Single Cell Epigenomics
scATAC sequencing

Fides Zenk

Learning Objectives of this week

Knowing different types of chromatin

How do identify/map accessible chromatin

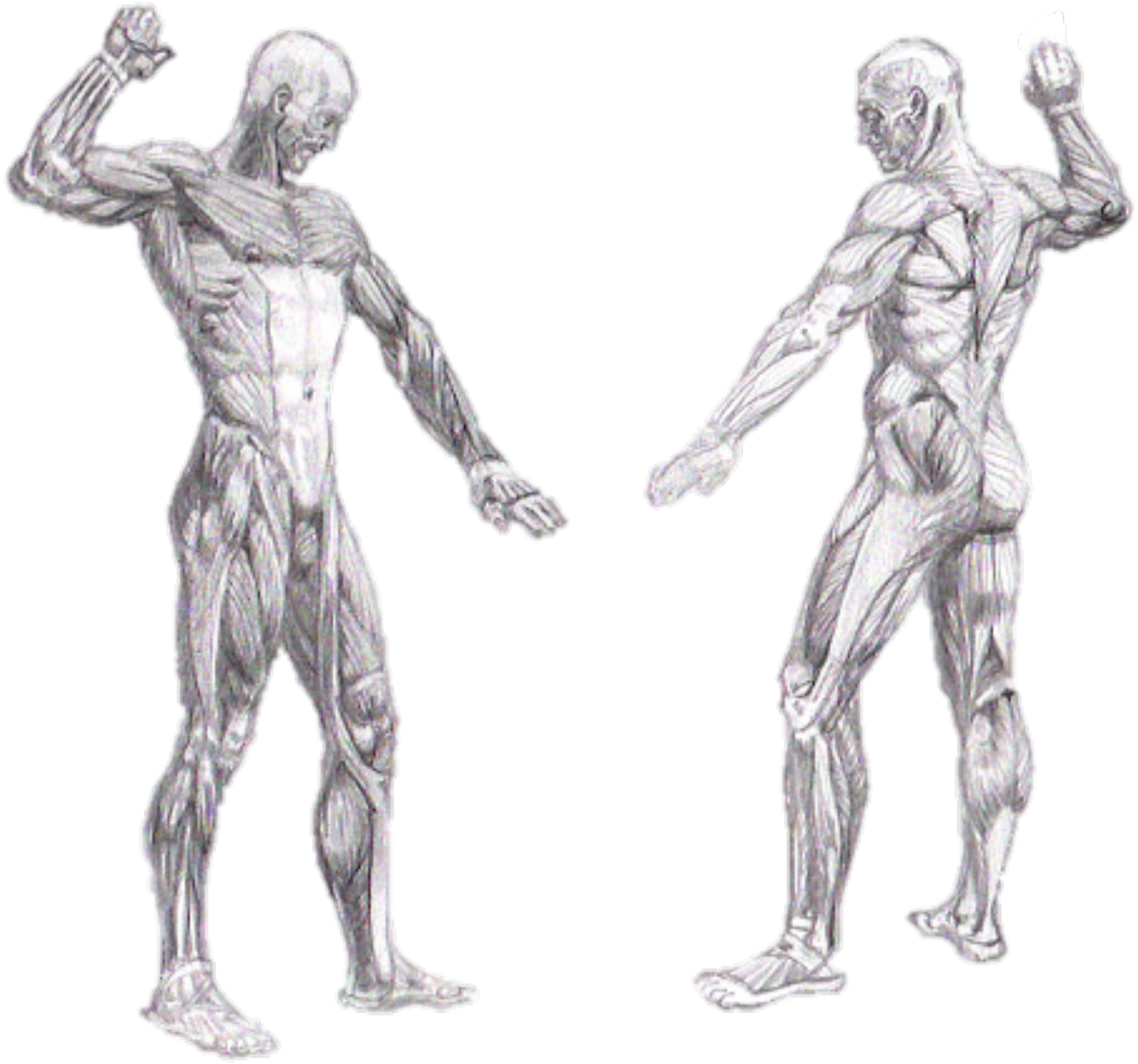
Isolating accessible chromatin in single cells

Sequencing libraries using NGS sequencing

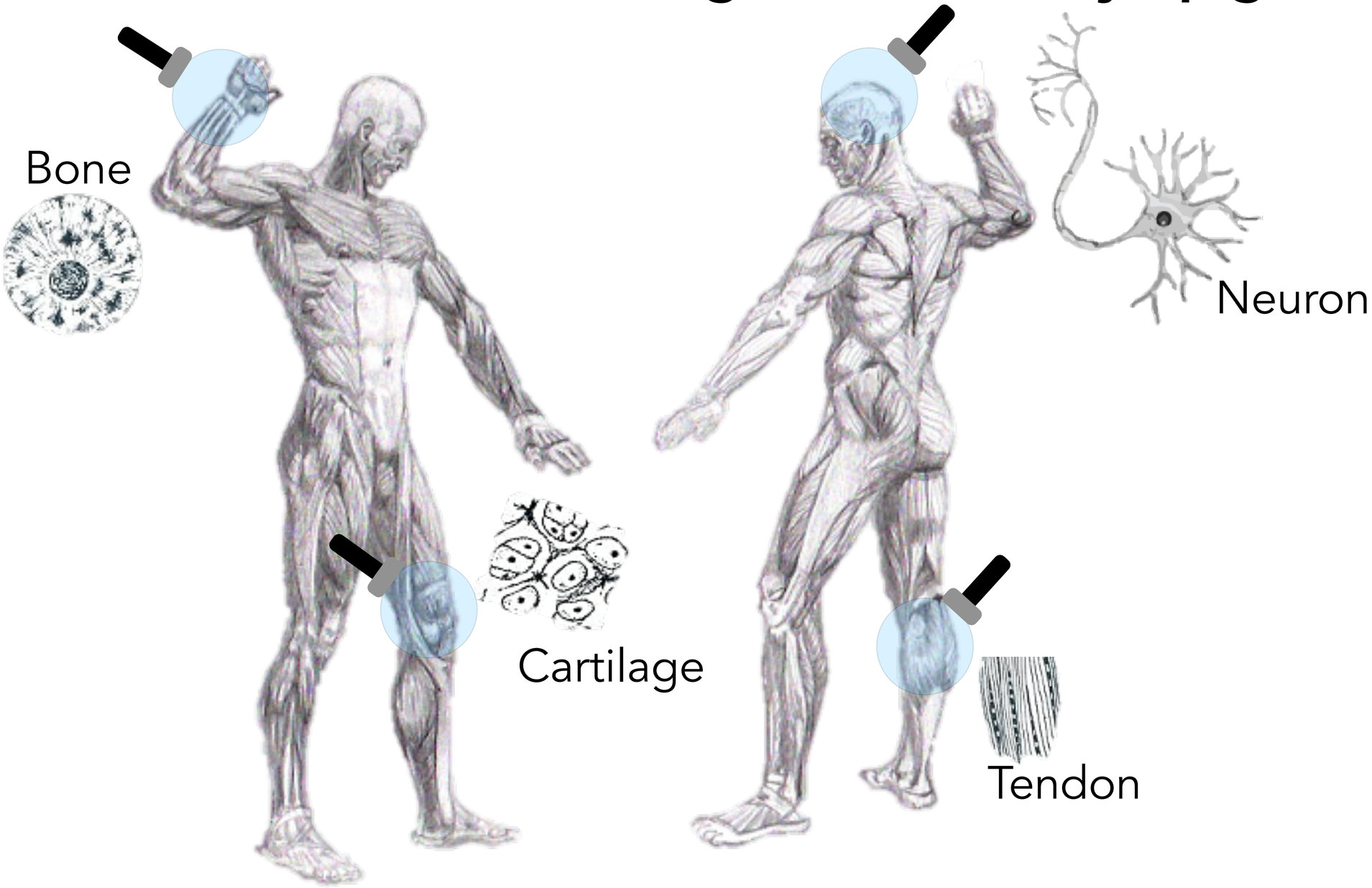
Visualizing chromatin accessibility in the genome browser

How to analyze and interpret single cell ATAC data - hands on tutorial will follow

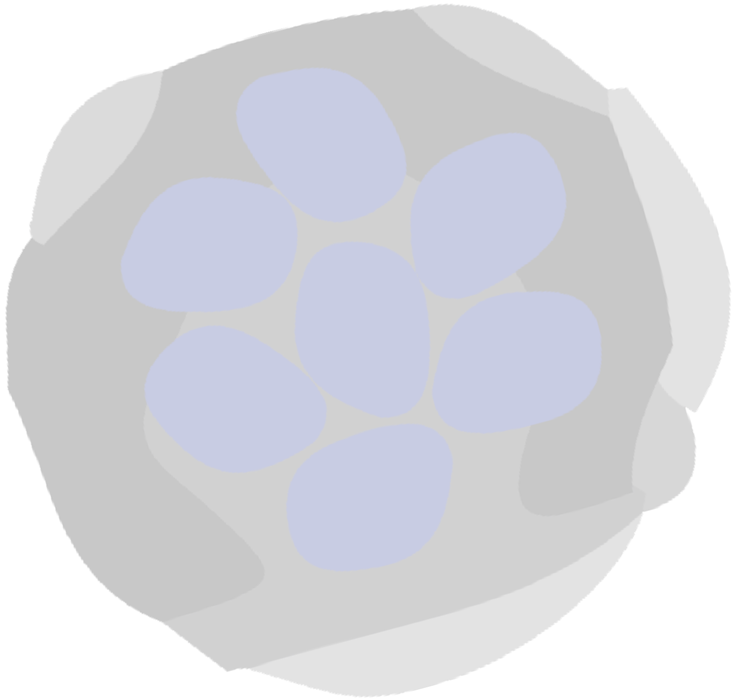
One genome - many epigenomes



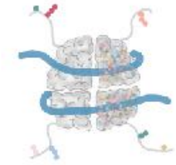
One genome - many epigenomes



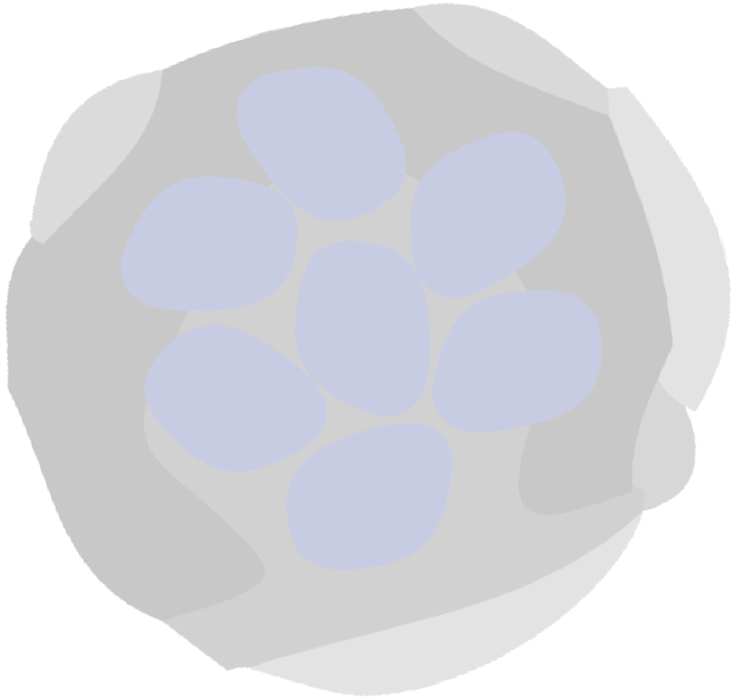
Packaging of Chromatin inside the Nucleus

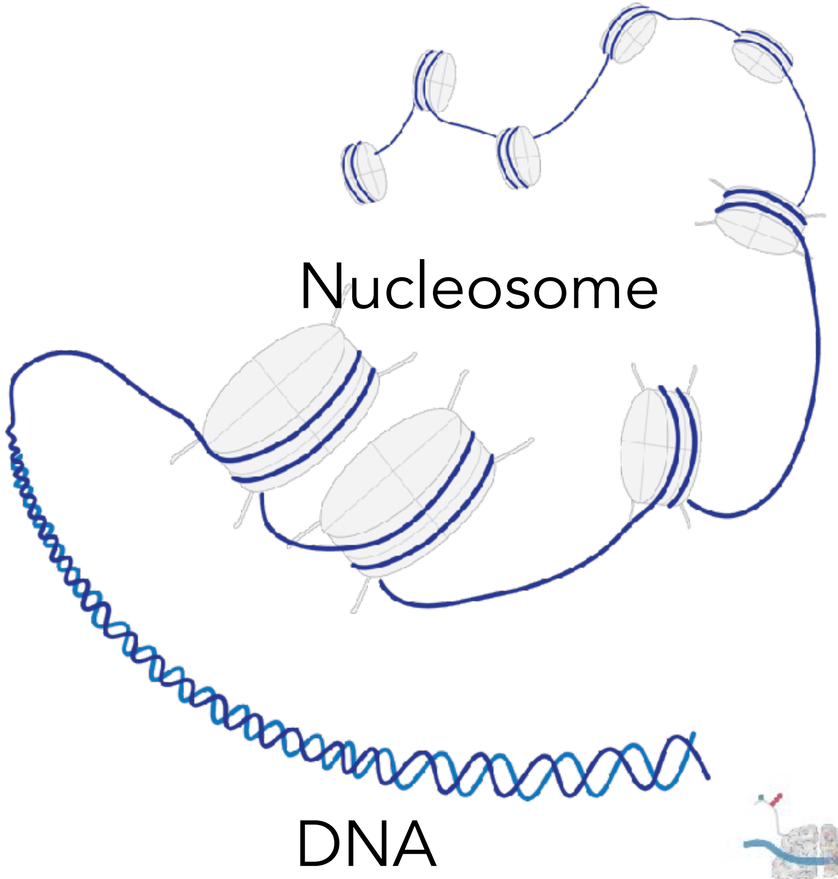


Chromosome Territories

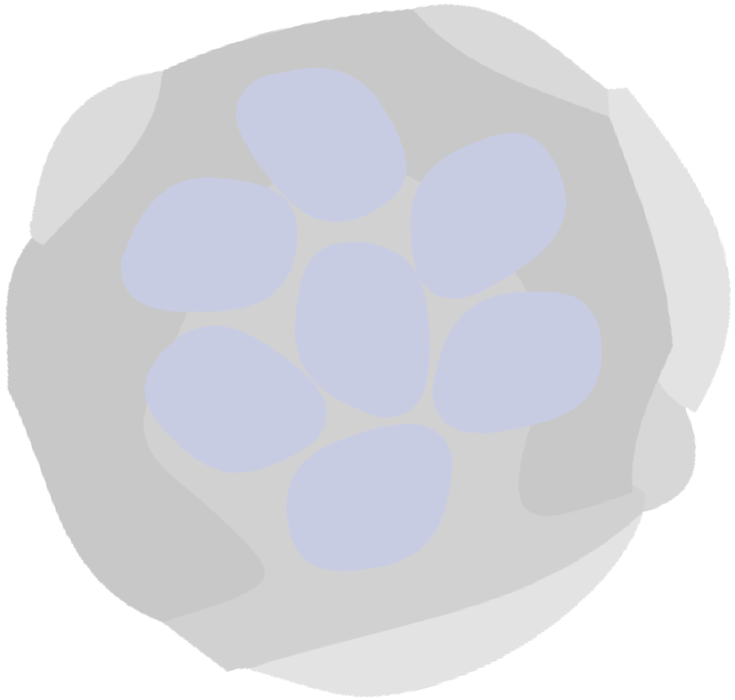


Packaging of Chromatin inside the Nucleus

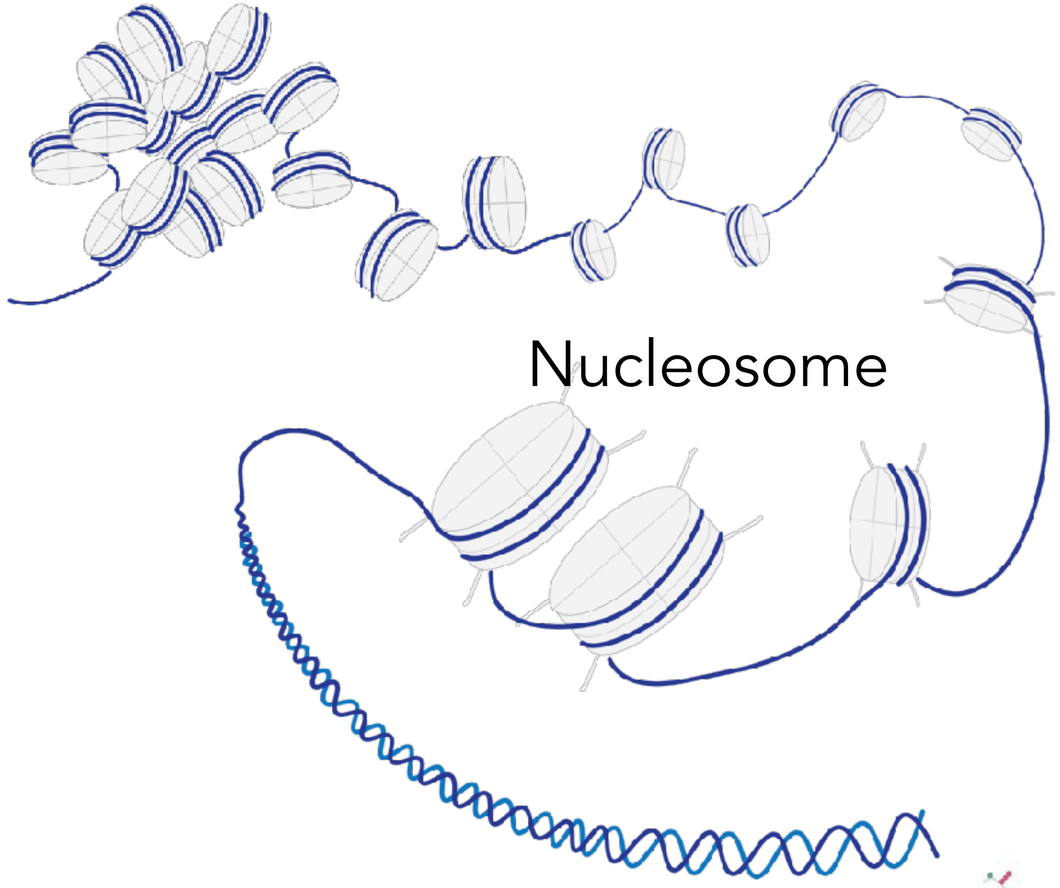




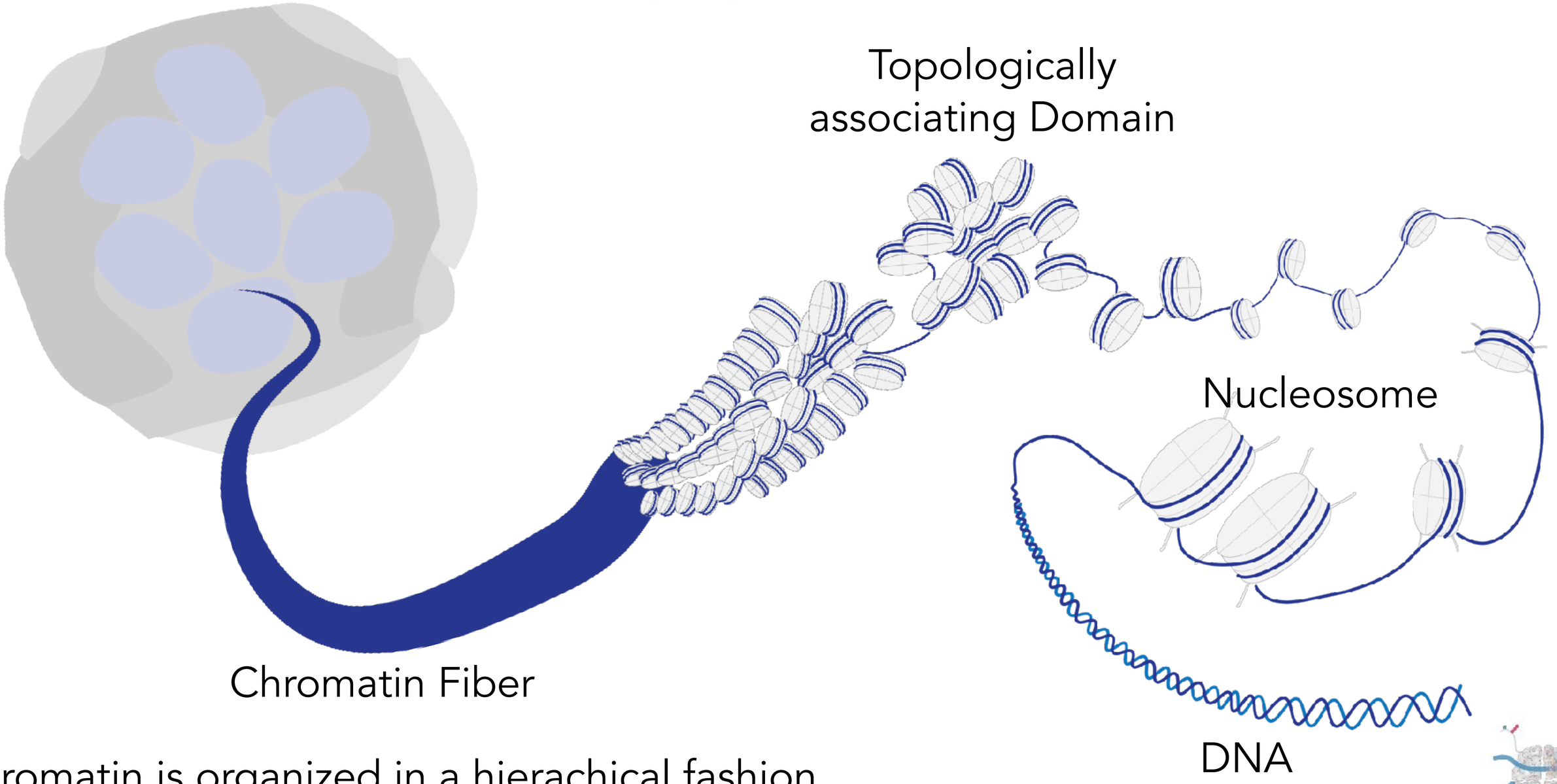
Packaging of Chromatin inside the Nucleus



