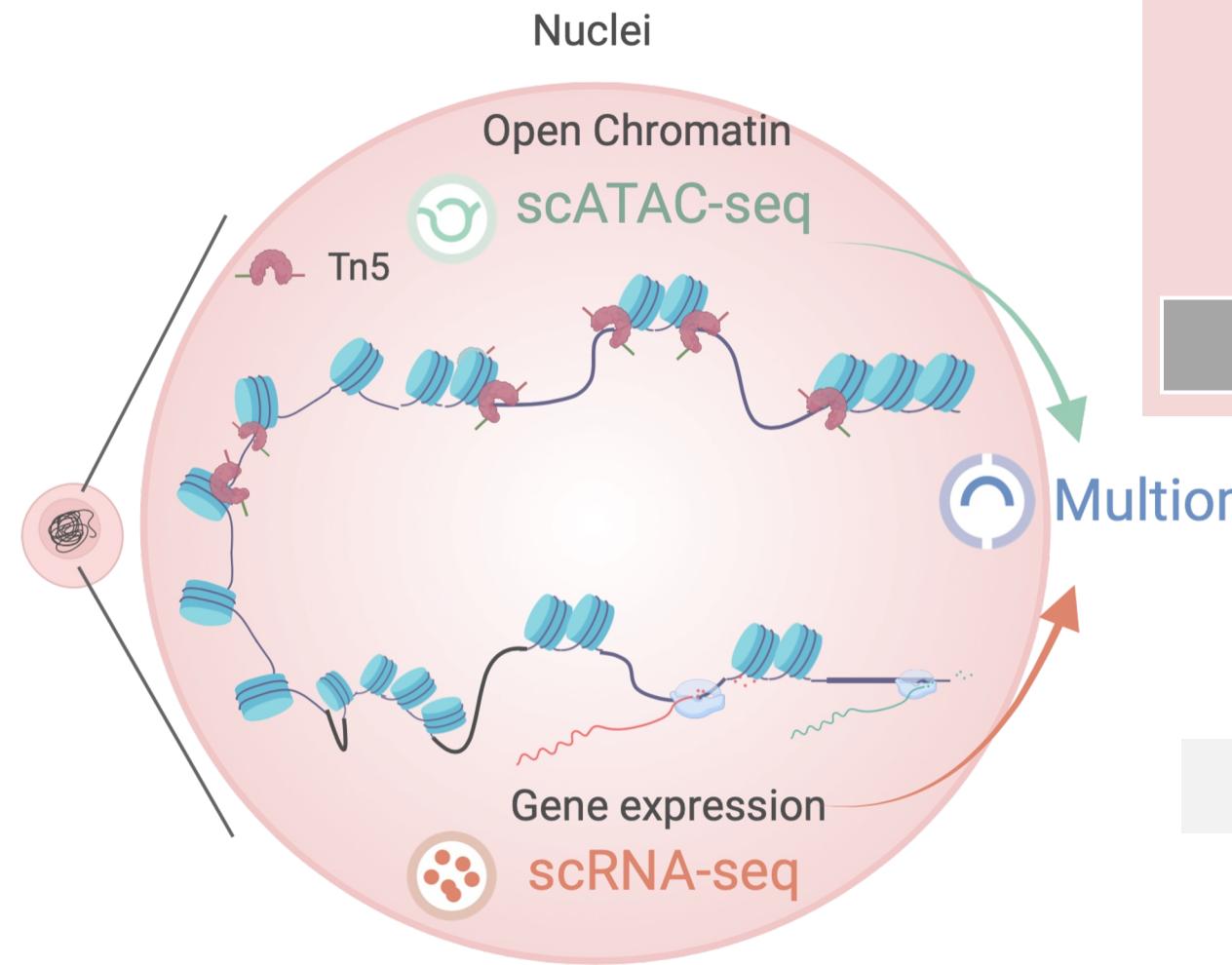


Workshop

Single-cell ATAC-seq



Part I: Introduction to scATAC-seq analysis

María Lucía Romero
PhD student in Bioinformatics, Epigenetics Biomedical, IDIBAPS

cnag

centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

CRG
Centre for Genomic Regulation

IDIBAPS
Institut D'Investigacions Biomèdiques August Pi i Sunyer

What are we going to learn?

What Single-Cell ATAC-seq (scATAC-Seq) is

What kind of information scATAC-seq can provide us?

Multiome (scRNA-seq+ scATAC-seq)

Why Multiome?

How Multiome works:

Focused on scATAC-seq

scATAC-seq Pipeline workflow analysis

1

2

3

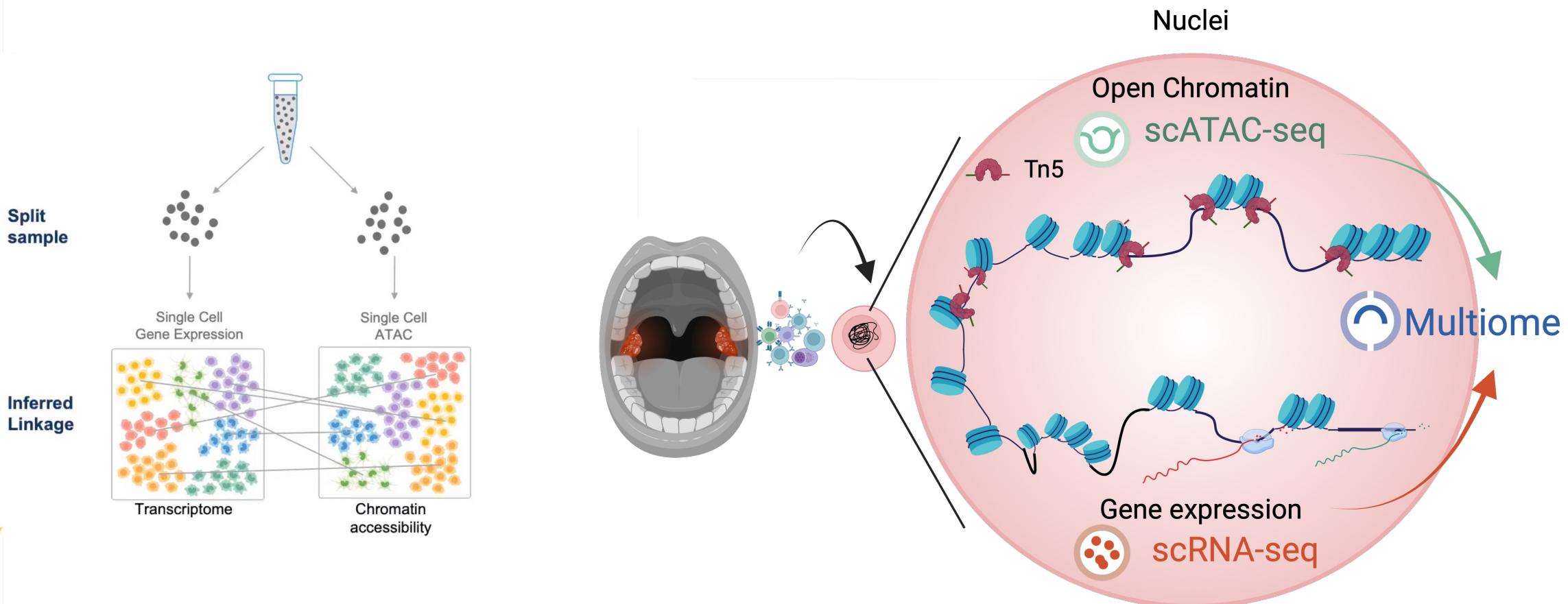
4

5

Introduction

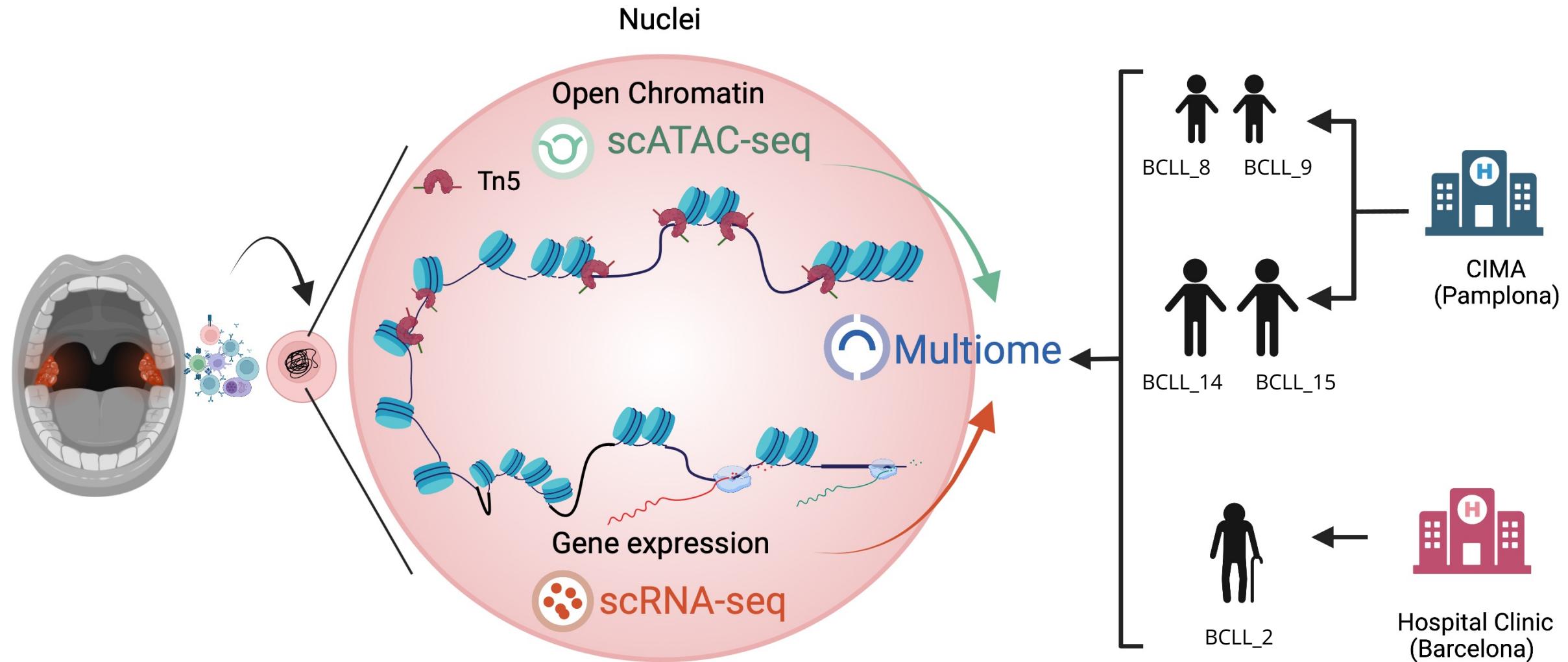
Single cell Multiome (Gene expression + ATAC-seq)

Consist in the simultaneous assessment of single cell RNA-seq and ATAC-seq for each individual cell



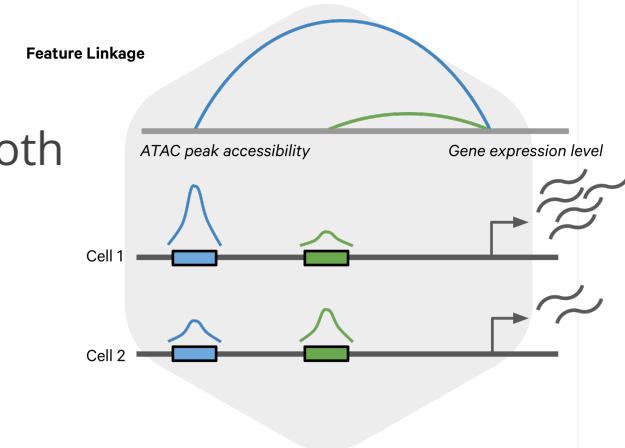
Introduction

Single cell Multiome (Gene expression + ATAC-seq)

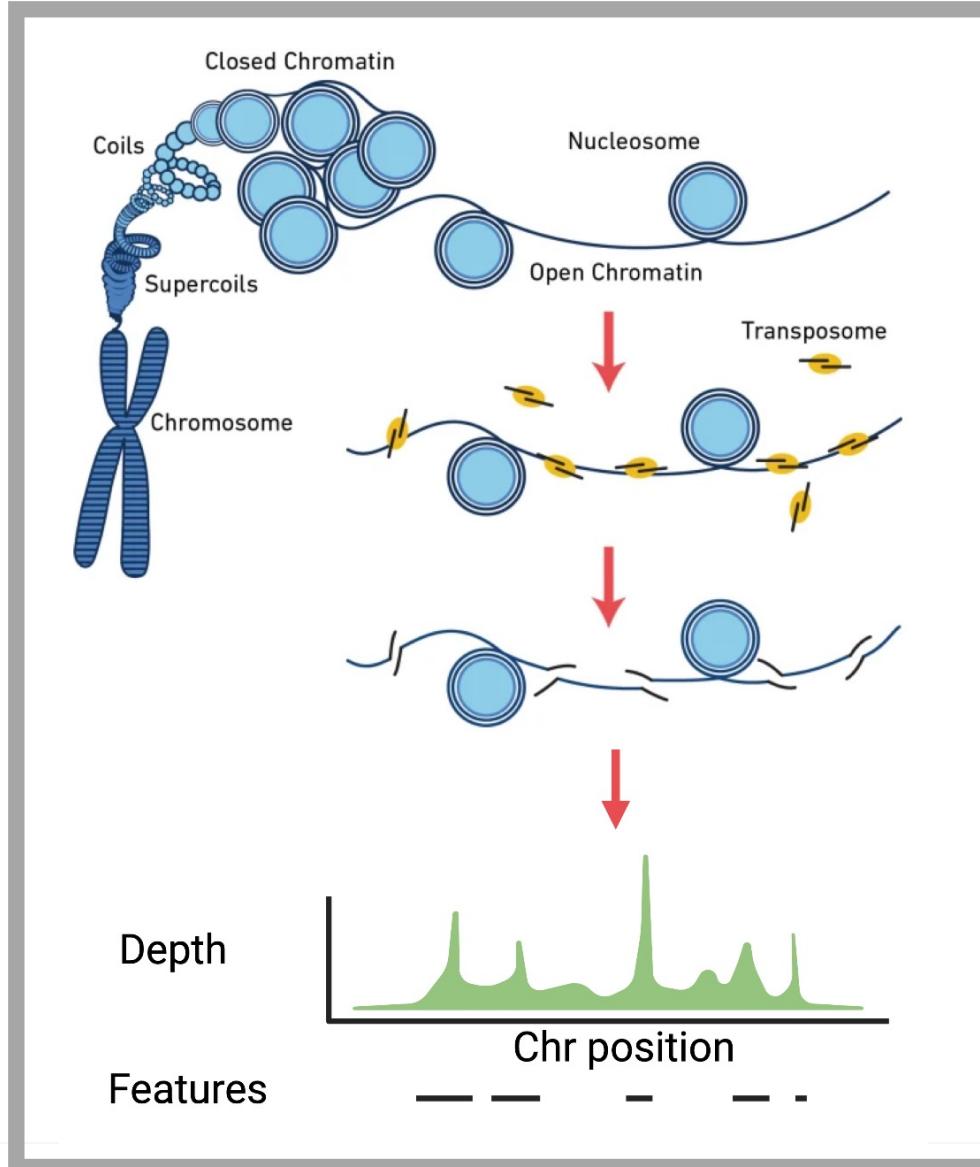


Why Single cell Multiome?

