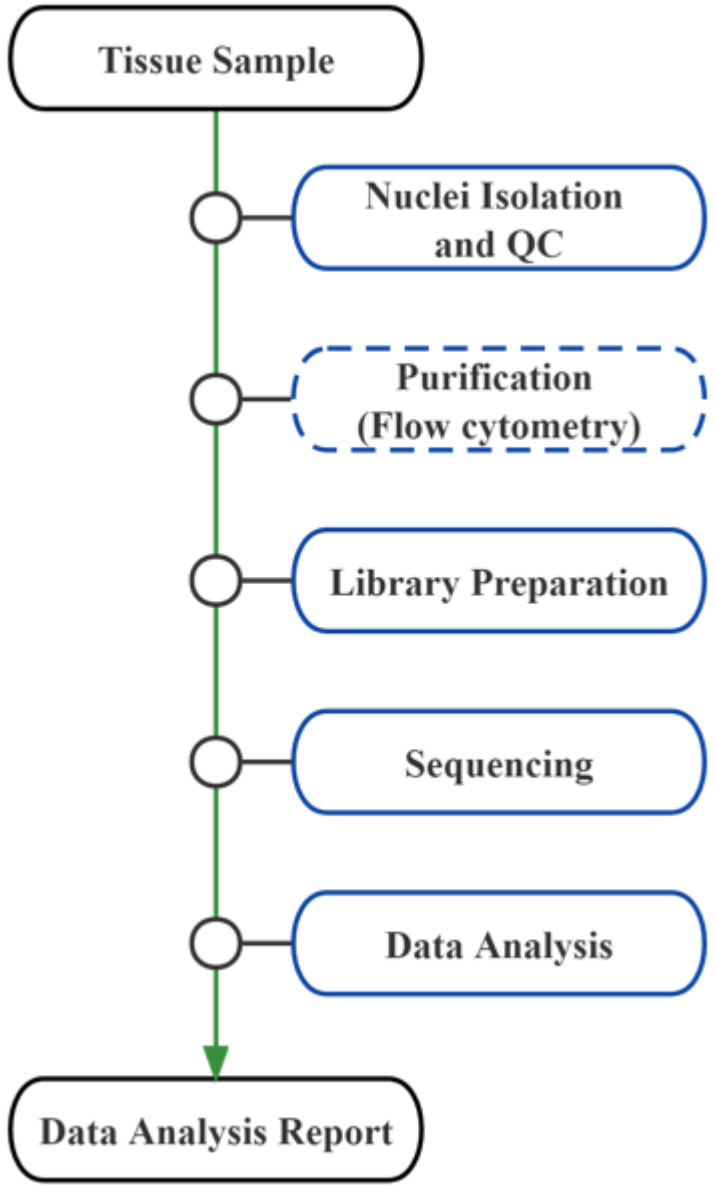
Sample of origin: With fragile tissue, single-nuclei approaches.



# Typical workflow




Sequencing using Illumina paired-end 150 bp (PE150)

*10X genomics*

# Pre-processing

# Flowchart: to count matrix

| | |
|---|---|
| Get bcl file | That is the role of the sequencing machine! |
| Create fastq files | **fastq files –** R1, R2 and I files showing raw reads. |
| QC: assess overall quality | Sequencing |

Cellranger

- QC
- Alignment
- Count matrix

# .bcl files

- **Raw data** output of a sequencing run
- Binary, non-human-readable file

- Contains the **base calling** and **quality score per sequencing lane**

# Flowchart: to count matrix

- Get bcl file — That is the role of the sequencing machine!
- Create fastq files — **fastq files –** R1, R2 and I files showing raw reads.
- QC: assess overall quality — Sequencing