Topologically
associating Domain

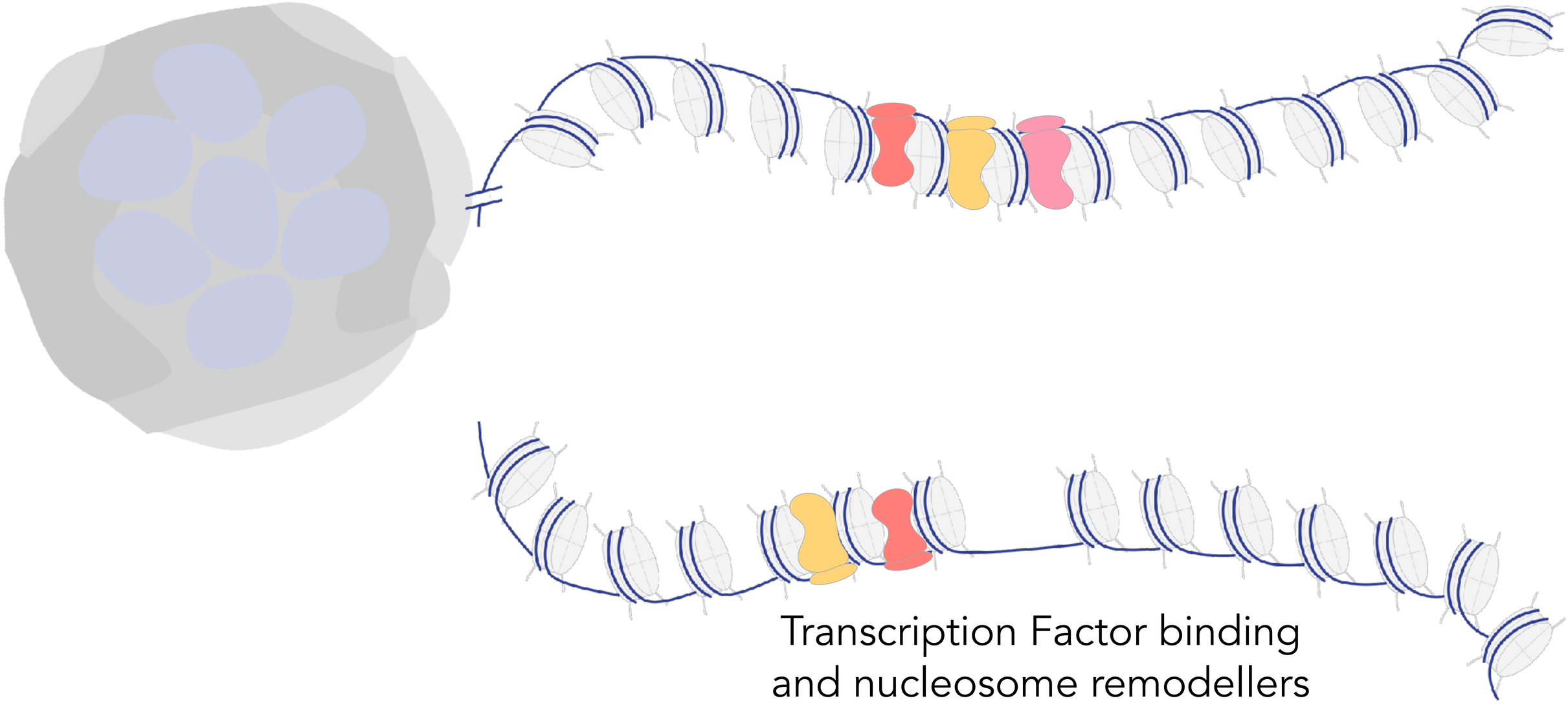


Packaging of Chromatin inside the Nucleus



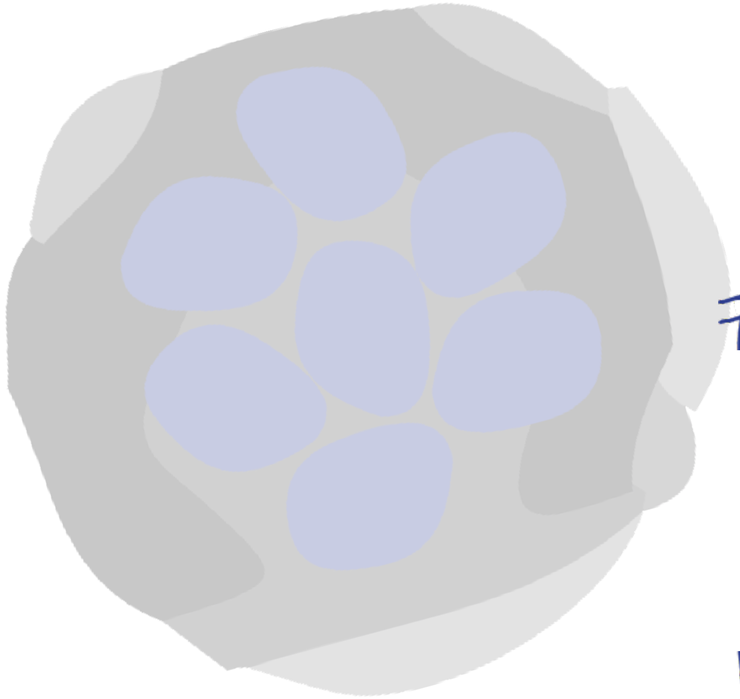
Chromatin is organized in a hierachical fashion

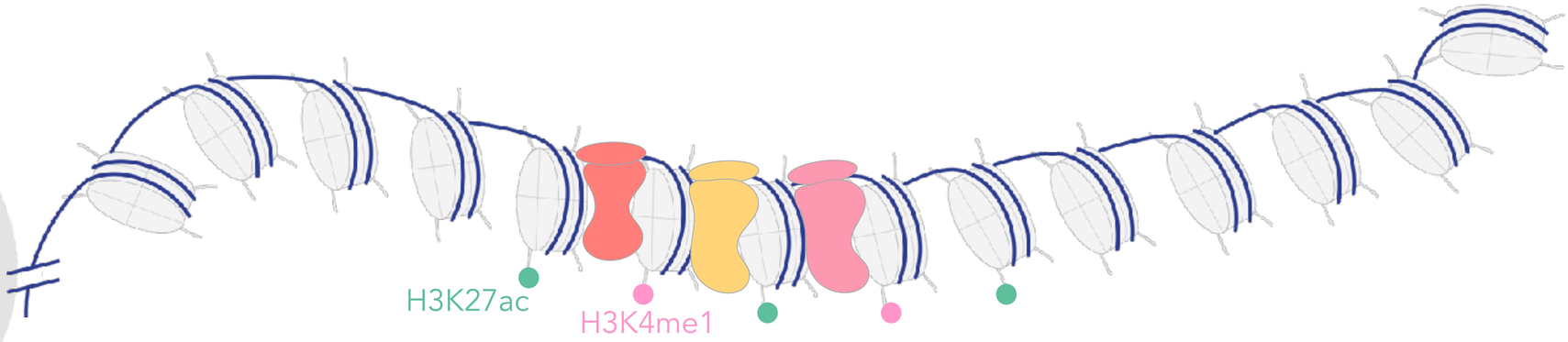
Epigenetic Mechanisms control Gene Activity



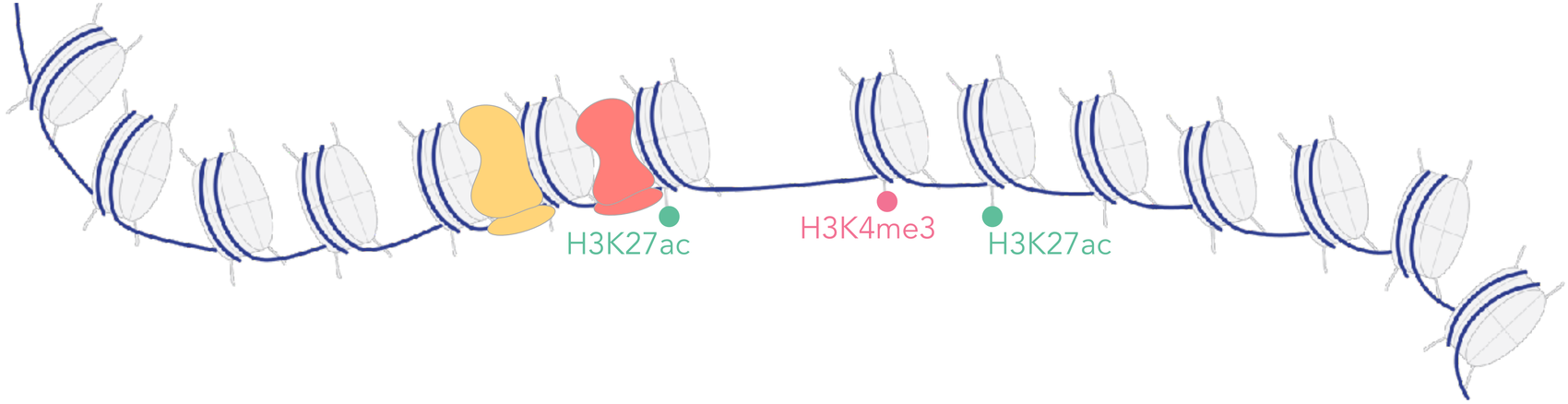
Transcription Factor binding
and nucleosome remodellers
impact chromatin accessibility

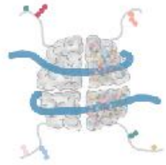
Epigenetic Mechanisms control Gene Activity



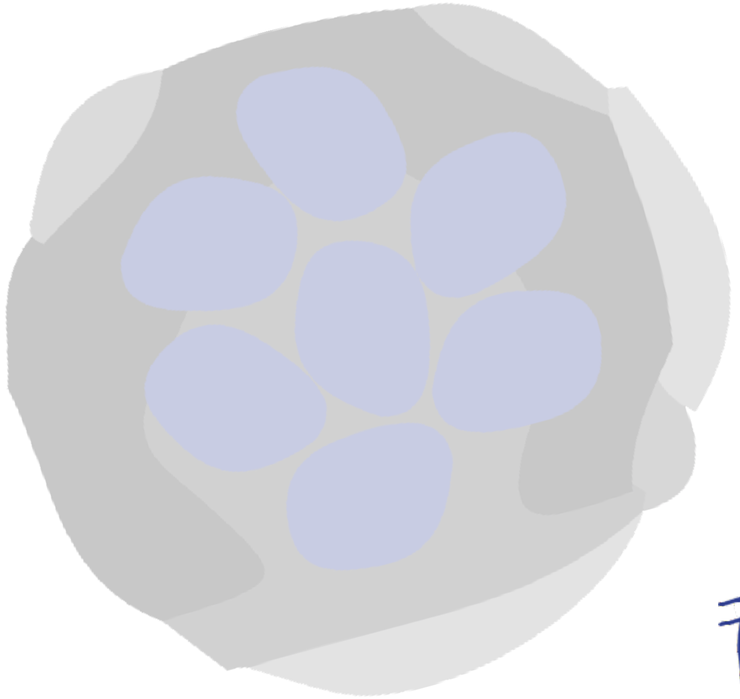


Histone modifications

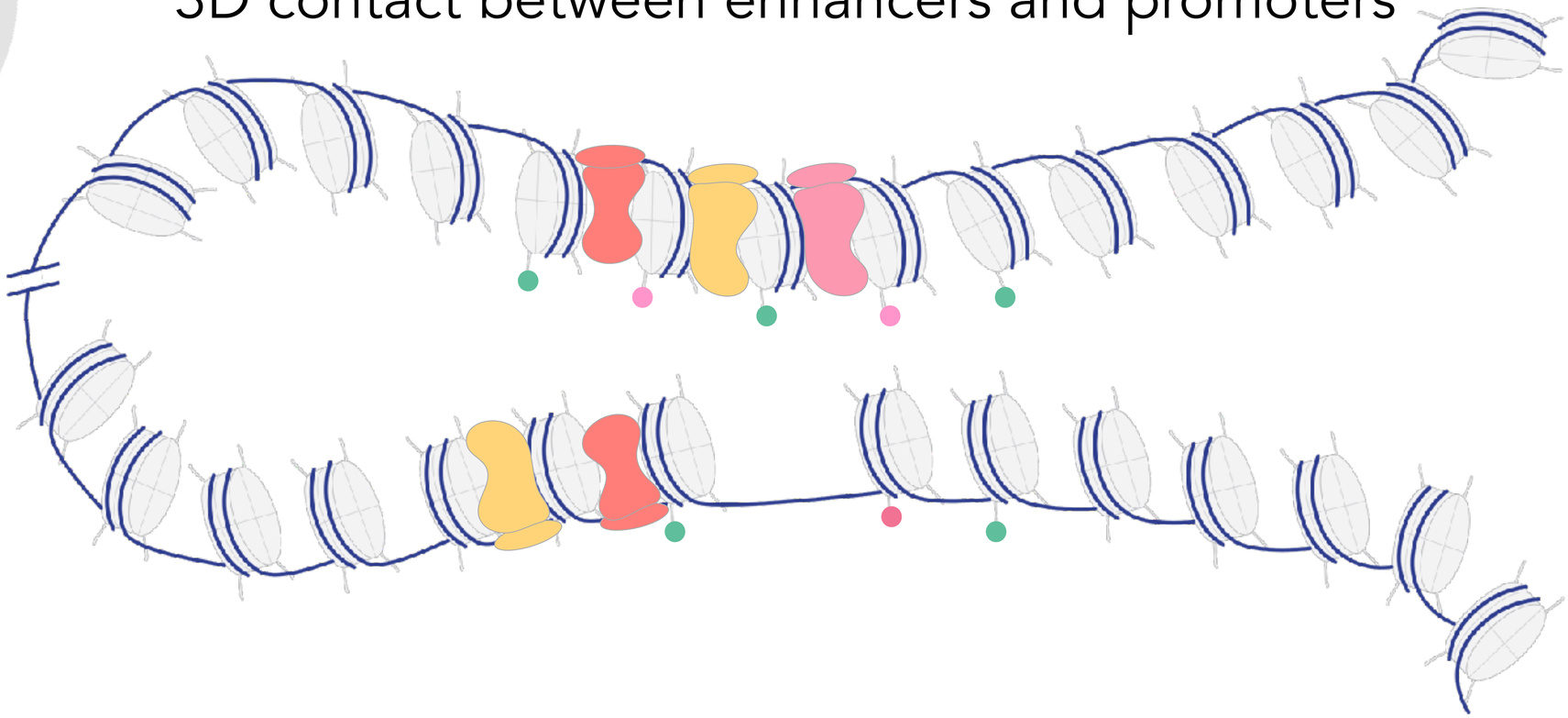


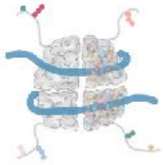


Epigenetic Mechanisms control Gene Activity

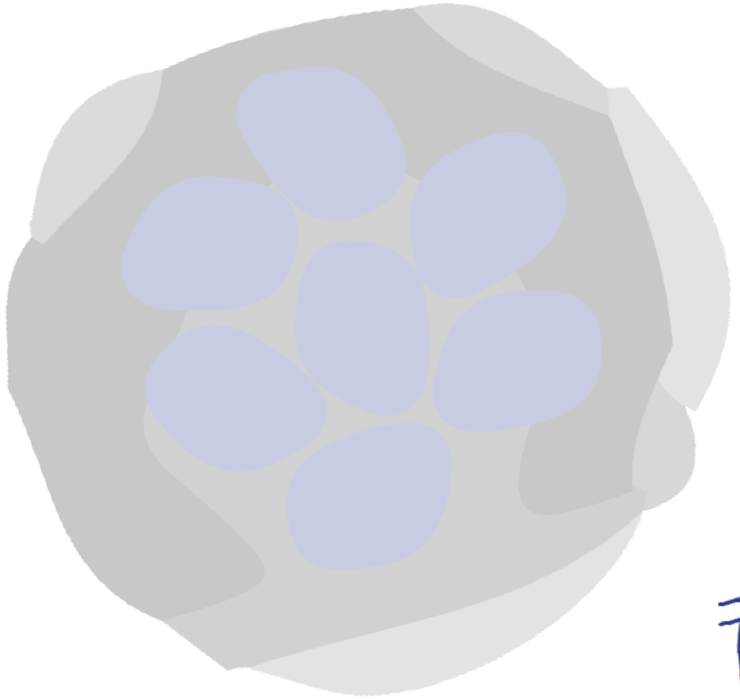


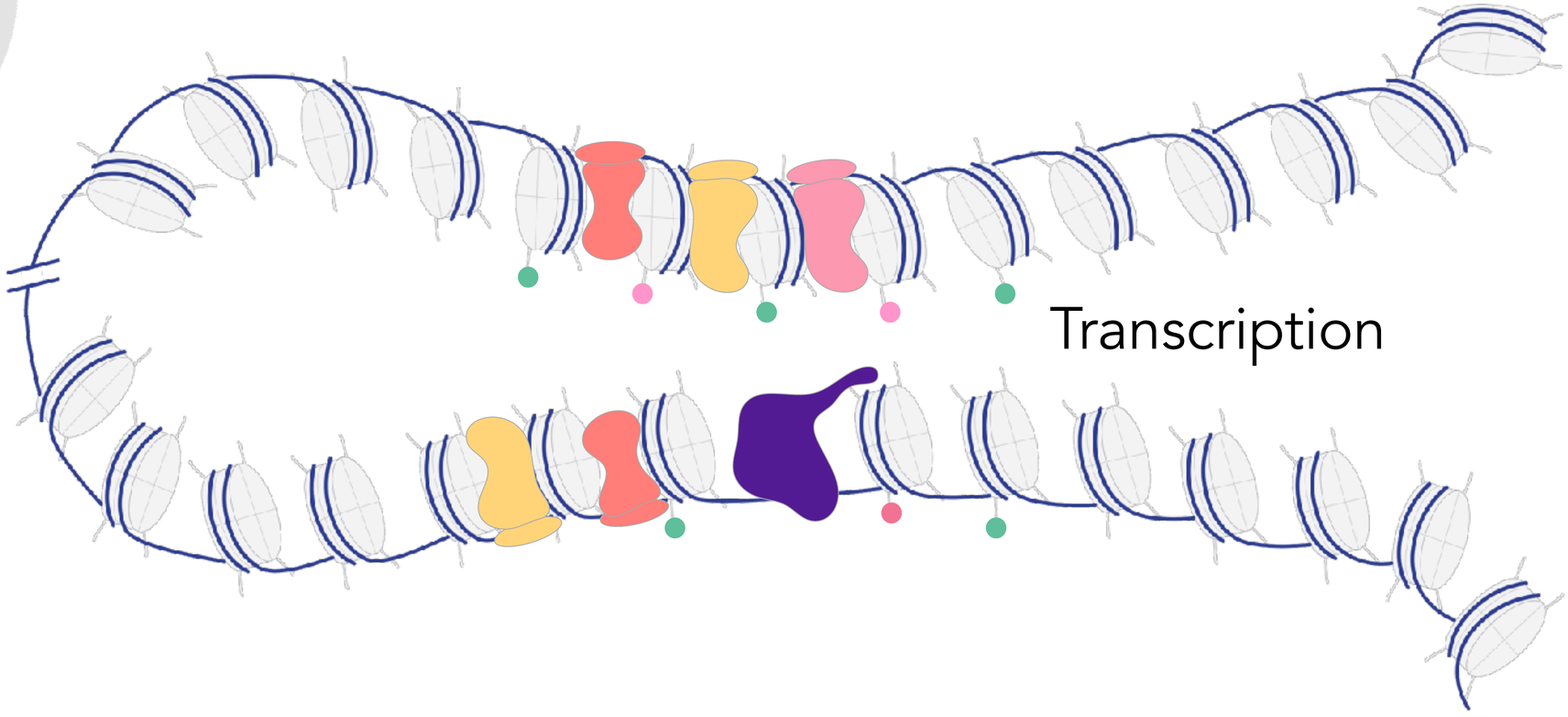
3D contact between enhancers and promoters

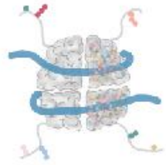




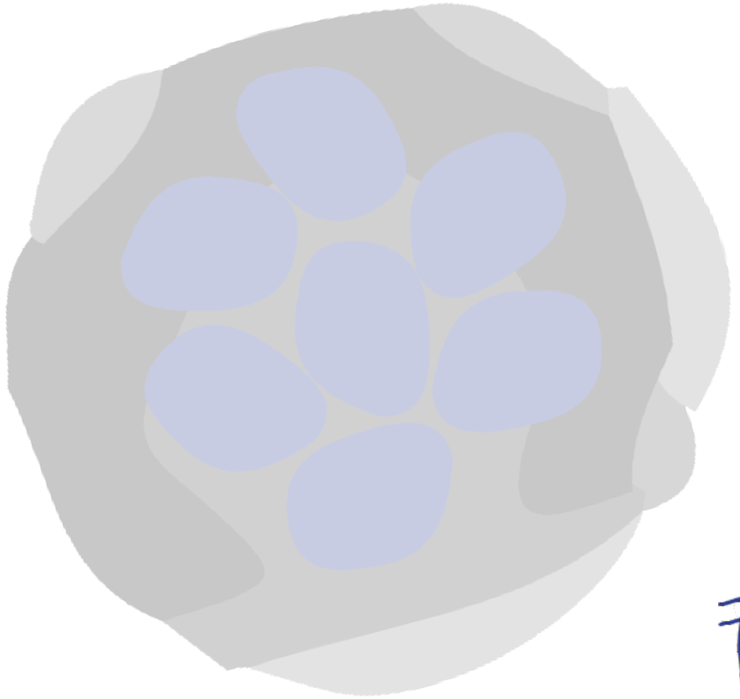
Epigenetic Mechanisms control Gene Activity