- Perform a **Peak-Gene Linkage analysis**:
 - Link gene expression to peaks (open chromatin regions) without inferring both methodology information
- Discover cells with **similar transcriptional profiles** but functionally **different chromatin landscapes**
- Discover new potential ***cis*-regulatory elements (CREs)** of target genes involved in cell differentiation, development and disease. (Need use chromatin state information)
- **Increase resolution** of complex and heterogeneous cell population
 - Improve the separation of cell subsets & identify rare population
 - Differentiate cell type depended on the cell state.



What scATAC-Seq is



Assay for Transposase Accessible Chromatin -

Allows the assessment of open and accessible chromatin region of the genome and assess the physical structure of the genome by identifying regions of open chromatin.

This technique uses a hyperactive transposase enzyme that cuts and inserts sequencing adapters into exposed DNA. The resulting sequencing library produces reads that are enriched in open chromatin regions.

What kind of information scATAC-seq can provide us?

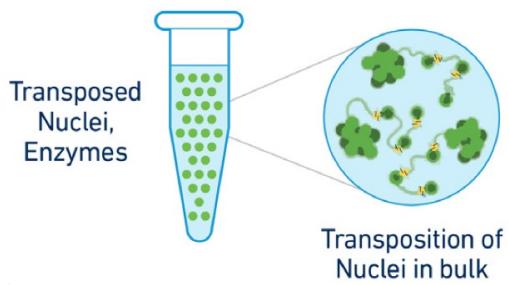
- Study the **single cell epigenetic states**:
 - Providing information about chromatin accessibility can **reveal areas of open chromatin that may correlate with gene expression (it should be validated by histone marks activity)**.
 - According to Signac: We can distinguish different cell subtype by identify cell-type-specific **TF enrichment**.
According to me: it is really difficult to differentiate cell subtypes.
- Increase our understanding of complex **regulatory networks**:
 - Where transcription factors binds and which genes TFs turn on or off depends on both the presence of a binding site in the DNA and whether that binding site is accessible. (Important to differentiate between active chromatin and open chromatin)
 - Reveal **regulatory patterns** that are complementary to gene expression
 - It enables the computational inference of **developmental**.
 - We can determine the temporal order of cells in a differentiation process.

What are the limitation of scATAC-seq?

- Only information of chromatin accessibility. Association with gene expression so difficult.
- Low cell subtype discriminatory power.
- More sensitive to number of principal component in lineal dimensionality reduction (LSI). (The first component correlate with sequencing depth).
- We need the histone marks bulk data to validate potential *cis*-Regulatory elements (CREs).

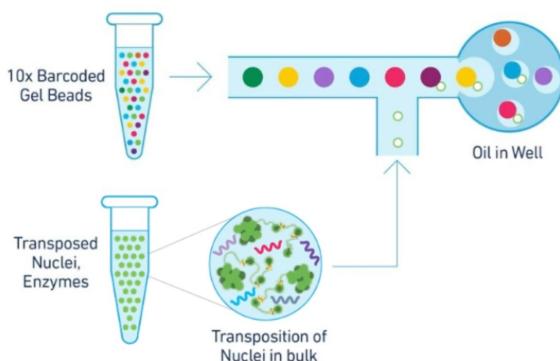
How single cell Multiome works

1. Bulk Nuclei



2. Chromium

(Reaction in GEM, Pre-amplification)



3. Library construction in parallel

ATAC library

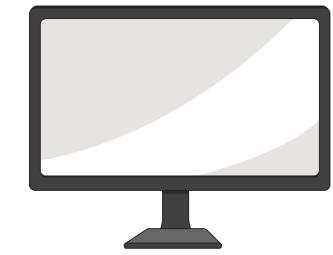


GEX library



4. Pre-processing data

(Sequencing, CellRanger)



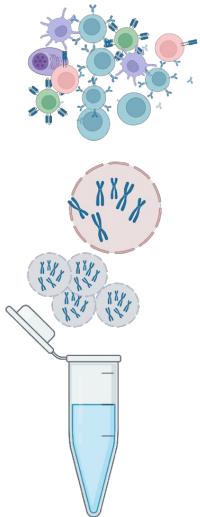
Multiome workflow

1. Bulk Nuclei

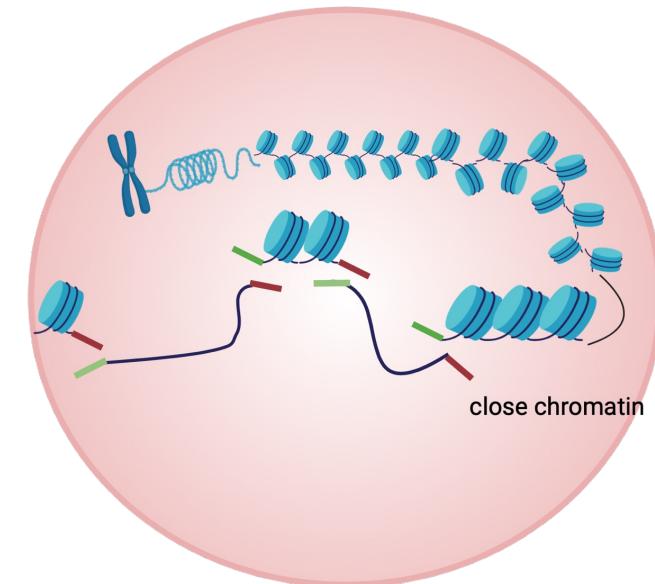
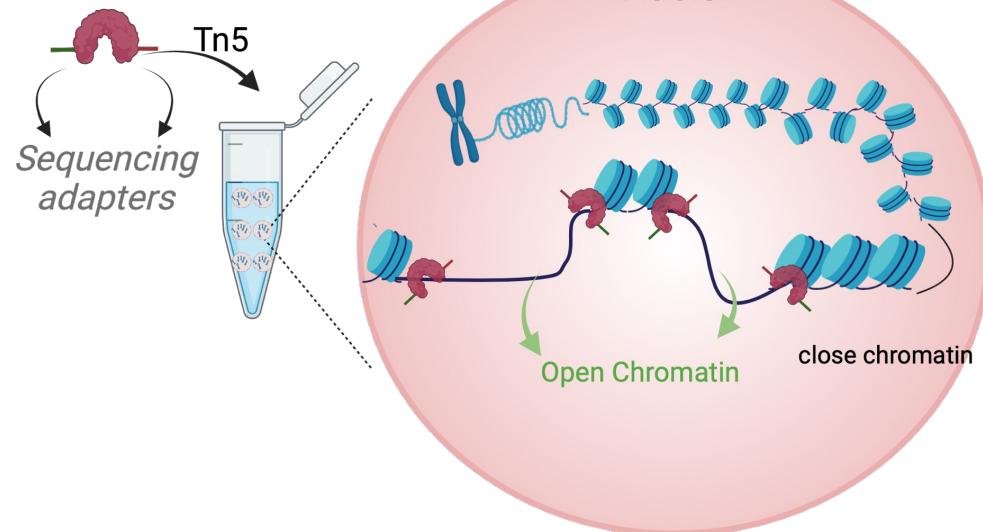


scATAC-seq

- 1
Assolation nuclei from 500 to 10,000 cells



- 2
Nuclei are incubated with Tn5 transposase.



Tn5 transposase targets chromatin open regions

Creating adapter tagged fragments

Multiome workflow

1. Nuclei

2. Chromium (Reaction in GEM, Pre-amplification)

3. Library construction in parallel

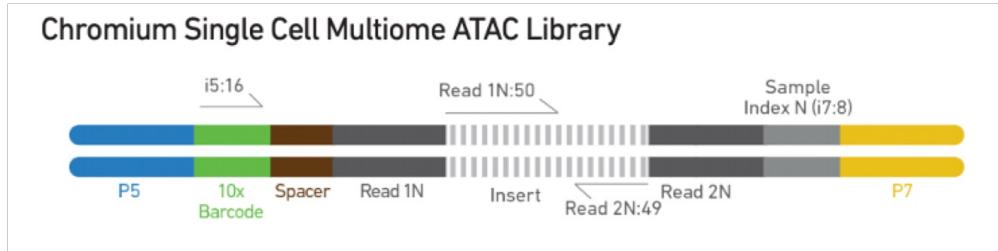
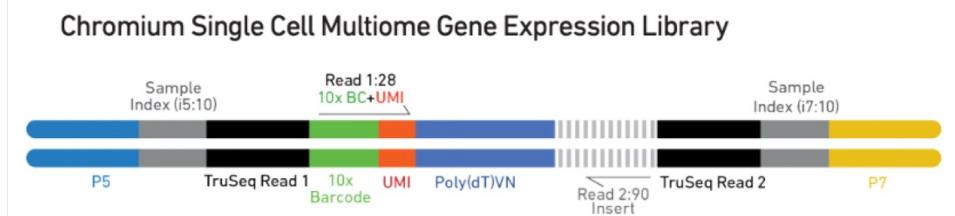
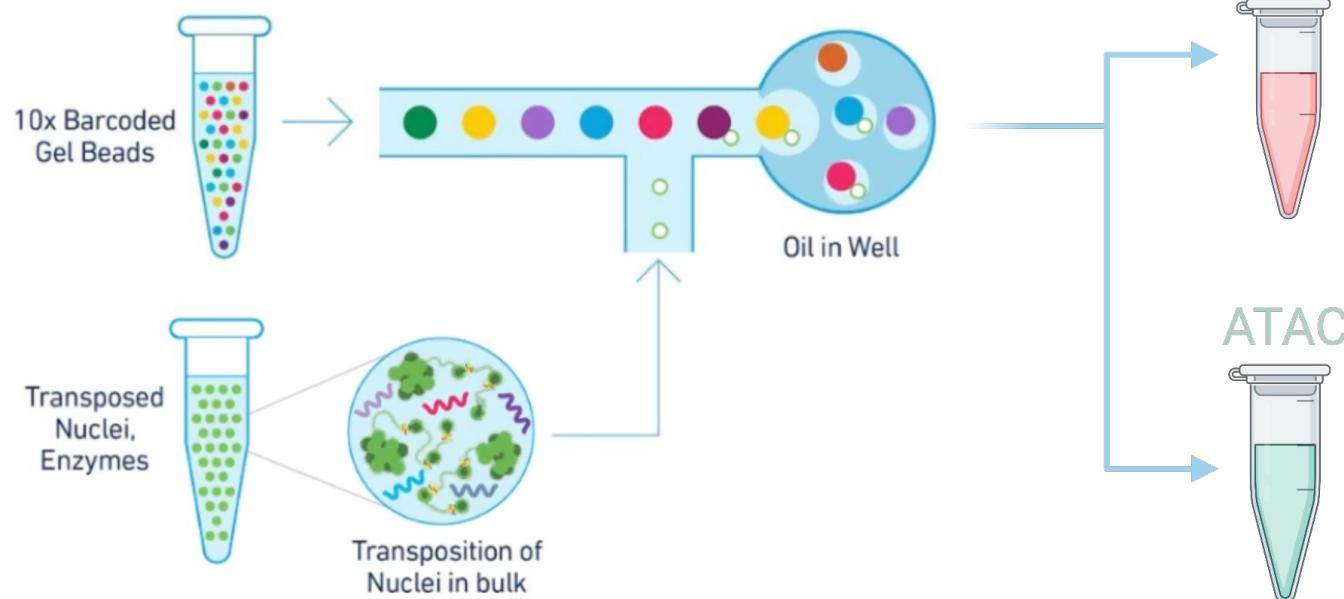
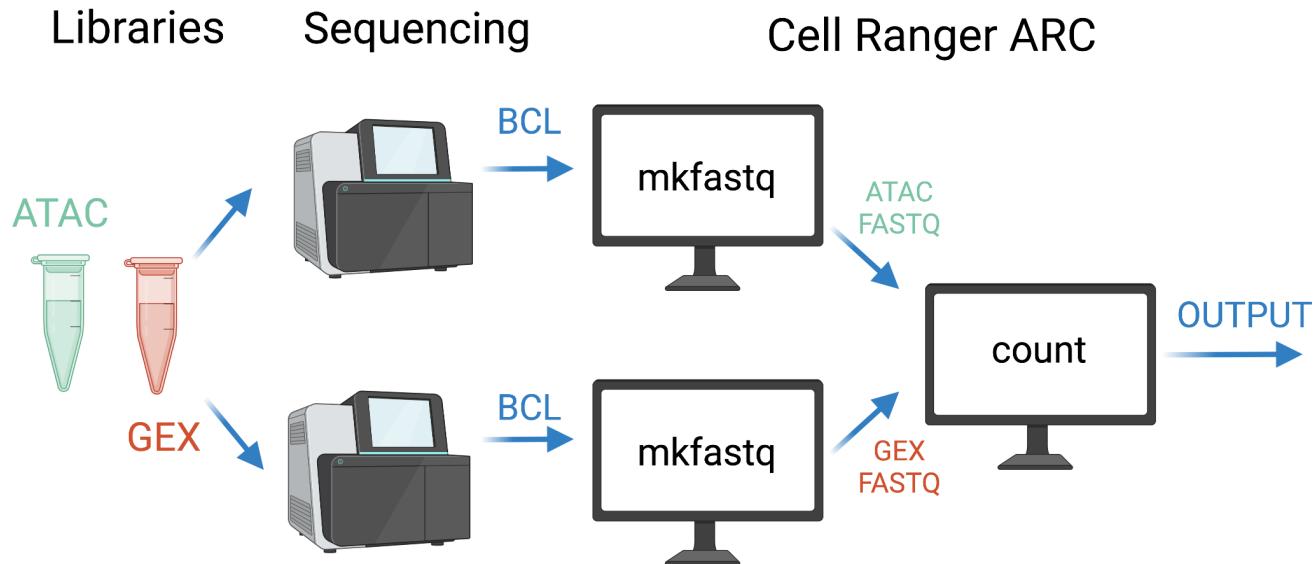


Image of 10x Genomics

- **Input:** Transposed nuclei
- **Output:**
 - One ATAC library and one Gene Expression library from each sample
 - Paired ATAC and Gene Expression data for each nucleus

Multiome workflow

4. Pre-processing



Pre-processed data

ATAC Per fragment information file:
atac_fragments.tsv.gv

Chrom	ChromStart	ChromEnd	barecode	readSupport
ch1	30015	31224	. ATGGCTAATCG	3
ch1	30126	31242	. ATGGCTCATGG	1
.

