Swiss Institute of Bioinformatics

BIOLOGY-INFORMED INTEGRATION AND VISUALIZATION OF MULTIOMICS DATA
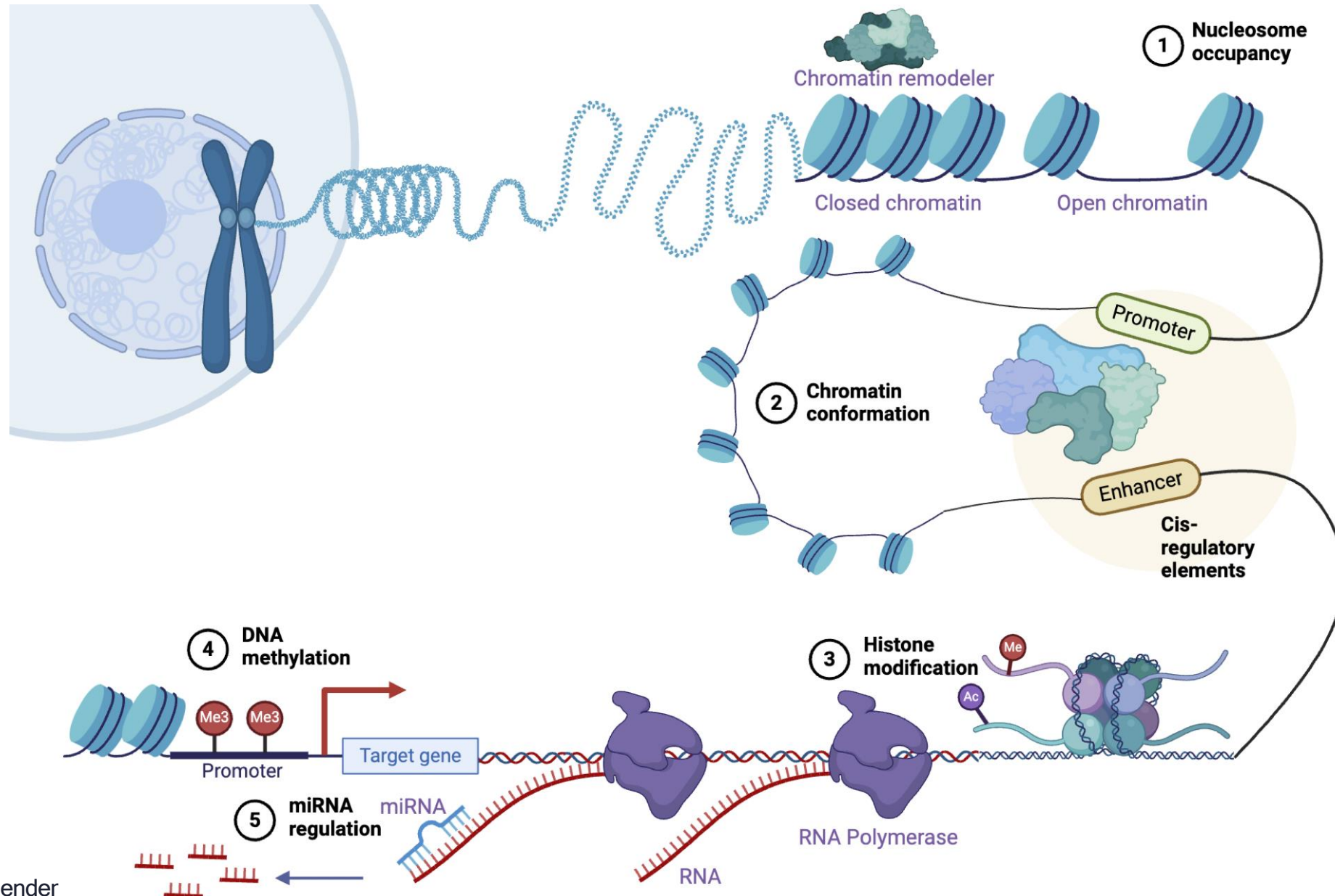
Multi-omics data
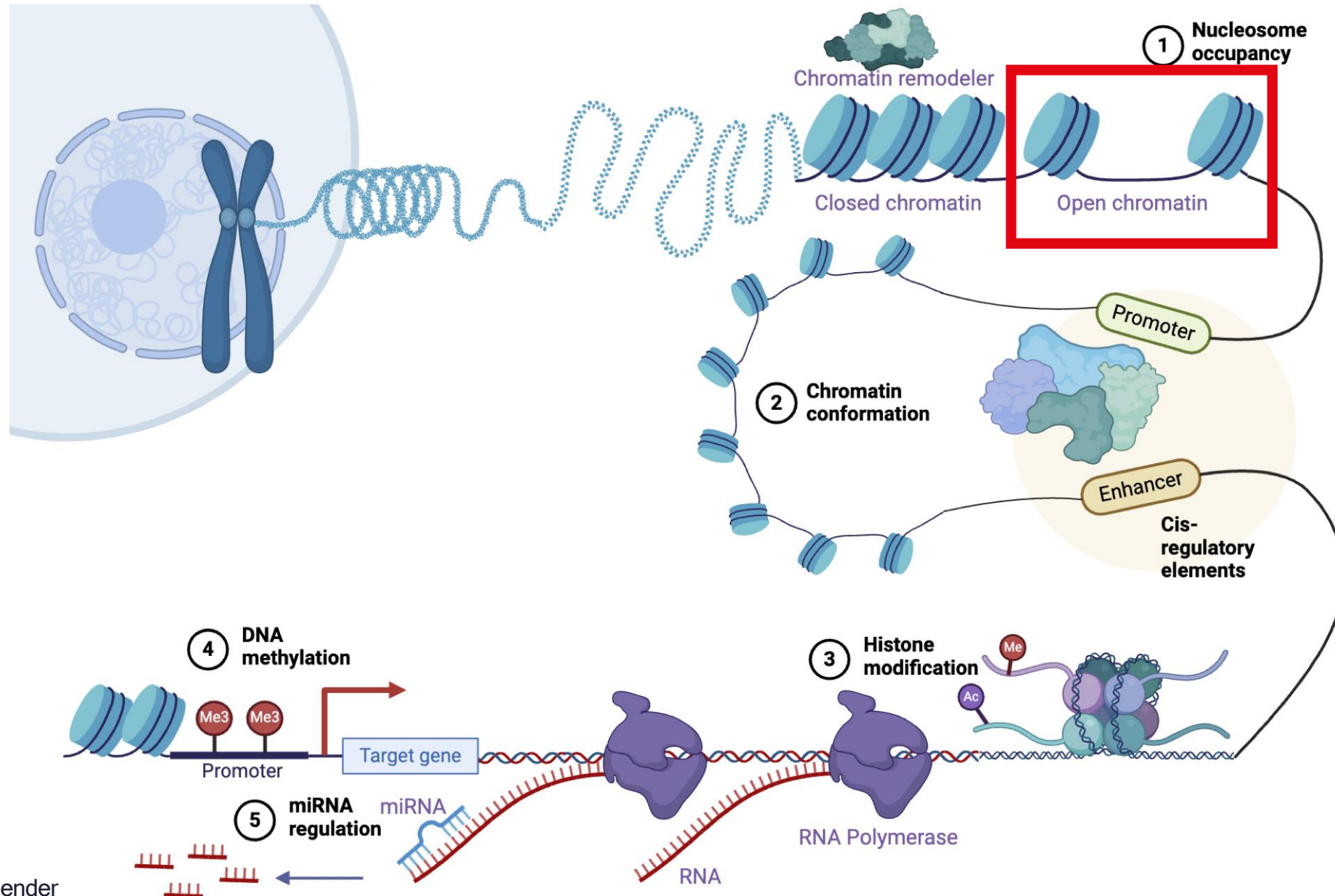
Deepak Tanwar

February 17-19, 2026

Learning outcomes

- Different layers of genome activity
- Basic pipeline for data generation
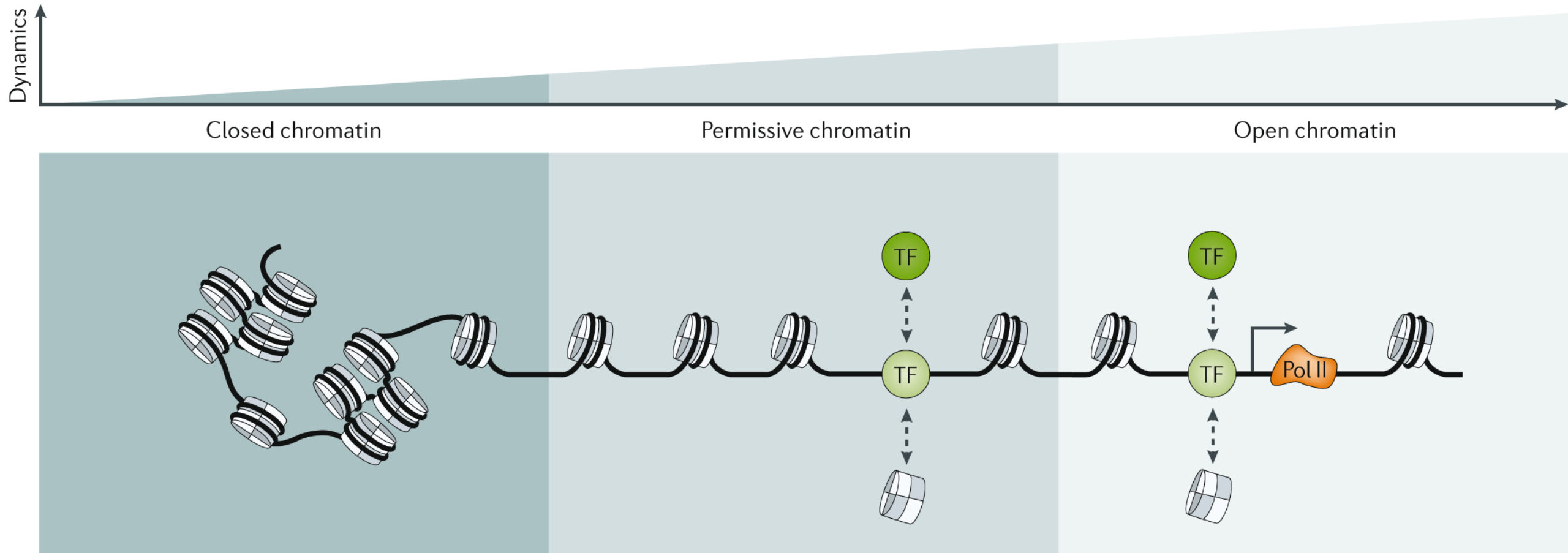- Basic pipeline for the data processing