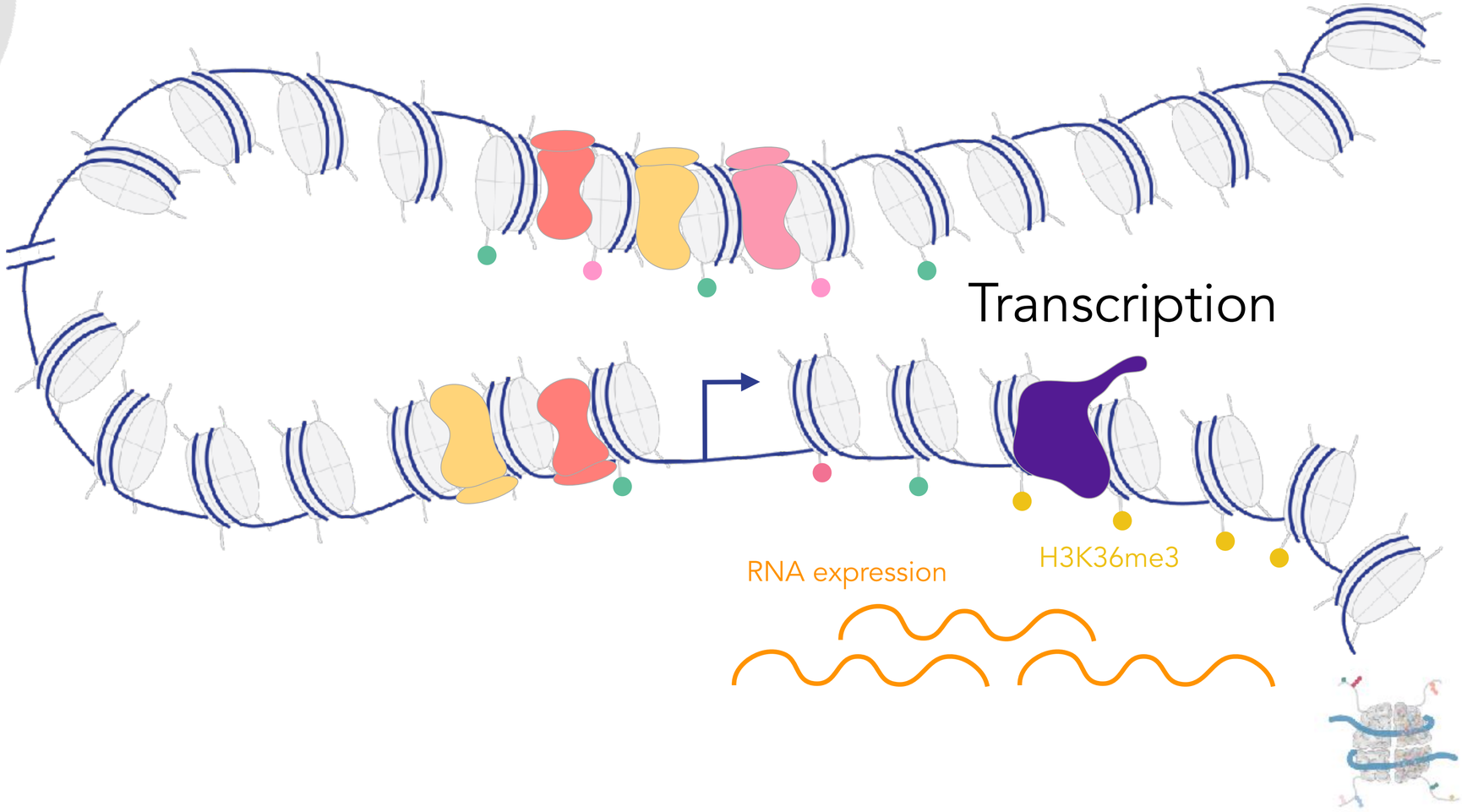




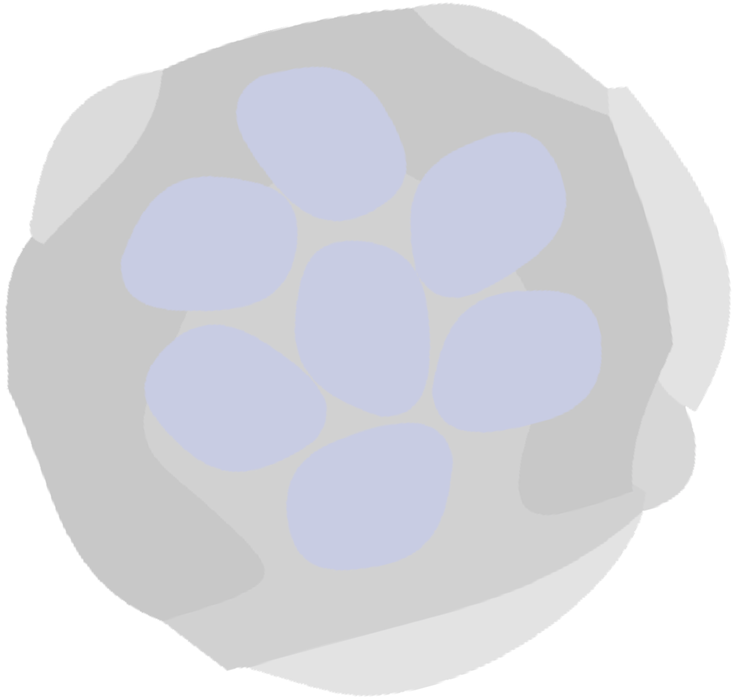


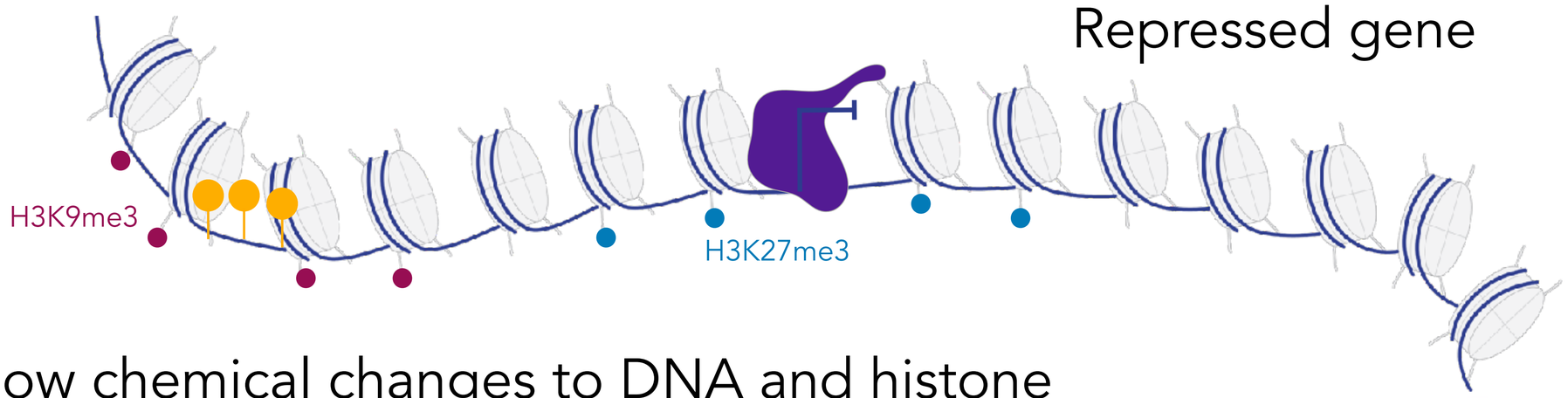
Epigenetic Mechanisms control Gene Activity



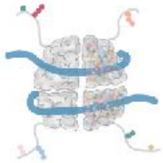


Epigenetic Mechanisms control Gene Activity

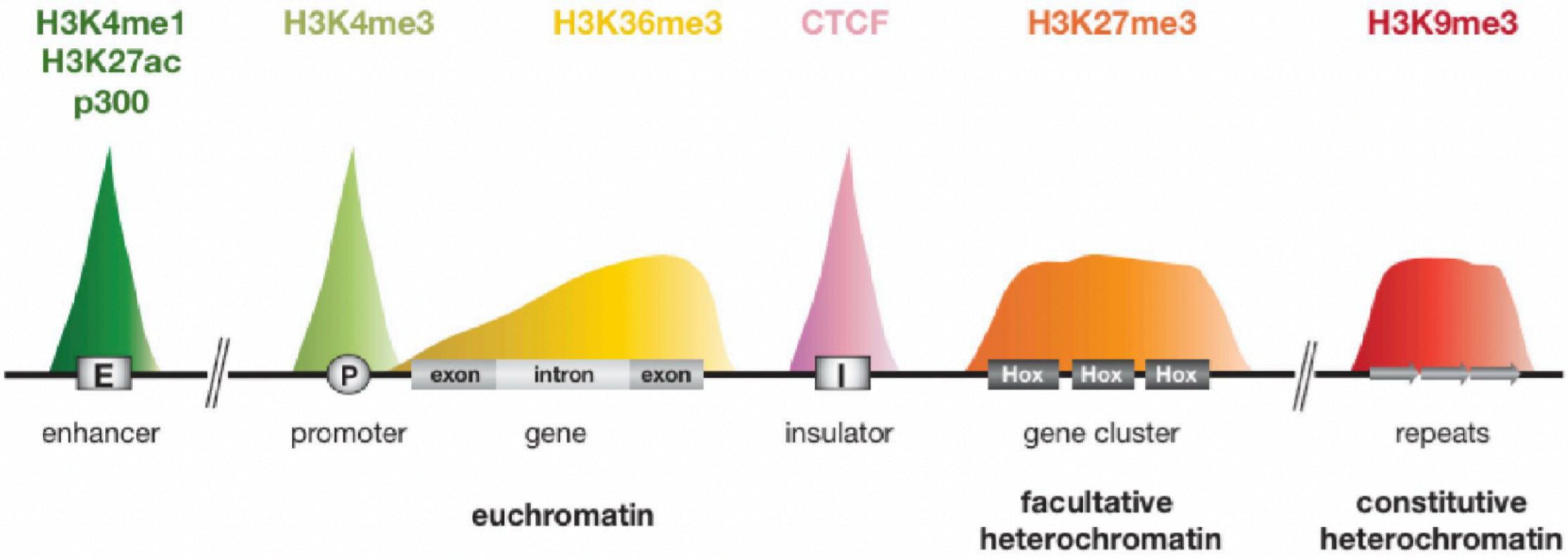




Epigenetics describes how chemical changes to DNA and histone proteins can influence gene expression

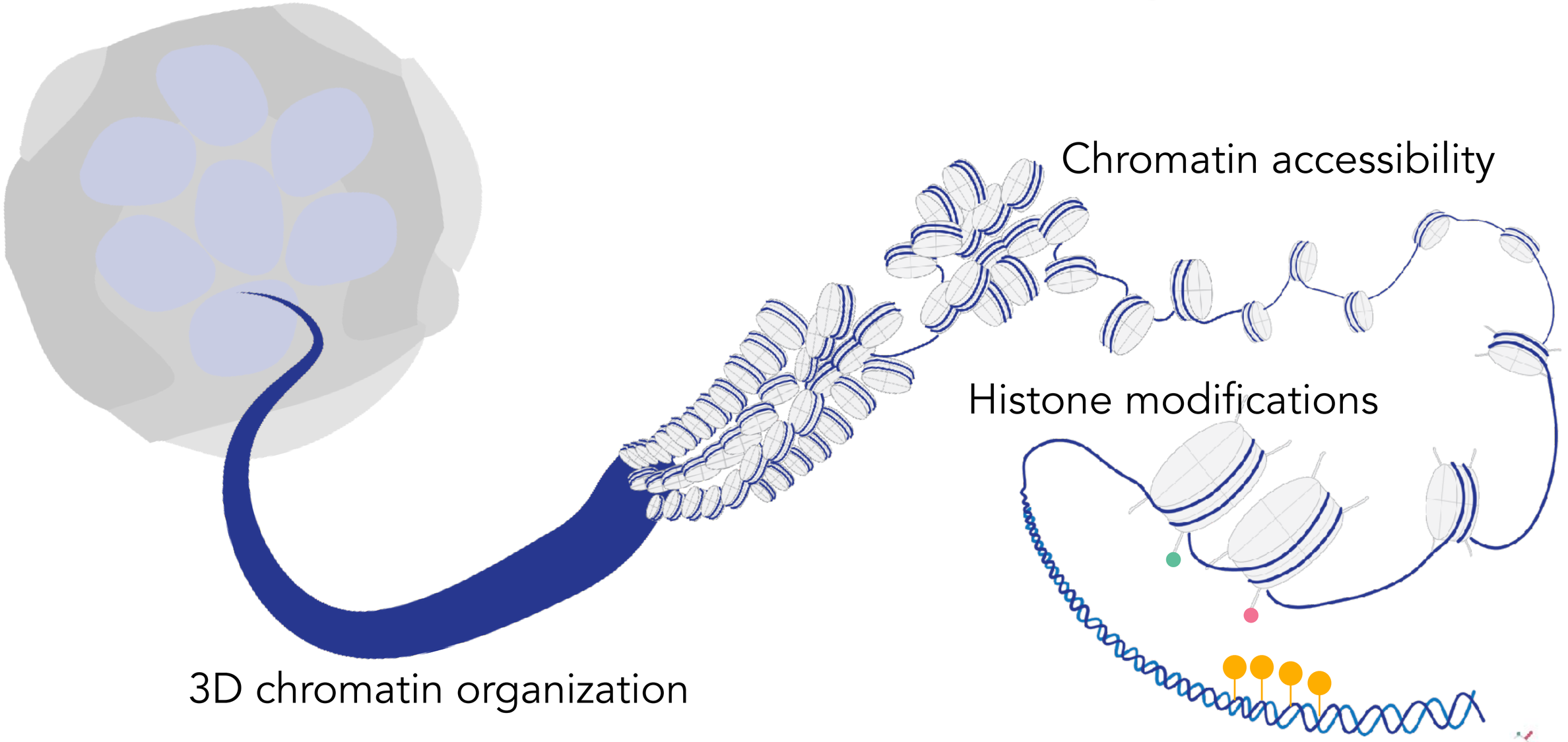


Histone modifications are differentially enriched in the genome

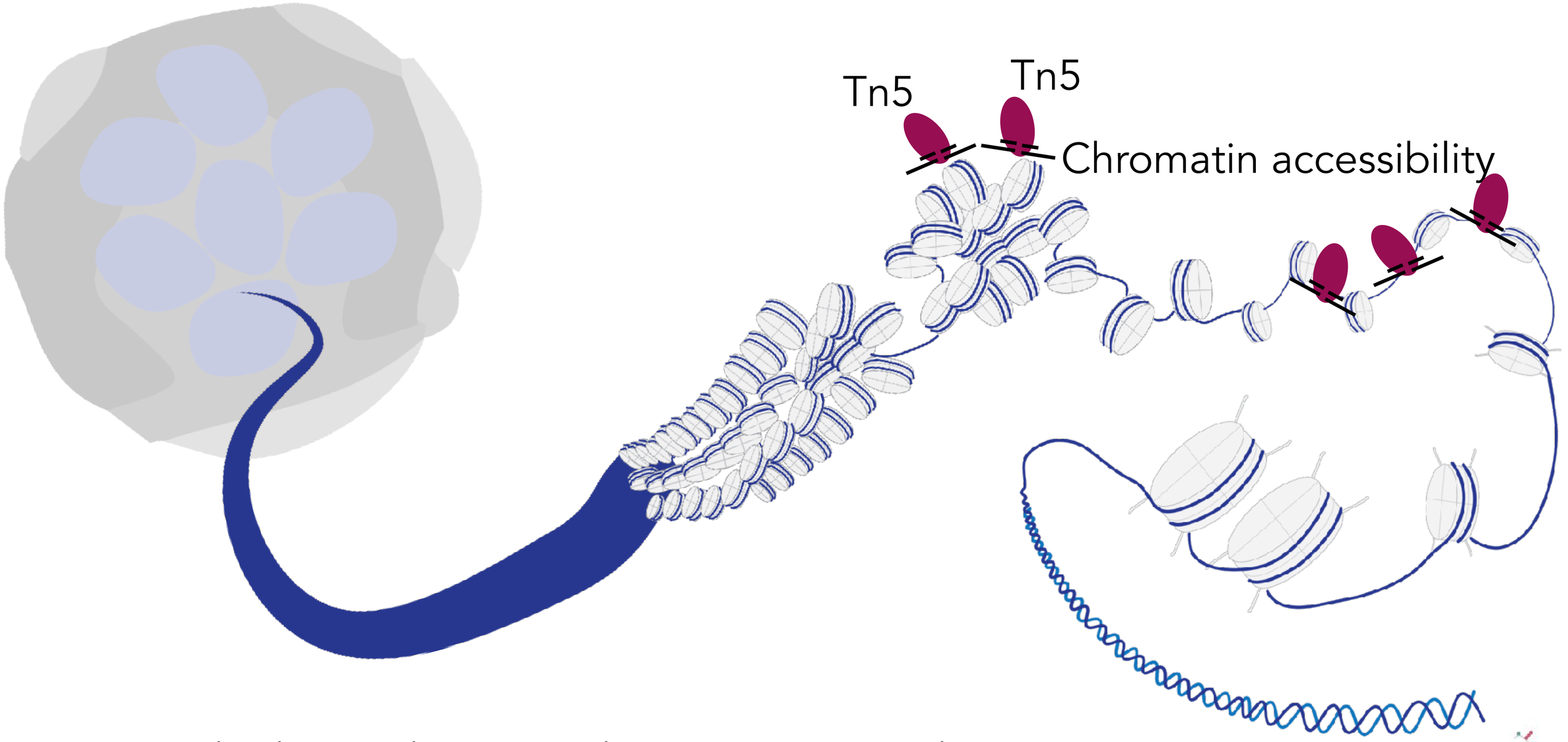


Jenuwein & Allis Epigenetics (2009)

How can we measure and study these features?

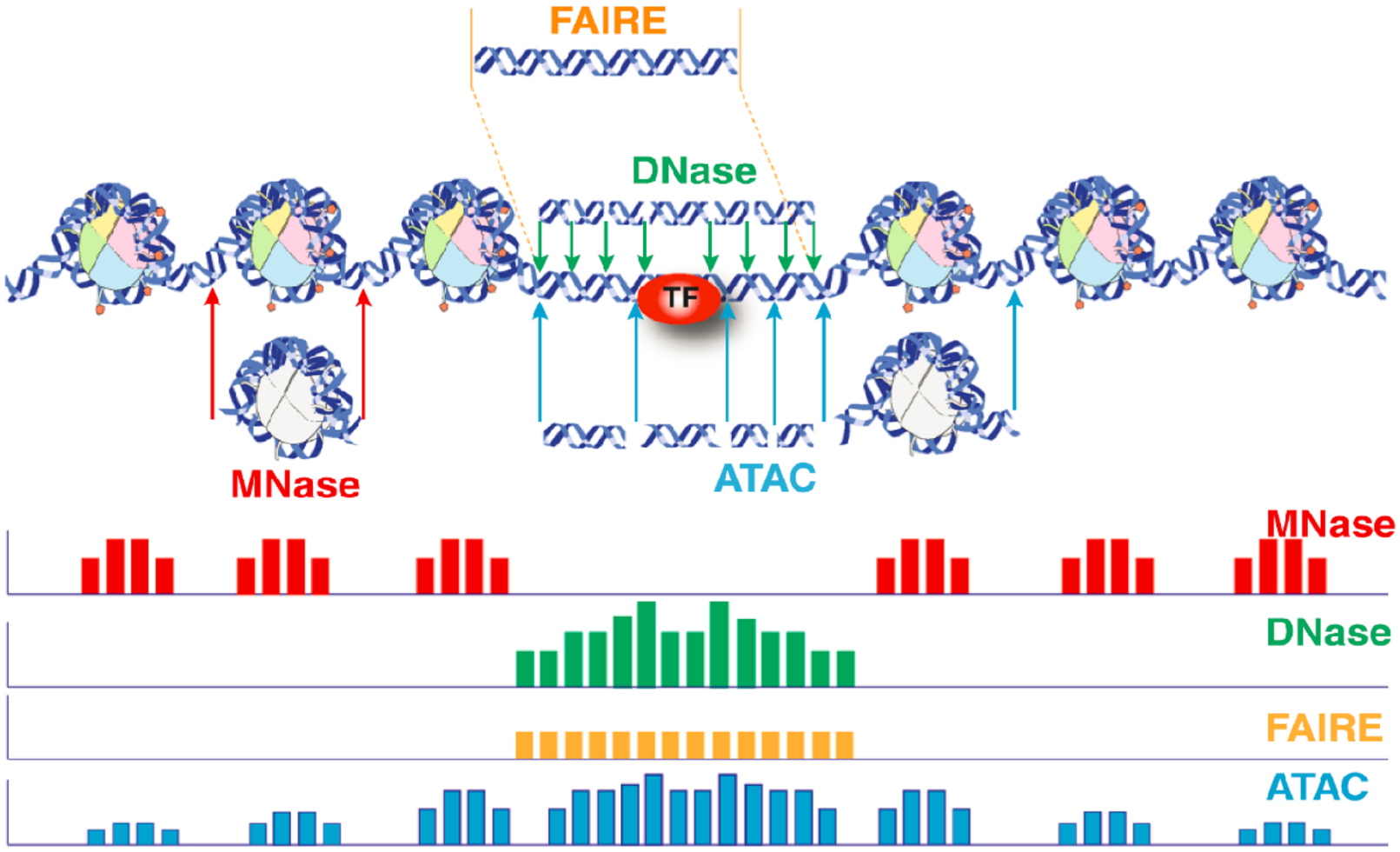


ATAC - assay for transposase-accessible chromatin



Tn5 (tagmentase) binds open chromatin and inserts sequencing adapters

Several different enzymes cut open chromatin regions



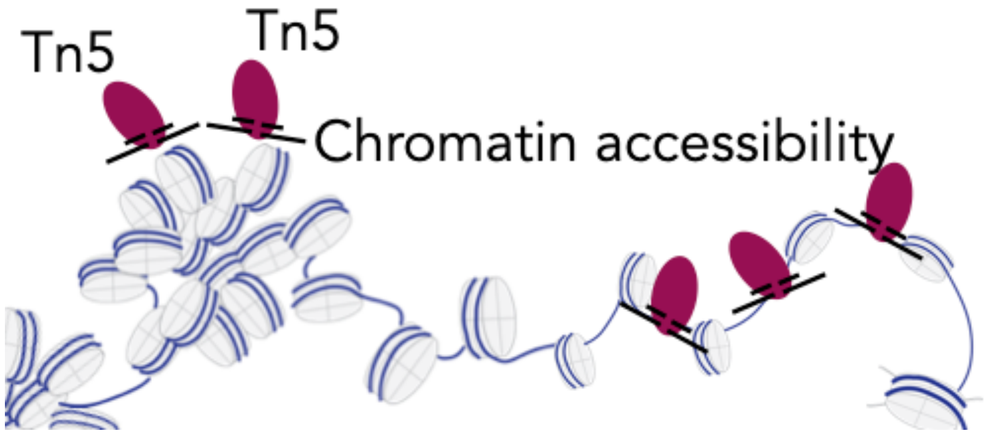
ATAC - assay for transposase-accessible chromatin