ATAC Per fragment information index:
atac_fragments.tsv.gv.tbi

Filtered feature barcode matrix HDF5:
filtered_feature_bc_matrix.h5

scRNA-seq

Gene count matrix

	Cell 1	Cell 1	...	Cell n
Gene 1	0	2	...	0
Gene 2	5	0		7
.
.
.
Gene m	0	20	...	0

scATAC-seq

Peak count matrix

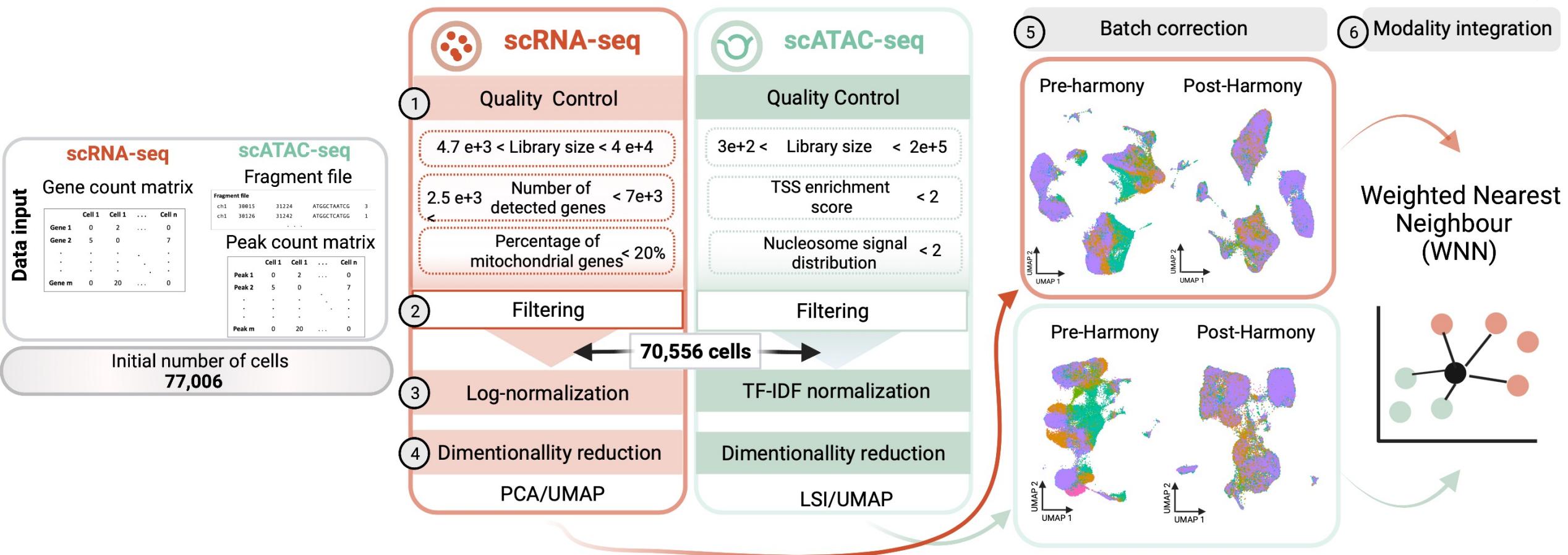
	Cell 1	Cell 1	...	Cell n
Peak 1	0	2	...	0
Peak 2	5	0		7
.
.
.
Peak m	0	20	...	0



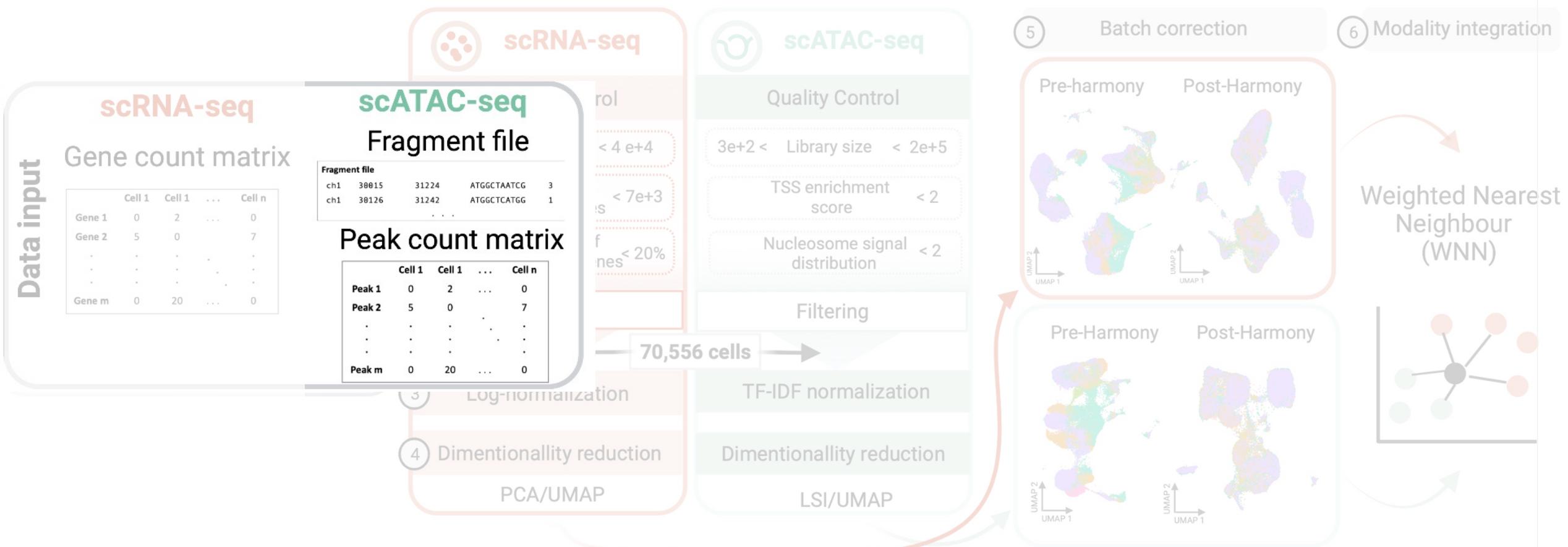
Remember: **tbi** fragment file has to be in the same directory that as you fragment file

Multiome analysis workflow

(Tools: Seurat, Signac, Harmony, Scrublet)



scATAC-seq analysis workflow



Data input

scATAC-seq

Fragment file

Chrom	ChromStart	ChromEnd	barcode	readSupport
ch1	30015	31224	.	ATGGCTAATCG
ch1	30126	31242	.	ATGGCTCATGG

BED-like format

Each line represents a unique
ATAC-seq fragment captured
by assay

Peak count matrix

	Cell 1	Cell 1	...	Cell n
Peak 1	0	2	...	0
Peak 2	5	0		7
.	.	.		.
.	.	.		.
.	.	.		.
Peak m	0	20	...	0

scATAC-seq workflow

Data input

scATAC-seq

Fragment file

Chrom	ChromStart	ChromEnd	barcode	readSupport
ch1	30015	31224	ATGGCTAATCG	3
ch1	30126	31242	ATGGCTCATGG	1
		

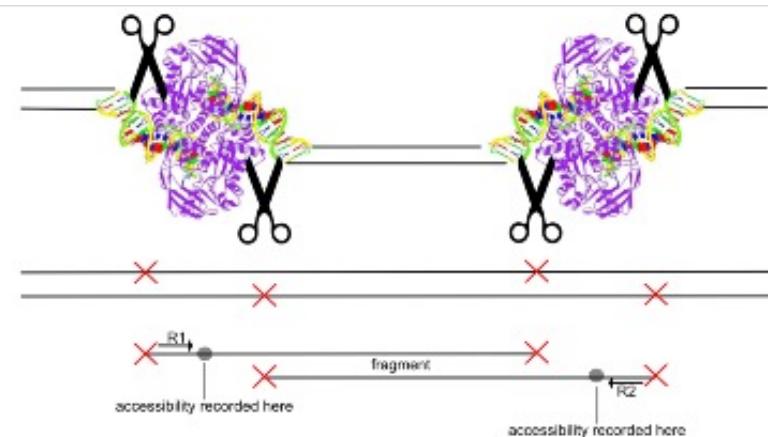
BED-like format

Peak count matrix

	Cell 1	Cell 1	...	Cell n
Peak 1	0	2	...	0
Peak 2	5	0		7
.	.	.		.
.	.	.		.
Peak m	0	20	...	0

Each line represents a unique **ATAC-seq fragment** captured by assay

Each fragment was created by 2 separated transposition events



Transposase image accessed from the Protein Data Bank, <https://rsb.org/structure/7M0H>

scATAC-seq workflow

Data input

scATAC-seq

Fragment file

Chrom	ChromStart	ChromEnd	barcode	readSupport
ch1	30015	31224	.	ATGGCTAATCG
ch1	30126	31242	.	ATGGCTCATGG

				3
				1

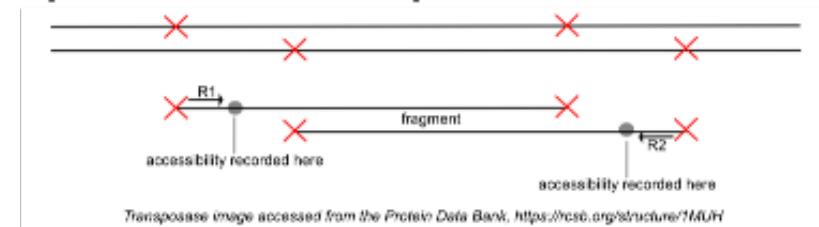
BED-like format

Peak count matrix

	Cell 1	Cell 1	...	Cell n
Peak 1	0	2	...	0
Peak 2	5	0		7
:	:	:		:
:	:	:		:
Peak m	0	20	...	0

Each line represents a unique **ATAC-seq fragment** captured by assay

Each fragment was created by 2 separated transposition events



Duplicate reads are collapsed into a single fragment record (readSupport)

scATAC-seq workflow

Data input

scATAC-seq

Fragment file

Chrom	ChromStart	ChromEnd	barcode	readSupport
ch1	30015	31224	.	ATGGCTAATCG
ch1	30126	31242	.	ATGGGCTCATGG
	...			

Peak count matrix

	Cell 1	Cell 1	...	Cell n
Peak 1	0	2	...	0
Peak 2	5	0		7
.
.
.
Peak m	0	20	...	0

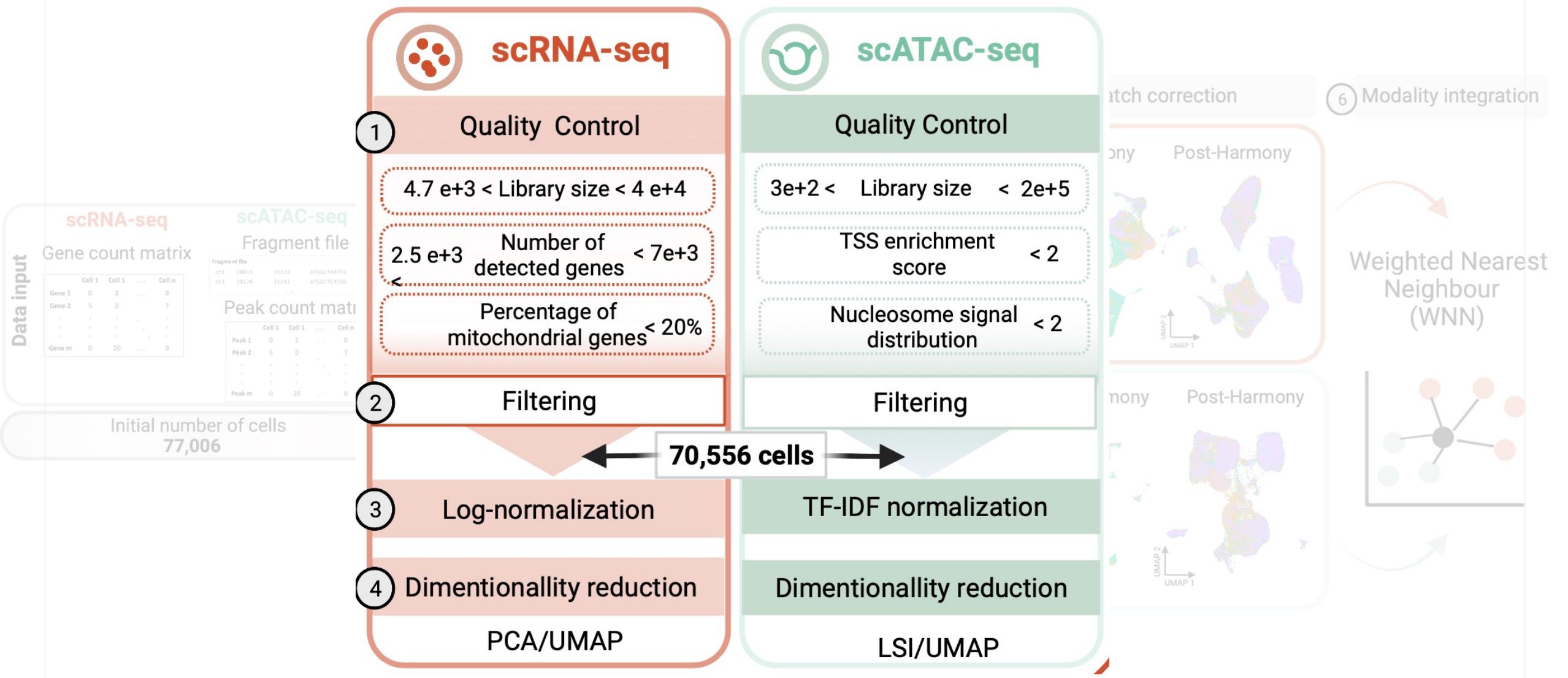
Each row represents a open chromatin region of the genome (a peak).