Cellranger

- QC
- Alignment
- Count matrix

# Follow the workflow directly from 10x with our inputs (fastq.gz)

**workflow reference:** 10x GENOMICS™

**Requirements:**

- Fastq.gz
- Genome
- GTF

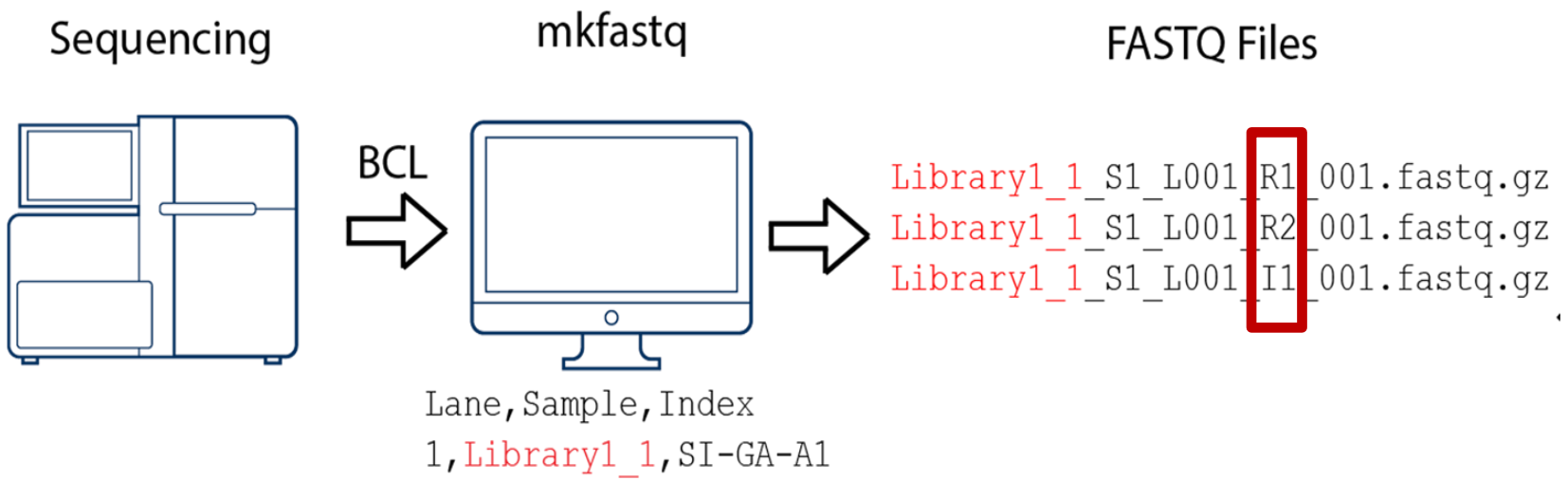
BASH
THE BOURNE-AGAIN SHELL

# Cellranger-arc

# bcl2fastq

Command:

**bcl2fastq --run-folder-dir <bcl_files_folder> -p 12 --output-dir <fastq_files_folder>**



File Requirements:

- raw data fastq files

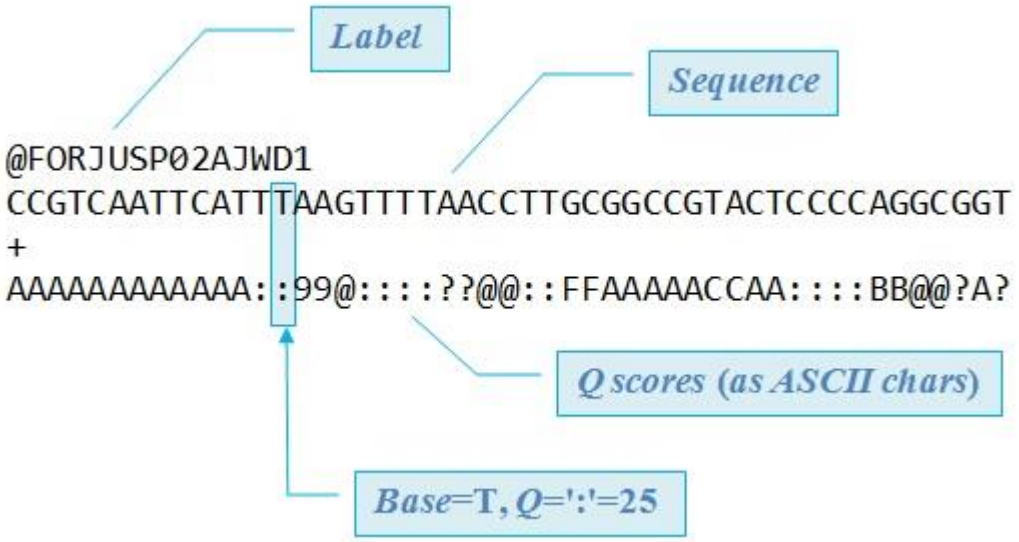
| Name | Size | File Ty... |
|---|---|---|
| scRNA_E5_PBS-SI_TT_D10_235GF3LT3_S3_L001_R2_001.fastq.gz | 18.46 GB | GZ File |
| scRNA_E5_PBS-SI_TT_D10_235GF3LT3_S3_L001_R1_001.fastq.gz | 24.47 GB | GZ File |
| scRNA_E5_PBS-SI_TT_D10_235GF3LT3_S3_L001_I2_001.fastq.gz | 1.71 GB | GZ File |
| scRNA_E5_PBS-SI_TT_D10_235GF3LT3_S3_L001_I1_001.fastq.gz | 1.82 GB | GZ File |
| MD5.txt | 364 Bytes | TXT File |