Article | Published: 06 October 2013

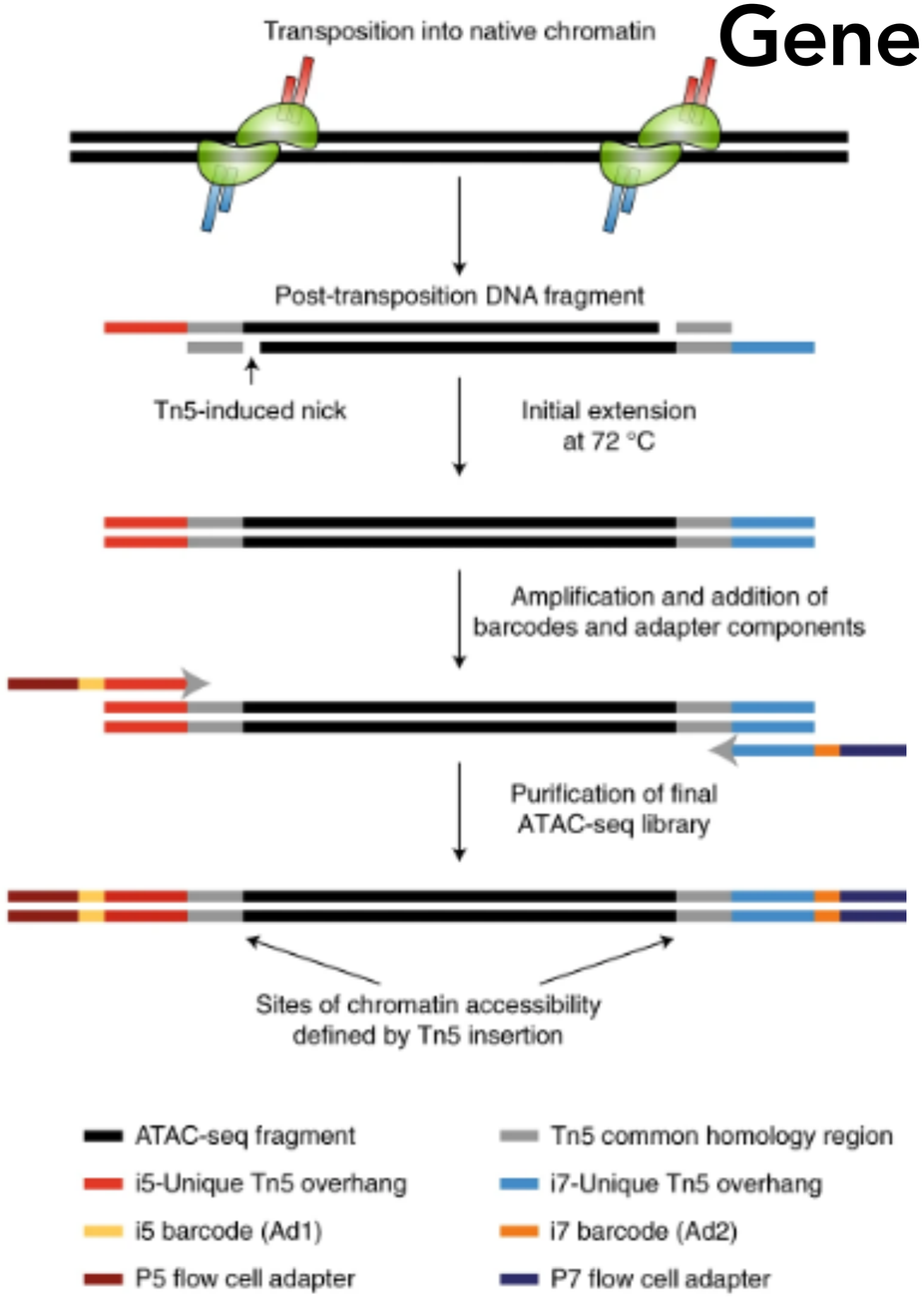
Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position

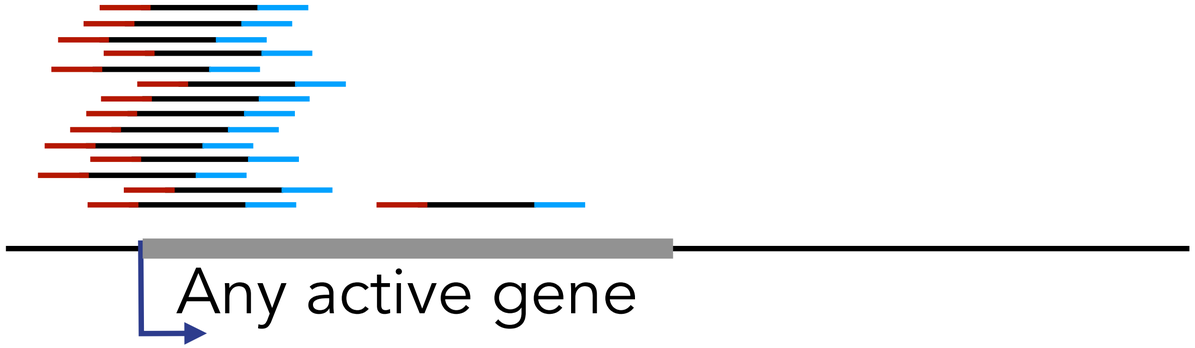
Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang ✉ & William J Greenleaf ✉

Nature Methods **10**, 1213–1218 (2013) | Cite this article

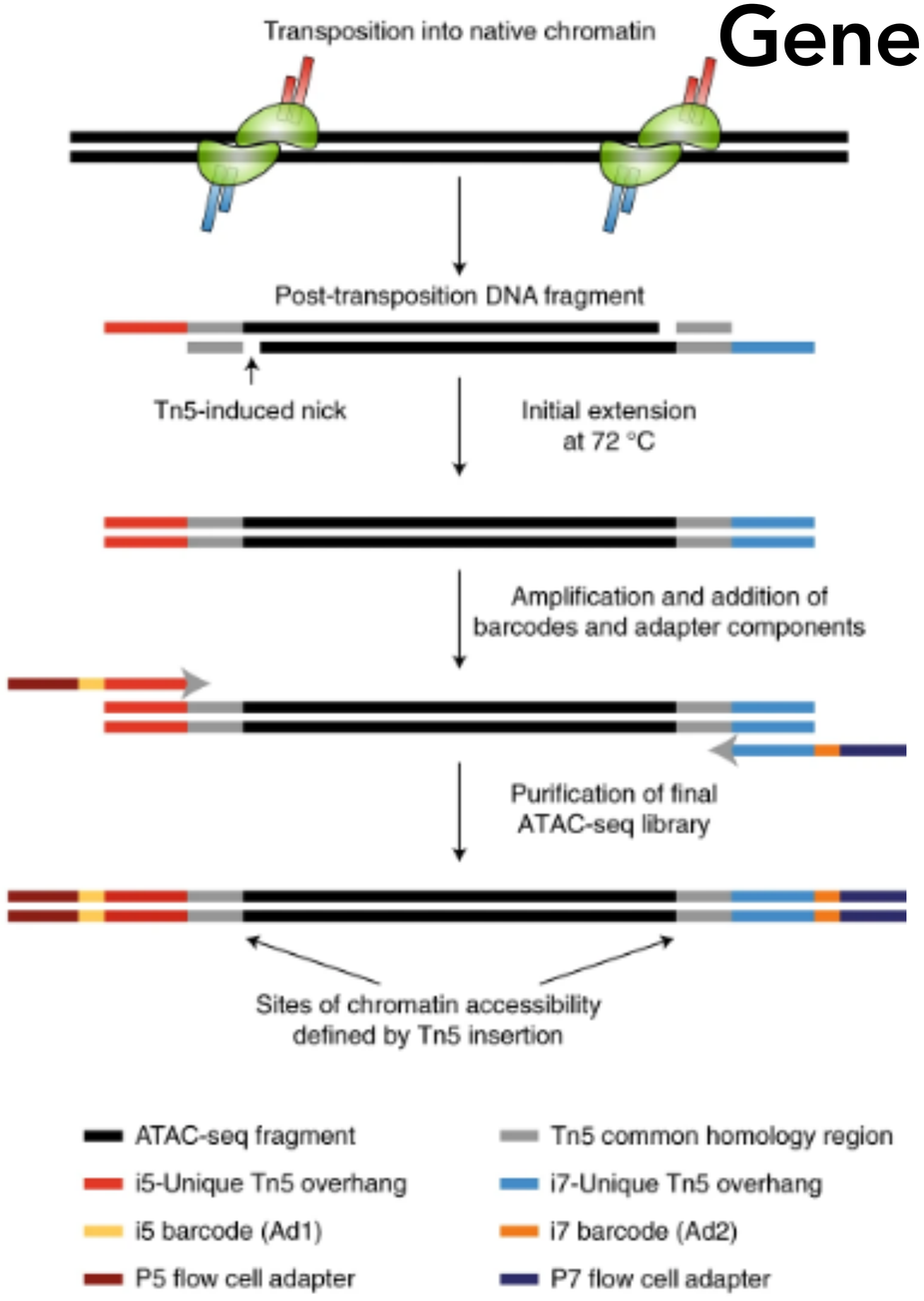


Generation of adapter-flanked DNA fragments



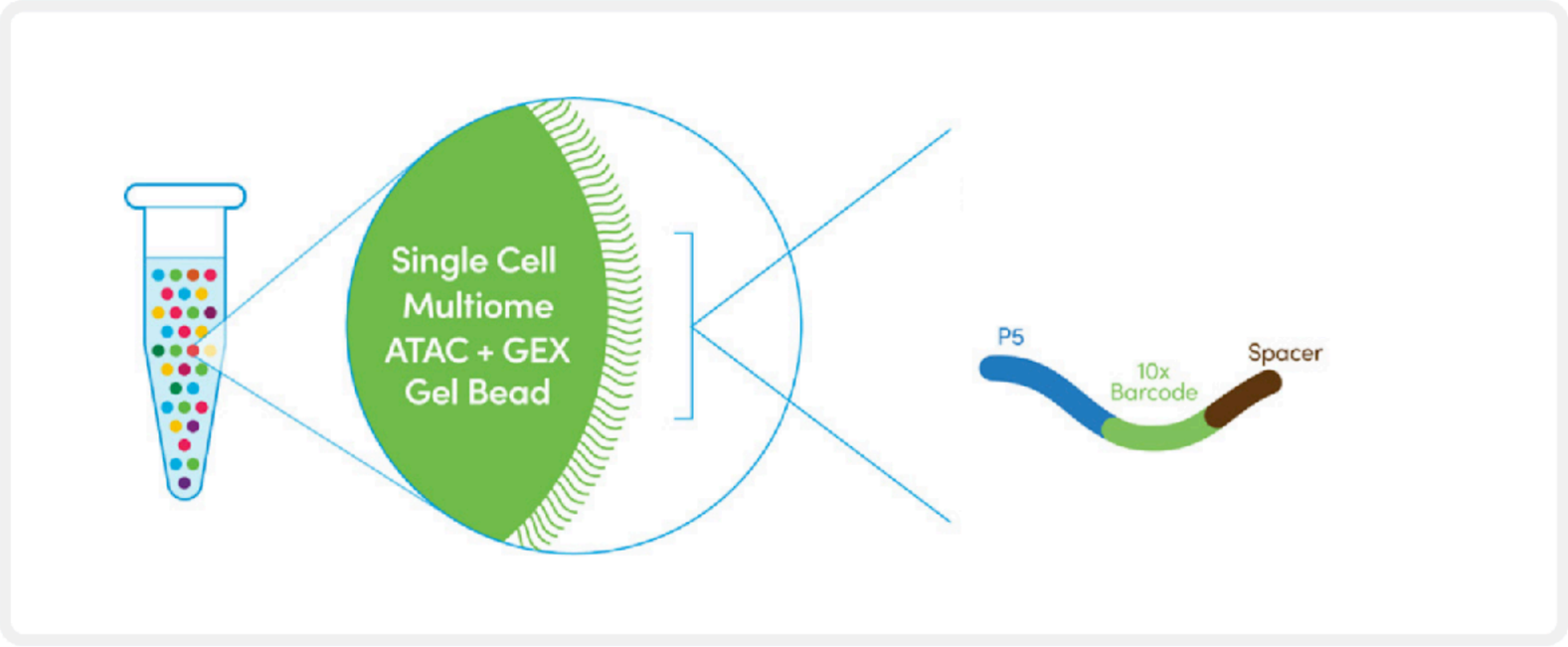


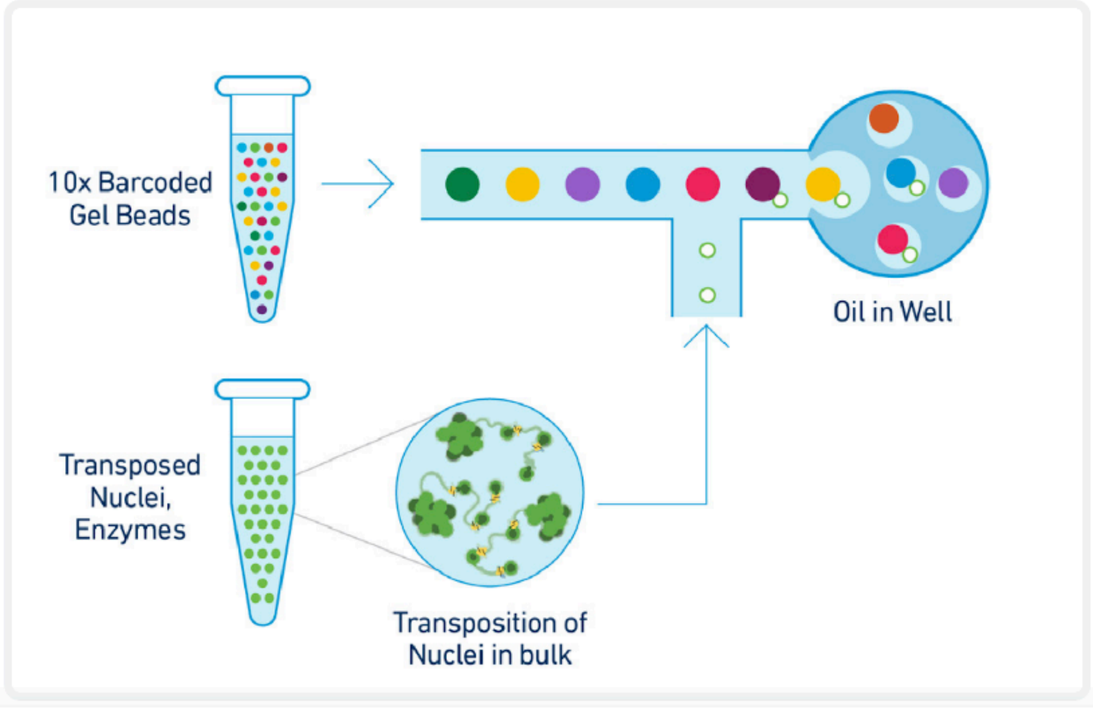
Generation of adapter-flanked DNA fragments



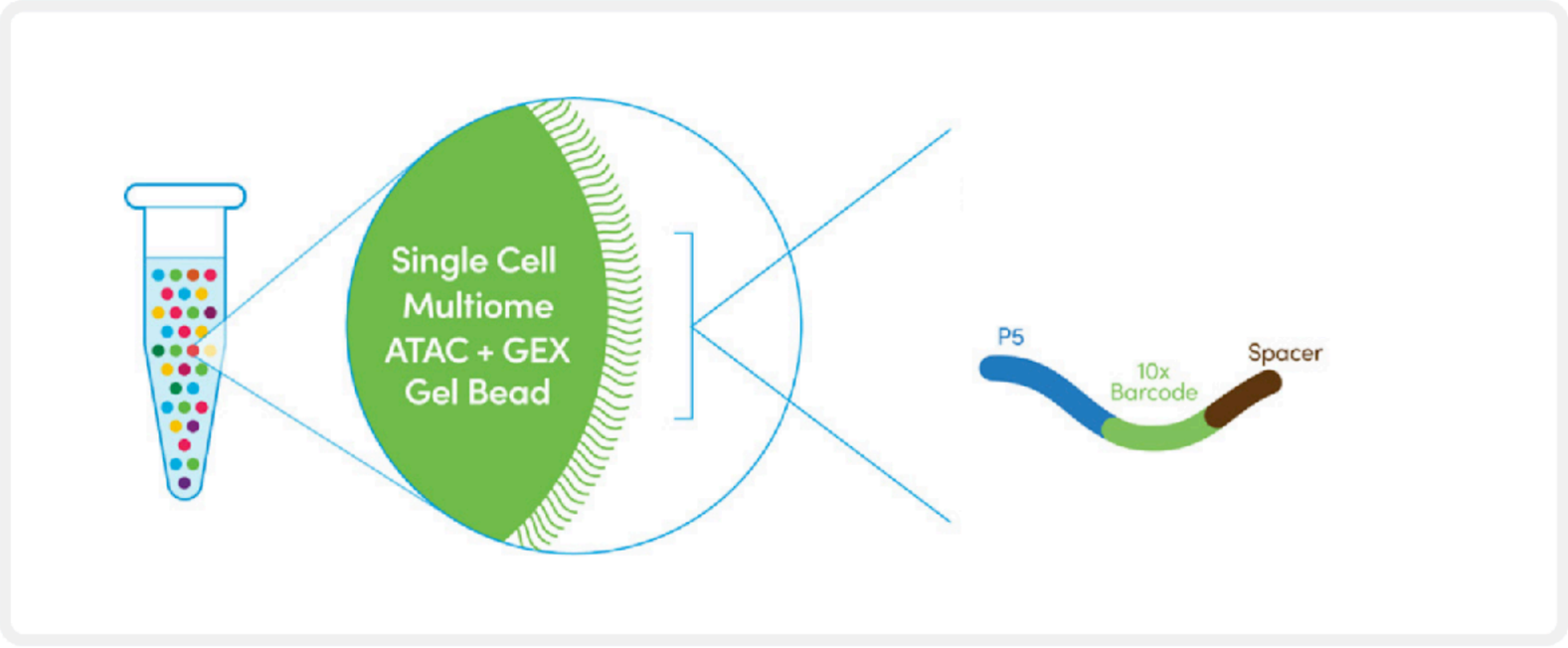
Counting sequencing reads in open chromatin and quantify the signal