Each value represents the number of Tn5 integration sites for each single barcode (i.e. a cell) that map within each peak.

Multiome analysis workflow

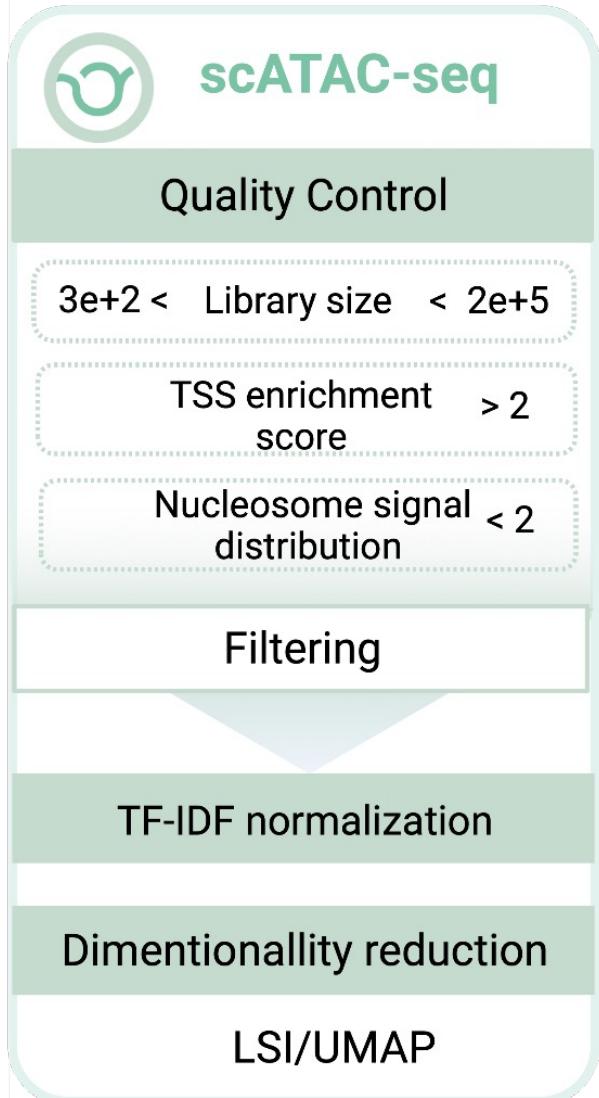
Quality Control

(Tools: Seurat, Signac,
Harmony, Scrublet)



scATAC-seq Workflow

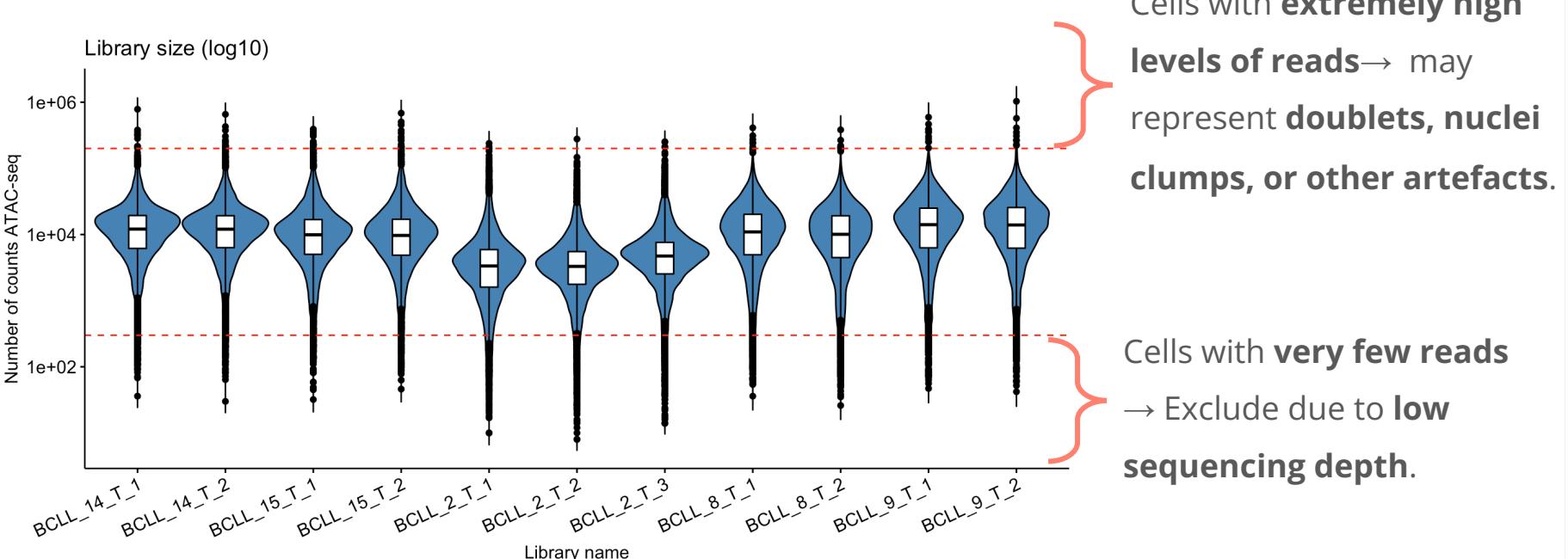
Quality control metrics



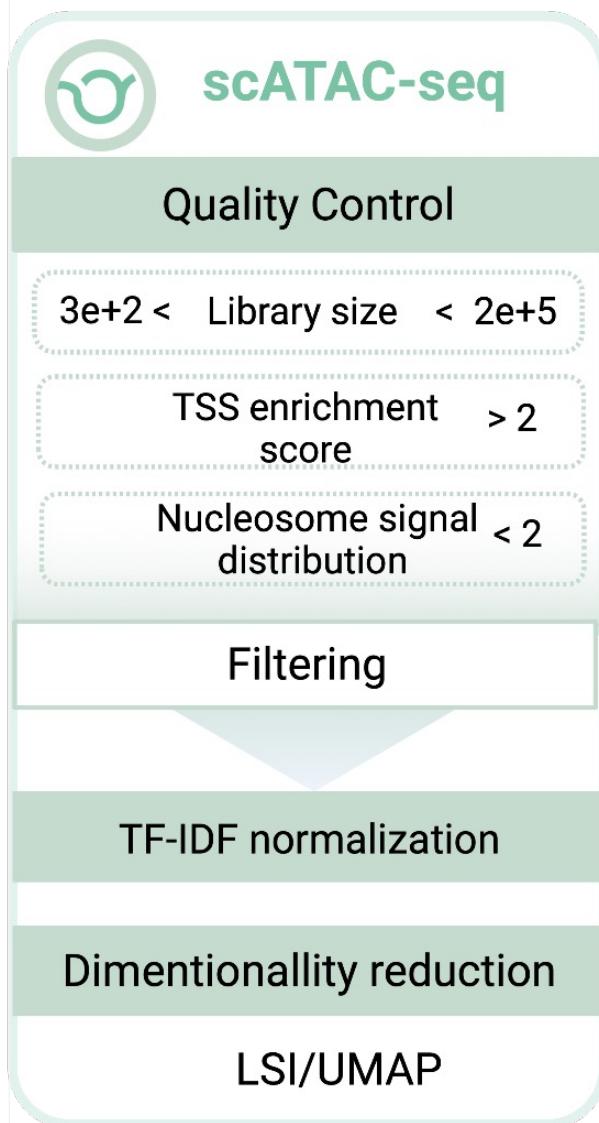
The expected range of values for these parameters will vary depending on your biological system, cell viability, and other factors.

Library sizes = sequencing depth / complexity

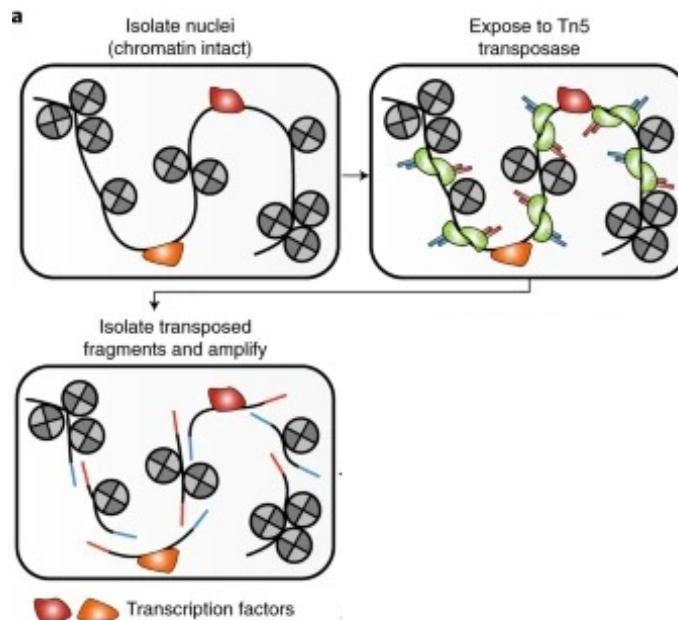
Total number of fragments in peaks:



Quality control metrics



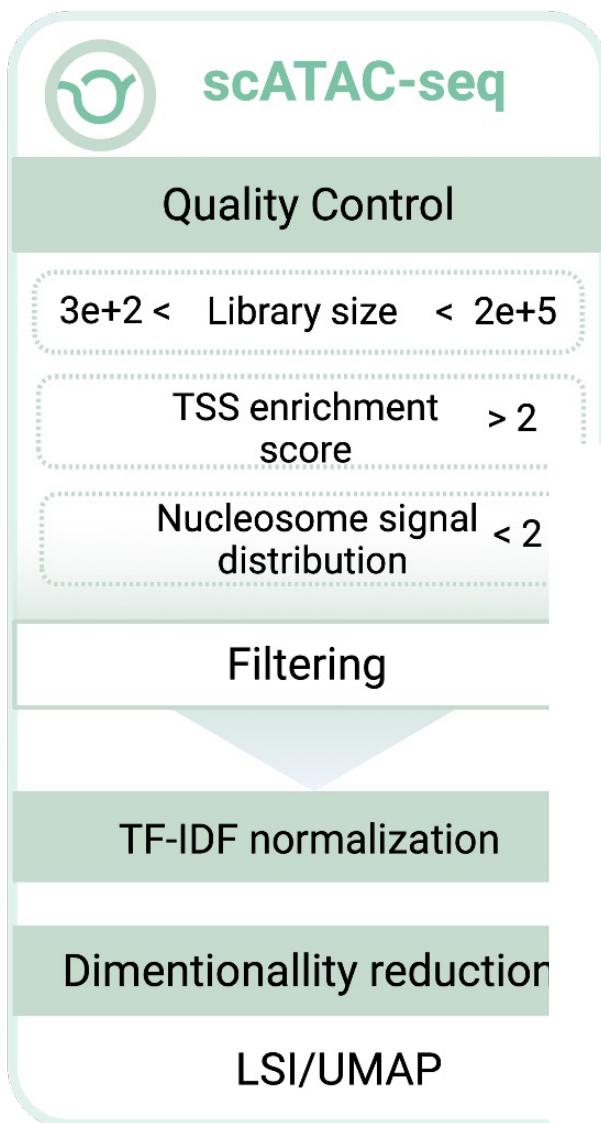
Nucleosome signal distribution



- It quantifies the approximate ratio of Di-nucleosome (DI) and mono-nucleosome (MONO) to nucleosome-free region (NFR) fragments.
- It is used to evaluate the quality of transposase reaction: We expect to **find half or more of the fragments within the NFR** to confirm that our data is high quality and the transposase worked properly.