| Name | Size | File Ty... |
|---|---|---|
| scATAC_E5_PBS-SI_3A_G11_235G7CLT3_S3_L001_R3_001.fastq.gz | 16.74 GB | GZ File |
| scATAC_E5_PBS-SI_3A_G11_235G7CLT3_S3_L001_R2_001.fastq.gz | 4.22 GB | GZ File |
| scATAC_E5_PBS-SI_3A_G11_235G7CLT3_S3_L001_R1_001.fastq.gz | 17.09 GB | GZ File |
| scATAC_E5_PBS-SI_3A_G11_235G7CLT3_S3_L001_I1_001.fastq.gz | 1.7 GB | GZ File |
| MD5.txt | 368 Bytes | TXT File |

# How is a .fastq organized?

Each fastq file contains reads, each read is composed of 4 lines:

1. A sequence identifier with information about the sequencing run
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores, using ASCII characters to represent the numerical quality scores.



# Why do we end up with so many fastq files?

We sequence the paired ends of each DNA fragment molecule, in 3|4 different sequencing “runs”


I1.fastq contains sample index
R1.fastq contains cell barcode + UMI
R2.fatsq contains transcript information

# Flowchart: to count matrix

- Get bcl file — That is the role of the sequencing machine!
- Create fastq files — **fastq files –** R1, R2 and I files showing raw reads.
- QC: assess overall quality — Sequencing

