Single-cell RNA-Seq Analysis

生物資訊與數據分析

10/14/2025

余承欣 (Chen-Hsin Albert Yu)

IMB Bioinformatics Core

Outline

1. Introduction of Single-cell RNA-Seq experiment and application
2. Analysis of 10X scRNA-seq data

   -- Cell Ranger command line

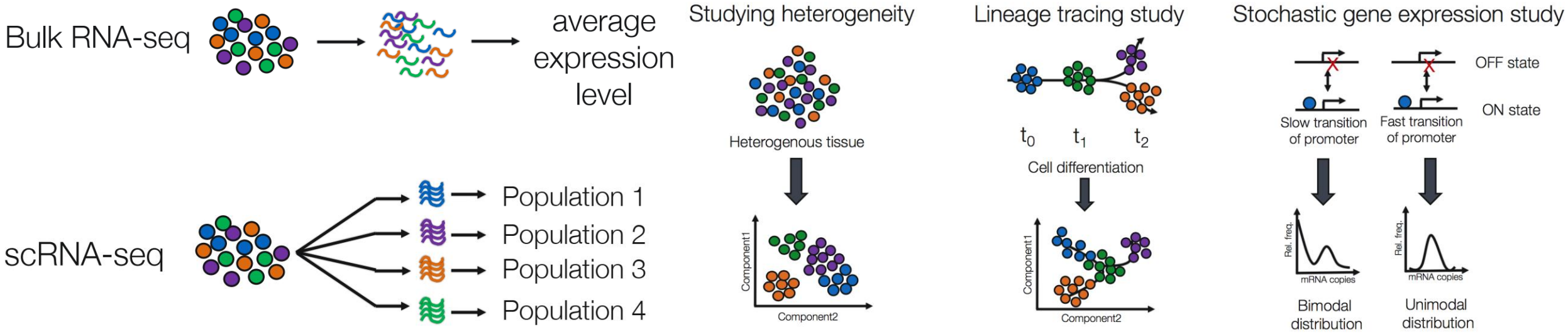
   -- 10X Cloud
3. Data visualization and advanced analysis

   -- Loupe Browser

   -- Further analysis: Seurat (R package) or Scanpy (python)

Why using single-cell RNA-Seq

- Cell heterogeneity (discovering rare cell types, ex: cancer stem cells)
- Cell population dynamics
- Clonal evolution (developmental trajectories)
- Cell regulatory networks



Popular commercial single cell RNA-Seq platforms

10x Genomics: Chromium

BD (Becton, Dickinson and Company): BD Rhapsody

Takara Bio: SMART-seq

Parse Biosciences: Evercode Technology

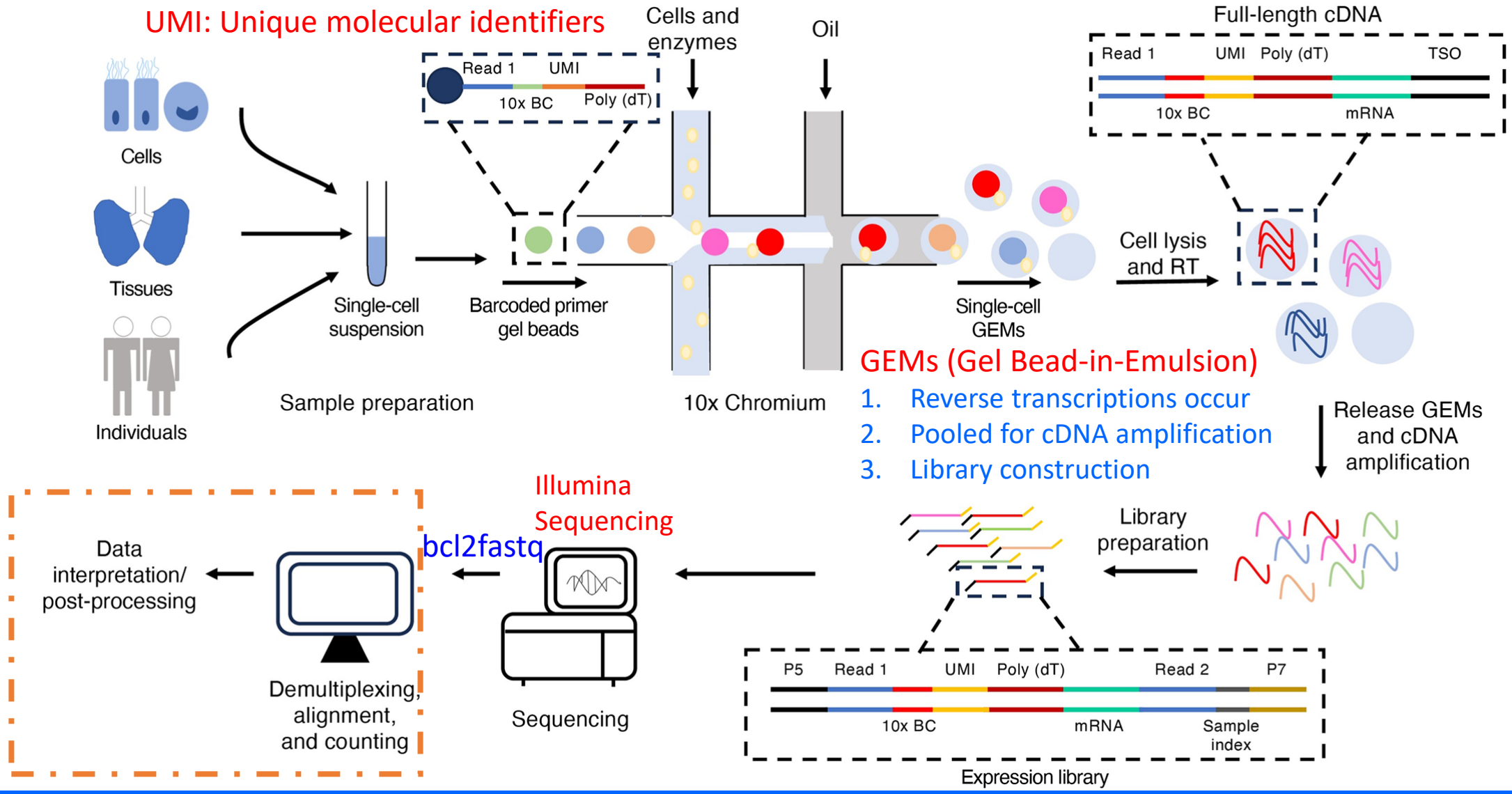
Singleron Bio: SCOPE Kits

Bio-Rad: ddSEQ Single-Cell Isolator

Mission Bio: Tapestri Platform

…

10X Genomic Single-cell sequencing workflow



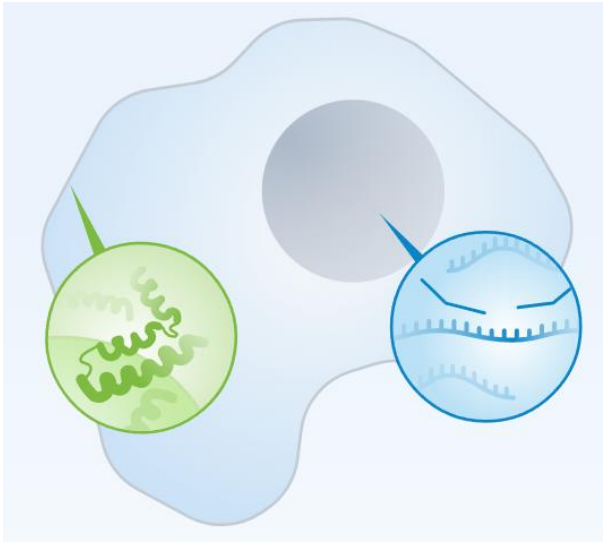
https://doi.org/10.1371/journal.ppat.1011898

10x Chromium Single-cell Gene Expression

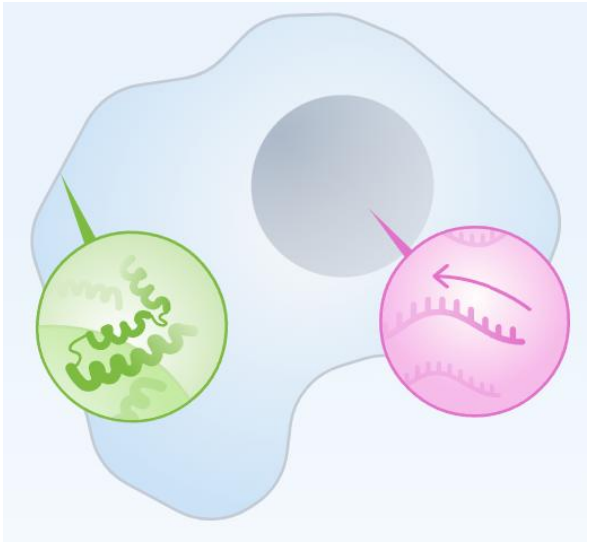
Probe-based

Reverse transcription-based

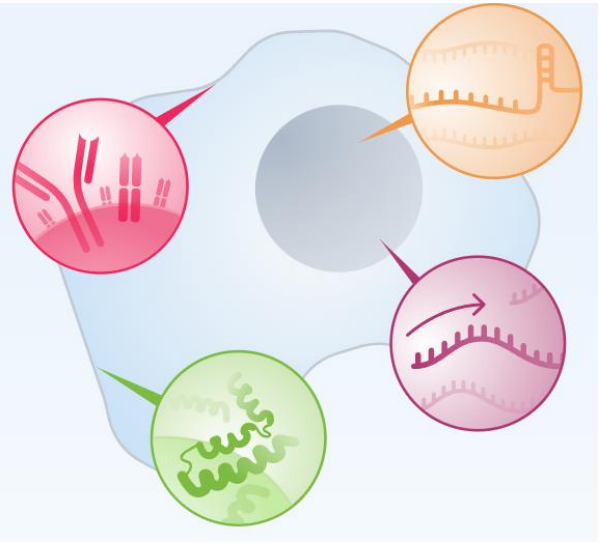
GEM-X Flex Gene Expression



GEM-X Universal 3’ Gene Expression



GEM-X Universal 5’ Gene Expression



Compatible samples

GEM-X Flex Gene Expression

GEM-X Universal 3’ Gene Expression

GEM-X Universal 5’ Gene Expression

Multiomic capabilities

GEM-X Flex Gene Expression

Cell surface protein

CRISPR screening

Intracellular protein

GEM-X Universal 3’ Gene Expression

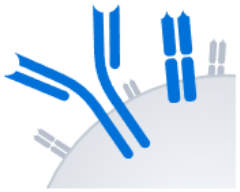
Cell surface protein

GEM-X Universal 5’ Gene Expression

Cell surface protein

Antigen specificity

CRISPR screening



TCR/BCR sequencing

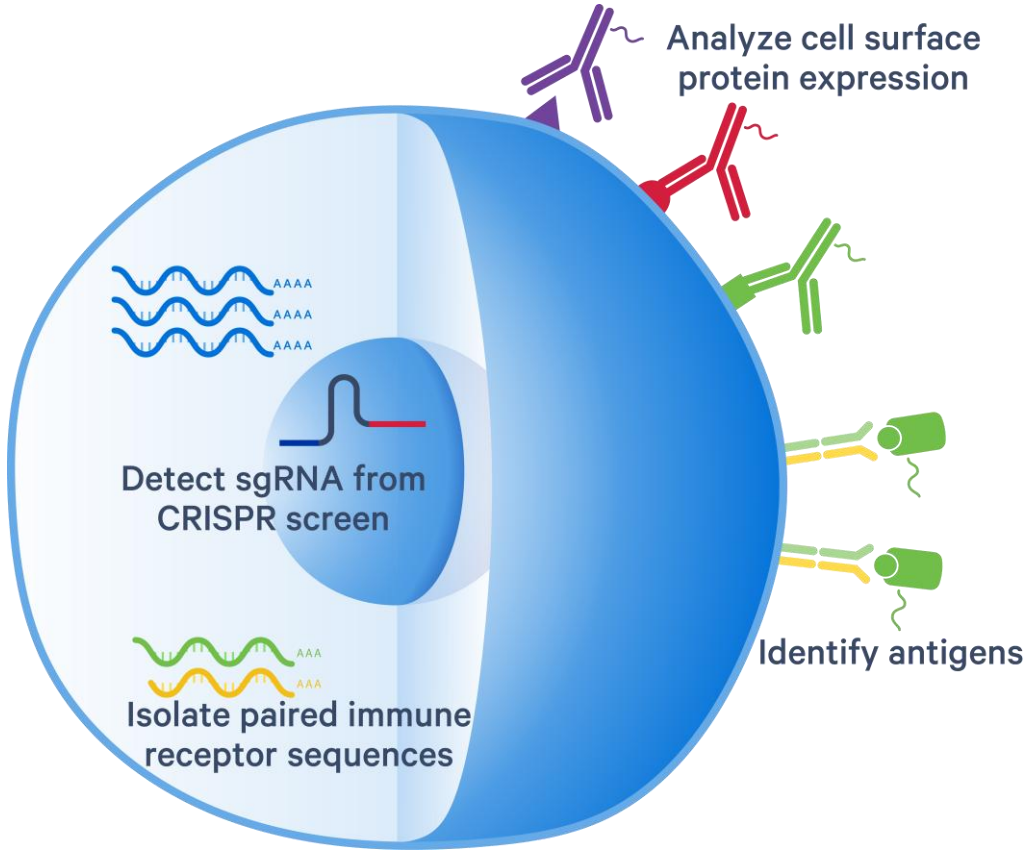
Intracellular protein

Feature Barcode Technology

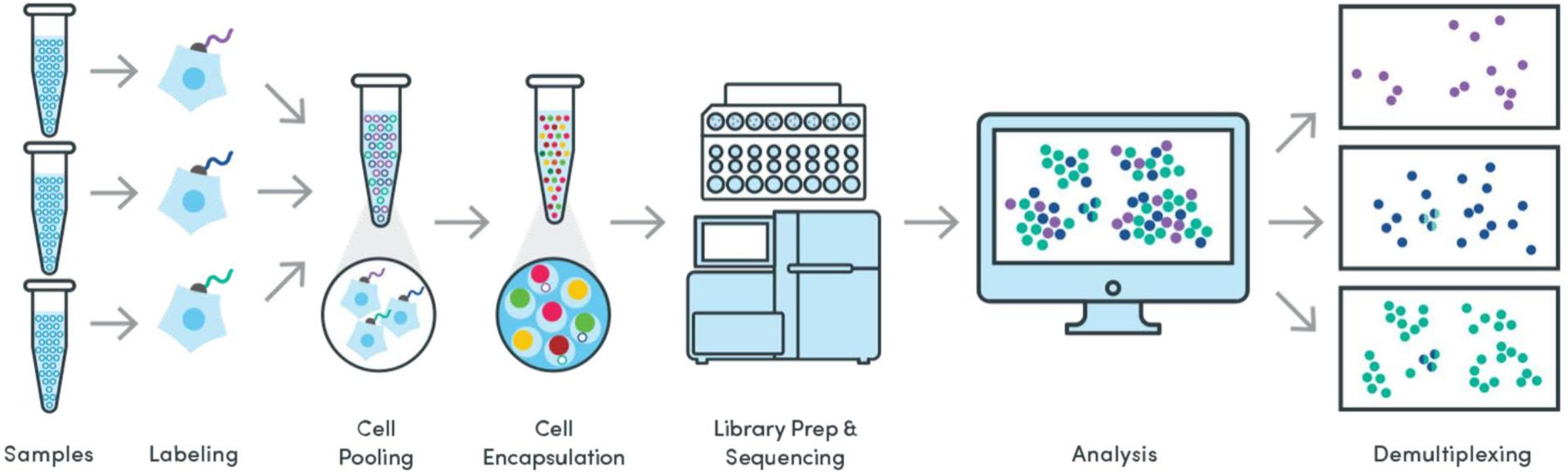
Antibody Capture (cell surface protein measurement): A method to label cell surface proteins using a specific protein binding molecule, such as an antibody conjugated to a Feature Barcode oligonucleotide.

CRISPR Guide Capture: A method to analyze gene expression changes caused by the presence of CRISPR perturbations. The assay captures the full transcriptome and transfected guide RNAs from the same cell to correlate changes in the transcriptome to CRISPR perturbations.

3' Cell Multiplexing: A method to label a given cell (or nuclei) sample with a molecular tag and subsequently pooling this sample with other labeled samples.