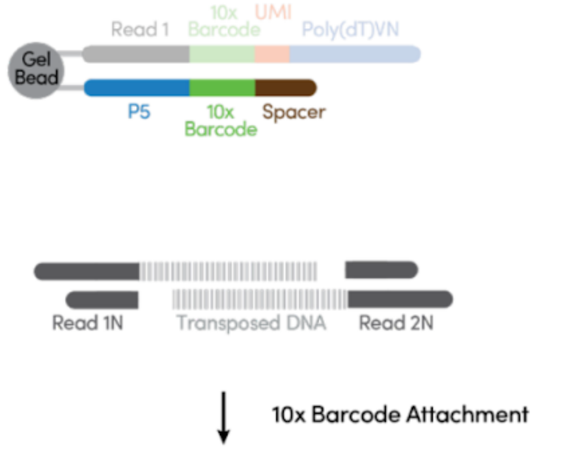
Single-cell ATAC Seq

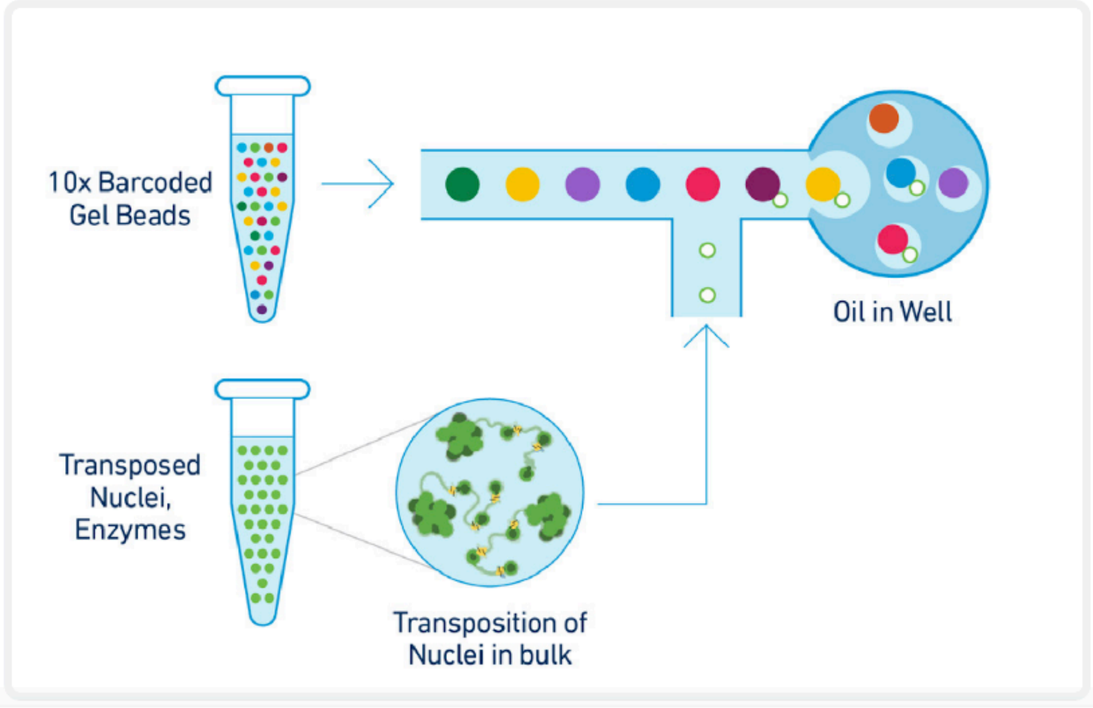


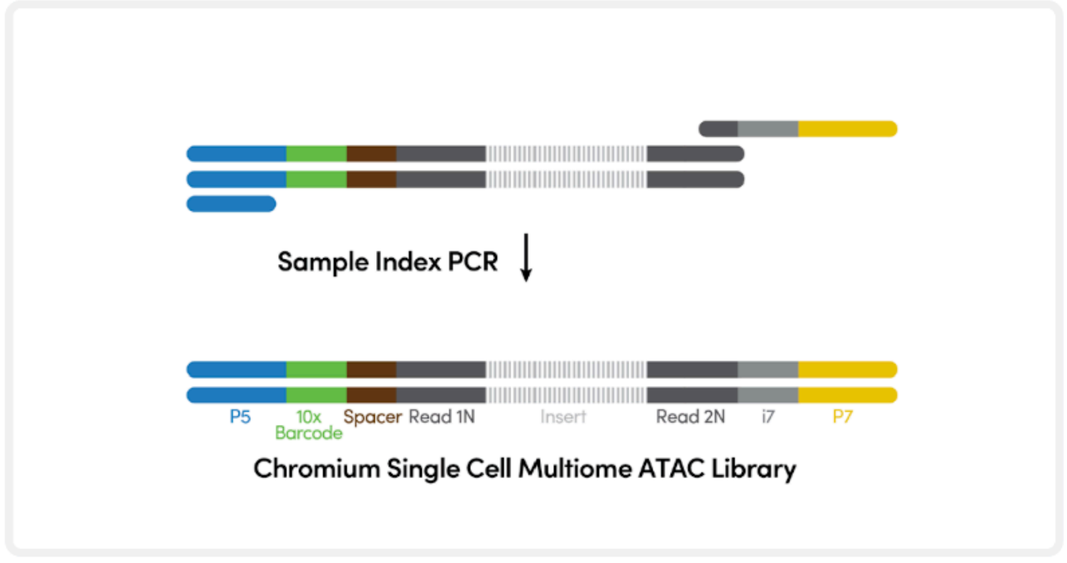


Single-cell ATAC Seq

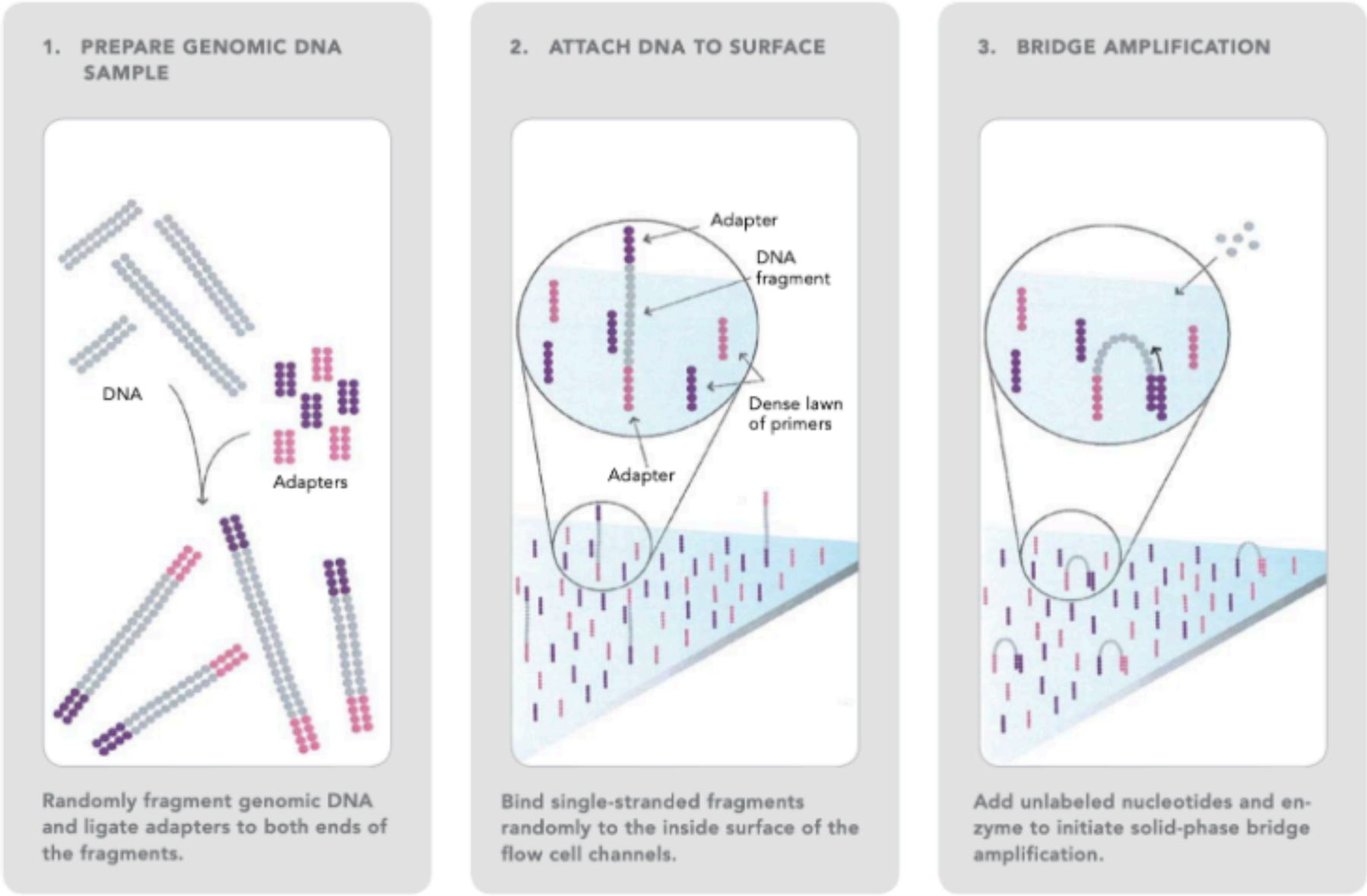






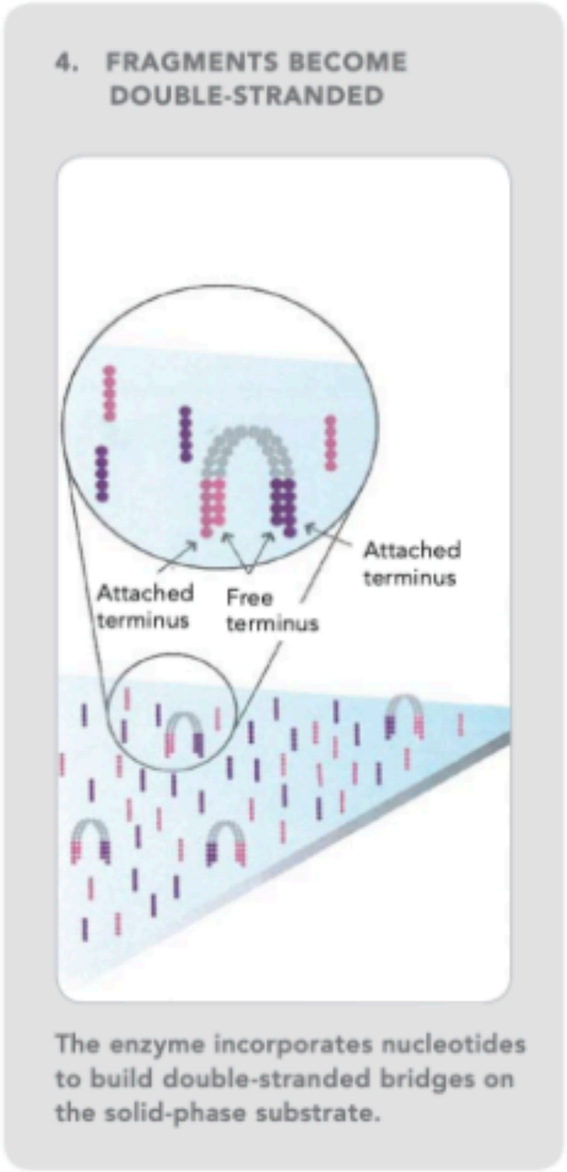


How does next generation sequencing work?



https://www.futurelearn.com/info/courses/a-practical-guide-for-sars-cov-2-whole-genome-sequencing/0/steps/338178

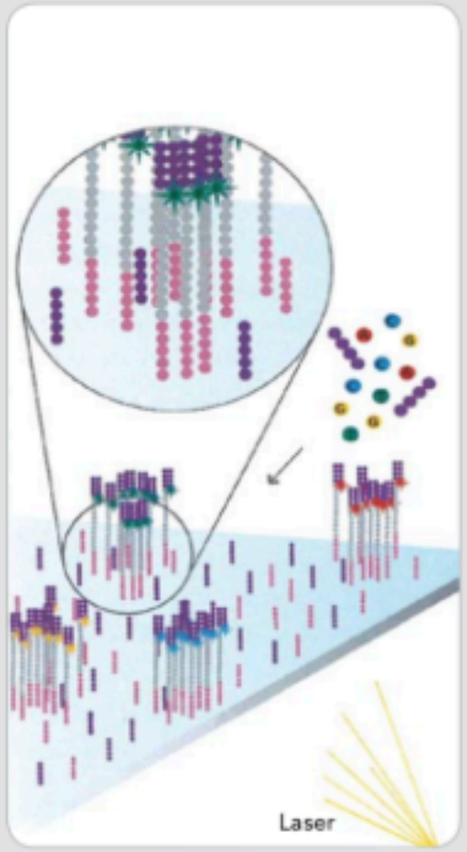
How does next generation sequencing work?



https://www.futurelearn.com/info/courses/a-practical-guide-for-sars-cov-2-whole-genome-sequencing/0/steps/338178

How does next generation sequencing work?

7. DETERMINE FIRST BASE

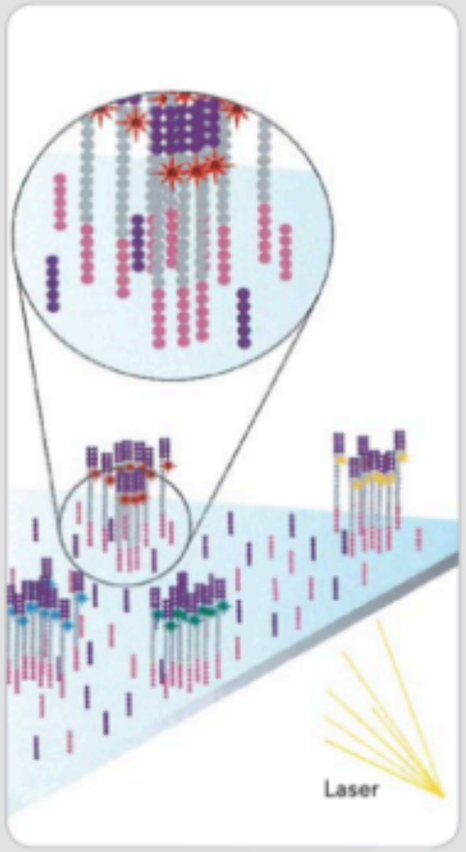


The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

8. IMAGE FIRST BASE

After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

9. DETERMINE SECOND BASE

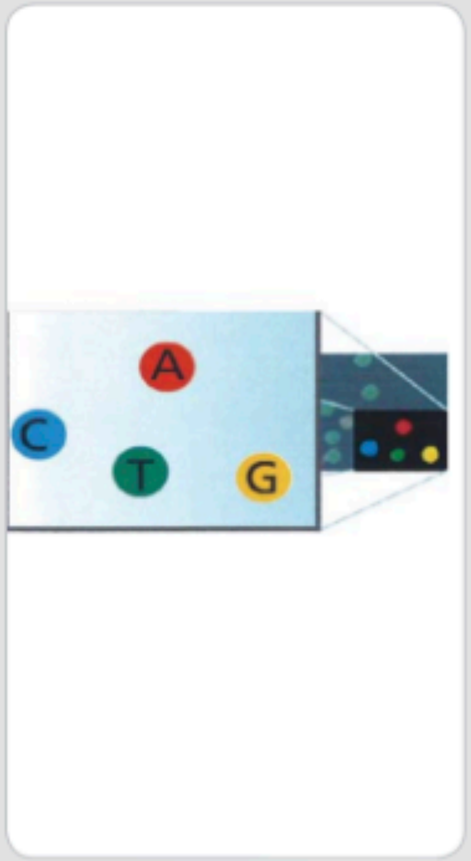


The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

https://www.futurelearn.com/info/courses/a-practical-guide-for-sars-cov-2-whole-genome-sequencing/0/steps/338178

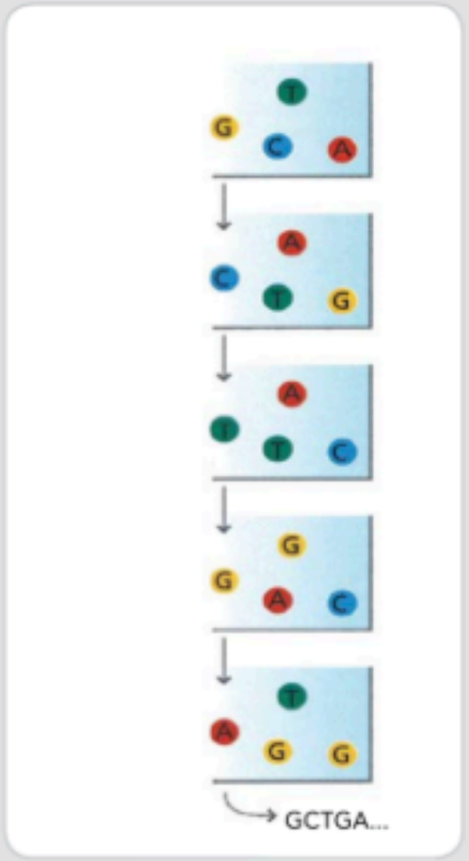
How does next generation sequencing work?

10. IMAGE SECOND CHEMISTRY CYCLE



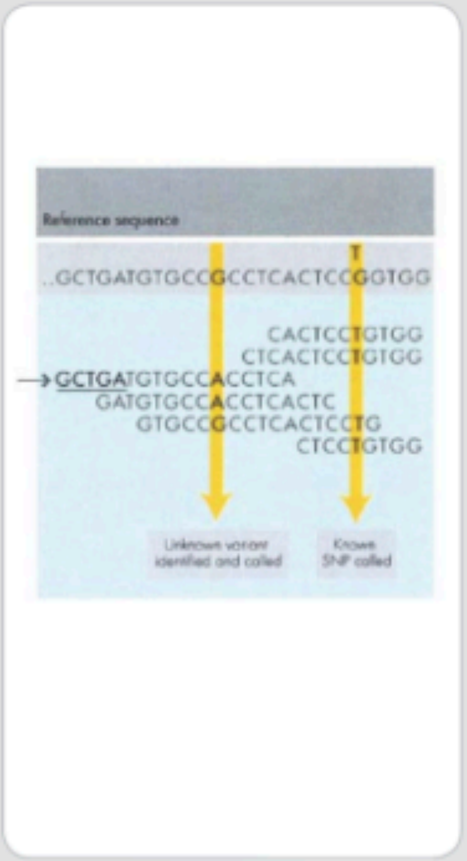
After laser excitation, the image is captured as before, and the identity of the second base is recorded.

11. SEQUENCING OVER MULTIPLE CHEMISTRY CYCLES



The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

12. ALIGN DATA



The data are aligned and compared to a reference, and sequencing differences are identified.

https://www.futurelearn.com/info/courses/a-practical-guide-for-sars-cov-2-whole-genome-sequencing/0/steps/338178

Sequencing data has to be mapped against the reference genome

EPFL

What is a .fastq file?

- Text file containing (short) nucleotide sequences (reads)

First 12 lines of a *.fastq* file

4 lines = 1 read

```
@SRR038845.3 HWI-EAS038:6:1:0:193
CAACGAGTTCACACCTTGGCCGACAGGCCCGGG
+SRR038845.3 HWI-EAS038:6:1:0:193
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8
@SRR038845.41 HWI-EAS038:6:1:0:14
CCAATGATTTTTTTCCGTGTTTCAGAATACGGT
+SRR038845.41 HWI-EAS038:6:1:0:14
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:
@SRR038845.53 HWI-EAS038:6:1:1:36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAA
+SRR038845.53 HWI-EAS038:6:1:1:36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBB
```

BIOENG-420 SINGLE-CELL BIOLOGY

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A

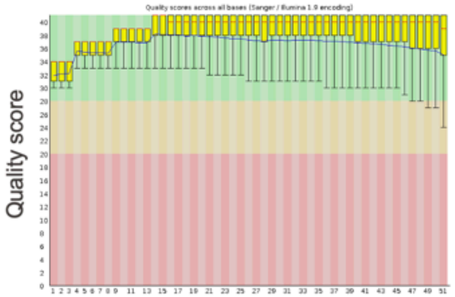
BIOENG-420 SINGLE-CELL BIOLOGY

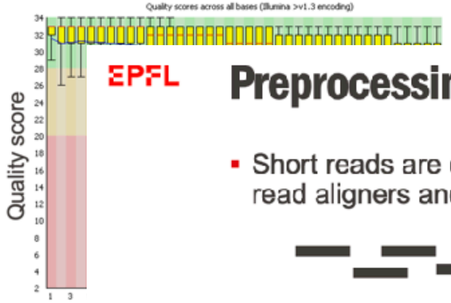
Vincent Gardeux 13

EPFL

Preprocessing: Sequencing quality

- The raw .fastq file is checked for:
  - Number of reads
  - Read length
  - Base quality & distribution
  - Duplication level
  - Overrepresented sequences
  - GC content
  - etc.





GOOD: Process without trimming

- Quality typically decreases towards the
  - Remove or trim reads with low quality scores be

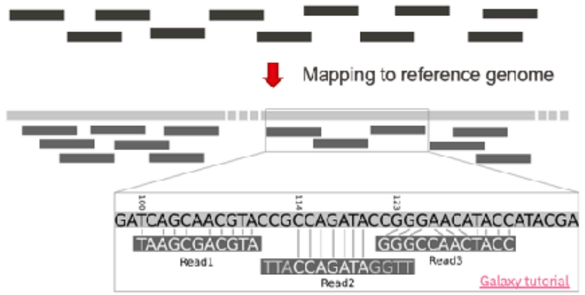
BIOENG-420 SINGLE-CELL BIOLOGY

Vincent Gardeux 14

EPFL

Preprocessing: Mapping to genome/transcriptome

- Short reads are efficiently mapped to the genome or transcriptome using short read aligners and known genomic annotations (e.g. Ensembl or UCSC)



- Efficient alignment with specialized tools
  - Large number of aligners available
    (>70 on Wikipedia 20.02.2022)
  - Considerations: accuracy, gap-aware, speed, memory, …

Tools:

- *BWA*
- *Kallisto*
- *STAR*
  - *Now **STARsolo** is specifically designed for single-cell and multiplexed exp.*

BIOENG-420 SINGLE-CELL BIOLOGY

Vincent Gardeux 15

Mapping ATAC Data:
BWA
BOWTIE2
Taken care of by CellRanger

What analysis packages can you use?

nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-021-00790-6

Check for updates

OPEN

ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis

Jeffrey M. Granja[1,2,3,12] ✉, M. Ryan Corces[3,4,5,6,12], Sarah E. Pierce[1,7], S. Tansu Bagdatli[1], Hani Choudhry[8], Howard Y. Chang[1,3,9] ✉ and William J. Greenleaf[1,3,10,11] ✉

nature methods

ARTICLES

https://doi.org/10.1038/s41592-021-01282-5

Check for updates

Single-cell chromatin state analysis with Signac

Tim Stuart[1,2] ✉, Avi Srivastava[1,2], Shaista Madad[1,2], Caleb A. Lareau[3] and Rahul Satija[1,2] ✉

Wolf *et al. Genome Biology* (2018) 19:15
https://doi.org/10.1186/s13059-017-1382-0