Cellranger

- QC
- Alignment
- Count matrix

## scATAC

| Sample | Library-index | Yield (G) | Total yield (G) | Q30 % |
|---|---|---|---|---|
| scATAC_E4_PBS | scATAC_E4_PBS-SI_3A_E9_232FTKLT3_S1_L003_S1 | 23.3 | 93.7 | 89.10 |
| scATAC_E4_PBS | scATAC_E4_PBS-SI_3A_E9_232FTKLT3_S1_L003_S2 | 27.1 | | 87.49 |
| scATAC_E4_PBS | scATAC_E4_PBS-SI_3A_E9_232FTKLT3_S1_L003_S3 | 23.4 | | 87.32 |
| scATAC_E4_PBS | scATAC_E4_PBS-SI_3A_E9_232FTKLT3_S1_L003_S4 | 19.9 | | 89.78 |
| scATAC_E5_FGF | scATAC_E5_FGF-SI_3A_E12_232FTKLT3_S4_L003_S13 | 22.0 | 159.4 | 89.25 |
| scATAC_E5_FGF | scATAC_E5_FGF-SI_3A_E12_232FTKLT3_S4_L003_S14 | 23.5 | | 89.32 |
| scATAC_E5_FGF | scATAC_E5_FGF-SI_3A_E12_232FTKLT3_S4_L003_S15 | 19.8 | | 89.64 |
| scATAC_E5_FGF | scATAC_E5_FGF-SI_3A_E12_232FTKLT3_S4_L003_S16 | 19.6 | | 89.14 |
| scATAC_E5_FGF | scATAC_E5_FGF-SI_3A_E8_232FTKLT3_S5_L003_S17 | 19.6 | | 89.75 |
| scATAC_E5_FGF | scATAC_E5_FGF-SI_3A_E8_232FTKLT3_S5_L003_S18 | 17.1 | | 90.09 |
| scATAC_E5_FGF | scATAC_E5_FGF-SI_3A_E8_232FTKLT3_S5_L003_S19 | 19.2 | | 87.83 |
| scATAC_E5_FGF | scATAC_E5_FGF-SI_3A_E8_232FTKLT3_S5_L003_S20 | 18.6 | | 88.34 |
| scATAC_E4_FGF | scATAC_E4_FGF-SI_3A_E10_232FTKLT3_S2_L003_S5 | 24.5 | 106.0 | 87.89 |
| scATAC_E4_FGF | scATAC_E4_FGF-SI_3A_E10_232FTKLT3_S2_L003_S6 | 27.0 | | 88.21 |
| scATAC_E4_FGF | scATAC_E4_FGF-SI_3A_E10_232FTKLT3_S2_L003_S7 | 30.3 | | 88.96 |
| scATAC_E4_FGF | scATAC_E4_FGF-SI_3A_E10_232FTKLT3_S2_L003_S8 | 24.2 | | 89.77 |
| scATAC_E5_PBS | scATAC_E5_PBS-SI_3A_E11_232FTKLT3_S3_L003_S9 | 24.1 | 87.4 | 87.64 |
| scATAC_E5_PBS | scATAC_E5_PBS-SI_3A_E11_232FTKLT3_S3_L003_S10 | 17.9 | | 89.12 |
| scATAC_E5_PBS | scATAC_E5_PBS-SI_3A_E11_232FTKLT3_S3_L003_S11 | 19.6 | | 88.02 |
| scATAC_E5_PBS | scATAC_E5_PBS-SI_3A_E11_232FTKLT3_S3_L003_S12 | 25.8 | | 88.77 |

## GEXs

| | Effective(%) | Error(%) | Q30(%) |
|---|---|---|---|
| scRNA_E4_PBS | 100.00 | 0.08 | 95.07 |
| scRNA_E4_FGF | 100.00 | 0.08 | 95.30 |
| scRNA_E5_PBS | 100.00 | 0.08 | 95.15 |
| scRNA_E5_FGF | 100.00 | 0.09 | 95.70 |

**Quality control summary**
Sequencing yield (G) and read quality (Q30) were strong, supporting reliable downstream analysis.

# Flowchart: to count matrix

- Get bcl file — That is the role of the sequencing machine!
- Create fastq files — **fastq files –** R1, R2 and I files showing raw reads.
- QC: assess overall quality — Sequencing

