Single-cell RNA-seq Analysis in Python

June 2024

Course Overview

Day 1: Introduction to Single-cell RNA-Sequencing:

- Experimental design
- Processing raw reads to derive gene count matrices
- Setting up the environemnt
- Scverse ecosystem

Day 2: Data Processing and Analysis:

- Quality control
- Data normalization
- Feature selection
- Dimensionality reduction
- Clustering
- Batch correction and data integration

Day 3: Biological Data Interpretation

- Manual and automated cell type annotation
- Reference mapping
- Differential gene expression
- Gene set enrichment analyses

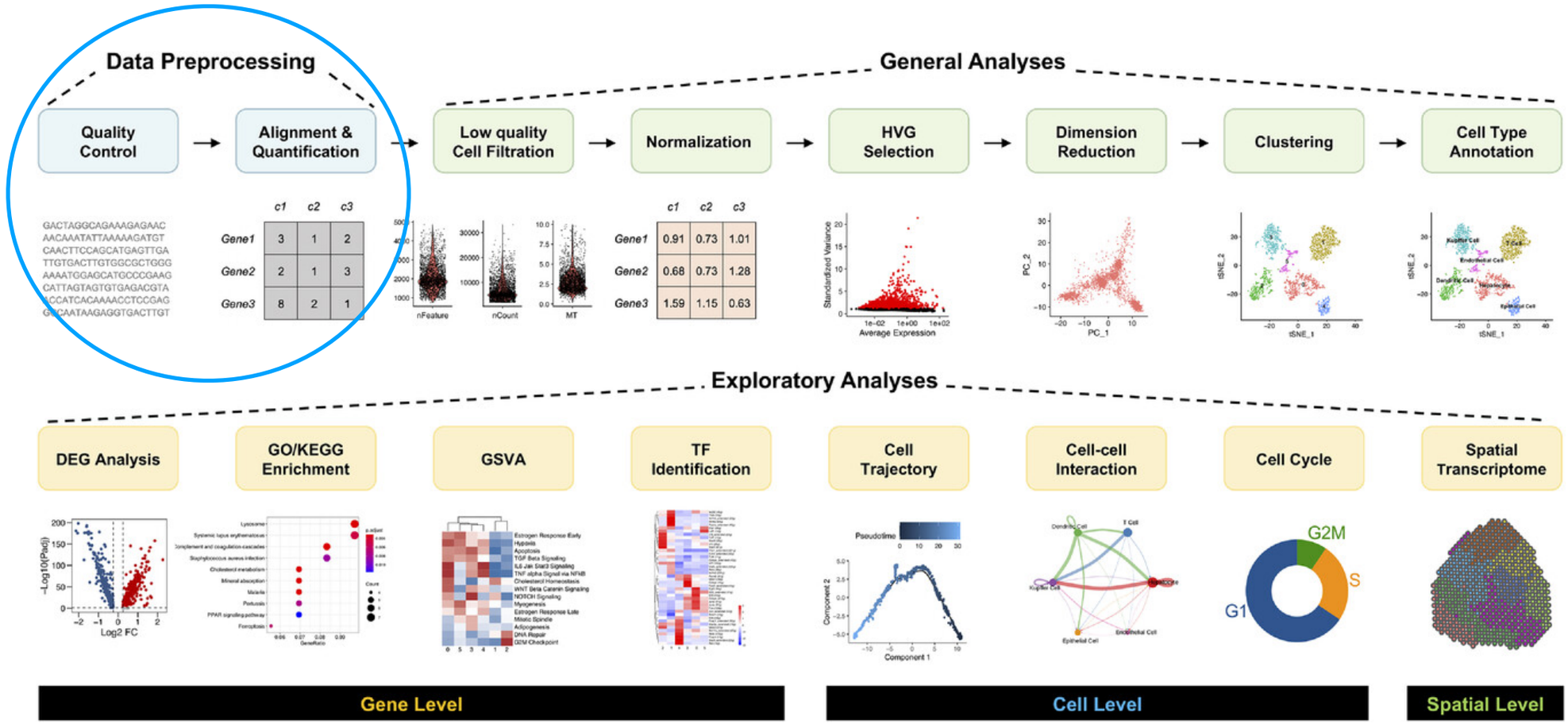
Day 4: Cellular Dynamics and Differentiation

- Peudotime analysis
- RNA velocity

References

- https://www.sc-best-practices.org/
- https://www.10xgenomics.com/support/software/cell-ranger/latest
- https://scverse.org/

Roadmap



Jovic et al., 2022, Clinical and Translation Medicine

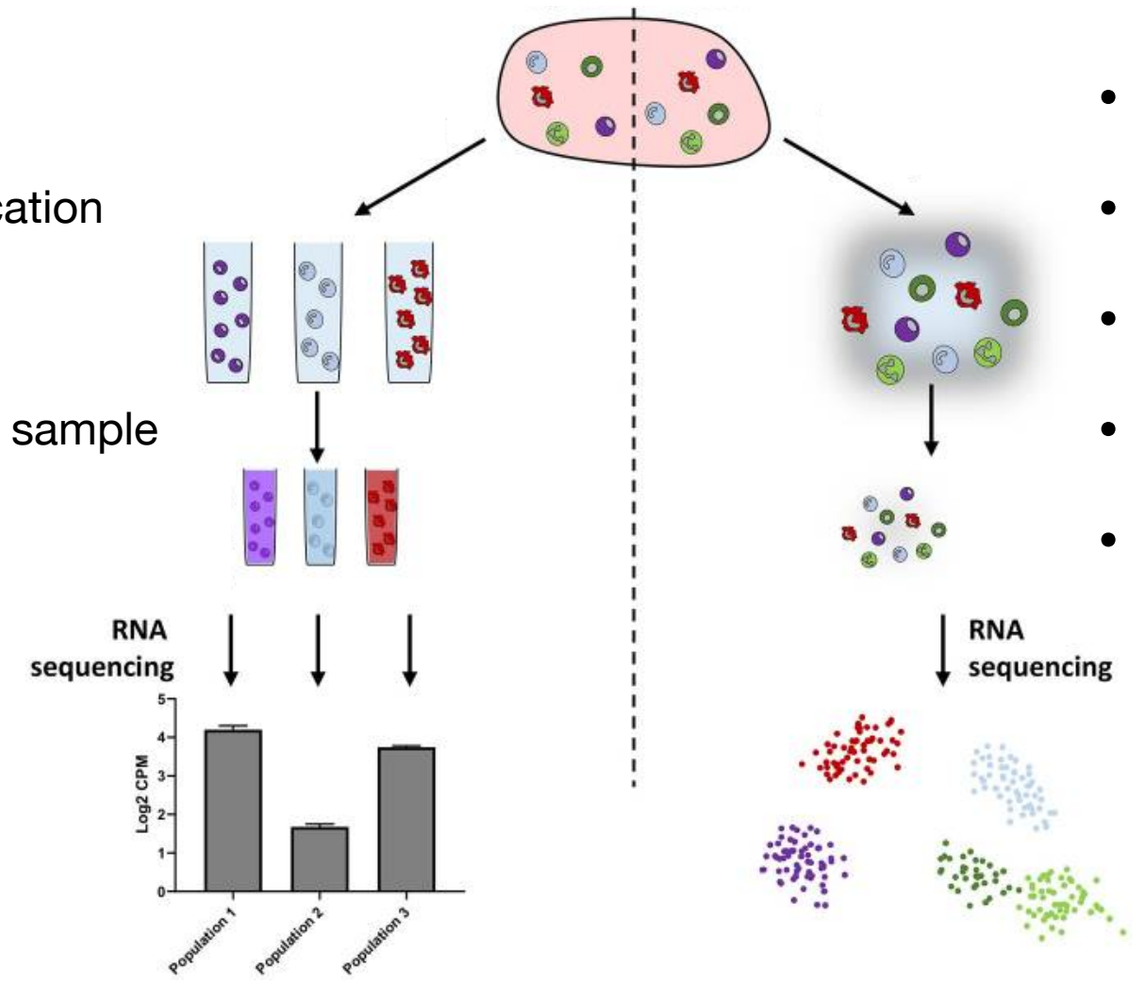
Single-cell vs. Bulk RNA-seq

Bulk

- No cell-type distinctions
- No heterogeneity quantification
- Small number of samples
- High number of genes per sample
- Optional dissociation

Single-cell

- Cell type detection
- heterogeneity quantification
- Lower number of genes per sample
- Many zeros
- Required dissociation



Carr et al., 2020, Front Med

Sample preparation

Cell condition

- Fresh viable cells
- Preserved sample (e.g. Cryopreserved or methanol fixated)
- Nuclear RNA from frozen tissues