Genome Biology

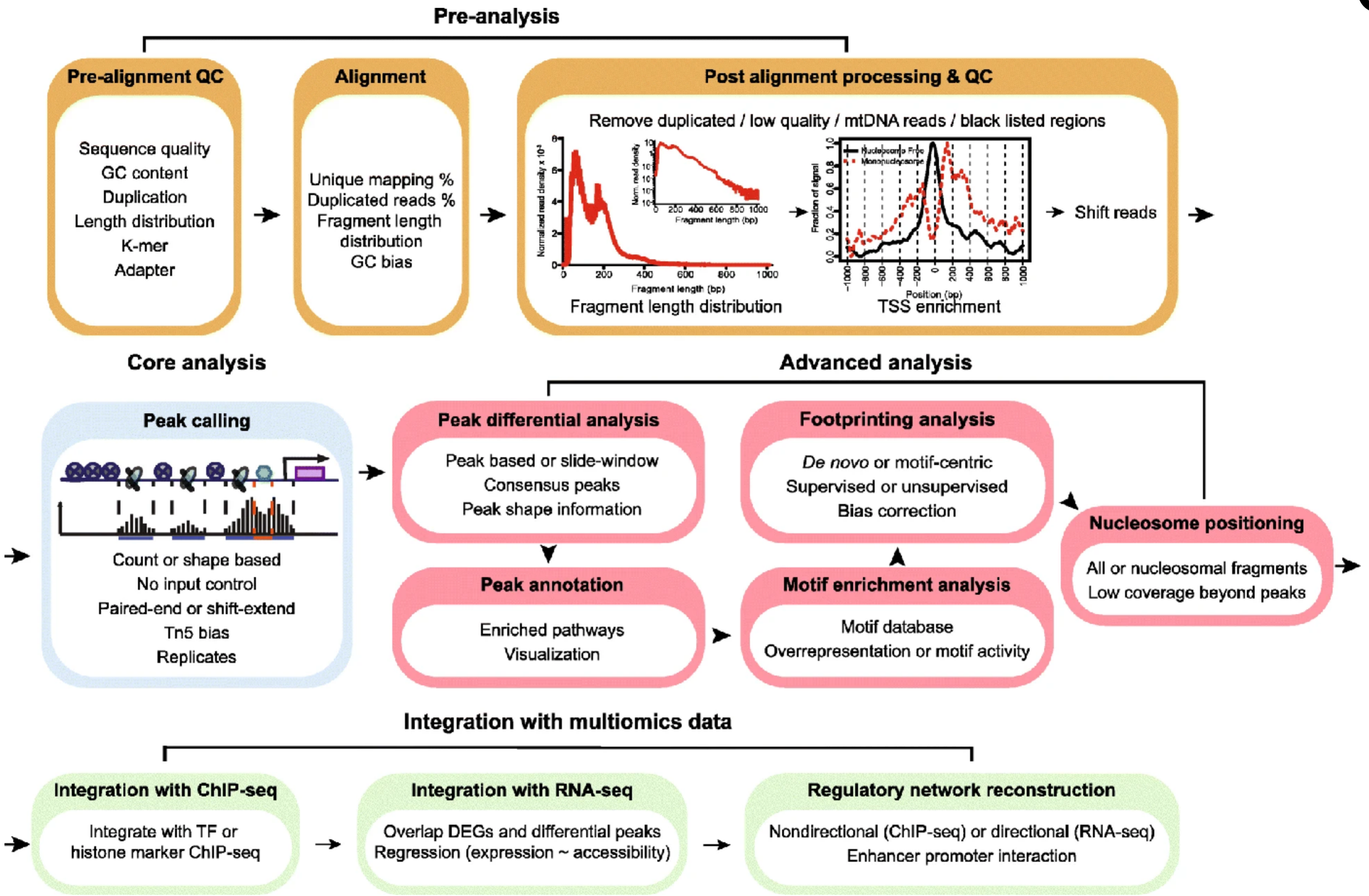
SOFTWARE Open Access

CrossMark

SCANPY: large-scale single-cell gene expression data analysis

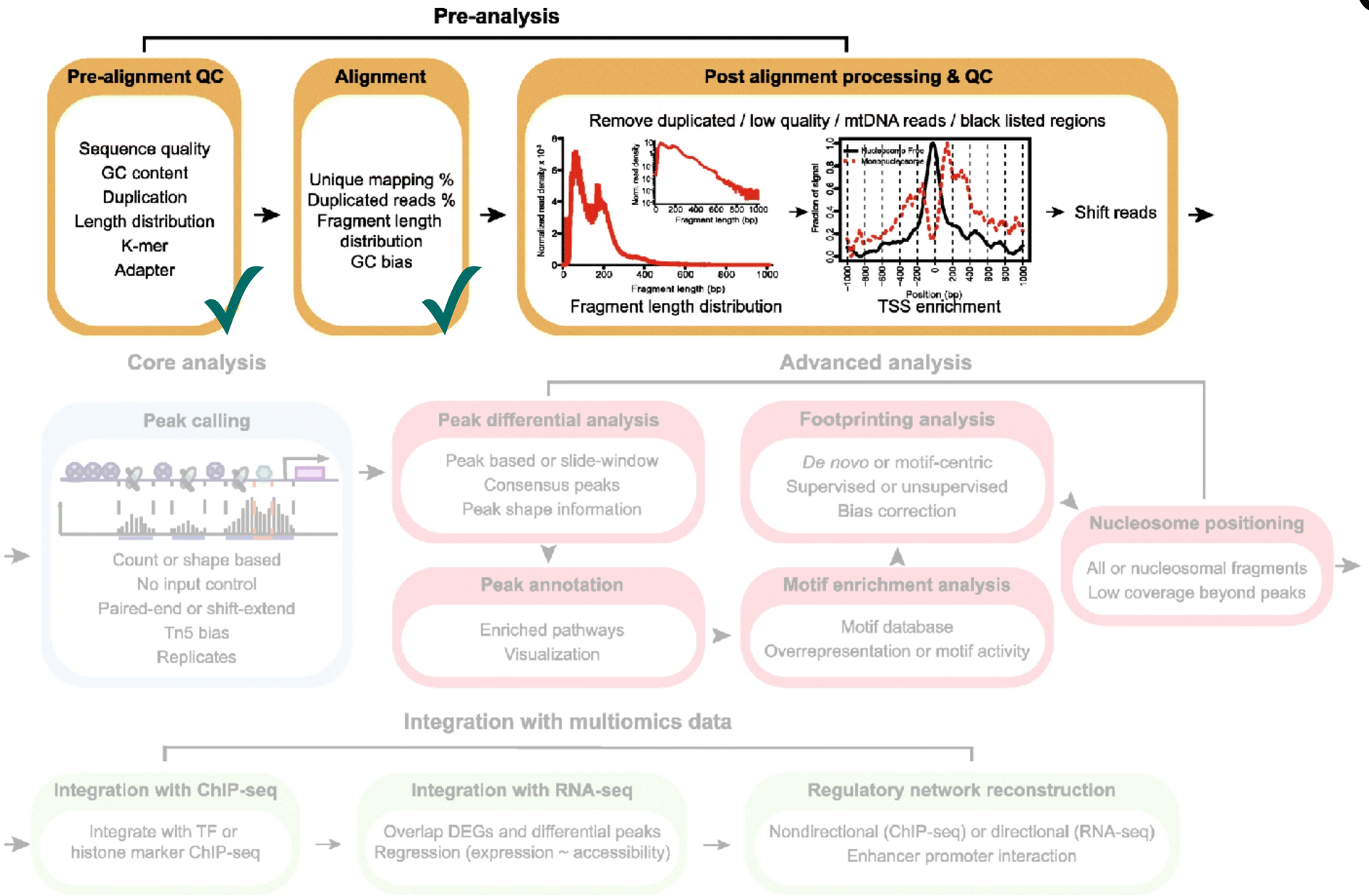
F. Alexander Wolf[1*], Philipp Angerer[1] and Fabian J. Theis[1,2*]

QC of ATAC data



Yan et al. (2020) Genome Biology

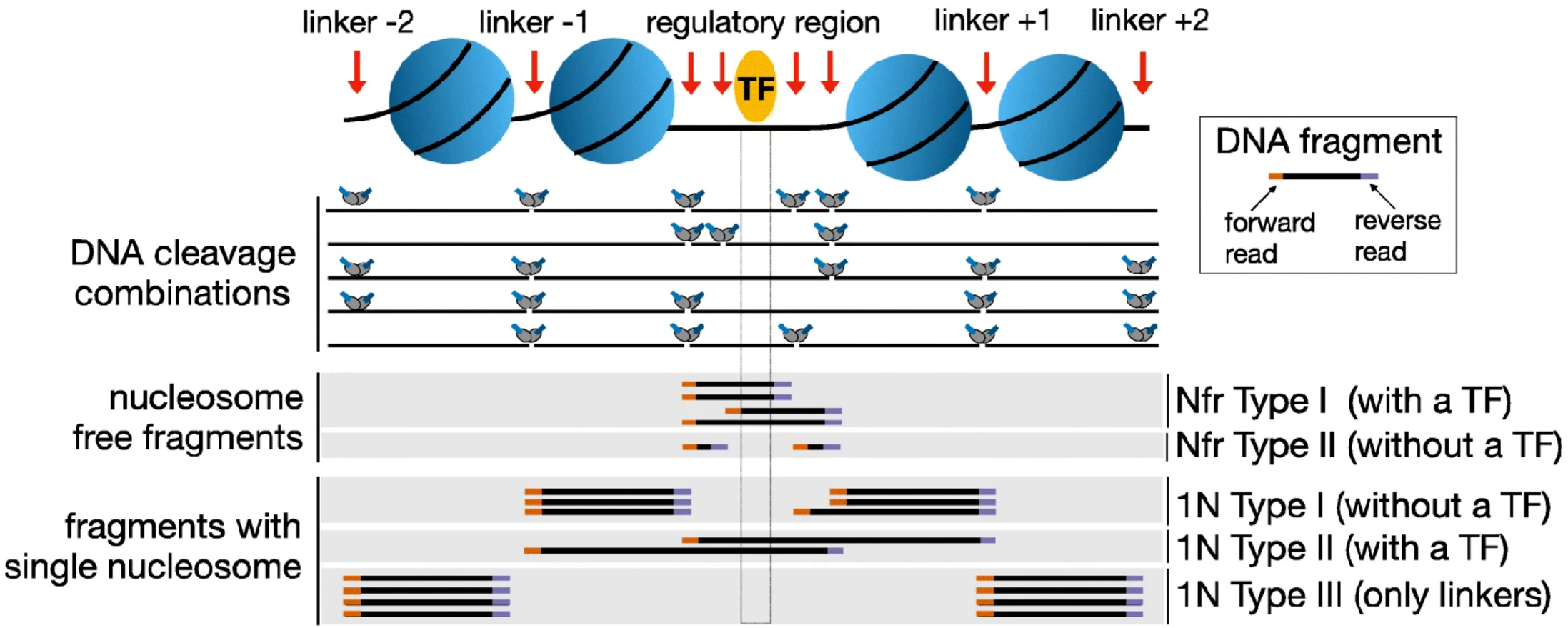
QC of ATAC data



Yan et al. (2020) Genome Biology

QC of ATAC data

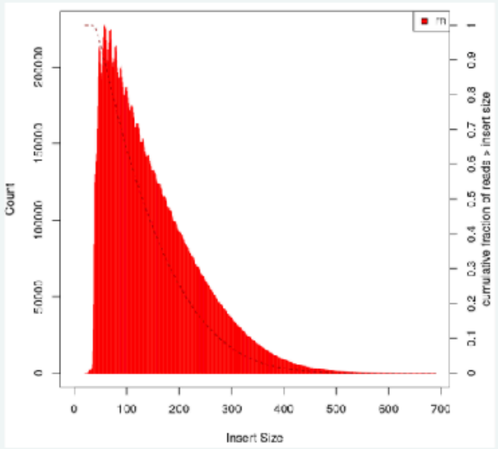
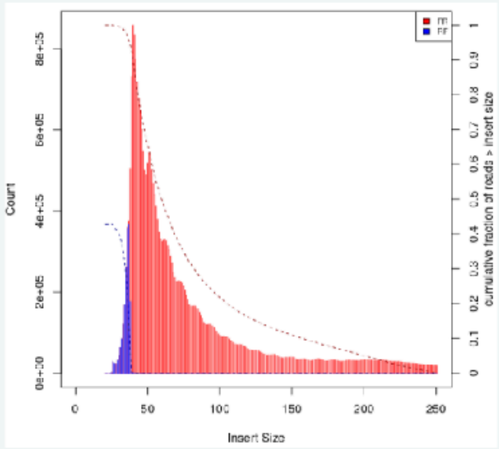
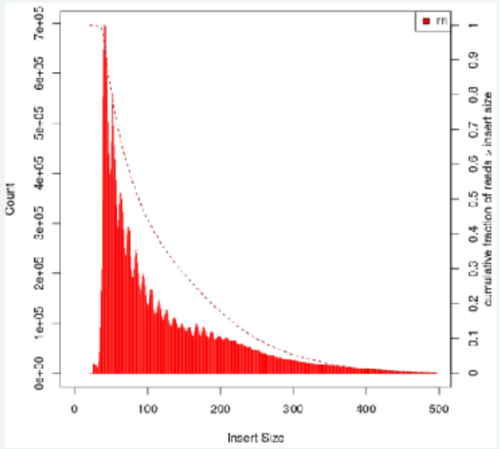
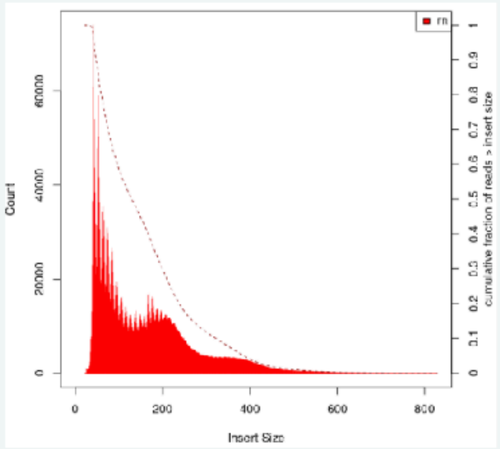
Digesting Chromatin with Tn5 results in very regular DNA insert sizes, this can be used as an indicator of library quality



QC of ATAC data

With paired-end sequencing data the insert sizes of an ATAC library can be checked.

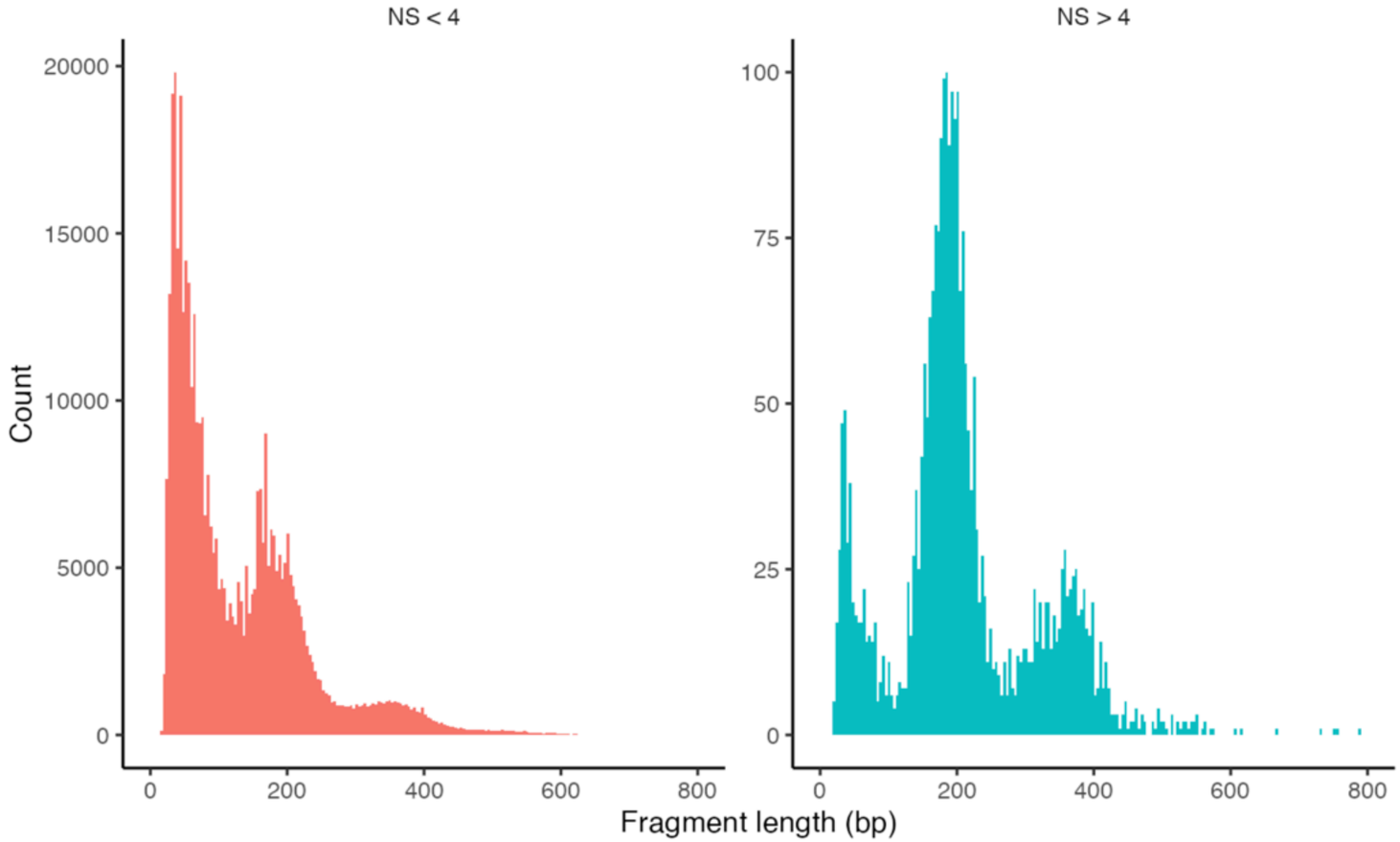
Which library looks good and how can you explain the pattern?

Naked DNA	Failed ATAC-seq	Noisy ATAC-seq	Successful ATAC-seq
			

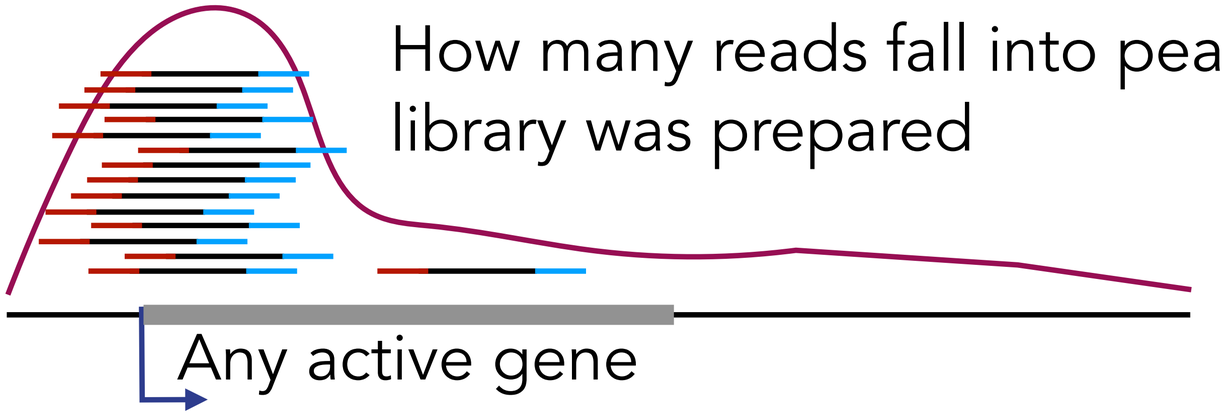
QC of ATAC data - Nucleosome signal

The nucleosome signal reflects the ratio between mononucleosomes and nucleosome-free regions

```
pbmc$nucleosome_group <- ifelse(pbmc$nucleosome_signal > 4, 'NS > 4', 'NS < 4')
FragmentHistogram(object = pbmc, group.by = 'nucleosome_group')
```



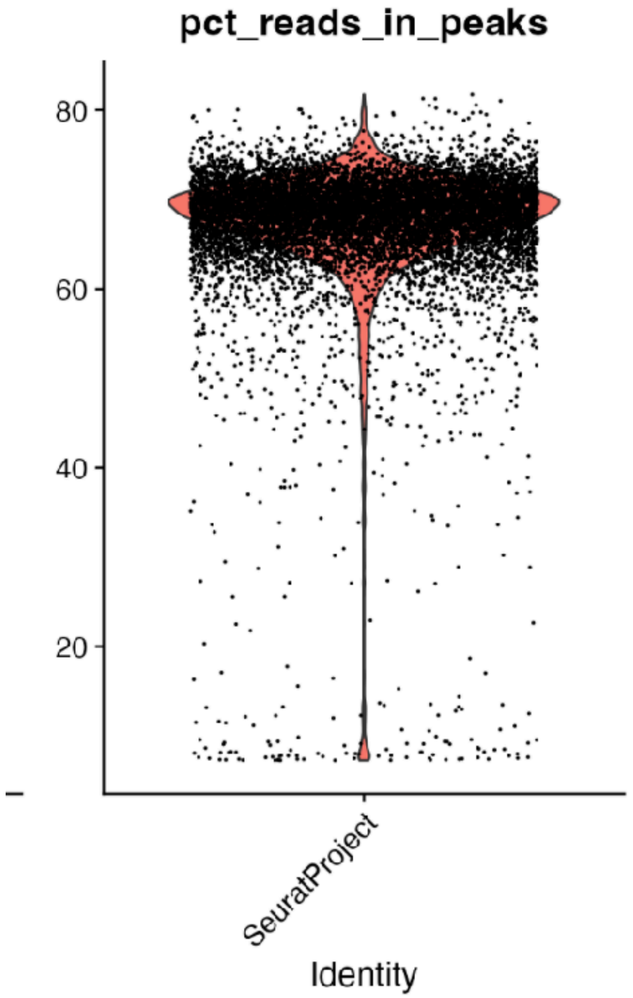
QC of ATAC data - Fraction of Fragments in peaks (FrIP)



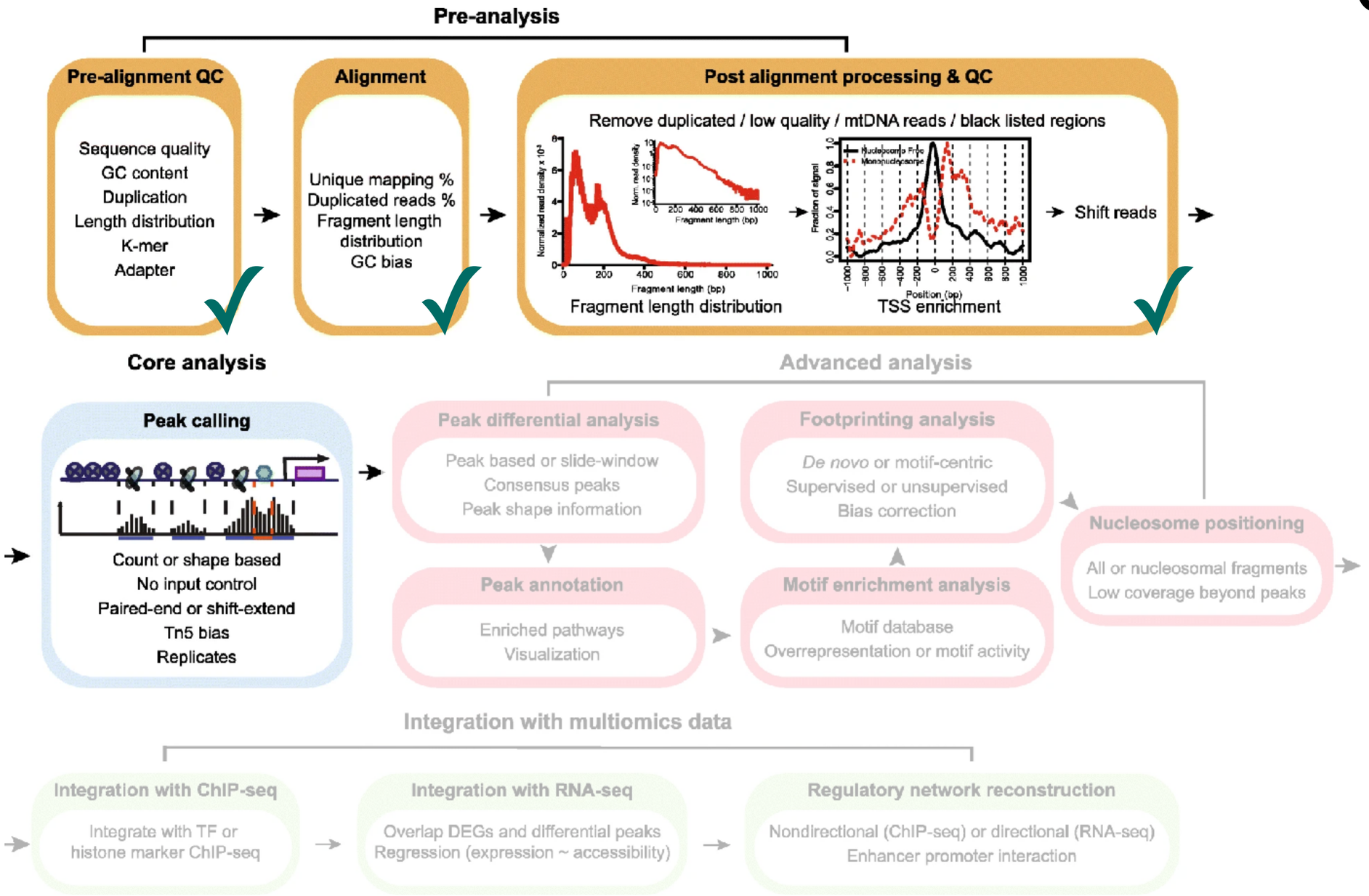
How many reads fall into peaks is a good indicator of how carefully the library was prepared

```
Signac code:

pbmc$pct_reads_in_peaks <-
pbmc$peak_region_fragments /
pbmc$passed_filters * 100
```