Cellranger

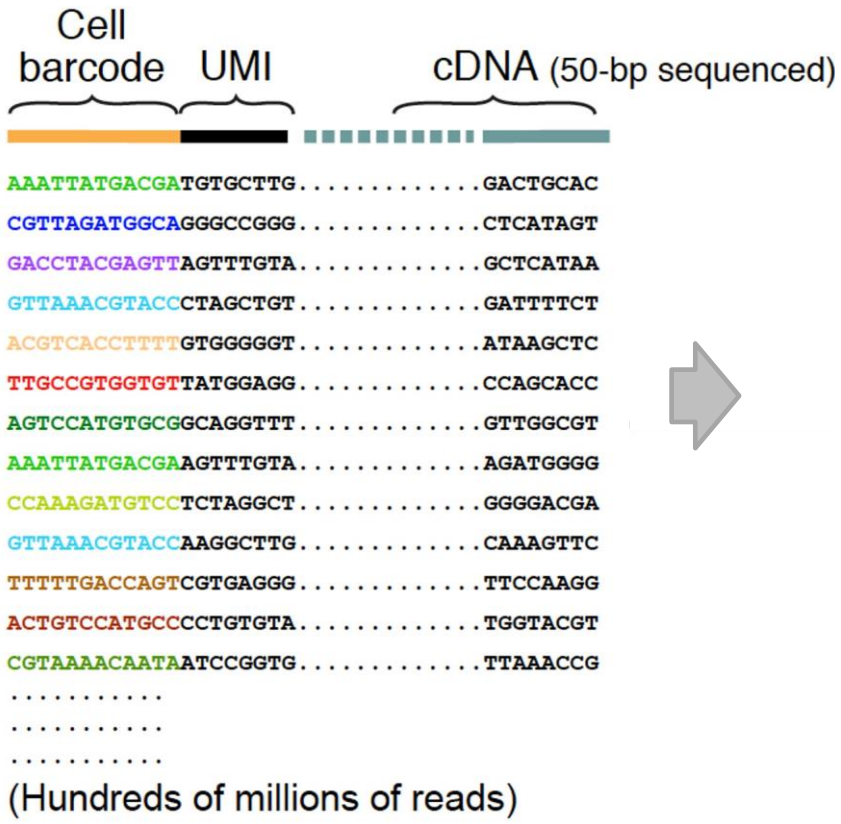
- QC
- Alignment
- Count matrix

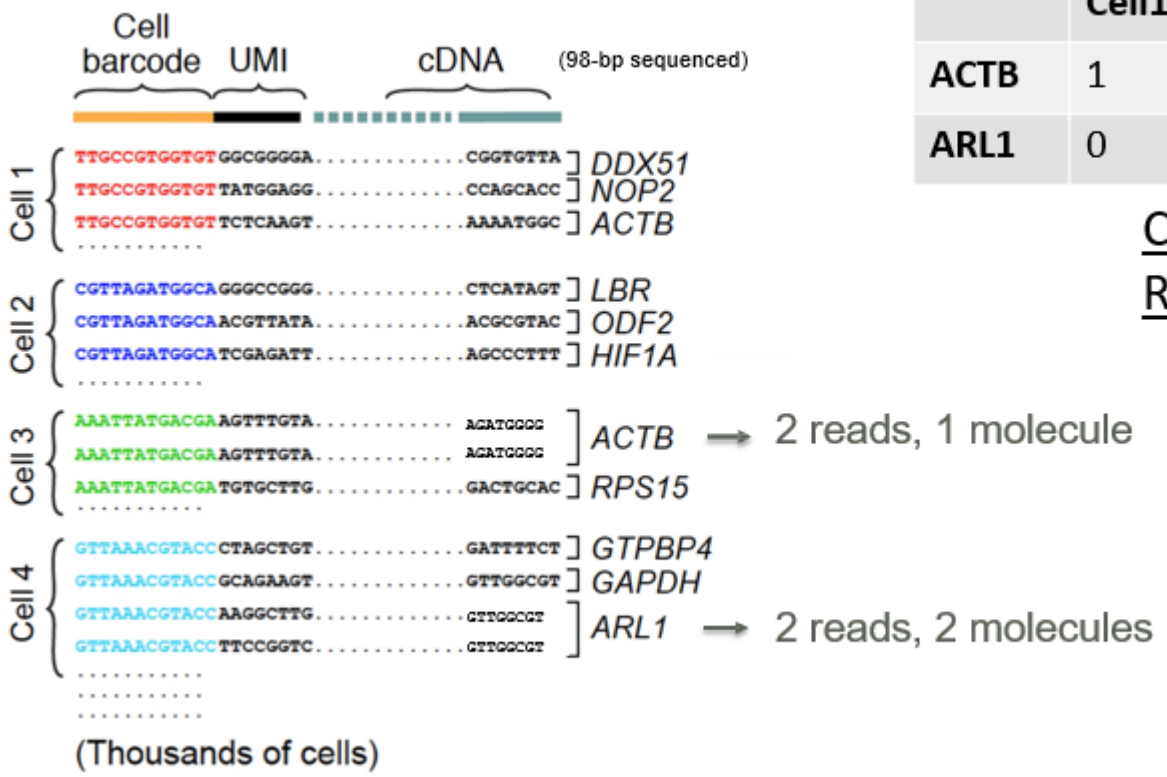
# Which reads are considered for UMI counting by Cell Ranger?