Preparing cell suspensions

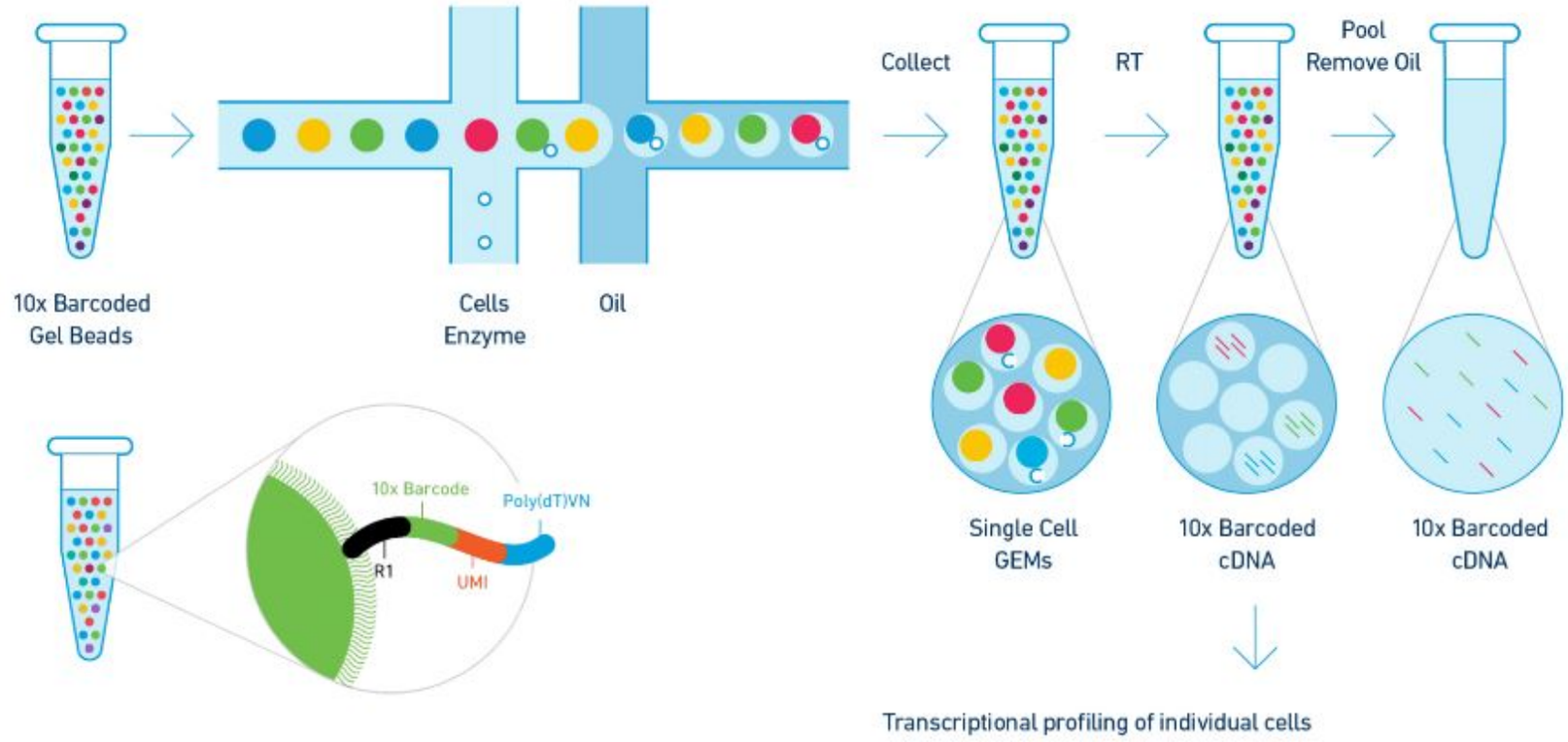
- Suspensions (e.g. blood samples) -> density centrifugation
- Solid tissues -> mechanical and enzymatic treatment

Single-cell capture

- Microfluidics
  - Drop-seq
  - inDrop
  - 10x Genomics Chromium
- Microwell
  - Seq-well
  - Smart-seq and Smart-seq2
- Microdissection or pipetting
- Fluorescence-activated cell sorting (FACS)

Sample Preparation

Single-cell Capture -> 10x Genomic Chromium



Transcriptome profiling

1) RNA molecule capture Poly(A)-tailed RNA

2) Reverse transcription Accompanied by adding single-cell-specific barcodes

3) Amplification

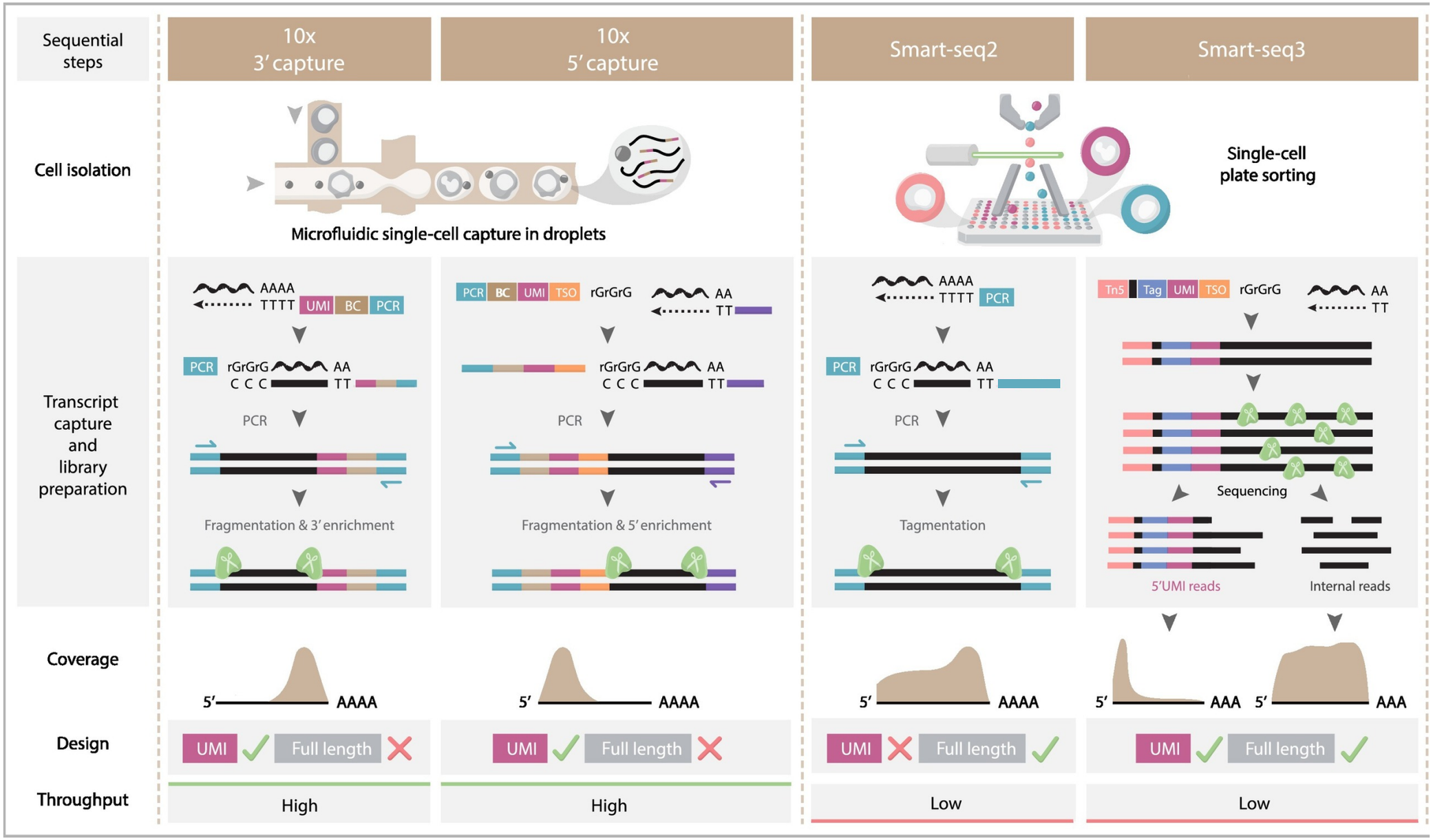
- PCR (e.g. Drop-seq, 10x, Smart-seq)
- In vitro transcription (e.g. CEL-seq, inDrop)