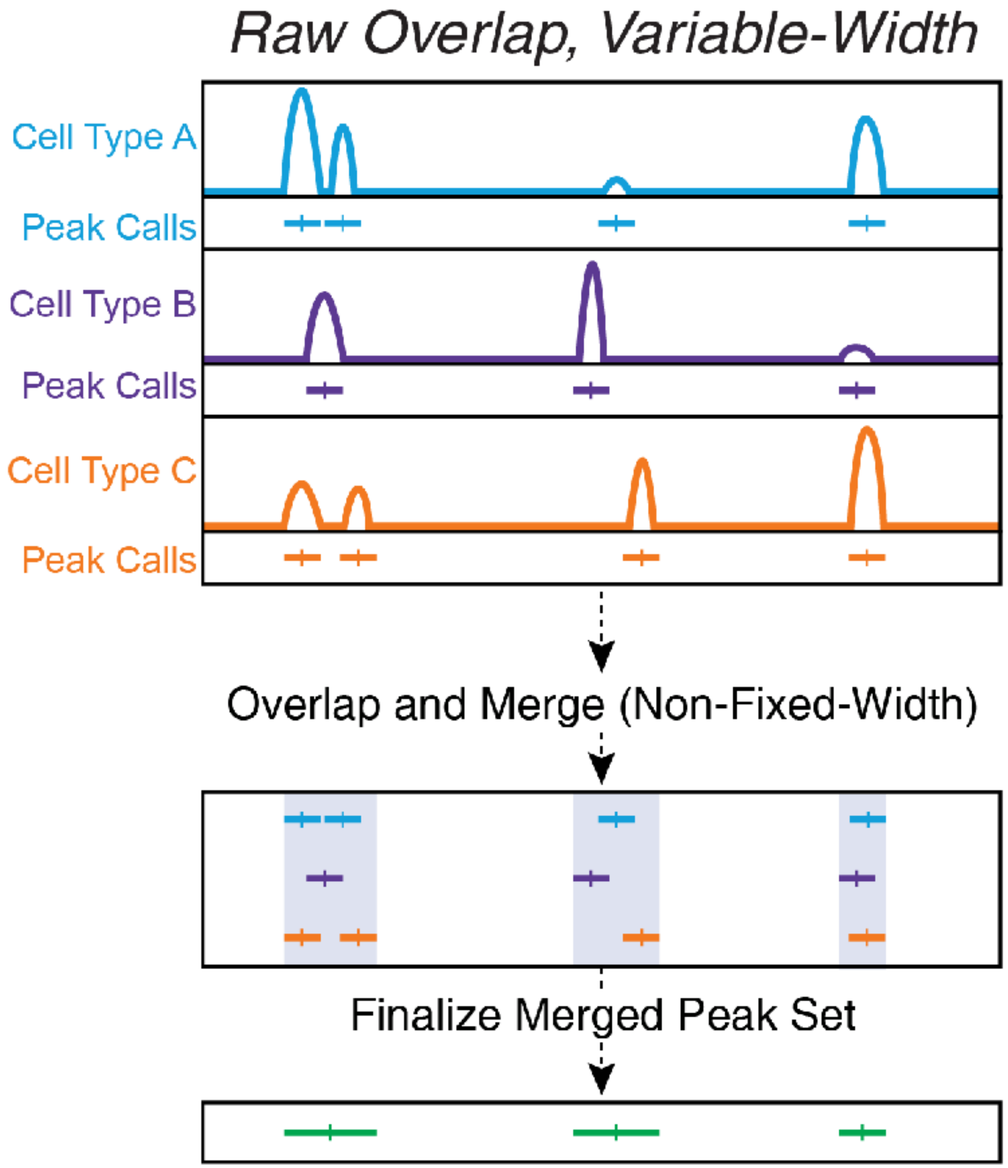


QC of ATAC data



Yan et al. (2020) Genome Biology

How to identify peaks?



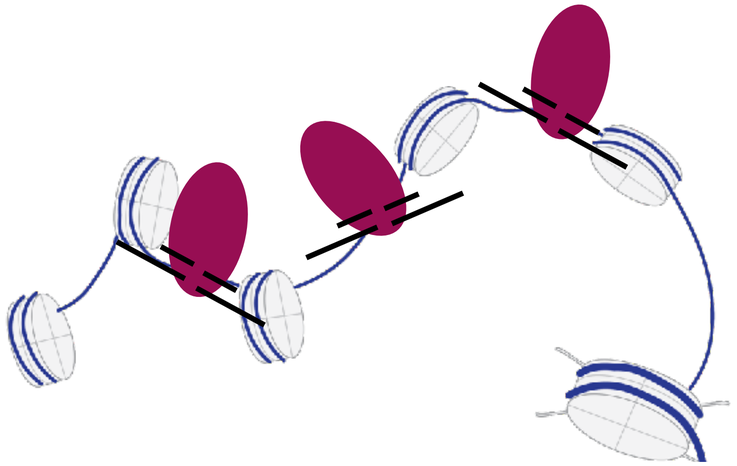
https://www.archrproject.com/

There are various algorithms to call peaks.
The most commonly used in MACS2 (now 3).
Several options for broad and narrow peaks.

CellRanger from 10x Genomics uses a proprietary peak caller.

Bedtools can merge peaks

TF-IDF Normalization



For ATAC sequencing we can only obtain two fragments per locus per cell (lower dynamic range compared to RNA exp)

Signac performs Term Frequency-Inverse Document Frequency (TF-IDF) normalization. This is a two-step normalization procedure:

Term frequency (TF) normalizes for sequencing depth by scaling peak accessibility counts within each cell

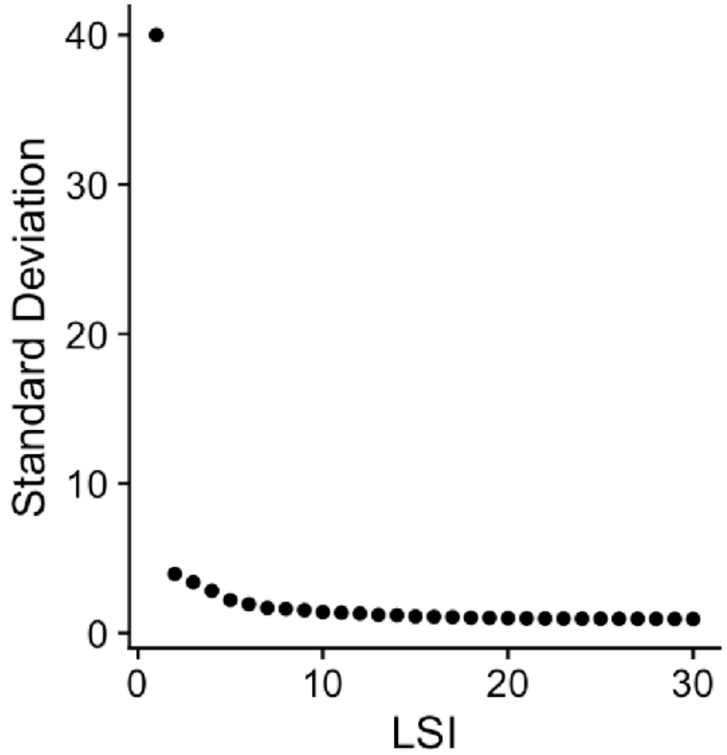
Inverse Document Frequency (IDF) down-weights these frequently accessible regions (i.e, peaks that are open across many cells but might not be biologically informative), ensuring that rare but cell-type-specific peaks get more importance.

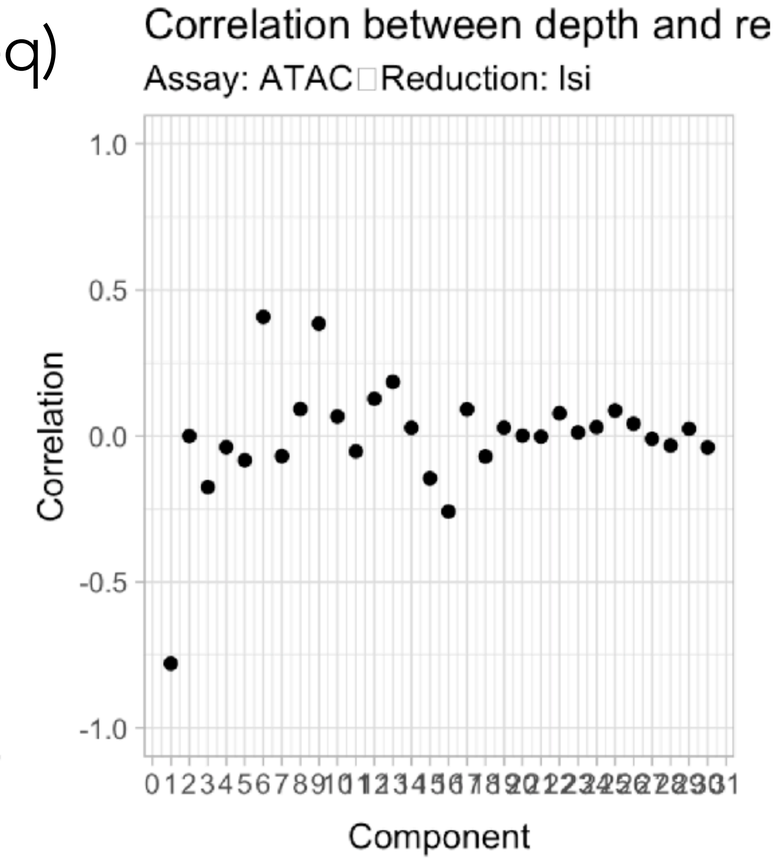
Cusanovich et al. (2015) Science

Dimensionality reduction

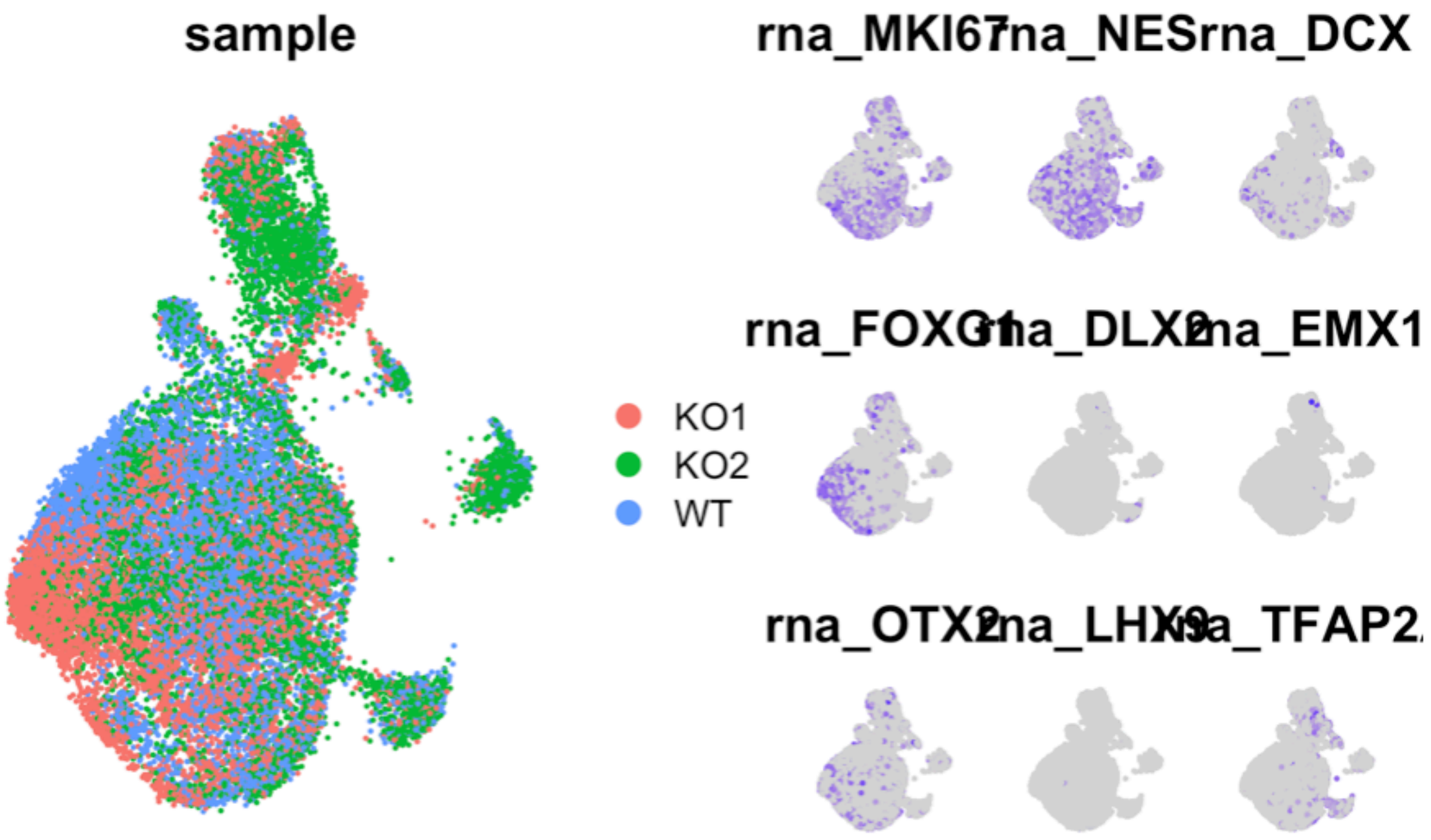
Perform Singular Value Decomposition (SVD) on the TD-IDF matrix, which gives Latent Semantic Indexing (LSI) components. This is very similar to PCA (but better for sparse data). The first singular vector often captures sequencing depth (technical variation) rather than biological variation.

Elbow plot (like for scRNA-seq)





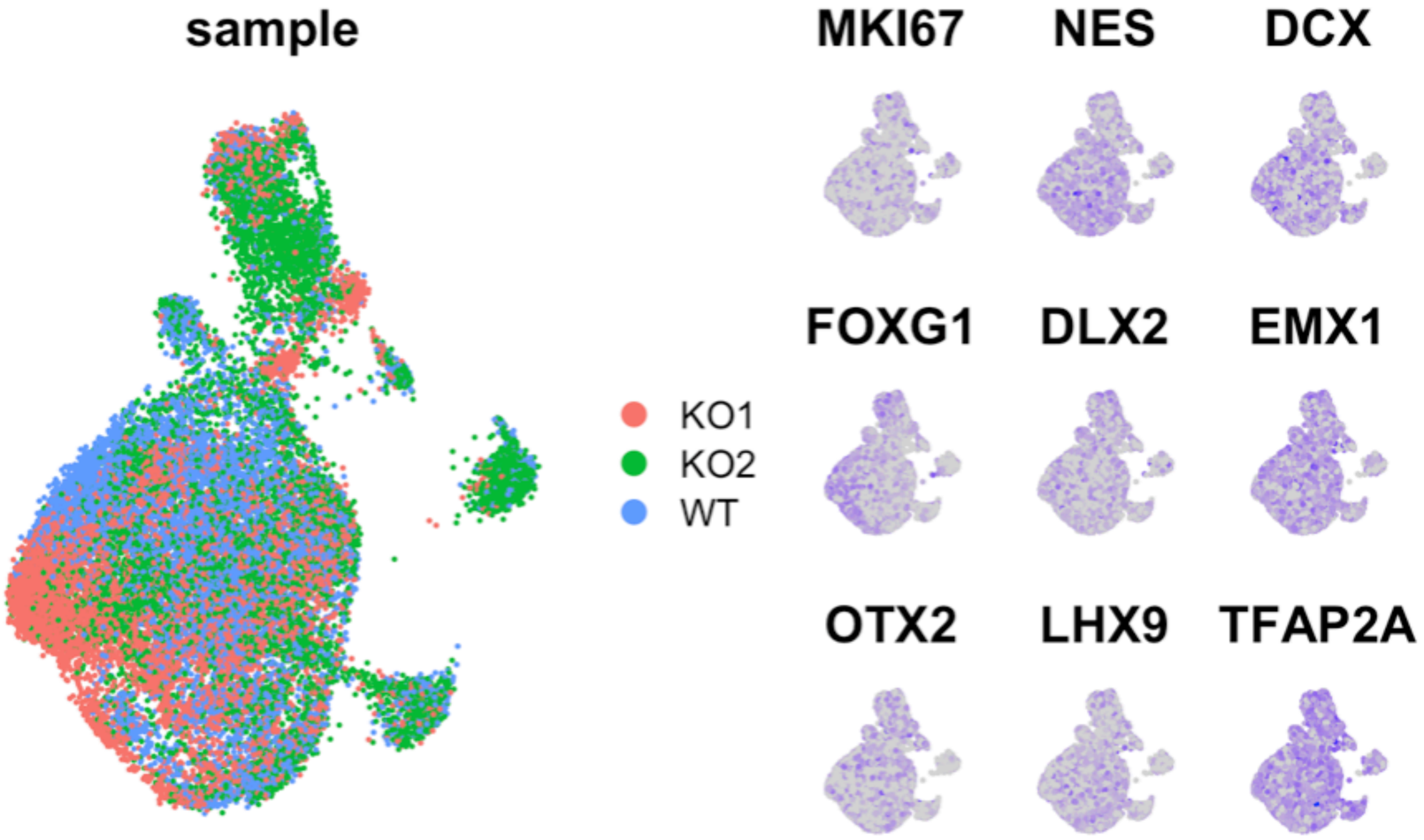
Dimensionality reduction



Run UMAP on the LSI to obtain the embedding

In this special case we have linked RNA-seq data for each cell and we can use it to check the embedding

Annotating scATAC-data - obtaining “gene activities”

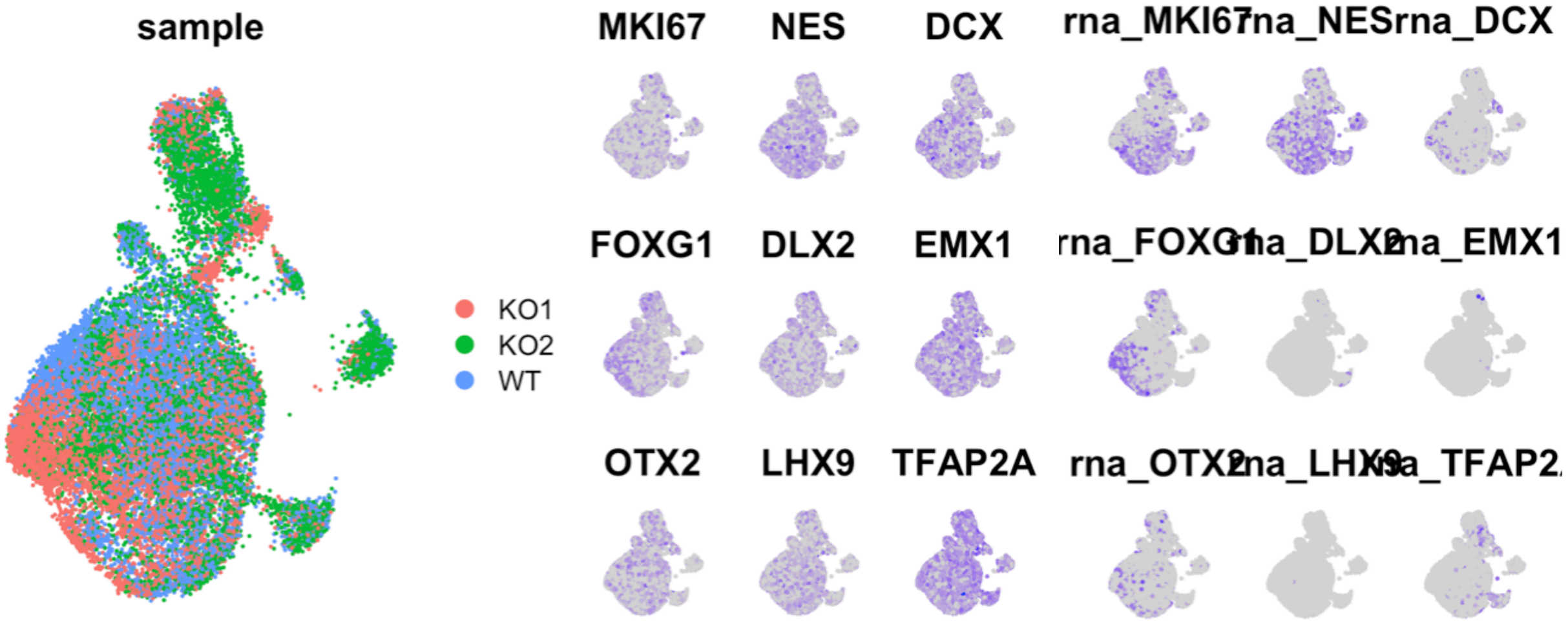


Signac uses gene activities to link detected fragments to genes.

It aggregates all fragments on the gene body.

