Single-cell Data Analysis Workshop

Lecture 1

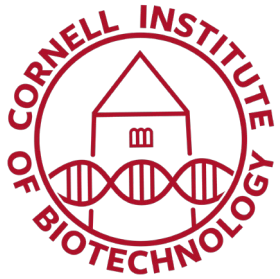
Single-cell profiling technologies

10x Genomics overview

Data processing and QC with ‘cellranger count’

Jen Grenier, PhD BRC Genomics Facility Director

Qi Sun, PhD BRC Bioinformatics Facility Co-Director



Workshop Logistics

February 12 – March 10, 2024 **All meetings on Zoom**

Mondays	3-4 pm	Lecture
	4-5 pm	Hands-on BioHPC workshop†*
Wednesdays	1-2 pm	Discussion and office hours*
Thursdays	3-4 pm	Discussion and office hours*

† *computer assignments: https://biohpc.cornell.edu/ww/machines.aspx?i=165*

** Zoom break out rooms will be used to cover different needs and topics*

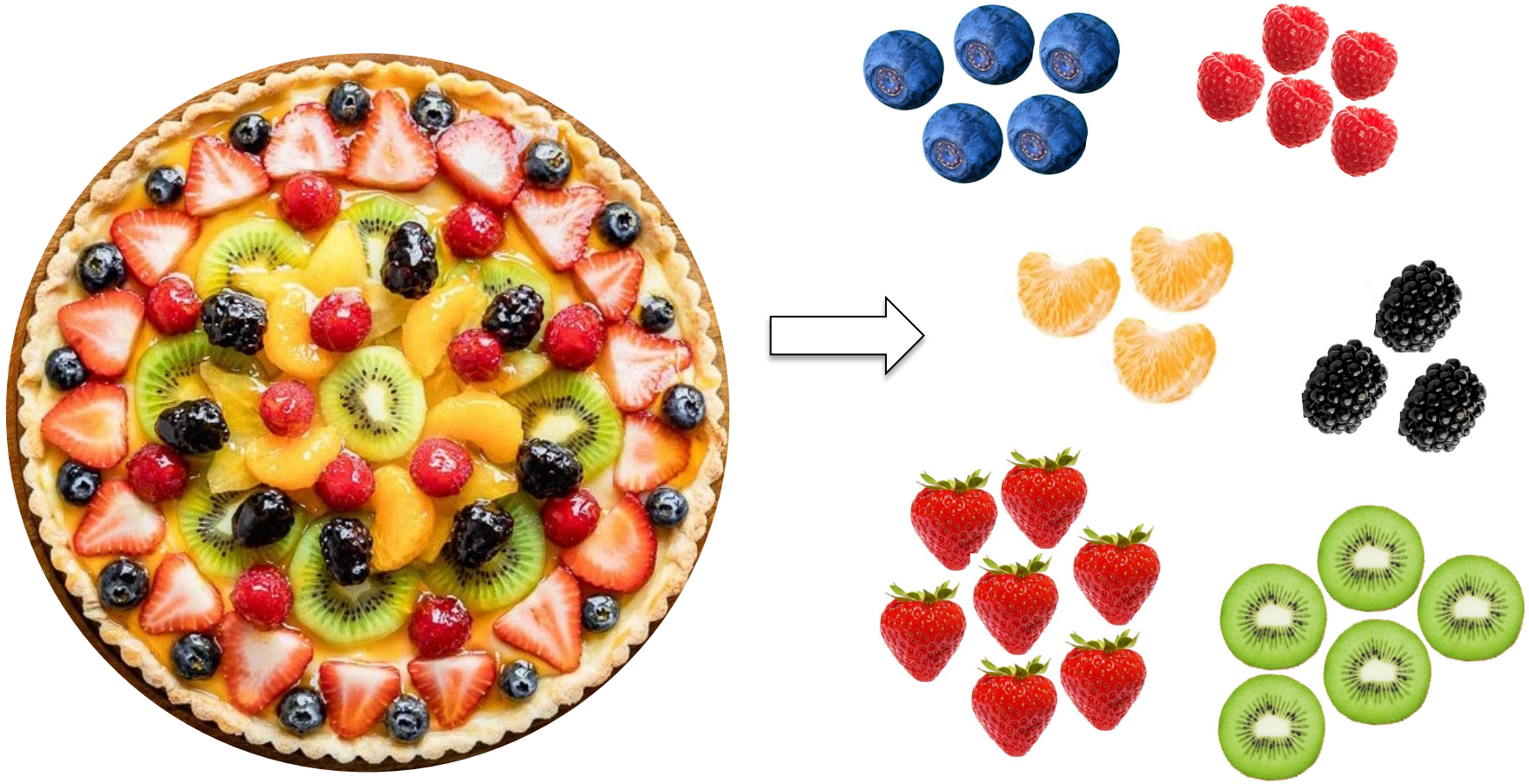
Single-cell Profiling

In multicellular organisms, most tissues are not homogeneous.

Therefore, **bulk** profiling represents a **composite/average** of all cells in a sample.

Single-cell Profiling

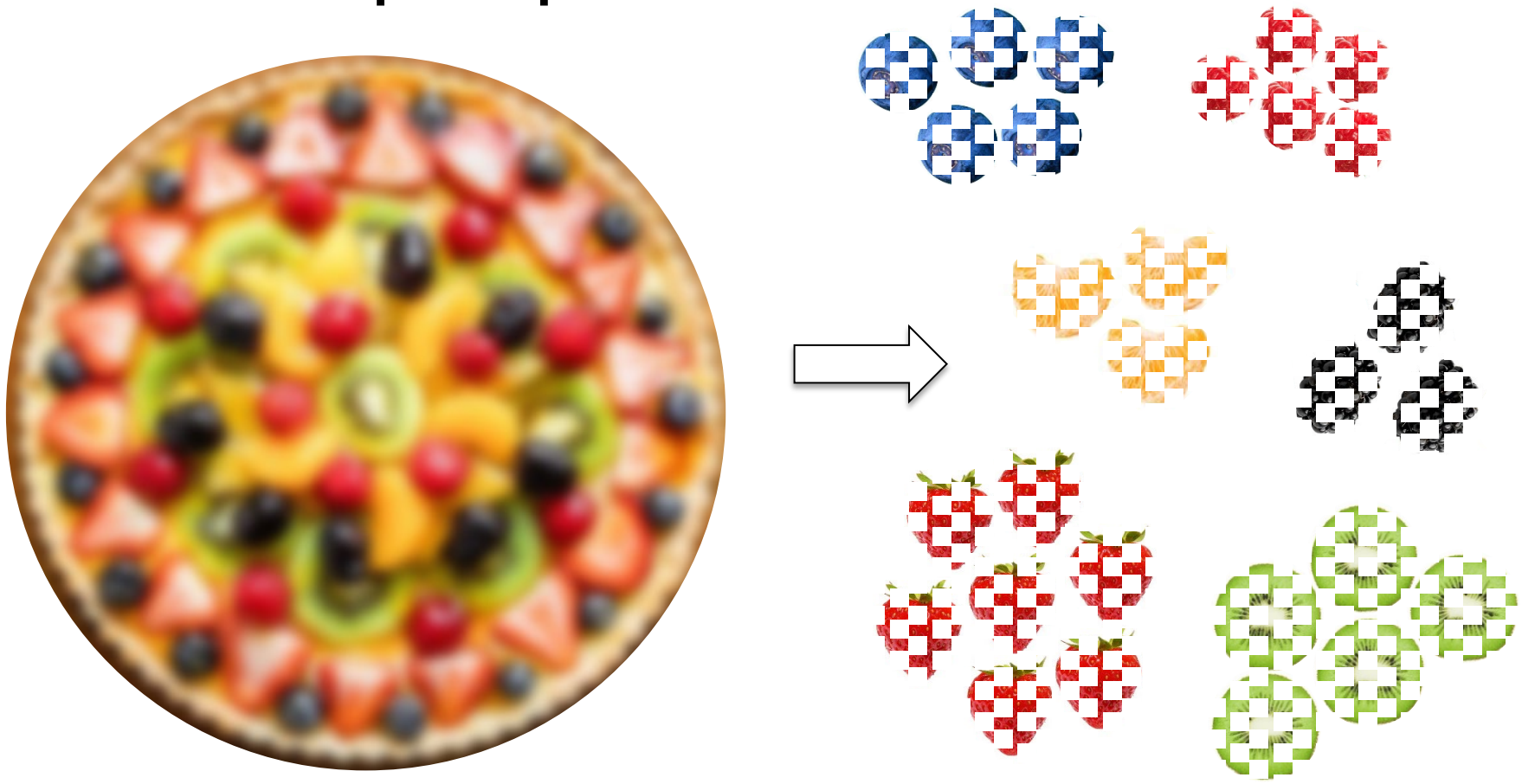
Single-cell profiling methods preserve information from individual cells



Single-cell Profiling

Single-cell profiling methods preserve information from individual cells

…but often results in a sparse profile



Single-cell Profiling Methods

Given a single-cell suspension….