4) Library preparation

- 3’- or 5’-based libraries
- Full-length libraries

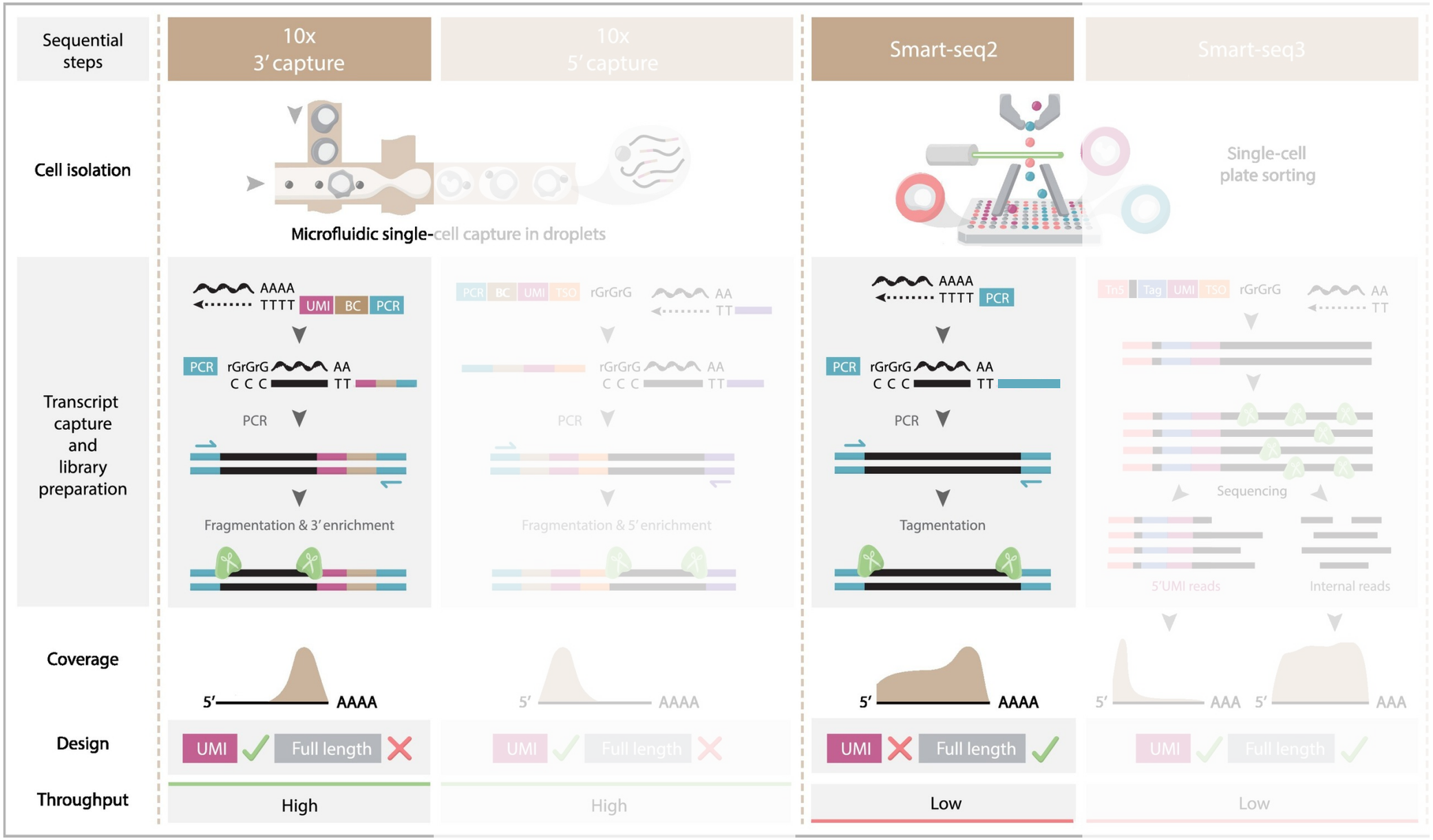
5) Sequencing

Library prep. of tag-based vs. full-length methods



Conte et al., 2023, TiG

Library prep. of tag-based vs. full-length methods



Conte et al., 2023, TiG

Tradeoff between throughput and sequencing coverage

↑ **Throughput** ↓ **Coverage**

- Generating a single-cell atlas of a biological specimen
- Using Unique Molecular Identifier (UMI) to reduce amplification noise

↓ **Throughput** ↑ **Coverage**

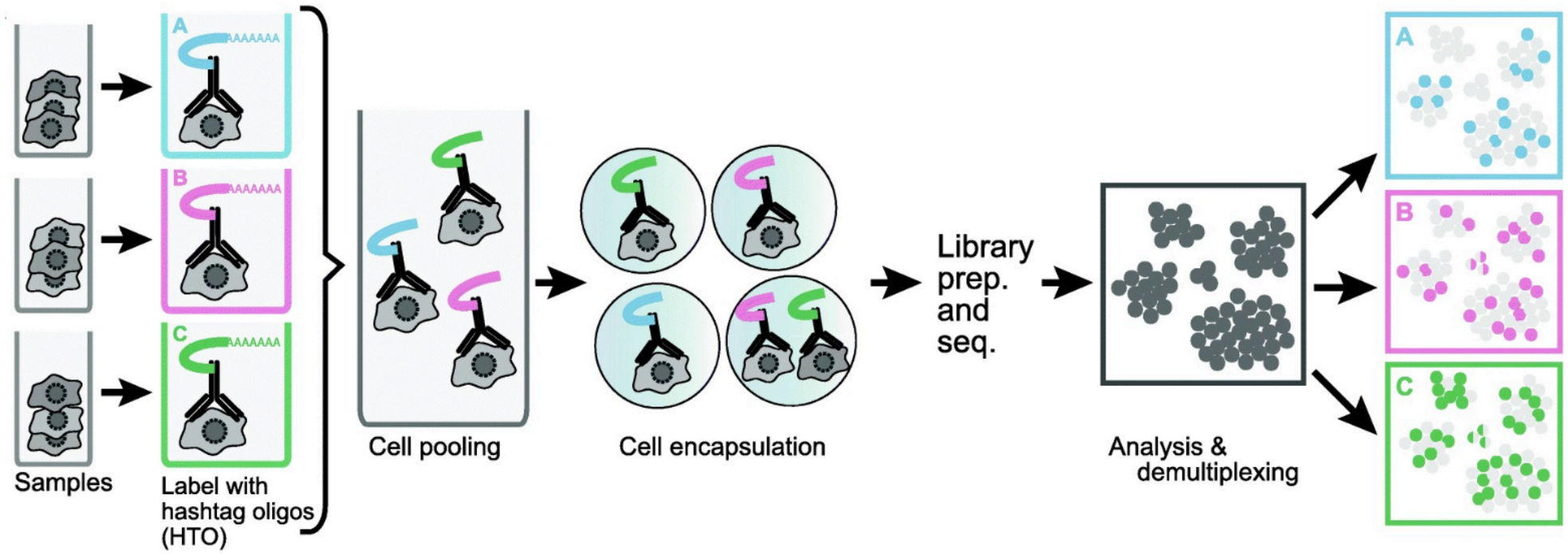
- Studying alternative splicing
- High power to detect somatic mutations/genetic variants
- Costly
- Suboptimal quantification due to PCR amplification bias

Conte et al., 2023, TiG

Cell Hashing

It resolves problems of

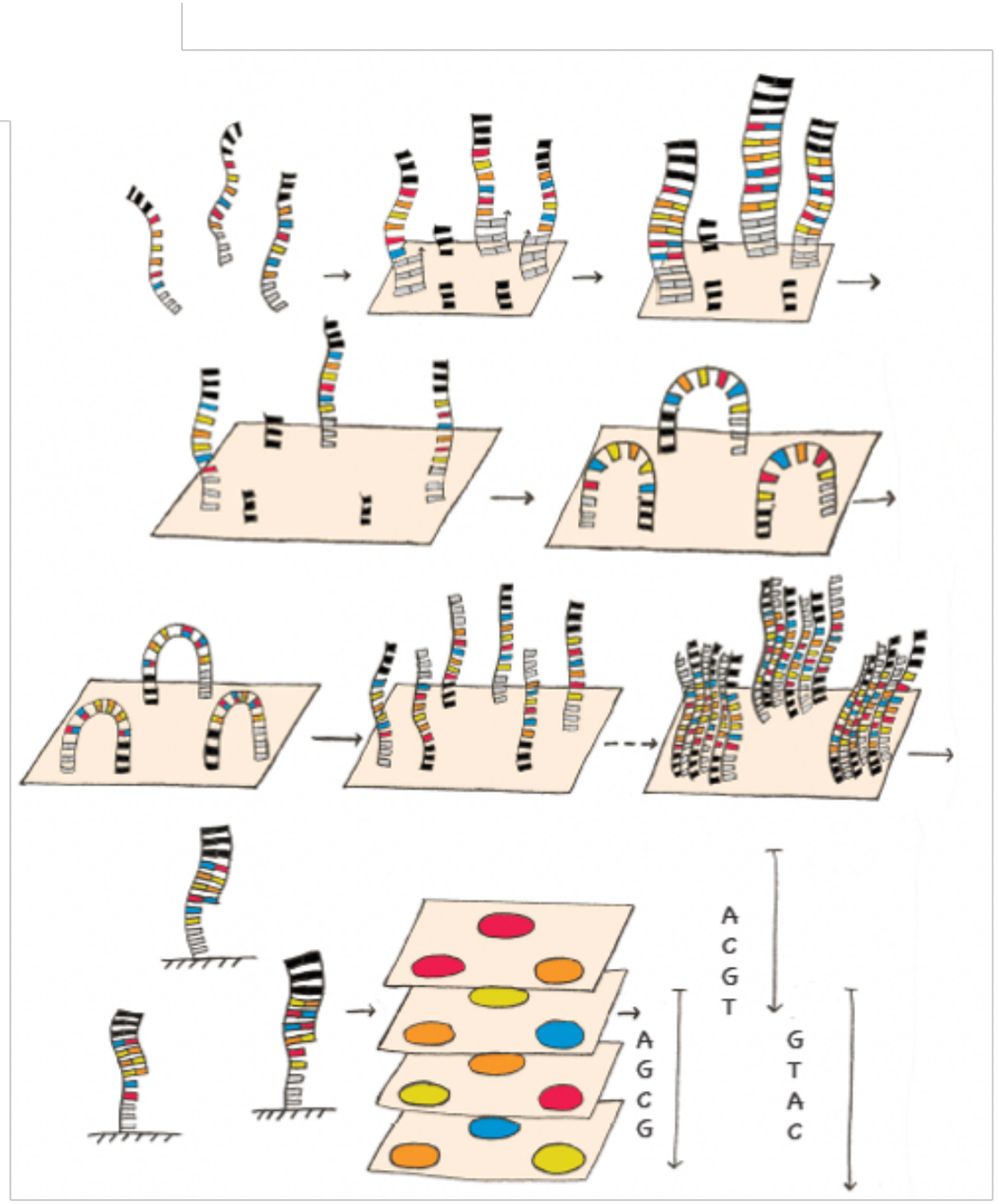
- sample-specific batch effects
- cell multiplets
- experimental costs