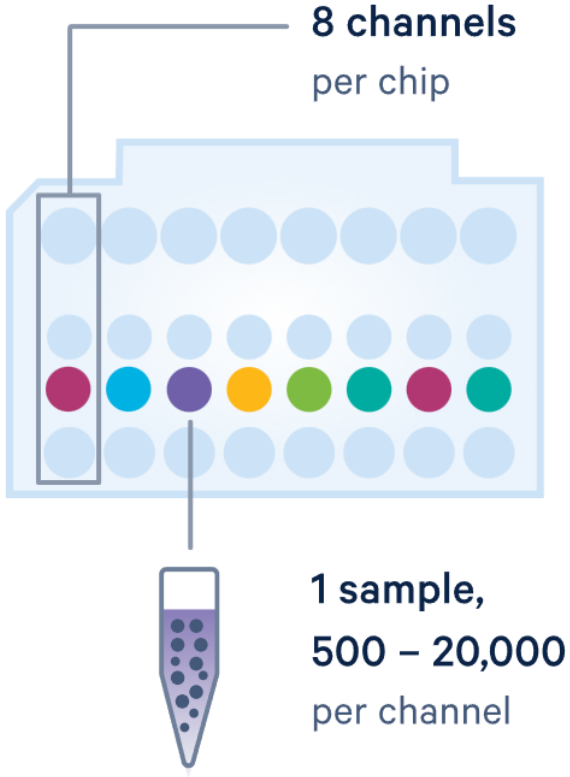


Feature Barcode Technology



Multiplexed GEM-X Gene Expression

Singleplex

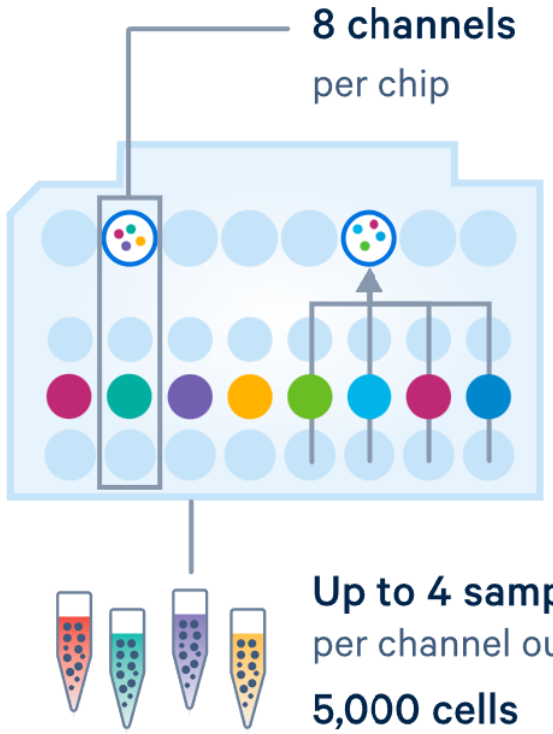


8 channels
per chip

1 sample,
500 – 20,000 cells
per channel

8 samples,
160,000 cells
per chip, maximum

On-chip multiplex



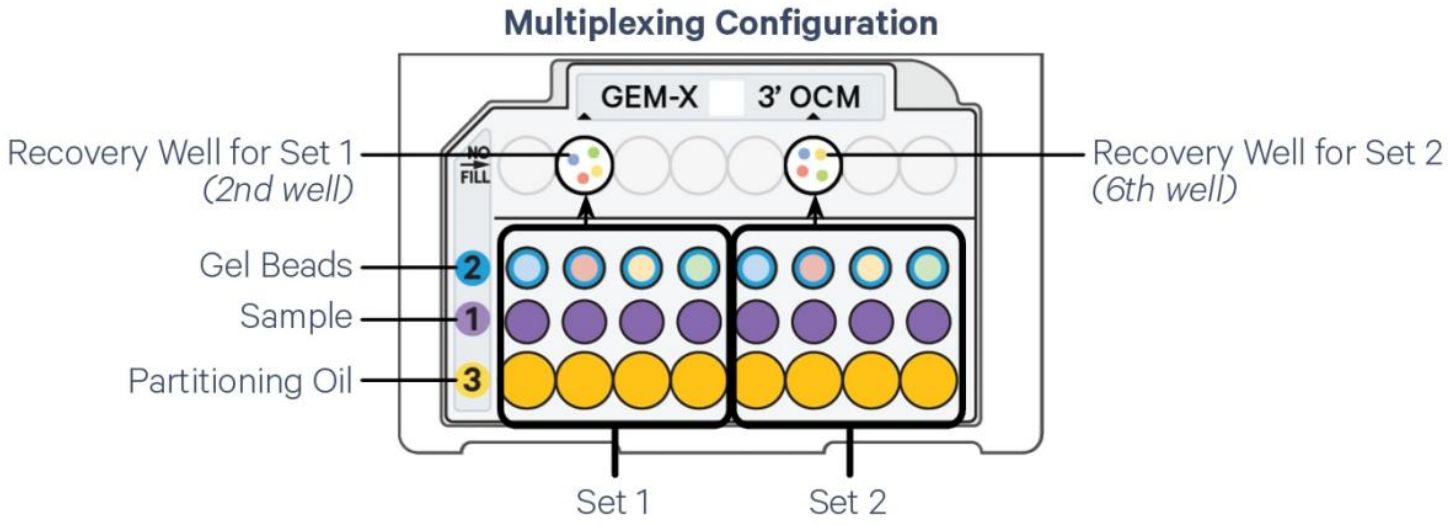
8 channels
per chip

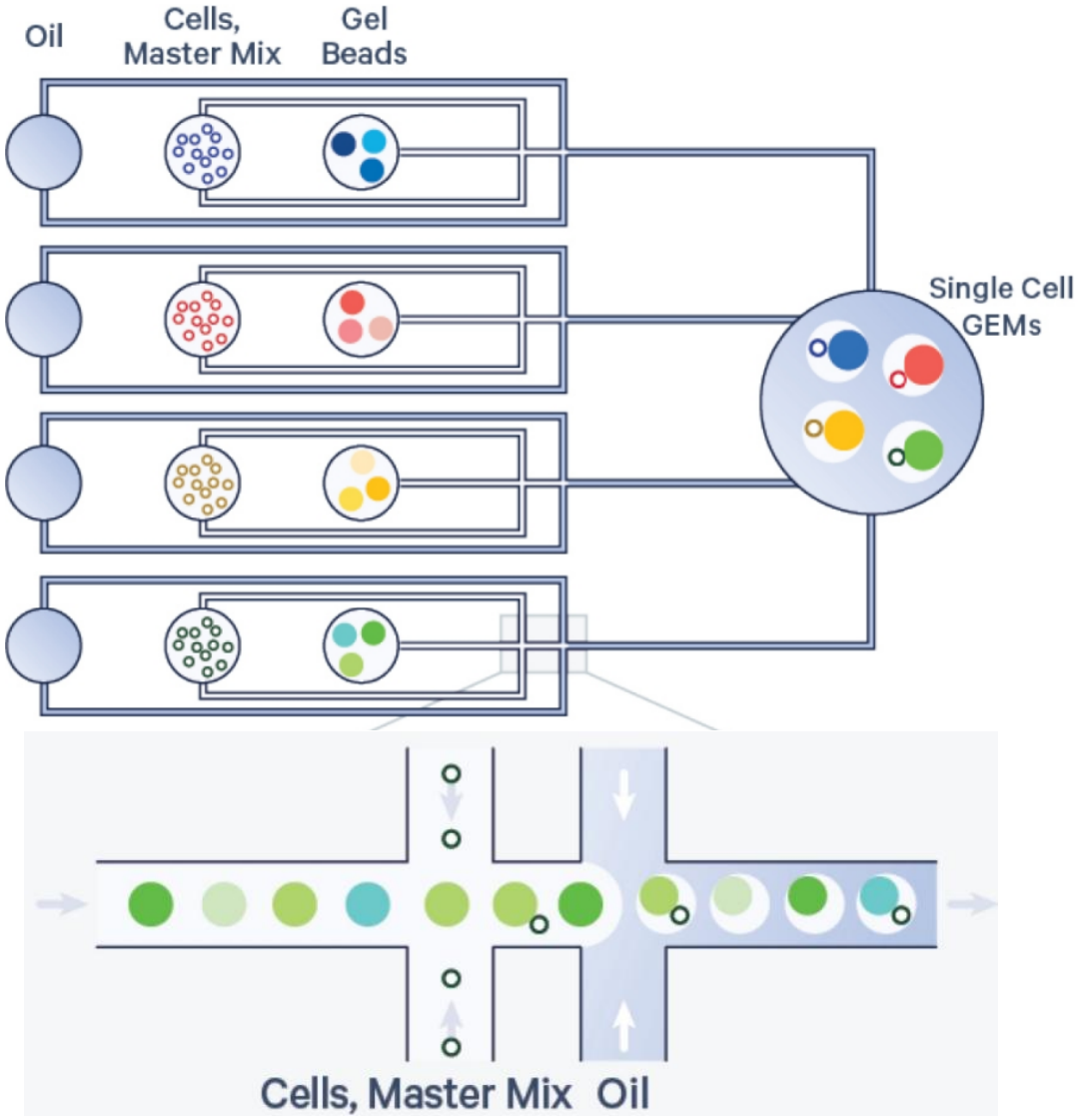
Up to 4 samples
per channel output

5,000 cells
per sample

8 samples,
40,000 cells
per chip, maximum

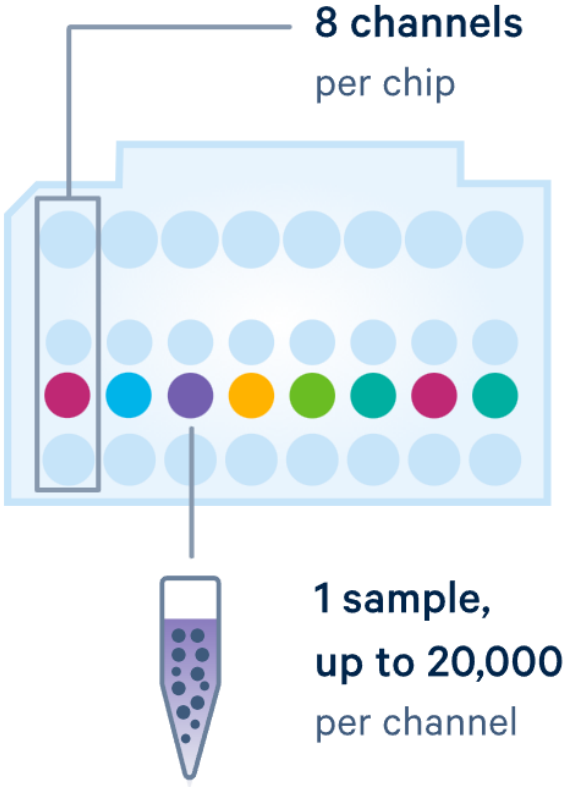
Chromium Single Cell - Gene Expression





Multiplexed GEM-X Gene Expression

Singleplex

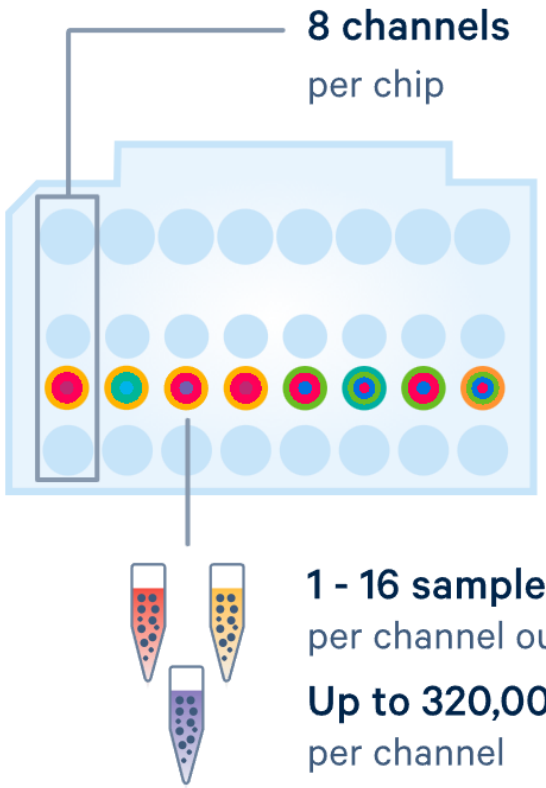


8 channels
per chip

1 sample,
up to 20,000 cells
per channel

8 samples,
160,000 cells
per chip, maximum

In-line multiplex



8 channels
per chip

1 - 16 samples
per channel output

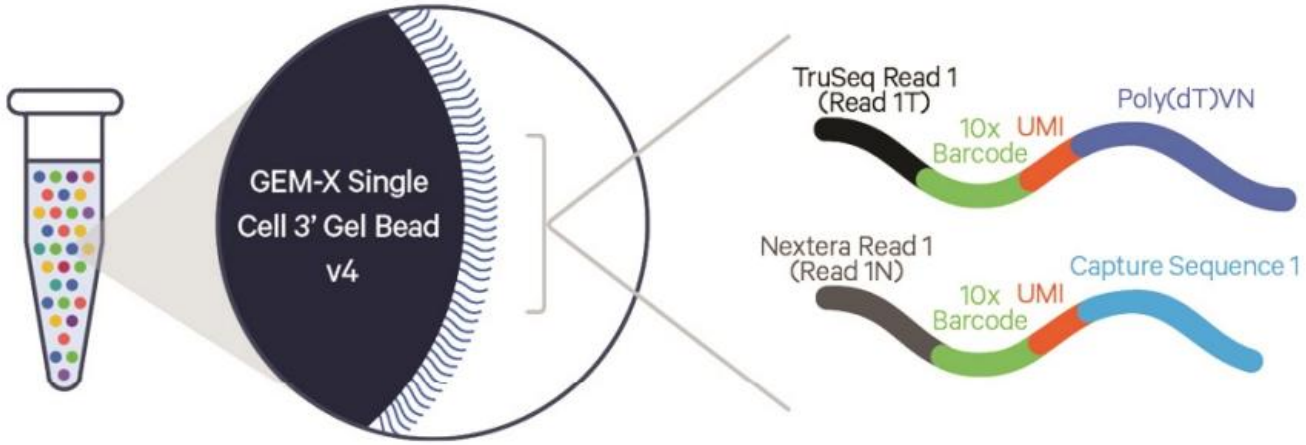
Up to 320,000 cells
per channel

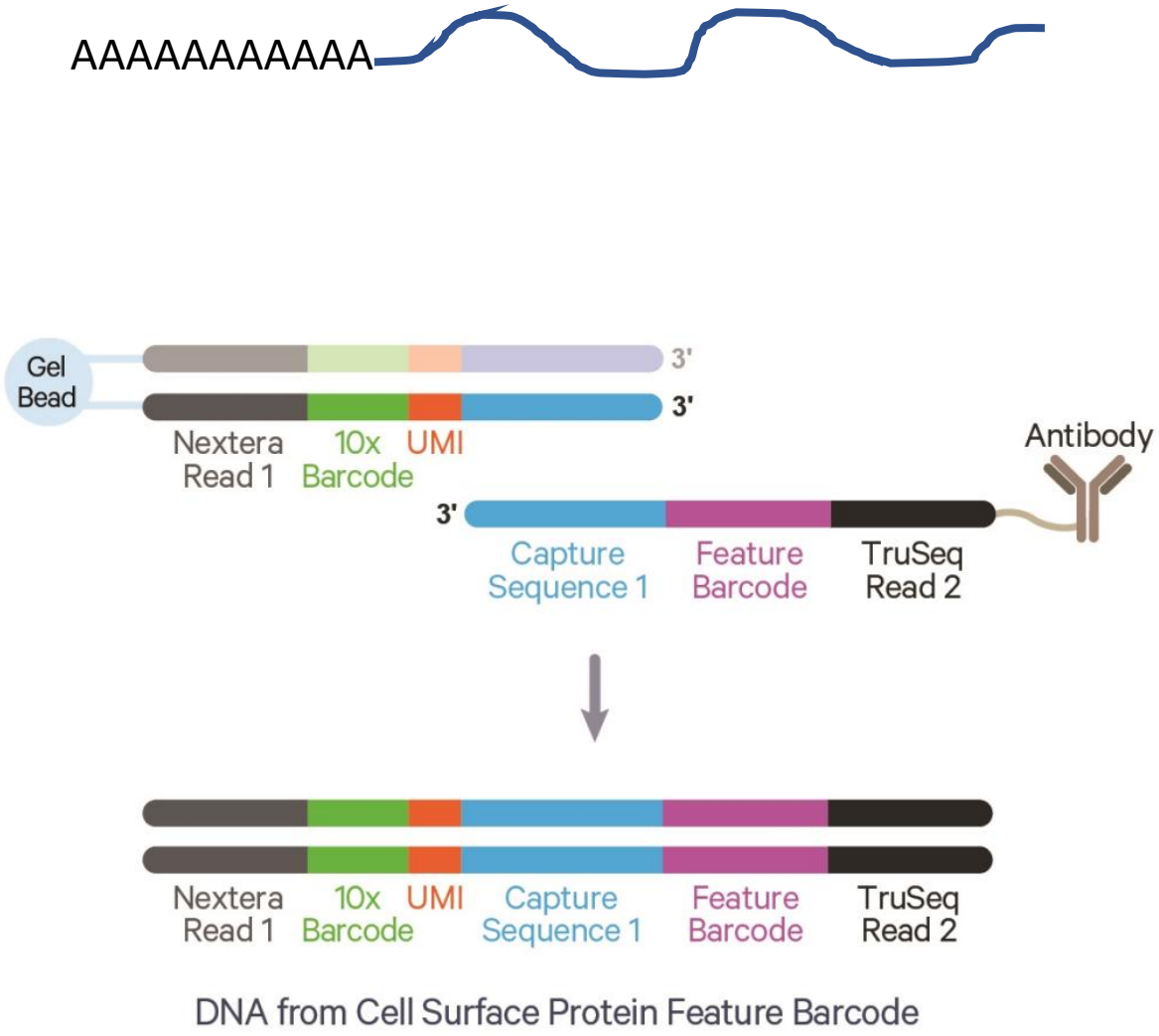
128 samples,
2,560,000 cells
per chip, maximum

GEM-X Gene Expression & Feature Barcode (e.g., Cell Surface Protein/ Antibody Capture)

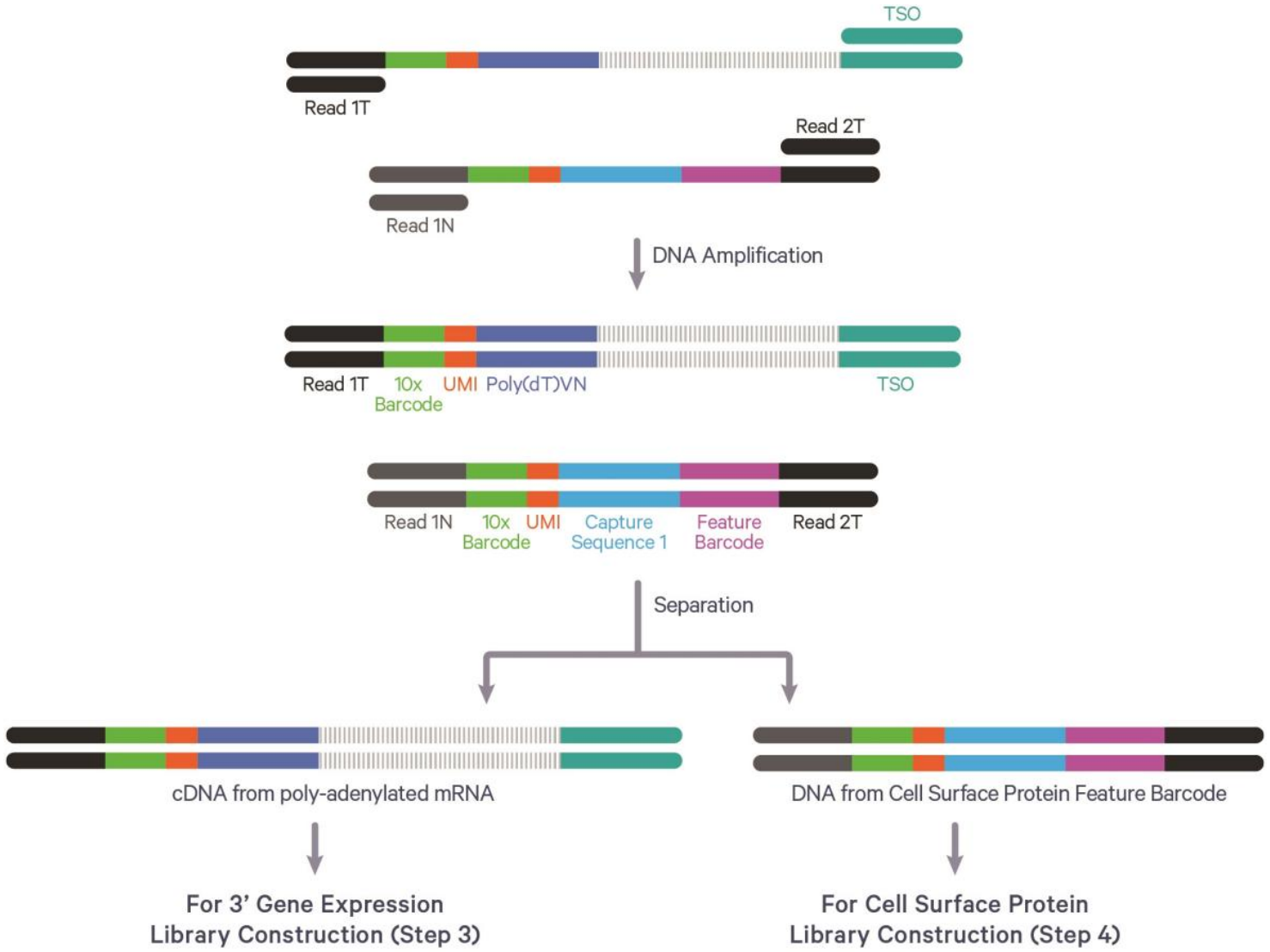
- GEM-X Universal 3’ Gene Expression & Feature Barcode
- GEM-X Flex Gene Expression & Feature Barcode

GEM-X 3’ Gene Expression & Feature Barcode

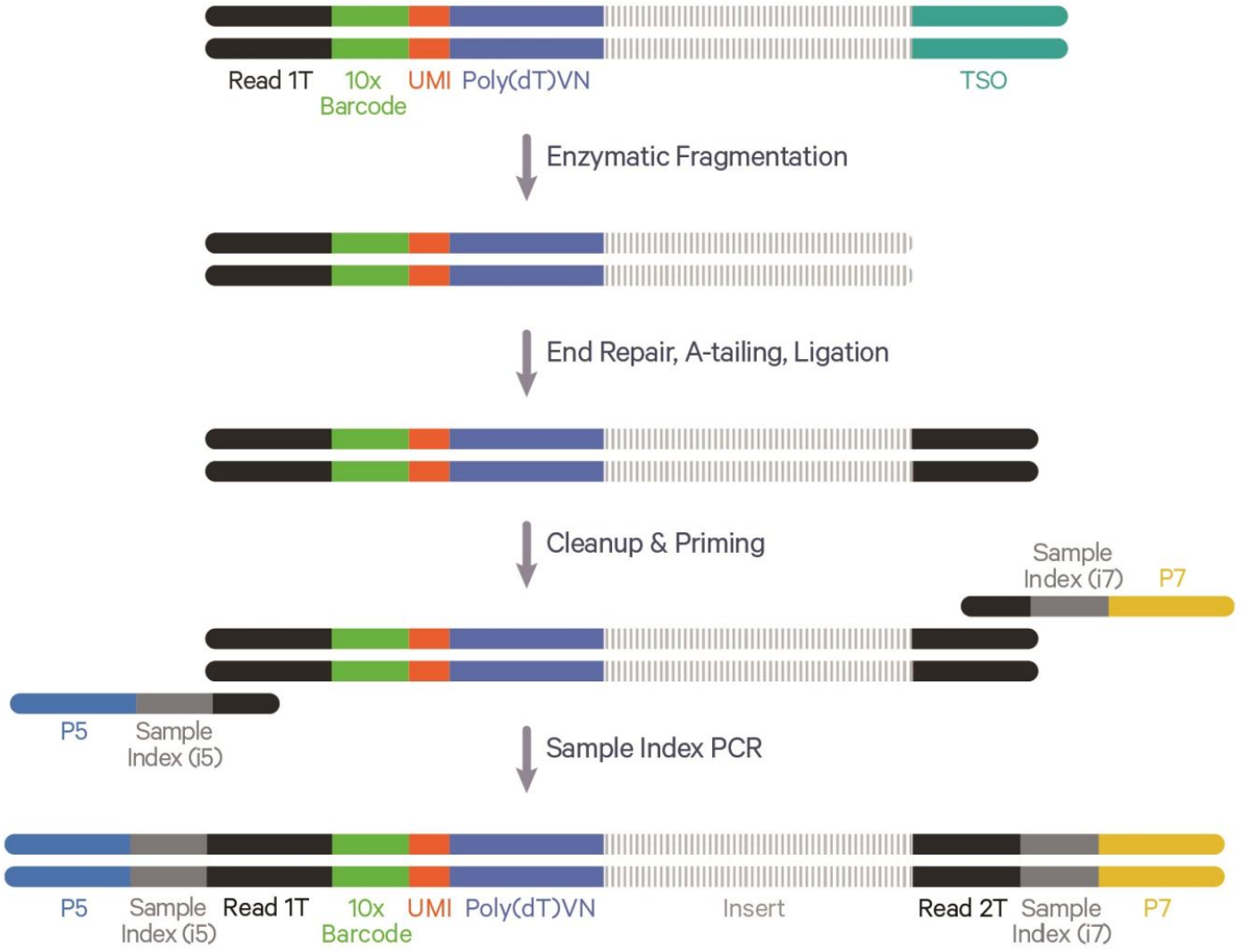




Library Construction of GEM-X 3’ Gene Expression & Feature Barcode



Library Construction of GEM-X 3’ Gene Expression & Feature Barcode

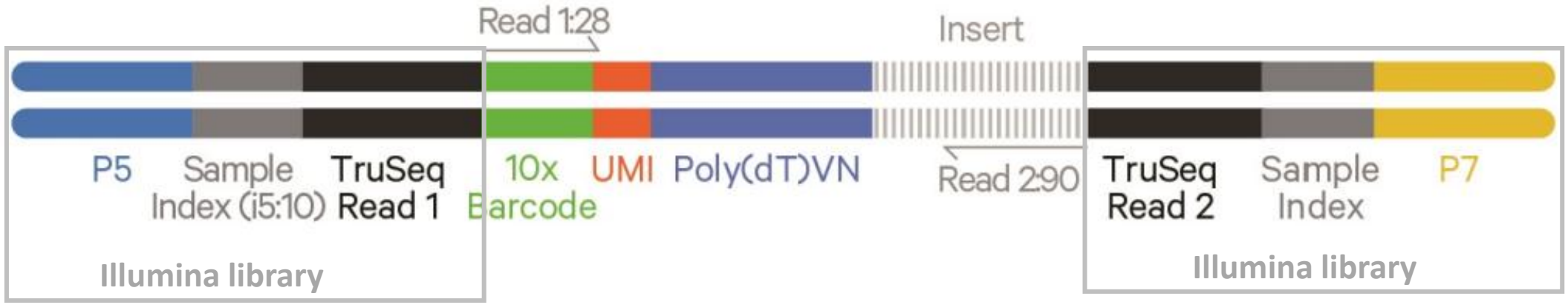


Library Construction of GEM-X 3’ Gene Expression & Feature Barcode

Chromium Single Cell 3' Gene Expression Library

Chromium Single Cell 3' Cell Surface Protein Library

Sequencing Library Structure



Index	Purpose
P5/P7	the library fragment to attach to the flow cell surface (Illumina)
Sample Index (i5/i7)	demultiplexing sequencing runs when running more than one sample per lane (Illumina)
Seq Read1/2	Illumina sequencing adaptors
10x Barcode **(16 bp)**	Cell Barcode (Bead specific) --- each **Cell**
UMI **(12 bp)**	12 nt Unique Molecular Identifier --- each **RNA molecule**
Capture Seq	The reverse complement of the sequence inserted into the DNA or RNA based Feature
Feature Barcode	Barcode for Capture Seq

Cell Ranger count– GEM-X 3’ or 5’

```
cellranger count --id=sampleID \
        --transcriptome=/path/to/refdata-gex-GRCh38-2024-A \
        --create-bam=true \
        --libraries=library.csv \
        --feature-ref=feature_ref.csv
```

```
--id <ID>
--description <TEXT>
--transcriptome <PATH>
--fastqs <PATH>
--project <TEXT>
--sample <PREFIX>
--lanes <NUMS>
--libraries <CSV>
--feature-ref <CSV>
--expect-cells <NUM>
--force-cells <NUM>
--create-bam <true|false>
```

```
--localcores <NUM>
--localmem <NUM>
--localvmem <NUM>
--mempercore <NUM>

--maxjobs <NUM>
--jobinterval <NUM>
--overrides <PATH>
```

```
--nosecondary
--r1-length <NUM>
--r2-length <NUM>
--include-introns <true|false>
--chemistry <CHEM>
```

Library CSV (comma delaminated)
(GEX + Antibody Capture + CRISPR Guide Capture)

fastqs	sample	library_type
/opt/foo/	GEX_sample3	Gene Expression
/opt/foo/	Ab_sample3	Antibody Capture
/opt/foo/	CRISPR_sample3	CRISPR Guide Capture

Feature Reference CSV (comma delaminated)

id	name	read	pattern	sequence	feature_type
CD3	CD3_TotalSeqB	R2	5PNNNNNNNNN(BC)	AACAAGACCCTTGAG	Antibody Capture
CD4	CD4_TotalSeqB	R2	5PNNNNNNNNN(BC)	TACCCGTAATAGCGT	Antibody Capture
CD8a	CD8a_TotalSeqB	R2	5PNNNNNNNNN(BC)	ATTGGCACTCAGATG	Antibody Capture
CD14	CD14_TotalSeqB	R2	5PNNNNNNNNN(BC)	GAAAGTCAAAGCACT	Antibody Capture
CD15	CD15_TotalSeqB	R2	5PNNNNNNNNN(BC)	ACGAATCAATCTGTG	Antibody Capture
CD16	CD16_TotalSeqB	R2	5PNNNNNNNNN(BC)	GTCTTTGTCAGTGCA	Antibody Capture
CD56	CD56_TotalSeqB	R2	5PNNNNNNNNN(BC)	GTTGTCCGACAATAC	Antibody Capture
CD19	CD19_TotalSeqB	R2	5PNNNNNNNNN(BC)	TCAACGCTTGGCTAG	Antibody Capture

Reference transcriptome

1. Downloadable pre-built references transcriptomes: https://www.10xgenomics.com/support/software/cell-ranger/downloads#reference-downloads

```
refdata-cellranger-arc-GRCh38-2020-A-2.0.0/
refdata-cellranger-arc-mm10-2020-A-2.0.0/
refdata-cellranger-GRCh38-3.0.0/
refdata-cellranger-medaka/
refdata-cellranger-mm10-3.0.0/
refdata-cellranger-mm10-3.0.0_Ssna1e2/
refdata-cellranger-mm10-3.0.0_Ssna1ko/
refdata-cellranger-vdj-GRCh38-alts-ensembl-5.0.0/
refdata-cellranger-vdj-GRCm38-alts-ensembl-5.0.0/
refdata-cellranger-zebrafish-GRCz11/
refdata-gex-GRCh38-2020-A/
refdata-gex-GRCh38-2024-A/
refdata-gex-GRCh38_and_GRCm39-2024-A/
refdata-gex-GRCh38-and-mm10-2020-A/
refdata-gex-GRCm39-2024-A/
refdata-gex-mm10-2020-A/
```

2. mkgtf & mkref

```
cellranger mkgtf ref_genome.gtf \
        ref_genome.filtered.gtf \
        --attribute=gene_biotype:protein_coding
```

```
cellranger mkref \
   --genome=ref_genome \
   --fasta=ref_genome.primary_assembly.fa \
   --genes=ref_genome.filtered.gtf
```

Cell Ranger GEM-X 3’ or 5’ data analysis pipeline

Stage 1: Pre-processing with Cell Ranger

- Raw Seq Data - fastq
- Demultiplexing - fastq
- Cell Barcode & UMI Filtering
- Mapping to Reference (STAR)
- Generate feature count matrix

Cell Ranger

Cloud Analysis

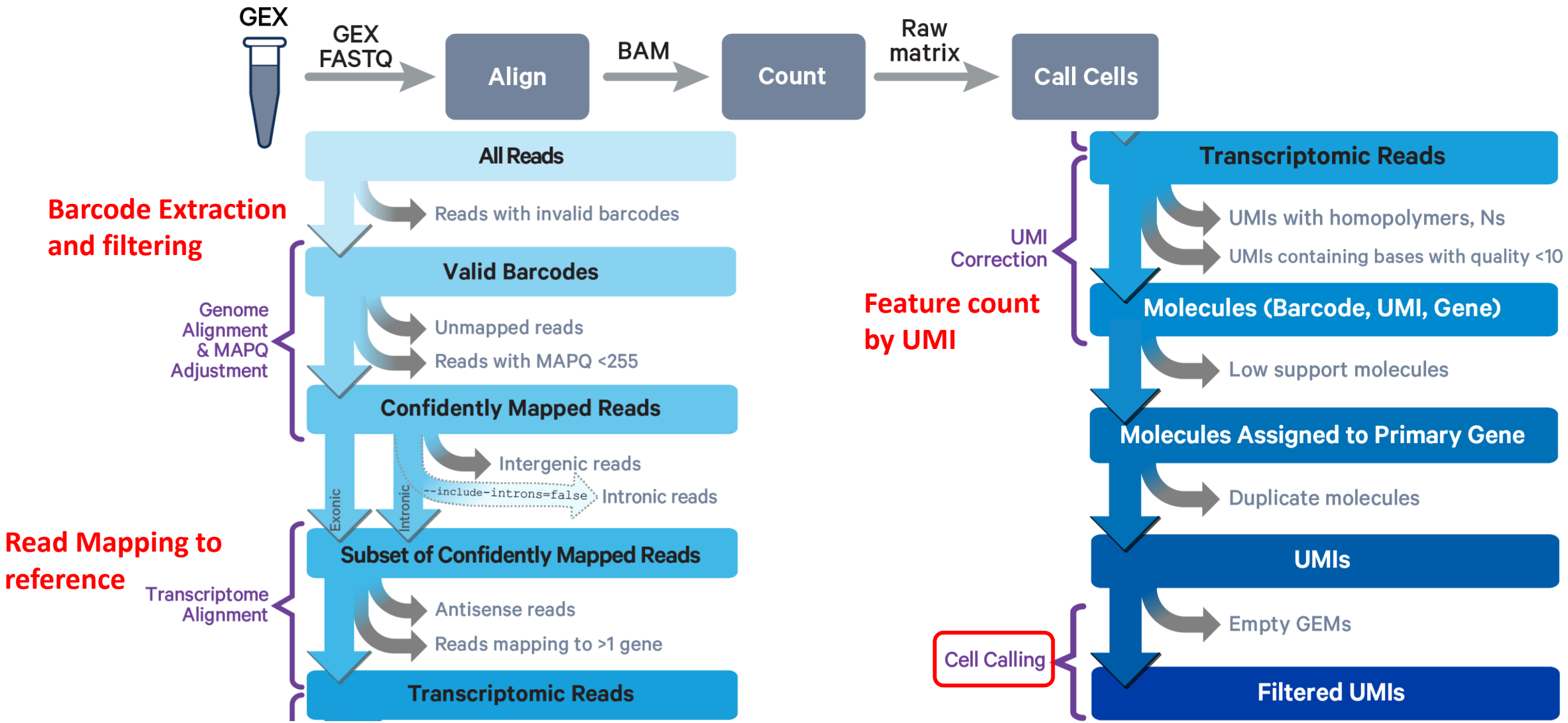
Loupe Browser

Stage 2: Downstream Analysis with Seurat or Scanpy

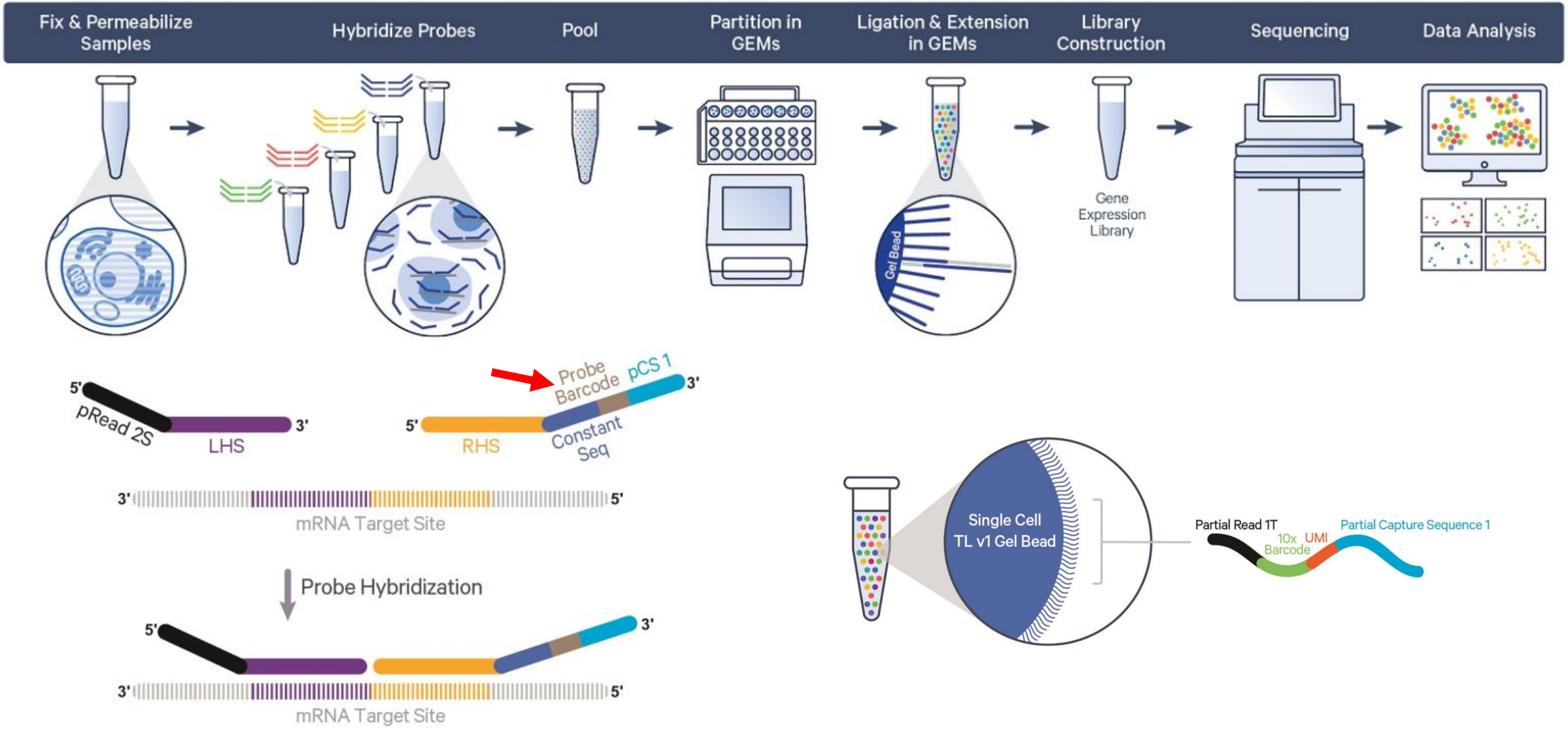
- Count matrix (cellranger output)
- Quality Control (QC) nfeature, UMI, MT genes
- Normalization and Scaling
- Feature Selection
- Dimensionality Reduction & Clustering