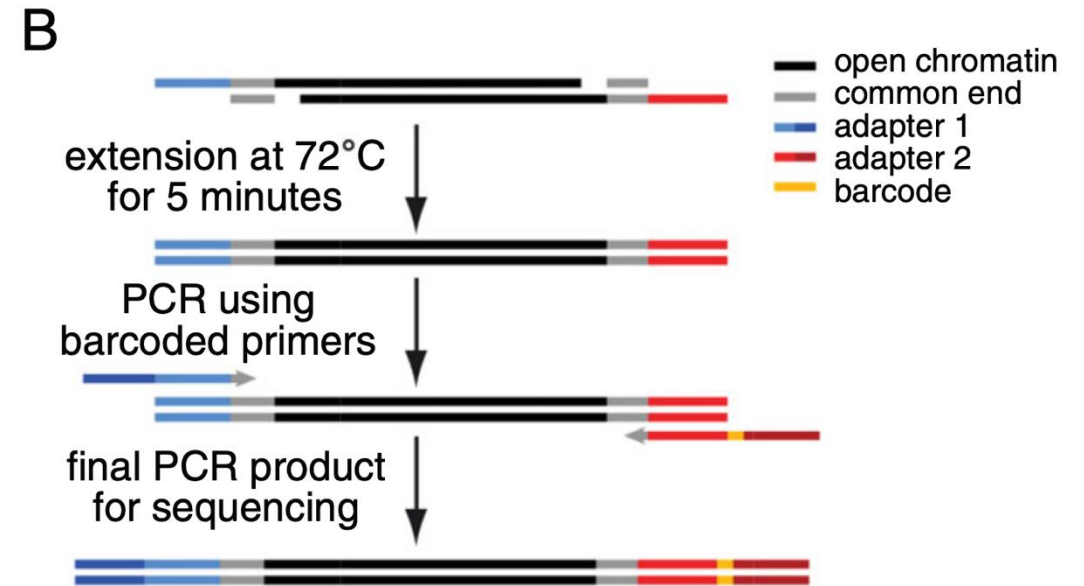# Simplified illustration of transcriptional regulation
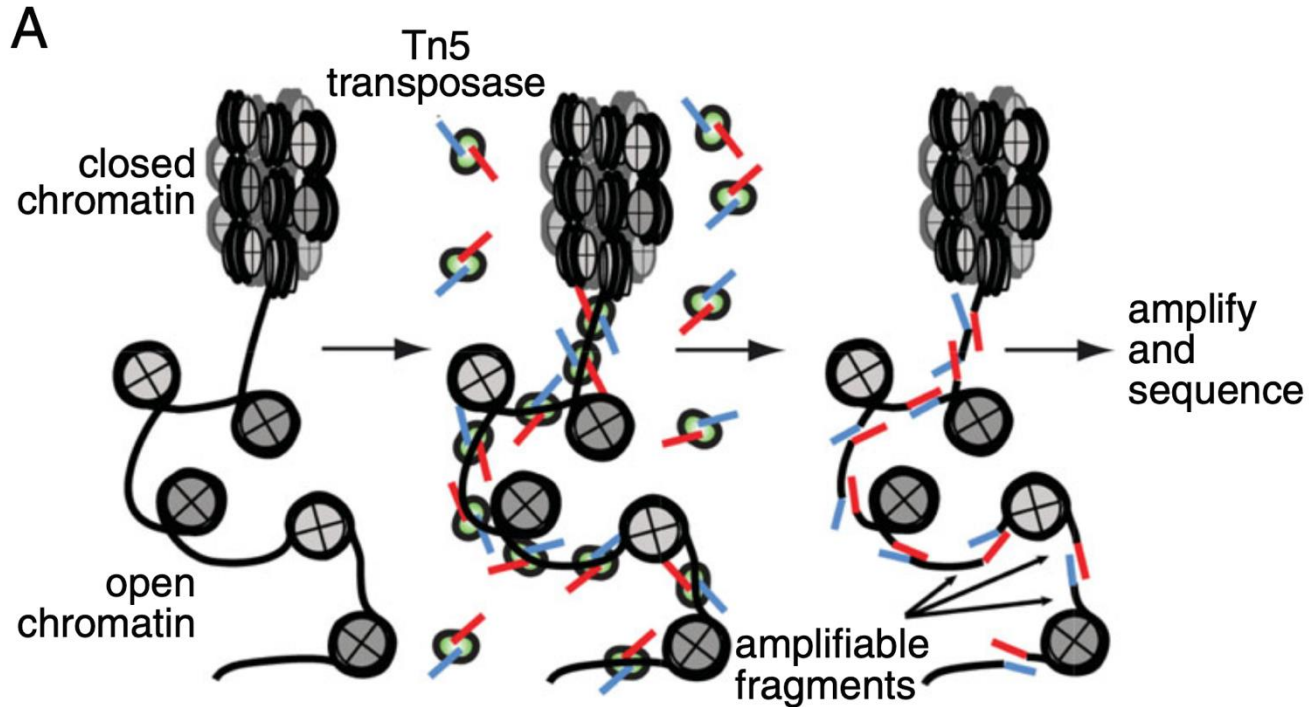

Generated using Biorender

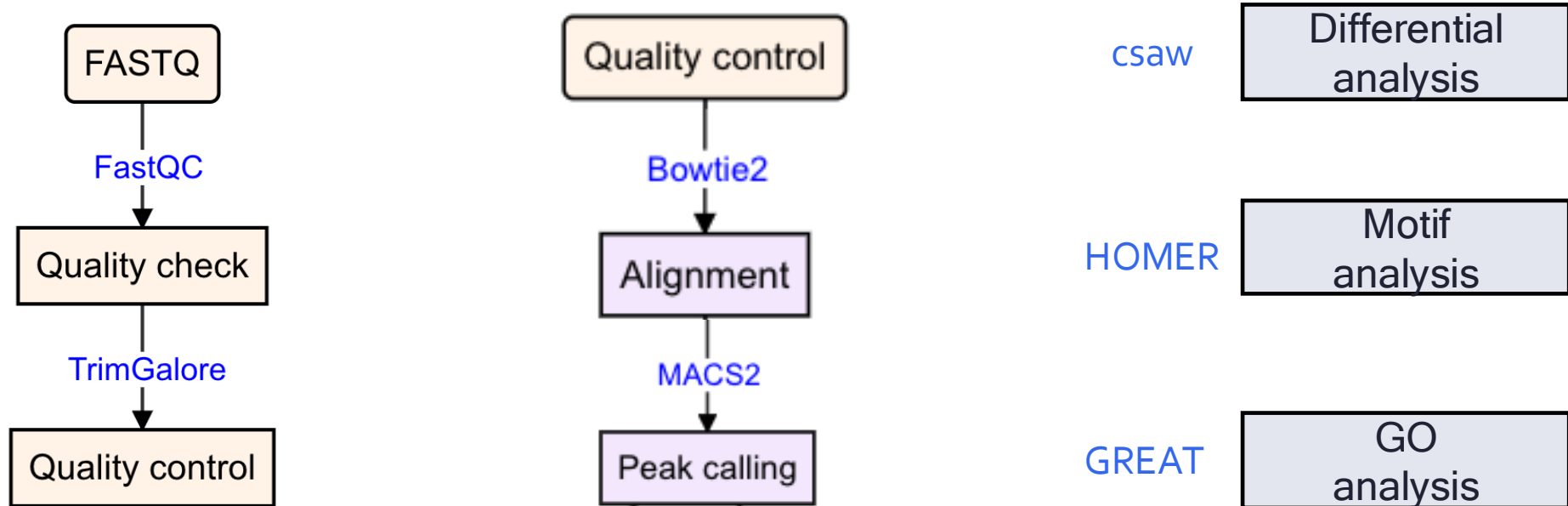# Simplified illustration of transcriptional regulation

Chromatin accessibility

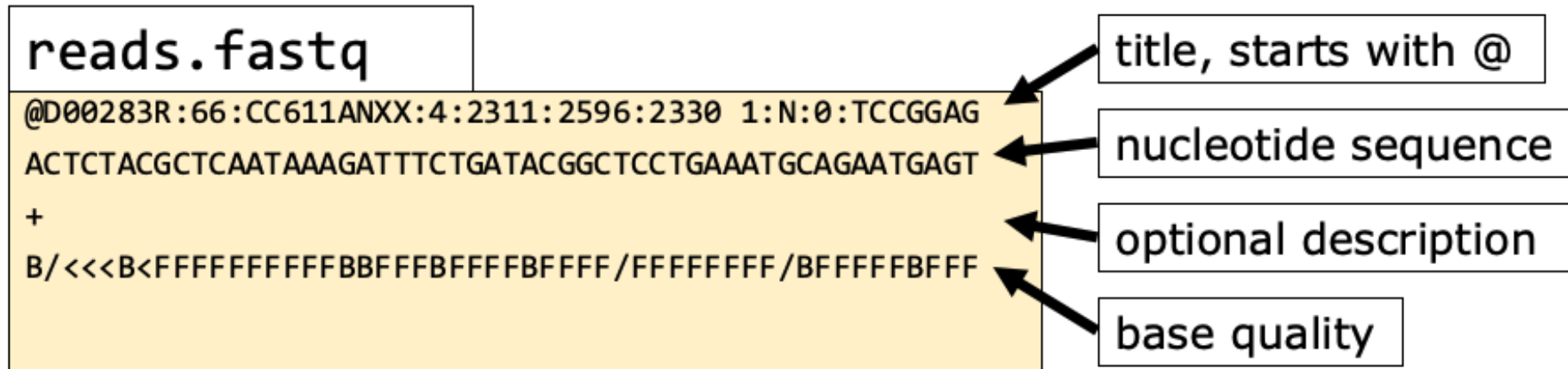# Library preparation for ATAC-seq

ATAC-seq data analysis pipeline

FASTA file

sequence.fasta

```
>sequence title1
ATCGTATCTATCGTATCT
GGTTTATCGTATCT
>sequence title2
ATGATGACGT
```

FASTQ file


```
reads.fastq
@D00283R:66:CC611ANXX:4:2311:2596:2330 1:N:0:TCCGGAG
ACTCTACGCTCAATAAAGATTTCTGATACGGCTCCTGAAATGCAGAATGAGT
+
B/<<<B<FFFFFFFFFFFBBFFFBFFFFBFFFF/FFFFFFFF/BFFFFFBFFF
```

title, starts with @

nucleotide sequence

optional description

base quality

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|                              |   |        |
0.2...........................26...31........41
```

SAM/ BAM file

```
@HD VN:1.5 SO:coordinate                                              Header
@SQ SN:ref LN:45                                                      section
r001     99 ref   7 30 8M2I4M1D3M = 37   39 TTAGATAAAGGATACTG *
r002      0 ref   9 30 3S6M1P1I4M * 0      0 AAAAGATAAGGATA       *
r003      0 ref   9 30 5S6M        * 0      0 GCCTAAGCTAA          * SA:Z:ref,29,-,6H5M,17,0;   Alignment
r004      0 ref  16 30 6M14N5M     * 0      0 ATAGCTTCAGC          *                           section
r003   2064 ref  29 17 6H5M        * 0      0 TAGGC                * SA:Z:ref,9,+,5S6M,30,1;
r001    147 ref  37 30 9M          = 7   -39 CAGCGGCAT            * NM:i:1
```

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

BigWig file

aim: **reduce the vast amount of information** from the BAM file to the simple information: **How many reads do I have** (per bp/genomic bin/…)?

BAM file

```
39V34V1:38:C0RLHACXX:4:1216:16137:31969 163 chr1 3000307 42 51M
= 3000408 152
CTGTAGTTACTGTTTGCTTACCTAGATTCTTCTTTTCCAGAATTCTCTTAG
CCCFFFFFHHHGHIIJIJJJJIIGHFGIGIJIIJJJHIHEHIGIIIIJJGF AS:i:0
XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:51 YS:i:0 YT:Z:CP
```
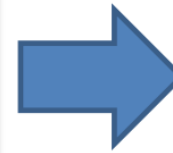
bedGraph/bigWig