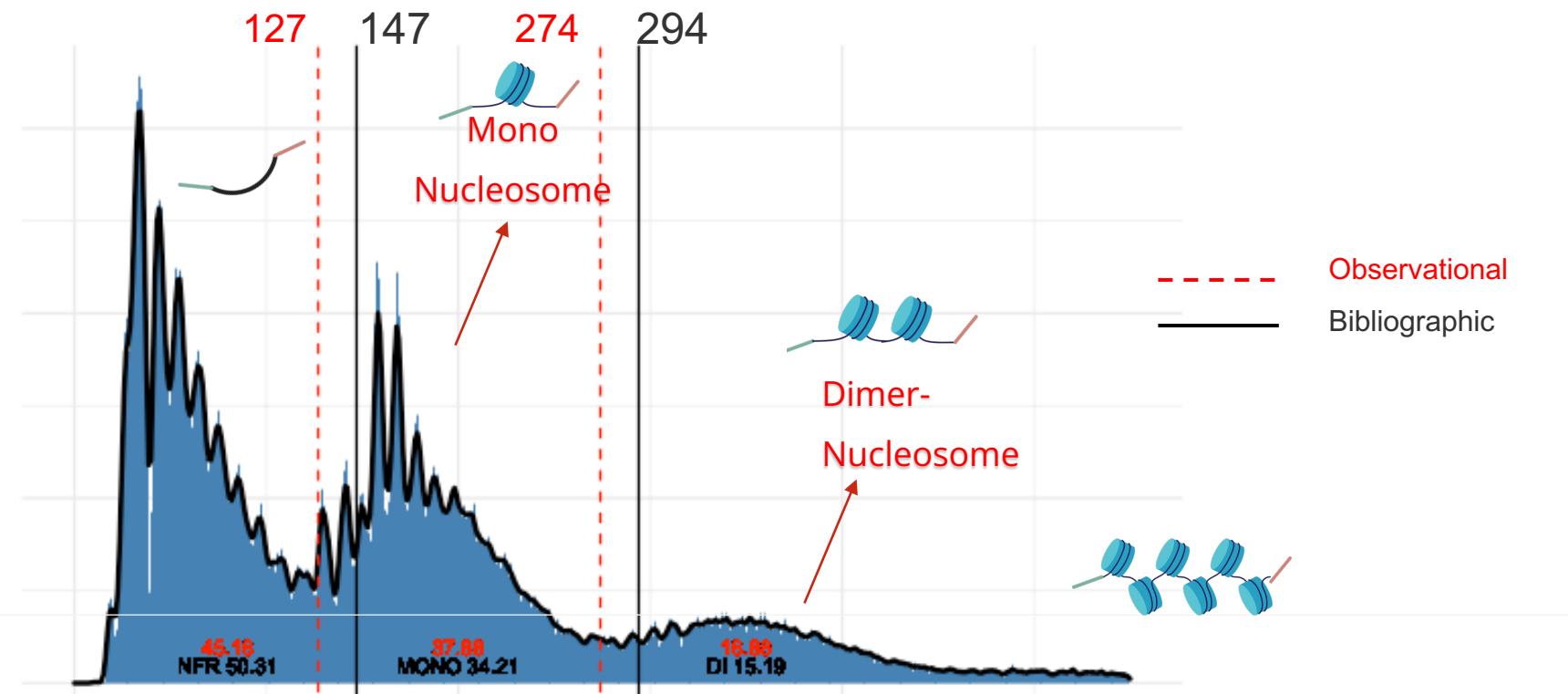
scATAC-seq Workflow

Quality control metrics



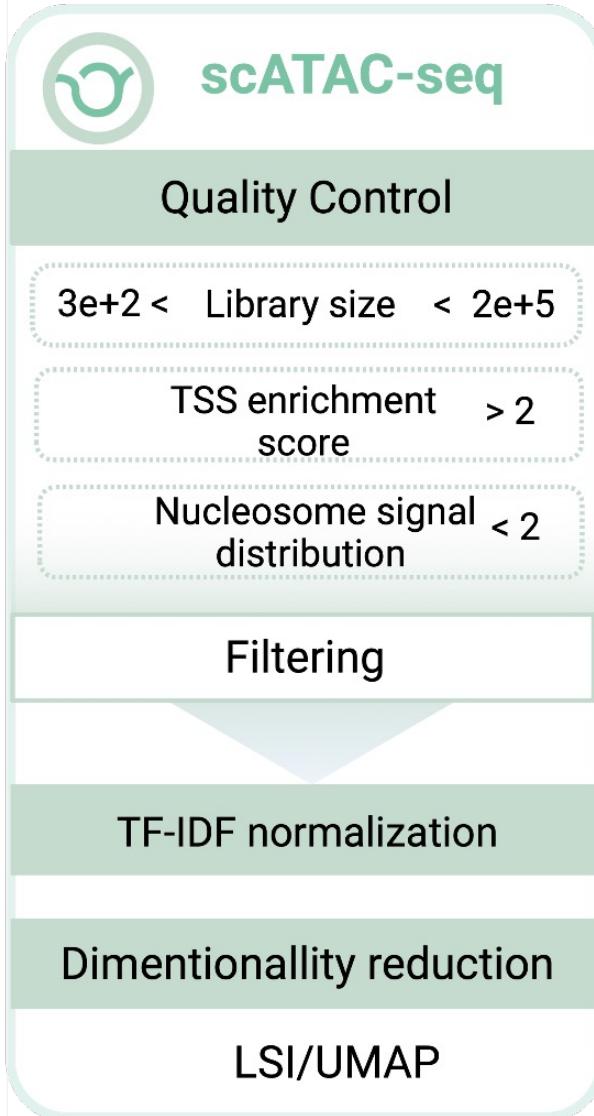
Nucleosome signal distribution

The histogram of DNA fragment sizes should exhibit a strong nucleosome banding pattern corresponding to the length of DNA wrapped around a single nucleosome.



scATAC-seq Workflow

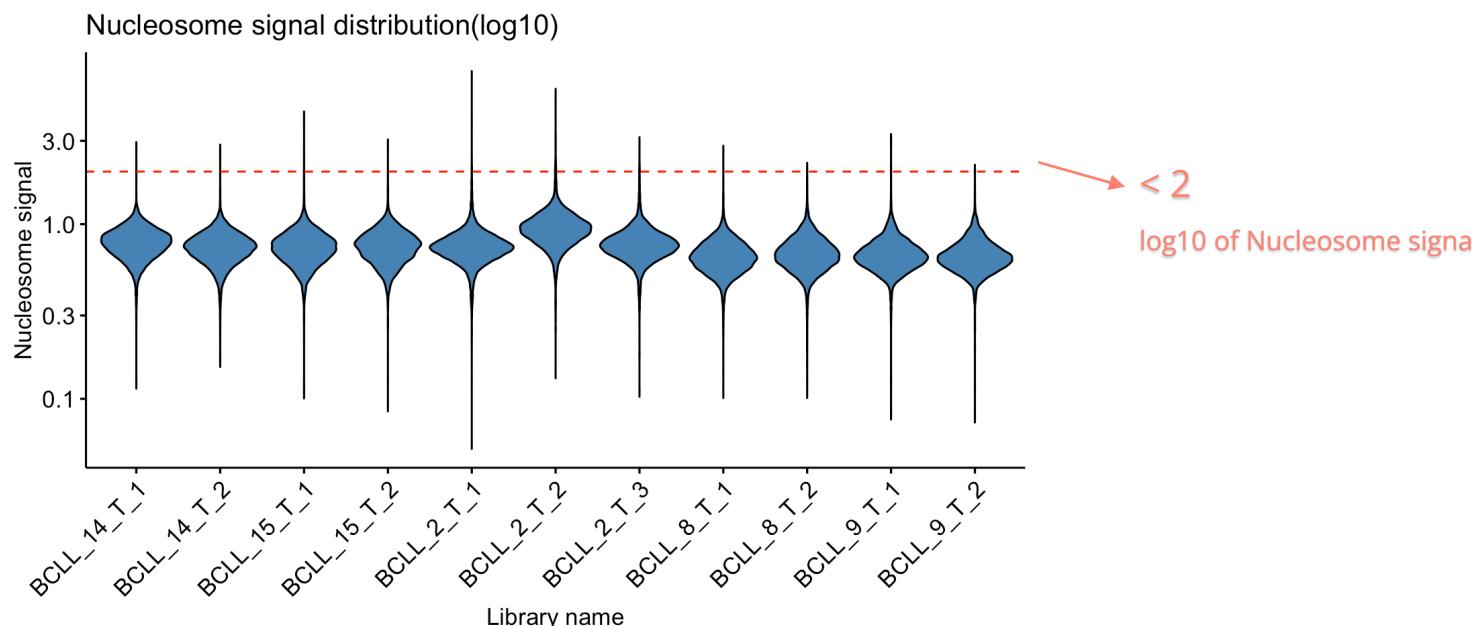
Quality control metrics



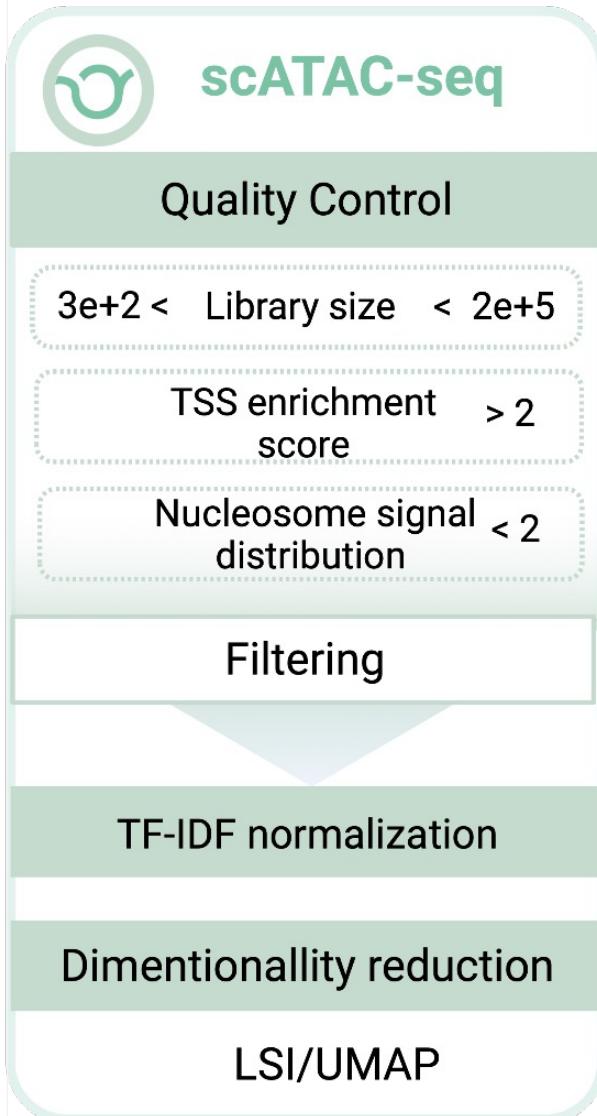
The expected range of values for these parameters will vary depending on your biological system, cell viability, and other factors.

Nucleosome signal distribution

The ratio of mononucleosomal (147–294 bp) to nucleosome-free (<147 bp) fragments sequenced for the cell is a way of quantifying the expected depletion of nucleosome-length DNA fragments.



Quality control metrics

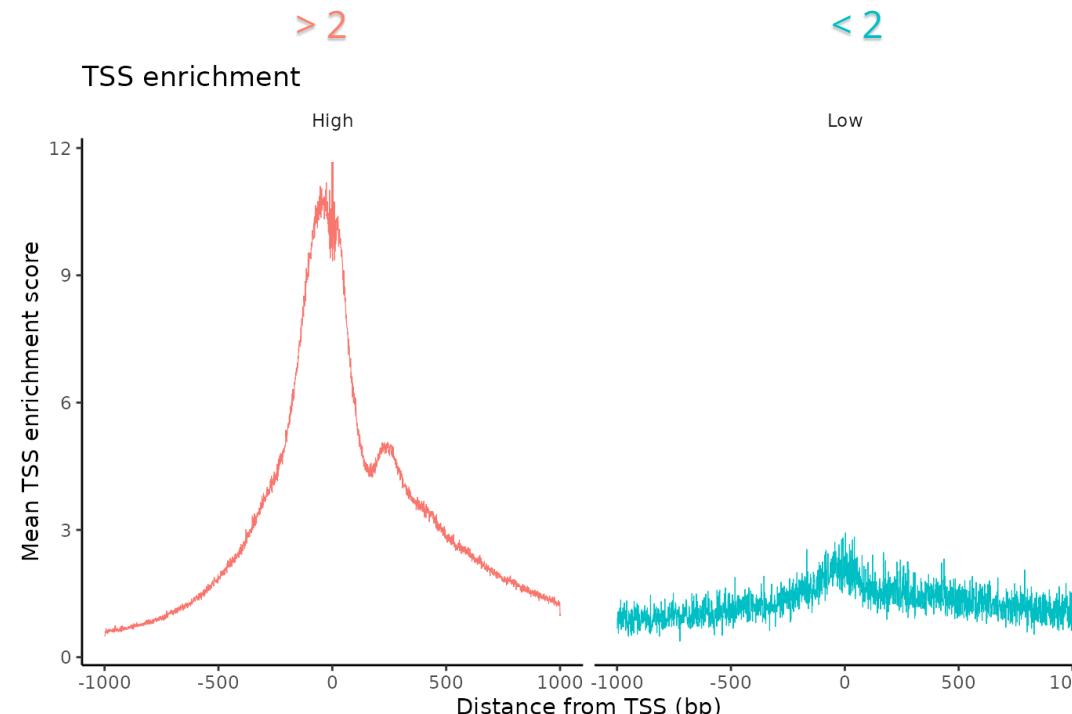


Transcriptional start site (TSS) enrichment score:

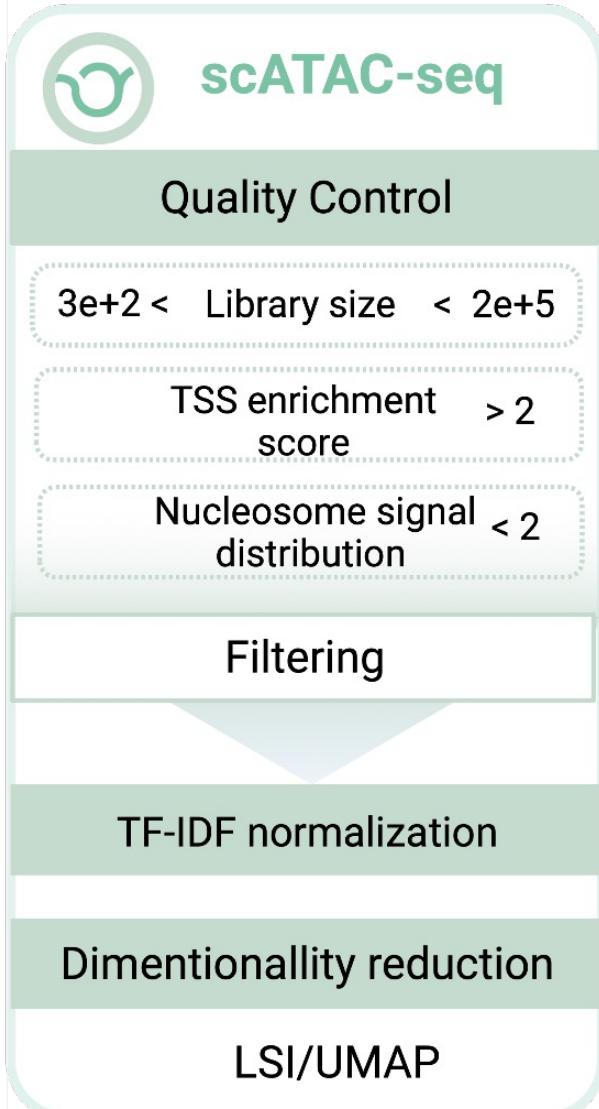
Score based on the ratio of fragments centered at the TSS to fragments in TSS-flanking regions.

- **Low TSS enrichment score** → Poor ATAC-seq experiments.

`TSSEnrichment()` function: computes this metric → stored in metadata under the column = `TSS.enrichment`.