- Make a library for each cell in separate tubes

  ~10-100 cells

  Fluidigm/Standard Biotools: C1

- Isolate individual cells in droplets

  ~1,000 – 10,000^{+} cells

  10x Genomics: Chromium, Fluent BioSciences: PIPseq

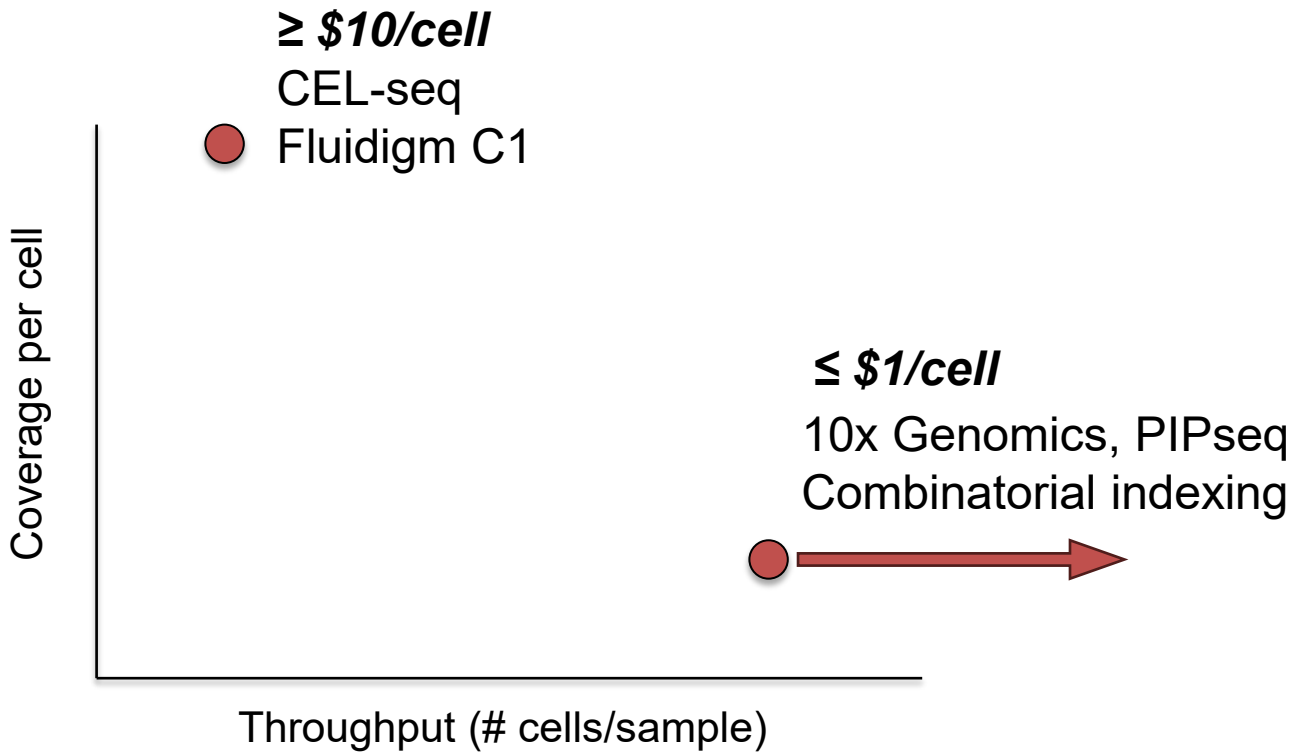
- Use a combinatorial indexing strategy

  ~1,000 – 1M cells

  Parse Biosciences: Evercode, Scale Biosciences, GIH

Split-seq: The movie

Single-cell Profiling Methods



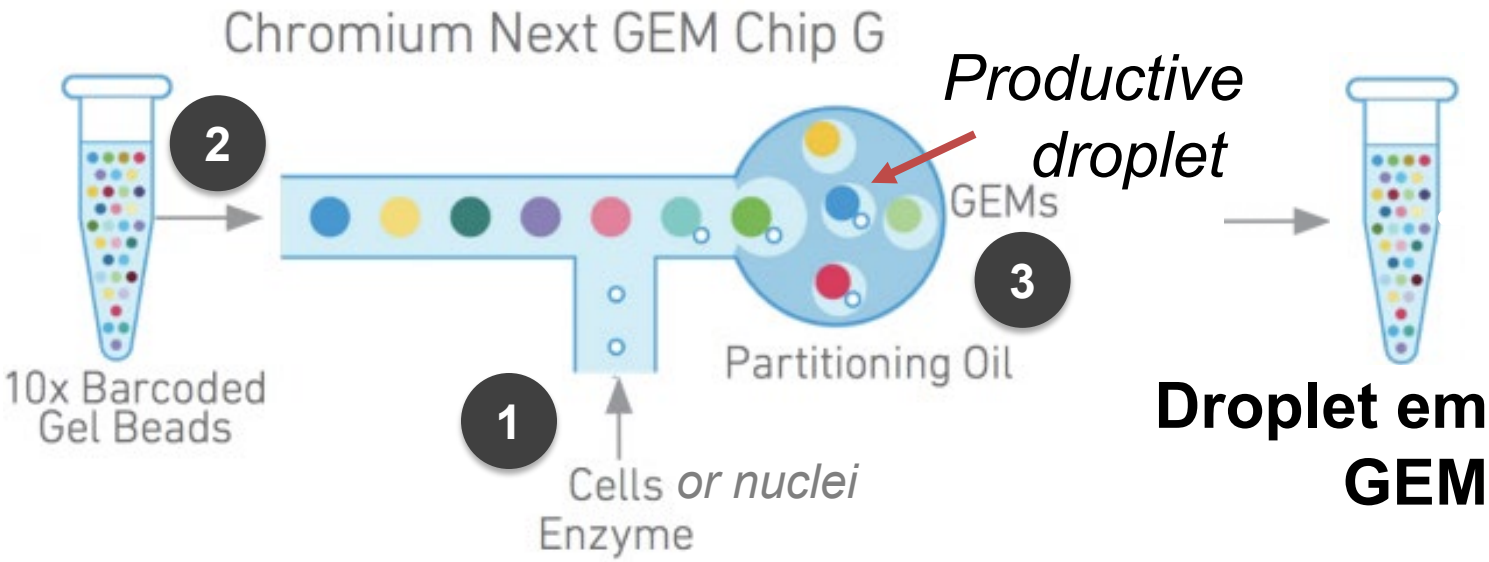
Single-cell Profiling Experiments

Experimental design considerations

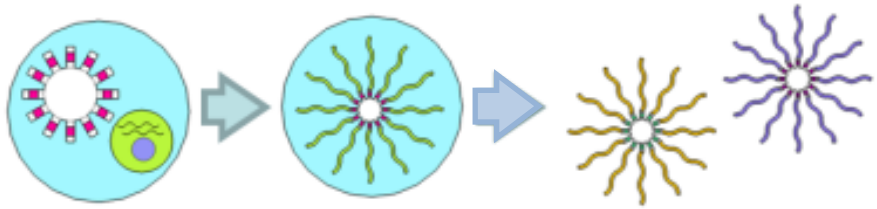
- Technology — *# of cells, coverage per cell*
- Sample prep — *high quality single cell/nucleus suspension*
  MACS/FACS for pre-enrichment or clean up
- Readout(s) — *gene expression, ATACseq, cell-surface markers, …*
- Replicates — *reproducibility/power, scaling sample prep, budget*
- Batching — *don't group samples by experimental condition!*
- Expertise — *multidisciplinary technology*