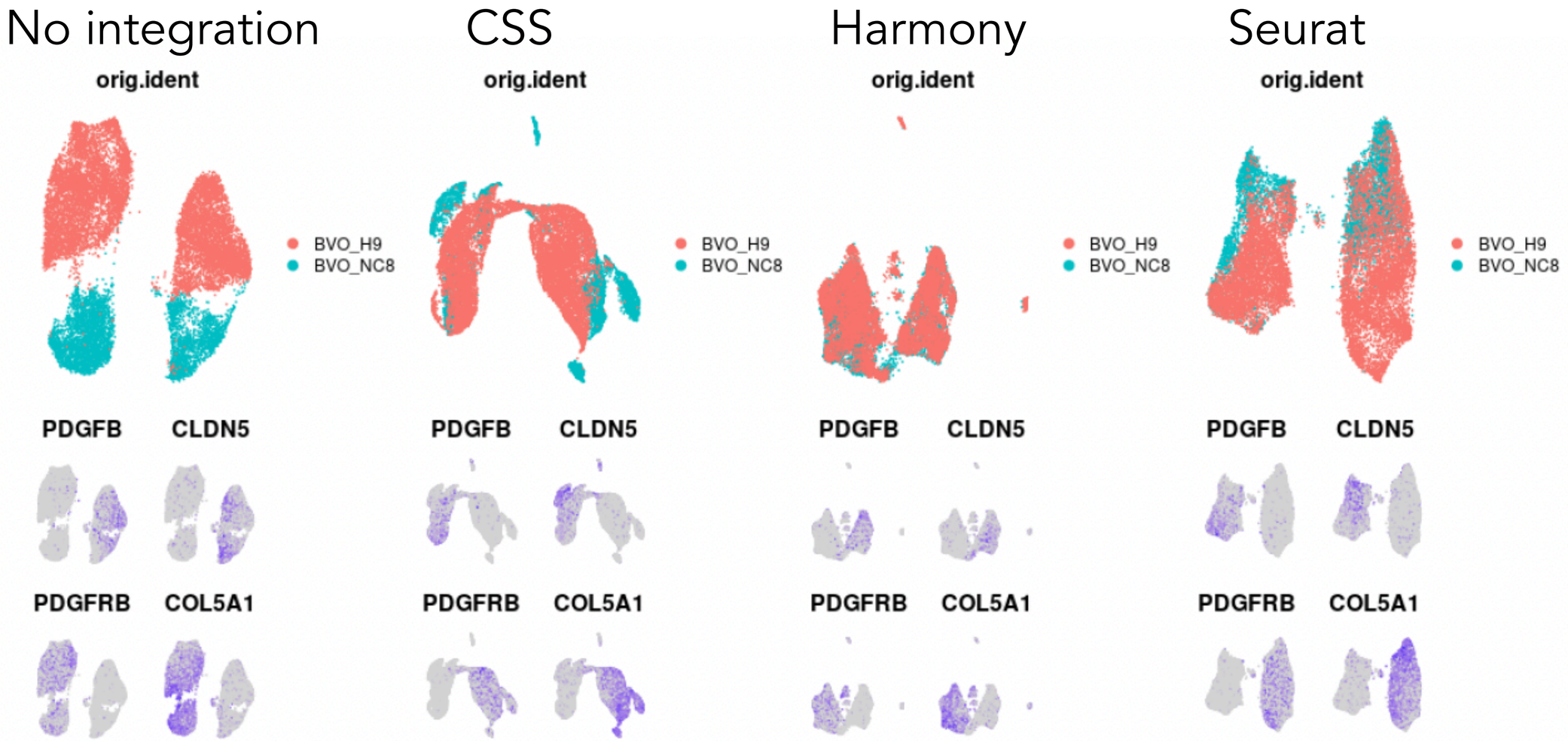
This can be visualized as a proxy of gene expression.

Annotating scATAC-data - obtaining "gene activities"



Comparing chromatin accessibility and RNA expression

Data integration



Similar integration tools as for RNA can be used - we have to perform the integration on the LSI (not PCA as default for RNA)

Integration for multi-omic measurement (more than one modality per cell)

In the example RNA and ATAC have been measures together from the same cell

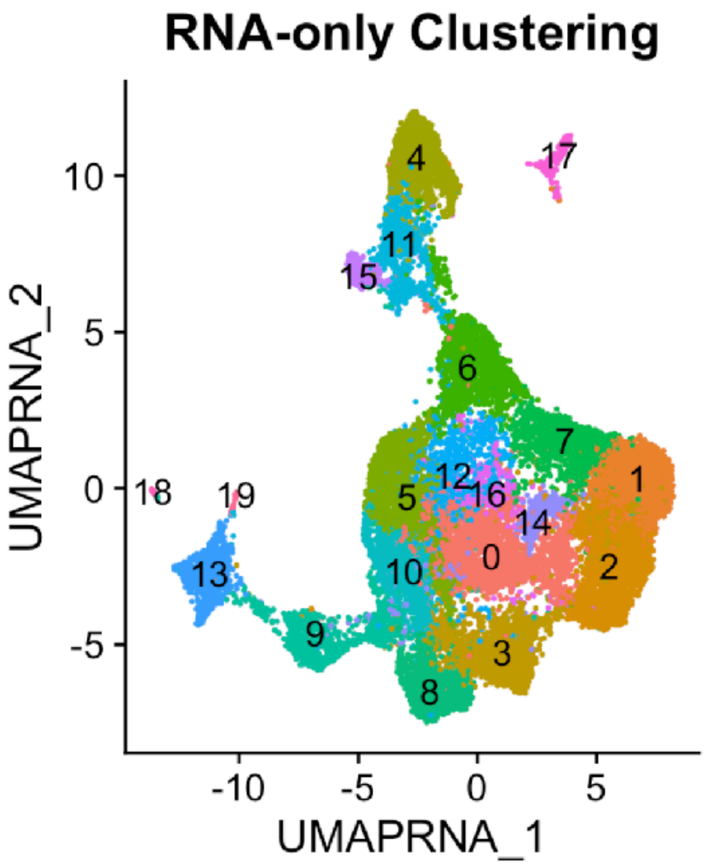
Seurat uses WNN (weighted nearest neighbour network) - this calculates the k-nearest neighbour of each cell taking both modalities into account.

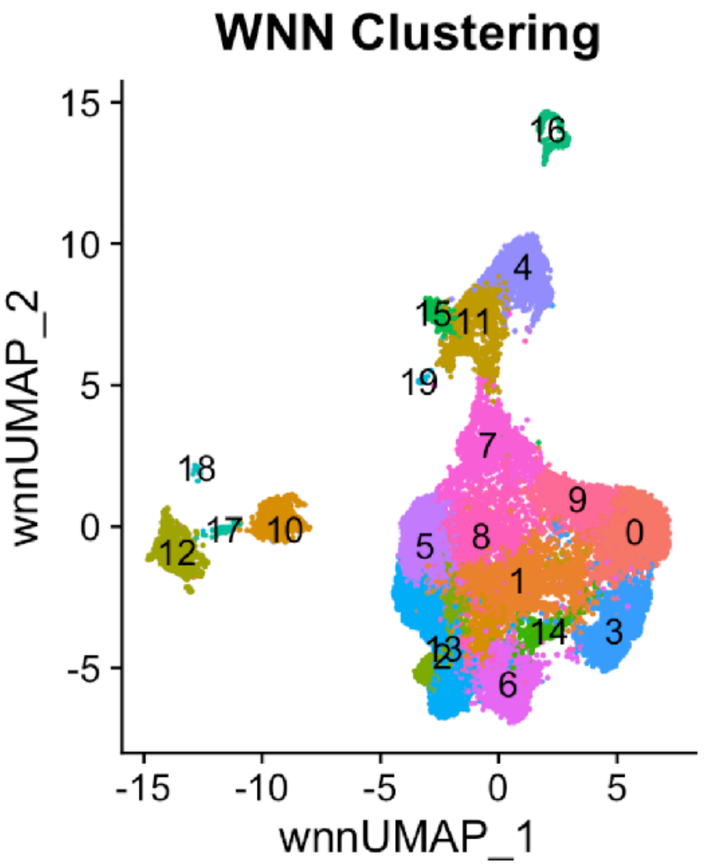
The weighted nearest neighbour network is used for embedding and clustering.

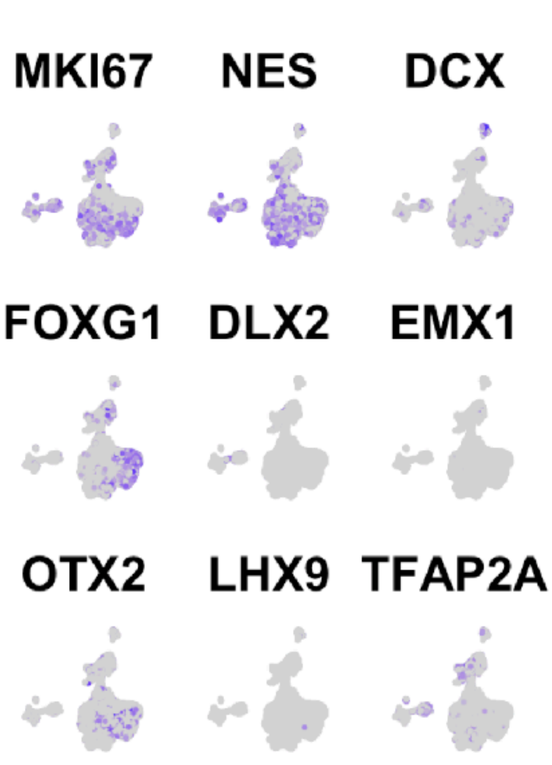
Multimodal integration does not always result in better cluster discrimination.

Hao et al. 2021, Cell

Integration for multi-omic measurement (more than one modality per cell)







In this case slight refinements of the clustering in the progenitors (NES positive)

Cell type annotation

Differential Expression (DE) analysis

Vincent Gardeux

- **Goal:** identification of genes (or transcripts, exons, …) that are expressed in significant different quantities in distinct groups

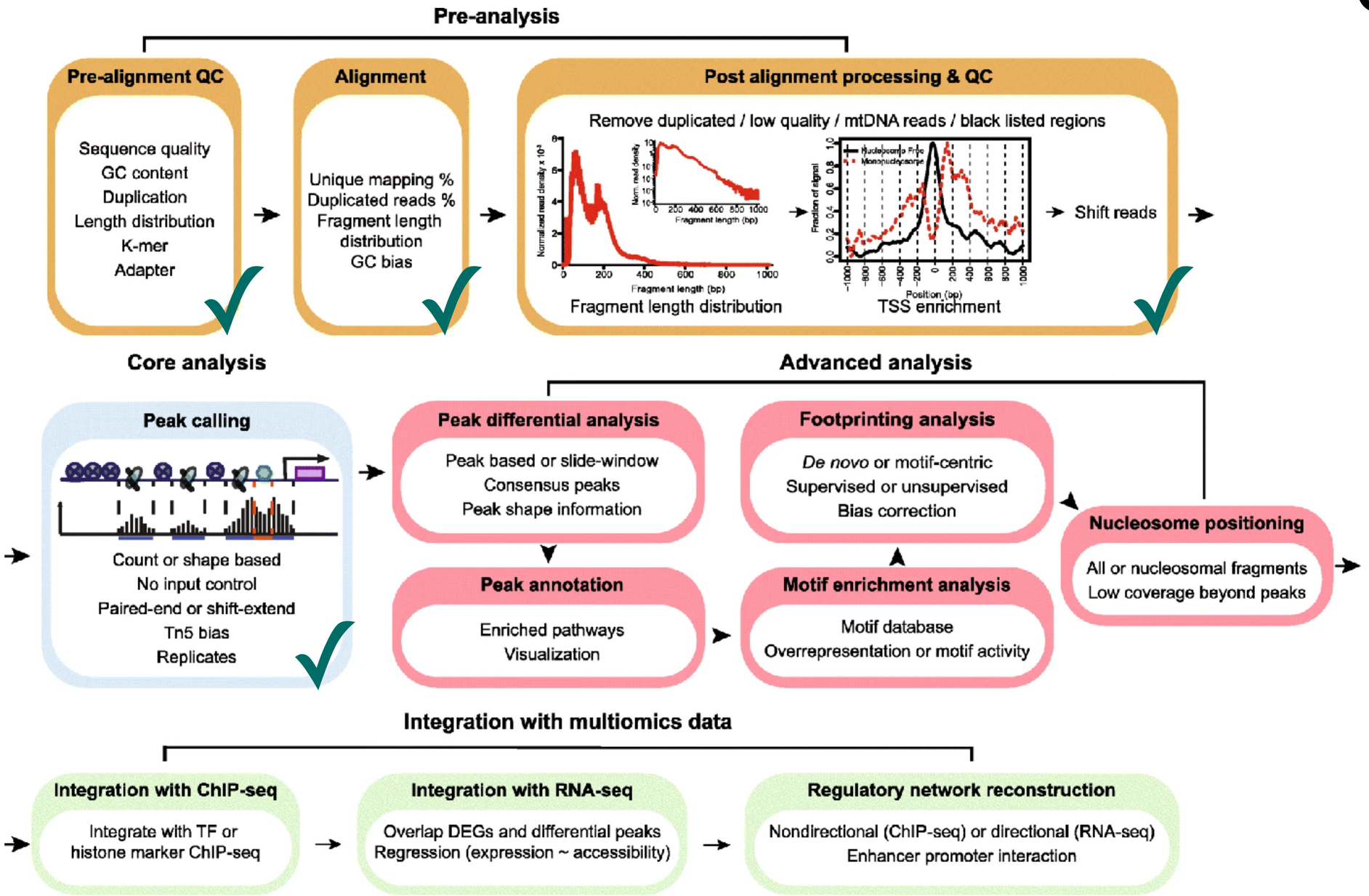
⇒e.g. drug-treated vs control, disease vs healthy, **cell-types**, tissues, development stages, ...

- What method to use for doing that?
  - Mostly statistical tests

- Why not simply using fold-change (FC)?
  - Many FP or FN would be expected, because does not take into consideration:
    - Low expressed genes that will tend to have higher FC (vs High)
    - The distribution of the data (inherent variance of gene expression)

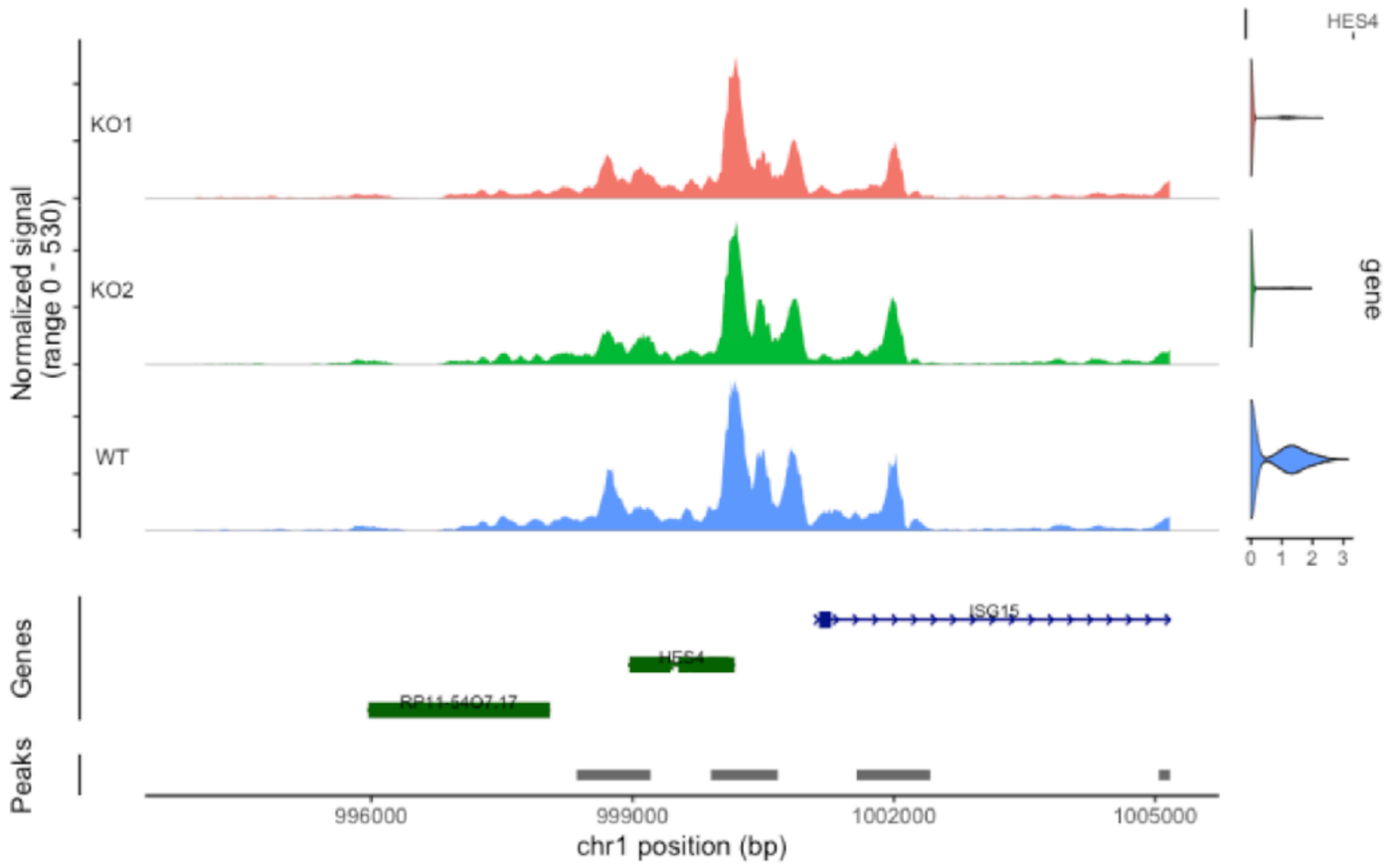
Really similar to what we do for scRNA-seq

QC of ATAC data



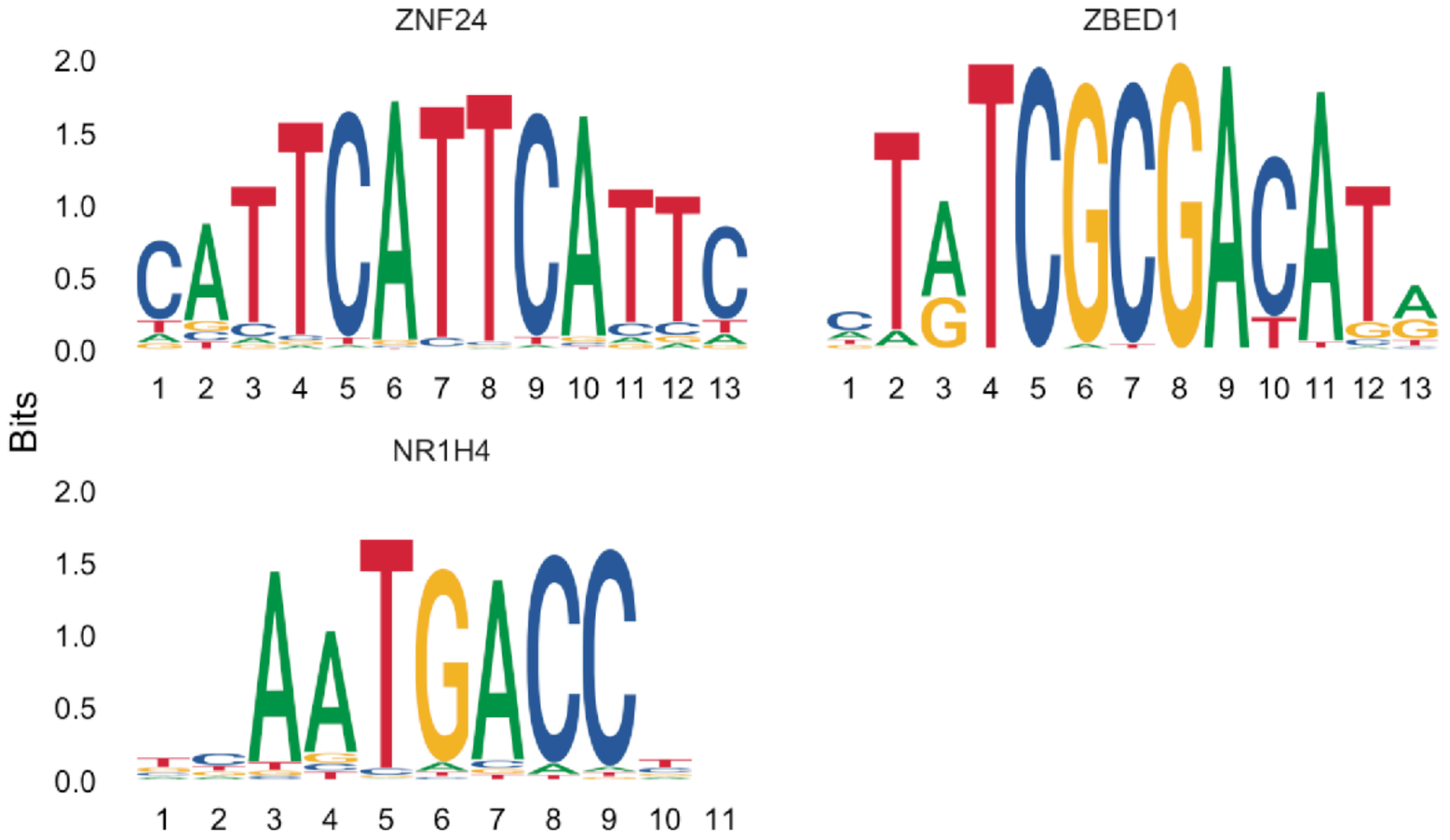
Yan et al. (2020) Genome Biology

What can we learn in addition from chromatin accessibility?



We can associate regulatory elements to expressed genes and use this to infer regulatory networks

We can identify transcription factor motifs that change upon perturbation



Motifs that are enriched in differentially accessible peaks.

Tools:
JASPAR
TFBSTools

Summary and Take Home

Knowing different types of chromatin

How do identify/map accessible chromatin

Isolating accessible chromatin in single cells

Sequencing libraries using NGS sequencing

Visualizing chromatin accessibility in the genome browser

How to analyze and interpret single cell ATAC data - hands on tutorial will follow