Quality control metrics



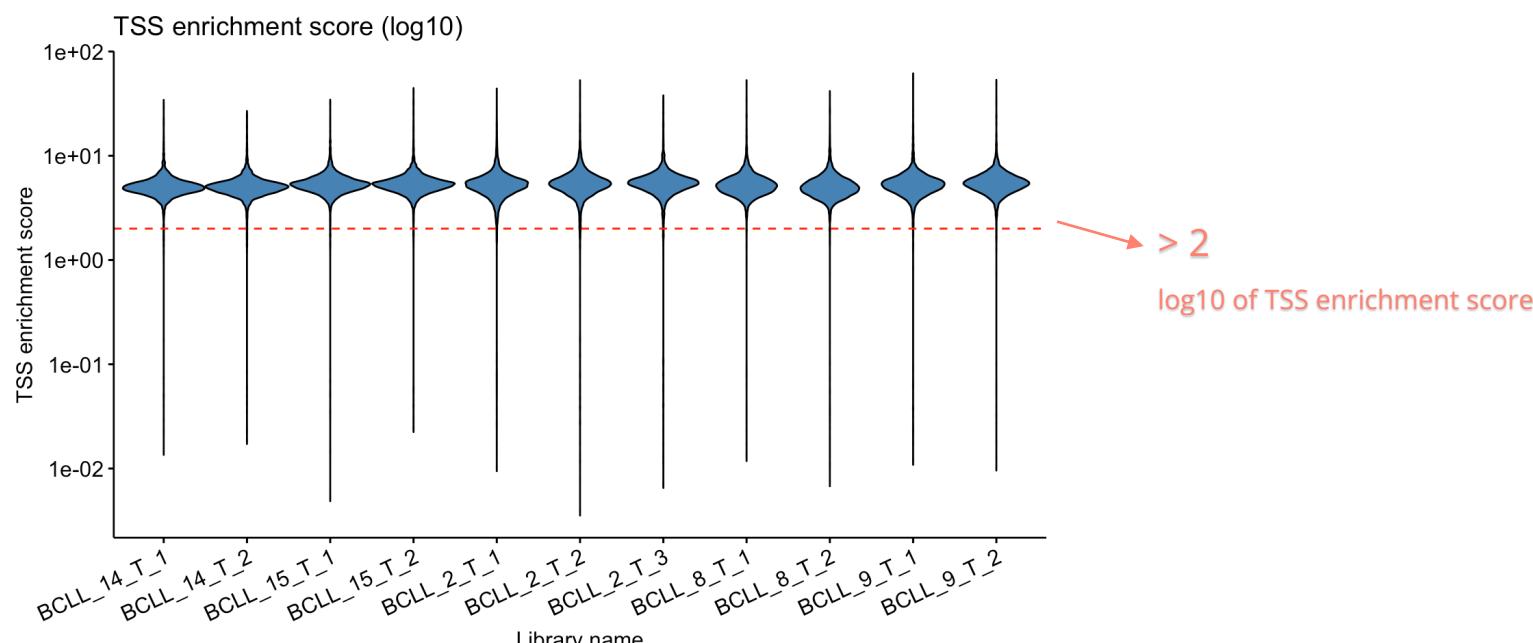
The expected range of values for these parameters will vary depending on your biological system, cell viability, and other factors.

Transcriptional start site (TSS) enrichment score:

Score based on the ratio of fragments centered at the TSS to fragments in TSS-flanking regions.

- **Low TSS enrichment score** → Poor ATAC-seq experiments.

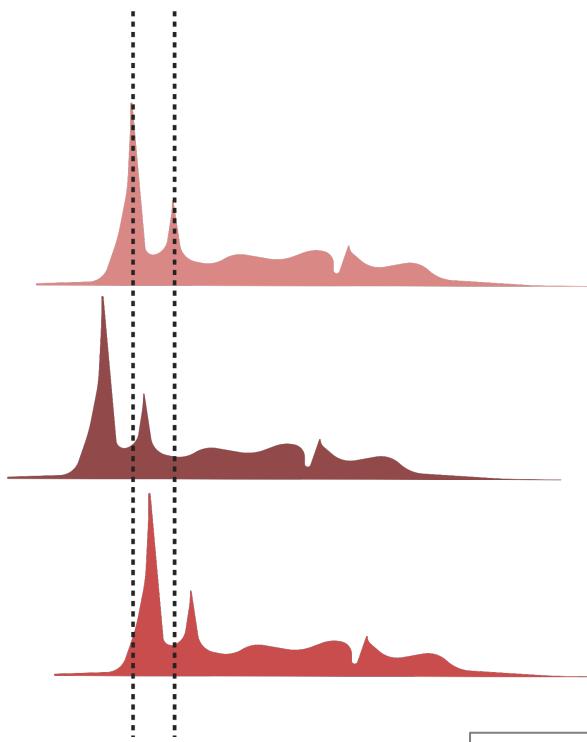
`TSSEnrichment()` function: computes this metric → stored in metadata under the column = `TSS.enrichment`.



scATAC-seq Workflow

Merging Seurat Objects

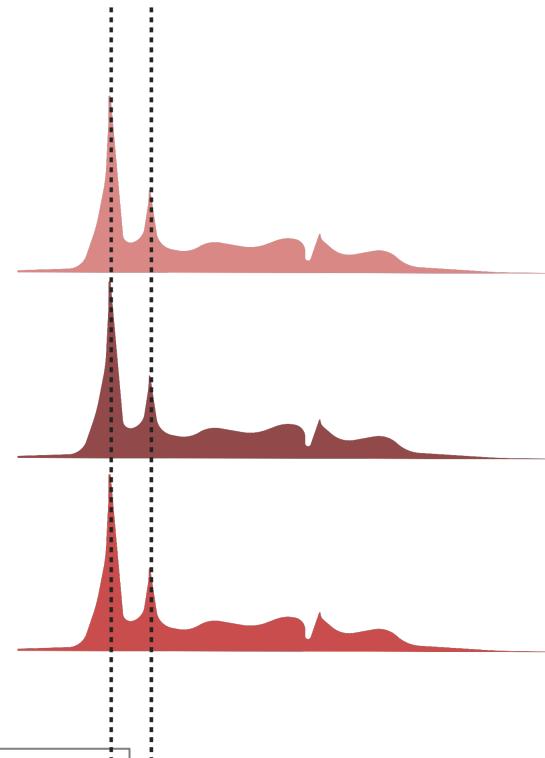
When merging multiple single-cell chromatin datasets, we must be aware **that each dataset is peak-called independently.** So, their peaks may not overlap perfectly → **create a common set of peaks before merging multiple datasets** using:



1

***Unifypeaks* function:**

- merge all the intersecting peaks
- create different sets of non-overlapping peaks.



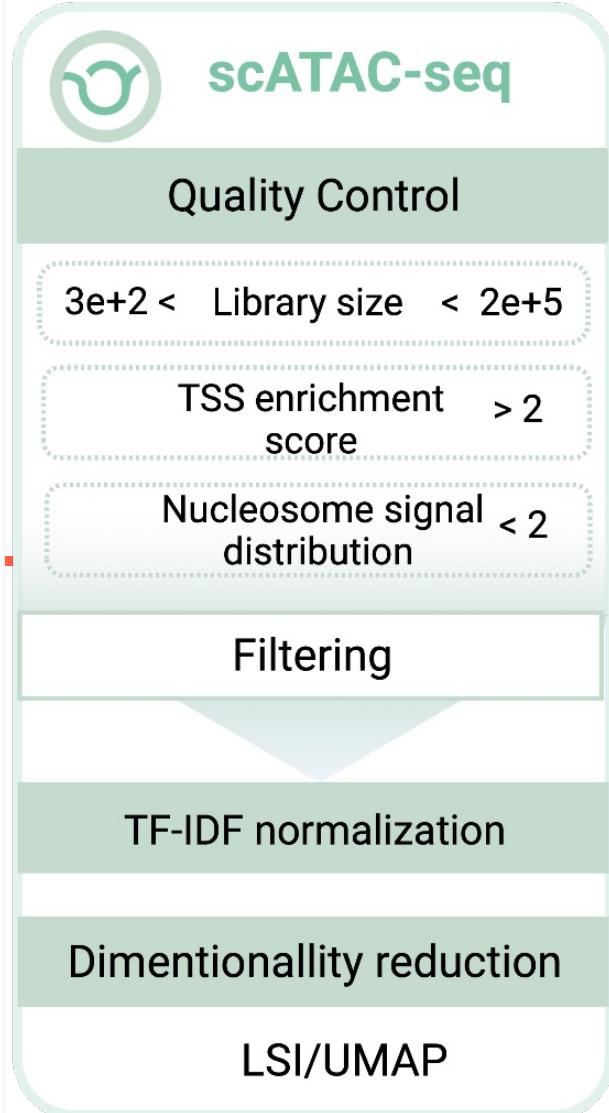
2

***FeatureMatrix* function :**

re-quantify the counts for each data set and it will create a new assay storing the Fragment object for each dataset.

scATAC-seq Workflow

Normalization



Signac performs a two step normalization technique:

TF-IDF: Term Frequency - Inverse Document Frequency (Signac modification where zero counts remain as zero after applying TF-IDF).

1. Normalize the cells: correcting possible differences in sequencing depth between cells
 2. Normalizes the peaks: increase the signal of unusual peaks

$$TF = \frac{C_{i,j}}{F_j} \quad \begin{aligned} &= \text{total number of counts for } \textbf{\textit{peak}} \ i \text{ in } \textbf{\textit{cell}} \ j. \\ &= \text{total number of counts for } \textbf{\textit{cell}} \ j. \end{aligned}$$

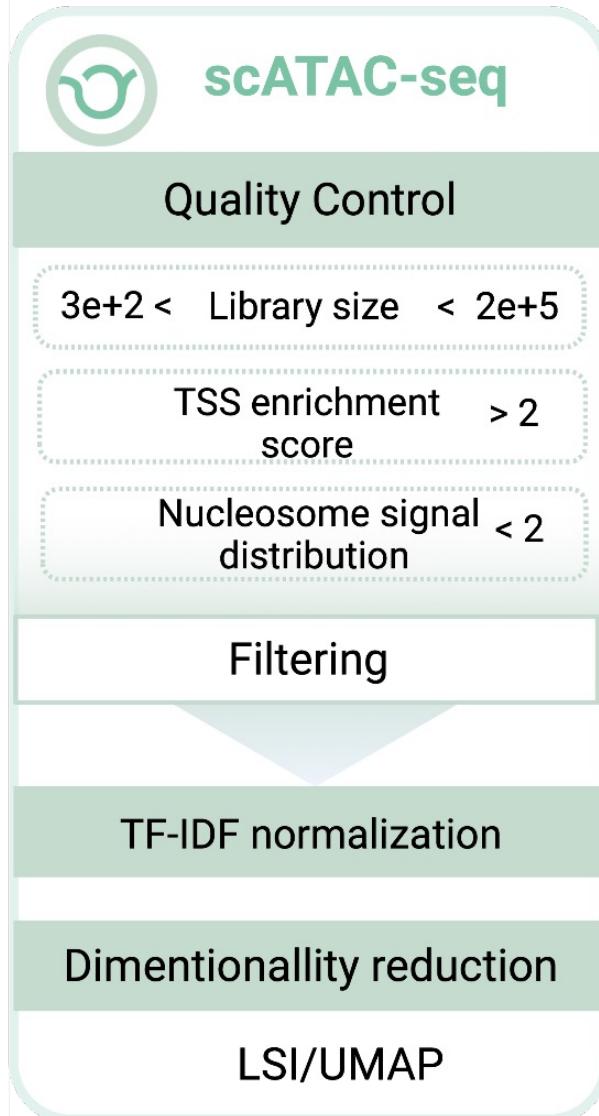
$$IDF = \frac{N}{n_i}$$

= total number of cells in the dataset
= total number counts for ***peak i*** across all cells

$$TF-IDF = \log(1 + (TF \times IDF) \times 10^4)$$

scATAC-seq Workflow

Lineal and non-lineal dimensionality reduction (UMAP)



Lineal dimensionality reduction

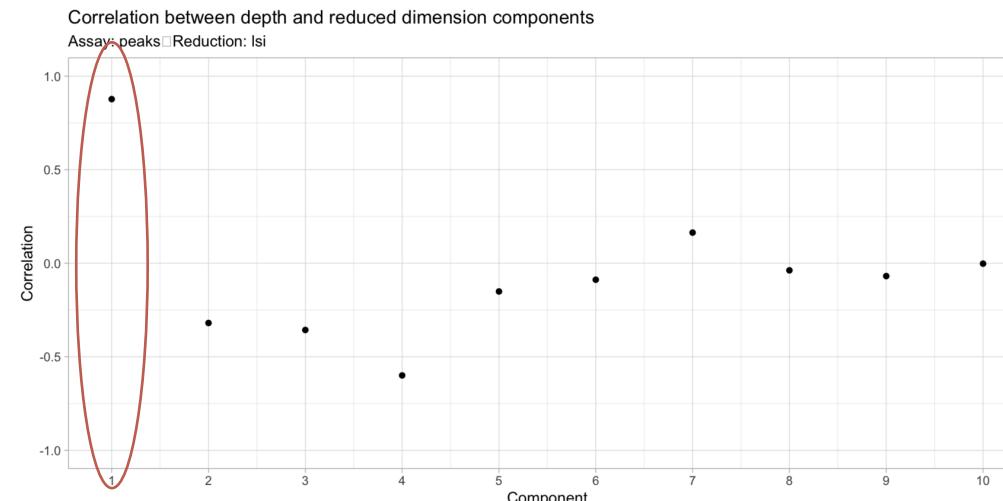
LSI: Latent Semantic Technique

It is scalable to large numbers of cells as it retains the data sparsity.

It performs the **SVD (Singular Value Decomposition)** on the **TF-IDF transformation matrix**.

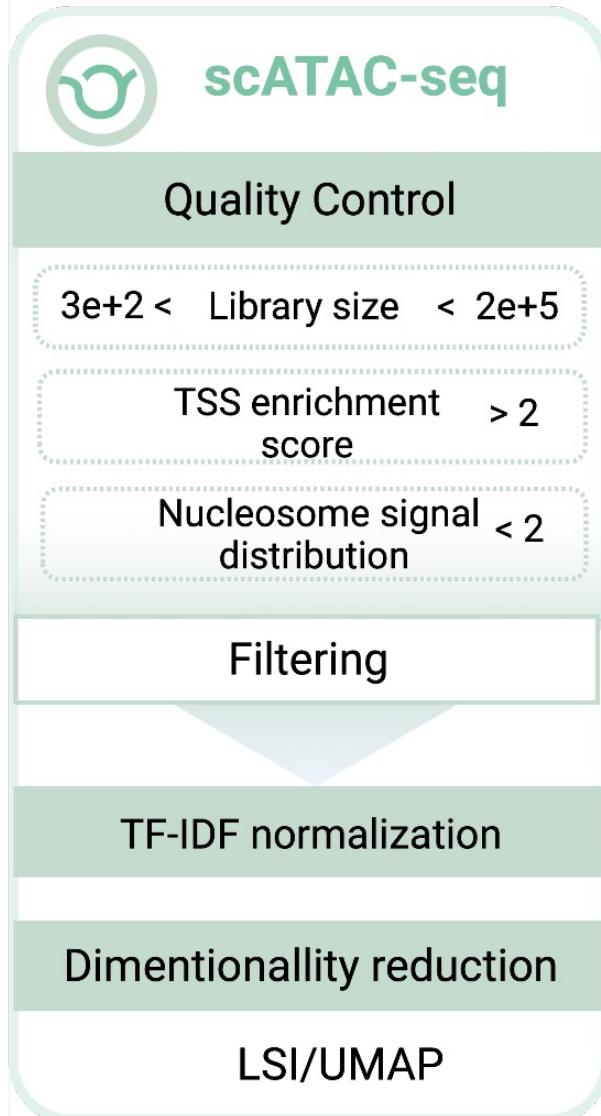
LSI technique that combine= TF-IDF step followed by an SVD step.