Stoeckius

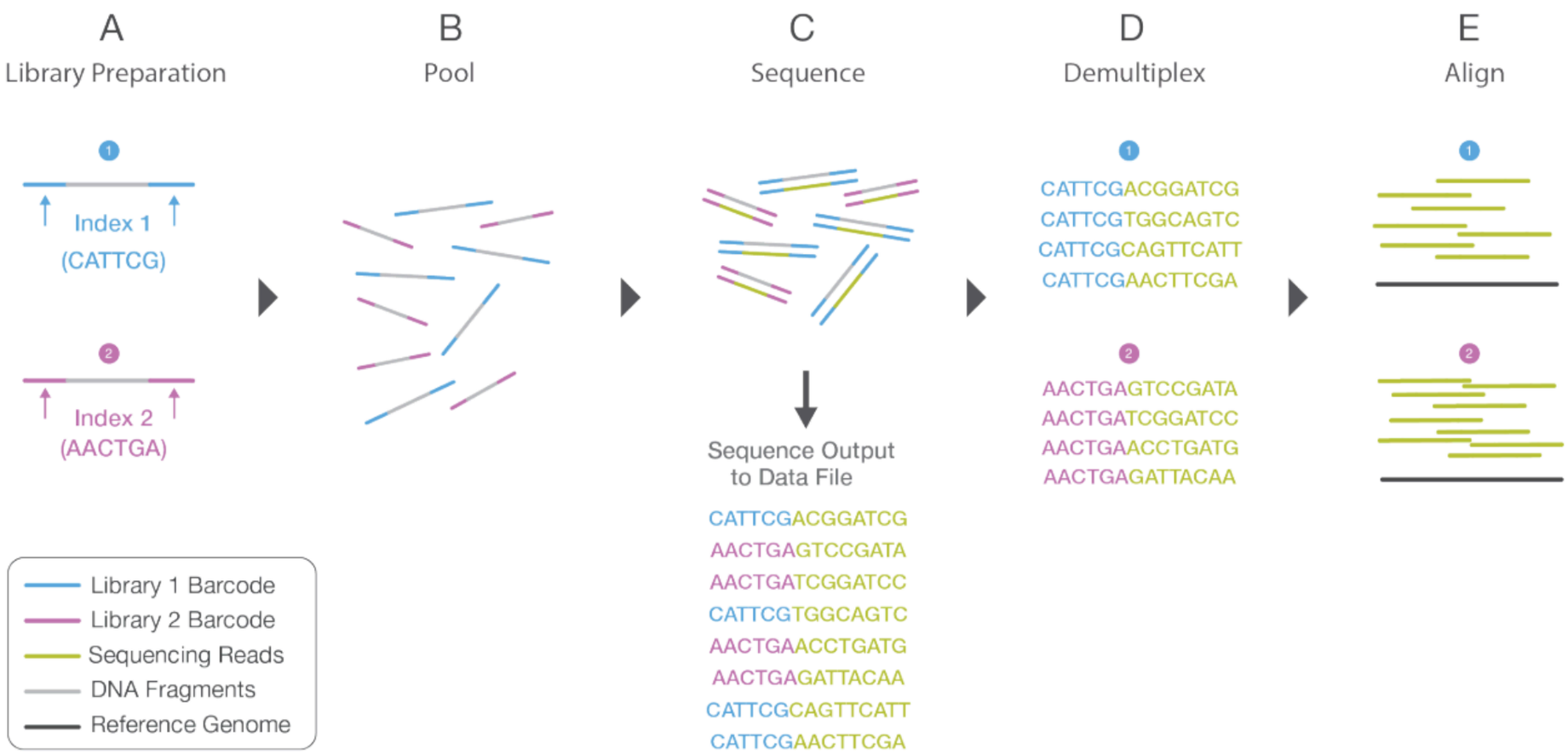
Stoeckius et al., 2018, Genome Biol.

Illumina Sequencing



Kvastad, L. (2021). The spatial context – through the lens of method development [PhD thesis]. KTH Royal Institute of Technology.

Overview



Illumina

Preprocessing

1. Assign reads to cells
2. Alignment
3. Quantification
4. Cell calling

Available tools

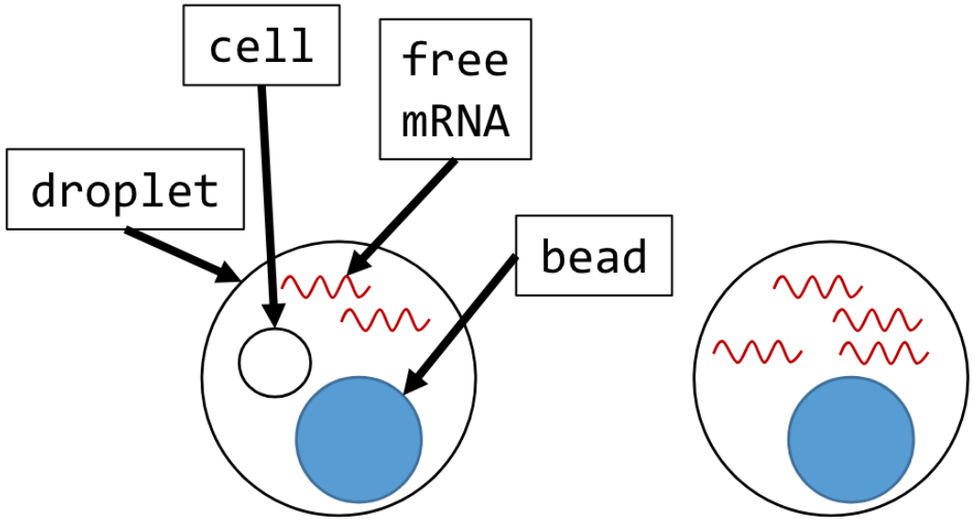
1. Cellranger (from 10x Genomics)
2. STARsolo
3. Alevin

Cell Ranger's Gene Expression Algorithm

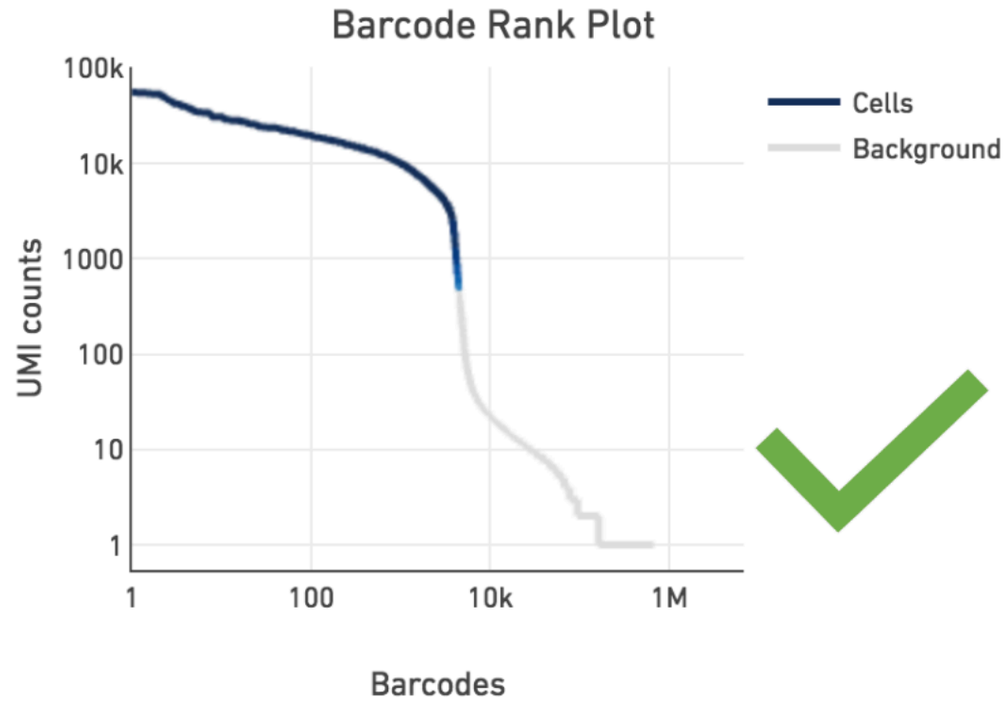
1. Read trimming -> to remove TSO and poly-A sequences
2. Genome alignment -> using the STAR aligner
    - categorizing reads into exonic, intornic and intergenic
3. MAPQ adjustment
4. Transcriptome alignment
5. 10x barcode correction
6. UMI counting
7. Calling cell barcode