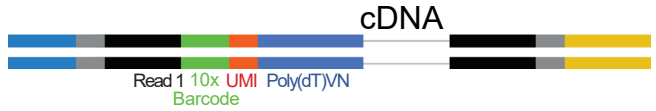
10x Genomics: Chromium

10x Genomics: Chromium



Droplet emulsion
GEM





cell lysis, mRNA capture, cDNA synthesis

break emulsion

cDNA amplification

library preparation

Droplet emulsion

Bulk sample

10x Genomics, DolomiteBio

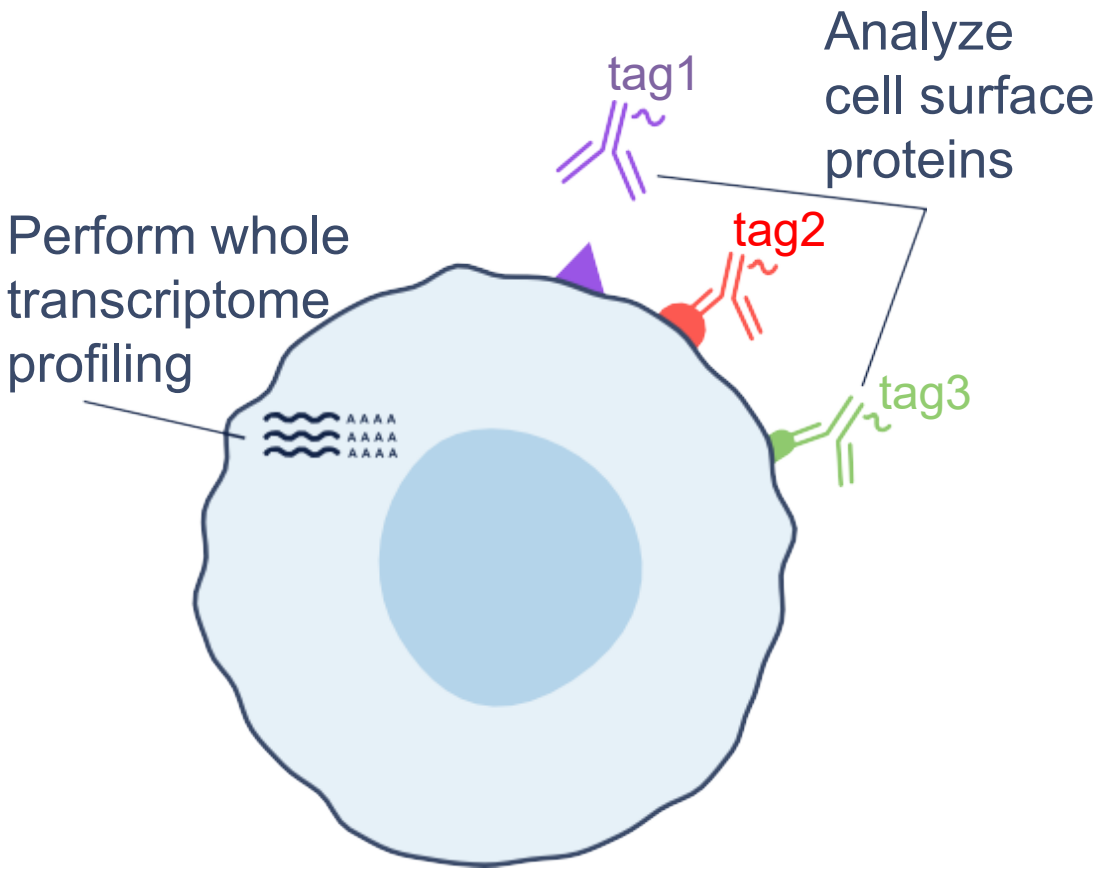
10x Genomics: Chromium

Similar cell barcoding strategy, different library content

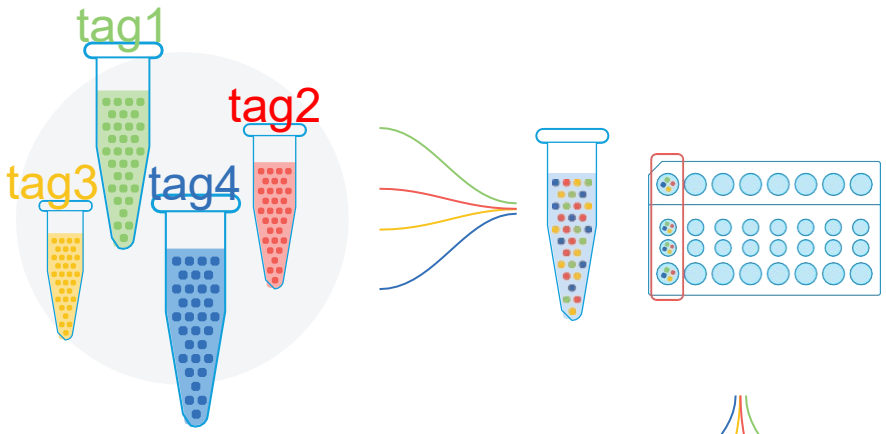
• scRNAseq	mRNA expression	*3' or 5' capture, fresh or fixed (Hs/Mm)*
	add: cell-surface markers	*oligo-tagged Antibodies*
	cell hashing	*oligo-tagged Antibodies or CellPlex*
	CRISPR	*sgRNAs with capture sequence*
	immune profiling	*VDJ/TCR repertoire profiling*
	antigen mapping	*barcoded antigen-specific Ab or MHC*
• scATACseq	accessible chromatin (enhancers, promoters)	
	stand-alone or paired with scRNAseq	

10x Genomics: Feature Barcoding

CITE-seq (Cell Surface Protein)

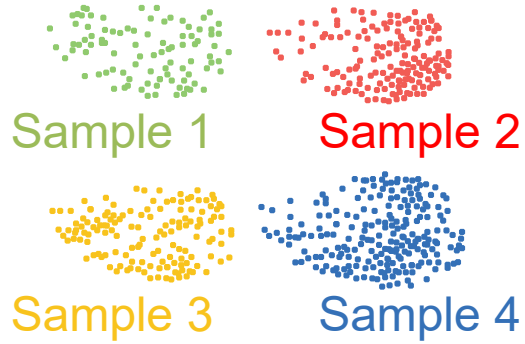


Sample Hashing (CellPlex)

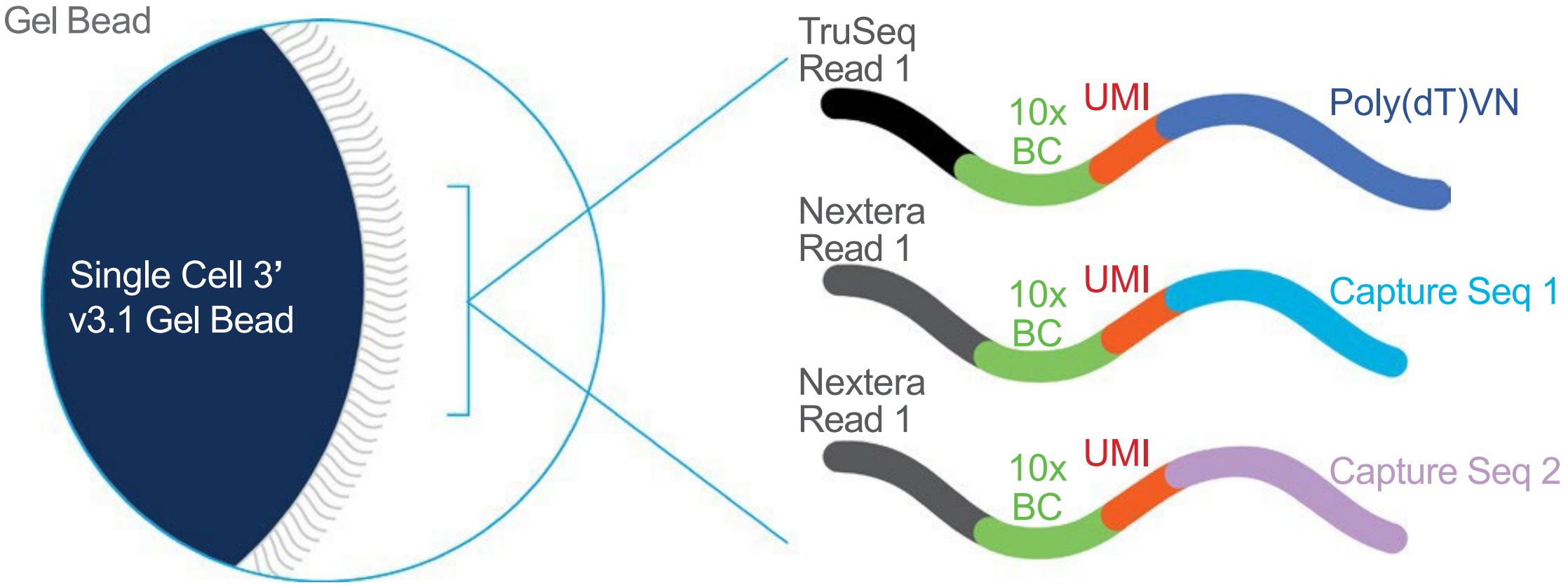


Label every cell in each sample with a unique tag, then pool

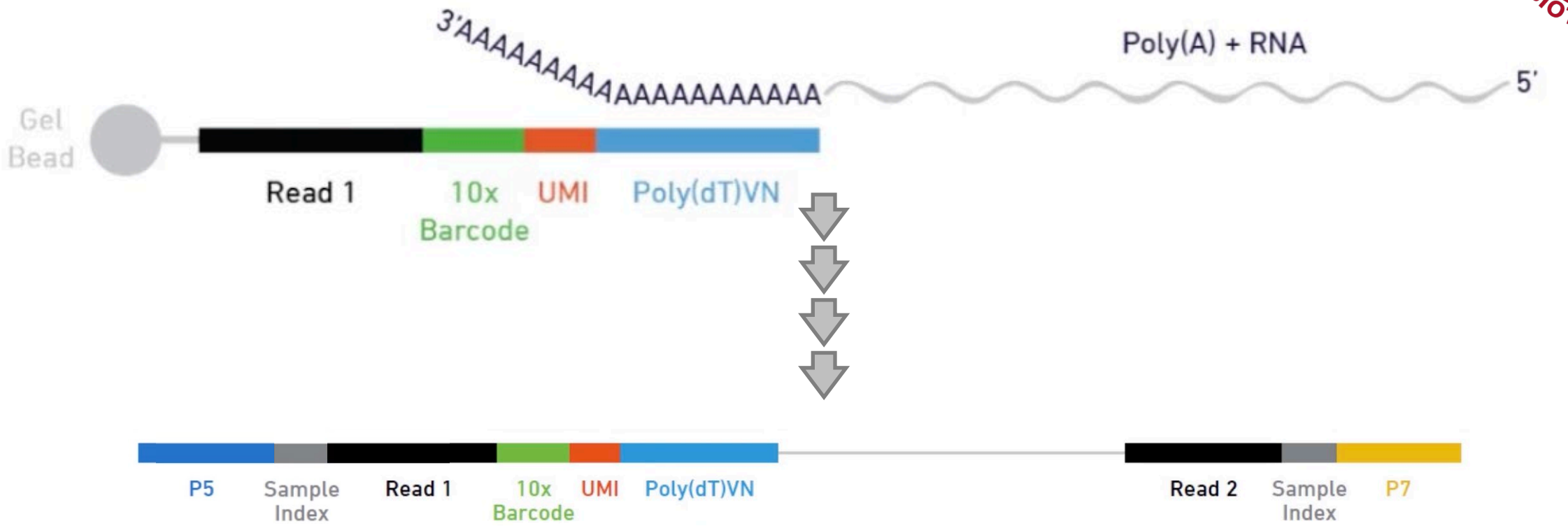
Demultiplex sequencing data based on hashtags & discard multiplets