1. Only reads with a valid UMI and a valid 10x barcode.
2. No bases with base quality < 10 (Q10)
1. Read maps to exactly one gene.
2. Overlaps an exon by at least 50% in a way consistent with annotated splice junctions and strand annotation.
3. Multiple reads that map to the same UMI will only count once.

# Aligning reads to a transcriptome reference

1. Group reads by cell-of-origin (using the cell barcodes)
2. **Recover which transcript the cDNA sequence aligns to**



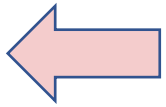




| | Cell1 | Cell2 | Cell3 | Cell4 |
|---|---|---|---|---|
| **ACTB** | 1 | 0 | **1** | 0 |
| **ARL1** | 0 | 0 | 0 | **2** |

Columns: cells
Rows: features


# Cellranger-arc

To make the **main outputs:**



- h5
- fragment.tsv
- fragment.tsv.tbi

| Name | Size | File Type |
|---|---|---|
| web_summary.html | 8 MB | HTML File |
| summary.csv | 1.8 KB | CSV File |
| raw_feature_bc_matrix.h5 | 130.9 MB | H5 File |
| per_barcode_metrics.csv | 78.3 MB | CSV File |
| gex_possorted_bam.bam.bai | 8.2 MB | BAI File |
| gex_possorted_bam.bam | 19.64 GB | BAM File |
| gex_molecule_info.h5 | 307 MB | H5 File |
| filtered_feature_bc_matrix.h5 | 109.2 MB | H5 File |
| cloupe.cloupe | 731.8 MB | CLOUPE File |
| atac_possorted_bam.bam.bai | 3.5 MB | BAI File |
| atac_possorted_bam.bam | 46.61 GB | BAM File |
| atac_peaks.bed | 1.4 MB | BED File |
| atac_peak_annotation.tsv | 3.8 MB | TSV File |
| atac_fragments.tsv.gz.tbi | 568.2 KB | TBI File |
| atac_fragments.tsv.gz | 1.77 GB | GZ File |
| atac_cut_sites.bigwig | 1.35 GB | BIGWIG File |
| raw_feature_bc_matrix | | Folder |
| filtered_feature_bc_matrix | | Folder |
| analysis | | Folder |

# Cellranger output

## Multiomix_E4_Dev - Multiomix_E4_Dev

### Alerts

The analysis detected ⚠ 2 warnings.

| | Alert | Value | Detail |
|---|---|---|---|
| ⚠ | GEX Reads mapping to reference is low | 79.2% | Ideal > 80%. This can be caused by the wrong reference genome being used or a poor quality genome assembly. Application performance may be affected. |
| ⚠ | ATAC TSS enrichment is low | 3.82 | Ideal > 5. A low TSS score can be caused by poor sample prep, poor sample quality, a population of cells with highly accessible DNA (e.g., activated granulocytes, dead or dying cells, unsupported organisms) or poor quality reference genome annotation. |

7,319
Estimated number of cells

11,576
ATAC Median high-quality fragments per cell

2,165
GEX Median genes per cell

## Multiomix_E5_Dev - Multiomix_E5_Dev

### Alerts

The analysis detected ⚠ 1 warning.

| | Alert | Value | Detail |
|---|---|---|---|
| ⚠ | ATAC TSS enrichment is low | 4.45 | Ideal > 5. A low TSS score can be caused by poor sample prep, poor sample quality, a population of cells with highly accessible DNA (e.g., activated granulocytes, dead or dying cells, unsupported organisms) or poor quality reference genome annotation. |

6,436
Estimated number of cells

12,110
ATAC Median high-quality fragments per cell

2,086
GEX Median genes per cell

# scATAC

## Sequencing

| | |
|---|---|
| Sequenced read pairs | 230,015,294 |
| Valid barcodes | 98.5% |

>85%

## Cells

| | |
|---|---|
| Estimated number of cells | 11,776 |
| Mean raw read pairs per cell | 19,532.55 |
| Fraction of high-quality fragments in cells | 81.1% |