10x Genomics

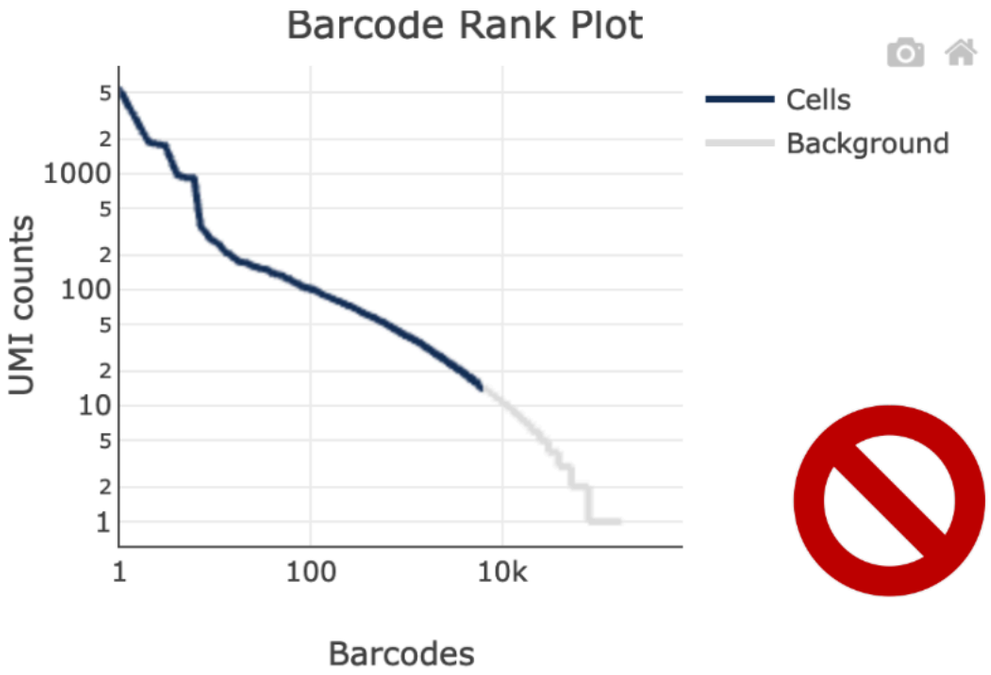
Cell calling



Cell
captured



Only
background
captured

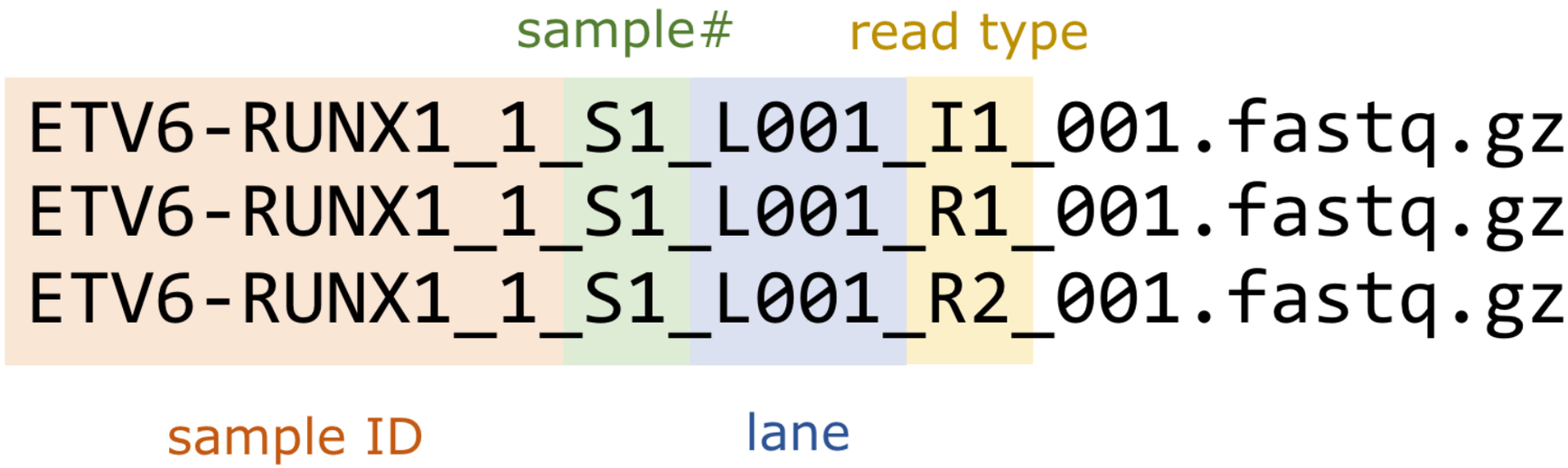


https://sib-swiss.github.io/single-cell-training

Practical guide to cellranger count

The input data should be generated by **10x Genomics Chromium** system

1. The FASTQ files



2. Reference transcriptome

Ref: https://www.10xgenomics.com/support/software/cell-ranger/latest/tutorials/cr-tutorial-ct
https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/inputs/cr-inputs-overview

Practical guide to cellranger count

The instruction for connecting to the remote Linux server is provided in **cellranger.txt**

1. The FASTQ files are located in /**course/reads/**

2. Reference transcriptome is located in **/course/chr21_ref/**

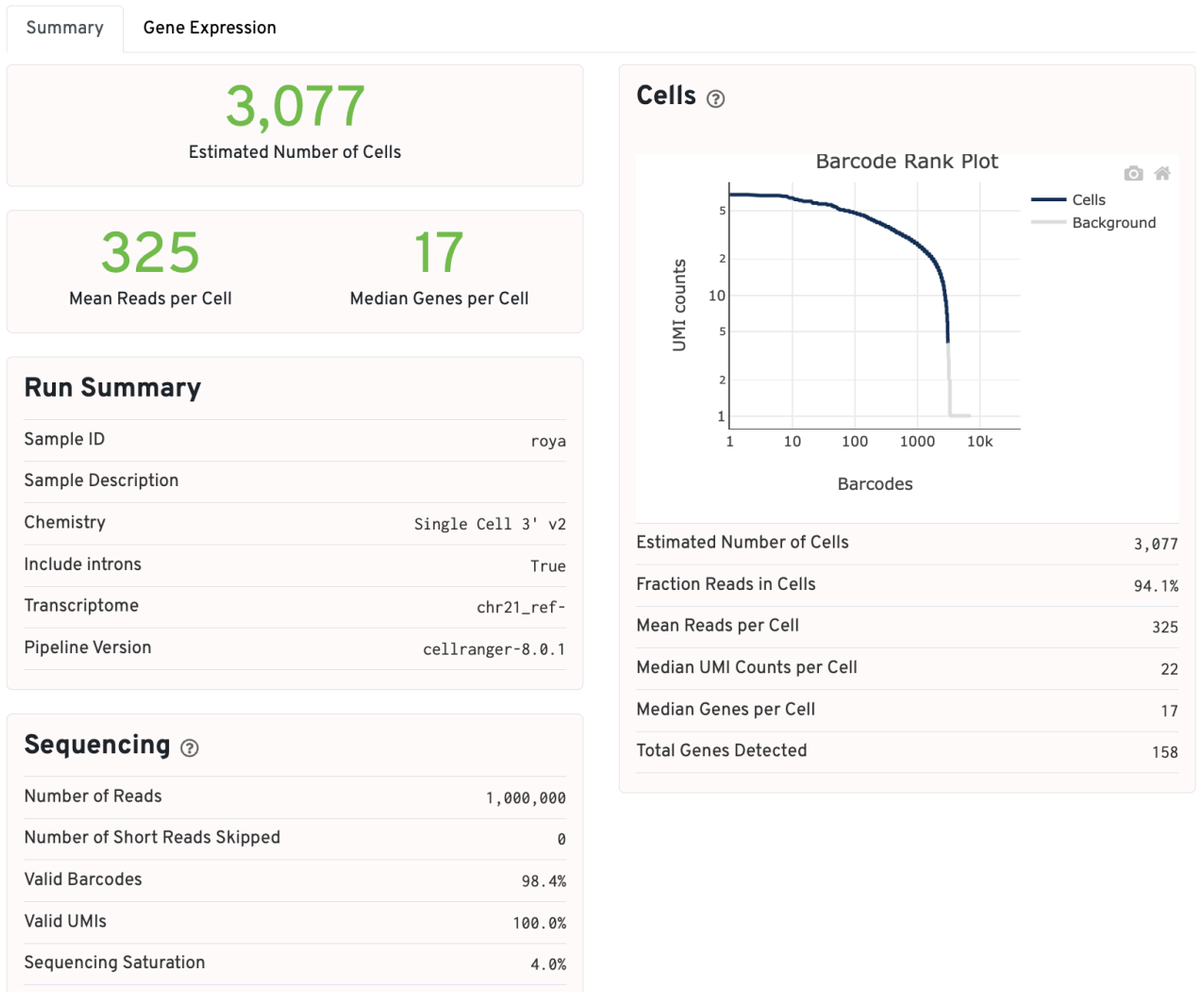
```
cellranger count --id <your_desired_name> --fastqs /course/
reads/ --transcriptome /course/chr21_ref/ --create-bam true
```

e.g. cellranger count --id course_count --fastqs /course/
reads/--transcriptome /course/chr21_ref –create-bam true