10x Genomics: Gel Beads



10x Genomics: Library preparation



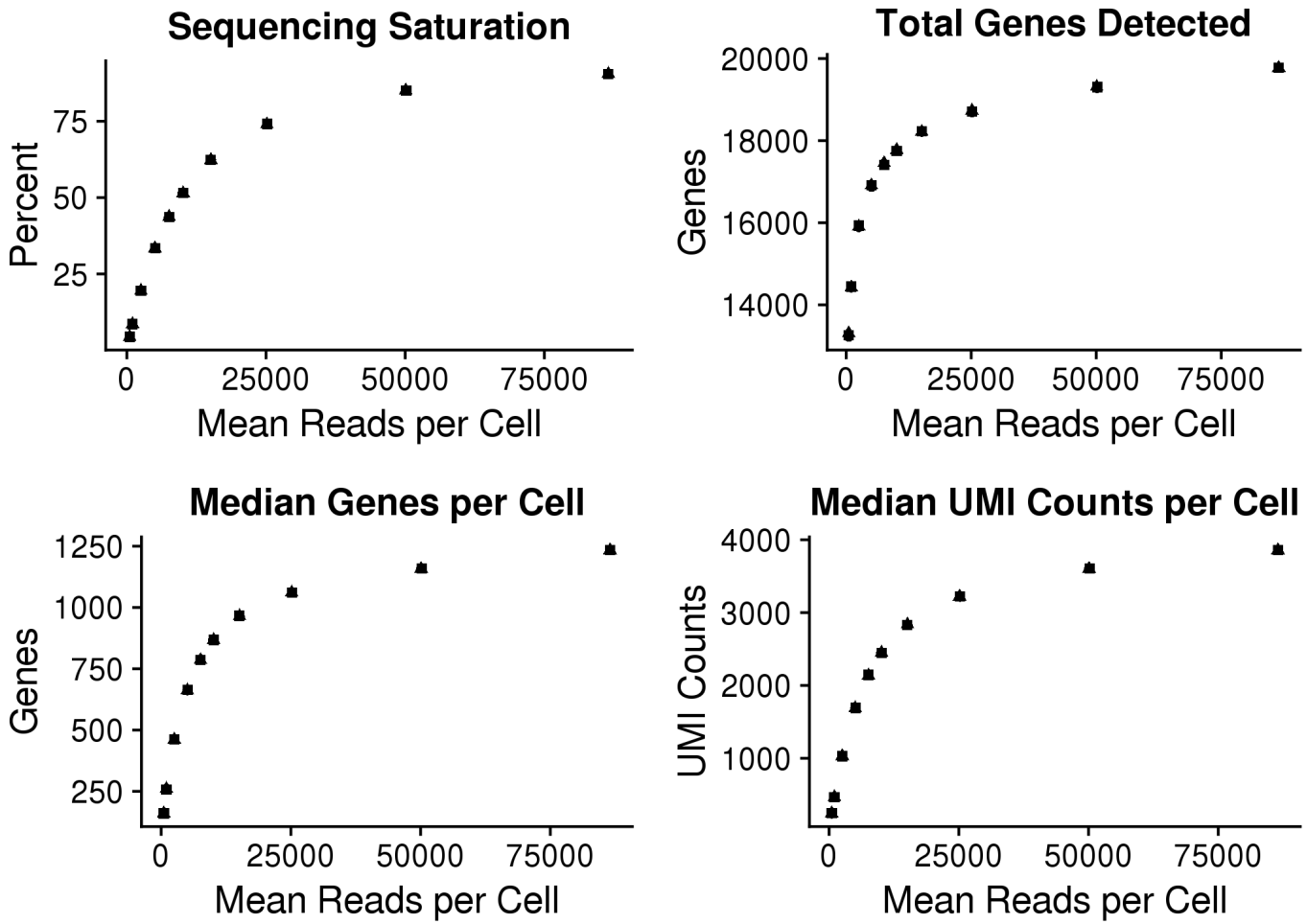
When multiple capture sequences are in use,
*multiple **libraries** are prepared from 1 **sample** (1 emulsion)*

https://www.10xgenomics.com/support

Sequencing 10x single-cell libraries

- Minimum recommended sequencing depth
  - Gene expression (GEX v3) *20,000 mean reads per cell*
  - Feature barcoding (CSP) *5,000 mean reads per cell*
  - **Use saturation metrics to guide target depth**
- Compatible with NovaSeq platform (2x150 PE reads) or NextSeq2000 (100bp kit)
  - Use UDIs when possible
  - Some library types can be pooled for sequencing, some cannot

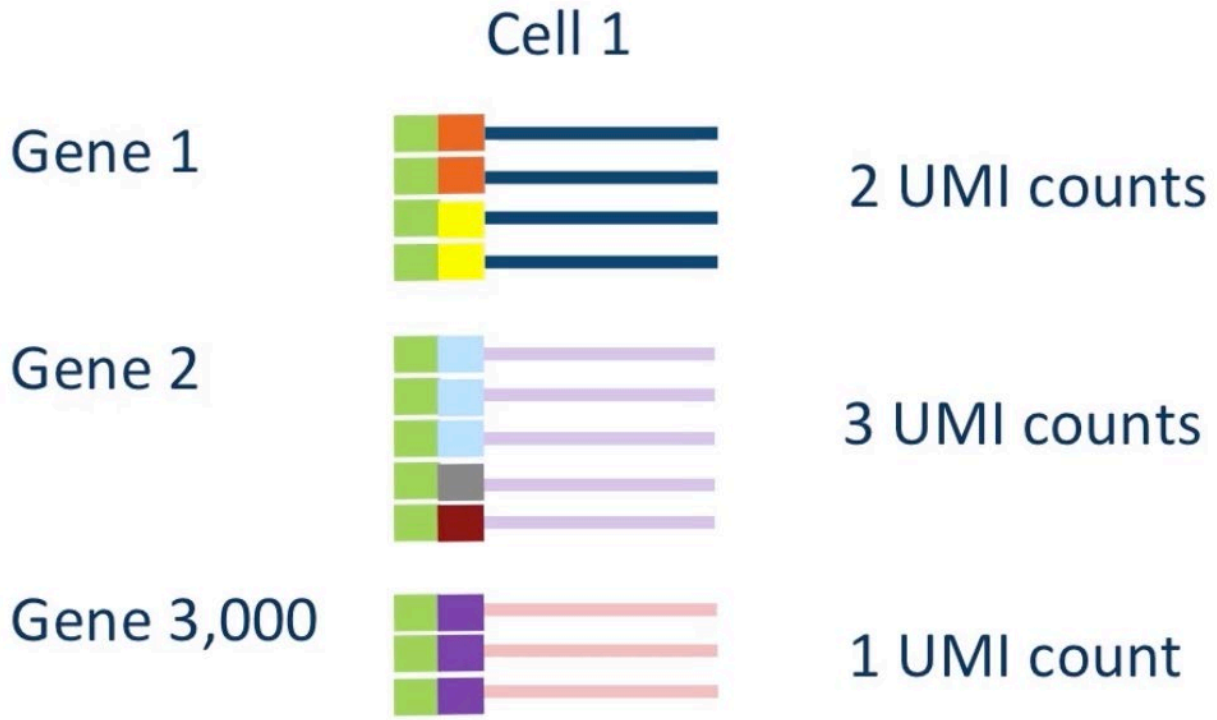
Sequencing 10x single-cell libraries



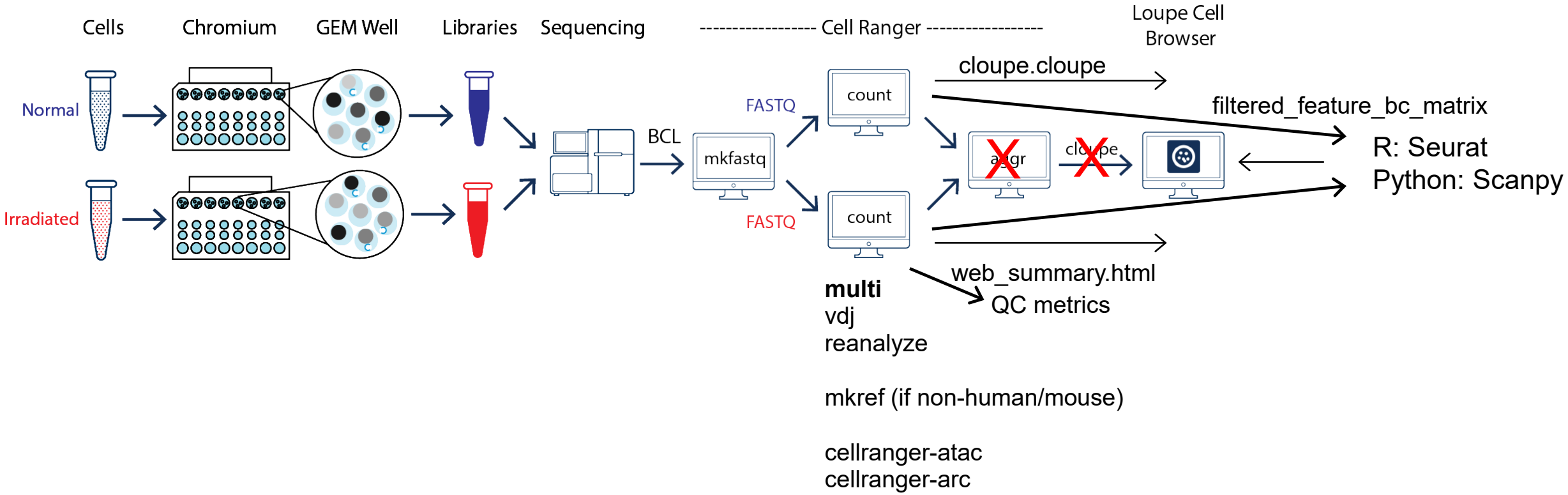
Accurate counting: Unique Molecular Identifier