- DEGs
- RNA Velocity
- Trajectories
- Cell Chat

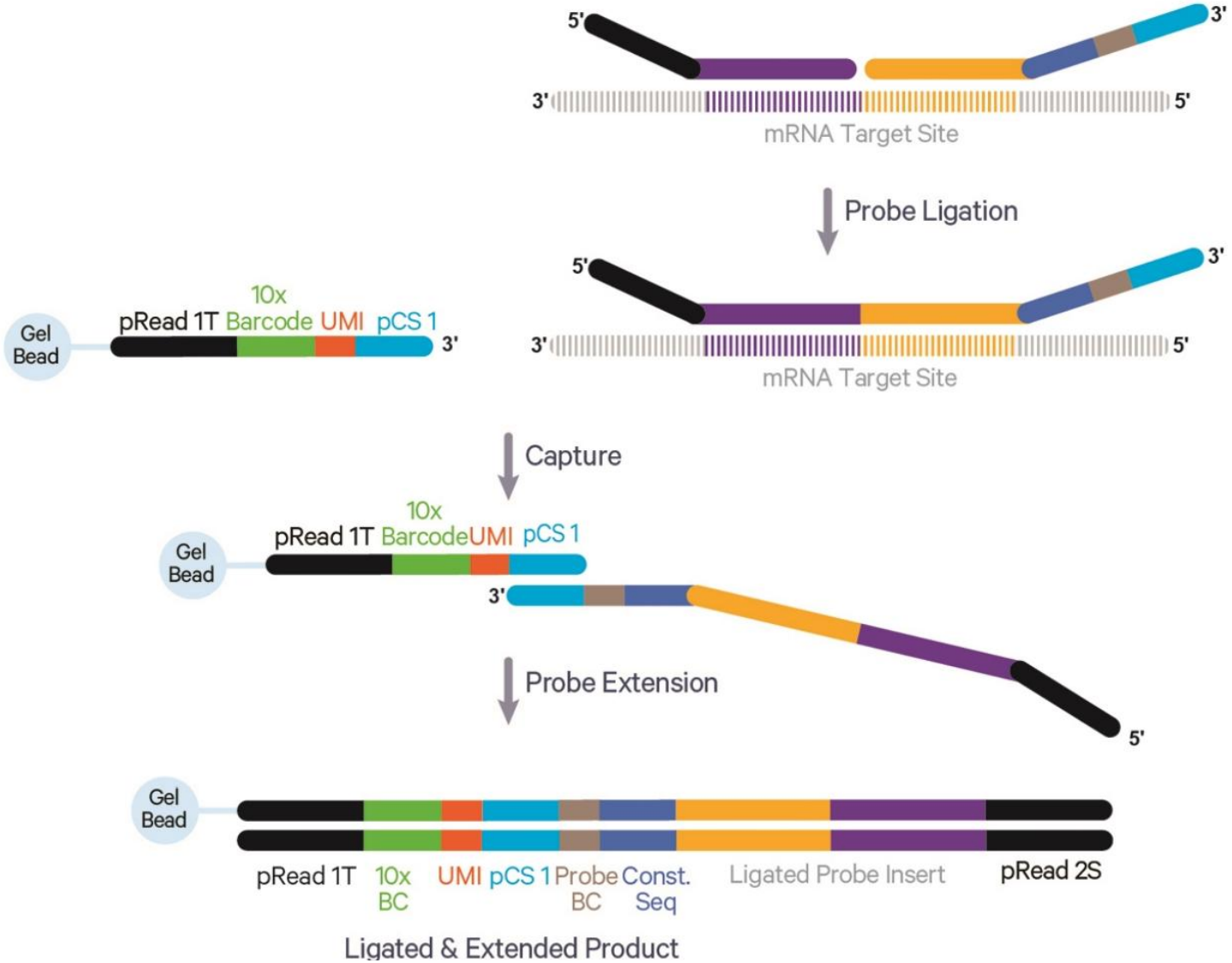
Data Analysis Algorithm



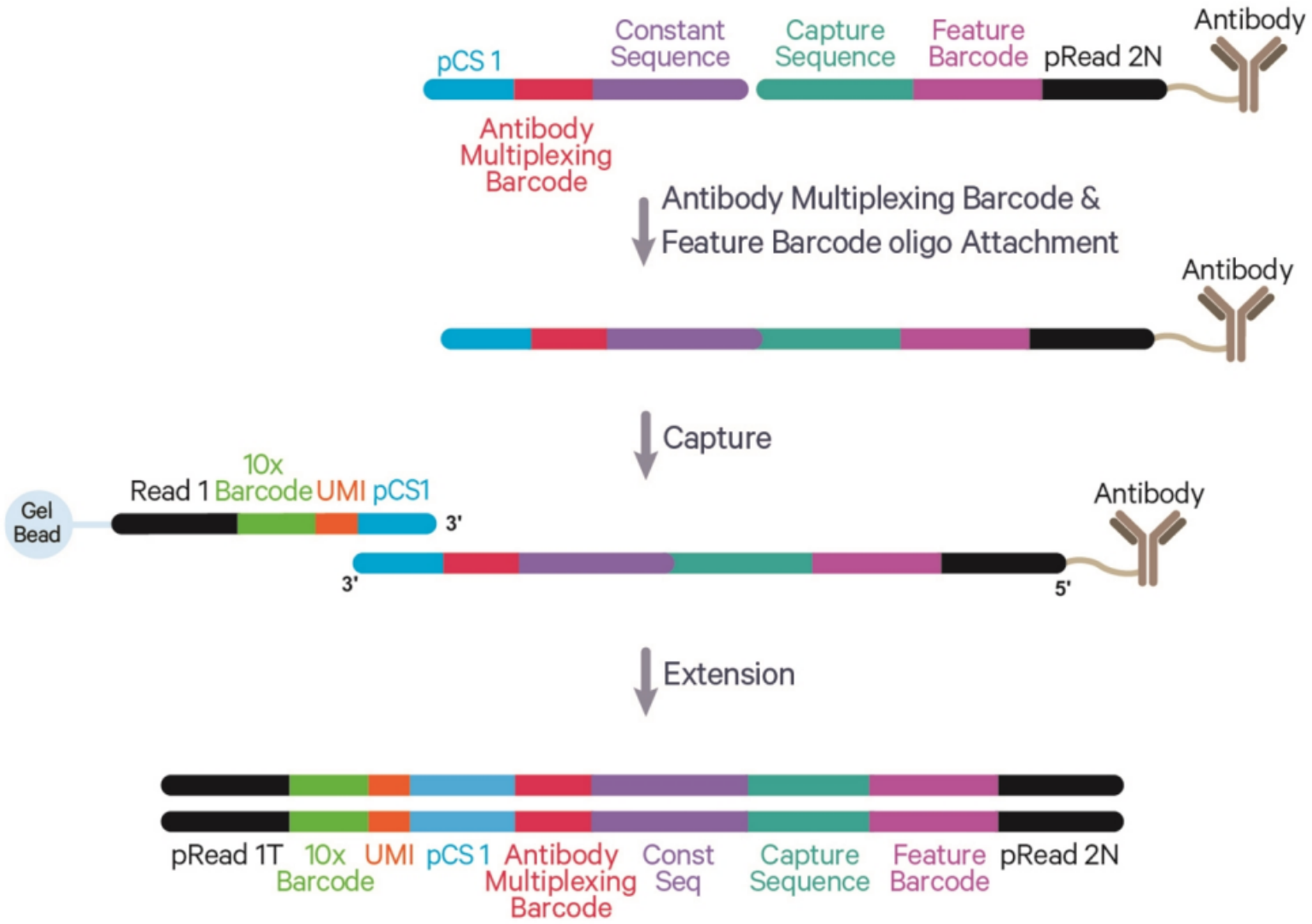
Flex Gene Expression & Feature Barcode



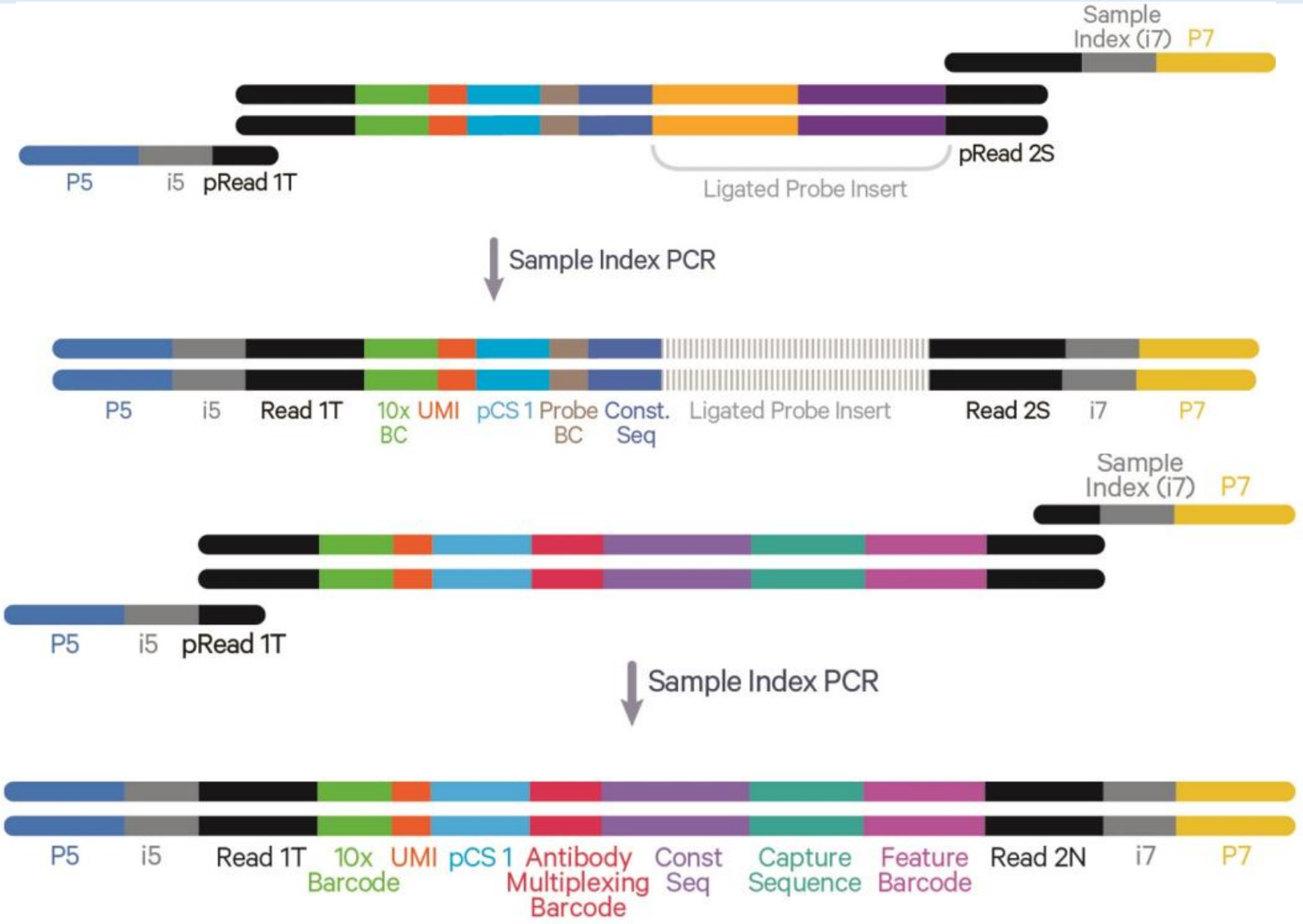
Library Construction of GEM-X Flex Gene Expression & Feature Barcode



Library Construction of GEM-X Flex Gene Expression & Feature Barcode



Library Construction of GEM-X Flex Gene Expression & Feature Barcode



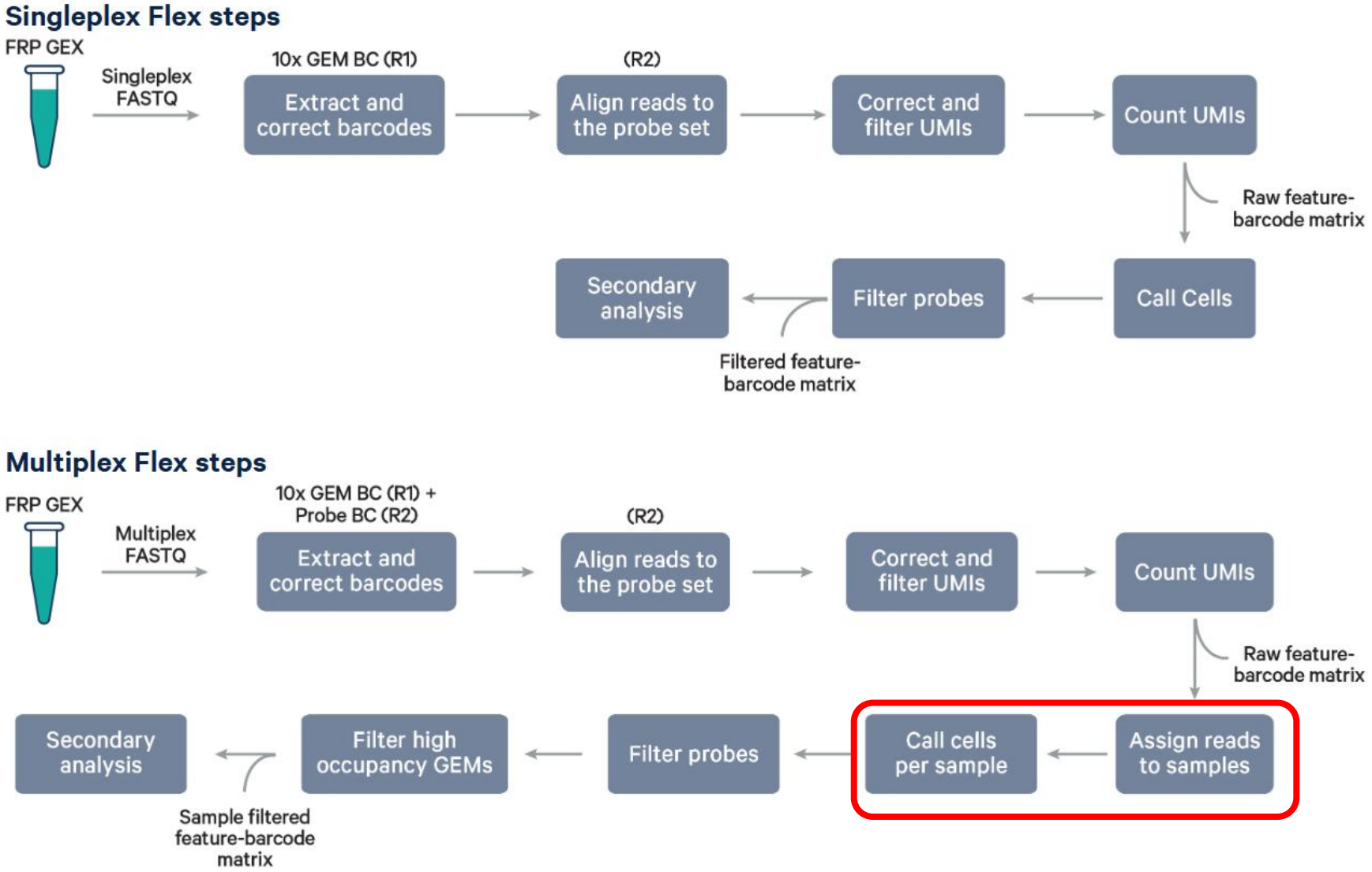
Cell Ranger count Pipeline – GEM-X Flex

```
cellranger multi --id=sampleID \
                 --csv=flex_multi_config.csv
```

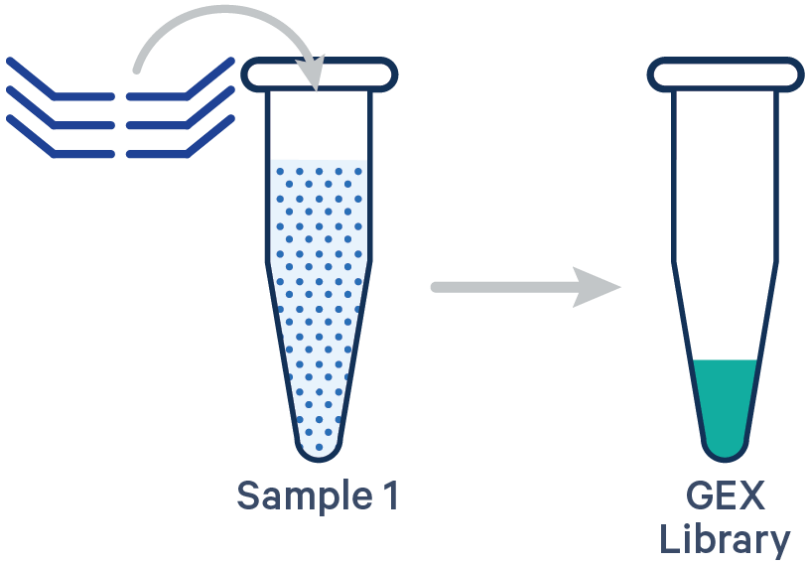
```
Usage: cellranger multi [OPTIONS] --id <ID> --csv <CSV>

Options:
      --id <ID>              A unique run id and output folder name [a-zA-Z0-9_-]+
      --description <TEXT>   Sample description to embed in output files [default: ]
      --csv <CSV>            Path of CSV file enumerating input libraries and analysis parameters
      --dry                  Do not execute the pipeline. Generate a pipeline invocation (.mro) file and stop
      --jobmode <MODE>       Job manager to use. Valid options: local (default), sge, lsf, slurm or path to a .template file. Search for help on "Cluster
                             Mode" at support.10xgenomics.com for more details on configuring the pipeline to use a compute cluster
      --localcores <NUM>     Set max cores the pipeline may request at one time. Only applies to local jobs
      --localmem <NUM>       Set max GB the pipeline may request at one time. Only applies to local jobs
      --localvmem <NUM>      Set max virtual address space in GB for the pipeline. Only applies to local jobs
      --mempercore <NUM>     Reserve enough threads for each job to ensure enough memory will be available, assuming each core on your cluster has at least
                             this much memory available. Only applies to cluster jobmodes
      --maxjobs <NUM>        Set max jobs submitted to cluster at one time. Only applies to cluster jobmodes
      --jobinterval <NUM>    Set delay between submitting jobs to cluster, in ms. Only applies to cluster jobmodes
      --overrides <PATH>     The path to a JSON file that specifies stage-level overrides for cores and memory. Finer-grained than --localcores, --mempercore
                             and --localmem. Consult https://10xgen.com/resource-override for an example override file
      --output-dir <PATH>    Output the results to this directory
      --uiport <PORT>        Serve web UI at http://localhost:PORT
      --disable-ui           Do not serve the web UI
      --noexit               Keep web UI running after pipestance completes or fails
      --nopreflight          Skip preflight checks
  -h, --help                 Print help
```

Cell Ranger analysis pipeline – GEM-X Flex



Singleplex Flex, 1 Probe Barcode

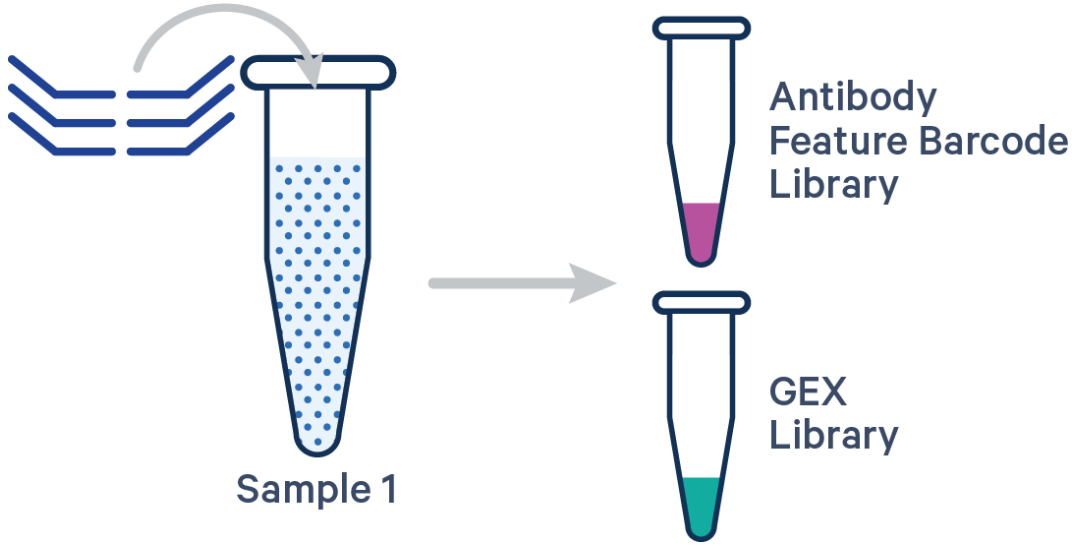



```
[gene-expression]
reference,/path/to/refdata-gex-GRCh38-2024-A
probe-set,/path/to/Transcriptome_Probe_Set.csv
create-bam,false

[libraries]
fastq_id,fastqs,feature_types
flex_gex,/path/to/fastqs,Gene Expression
```