```
chr2    100100    100120    5
chr2    100121    100141    3.2
chr2    100142    100163    13.8
```
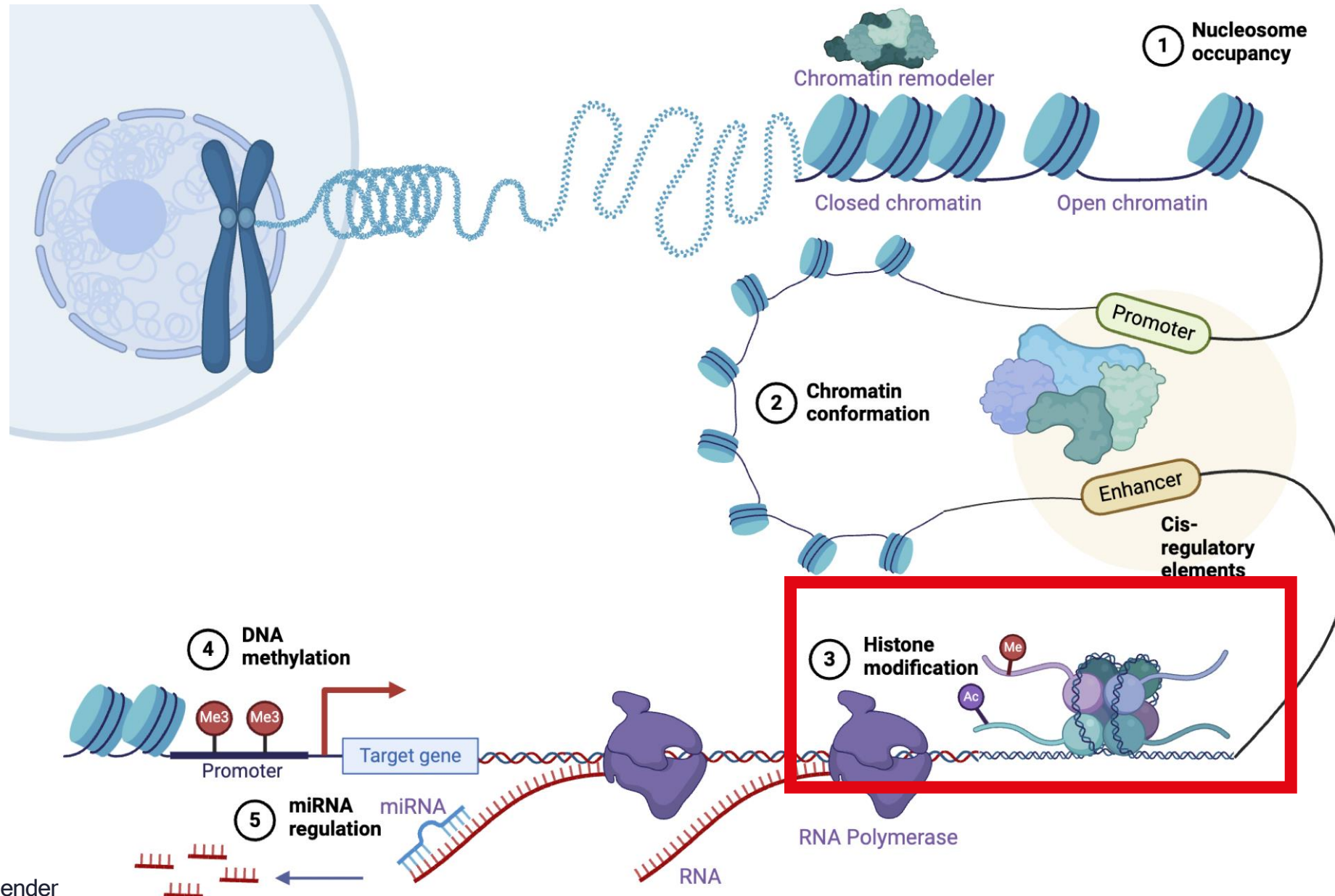
Position (bp)   ⊢ 50 bp ⊣

the reduced size leads to many advantages of bigwig files over BAMs:
- data storage
- data sharing!
- intuitive visualization via genome browsers
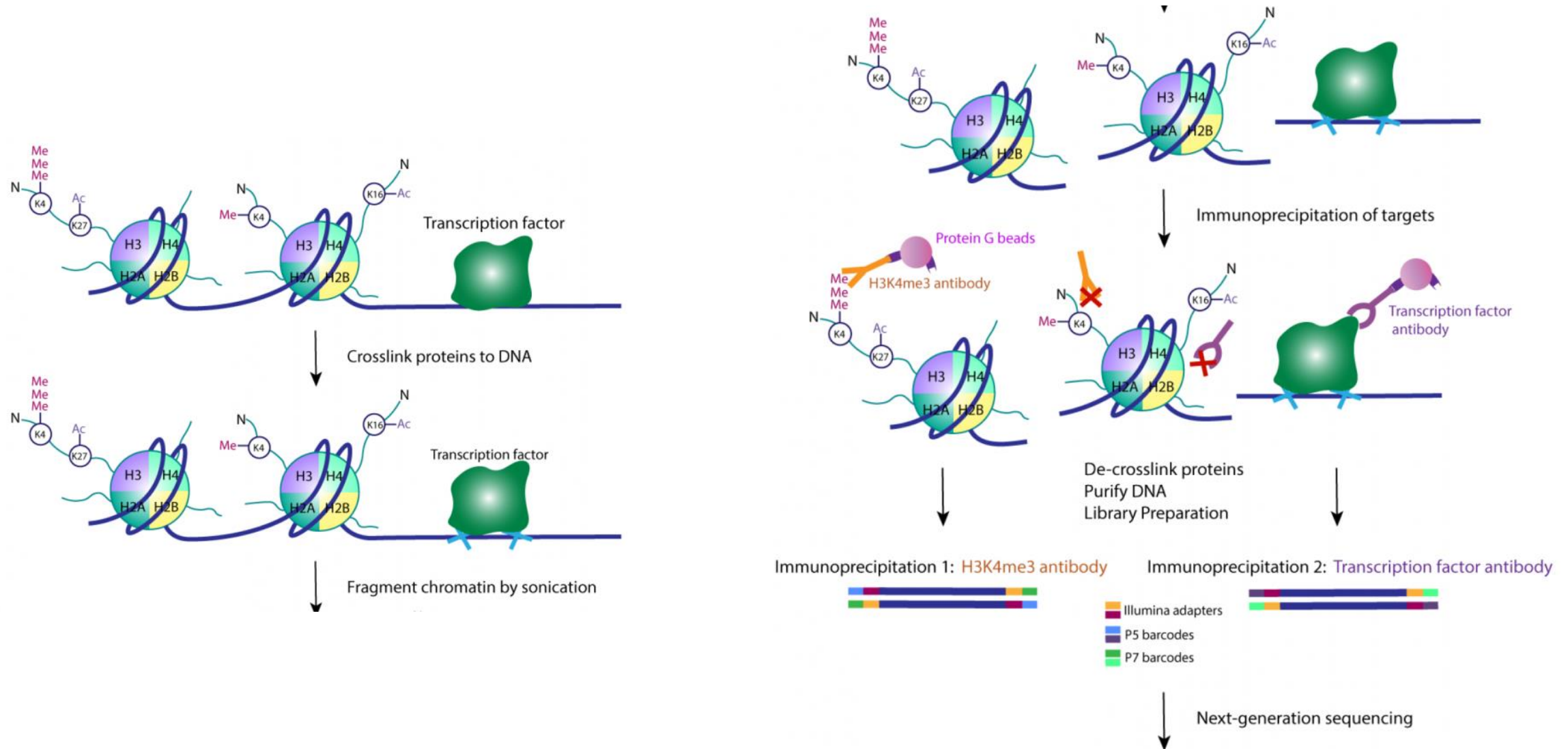- more efficient for down-stream analyses (…)

SIB

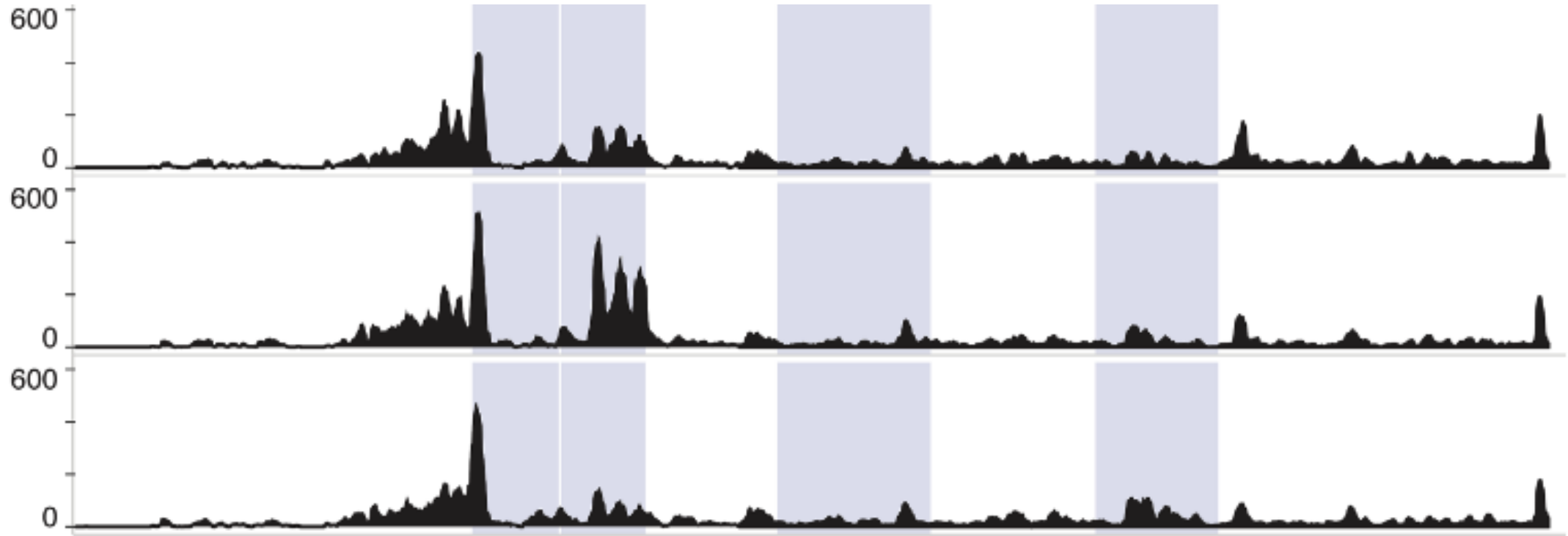# Example snapshot of chromatin accessibility
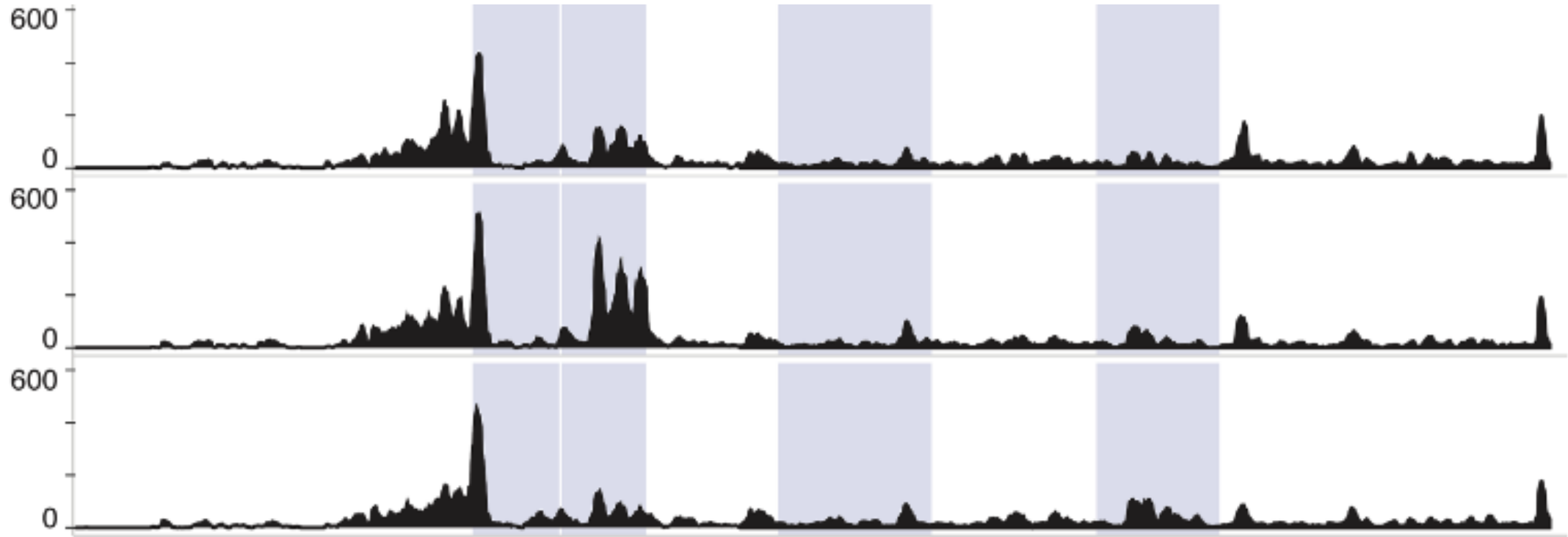
Simplified illustration of transcriptional regulation

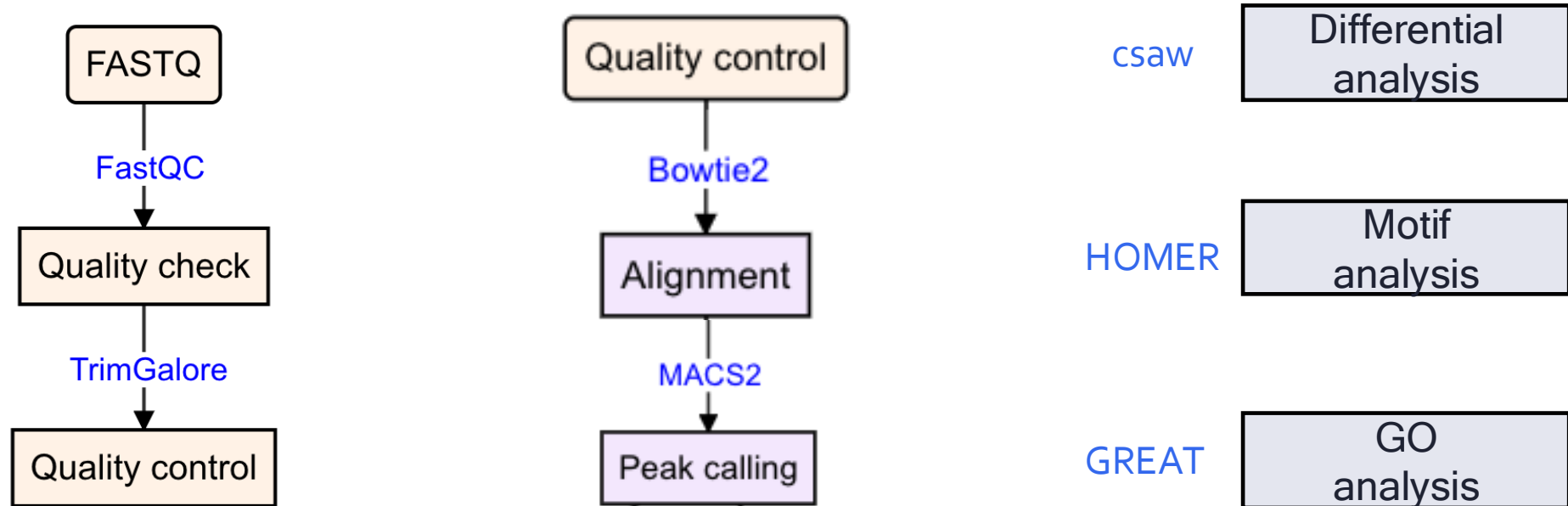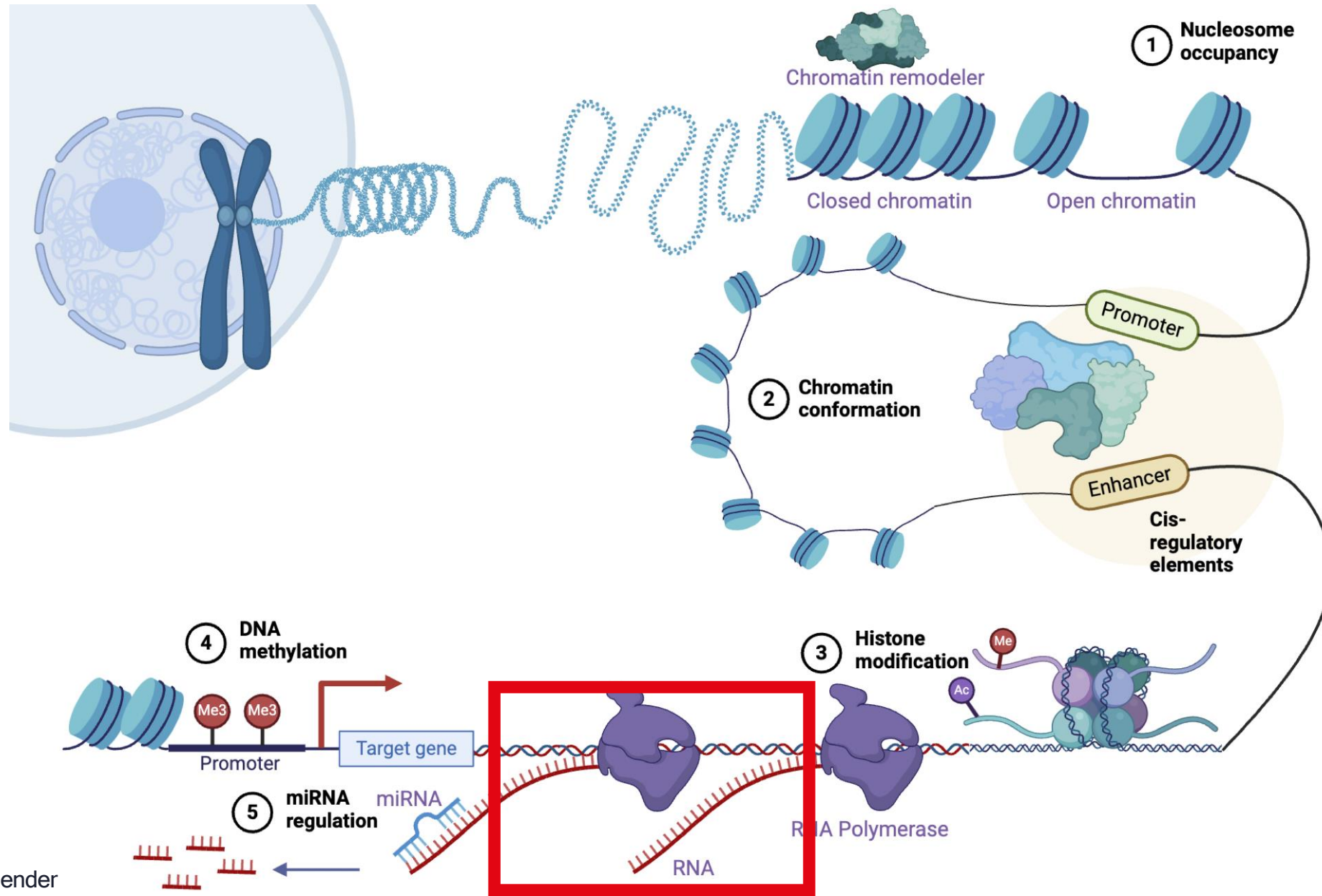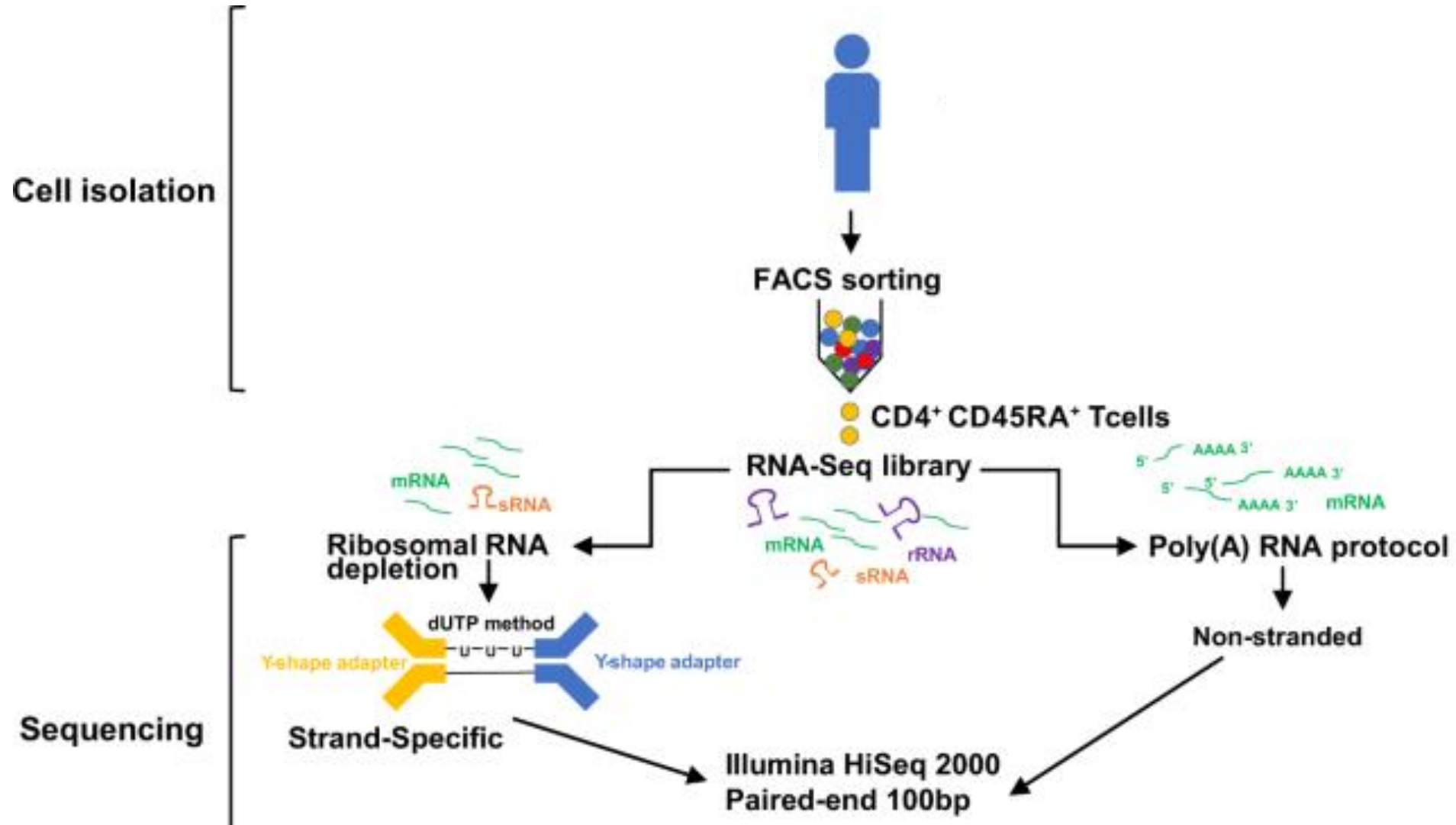# Chromatin Immunoprecipitation sequencing

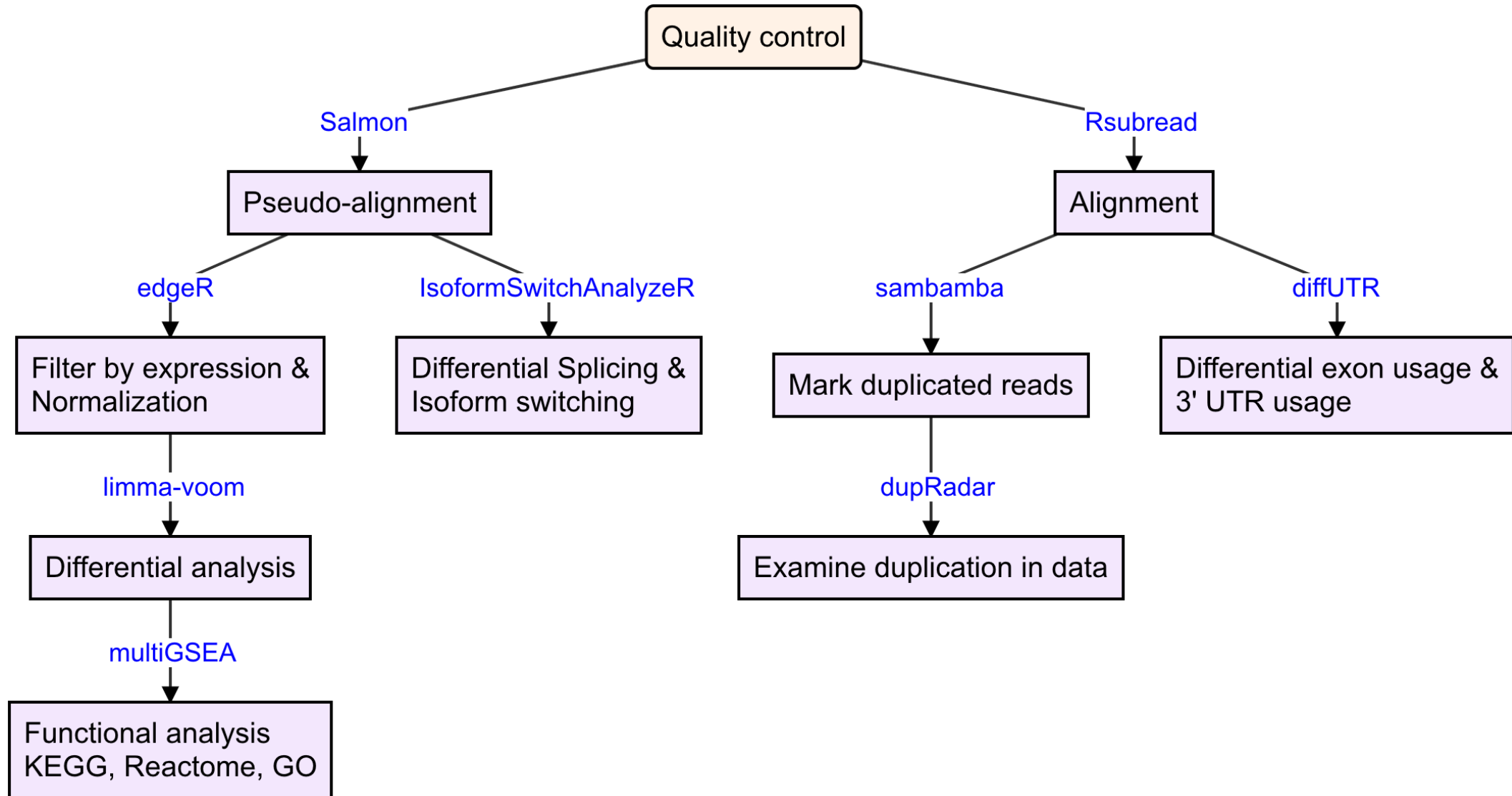# Example snapshot of enrichment

ChIP-seq data analysis pipeline