UMI sequence:
detect and remove
PCR duplicates

'UMI' = distinct transcript

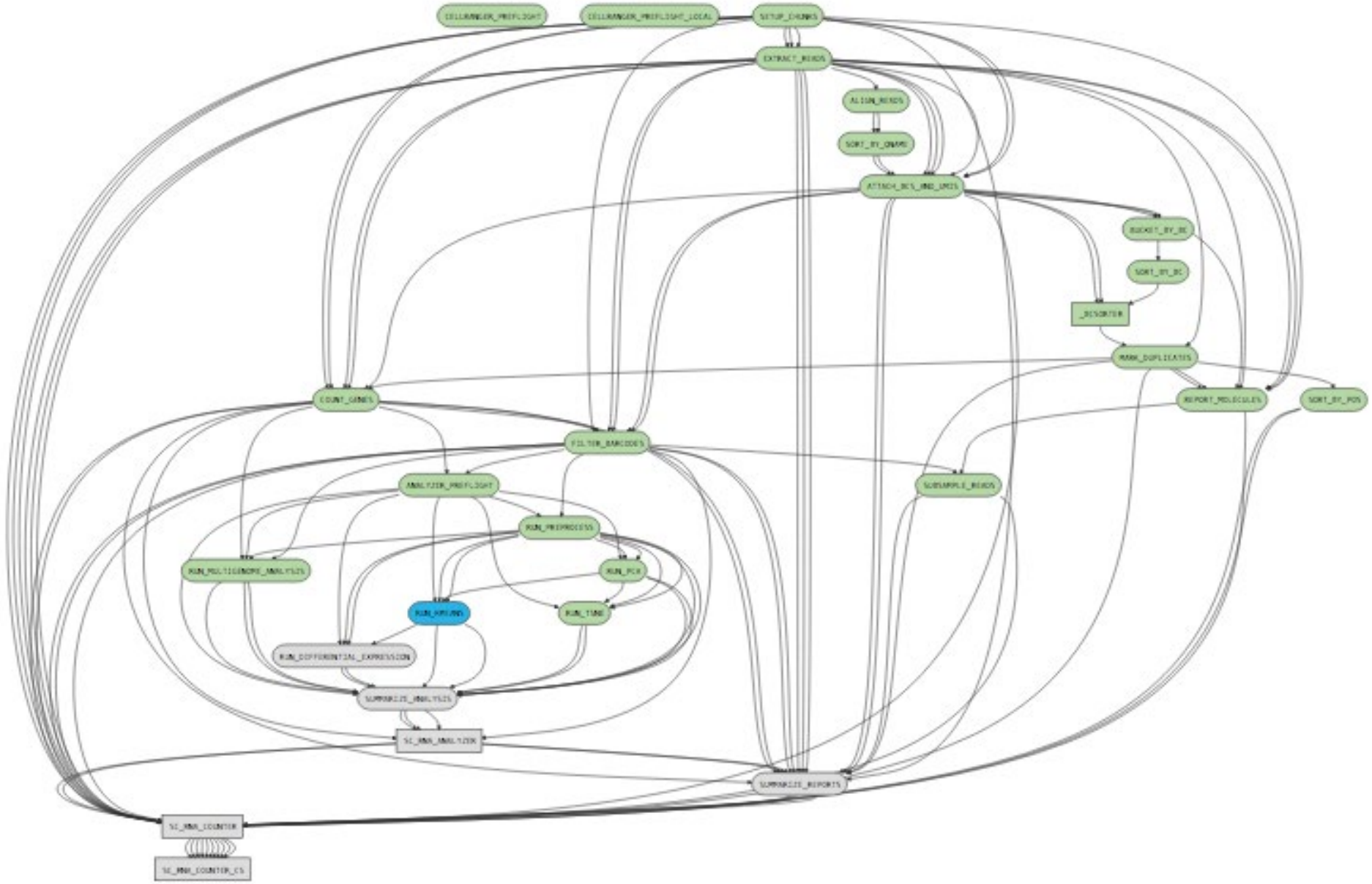


10x Genomics: Cell Ranger pipeline



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

10x Genomics: cellranger count pipeline



Cellranger

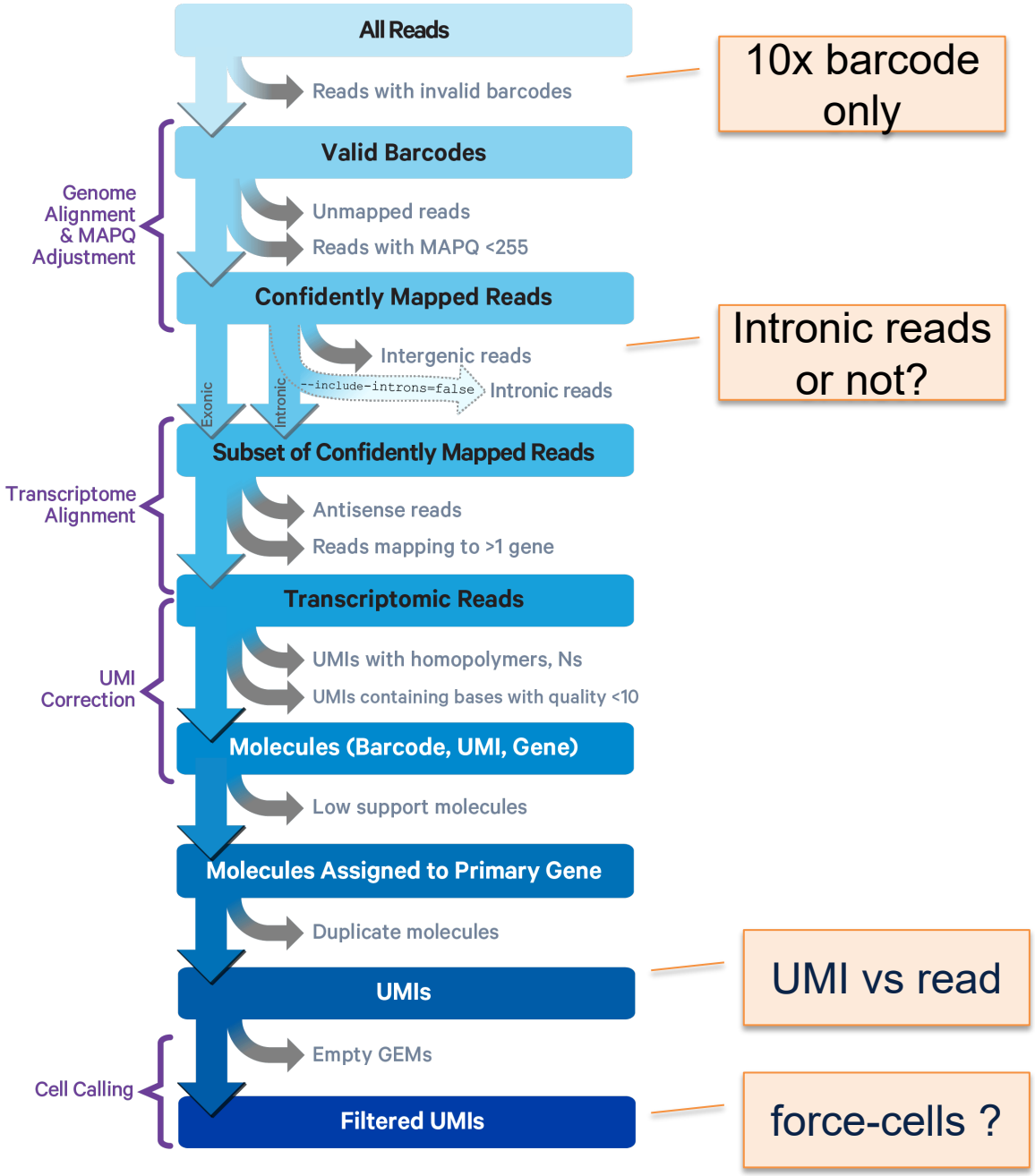
10x Genomics software to convert sequencing reads to gene expression matrix