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

Singleplex Flex with Antibody Capture, 1 Probe Barcode

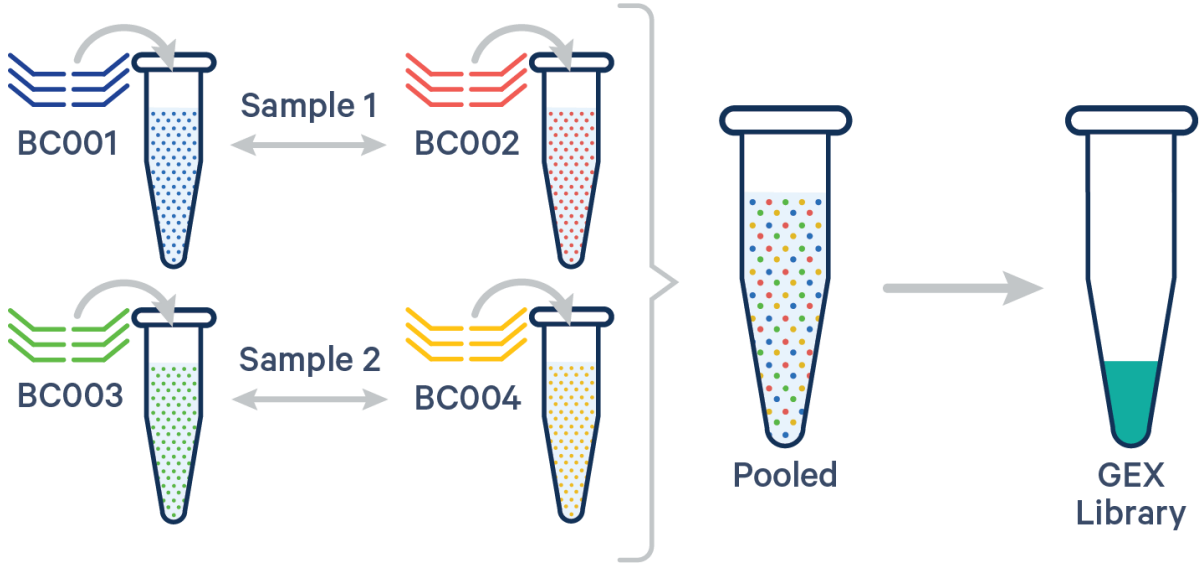



```
[gene-expression]
reference,/path/to/refdata-gex-GRCh38-2024-A
probe-set,/path/to/Transcriptome_Probe_Set.csv
create-bam,false

[libraries]
fastq_id,fastqs,feature_types
flex_gex,/path/to/fastqs,Gene Expression
flex_ab,/path/to/fastqs,Antibody Capture

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

Multiplex Flex, multiple Probe Barcodes/sample

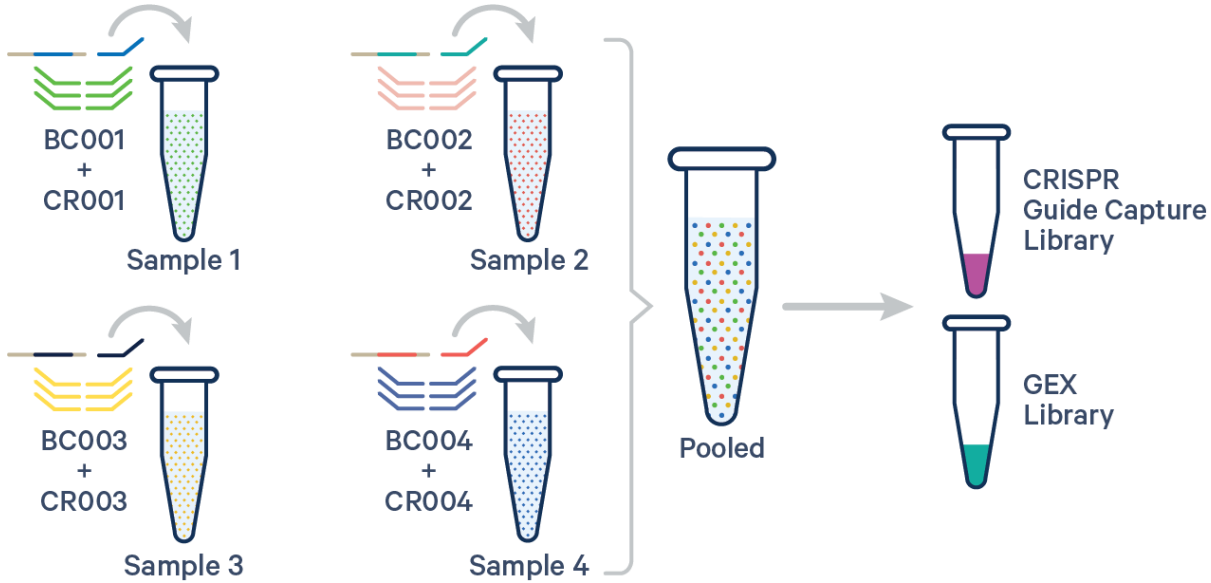



```
[gene-expression]
reference,/path/to/refdata-gex-GRCh38-2024-A
probe-set,/path/to/Transcriptome_Probe_Set.csv
create-bam,false

[libraries]
fastq_id,fastqs,feature_types
flex_gex,/path/to/fastqs,Gene Expression

[samples]
sample_id,probe_barcode_ids,description
sample1,BC001|BC002,Control
sample2,BC003|BC004,Treated
```

Multiplex Flex and CRISPR, 1 Probe Barcode pair/sample




```
[gene-expression]
reference,/path/to/refdata-gex-GRCh38-2024-A
probe-set,/path/to/Transcriptome_Probe_Set.csv
create-bam,false

[libraries]
fastq_id,fastqs,feature_types
flex_gex,/path/to/fastqs,Gene Expression
flex_cr,/path/to/fastqs,CRISPR Guide Capture

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

[samples]
sample_id,probe_barcode_ids
sample1,BC001+CR001
sample2,BC002+CR002
sample3,BC003+CR003
sample4,BC004+CR004
```

Cell Ranger commands

Cell Ranger 9.0.1 (Feb 6, 2025)

- **cellranger count**
  – Count gene expression and/or feature barcode reads from a single sample and GEM well: aligning reads, filtering low-quality cells, counting barcoded reads, and generating a feature-barcode matrix
- **cellranger multi**
  – analyzes multiplexed or combined 10x Genomics library types from a single GEM well, including Gene Expression, V(D)J, and Feature Barcode
- **cellranger mkref**
  – Prepare a custom reference for use with 10x analysis software

Primary

- **cellranger aggr**
  – Aggregate data from multiple Cell Ranger runs
- **cellranger annotate**
  – Annotate cell-types from outputs of a CellRanger run
- **cellranger reanalyze**
  – Reanalyze data using tunable parameters (dimensionality reduction, clustering, etc)

Secondary

cellranger aggr (merges multiple runs)

(After cellranger count)
After individual count processing for each run, aggregate the output results (molecule_info.h5)

```
$ cellranger aggr --id=run_ID \
                  --csv=libraries.csv \
                  --normalize=mapped
```

A unique run id and output folder name

libraries.csv

```
Sample_ID,molecule_h5
Sample1, Sample1/outs/molecule_info.h5
Sample2, Sample2/outs/molecule_info.h5
```

cellranger annotate (Cell type annotation)

(After cellranger count)
After individual count processing for each run, aggregate the output results (filtered_feature_bc_matrix.h5)

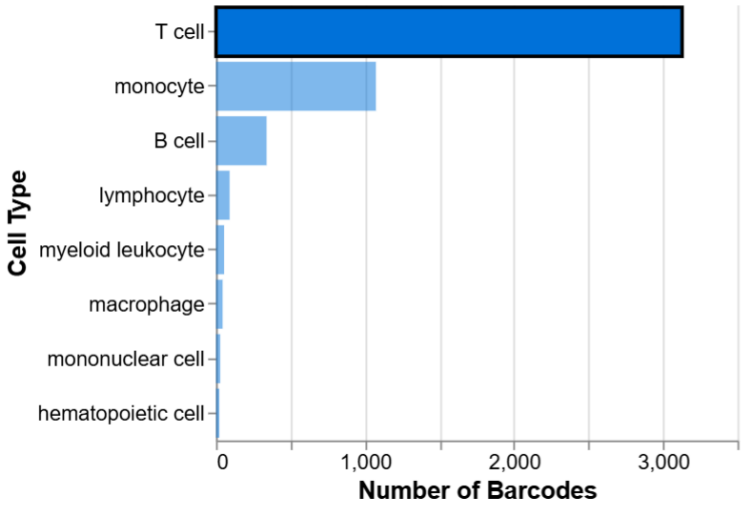
```
cellranger annotate --id=sampleID \
   --matrix=filtered_feature_bc_matrix.h5\
   --cell-annotation-model=auto \
   --tenx-cloud-token-path=/path/to/10xcloud_token.json
```

Cell type annotation is currently in beta and relies on the Chan Zuckerberg CELL by GENE reference. As this reference is community-driven it may not cover all tissue types, which could impact your results. For further details, please visit the 10x support site.

Annotation was performed using a model that was co-developed by 10x Genomics and the Cellarium AI Lab at the Data Sciences Platform of the Broad Institute.

cellranger annotate

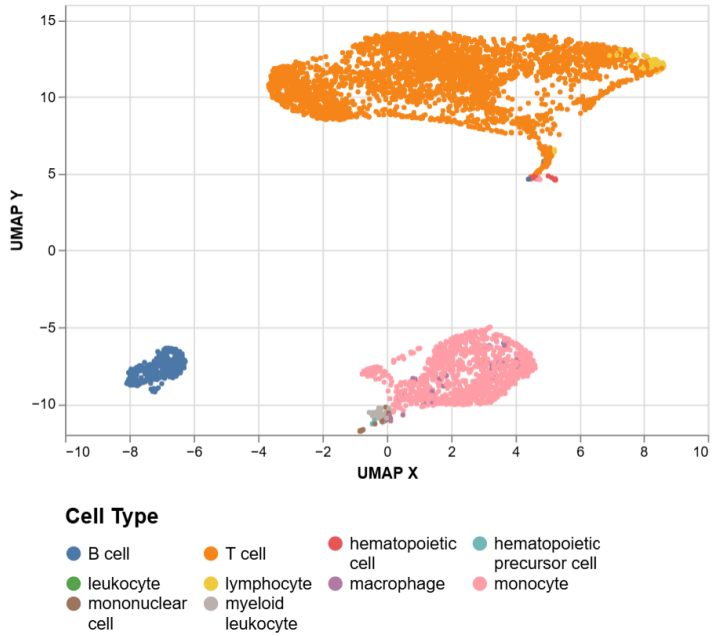
Cell Type Composition



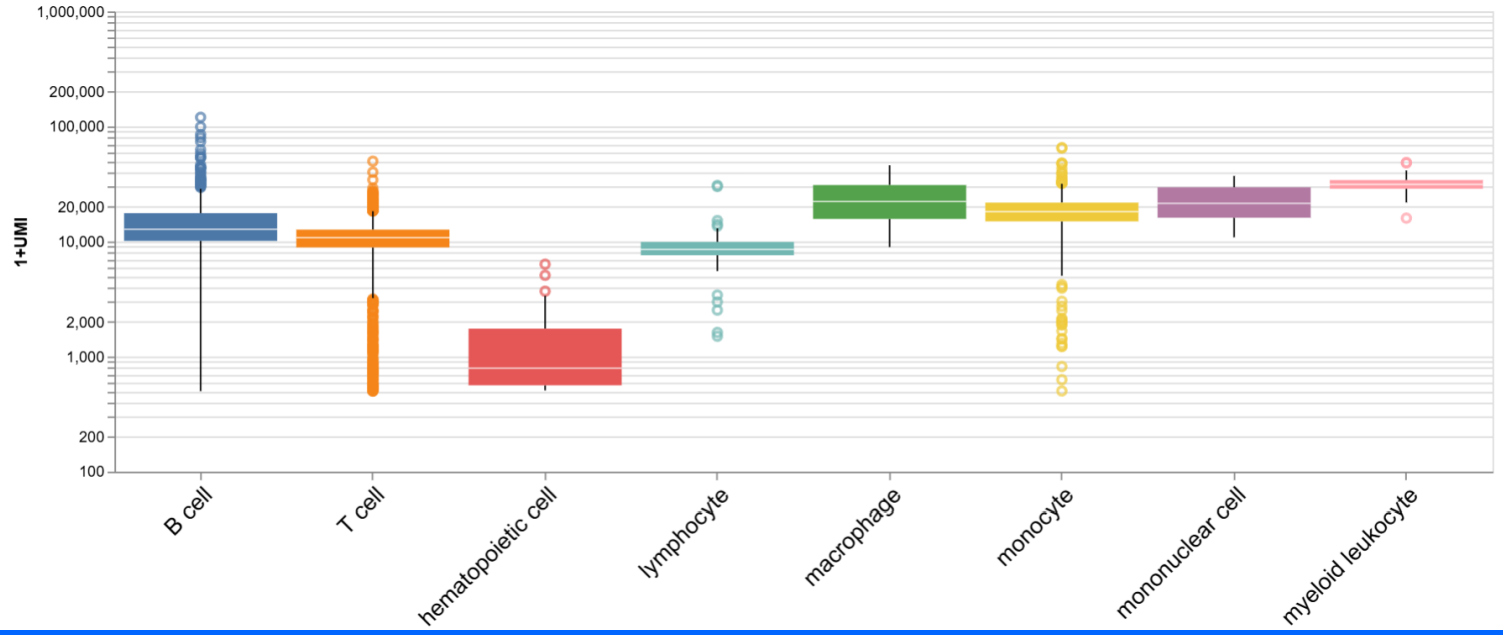
T Cell

Sub-type	Fraction BCs
mature T cell	50.22%
CD4-positive, alpha-beta …	19.44%
CD4-positive helper T cell	10.65%
T follicular helper cell	6.65%

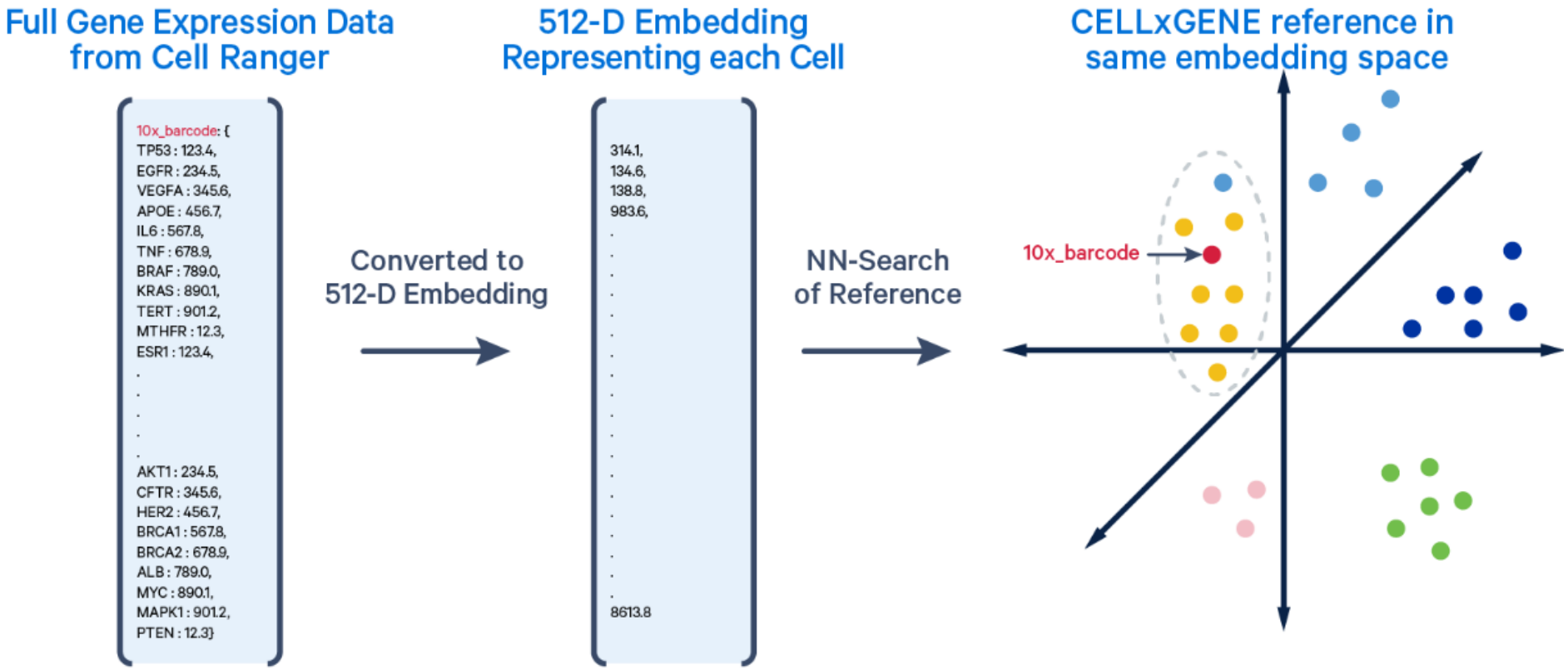
UMAP projection of cell types



UMI distribution by cell type



cellranger annotate(Cell type annotation)



The algorithm generates an embedding for each cell barcode. The gene expression profile of each cell barcode being analyzed is projected into the same embedding. To classify a cell, the algorithm performs an approximate nearest-neighbor (ANN) search, identifying the 500 most similar cells in the reference set based on these embeddings. The most common cell type among these nearest neighbors is then assigned to the query cell. Models may differ in their inputs (e.g., all genes vs specific genes) and how they calculate embeddings. The models are described in more detail below.

cellranger reanalyze

(After cellranger count)
reruns secondary analysis performed on the feature-barcode matrix (dimensionality reduction, clustering and visualization) using different parameter settings (filtered_feature_bc_matrix.h5)

```
$ cellranger reanalyze --id=runID_reanalysis \
                       --matrix=path-to-output/outs/filtered_feature_bc_matrix.h5 \
                       --params=reanalysis.csv
```

A unique run id and output folder name

reanalysis.csv