Ref: https://www.10xgenomics.com/support/software/cell-ranger/latest/tutorials/cr-tutorial-ct
https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/inputs/cr-inputs-overview

Output file of cellranger count

analysis	filtered_feature_bc_matrix.h5	possorted_genome_bam.bam
cloupe.cloupe	metrics_summary.csv	possorted_genome_bam.bam.bai
filtered_feature_bc_matrix	molecule_info.h5	raw_feature_bc_matrix
raw_feature_bc_matrix.h5	**web_summary.html**	



https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/outputs/cr-outputs-web-summary-count

Output file of cellranger count

```
analysis                                filtered_feature_bc_matrix.h5   possorted_genome_bam.bam
cloupe.cloupe                           metrics_summary.csv             possorted_genome_bam.bam.bai
filtered_feature_bc_matrix              molecule_info.h5                raw_feature_bc_matrix
raw_feature_bc_matrix.h5                web_summary.html
```

```
barcodes.tsv.gz  features.tsv.gz  matrix.mtx.gz
```

Analysis frameworks and tools

- Bioconductor
- Seurat

R

- Scverse

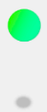
Python

CORE PACKAGES

anndata
Standard for annotated matrices

mudata
Multimodal data format

scanpy
Single-cell analysis framework

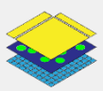


muon
Multi-omics analysis framework

scvi-tools
Single-cell machine learning framework

scirpy
Single-cell immune sequencing analysis framework

squidpy
Spatial single-cell analysis



spatialdata
Spatial data format

Setup

Install miniconda and create a conda environment

Choose the installer for Windows. Make sure to select the version that corresponds to your Python preference (**Python 3.11**) and your system architecture (64-bit or 32-bit). https://docs.anaconda.com/free/miniconda/miniconda-other-installer-links/

```
conda create -n newsc python=3.11
```

```
conda activate newsc
```

Install required packages

Download **requirement01.txt** file from env folder
For Mac users: Open the terminal and navigate to the directory containing the .txt file (e.g. `cd Downloads/`)

For Windows users: Open Anaconda Prompt (minoconda3) and navigate to the directory containing the .txt file (e.g. `cd Downloads/`)

```
pip install -r requirements01.txt
```

You might need to install jupyter lab separately: **jupyter lab**

```
conda install -c conda-forge jupyterlab
```

Setup

RESEARCH ARTICLE | 17 JUNE 2019

Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis

FREE

Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J. Theis, Heiko Lickert, Mostafa Bakhti

+ Author and article information

Development (2019) 146 (12): dev173849.

Download the example dataset from GEO - accession number: GSE132188

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132188

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE132188_RAW.tar	410.5 Mb	(http)(custom)	TAR (of TAR)
GSE132188_adata.h5ad.h5	318.2 Mb	(ftp)(http)	H5

SRA Run Selector

Raw data are available in SRA

Processed data provided as supplementary file

Processed data are available on Series record

Scverse community

Scverse community meetings happen **every second Tuesday at 6pm CET** and are open to everyone!

https://hackmd.io/VfVLKb3ETGKN2j_7tn8ZJQ?view

https://twitter.com/scverse_team

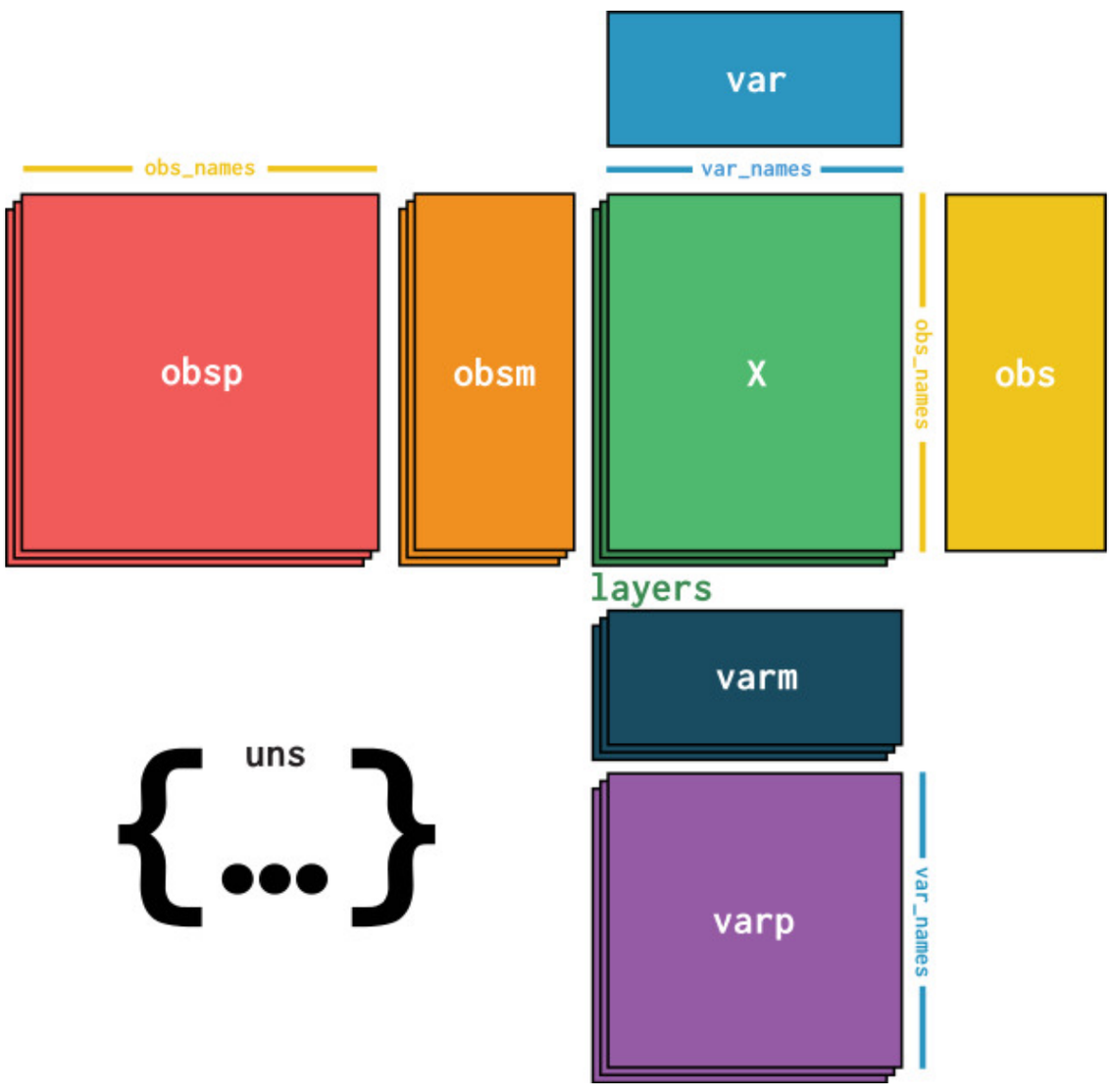
Scanpy

Scanpy toolkit

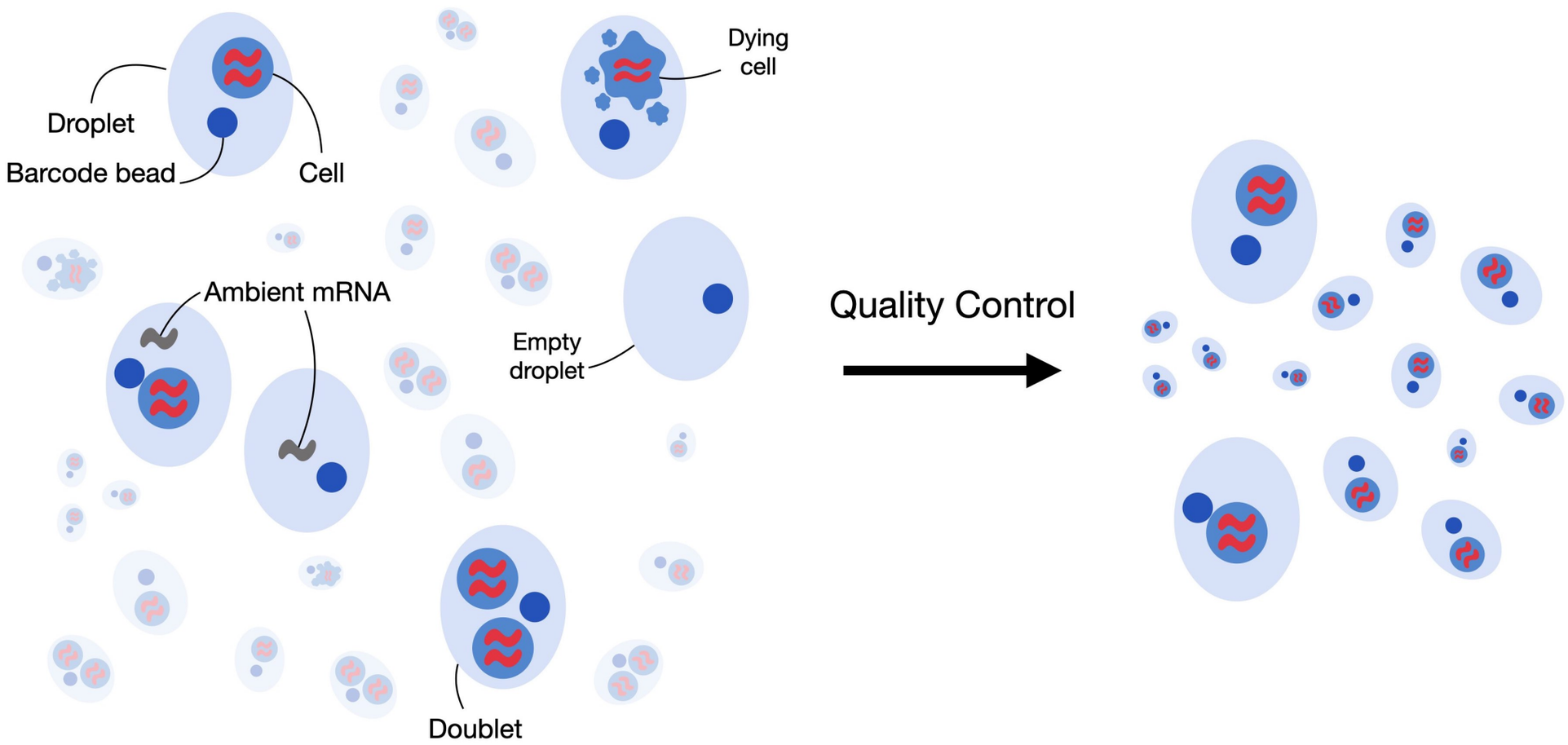
- Preprocessing `scanpy.pp`
- Tools `scanpy.tl`
- Plotting `scanpy.pl`