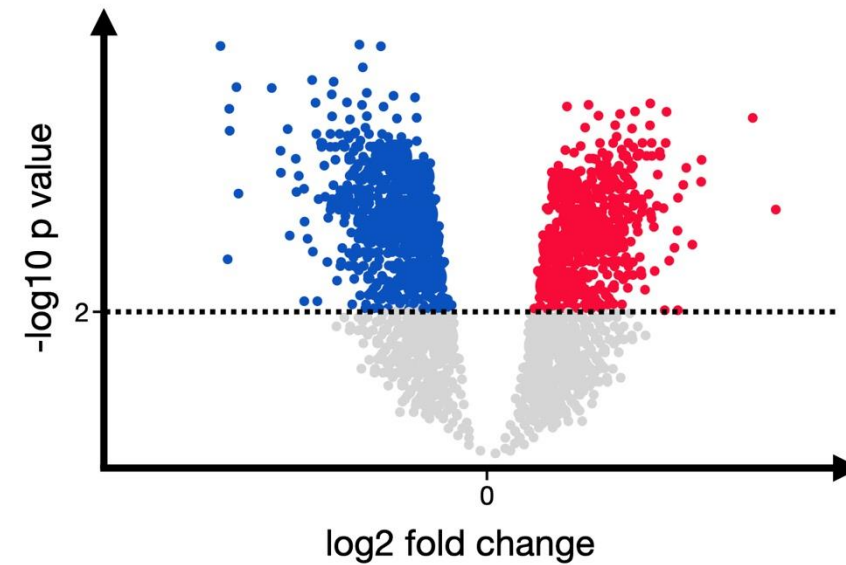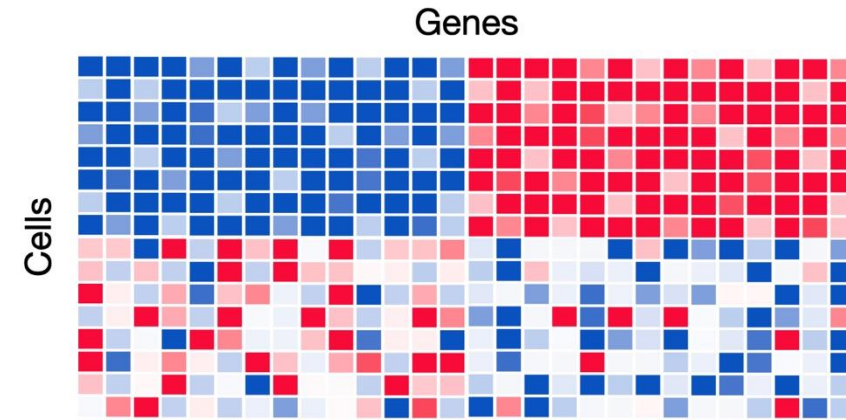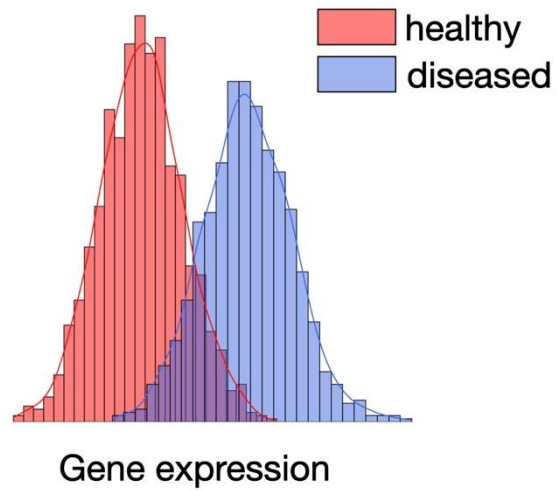# Simplified illustration of transcriptional regulation
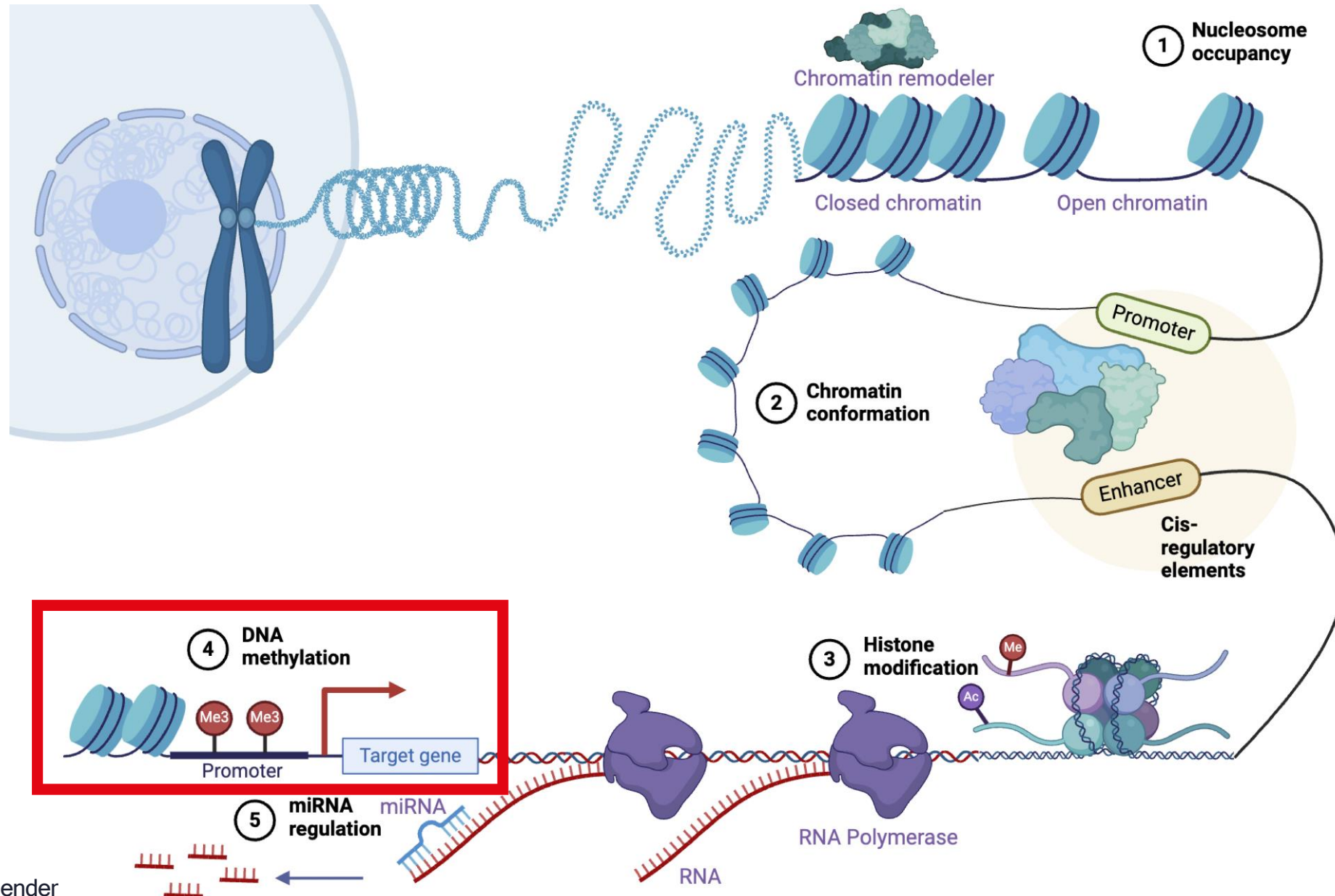

Generated using Biorender

Gene expression

RNA-seq data analysis pipeline

Differences in gene expression

Simplified illustration of transcriptional regulation

Generated using Biorender

DNA cytosine methylation is a biological process by which methyl groups are added to the DNA molecule
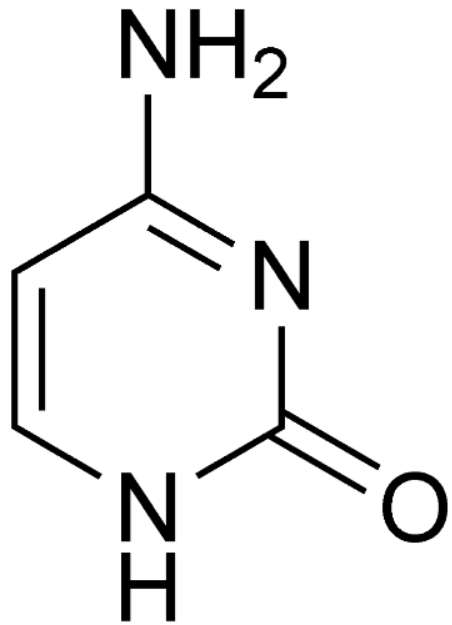
DNA methyl-transferases

DNA-demethylase(s)
TET enymes
Passive demethylation
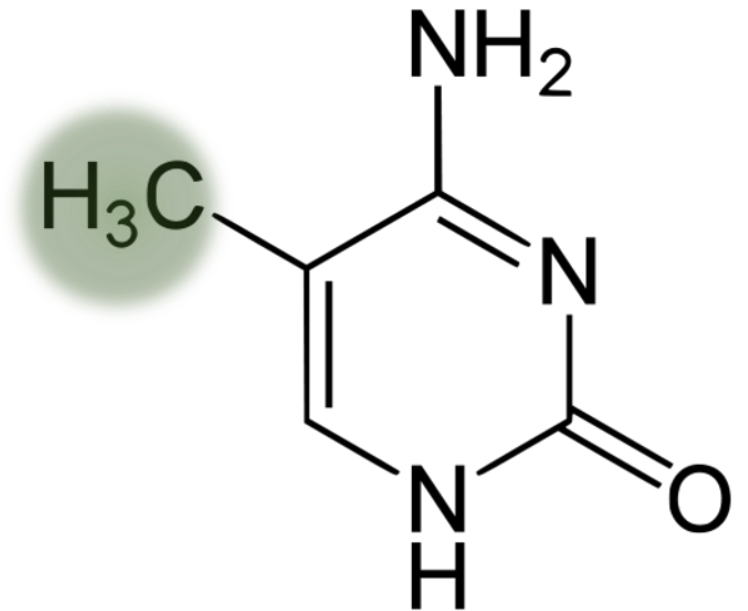
Cytosine

H3C — 5-methyl Cytosine

Role of DNA methylation in transcriptional regulation
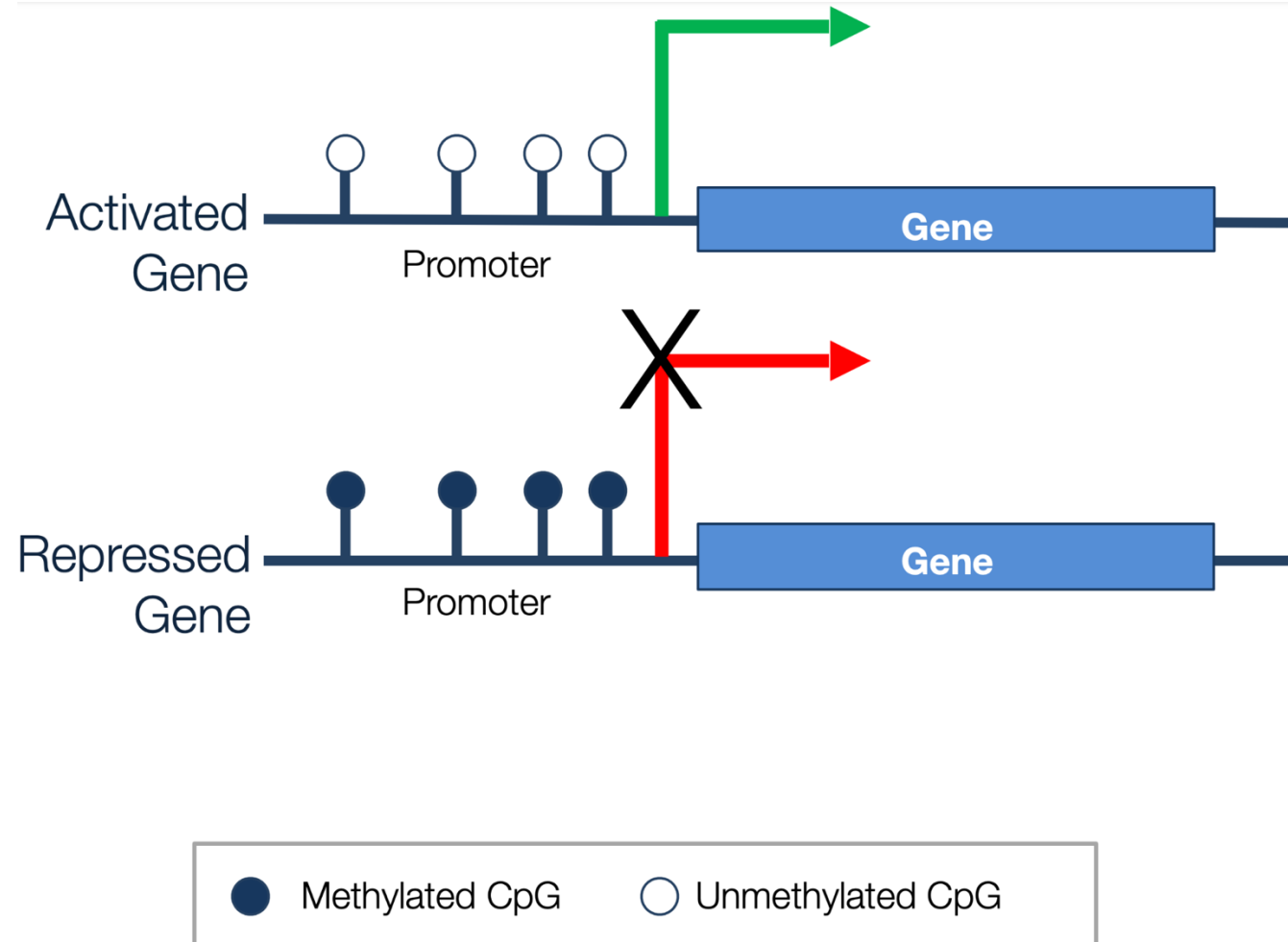
Gold standard to study DNA methylation patterns across the genome: WGBS