```
num_principal_comps,20
max_clusters,30
num_pca_bcs,50000
num_pca_genes,3000
```

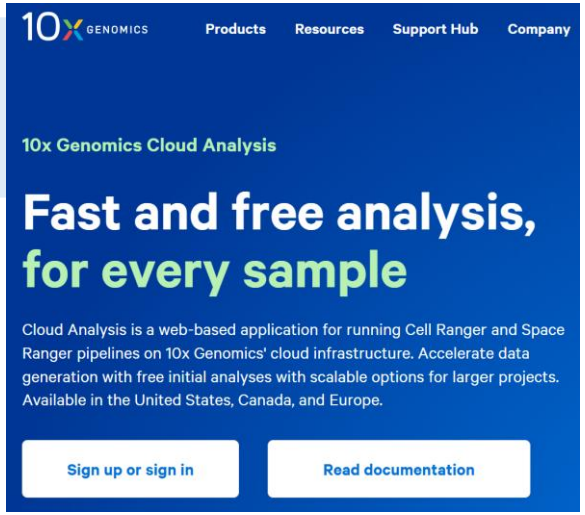
Randomly subset data to N barcodes when computing PCA projection

Subset data to the top N genes (ranked by normalized dispersion) when computing PCA

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

10X Cloud Analysis (demo)

https://cloud.10xgenomics.com/



What you can do	Included for free[1]
Run 10x analysis pipelines	**Run every 10x dataset you upload for free** For every dataset, run and re-run the pipeline up to 5 times each, up to a total of 50 free analyses per calendar month[2].
Store data	**90 days of free storage for all data**
Download analysis outputs	**Download analysis output files for free** For every analysis, we provide one free download for all files.
Download FASTQ files	**Download each uploaded FASTQ file once for free**

10X Cloud Analysis - 3’ Gene Expression

Datasets (Showing 9 datasets)	**Product**	**Species**	**Sample type**	**Cells or nuclei**	**Preservation**
5k Human PBMCs (Donor 2)	Universal 3' Gene Expression v4	Human	Blood	Cells	Cryopreserved

Peripheral blood mononuclear cells (PBMCs) from a healthy human male donor, aged 18-35 were obtained by 10x Genomics from Cellular Technologies Limited.

Gene Expression libraries were generated as described in the Chromium GEM-X Single Cell 3' Reagent Kits v4 User Guide (CG000731) and sequenced on an Illumina NovaSeq 6000 with approximately 34,000 read pairs per cell for the Gene Expression library.

Paired-end, dual indexing libraries were sequenced with this configuration:

- 28 cycles Read 1
- 10 cycles i7
- 10 cycles i5
- 90 cycles Read 2

Product

Universal 3' Gene Expression

Software

Cell Ranger

Subpipeline

cellranger multi

10x Instrument(s)

Chromium X Series

Output files – cellranger count (singleplex)

- **web_summary.html** – Web format QC report
- **filtered_feature_bc_matrix (for further analysis, ex: Seurat)**
  - barcodes.tsv.gz - cell level barcodes seen in this sample
  - features.tsv.gz - list of quantitated features (usually Ensembl genes)
  - matrix.mtx.gz - (sparse) matrix of counts for cells and features
- **possorted_genome_bam.bam** – BAM file of mapped reads
- **molecule_info.h5** – Details of the cell barcodes (for data merging - cellranger aggr)
- **cloupe.cloupe** – Analysis data for Loupe Cell browser

Output files – cellranger multi (per_sample_outs)

- **web_summary.html** – Web format QC report
- **sample_filtered_feature_bc_matrix (for further analysis, ex: Seurat)**
  - barcodes.tsv.gz - cell level barcodes seen in this sample
  - features.tsv.gz - list of quantitated features (usually Ensembl genes)
  - matrix.mtx.gz - (sparse) matrix of counts for cells and features
- **sample_alignments.bam** – BAM file of mapped reads
- **sample_molecule_info.h5**– Details of the cell barcodes (for data merging - cellranger aggr)
- **sample_cloupe.cloupe** – Analysis data for Loupe Cell browser

Output files

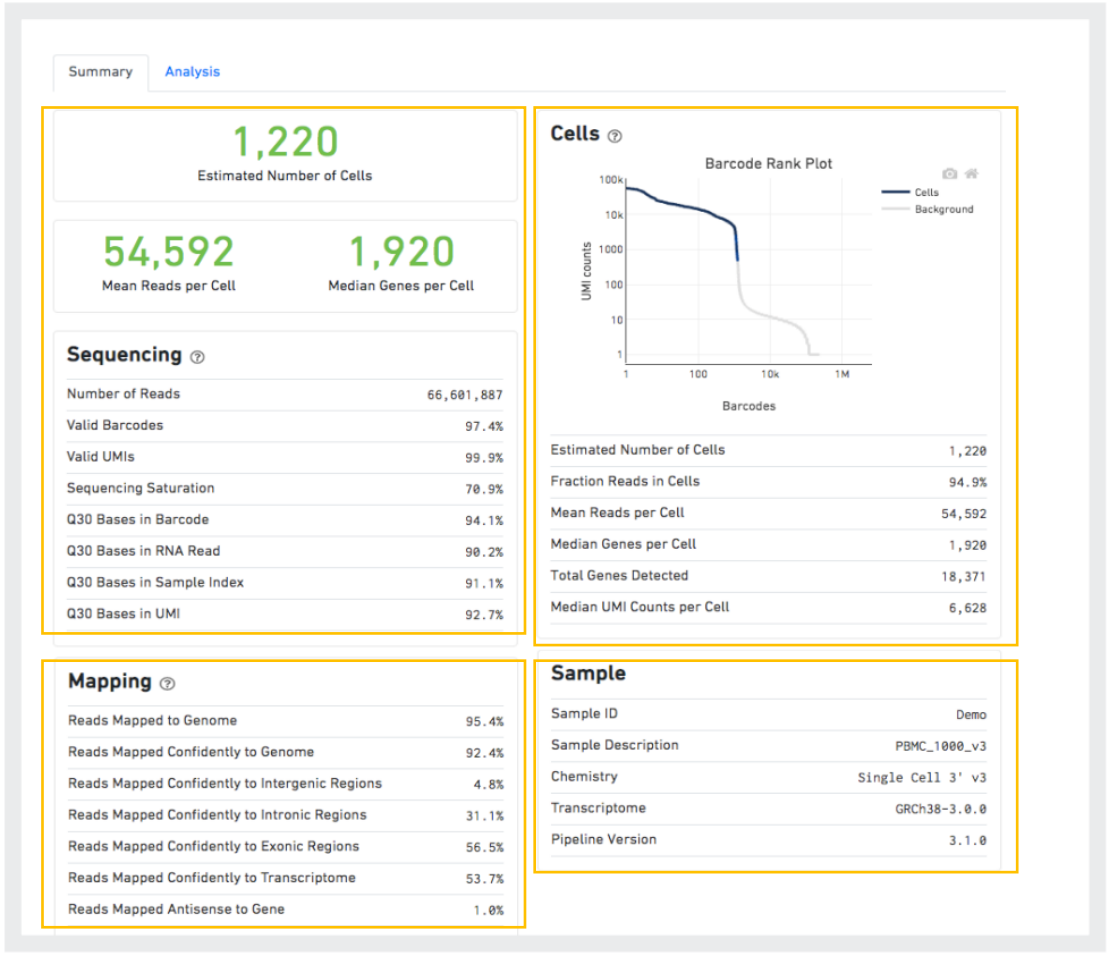
- **analysis:**
  - clustering (graphclust/kmeans)
  - diffexp (graphclust/kmeans)
  - pca
  - tsne
  - umap

Evaluating output data

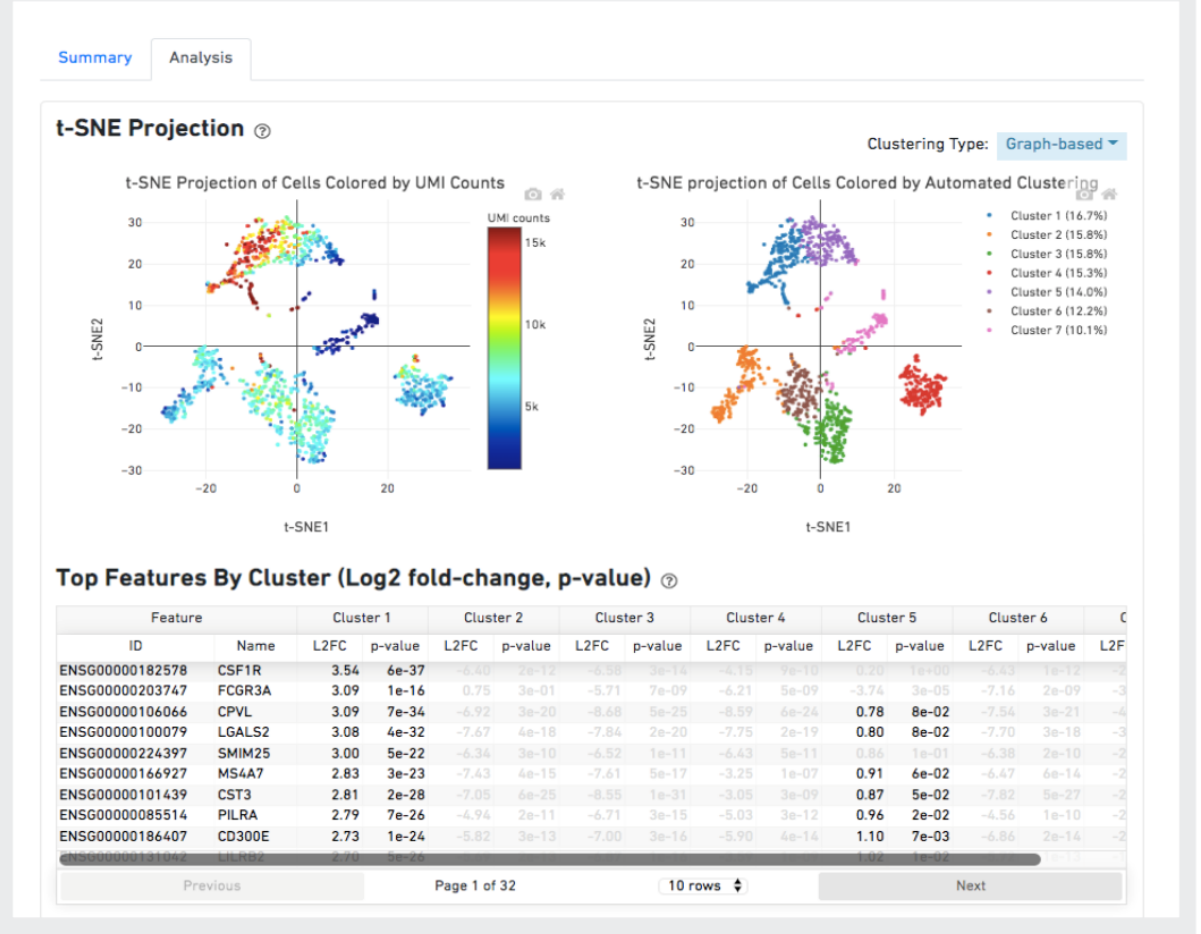
- **web_summary.html**
  - Check valid barcode
  - Check valid UMI
  - Check quality of data (Q30)
  - Check sequencing saturation
  - Check number of cells
  - Check mean reads per cell
  - Check library diversity

Evaluating output data

The summary metrics describe sequencing quality and various characteristics of the detected cells.



Analysis tab contains results from automated secondary analysis



Evaluating output data

Errors and Warnings

The analysis detected some issues with your sequencing run. Details »

Alert	Value	Detail
⚠ Low Fraction Reads Confidently Mapped To Transcriptome	51.5%	Ideal > 60%. This can indicate use of the wrong reference transcriptome, poor library quality, or poor sequencing quality. Application performance may be affected.

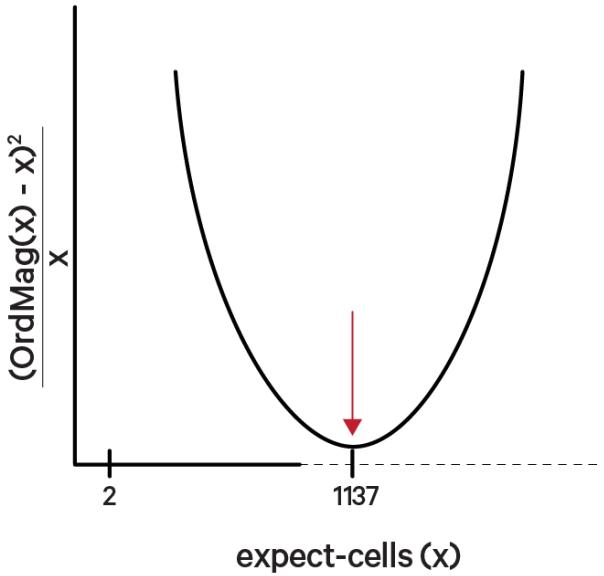
The analysis detected some serious issues with your sequencing run. Details »

Alert	Value	Detail
⊘ No Cells Detected	0	No valid sequencing data was detected. Please check the sequencing data.
⊘ Low Fraction Valid UMIs	0.0%	Ideal > 75%. This usually indicates a quality issue with the Ilumina R2 read. Application performance is likely to be affected.
⚠ Low Barcode Q30 Fraction (Illumina I7 Read)	67.5%	Ideal > 70%. Application performance may be affected.
⊘ Low UMI Q30 Fraction (Illumina R2 Read)	29.2%	Ideal > 80%. Application performance is likely to be affected.

Calling cell barcodes

Step 1: High RNA content cell identification (OrdMag)

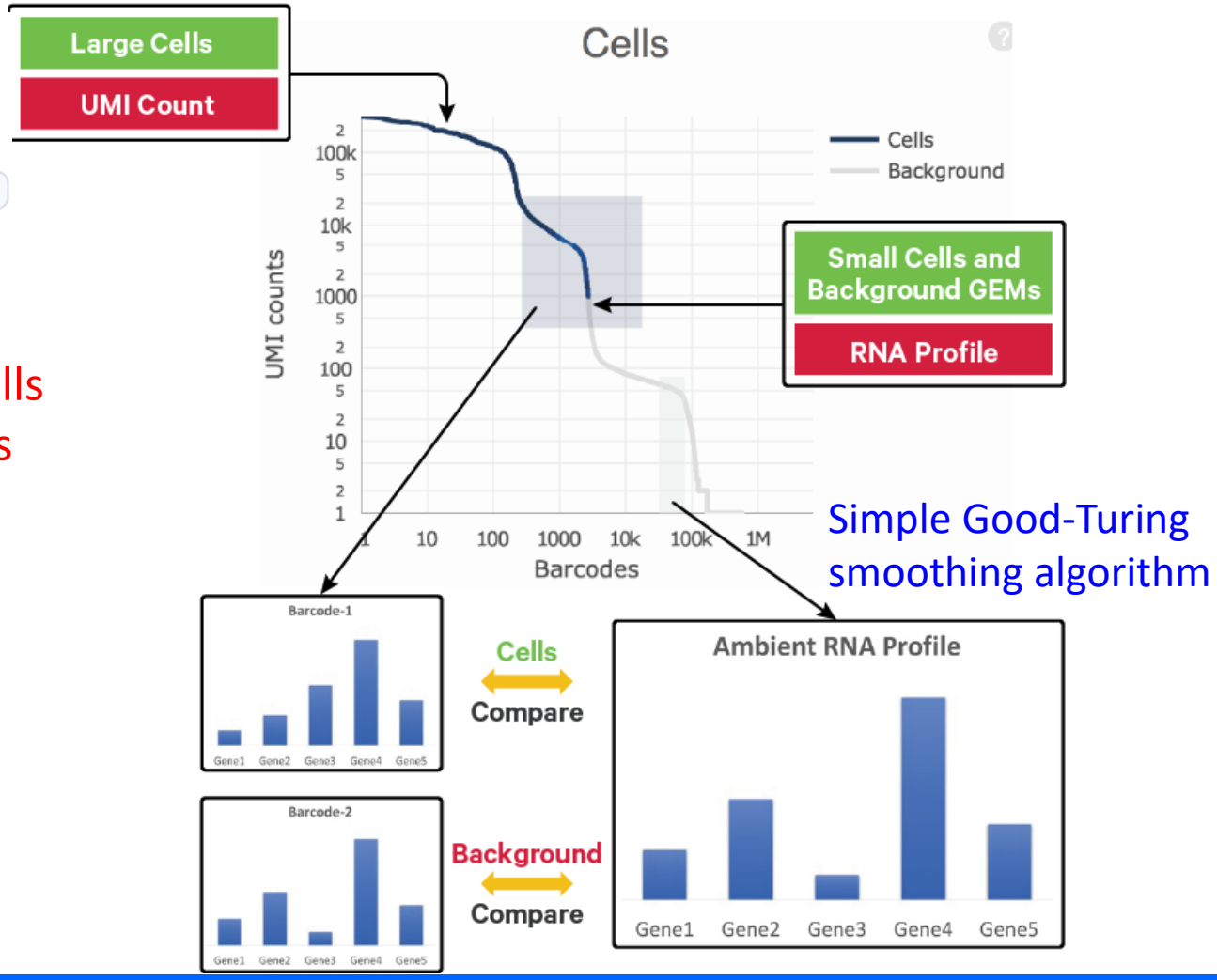
The OrdMag algorithm estimates the initial number of recovered cells by calling barcodes as cells if their total UMI counts exceed `m/10`, where `m` is the 99th percentile of the top N barcodes based on total UMI counts. The estimation process involves finding a value `x` that approximates `OrdMag(x)` by minimizing a loss function: (N = 3000)



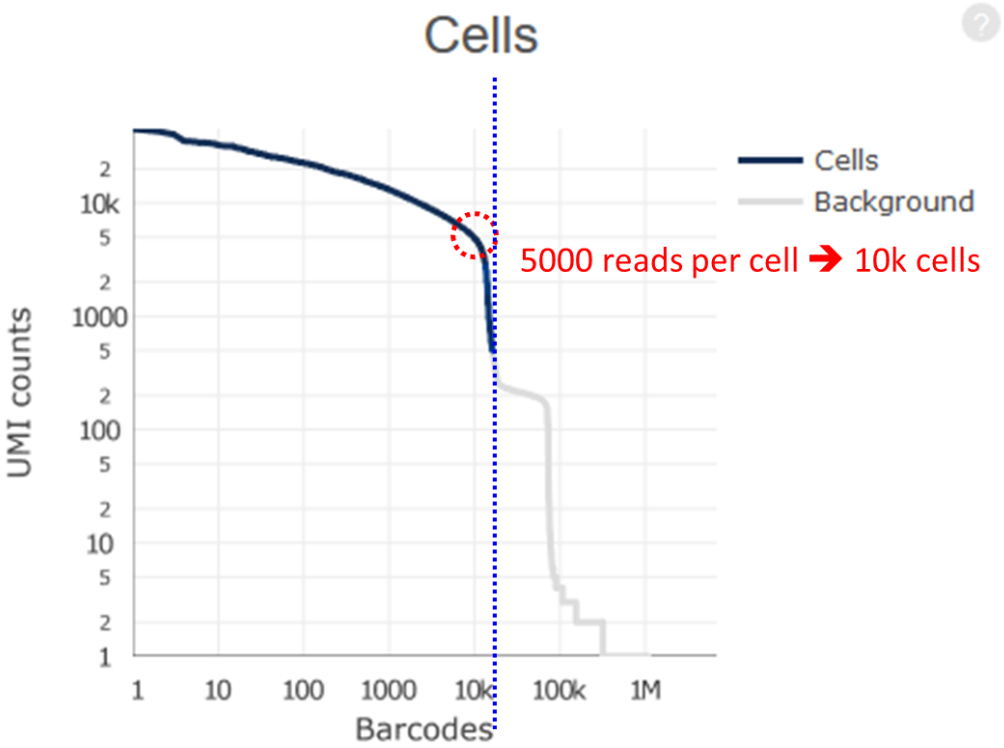
--expect-cells
--force-cells

$$expect-cells = min_x\left(\frac{(\mathrm{OrdMag}(x)-x)^2}{x}\right)$$

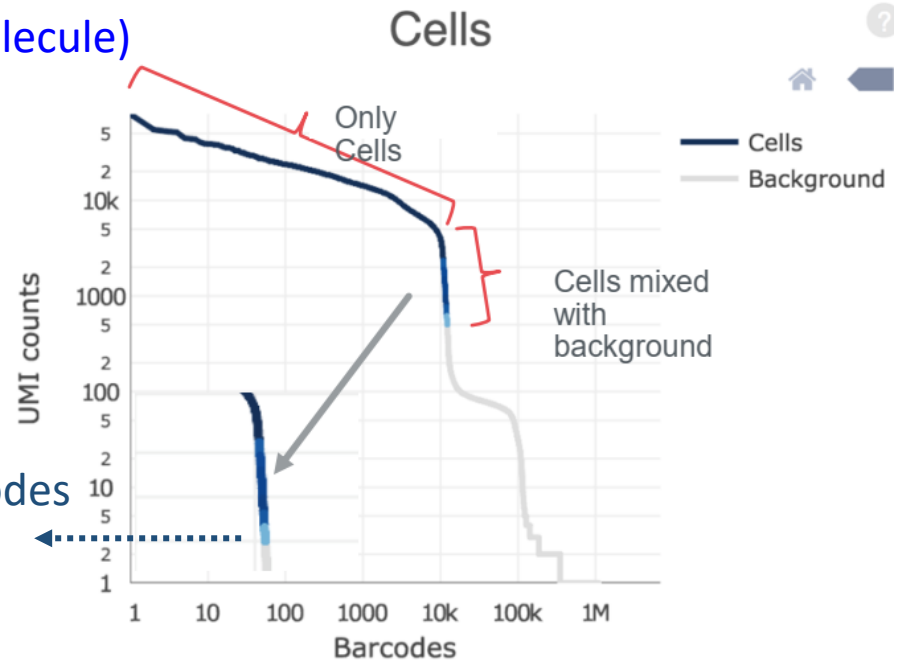
Step 2: Low RNA content cell identification (EmptyDrops)



Evaluating output data - Barcode rank plot



UMI - each read (RNA molecule)
Barcode - each cell
Feature - each gene

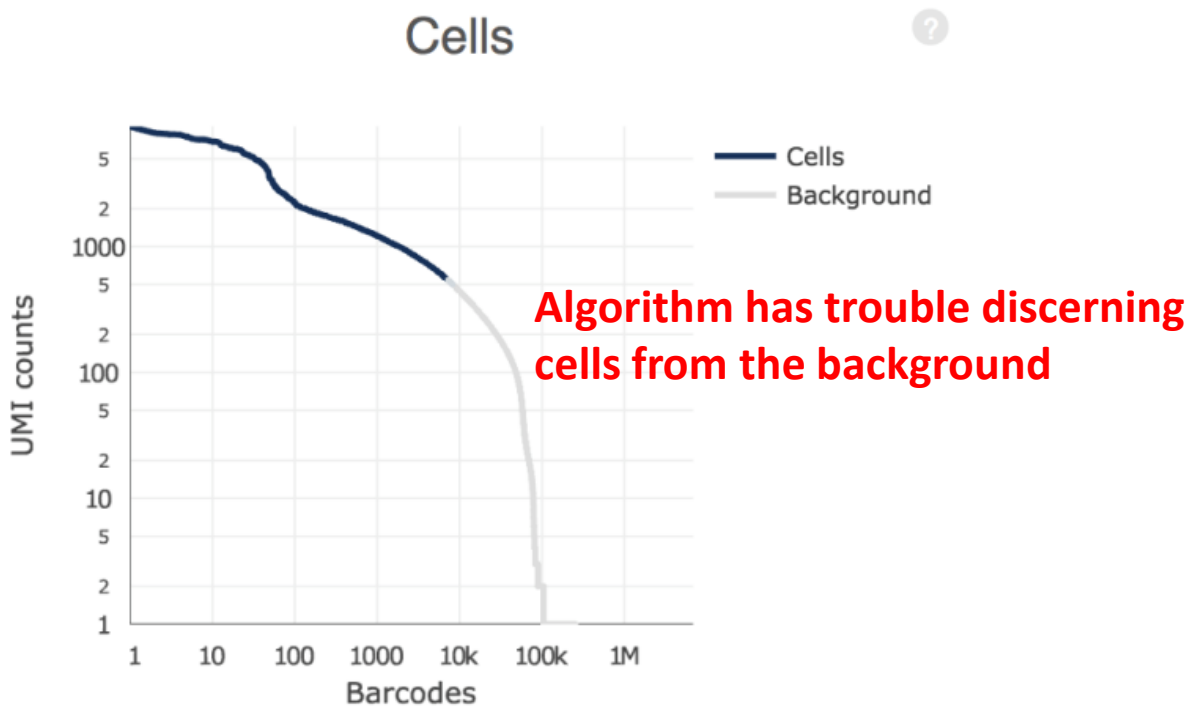


Cells and background barcodes have similar UMI counts

Y: the number of UMI counts mapped to each barcode
X: the number of barcodes below that value
A steep drop-off is indicative of good separation between the cell-associated barcodes and the background barcodes.
Barcode rank plot shows distribution of barcode counts and which barcodes were inferred to be associated with cells.
Cells are colored in “blue”
Blue color’s gradient is proportional to the fraction of cells in a given subset of barcodes

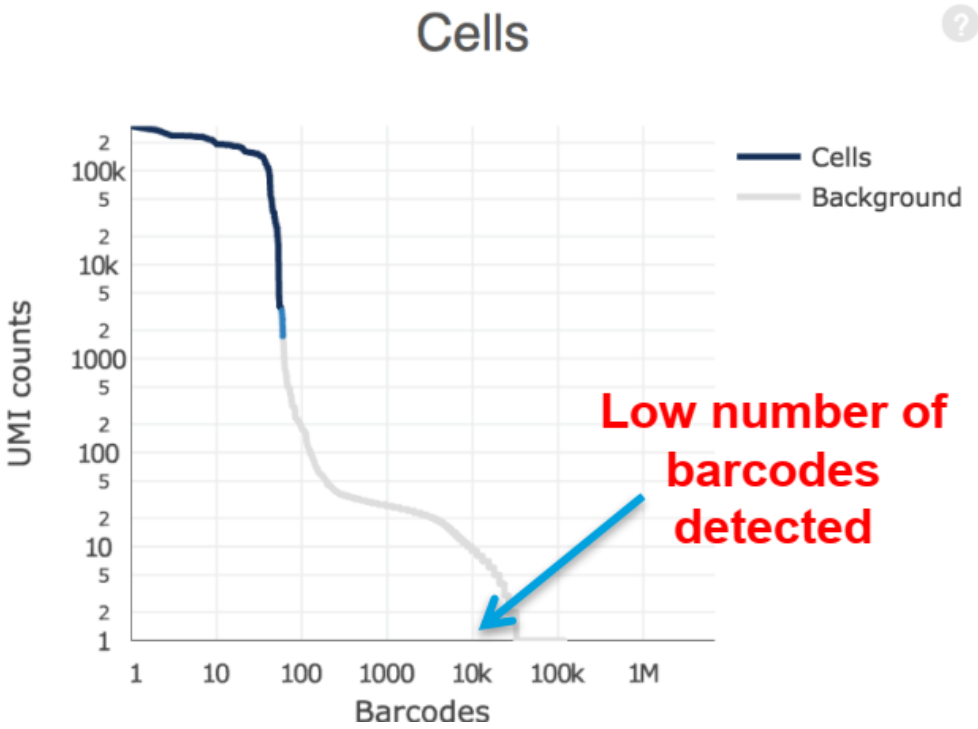
Trouble shooting - Cell calling

Wetting failure



Algorithm has trouble discerning cells from the background

Clogs, low sequencing depth



Low number of barcodes detected

The number of cells recovered will likely be lower than the targeted cell numbers. But the data for the cells called is usable

QC metrics - mapping

Sequencing

Number of Reads	66,601,887
Valid Barcodes	97.4%
Valid UMIs	99.9%
Sequencing Saturation	70.9%
Q30 Bases in Barcode (> 70%)	94.1%
Q30 Bases in RNA Read	90.2%
Q30 Bases in Sample Index	91.1%
Q30 Bases in UMI	92.7%

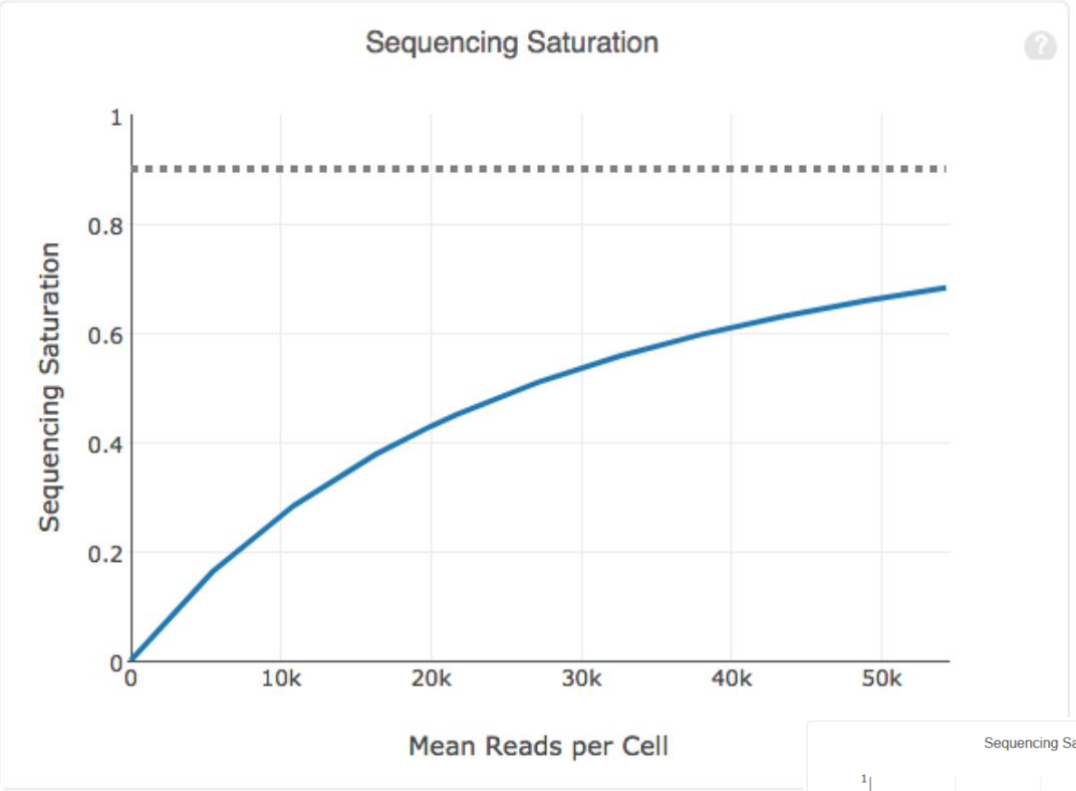
Mapping

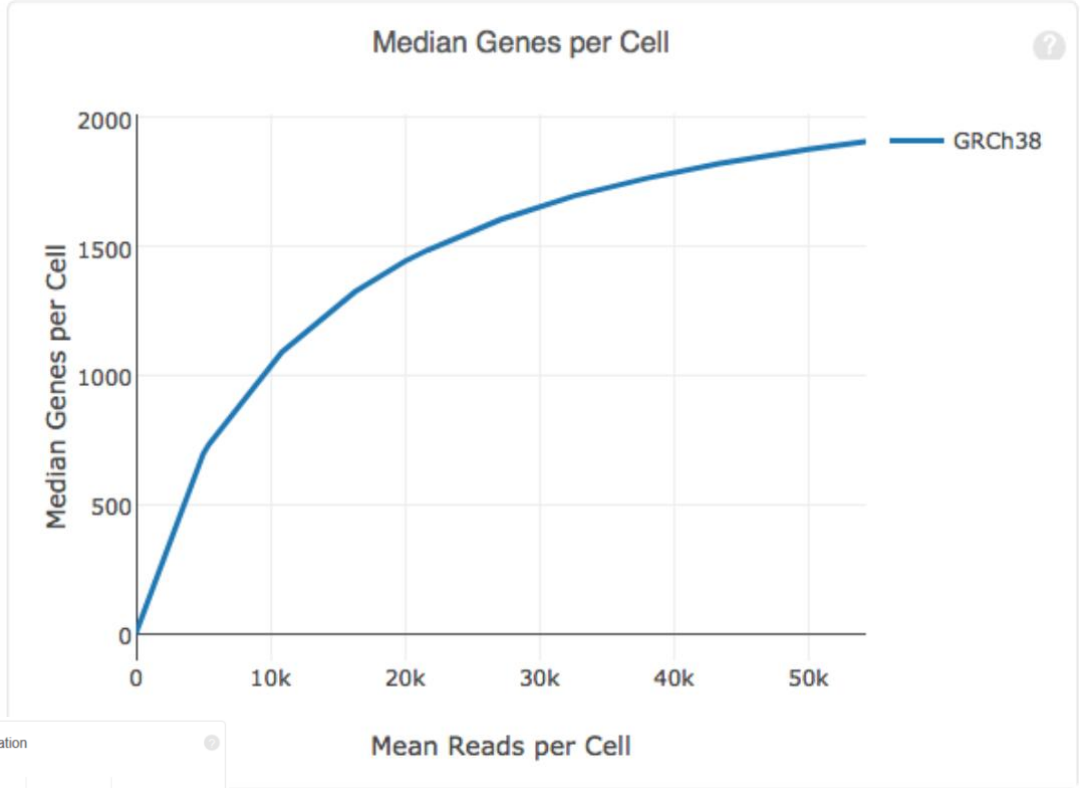
Reads Mapped to Genome	95.4%
Reads Mapped Confidently to Genome	92.4%
Reads Mapped Confidently to Intergenic Regions	4.8%
Reads Mapped Confidently to Intronic Regions	31.1%
Reads Mapped Confidently to Exonic Regions	56.5%
Reads Mapped Confidently to Transcriptome (> 30%)	53.7%
Reads Mapped Antisense to Gene	1.0%

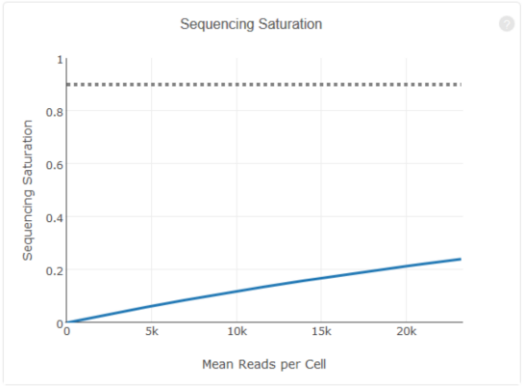
Possible causes for low mapping rates:

- Reads mapped to wrong genome or different strain
- Very short RNA read length
- Large stretches of no-calls in the reads
- Custom reference contains overlapping genes

QC metrics - sequencing



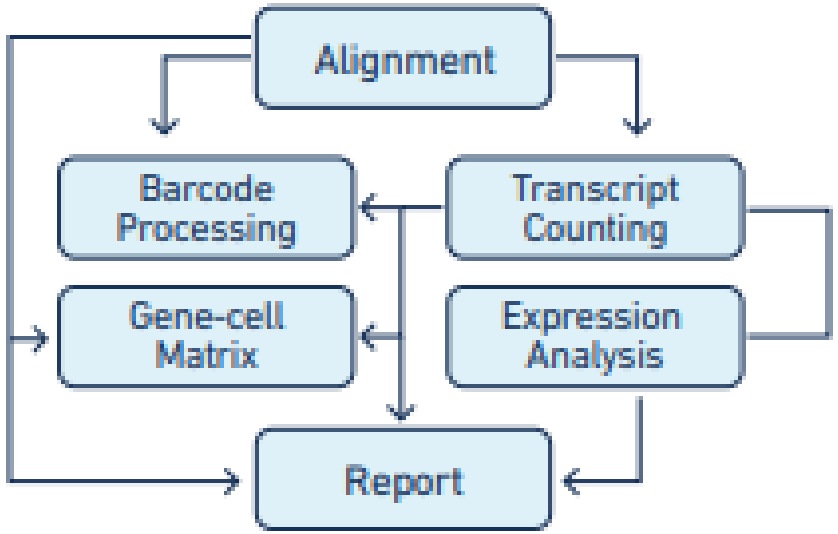




10x Loupe Browser (Demo)

Download
https://www.10xgenomics.com/products/loupe-browser

Statistics analysis



https://satijalab.org/seurat/

Clustering & Feature anchors to match datasets

https://cole-trapnell-lab.github.io/monocle3/

Trajectory mapping

1. Transcript Counts
2. Data Normalizations
3. Metadata
4. Data visualization & clustering- dimension reduction
   - t-SNE (t-distributed stochastic neighbor embedding)
   - UMAP (Uniform Manifold Approximation and Projection)

Seurat

Seurat is an **R package** designed for QC, analysis, and exploration of single-cell RNA-seq data. Seurat aims to enable users to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements, and to integrate diverse types of single-cell data.

Seurat