https://scanpy.readthedocs.io/en/latest/

AnnData


var
obs_names
var_names
obsp
obsm
X
obs_names
obs
layers
uns
{...}
varm
varp
var_names

Quality Control

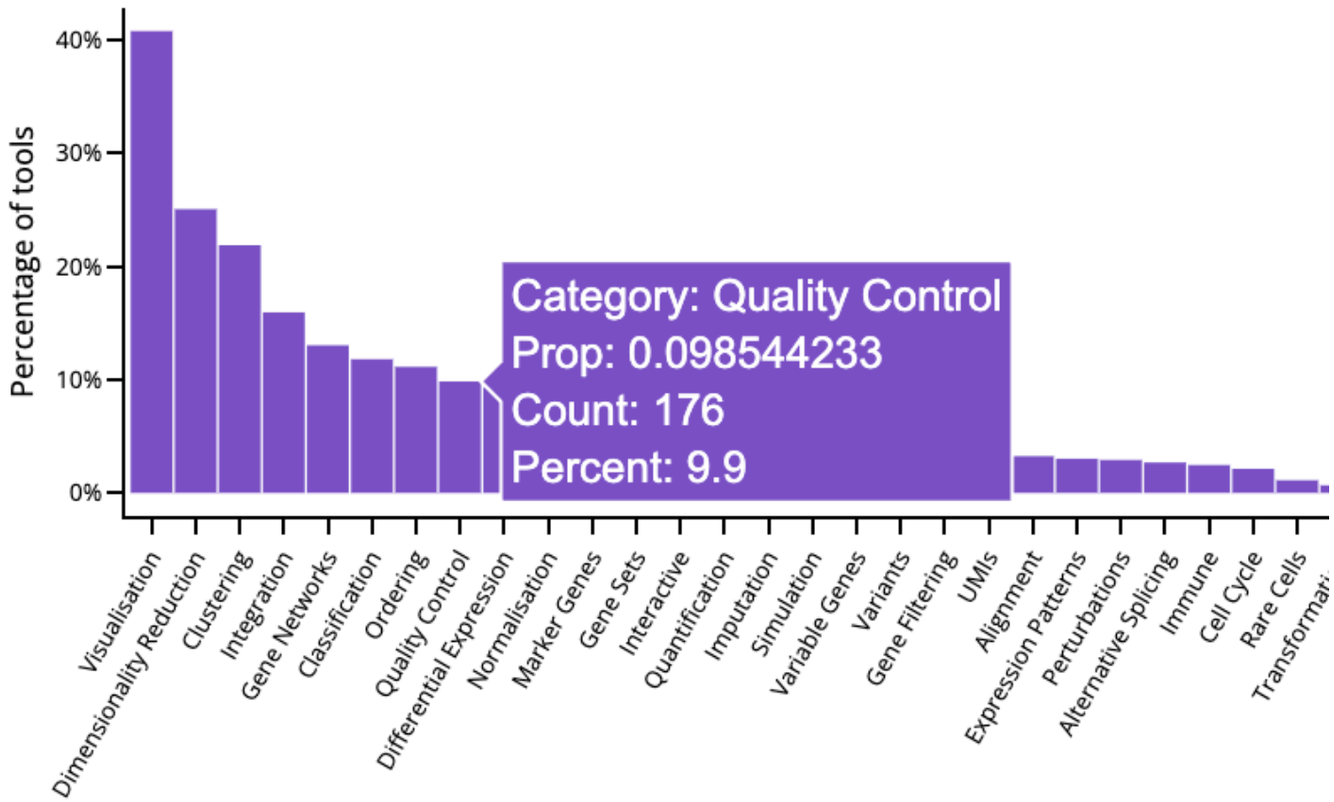


https://www.sc-best-practices.org/

Quality Control -> 1. Filtering Low-quality Cells

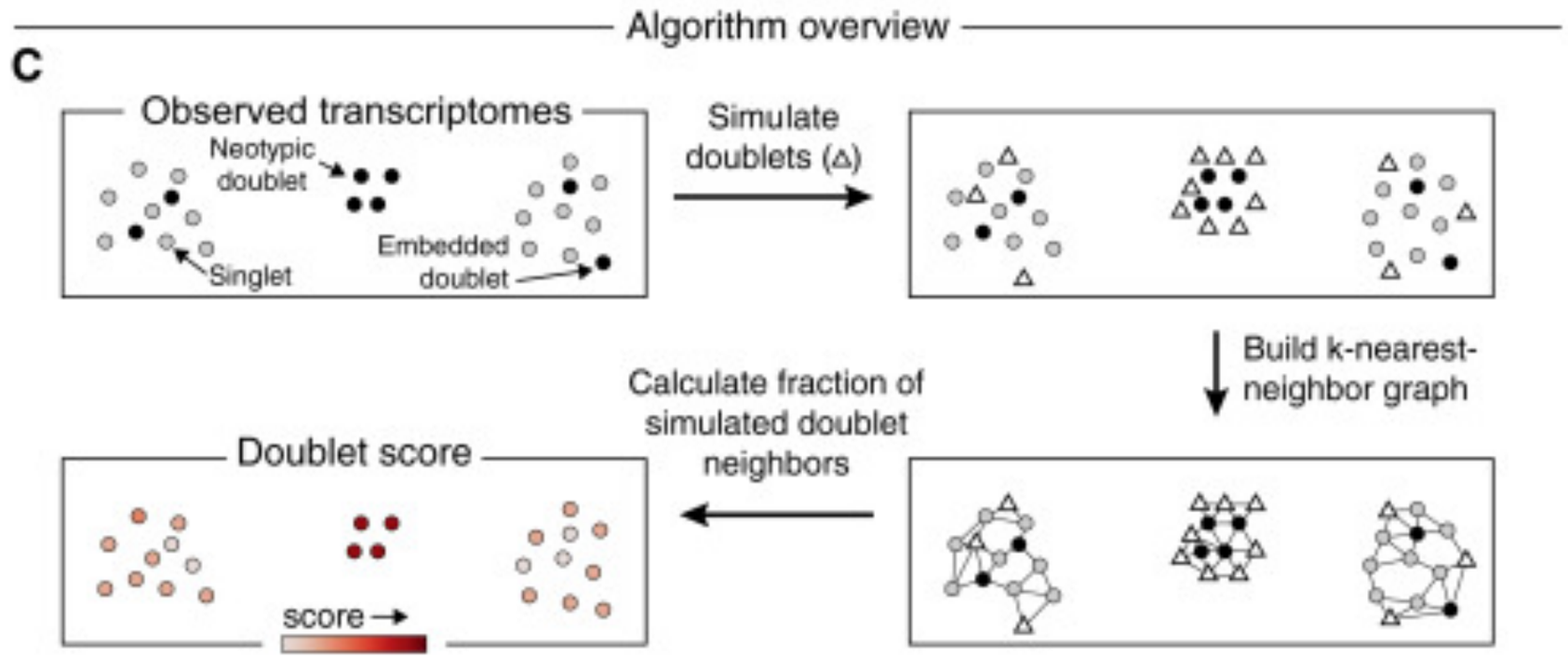
- Number of counts per barcode (count depth)
- Number of genes per barcode
- Fraction of counts from mitochondrial genes per barcode