LSI typically captures sequencing depth (technical variance) and not biological variance. We can see the fist component is correlated with the sequence depth ($R \approx 1$)



scATAC-seq Workflow

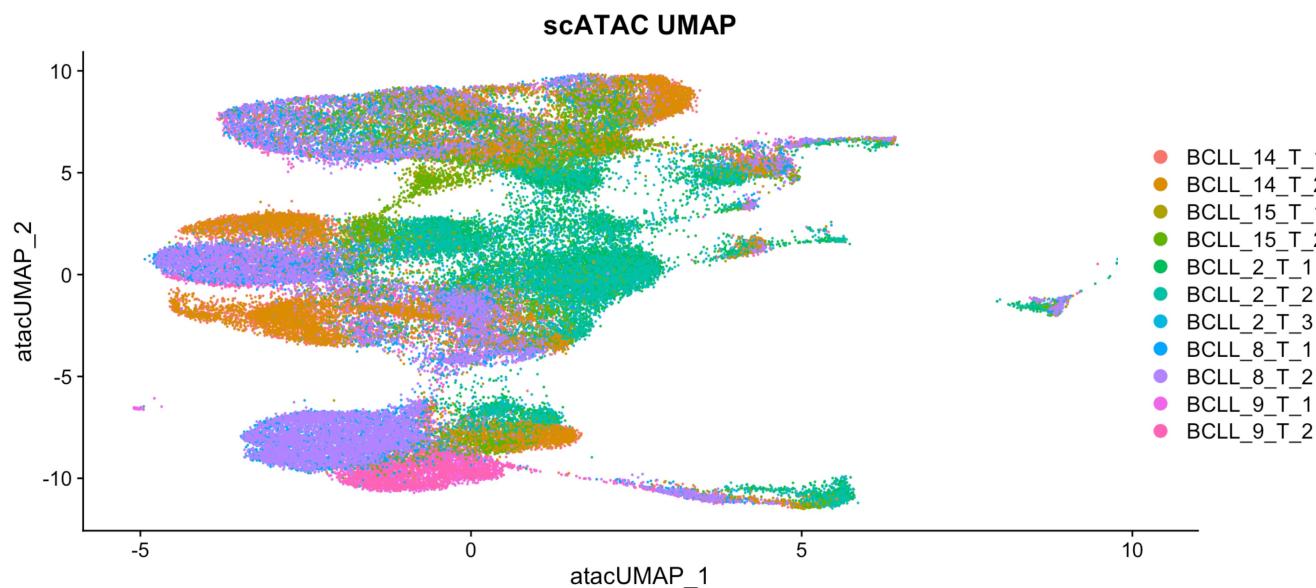
Lineal and non-lineal dimensionality reduction (UMAP)



Non-lineal dimensionality reduction

UMAP: Uniform Manifold Approximation and Projection

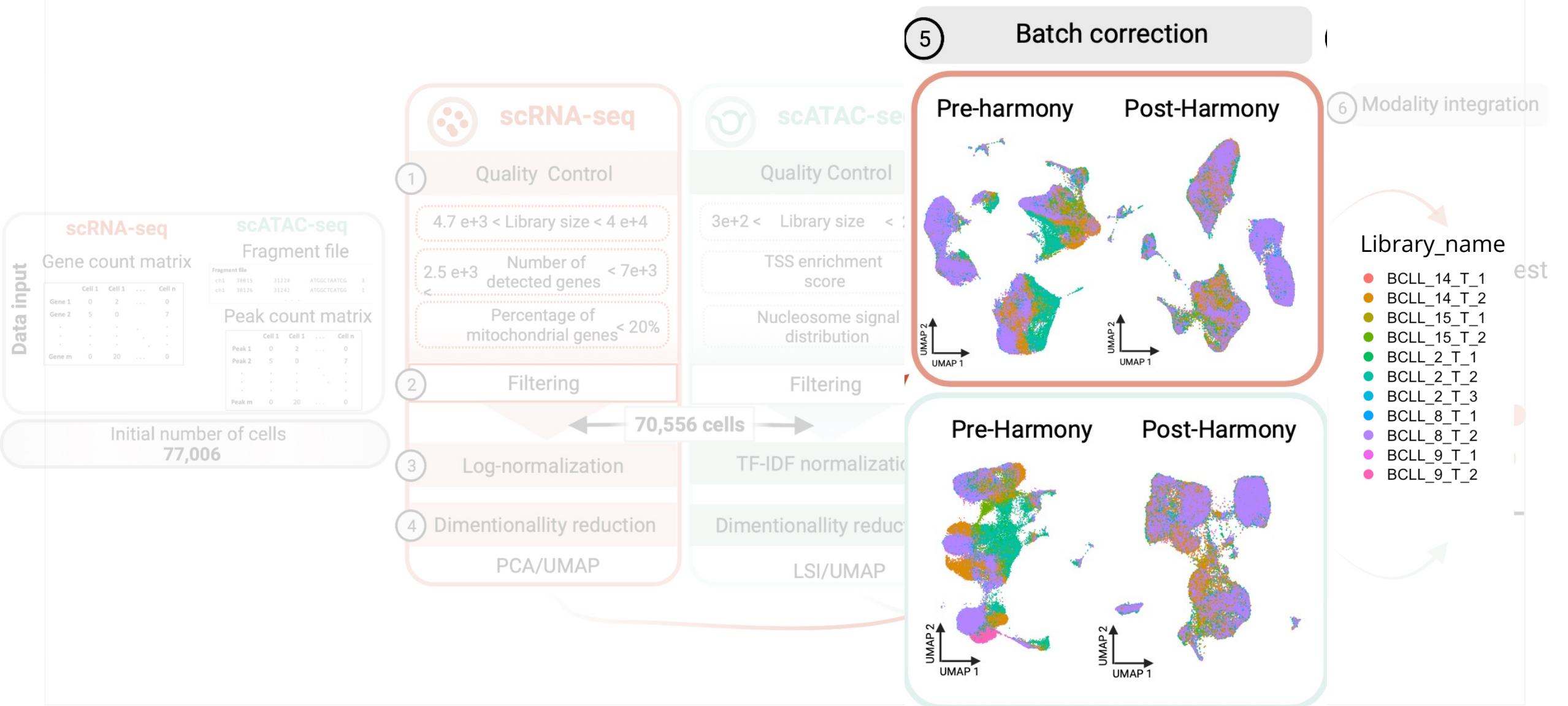
This step is like the scRNA-seq but in this case we will omit the first component of the LSI (dims = 2:40).



Multiome analysis workflow

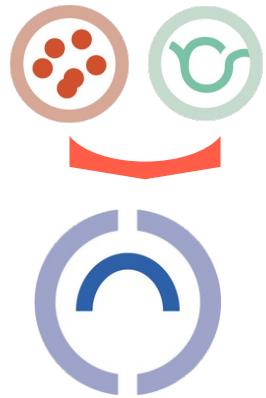
Batch Correction (Harmony)

(Tools: Harmony)

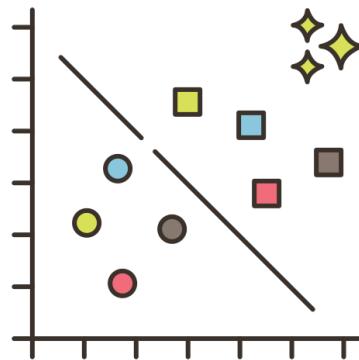


In the following session

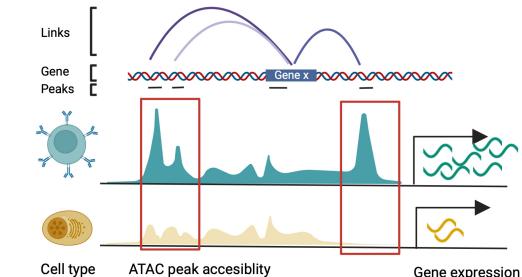
Next steps



Integration of
scATAC-seq + scRNA-seq by
Weighted Nearest Neighbour (WNN)



**Clustering &
Peak calling**

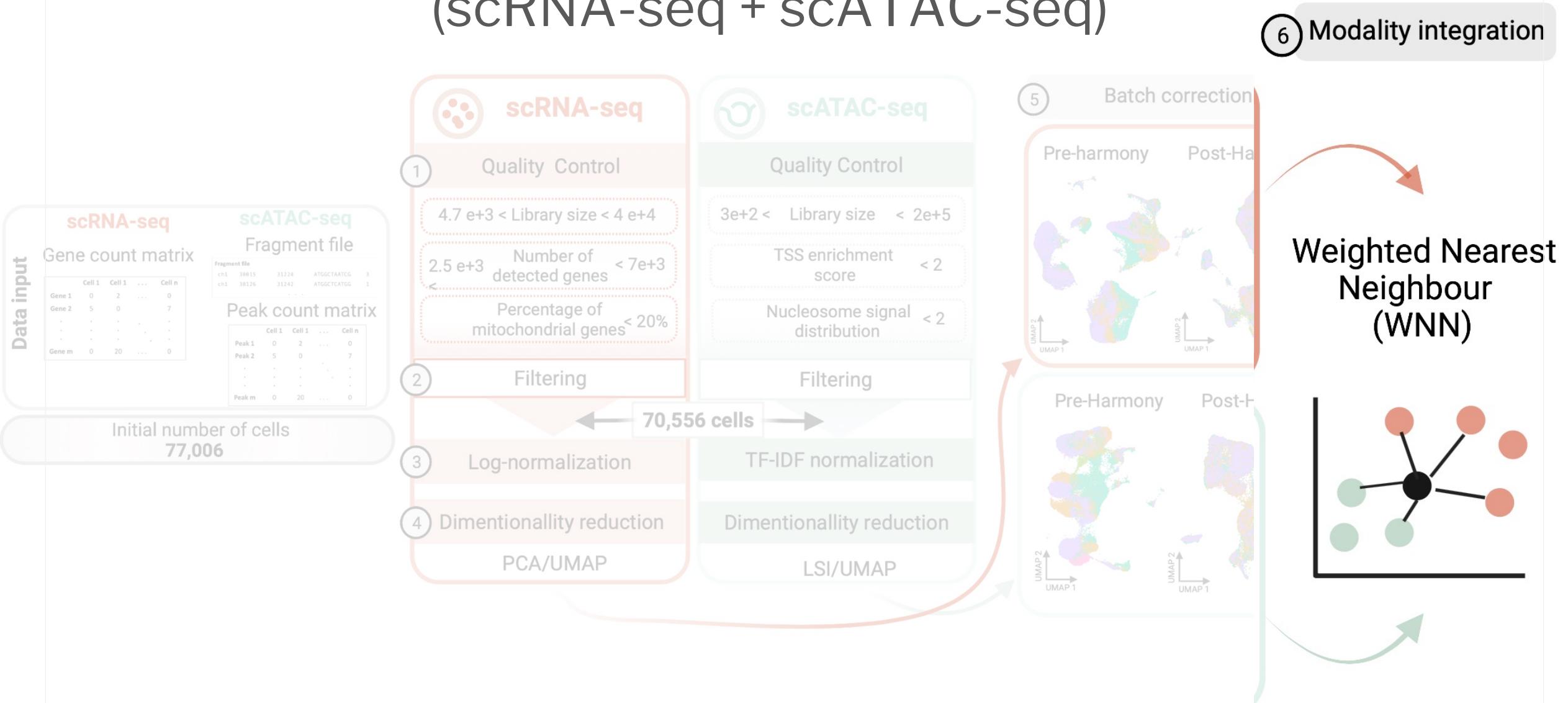


**Peak-genes
Linkages analysis**

Multiome analysis workflow

Modality integration

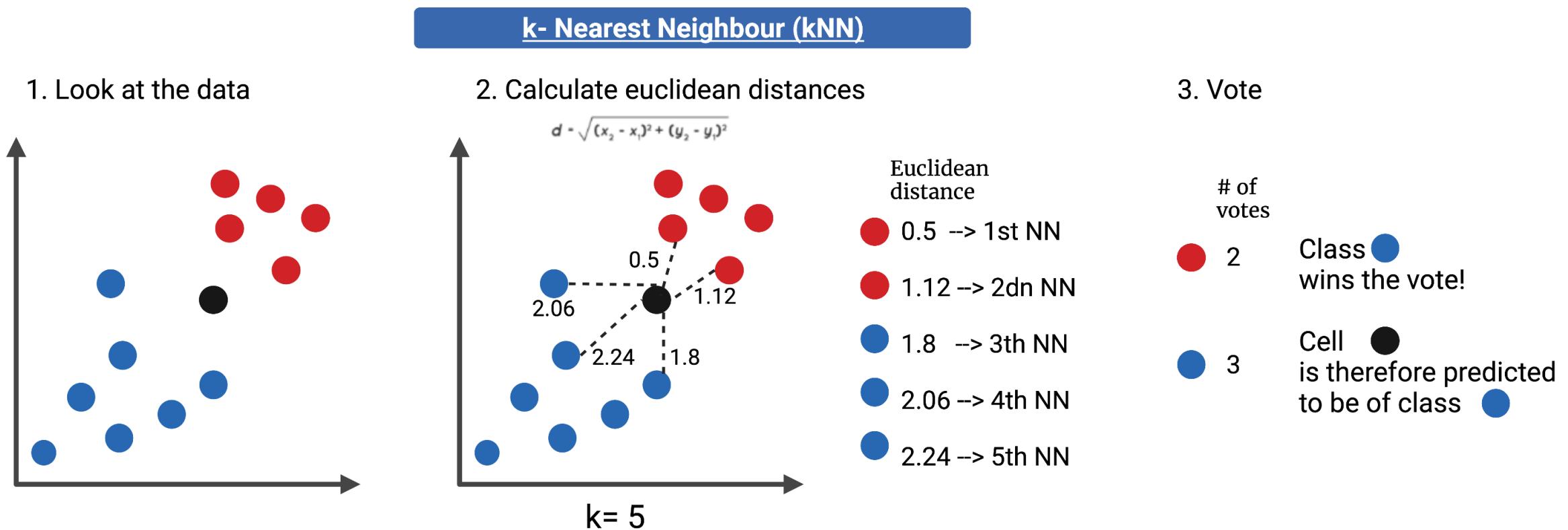
(scRNA-seq + scATAC-seq)



Weighted Nearest Neighbour

It is an unsupervised framework that learns the relative utility of each data type in each cell and allows integrative analysis of multiple modalities.