Genomic DNA

+ Bisulfite

Bisulfite conversion

Ligate adpaters

Amplify and Sequence
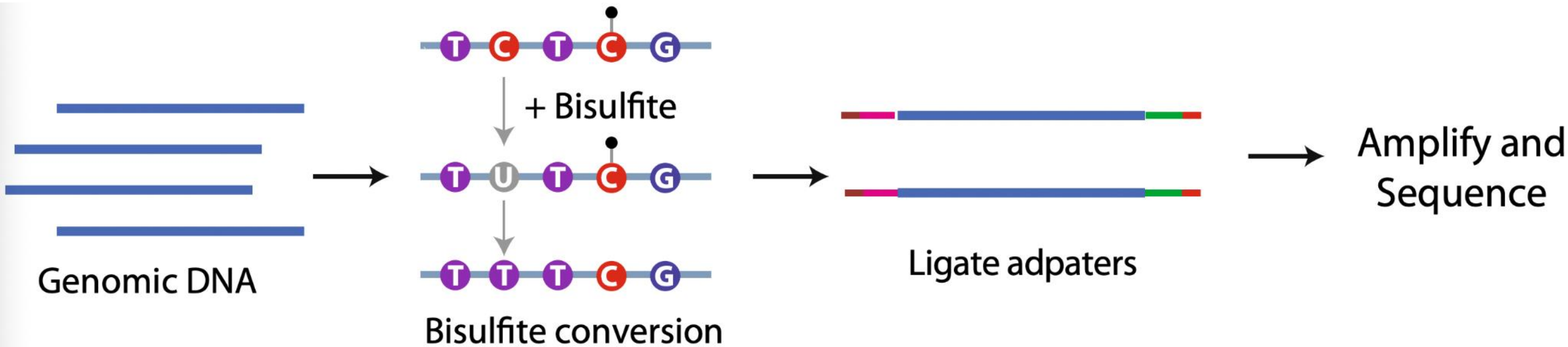
Example plot for DNA methylation

chr2: 4,799,675 − 4,802,240 (width = 2,566)
Stat: −10.552, FDR: 0.522
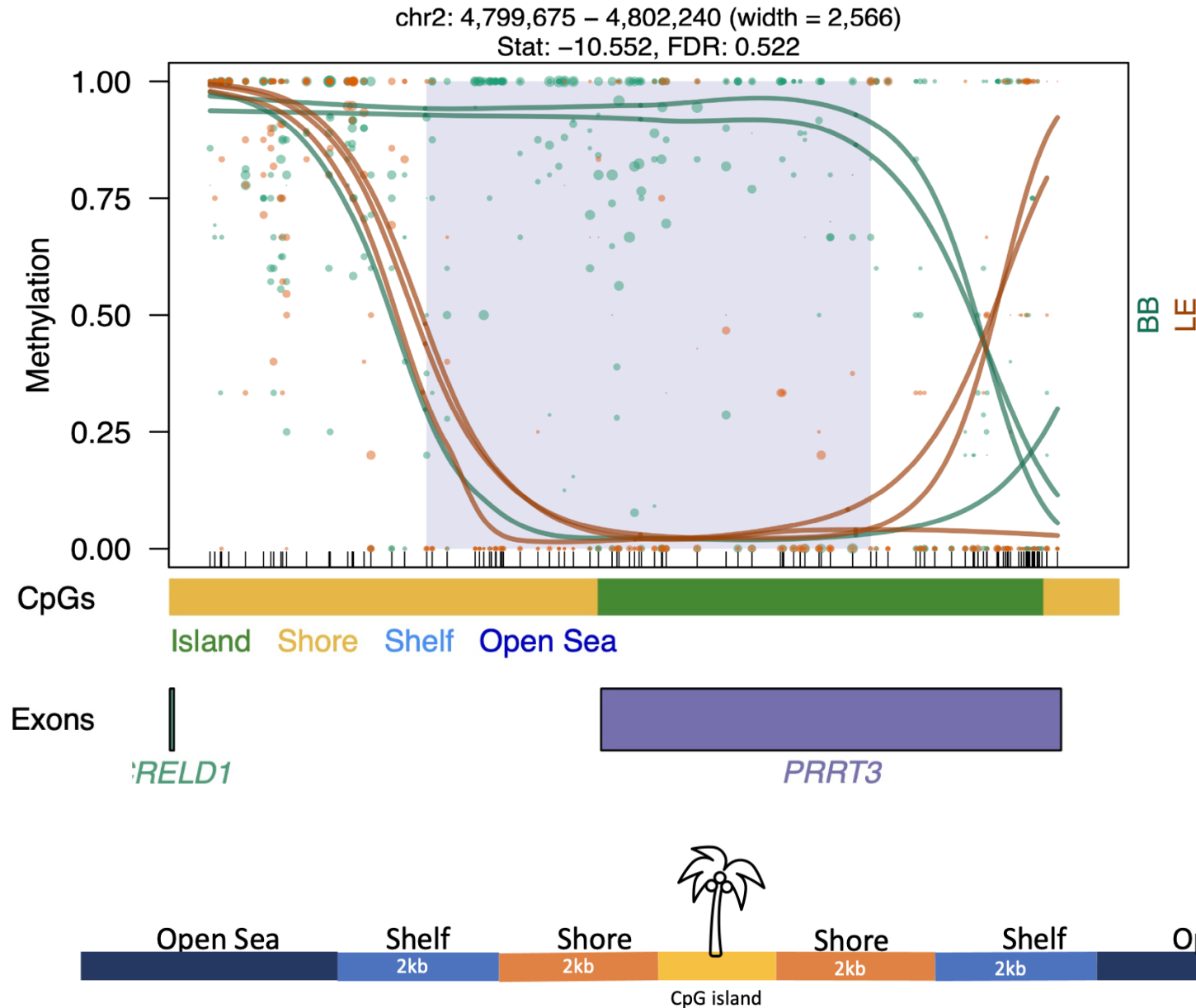
DNAme data analysis pipeline
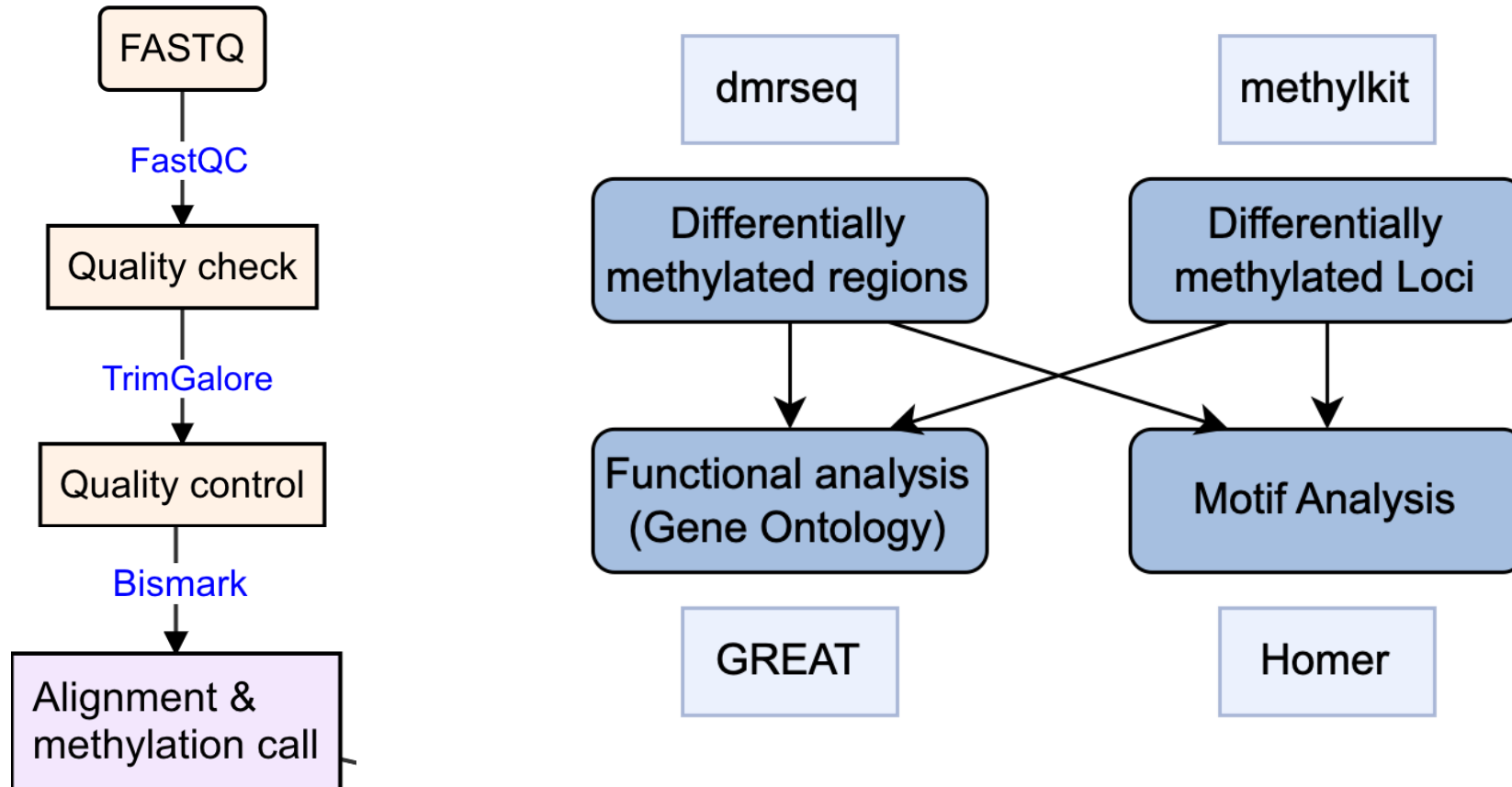
File structures

FASTQ

Integration of RNA-seq and WGBS
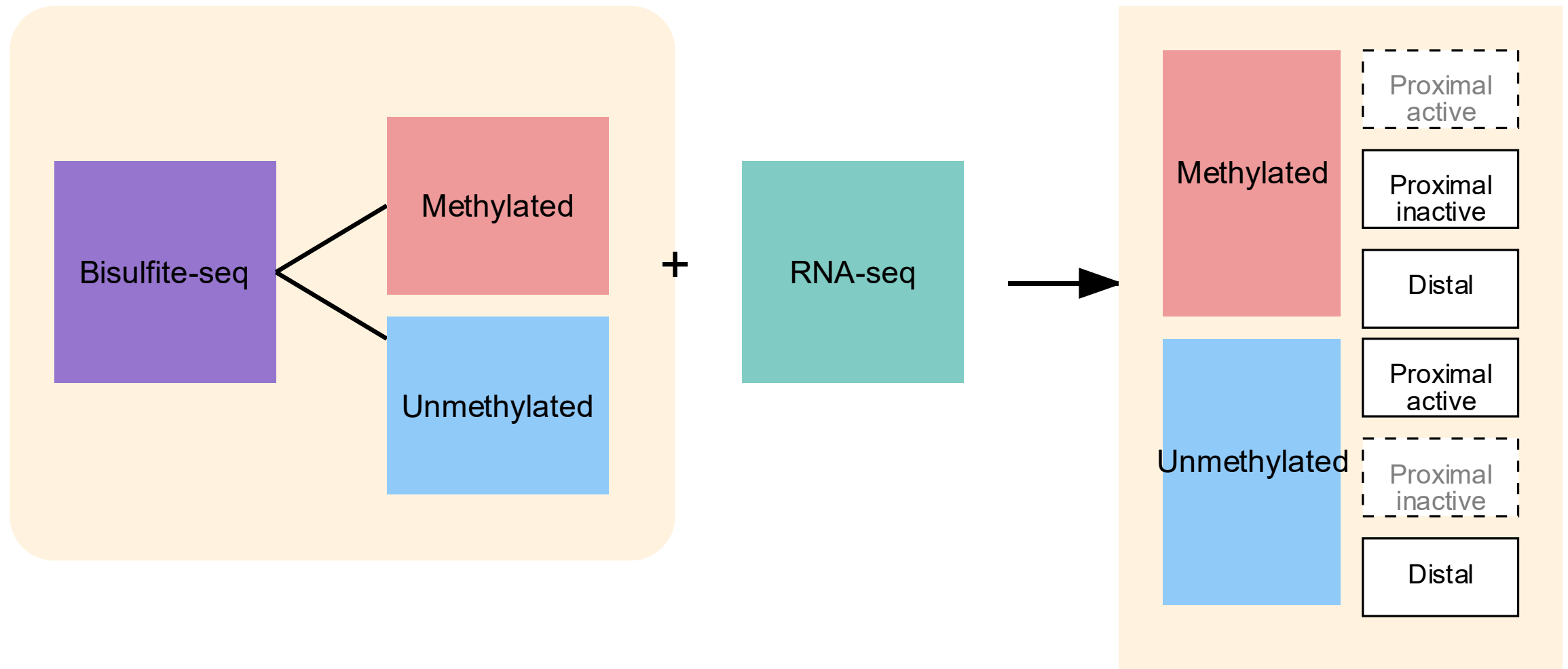
Integration of RNA-seq and ATAC-seq
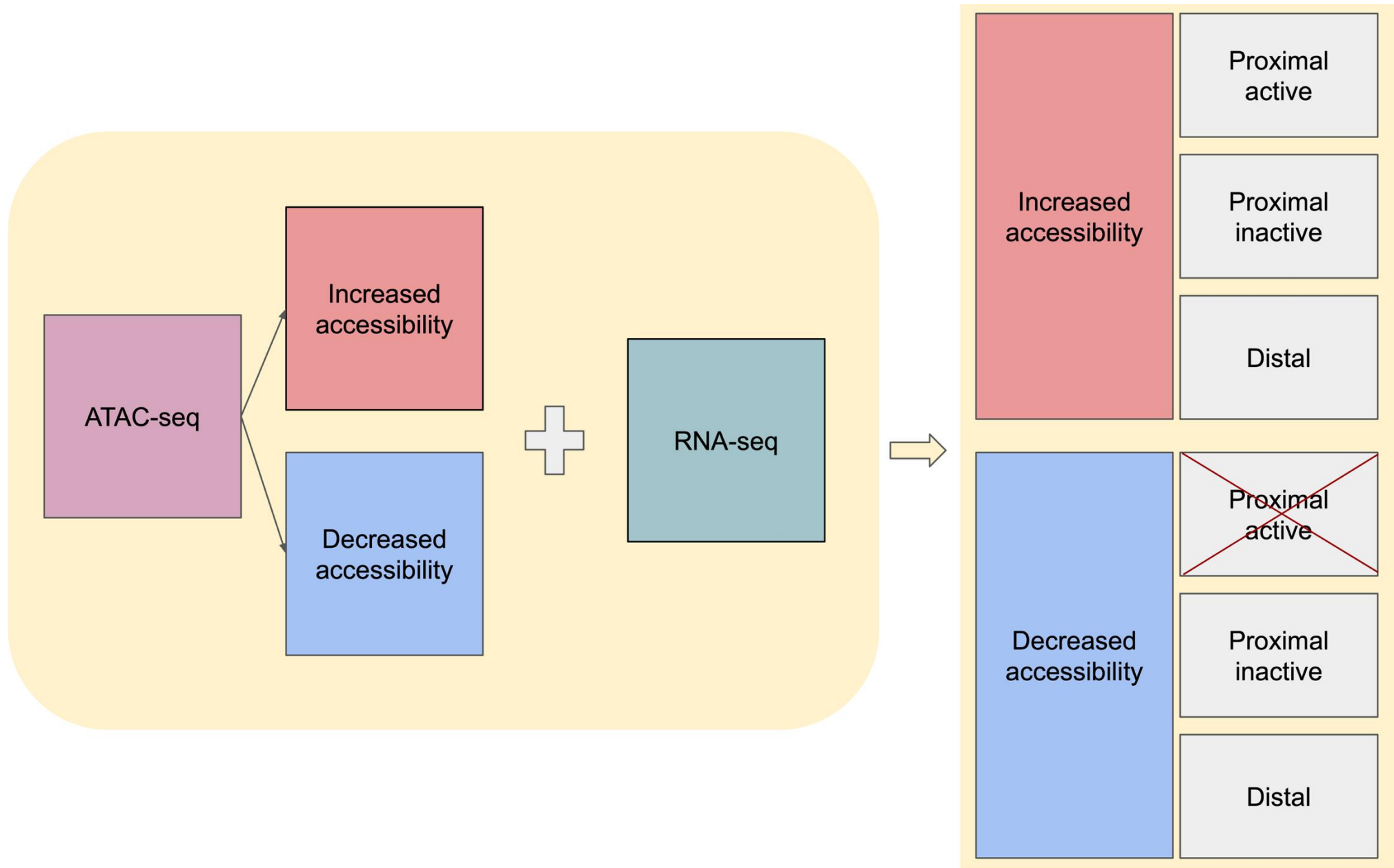
Integration of RNA-seq and ATAC-seq: diffTF

A Schematic of diffTF - basic mode

ATAC-Seq in group 1 (e.g. U-CLL)

Coverage

ATAC-Seq in group 2 (e.g. M-CLL)

Coverage

For each TF:
calculate fold changes at each TFBS and compare to background

TFBS of given TF

Background

Density

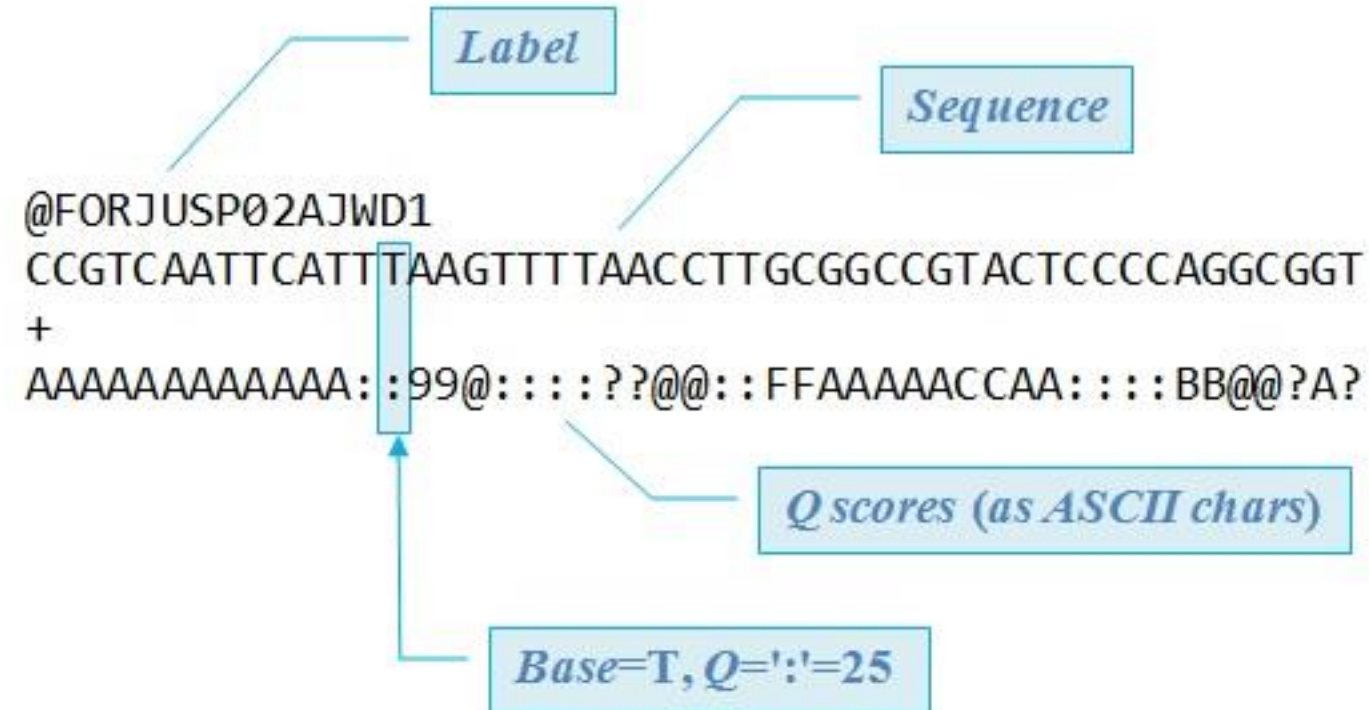Log2 fold changes [group 1 / group 2]