Quality Control -> 2. Doublet Detection



https://www.scrna-tools.org/

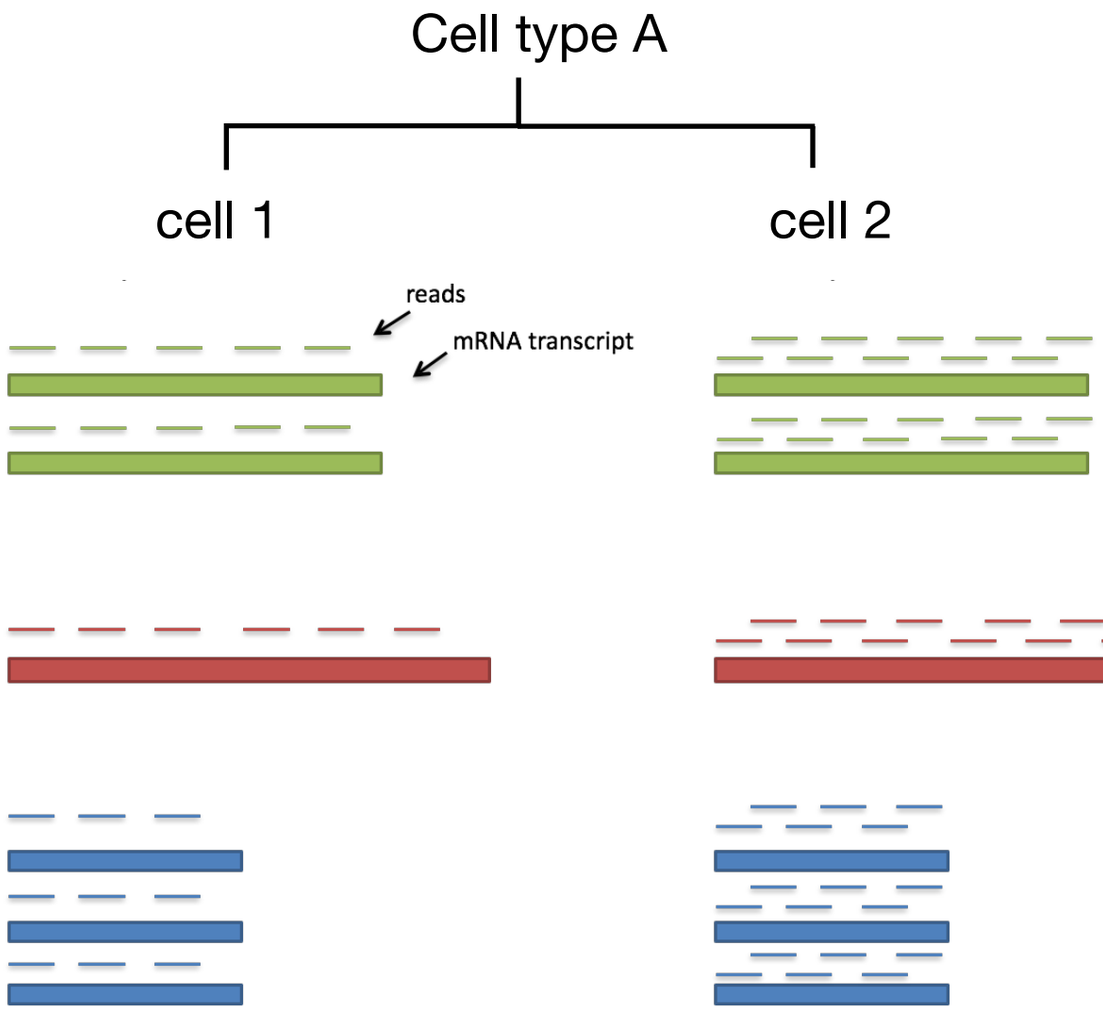
Quality Control -> 2. Doublet Detection -> Scrublet



Wolock et al., 2019, Cell Syst.

Normalization

Why do we need normalization?



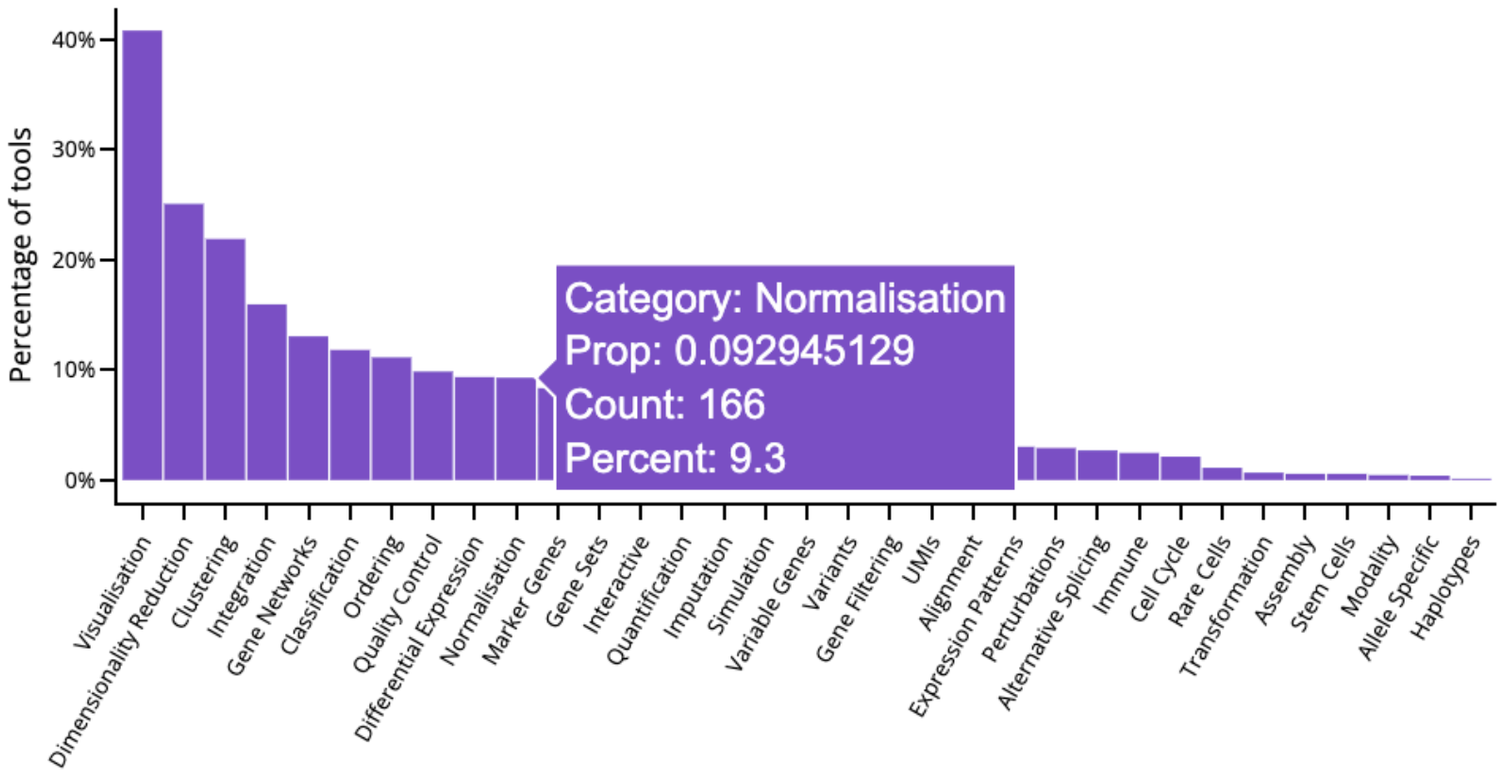
MI Love: RNA-seq statistical analysis

https://hbctraining.github.io/

Normalization

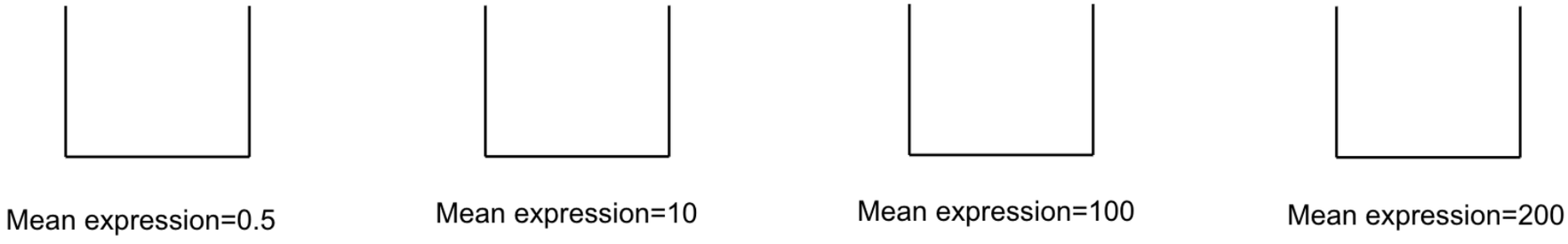
Shifted logarithm transformation $$f(y) = \log(\frac{y}{s} + y_0)$$

$$s_c = \frac{\sum_g y_{gc}}{L}$$



Feature Selection

Dispersion-based (“Seurat”, “Cell-ranegr”)



A. Calculate dispersion of each gene in each bin
B. Calculate the mean and the standard deviation of the dispersions in each bin
C. Normalise the dispersion of each gene by using the mean and the standard deviation from b
D. Genes within each bin are ranked based on their normalized dispersion values --> Highly variable genes

Variance-based (“SeuratV3”)

A. Expects raw counts (not normalised or log-transformed)
B. Variance-stabilising transformation is applied to the raw data.
C. Highly variable genes are selected based on the variance of the standardised values (mean-variance relationship is taken into account)