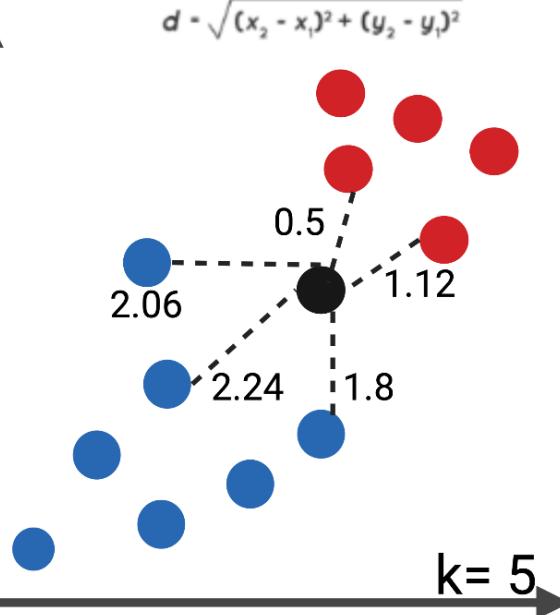
Weighted Nearest Neighbour is a modified version of k-Nearest Neighbors (kNN) where we take an assumption that the impact of a nearer neighbour on the query point is more than the farther away point



Weighted Nearest Neighbour

2. Calculate euclidean distances

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



3. Weight of each class

Euclidean distance (d)	Weight $1/d$
0.5	2.0
1.12	0.89
1.8	0.56
2.06	0.48
2.24	0.45

Diagram illustrating the calculation of weights:

- Red bracket: $2.0 + 0.89 = 2.89$
- Blue bracket: $0.56 + 0.48 + 0.45 = 1.49$

$= 2.89$

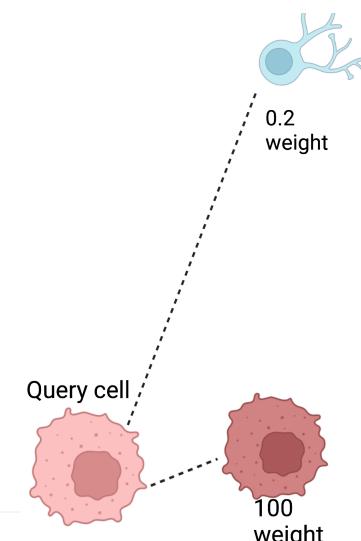
Class has more weight.

Cell is therefore predicted to be of class

In that way we are **giving more weight to smaller distances**.

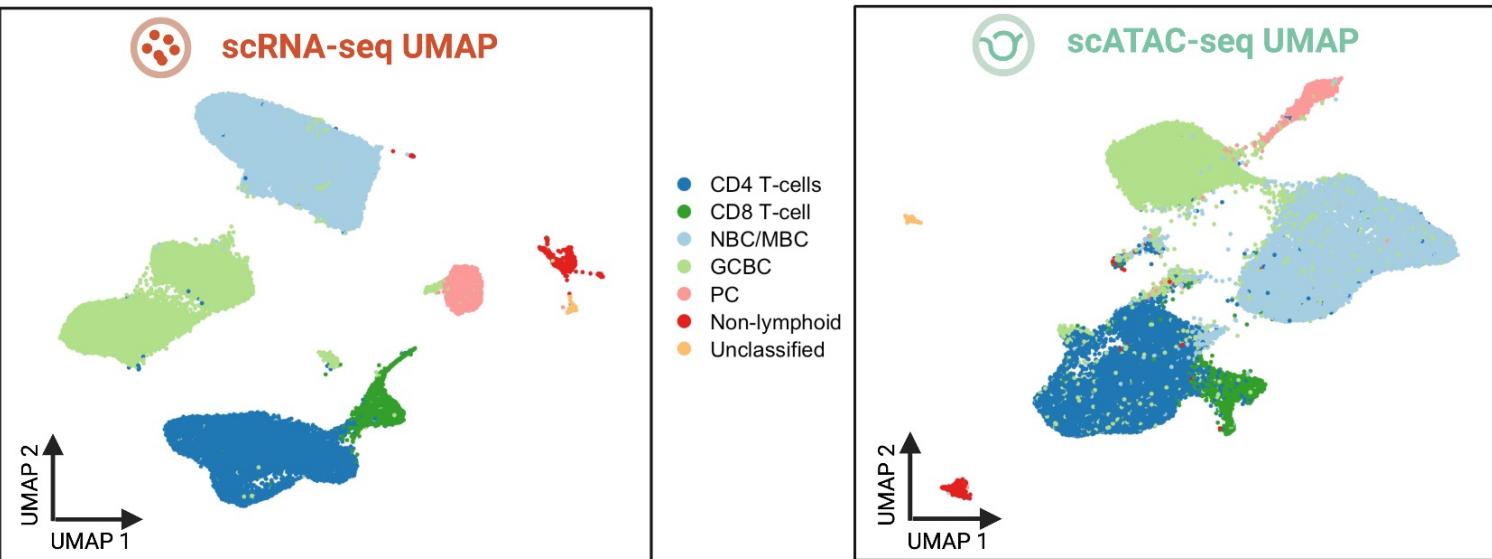
So this methods identify similarities depended on the weight:

- **High weight** = more similarity.
- **Low weight** = less similarity

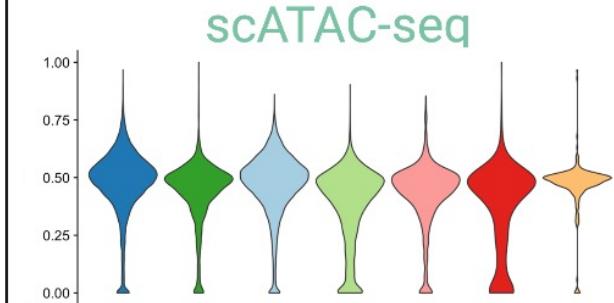
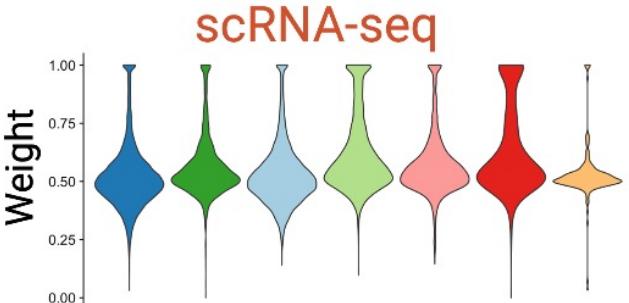
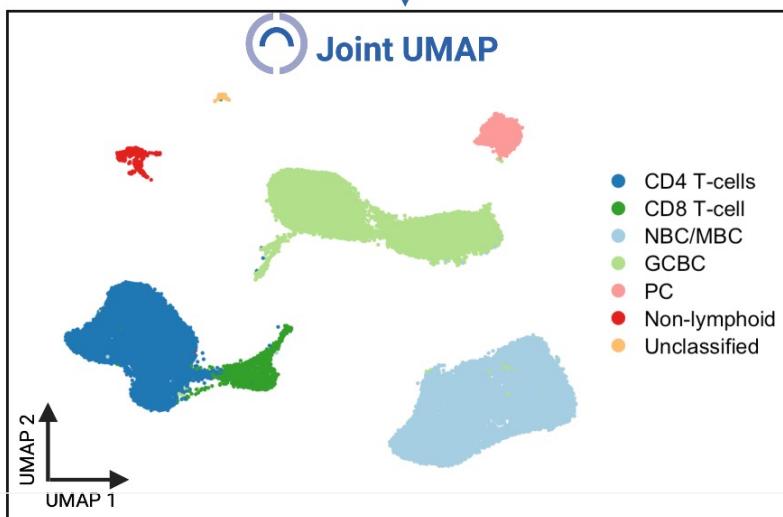


Multiome analysis workflow

Modality integration (scRNA-seq + scATAC-seq)



Weight Nearest
Neighbour
(WNN)

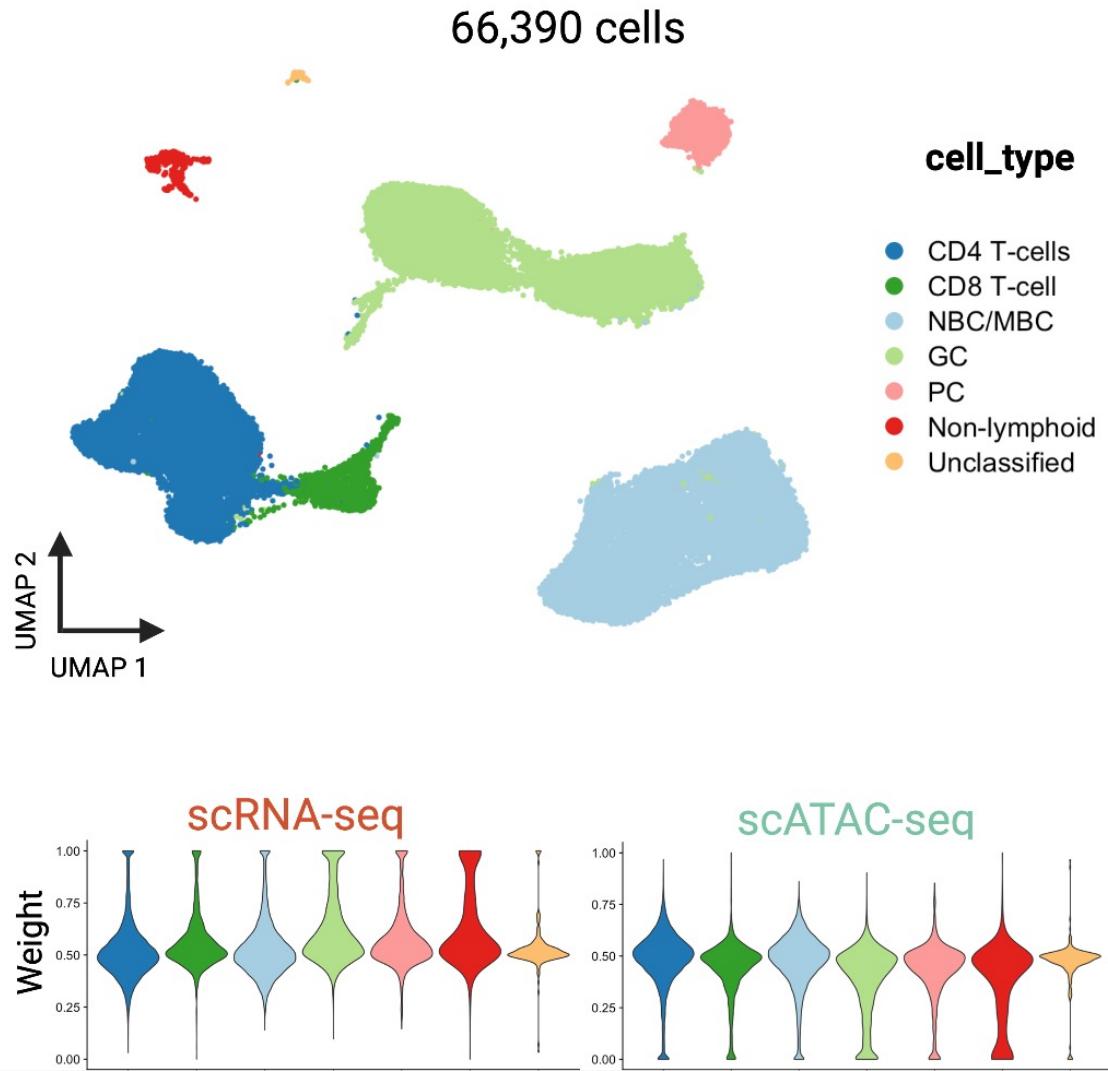


Multiome analysis workflow

Modality integration

(scRNA-seq + scATAC-seq)

Joint UMAP



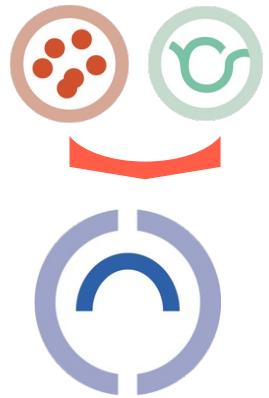
1° ***FindModalNeighbors***

This function will construct a weighted nearest neighbor (WNN) graph.

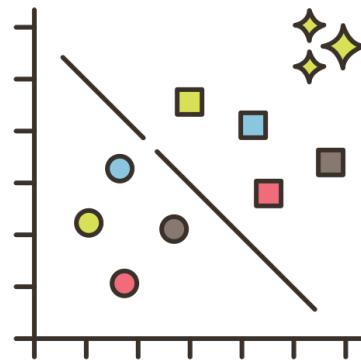
2° **RunUMAP**: as we do in scRNA-seq

In the following session

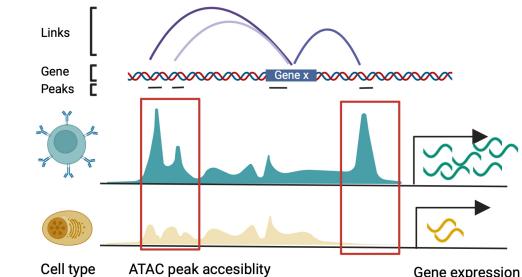
Next steps



Integration of
scATAC-seq + scRNA-seq by
Weighted Nearest Neighbour (WNN)



Clustering &
Peak calling



Peak-genes
Linkages analysis

In the following session

Homeworks

You will find 2 pipeline vignette we you will use a data set of 2 tonsil Multiome data and you will see all the functions and outputs to perform step by step all the downstream analysis of Multiome data .

- **Part_I_Seminar_QC_Filtering_Merging.Rmd:** you will perform the quality control and visualization of the results, the filtering step and the merging of both methodologies (scATAC-seq and scRNA-seq)
- **Part_I_Seminar_I_Normalization_Batch.Rmd :**you will normalize and correct the biases of the data caused by batch effect.

The image features a large, bold red word "thank you" centered in the middle. Surrounding it are numerous other words in various colors and fonts, each representing a different language's way of saying "thank you". The surrounding words include: "danke" (German), "謝謝" (Chinese), "ngiyabonga" (Swahili), "teşekkür ederim" (Turkish), "спасибо" (Russian), "apărat" (Romanian), "dank je" (Dutch), "gracias" (Spanish), "mochchakkeram" (Korean), "go raibh maith agat" (Irish), "obrigado" (Portuguese), "dziekuje" (Polish), "sukriya" (Hindi), "kop khun krap" (Thai), "terima kasih" (Indonesian), "감사합니다" (Korean), "grazie" (Italian), "arigatō" (Japanese), "takk" (Norwegian), "dakujem" (Czech), and "мерси" (Ukrainian). Each word is written in its respective language's script and color.

For your attention