```
library(Seurat)
library(SeuratData)
library(SeuratWrappers)
```

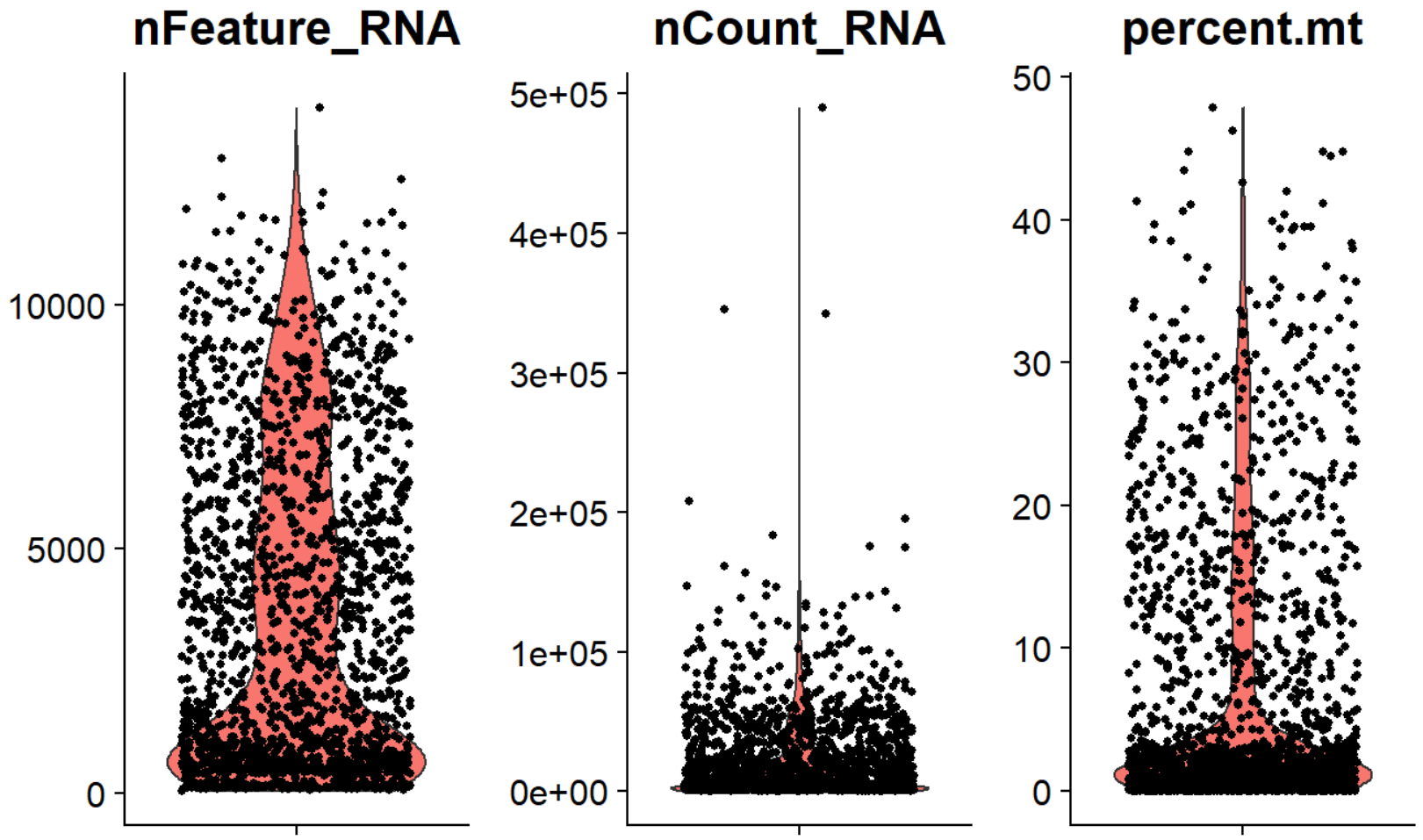
Seurat (for general single cell loading and processing)

Input data

```
pbmc.data <- Read10X(data.dir = data_dir)
pbmc <- CreateSeuratObject(counts = pbmc.data, min.cells = 3, min.features = 200, project = “pbmc")
pbmc[["percent.mt"]] <- PercentageFeatureSet(pbmc, pattern = "^MT-")
```

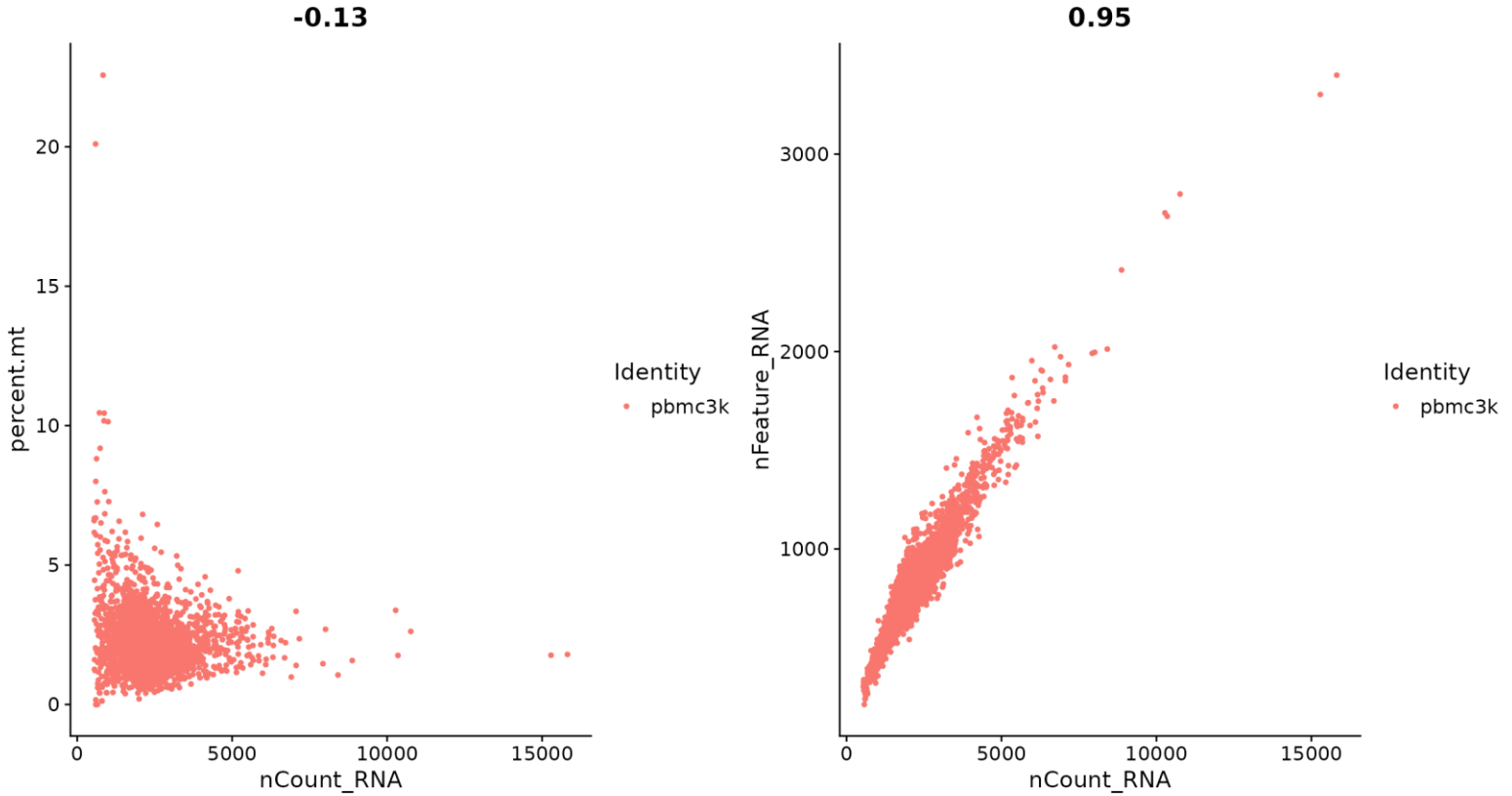
Violin plot

```
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
```



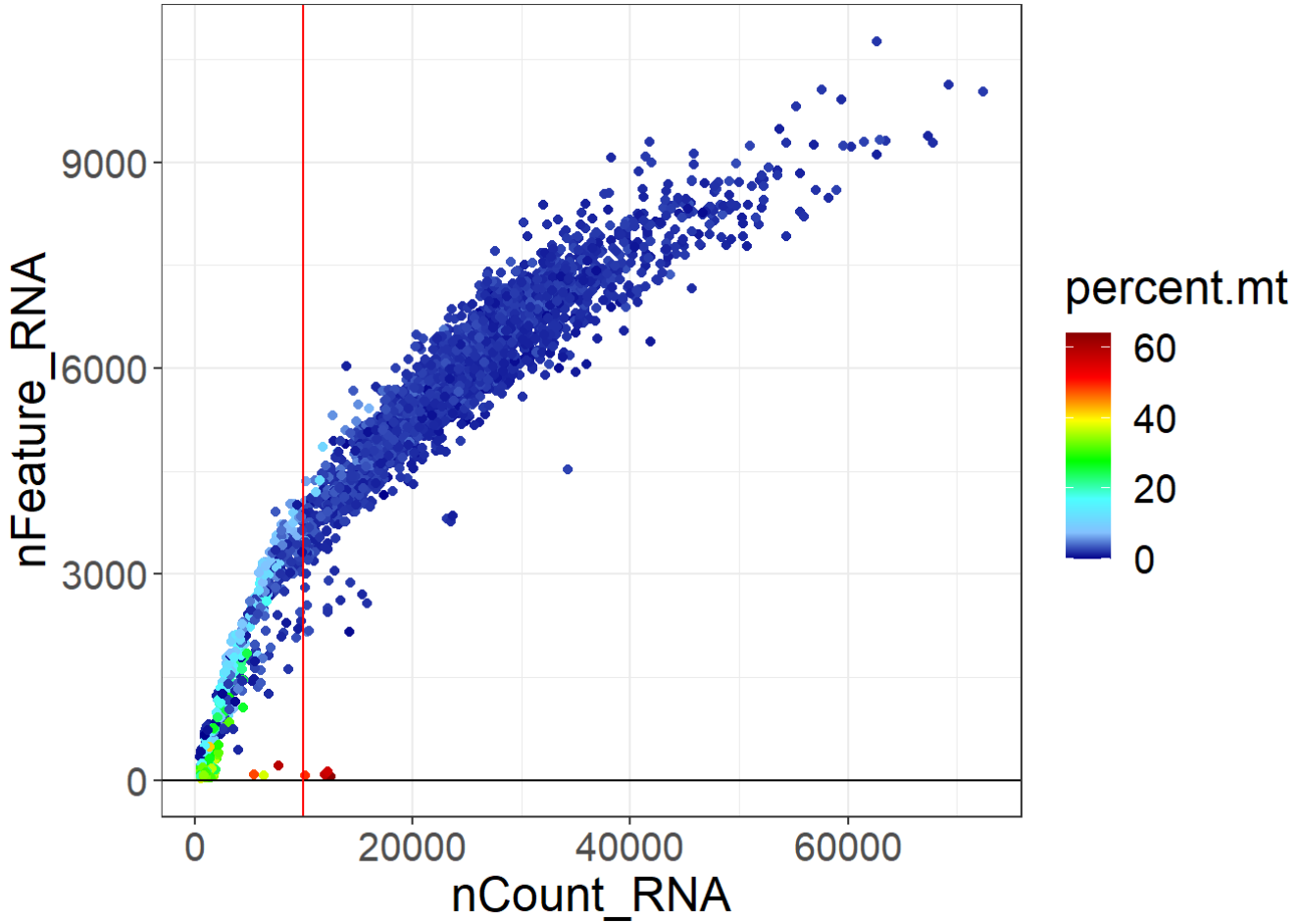
Quality control

```
plot1 <- FeatureScatter(pbmc, feature1 = "nCount_RNA", feature2 = "percent.mt")
plot2 <- FeatureScatter(pbmc, feature1 = "nCount_RNA", feature2 = "nFeature_RNA")
plot1 + plot2
```



Data filtering

```
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)
```



ggplot

Data Normalization

```
pbmc <- NormalizeData(pbmc, normalization.method = "LogNormalize", scale.factor = 10000)
```

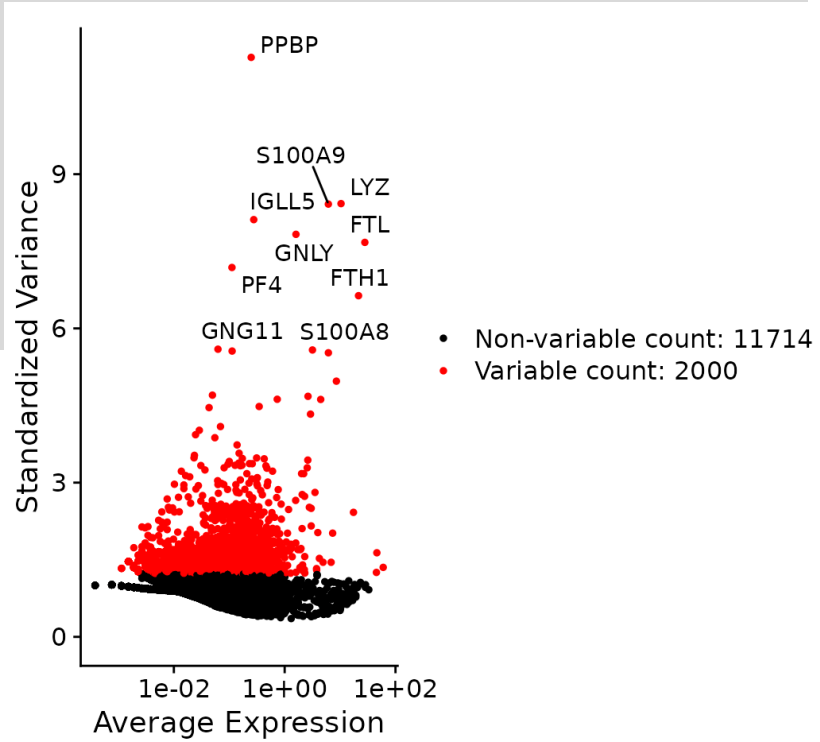
Default values

normalization.method =

- **LogNormalize**: Feature counts for each cell are divided by the total counts for that cell and multiplied by the scale.factor. This is then natural-log transformed using log1p.
- **CLR**: Applies a centered log ratio transformation
- **RC**: Relative counts. Feature counts for each cell are divided by the total counts for that cell and multiplied by the scale.factor. No log-transformation is applied. For counts per million (CPM) set scale.factor = 1e6

Selected features/genes