Keep reads with valid cell barcode

Map reads to reference genome and transcriptome

Resolve cell barcode, UMI, gene

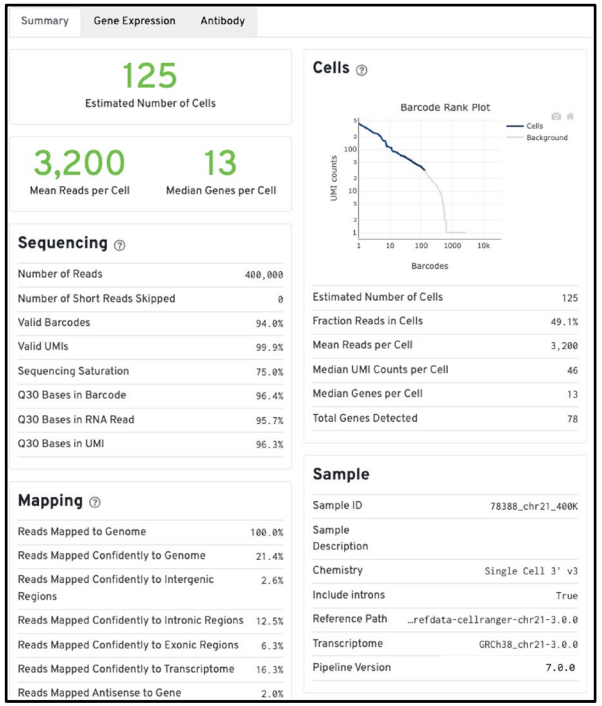
Count UMI



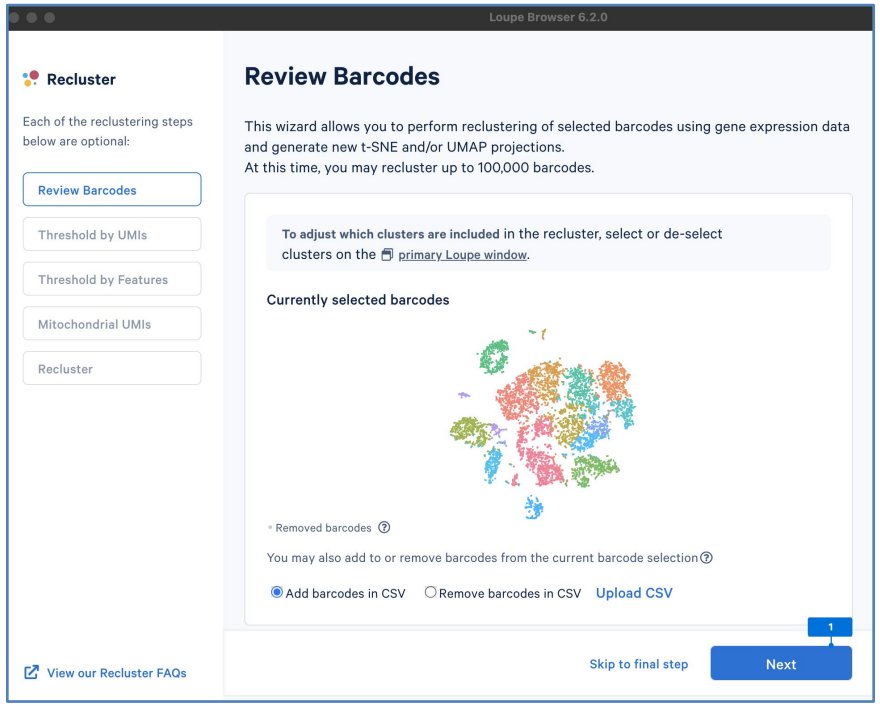
For details, see https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-gex-algorithm

Cellranger outputs

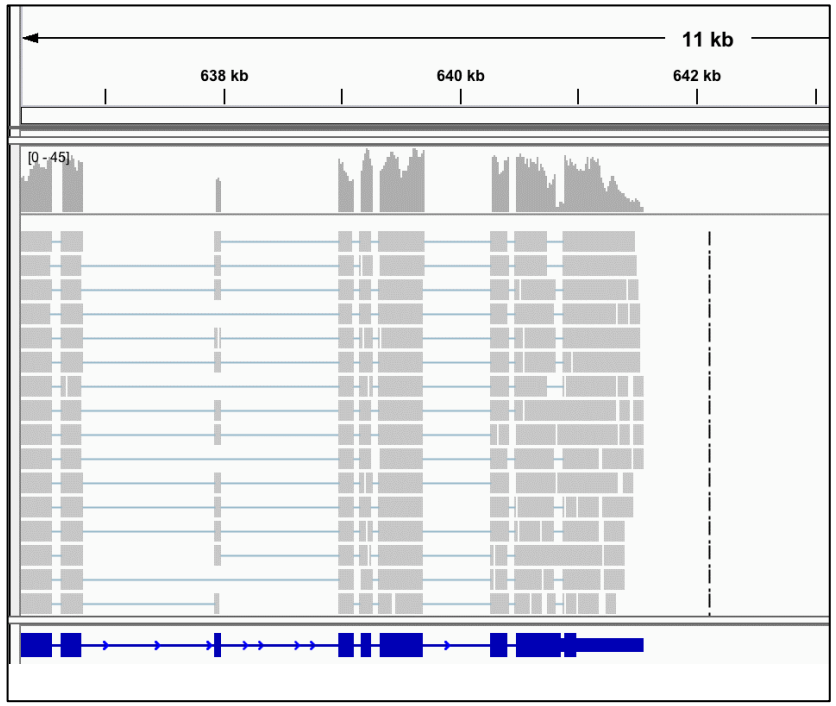
web_summary.html



cloupe.cloupe (for Loupe Browser)



possorted_genome_bam
sample_alignments.bam



(Use bam files to deposit multiplex data to NCBI)

For downstream data analysis (Seurat or Scanpy)

filtered_feature_bc_matrix.h5: HDF5 formatted
filtered_feature_bc_matrix: MEX formated

Cellranger tools

- **mkref**
- mkfastq
- count
- multi
- reanalyze
- aggr

Prebuilt references

- Human
- Mouse

URL: https://www.10xgenomics.com/support/software/cell-ranger/downloads#reference-downloads

Custom references

https://www.10xgenomics.com/support/software/cell-ranger/latest/tutorials/cr-tutorial-mr

```
cellranger mkgtf \
  Danio_rerio.GRCz11.105.gtf \
  Danio_rerio.GRCz11.105.filtered.gtf \
  --attribute=gene_biotype:protein_coding
```

Filter GTF (optional):
e.g. keep protein coding gene only

```
Add marker gene to GTF
(instructions in mkref tutorial page)
```

e.g. GFP gene

```
cellranger mkref \
--genome=Drerio_genome \
--fasta=Danio_rerio.GRCz11.dna.primary_assembly.fa \
--genes=Danio_rerio.GRCz11.105.filtered.gtf
```

Build reference database from genome FASTA + GTF

10x Cellranger pipelines

- mkref
- **mkfastq**
- count
- multi
- reanalyze
- aggr

mkfastq: convert Illumina .bcl file to .fastq.gz file

- Most vendors would do this step for you;
- FASTQ file naming convention
  - Files can be in multiple levels of sub-directories.
    Cellranger recursively locates files with expected sample names;
  - One sample (library) can have multiple files;
  - **But file names must be in this format.**

IgG1d_S1_L001_R1_001.fastq.gz ← cell barcode + UMI

IgG1d_S1_L001_R2_001.fastq.gz ← RNAseq read

Sample name S# Lane Read type (R1, R2 required; I1, I2 optional)

https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/inputs/cr-mkfastq

Cellranger tools

- mkref
- mkfastq
- **count**
- **multi**
- reanalyze
- aggr

cellranger count

Types of libraries

- Singleplex
- Singleplex + feature barcode

cellranger multi

(one or more fastq sets containing multiple samples or libraries)

- Multiplex
- Multiplex + feature barcode

❖ Multiplex: multiple samples in a single GEM well, each sample tagged with barcoded antibody or lipid.

❖ Feature barcode: for example, CITE-seq, which uses DNA-barcoded antibodies targeting cell surface proteins

Input fastq data files

generated by mkfastq

Singleplex data

```
IgG1d_S1_L001_R1_001.fastq.gz
IgG1d_S1_L001_R2_001.fastq.gz
```

(R1 & R2 files required, I1 & I2 files optional)

Multiplex data

```
PBMC_gex
    PBMC_gex_S2_L001_I1_001.fastq.gz
    PBMC_gex_S2_L001_I2_001.fastq.gz
    PBMC_gex_S2_L001_R1_001.fastq.gz
    PBMC_gex_S2_L001_R2_001.fastq.gz
PBMC_plex
    PBMC_plex_S1_L001_I1_001.fastq.gz
    PBMC_plex_S1_L001_I2_001.fastq.gz
    PBMC_plex_S1_L001_R1_001.fastq.gz
    PBMC_plex_S1_L001_R2_001.fastq.gz
```

RNAseq

Multiplex barcode

Multiplex data + feature barcode

```
antibody_fastqs
    20k_NSCLC_DTC_3p_nextgem_antibody_S6_L001_I1_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_antibody_S6_L001_I2_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_antibody_S6_L001_R1_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_antibody_S6_L001_R2_001.fastq.gz
cmo_fastqs
    20k_NSCLC_DTC_3p_nextgem_cmo_S5_L001_I1_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_cmo_S5_L001_I2_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_cmo_S5_L001_R1_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_cmo_S5_L001_R2_001.fastq.gz
gex_fastqs
    20k_NSCLC_DTC_3p_nextgem_gex_S4_L001_I1_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_gex_S4_L001_I2_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_gex_S4_L001_R1_001.fastq.gz
    20k_NSCLC_DTC_3p_nextgem_gex_S4_L001_R2_001.fastq.gz
```

Feature barcode

Multiplex barcode

RNAseq

Cellranger tools

- mkref
- mkfastq
- **count**
- **multi**
- reanalyze
- aggr

cellranger count

```
cellranger count \
      --id=sample345 \
      --transcriptome=/workdir/jdoe/refdata-gex-GRCh38-2020-A \
      --fastqs=/workdir/jdoe/fastq_path \
      --sample=mysample \
      --localcores=8 \
      --localmem=64
```

- --id: output directory
- --sample: sample name, must matching the fastq file names
- --localcores and --localmem: if not specified, one cellranger run will take over all available on the computer. There is no benefit from localcores > 32. You might want to parallelize the run, with 4 samples at a time, 16 cores per sample.

https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/running-pipelines/cr-gex-count

Cellranger tools

cellranger count

A few parameters to consider:

--include-introns: Include intronic reads
default: true in v7

--force-cells: Specify number of cells
default: call automatically

Cellranger tools

- mkref
- mkfastq
- **count**
- **multi**
- reanalyze
- aggr

cellranger multi

```
cellranger multi --id=sample345 \
     --csv=/workdir/jdoe/sample345.csv \
     --localcores=8 \
     --localmem=64
```

- --id : output directory
- --csv: a figuration file in csv format

https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/running-pipelines/cr-3p-multi

Cellranger tools

- mkref
- mkfastq
- **count**
- **multi**
- reanalyze
- aggr

cellranger multi

.csv file format

```
[gene-expression]
reference,/path/to/transcriptome

[feature]
reference,/path/to/feature_reference.csv

[libraries]
fastq_id,fastqs,feature_types
gex1,/path/to/fastqs,Gene Expression
abc1,/path/to/fastqs,Antibody Capture
mux1,/path/to/fastqs,Multiplexing Capture

[samples]
sample_id,cmo_ids
sample1,CMO301
sample2,CMO303
```