>40%

## Targeting

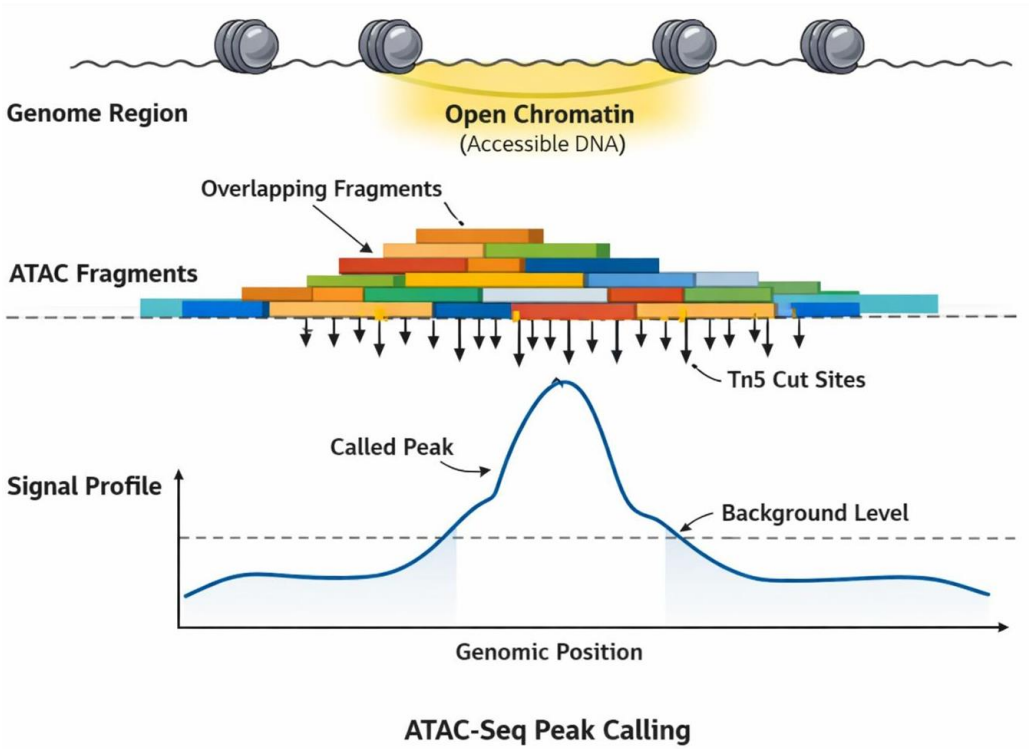
| | |
|---|---|
| Number of peaks | 96,302 |
| Fraction of genome in peaks | 7.9% |
| TSS enrichment score | 5.02 |

>5

## Mapping

| | |
|---|---|
| Confidently mapped read pairs | 92.1% |

>80%




ATAC-Seq Peak Calling

## scATAC

### Sequencing

| | | |
|---|---|---|
| Sequenced read pairs | 230,015,294 | |
| Valid barcodes | 98.5% | >85% |

### Cells

| | | |
|---|---|---|
| Estimated number of cells | 11,776 | |
| Mean raw read pairs per cell | 19,532.55 | |
| Fraction of high-quality fragments in cells | 81.1% | >40% |

### Targeting

| | | |
|---|---|---|
| Number of peaks | 96,302 | |
| Fraction of genome in peaks | 7.9% | |
| TSS enrichment score | 5.02 | >5 |

### Mapping

| | | |
|---|---|---|
| Confidently mapped read pairs | 92.1% | >80% |

## GEXs

### Sequencing

| | | |
|---|---|---|
| Sequenced read pairs | 213,585,631 | |
| Valid barcodes | 90.9% | >80% |

### Cells

| | | |
|---|---|---|
| Estimated number of cells | 5,776 | |
| Mean raw reads per cell | 36,978.12 | |
| Fraction of transcriptomic reads in cells | 60.7% | >60% |

### Mapping

| | | |
|---|---|---|
| Reads mapped to genome | 84.9% | |
| Reads mapped confidently to genome | 84.0% | >80% |