```
# example: Select highly variable features
pbmc <- FindVariableFeatures(pbmc, selection.method = "vst", nfeatures = 2000)
# Identify the 10 most highly variable genes
vst_genes_top10 <- head(VariableFeatures(pbmc), 10)
# plot variable features with and without labels
plot1 <- VariableFeaturePlot(pbmc)
plot2 <- LabelPoints(plot = plot1, points = vst_genes_top10 , repel = TRUE)
plot1 + plot2
```



Data Scaling

linear transformation

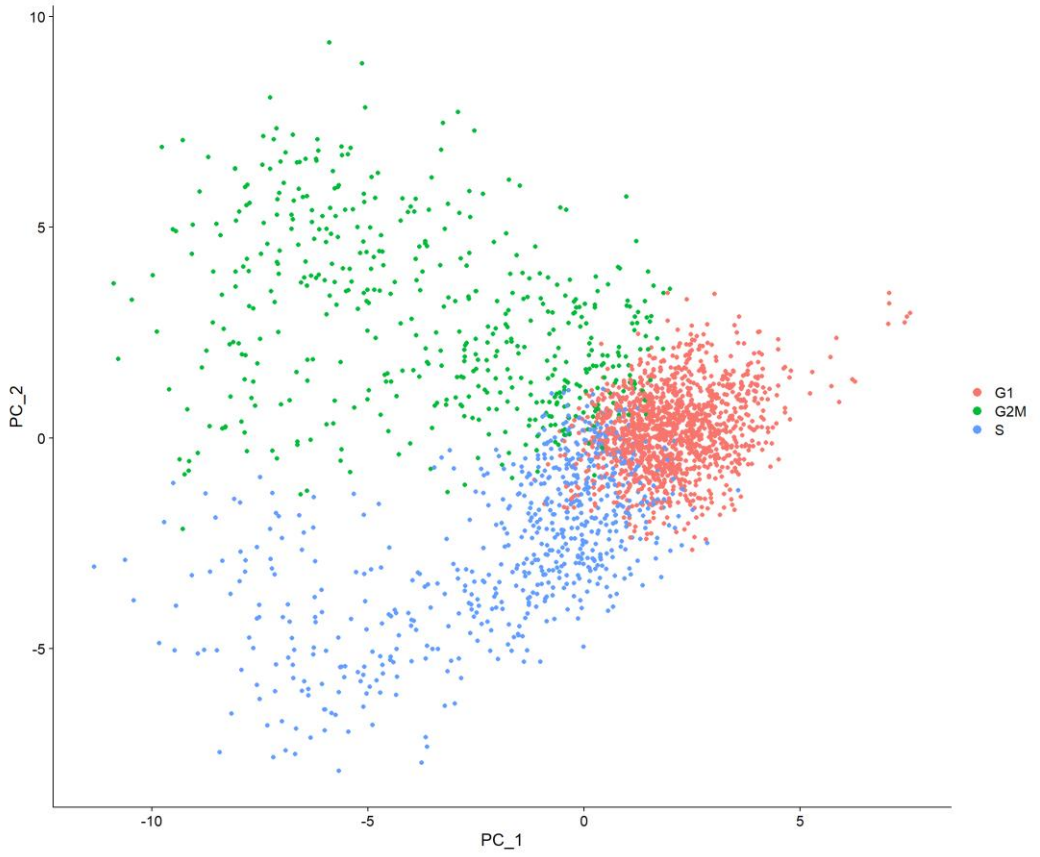
```
pbmc <- ScaleData(pbmc, features = rownames(pbmc))
```

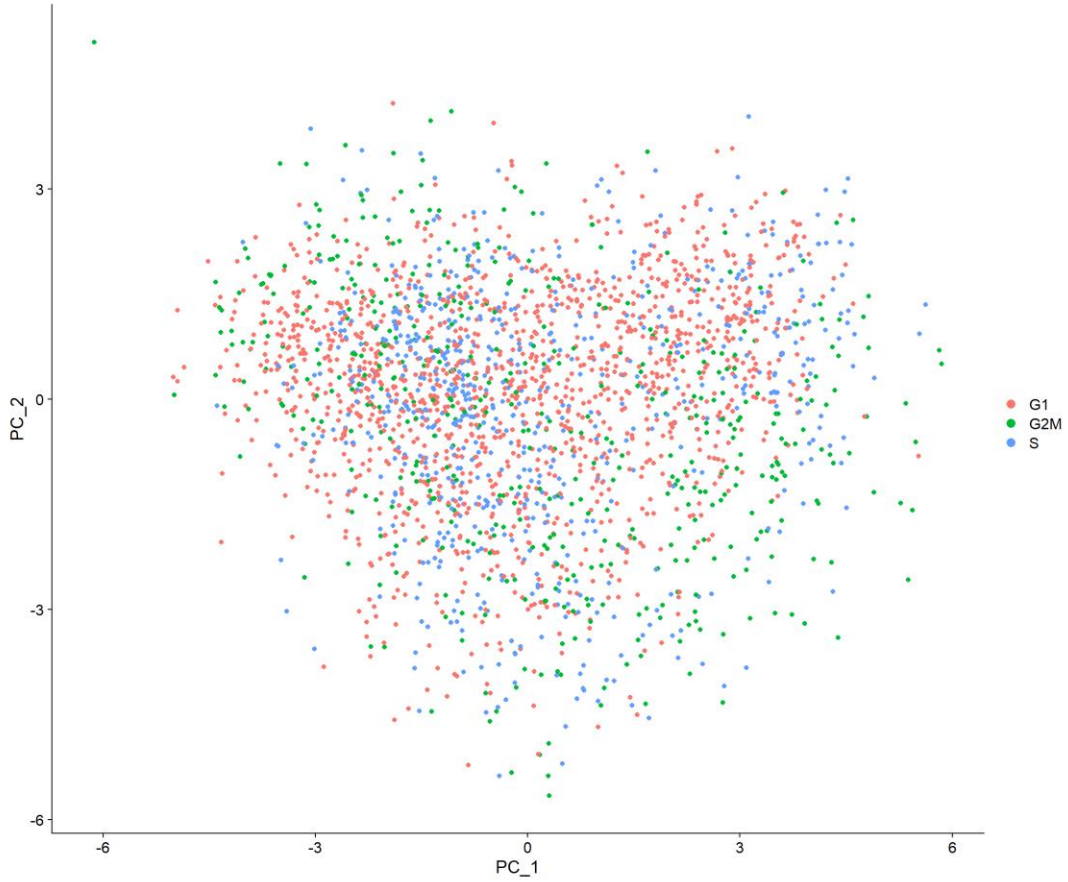
- Required for the dimensional reduction techniques
- Shifts the expression of each gene, so that the mean expression across cells is 0
- Scales the expression of each gene, so that the variance across cells is 1

- If this step takes too long, you could only select genes that will be used as input to dimensional reduction
- By default, **ScaleData** use top 2000 highly variable features (if no “features” parameter)

Cell-cycle genes (regressing out cellcycle)

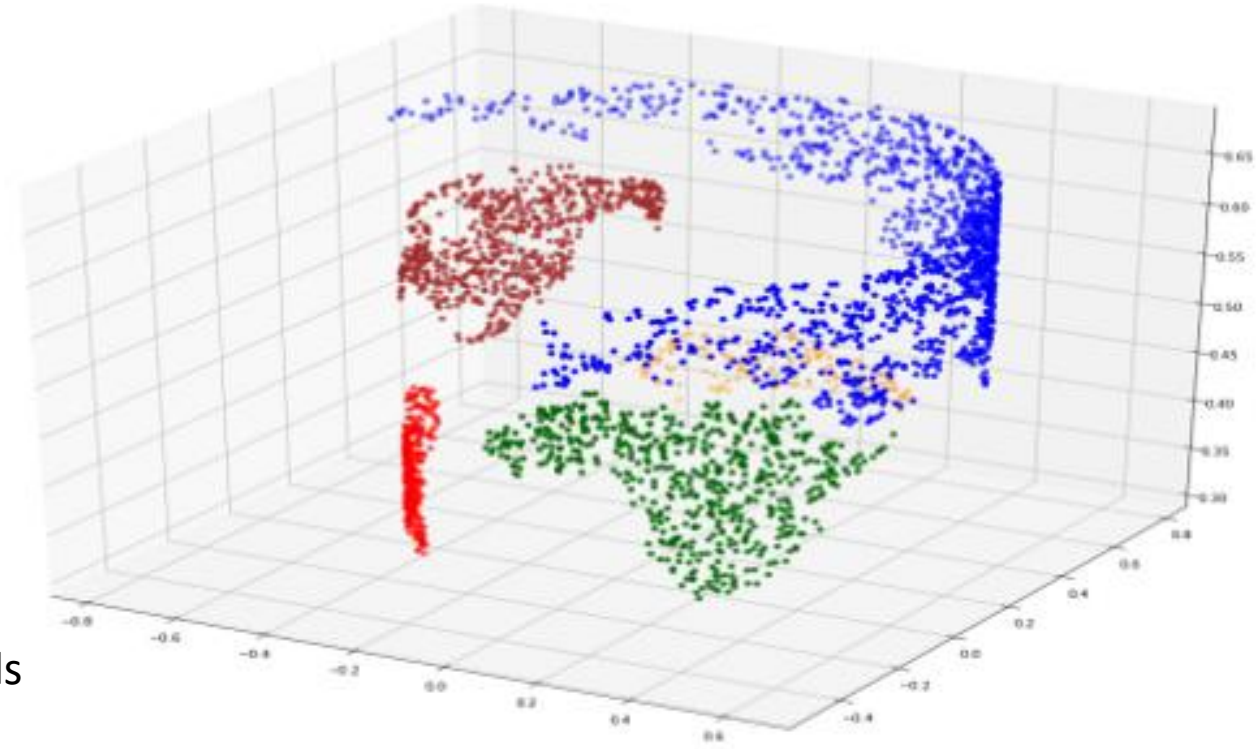
```
pbmc <- CellCycleScoring(pbmc, s.features = s.genes, g2m.features = g2m.genes, set.ident = TRUE)
pbmc <- ScaleData(Sc pbmc, vars.to.regress = c("S.Score", "G2M.Score"), features = rownames(pbmc))
```





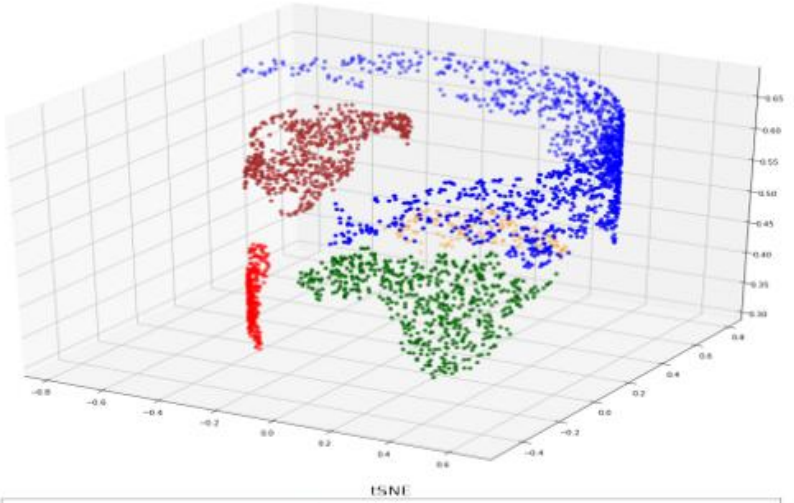
Dimensional reduction

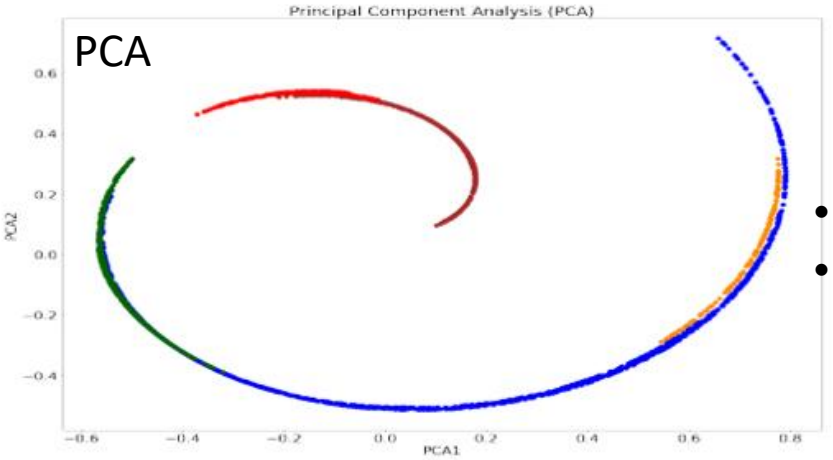
- Too many cells and measured genes (space is huge)
- Realistically only 2 dimensions we can easily understand and view

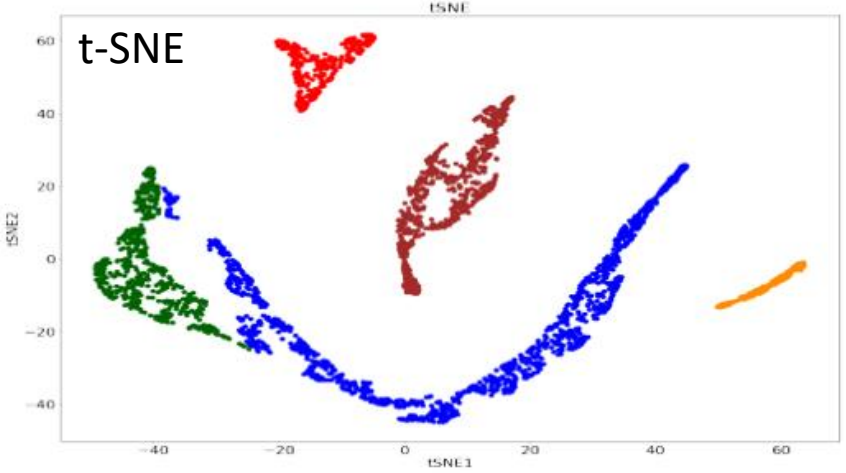


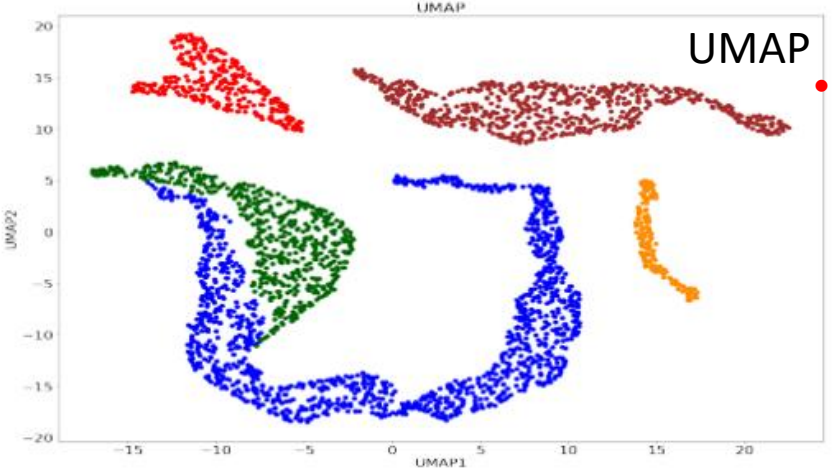
Each dot is a cell
Groups of dots are similar cells

Dimensional reduction









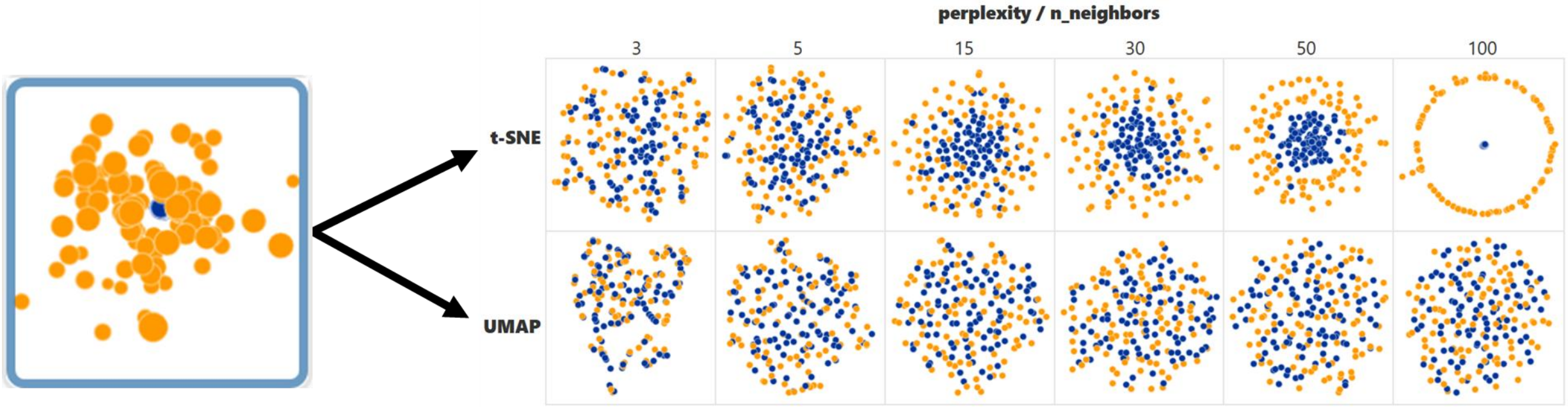
- PCA failed to reconstruct the data sets.
- PCA as a **linear algorithm** assigns equal weights to all pairwise distances

- **Non-linear** manifold learners, such as tSNE and UMAP, prioritize distances between neighbors. This strategy allows them to figure out the intrinsic 2D dimensionality of the data.

UMAP

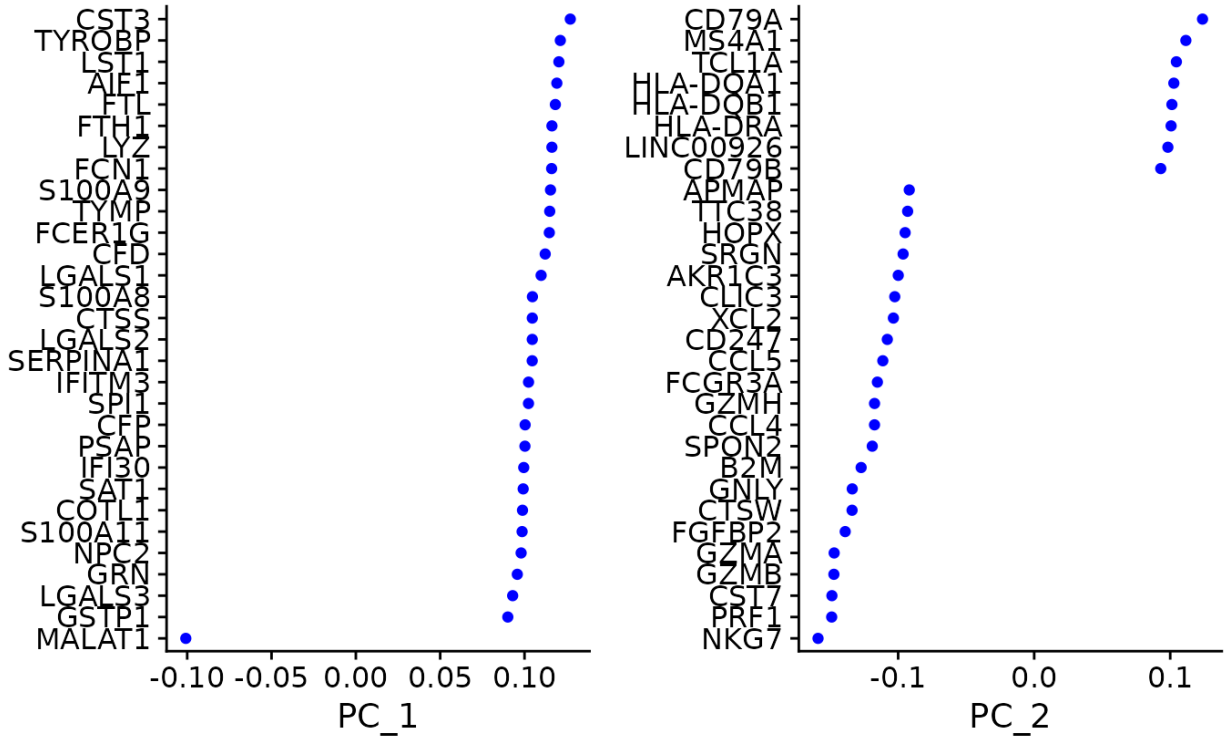
- Considering both **Nearest neighbours** and **Minimum distance**
- Can preserve more global structure than tSNE
- Quicker than t-SNE
  - UMAP skips a normalization step in the calculation of high dimensional distances which speeds it up
  - In the 2D projection UMAP uses a more efficient method to shuffle the cells into their final position

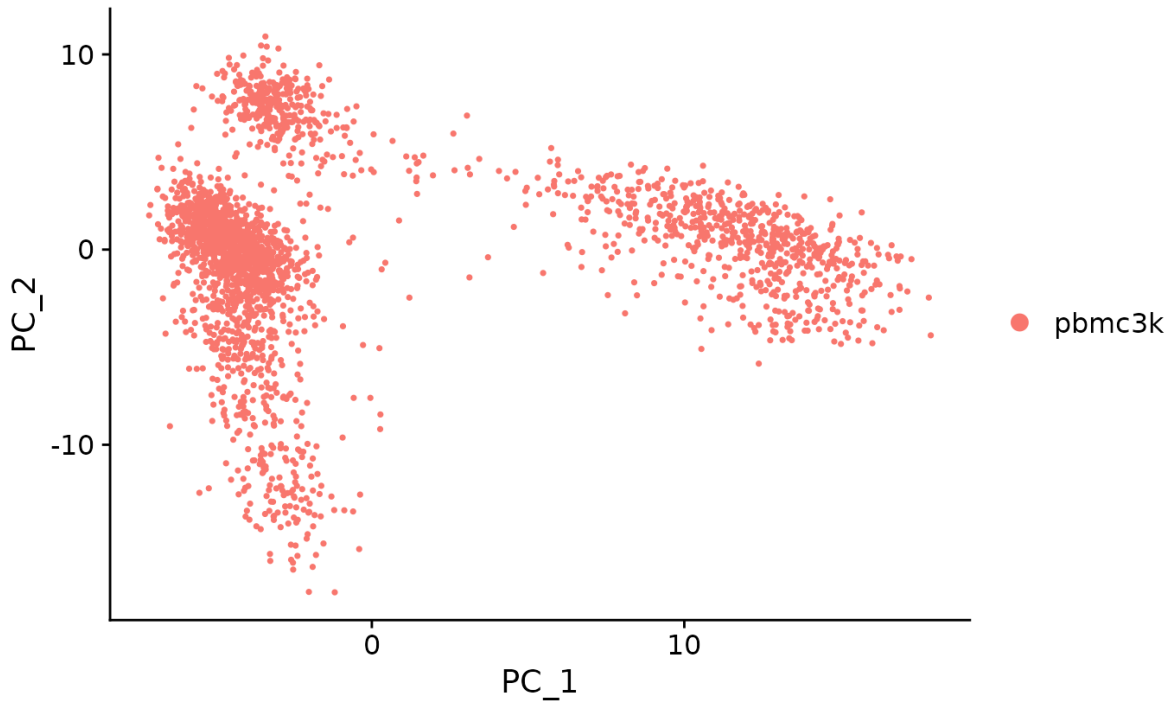
Conclusion: it still depends on the datasets



Perform linear dimensional reduction (PCA)

```
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
VizDimLoadings(pbmc, dims = 1:2, reduction = "pca")
DimPlot(pbmc, reduction = "pca")
```

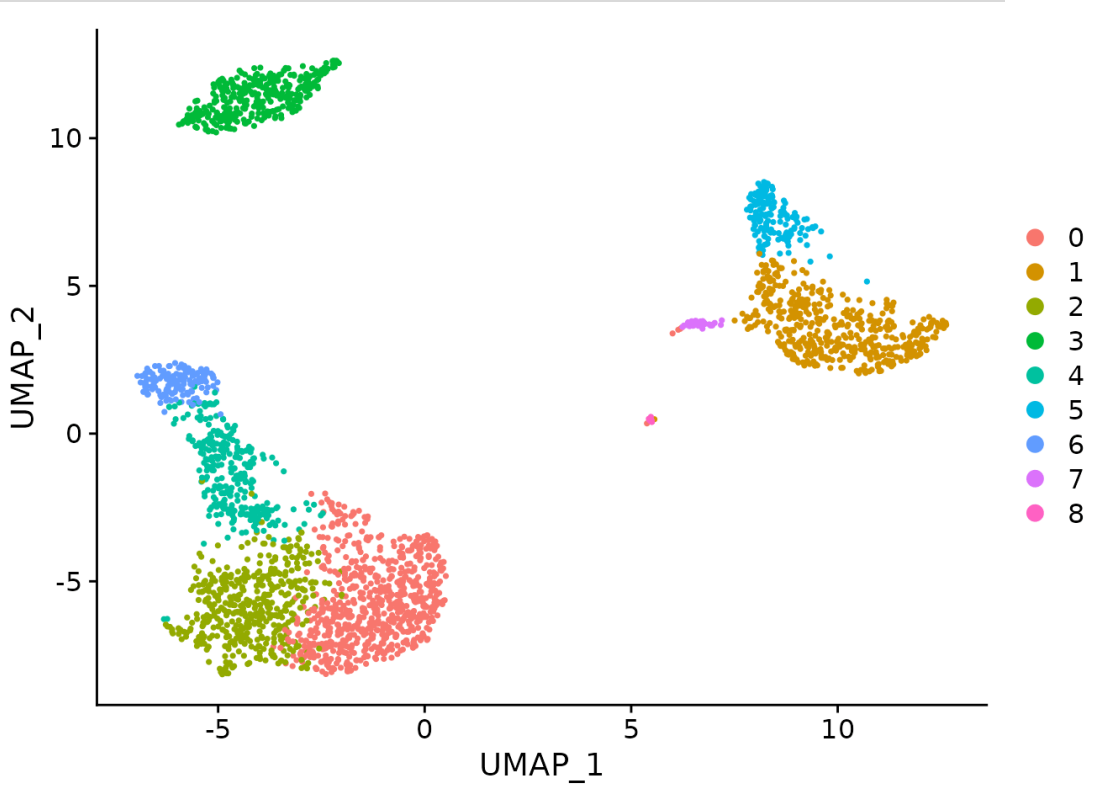




Perform nonlinear dimensional reduction (t-SNE/UMAP)

```
obj <- FindNeighbors(obj, dims = 1:30, reduction = "pca")
obj <- FindClusters(obj, resolution = 2, cluster.name = "unintegrated_clusters")

## t-SNE
pbmc <- RunTSNE(pbmc, dims = 1:10)
DimPlot(pbmc, reduction = "tsne")
## UMAP
pbmc <- RunUMAP(pbmc, dims = 1:10)
DimPlot(pbmc, reduction = "umap")
```



Finding markers/features of each cluster

By default, it identifies positive and negative markers of a single cluster (specified in ident.1), compared to all other cells

```
cluster2.markers <- FindMarkers(pbmc, ident.1 = 2, min.pct = 0.25, logfc.threshold = 0.25 )
                                                   cluster

# find all markers distinguishing cluster 5 from clusters 0 and 3
cluster5.markers <- FindMarkers(pbmc, ident.1 = 5, ident.2 = c(0, 3), min.pct = 0.25)
```

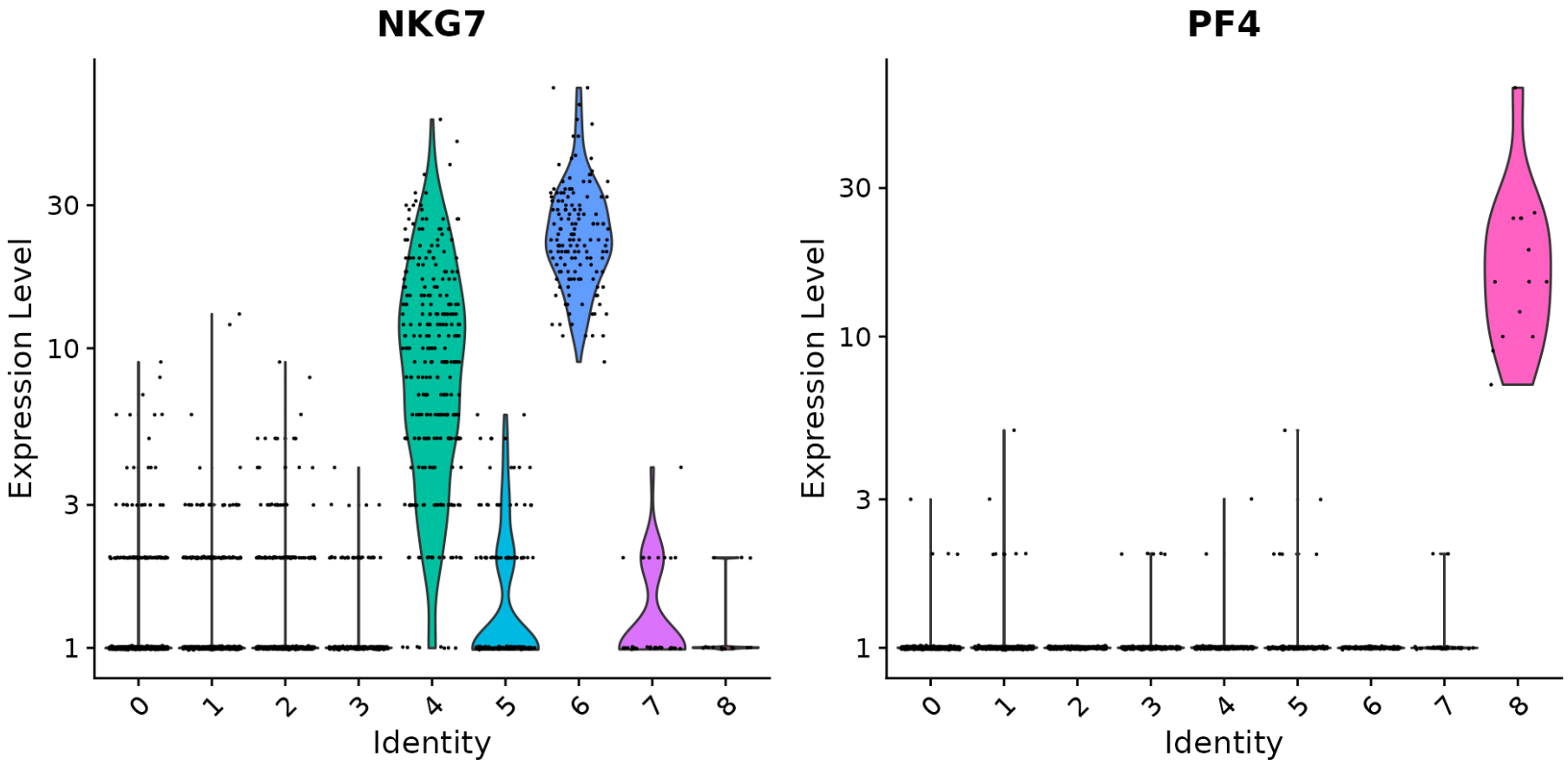
only test genes that are detected in a minimum fraction of min.pct cells in either of the two populations.

test.use =

- "wilcox“ (Default)
- "bimod"
- "roc"
- "t"
- "negbinom"
- "poisson"
- "LR"
- "MAST"
- "DESeq2"

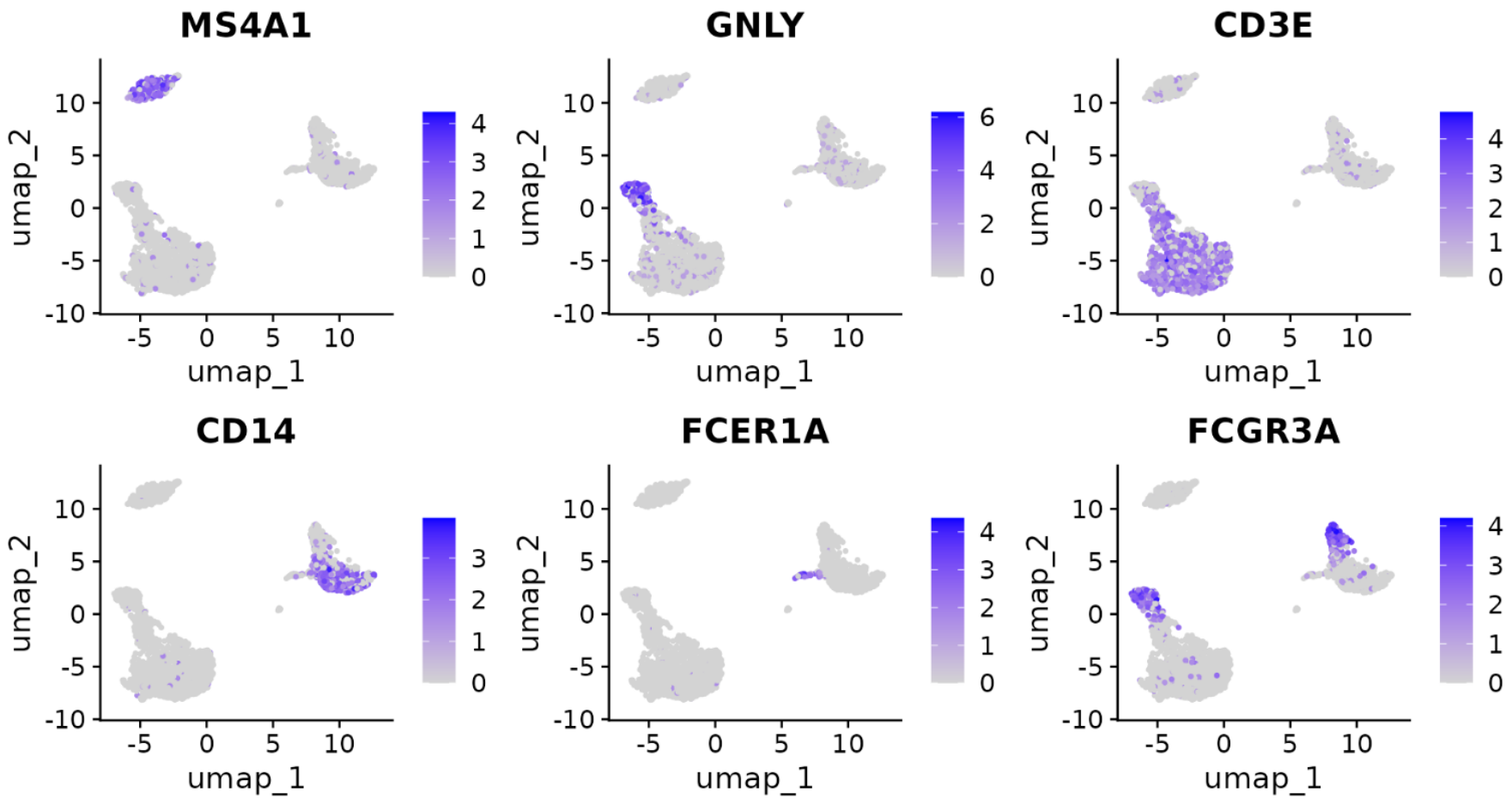
Expression probability distributions across clusters

```
VlnPlot(pbmc, features = c("NKG7", "PF4"), slot = "counts", log = TRUE)
```



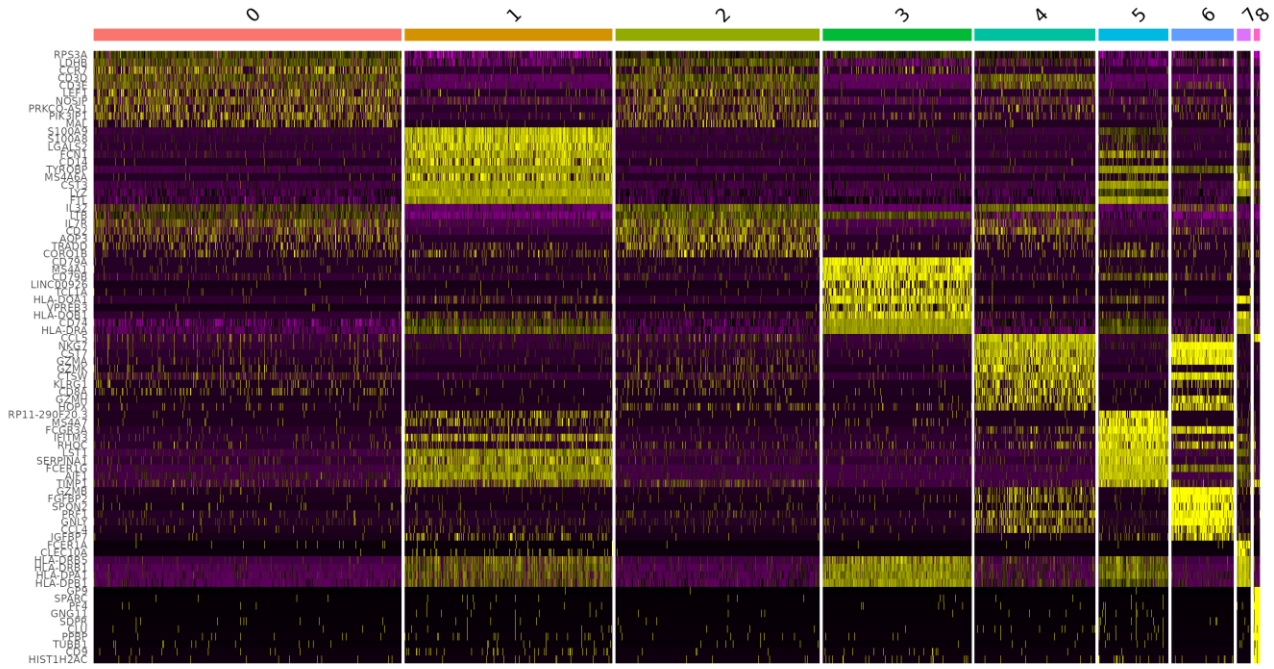
Expression probability distributions across clusters