Specify a file with feature barcode sequences.

feature_types must match one of supported types:

- Gene Expression
- Antibody capture
- Multiplexing Capture
- …

If using 3rd party multiplexing kit, include a “cmo-set” file under [gene-expression] section

https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/running-pipelines/cr-3p-multi

Cellranger tools

- mkref
- mkfastq
- **count**
- **multi**
- reanalyze
- aggr

Feature barcode .csv file format

```
id,name,read,pattern,sequence,feature_type
CD3,CD3,R2,^NNNNNNNNNN(BC)NNNNNNNNN,CTCATTGTAACTCCT,Antibody Capture
CD4,CD4,R2,^NNNNNNNNNN(BC)NNNNNNNNN,TGTTCCCGCTCAACT,Antibody Capture
CD8,CD8,R2,^NNNNNNNNNN(BC)NNNNNNNNN,GCGCAACTTGATGAT,Antibody Capture
CD11c,CD11c,R2,^NNNNNNNNNN(BC)NNNNNNNNN,TACGCCTATAACTTG,Antibody
Capture
CMO301,CMO301,R2,5P(BC),ATGAGGAATTCCTGC,Multiplexing Capture
CMO302,CMO302,R2,5P(BC),CATGCCAATAGAGCG,Multiplexing Capture
CMO303,CMO303,R2,5P(BC),CCGTCGTCCAAGCAT,Multiplexing Capture
```

Output gene expression matrix

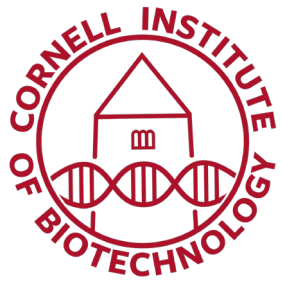
…		
ENSG00000277836	AC141272.1	Gene Expression
ENSG00000278633	AC023491.2	Gene Expression
ENSG00000276017	AC007325.1	Gene Expression
ENSG00000278817	AC007325.4	Gene Expression
ENSG00000277196	AC007325.2	Gene Expression
CD3	CD3	Antibody Capture
CD4	CD4	Antibody Capture
CD8	CD8	Antibody Capture

“cellranger count” outputs

```
outs
├── analysis
├── cloupe.cloupe
├── filtered_feature_bc_matrix
├── filtered_feature_bc_matrix.h5
├── metrics_summary.csv
├── molecule_info.h5
├── possorted_genome_bam.bam
├── possorted_genome_bam.bam.bai
├── raw_feature_bc_matrix
├── raw_feature_bc_matrix.h5
└── web_summary.html
```

“cellranger multi” outputs

```
outs
├── config.csv
└── per_sample_outs
    ├── donor_1
    │   ├── count
    │   │   ├── aggregate_barcodes.csv
    │   │   ├── analysis
    │   │   ├── feature_reference.csv
    │   │   ├── sample_alignments.bam
    │   │   ├── sample_alignments.bam.bai
    │   │   ├── sample_cloupe.cloupe
    │   │   ├── sample_filtered_barcodes.csv
    │   │   ├── sample_filtered_feature_bc_matrix
    │   │   ├── sample_filtered_feature_bc_matrix.h5
    │   │   └── sample_molecule_info.h5
    │   ├── metrics_summary.csv
    │   └── web_summary.html
    └── donor_2
        ├── count
        │   ├── aggregate_barcodes.csv
        │   ├── analysis
        │   ├── feature_reference.csv
        │   ├── sample_alignments.bam
        │   ├── sample_alignments.bam.bai
        │   ├── sample_cloupe.cloupe
        │   ├── sample_filtered_barcodes.csv
        │   ├── sample_filtered_feature_bc_matrix
        │   ├── sample_filtered_feature_bc_matrix.h5
        │   └── sample_molecule_info.h5
        ├── metrics_summary.csv
        └── web_summary.html
```



Cellranger tools

- mkref
- mkfastq
- count
- multi
- **reanalyze**
- aggr

Not commonly used.

force-cell option

```
cellranger reanalyze \
  --id=10k_pbmc_reanalyze_pc_clust \
  --matrix=pbmc_10k_v3_filtered_feature_bc_matrix.h5 \
  --force-cell=5000
```

Cellranger tools

- mkref
- mkfastq
- count
- multi
- reanalyze
- **aggr**

cellranger aggr Aggregating Multiple Samples

Not commonly used.

Use Seurat instead to integrate samples

Data Sharing

Deposit 10x Genomics data to NCBI GEO/SRA

Singleplex: R1_001.fastq.gz & R2_001.fastq.gz

Multiplex: sample_alignments.bam

https://kb.10xgenomics.com/hc/en-us/articles/360024716391-What-format-of-10x-Genomics-data-should-I-submit-to-NCBI-GEO-SRA

Download public data from SRA

Files downloaded from SRA:

SRR9291388_1.fastq.gz
SRR9291388_2.fastq.gz

Sample name Read type

Change file names before running cellranger:

SRR9291388_S1_L001_R1_001.fastq.gz
SRR9291388_S1_L001_R2_001.fastq.gz

Sample name S# Lane Read type

--or--

SRR9291388_S1_R1_001.fastq.gz
SRR9291388_S1_R2_001.fastq.gz

Sample name S# Read type

Cellranger count QC

web_summary.html file: 1 per sample

- Alerts and warnings

- Summary tab

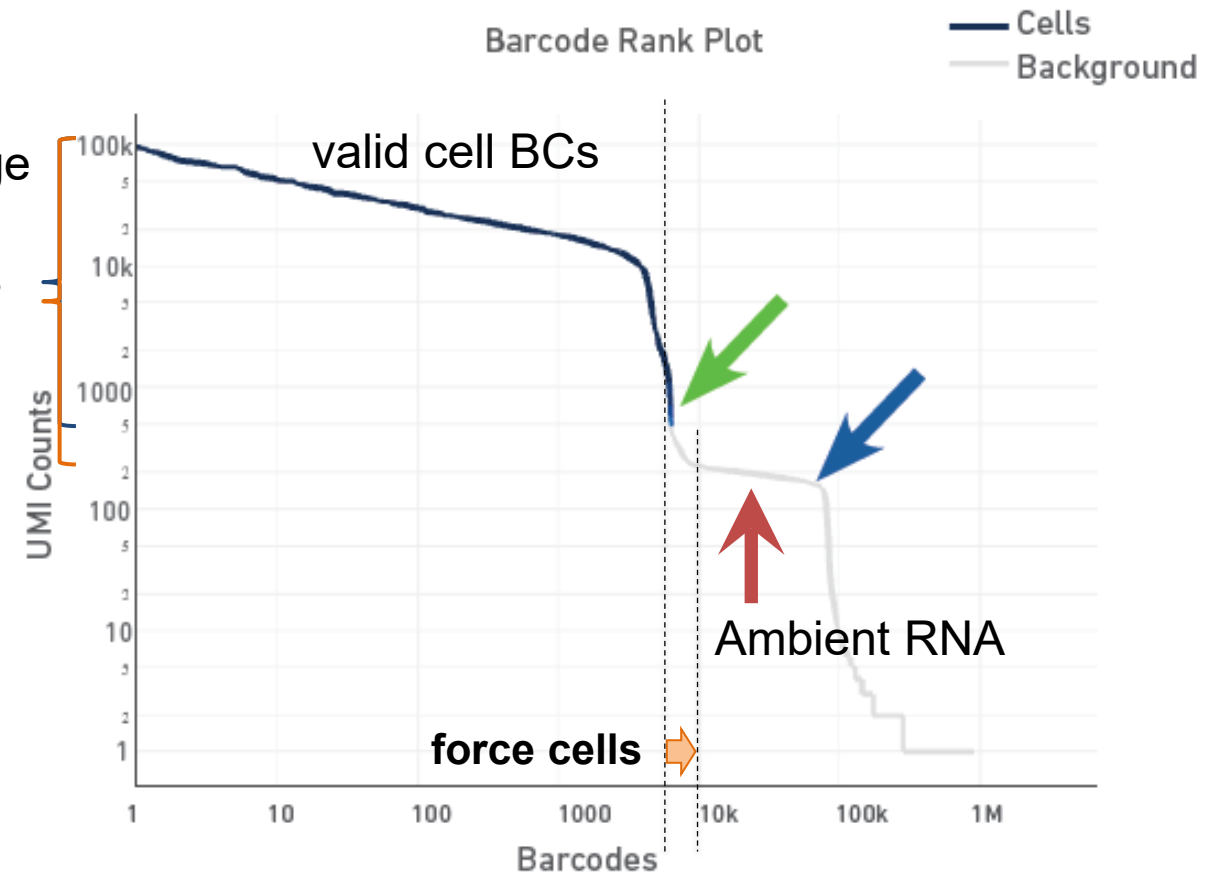
  Number of cells, mean reads per cell, median genes and UMIs per cell
  Read counts, mapping rates, and much more *(also in metrics_summary.csv)*

  Barcode Rank (Knee) plot

- Gene Expression tab

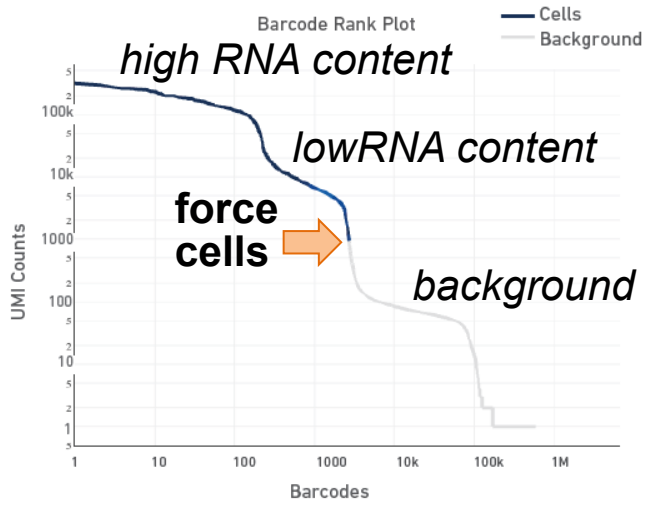
  tSNE clustering, top marker genes, saturation curves