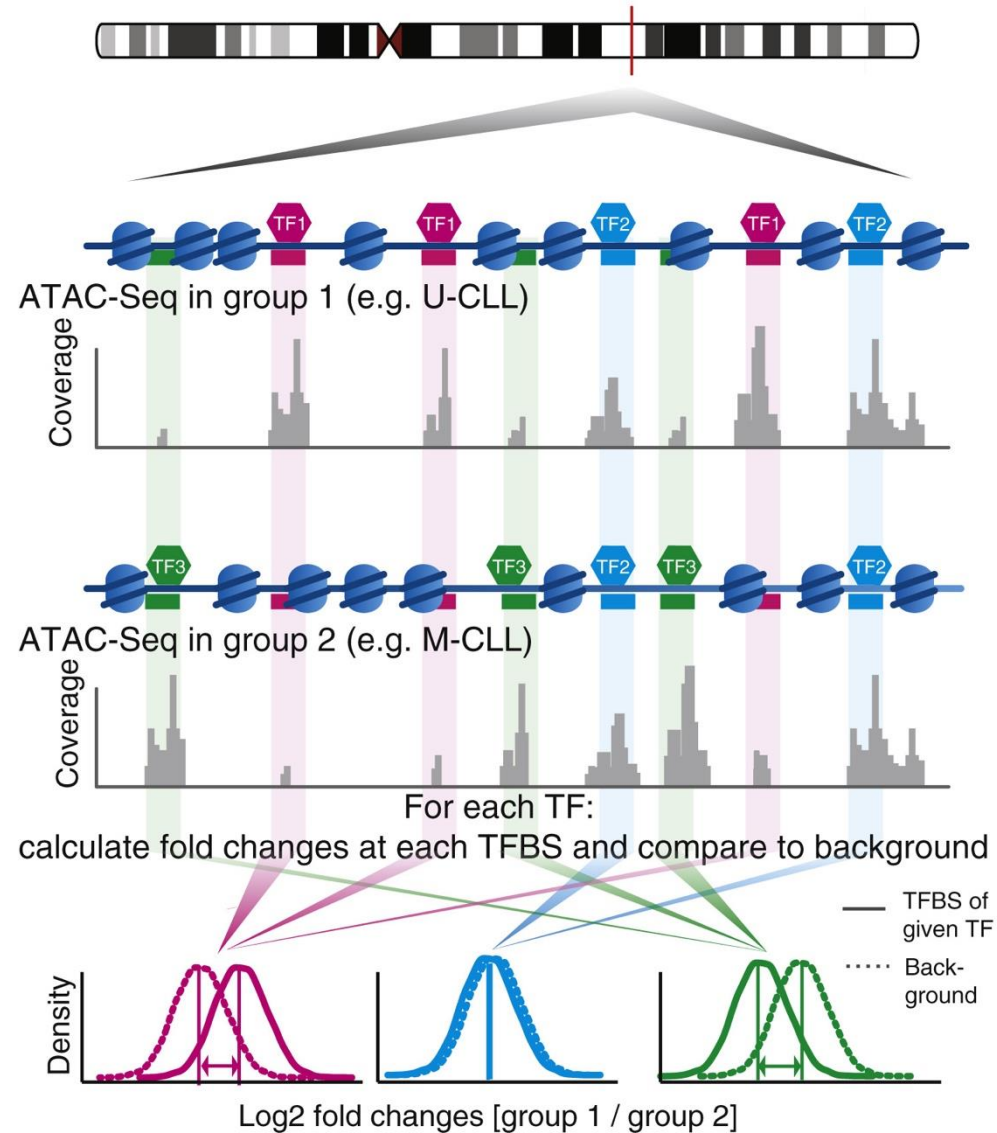
B Schematic of diffTF - classification mode

RNA-Seq

ATAC-Seq

Individuals

TF1

Repeat for all TFs

$R_{bg} \sim 0$

$Peak_x$ signal — TF_i expression

$Peak_y$ signal — TF_j expression

$Peak_z$ signal — TF_k expression

Repeat for all peaks

$R >> 0 \Rightarrow$ activator $R \sim 0 \Rightarrow$ undetermined $R << 0 \Rightarrow$ repressor

Density

Pearson correlation coefficient
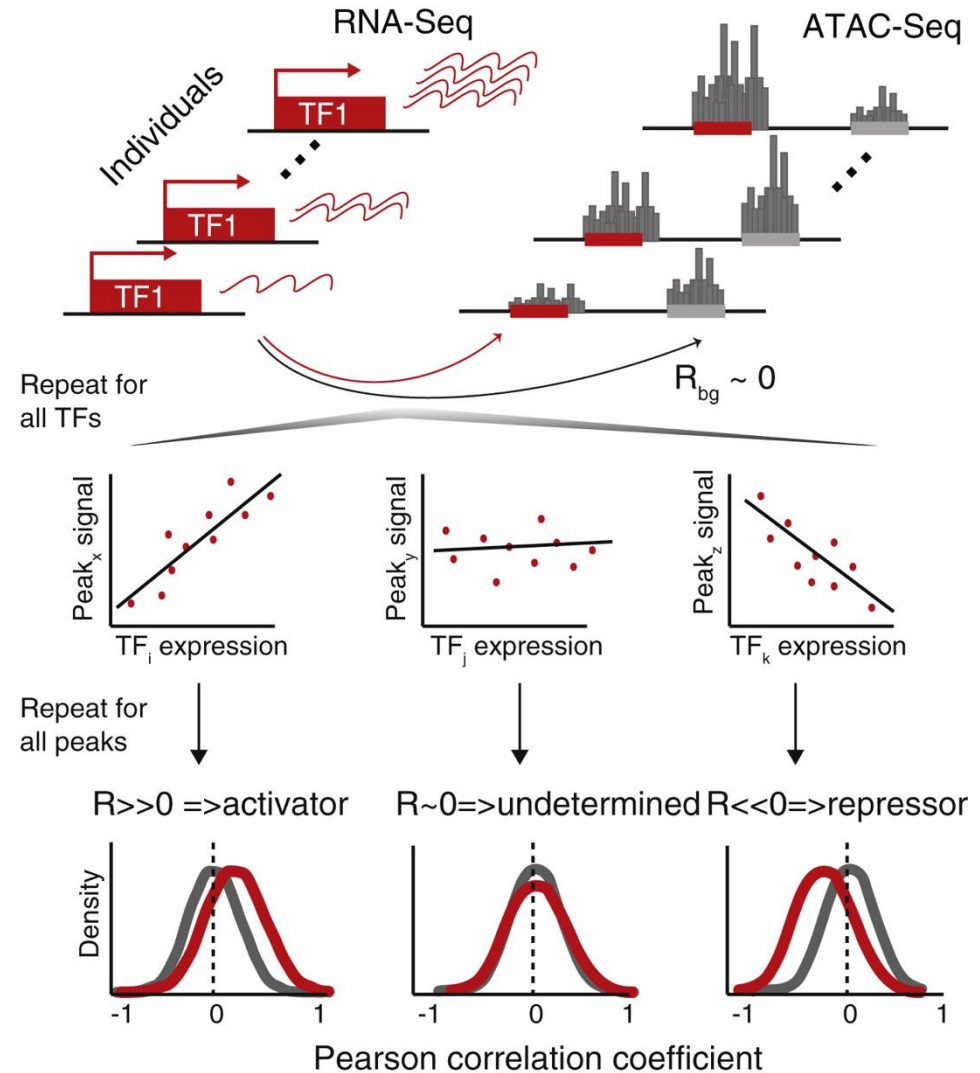
Multi-omics integration overview

Feature	Systematic Approaches	Ad-hoc Approaches
Definition	Algorithmic frameworks for integrating multiple omics datasets using statistical or machine learning models.	Custom, hypothesis-driven integration of 1–2 omics datasets based on specific biological questions.
Tools	- mixOmics, DIABLO - MOFA - MINT - sGCCA	- diffTF - Custom R/Python scripts