```
FeaturePlot(pbmc, features = c("MS4A1", "GNLY", "CD3E", "CD14", "FCER1A", "FCGR3A"))
```



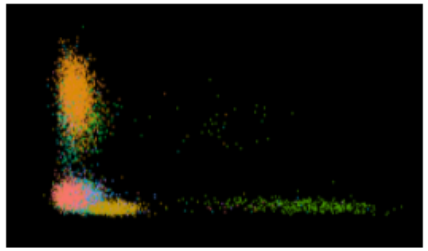
Heatmap

```
pbmc.markers %>%
    group_by(cluster) %>%
    dplyr::filter(avg_log2FC > 1) %>%
    slice_head(n = 10) %>%
    ungroup() -> top10
DoHeatmap(pbmc, features = top10$gene) + NoLegend()
```



Other analyses (satijalab.org/seurat/articles/get_started.html)

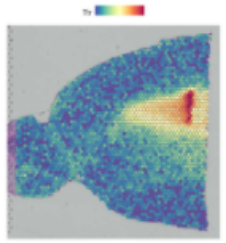
Multimodal analysis



An introduction to working with multi-modal datasets in Seurat.

GO

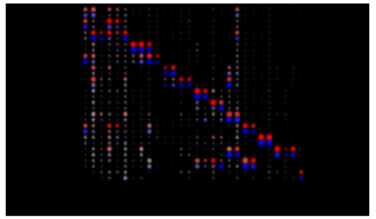
Analysis of spatial datasets



Learn to explore spatially-resolved transcriptomic data with examples from 10x Visium and Slide-seq v2.

GO

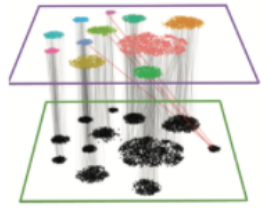
Introduction to scRNA-seq integration



An introduction to integrating scRNA-seq datasets in order to identify and compare shared cell types across experiments

GO

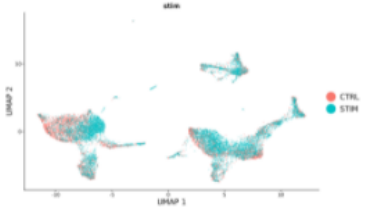
Mapping and annotating query datasets



Learn how to map a query scRNA-seq dataset onto a reference in order to automate the annotation and visualization of query cells

GO

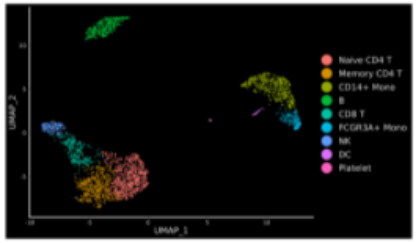
Fast integration using reciprocal PCA (RPCA)



Identify anchors using the reciprocal PCA (rPCA) workflow, which performs a faster and more conservative integration

GO

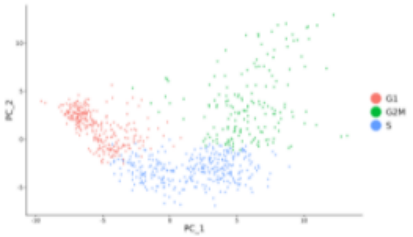
Visualization



An overview of the major visualization functionality within Seurat.

GO

Cell Cycle Regression



Mitigate the effects of cell cycle heterogeneity by computing cell cycle phase scores based on marker genes

GO

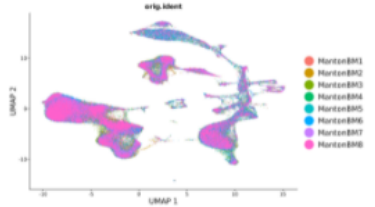
Differential Expression Testing

```
> FindMarkers(pbmc, ident.1 = "CD14+ Mono",
      ident.2 = "FCGR3A+ Mono", min.pct = 0.5)
               p_val avg_logFC pct.1 pct.2    p_val_adj
FCGR3A 1.193617e-101 -2.617707 0.131 0.975 1.636926e-97
LYZ     8.134552e-75  1.812078 1.000 0.988 1.115572e-70
RHOC    4.479768e-68 -1.611576 0.162 0.864 6.143554e-64
S100A8  7.471811e-65  2.610696 0.975 0.500 1.024684e-60
S100A9  1.318422e-64  2.286734 0.996 0.870 1.808084e-60
IFITM2  4.821669e-64 -1.445772 0.677 1.000 6.612437e-60
```

Perform differential expression (DE) testing in Seurat using a number of frameworks.

GO

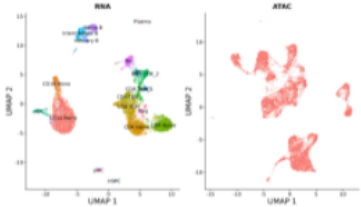
Tips for integrating large datasets



Tips and examples for integrating very large scRNA-seq datasets (including >200,000 cells)

GO

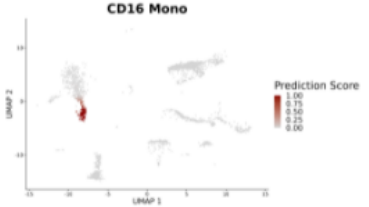
Integrating scRNA-seq and scATAC-seq data



Annotate, visualize, and interpret an scATAC-seq experiment using scRNA-seq data from the same biological system

GO

Multimodal Reference Mapping



Analyze query data in the context of multimodal reference atlases.

GO

Trajectories (Monocle3)

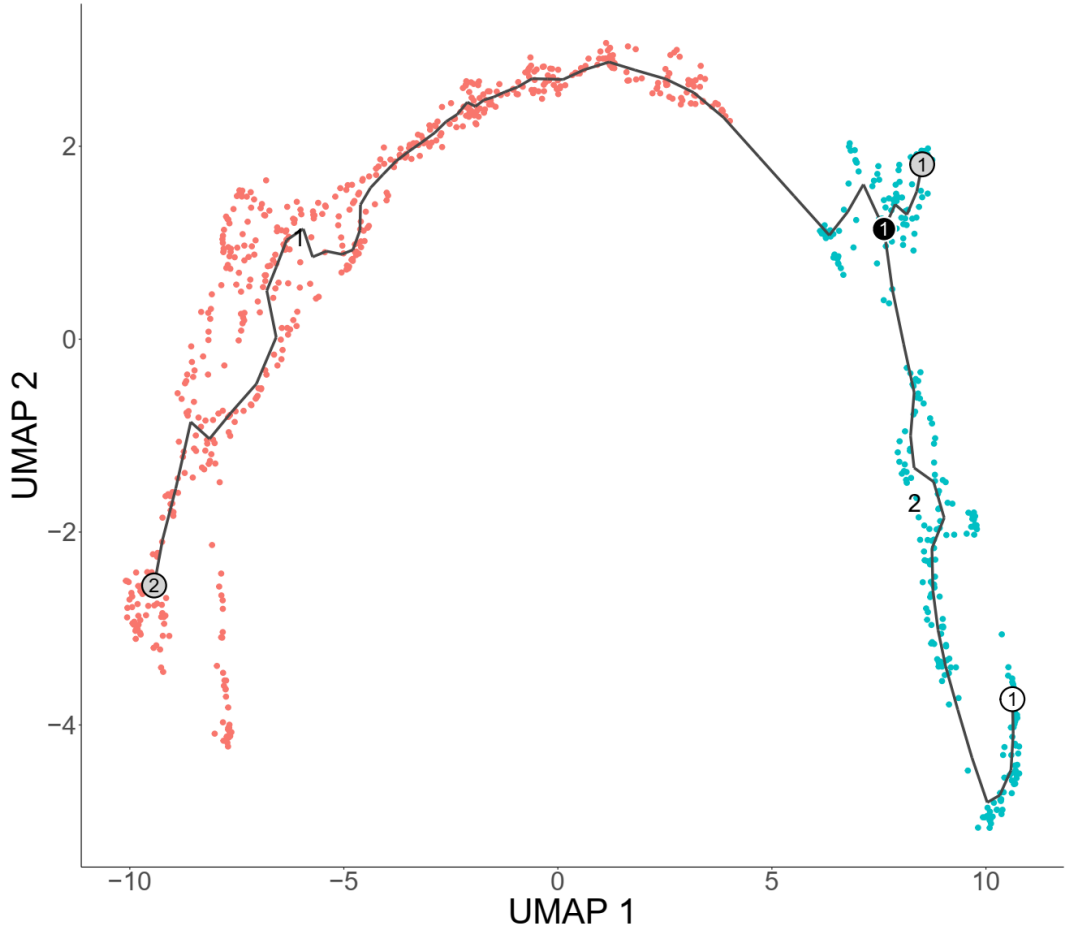
Clustering, classifying, and counting cell

Constructing single-cell trajectories

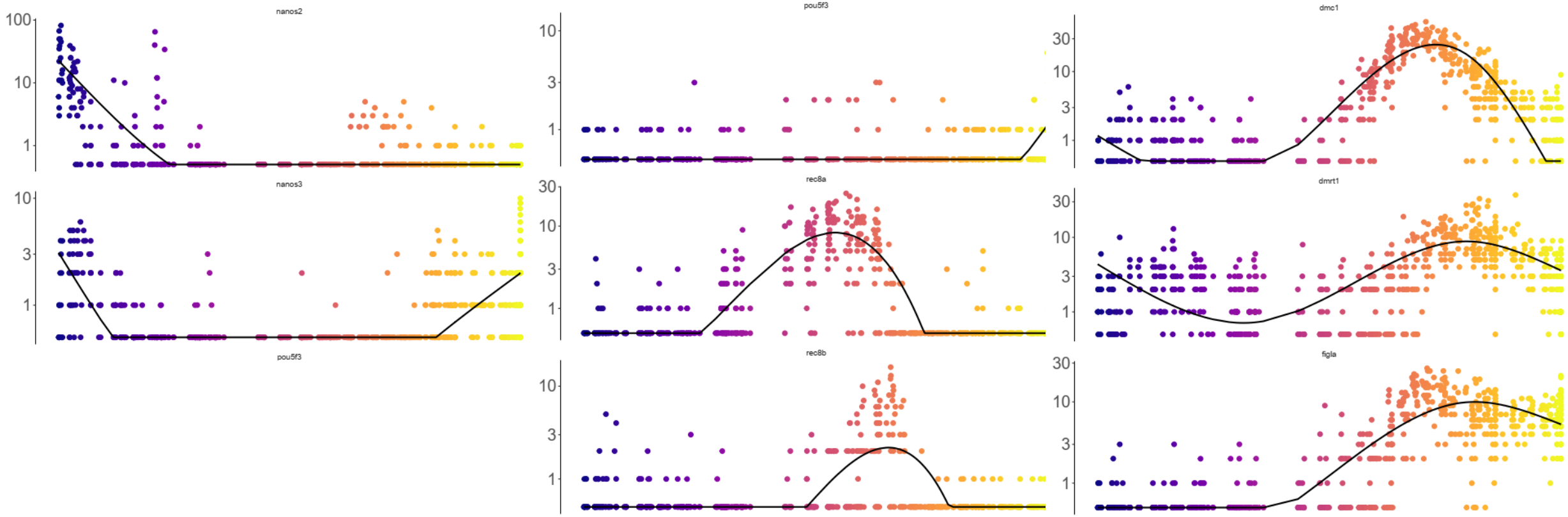
Differential expression analysis

In development, disease, and throughout life, cells transition from one state to another.

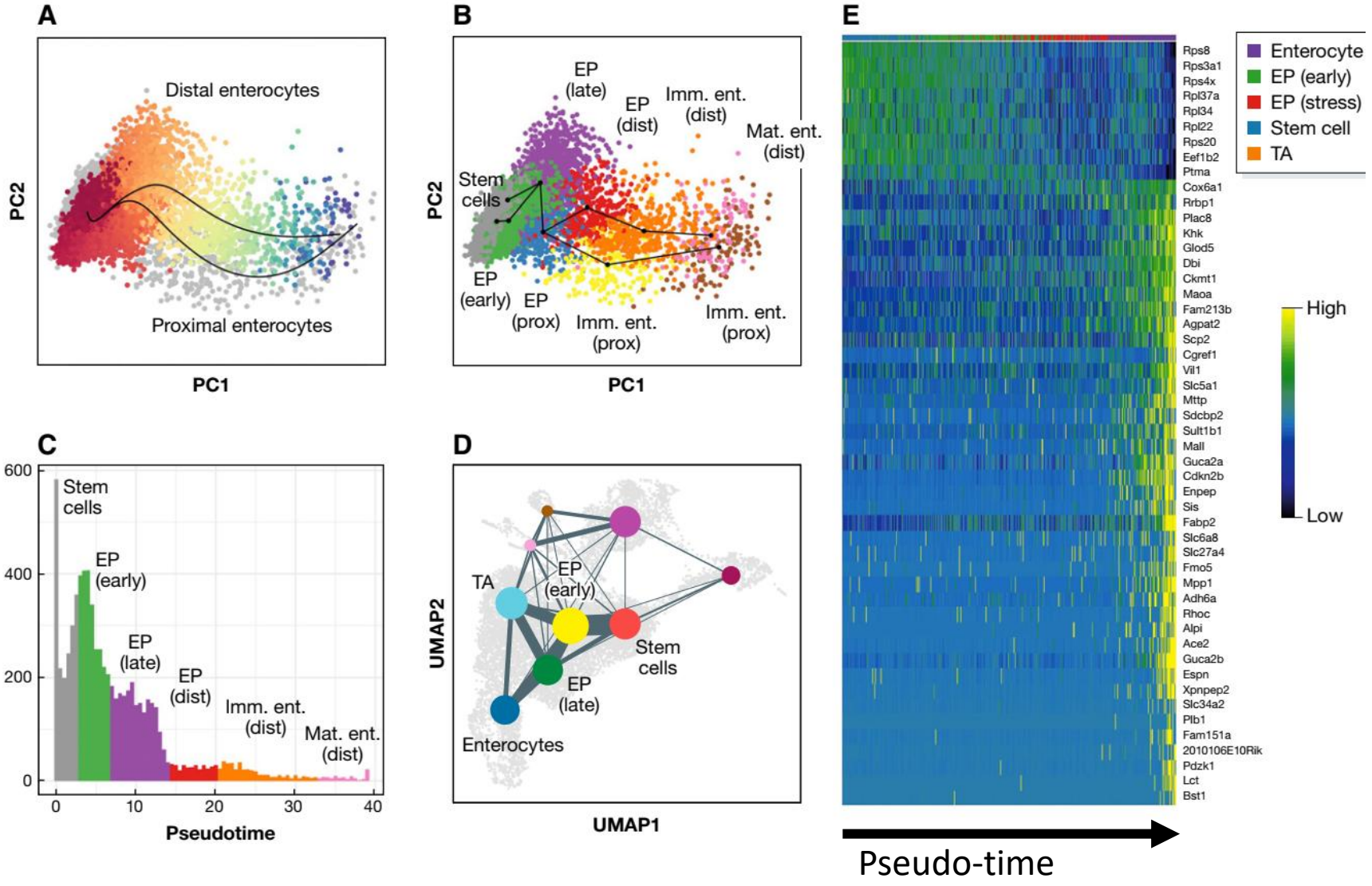
```
library(monocle3)
```



Trajectories (Monocle3)



Trajectory analysis

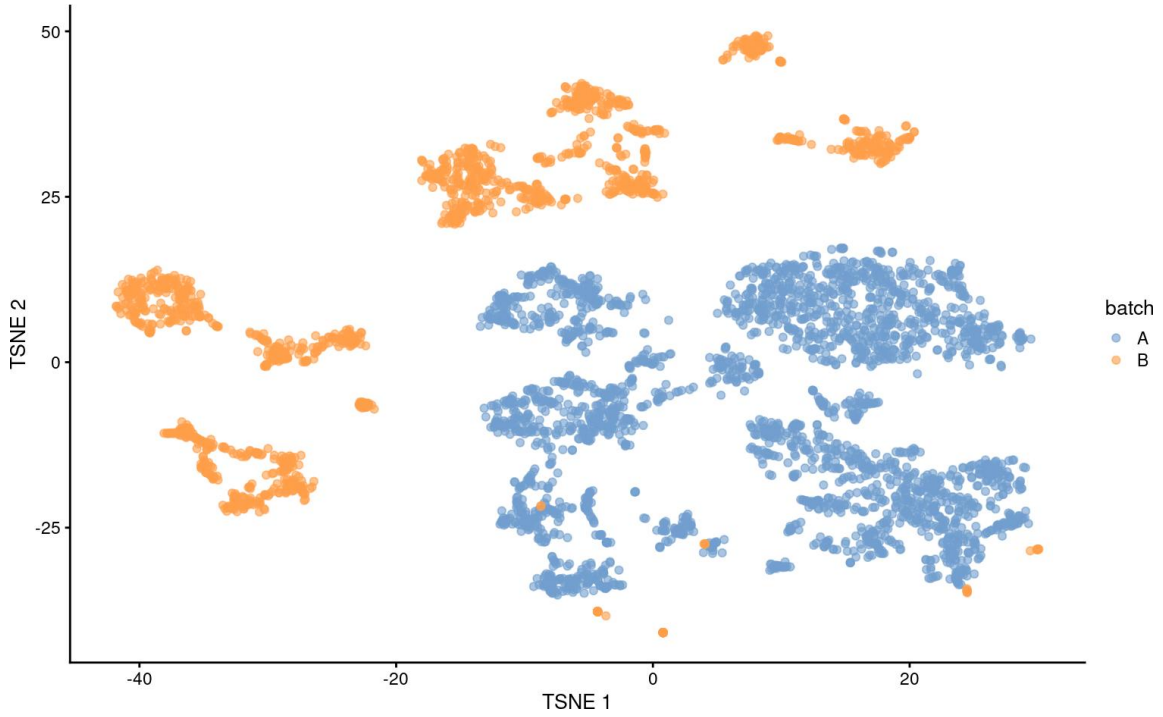


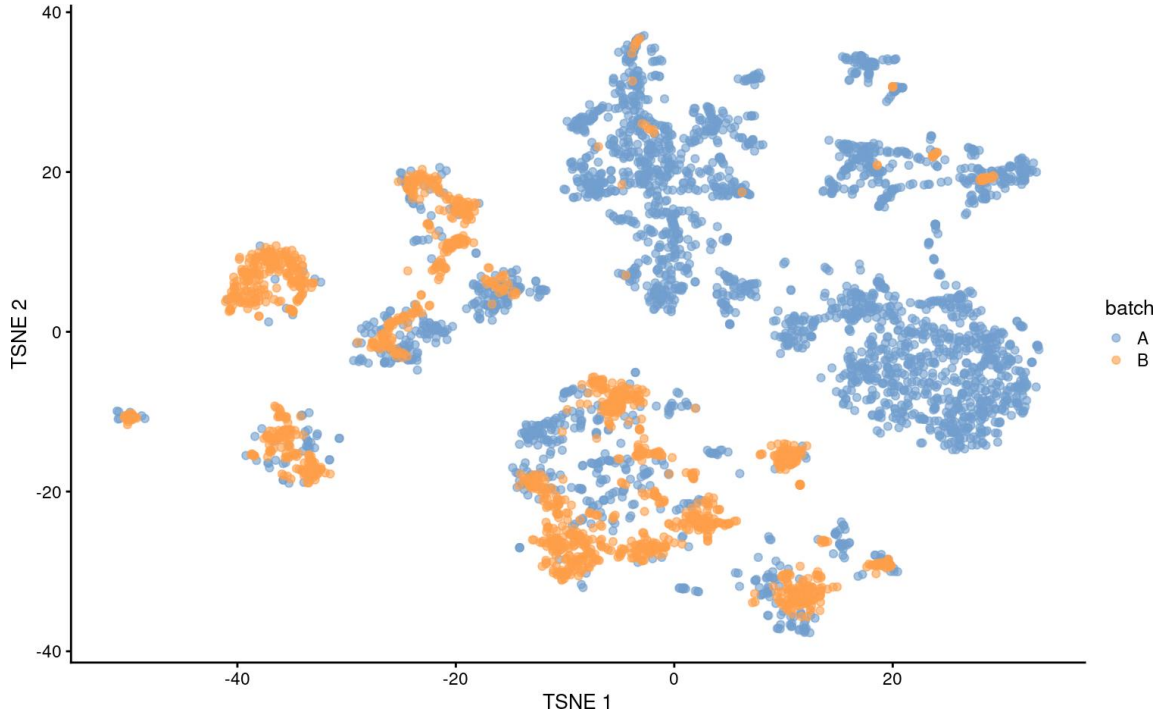
Batch correction (batchelor)

```
library(batchelor)
fastMNN()
mnnCorrect()
clusterMNN()
```

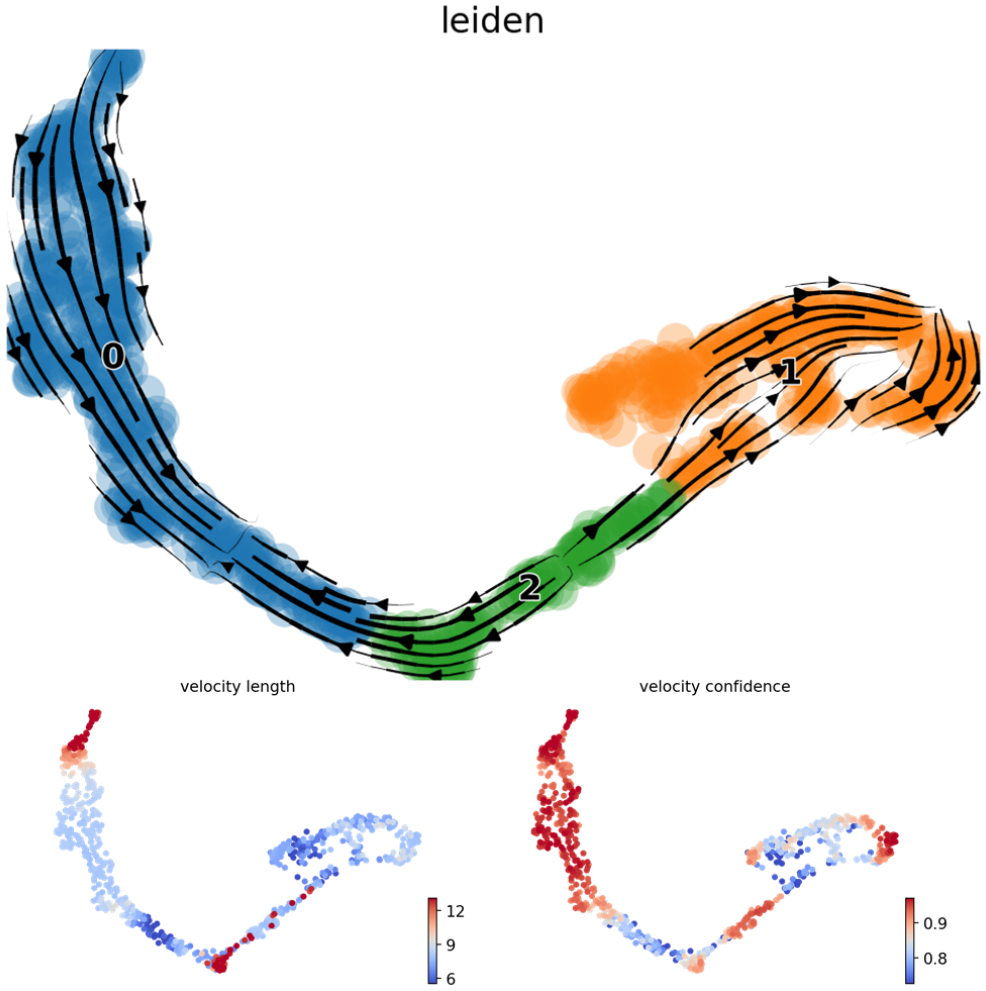
http://bioconductor.org/packages/release/bioc/html/batchelor.html

Mutual nearest neighbors (MNNs) are defined as pairs of cells - one from each batch - that are within each other's set of k nearest neighbors.





RNA Velocity Analysis (scVelo - python)



RNA velocity—the time derivative of the gene expression state—can be directly estimated by distinguishing between nascent (unspliced) and mature (spliced) mRNAs in common single-cell RNA sequencing protocols. RNA velocity is a high-dimensional vector that predicts the future state of individual cells on a timescale of hours.

RNA velocity aid the analysis of developmental lineages and cellular dynamics, particularly in humans.

La Manno et al., Nature, 2018.
Bergen et al. Molecular Systems Biology, 2021.

Single-cell RNA-Seq Resources

Datasets created by 10x Genomics: www.10xgenomics.com/datasets

Public available single-cell RNA-Seq databases:

- NCBI GEO/SRA: www.ncbi.nlm.nih.gov/geo
- EMBL Single Cell Expression Atlas: www.ebi.ac.uk/gxa/sc/home
- Single Cell Portal: singlecell.broadinstitute.org/single_cell
- CZ Cell x Gene Discover: cellxgene.cziscience.com
- PanglaoDB: panglaodb.se
- Allen Brain Cell Atlas: portal.brain-map.org/atlases-and-data/bkp/abc-atlas

References

https://www.10xgenomics.com
https://satijalab.org/seurat
https://scanpy-tutorials.readthedocs.io/en/latest/
https://cole-trapnell-lab.github.io/monocle3
https://github.com/theislab/scvelo
https://github.com/hbctraining/scRNA-seq
https://www.bioinformatics.babraham.ac.uk

IMB Bioinformatics Core
https://bc.imb.sinica.edu.tw