Barcode Rank (Knee) Plot

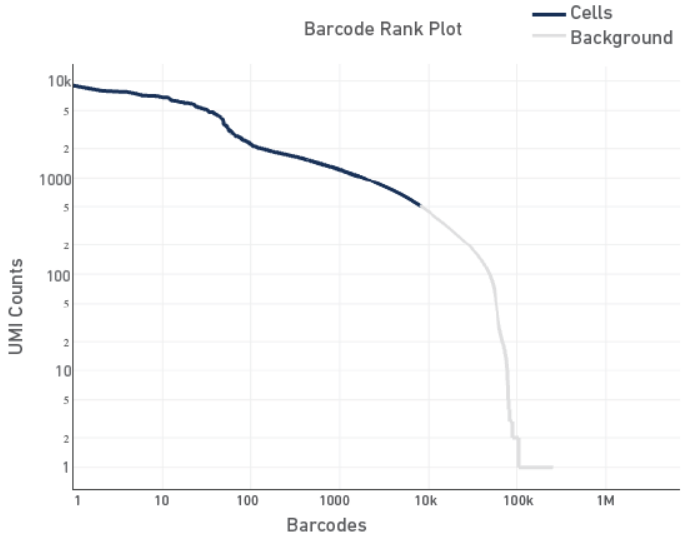


Range of UMIs per cell

ordered by UMI count per barcode (=gel bead)



Multiple knees: heterogeneous cell types



Shallow cliffs, rounded knees: low sample quality or loss of single-cell behavior

Cellranger count (or multi) QC

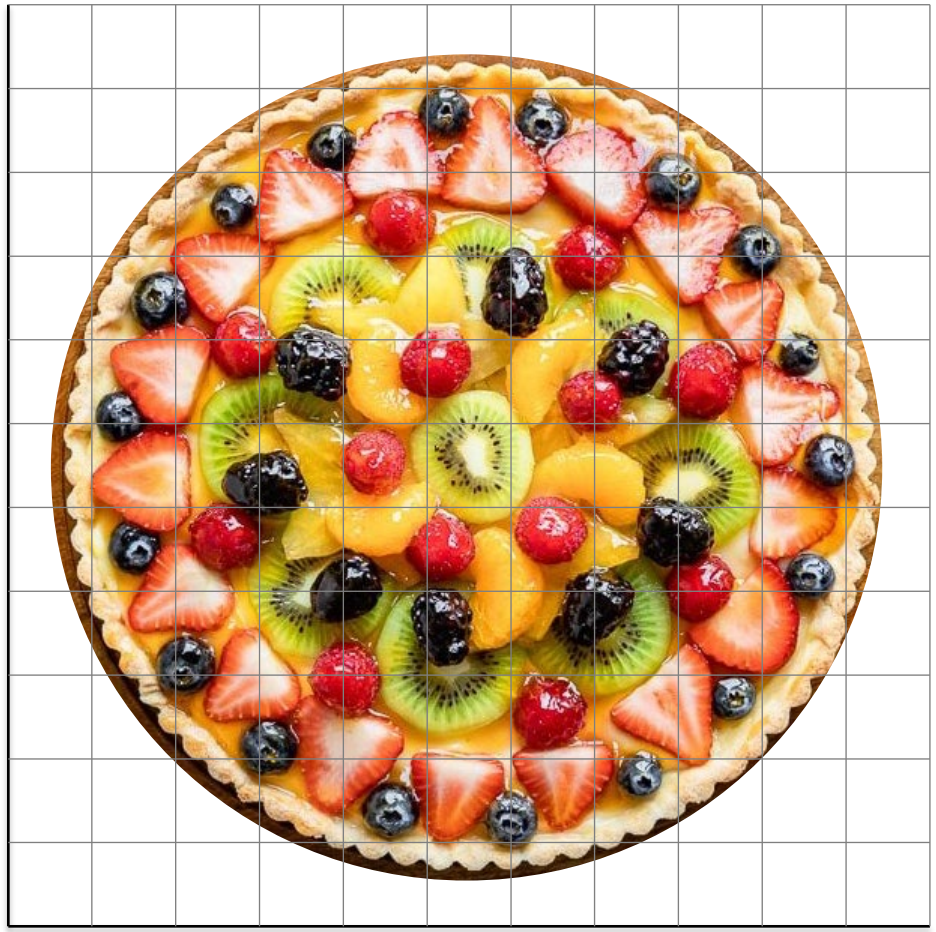
When cellranger count (or multi) has completed:

- Using Mozilla, find the web_summary.html and cloupe.cloupe files in the cellranger ‘outs’ directory and transfer to your laptop/computer
- web_summary.html files: open with a web browser
- cloupe.cloupe files: open with Loupe Browser v7

Links to detailed guides and tutorials at the 10x Genomics support web site can be found in the hands-on html document

Spatial Transcriptomics

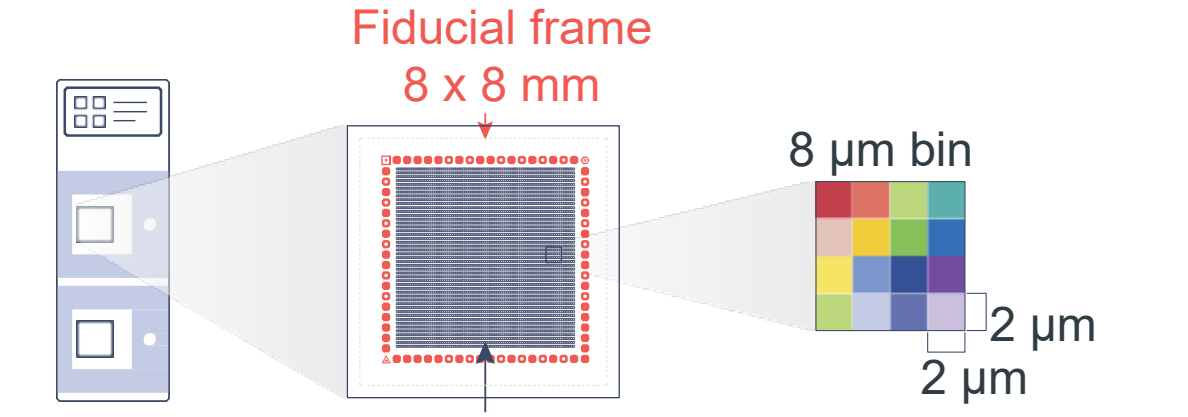
Preserve spatial organization



- **Imaging**
  - probe-based FISH for 100s of genes
  - sub-cellular resolution
  - **Merscope (Visgen), Xenium (10x Genomics)**

- **Genomics**
  - cell BC can be mapped to 2D coordinates
  - not true single-cell resolution
  - sequencing read-out
  - **Visium (10x Genomics), Slide-seq (Curio)**

10x Genomics: Visium HD

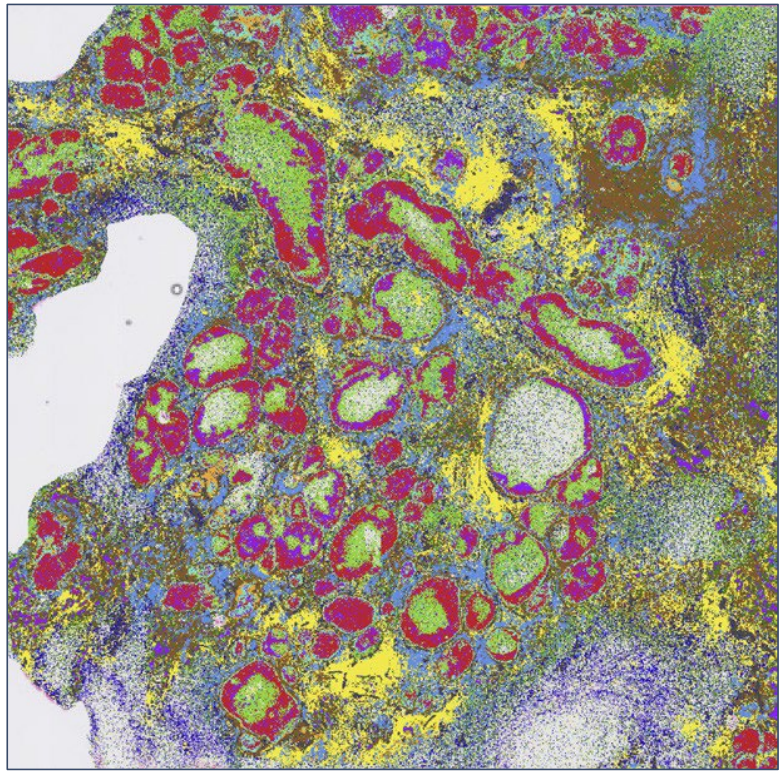


Visium HD Slide

Capture Area with continuous lawn of oligos, 6.5 x 6.5 mm

Grid of 2 x 2 µm barcoded squares, binned to 8 x 8 µm

Spatially mapped GEX clustering



- Breast glandular cells
- Breast myoepithelial cells
- Endothelial cells
- Adipocytes
- T cells
- B plasma cells
- Fibrolasts