Supervision	Supervised, unsupervised, or semi-supervised	Typically unsupervised or guided by prior biological knowledge
Omics Types	Multiple omics datasets (≥2), can be diverse (transcriptomics, epigenomics, proteomics, etc.)	Usually 2 omics types (e.g., RNA-seq + ATAC-seq)
Advantages	- Scalable and reproducible - Handles complex data - Reveals hidden patterns - Supports biomarker discovery	- Biologically intuitive - Easy to prototype - Focused and hypothesis-driven
Interpretability	Moderate (depends on model complexity)	High (based on direct biological rationale)

Multi-omics integration

Feature	Systematic Approaches	Ad-hoc Approaches
Definition	Algorithmic frameworks for integrating multiple omics datasets using statistical or machine learning models.	Custom, hypothesis-driven integration of 1–2 omics datasets based on specific biological questions.
Tools	- mixOmics, DIABLO - MOFA - MINT - sGCCA	- diffTF - Custom R/Python scripts
Supervision	Supervised, unsupervised, or semi-supervised	Typically unsupervised or guided by prior biological knowledge
Omics Types	Multiple omics datasets (≥2), can be diverse (transcriptomics, epigenomics, proteomics, etc.)	Usually 2 omics types (e.g., RNA-seq + ATAC-seq)
Advantages	- Scalable and reproducible - Handles complex data - Reveals hidden patterns - Supports biomarker discovery	- Biologically intuitive - Easy to prototype - Focused and hypothesis-driven
Interpretability	Moderate (depends on model complexity)	High (based on direct biological rationale)

We will learn ad-hoc approaches for multi-omics in this course

Quiz: 6-10
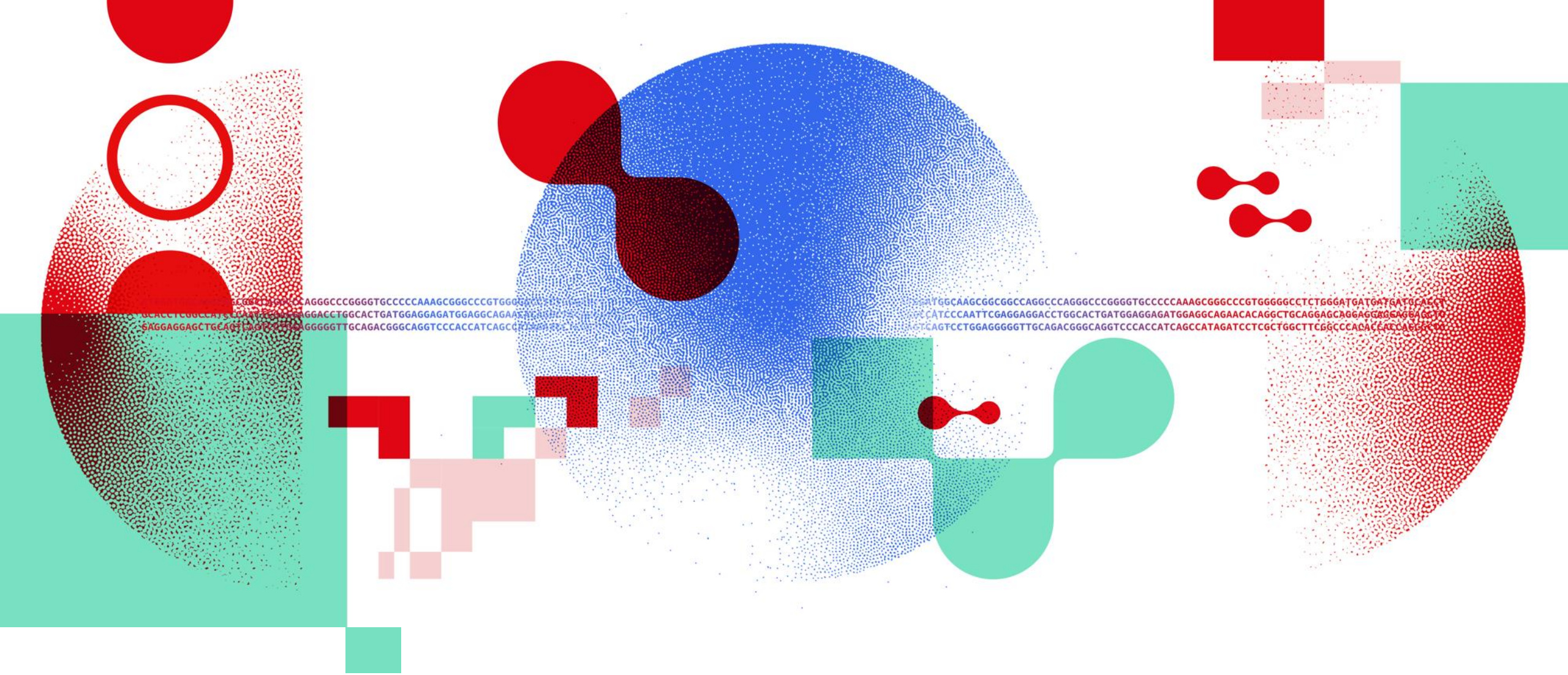
Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss