Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L - 2022 | week 07

Pierre-Luc Germain

ETH Zürich

Today’s plan

(launch install)

- Debriefing on the assignment
- DNA accessibility
- ATAC-seq analysis (practical)
- Nucleosome positioning

Debriefing on the assignment

TF: Nanog

Scanning for motif in peaks and answering question 1

```
moi <- memes::runFimo(peak_seqs,  convert_motifs(motif), meme_path = "/mnt/IM/conda/bin/")

length(peaks)
```

```
## [1] 16037
```

```
sum(overlapsAny(peaks, moi))
```

```
## [1] 4311
```

```
sum(overlapsAny(peaks, moi))/length(peaks)
```

```
## [1] 0.2688159
```

From all Nanog peaks (n = 16037), 4311 (26.88%) contain the chosen Nanog motif

Debriefing on the assignment

TF: REST

```r
peaks <- rtracklayer::import("REST_ENCFF368VWJ.bed.gz", format="NarrowPeak")
seqlevelsStyle(peaks) <- "Ensembl"
peaks_chr1 <- peaks[seqnames(peaks)=="1"]
peak_centers <- resize(peaks_chr1, fix="center", width=100)
peak_seqs <- memes::get_sequence(peak_centers, genome)
moi2 <- findMotifInstances(peak_seqs, motif, mc.cores=2) # running with 2 threads
sum(overlapsAny(peaks, moi2))
```

```r
percentage <- (sum(overlapsAny(peaks, moi2)))/length(peaks)*100
percentage
```

```
## [1] 2.531646
```

Result: Of all the 3555 peaks, 90 (2.531646%) contain a motif.

What’s the problem here?

Debriefing on the assignment

TF: MYOD1

```
# instances of motif bound by MYOD1 in chr1

motif_bound_by_MYOD1 <- findOverlaps(motif_in_chr1, motif_in_MOYD1_peaks)

length(motif_bound_by_MYOD1)
```

```
## [1] 29810
```

Of all the peaks, what proportion contains a motif for the factor?

Of the 22641 peaks, 24495 contain a motif

–> this is very odd because it appears as if there are more sequences containing MYOD1 peak than there are peaks in total. How can this be explained?

Debriefing on the assignment

```
table(overlapsAny(motifs, peaks))
```

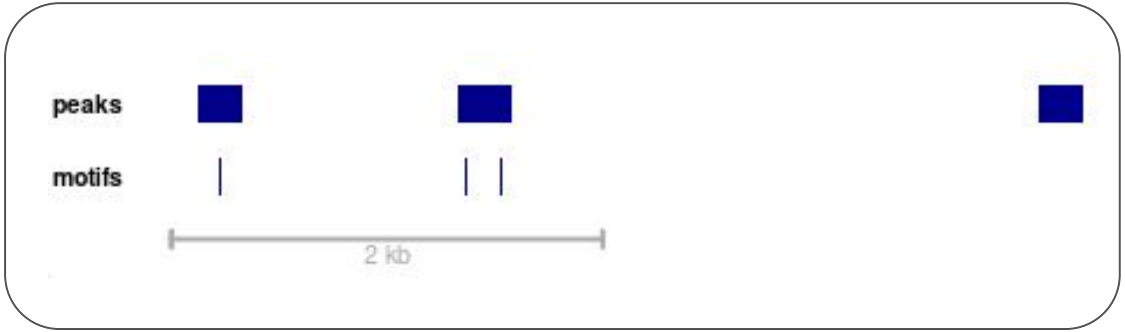
```
##
## TRUE
##    3
```

```
table(overlapsAny(peaks, motifs))
```

```
##
## FALSE  TRUE
##     1     2
```

```
findOverlaps(peaks, motifs)
```

```
## Hits object with 3 hits and 0 metadata columns:
##       queryHits subjectHits
##       <integer>   <integer>
##   [1]         1           1
##   [2]         2           2
##   [3]         2           3
##   -------
##   queryLength: 3 / subjectLength: 3
```



Of all instances of that motif in the genome, what proportion is bound by the factor (i.e. has a peak)?

```
mmusculus <- import(genome, "2bit", which = as(seqinfo(genome), "GenomicRanges"))
motif_instances_genome <- findMotifInstances(mmusculus, motif, mc.cores=2)
```

```
## Note: motif [motif] has an empty nsites slot, using 100.
```

```
length(motif_instances_genome)
```

```
## [1] 6544422
```

```
motif_with_peaks = overlapsAny(motif_instances_genome, peaks)
sum(motif_with_peaks)
```

```
## [1] 16856
```

```
percentage2 <- sum(motif_with_peaks)/length(motif_instances_genome)*100
percentage2
```

```
## [1] 0.2575629
```

Of the 6544422 motif instances, 16856 (0.2575629%) overlap a peak.

TF: GATA1

Of the 9675 GATA1 peaks, 7277 (~75%) contain a GATA1 motif, but…

Debriefing on the assignment – wrapping up

- The proportion of binding sites that show the motif depends on the TF (ranging from roughly 20 to 95%)
- The proportion of (genome-wide) motif instances that are bound by the factor is typically very very small
- This means that something else determines whether a stretch of DNA will be bound

DNA accessibility, which is associated to lower nucleosome density, **reflects activity**

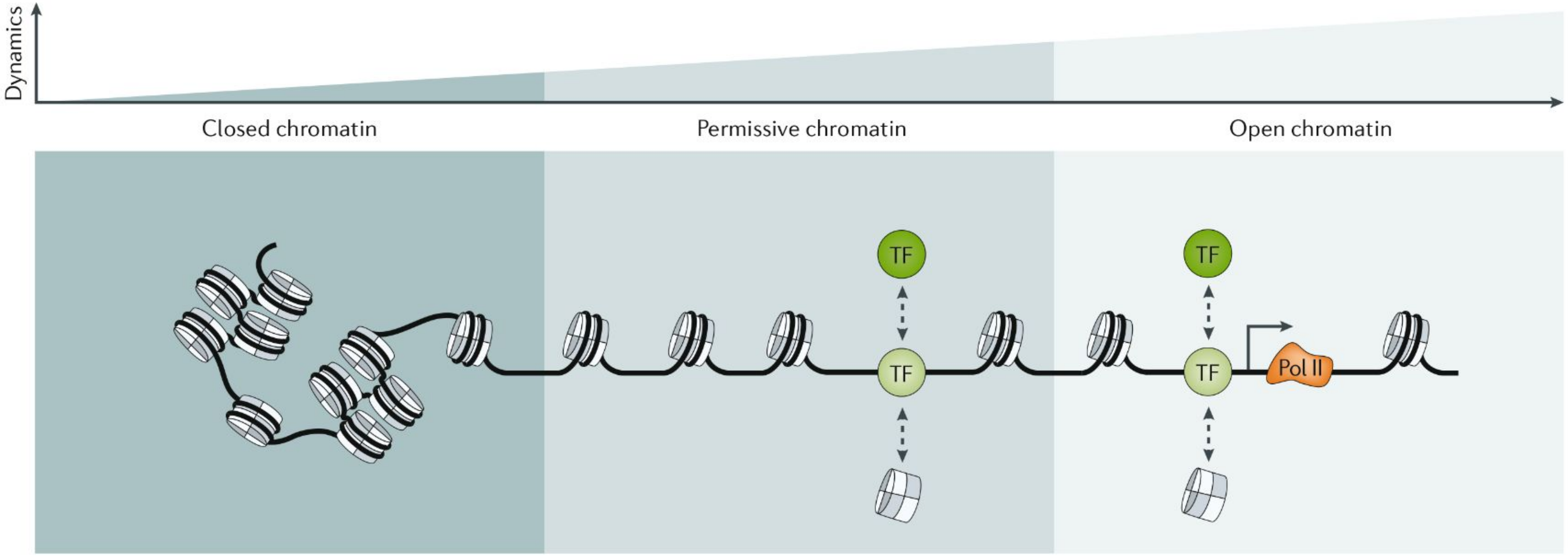


Fig. 1 | **A continuum of accessibility states broadly reflects the distribution of chromatin dynamics across the genome.** In contrast to closed chromatin, permissive chromatin is sufficiently dynamic for transcription factors to initiate sequence-specific accessibility remodelling and establish an open chromatin conformation (illustrated here for an active gene locus). Pol II, RNA polymerase II; TF, transcription factor.

(Klemm, Shipony and Greenleaf, 2019)

Returning to our very brief history of genetics & genomics

…

1900 - Rediscovery of Mendel's work (1860s)

1913 - Chromosomes are linear arrays of genes

1941 - the one-gene-one-enzyme hypothesis

1944 - DNA is the genetic material

1951 - First protein sequenced

1977 - DNA sequencing

1977 - Eukaryotic genes are spliced

1995 - First bacterial genomes sequenced

2000 - Next Generation Sequencing (NGS)

2001 - Draft of the human genome

2003 - RNA-seq

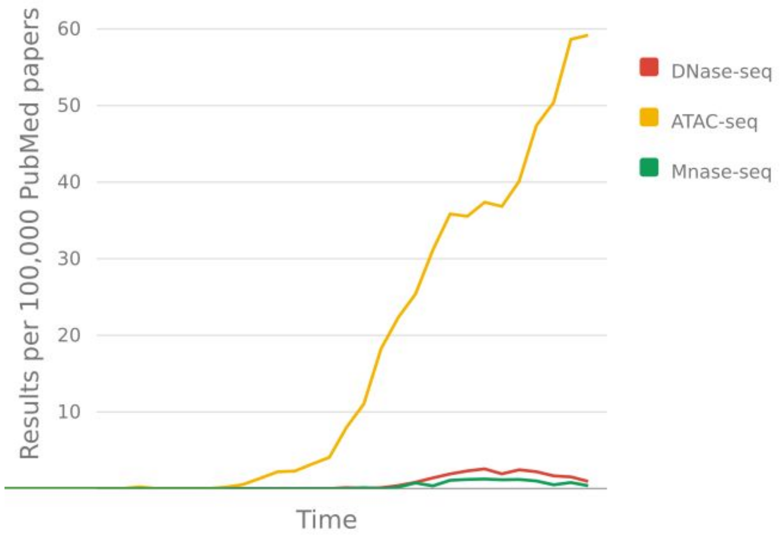
2006 - ChIP-seq

2008 - DNAse-seq, MNase-seq

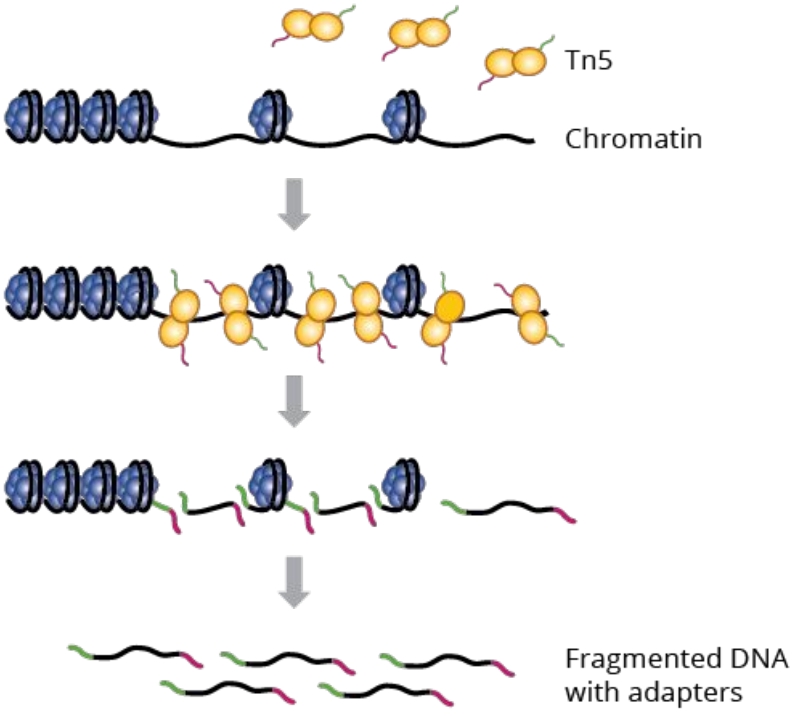
2012 - ATAC-seq

…

Accessibility assays

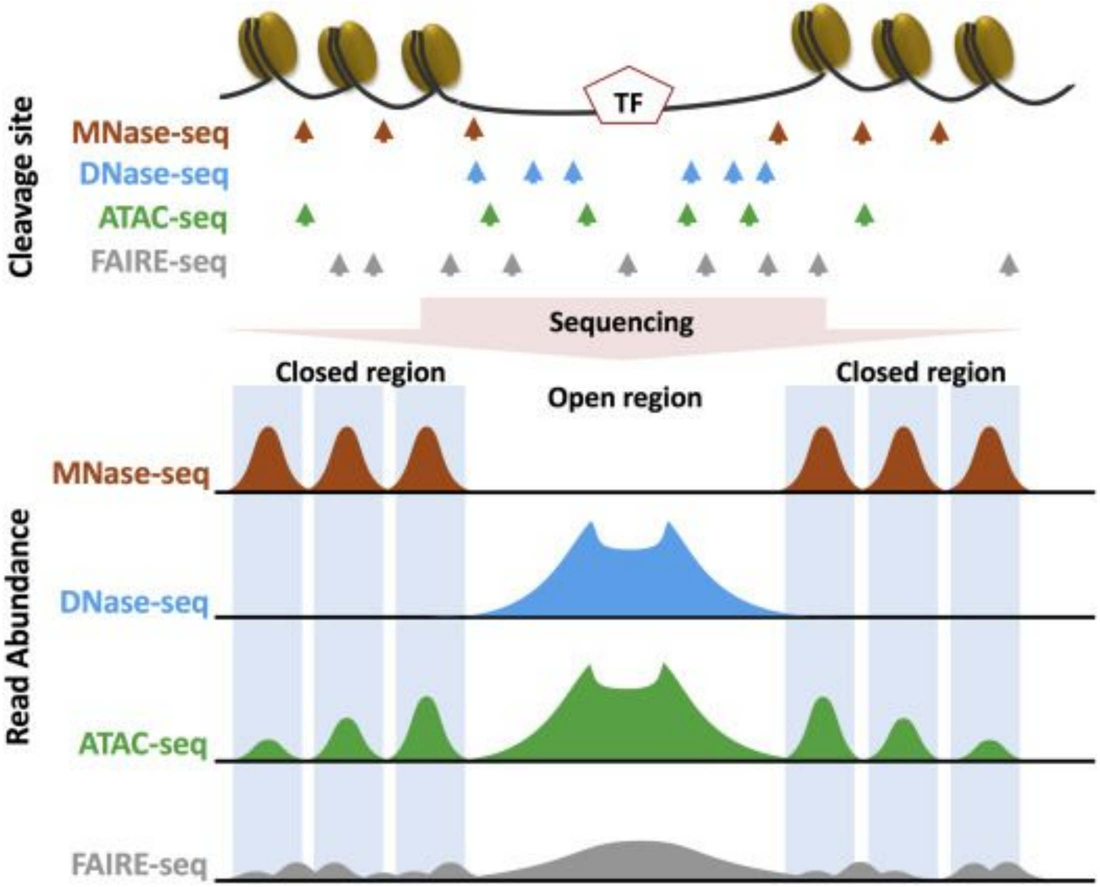


ATAC-seq



ATAC-seq recently became extremely popular due to its information content and low material requirement (i.e. # cells)

Chromatin accessibility assays

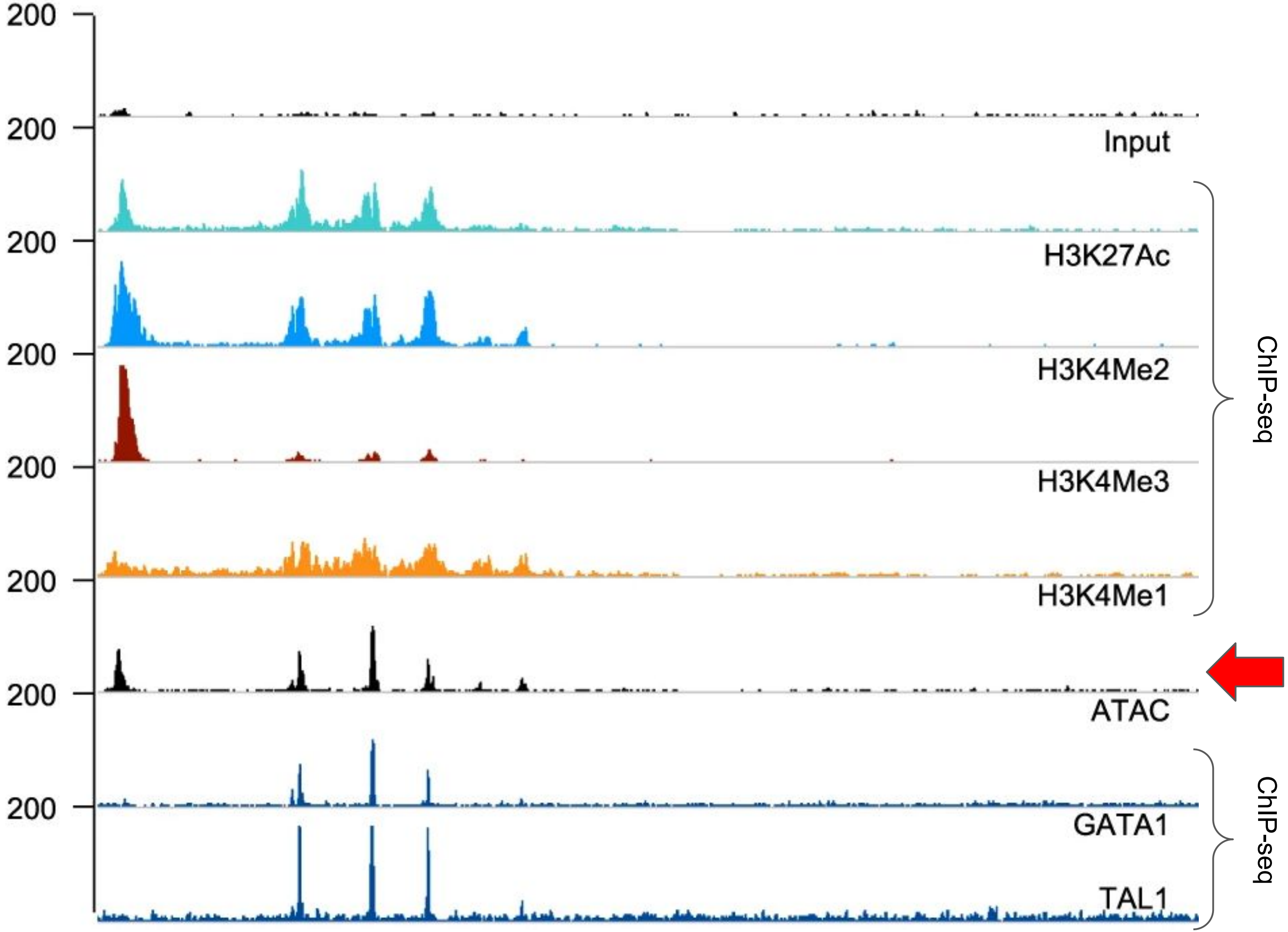


→ Nucleosome positions

→ DNase1-hypersensitive sites (DHS)
→ regulatory elements

} Both

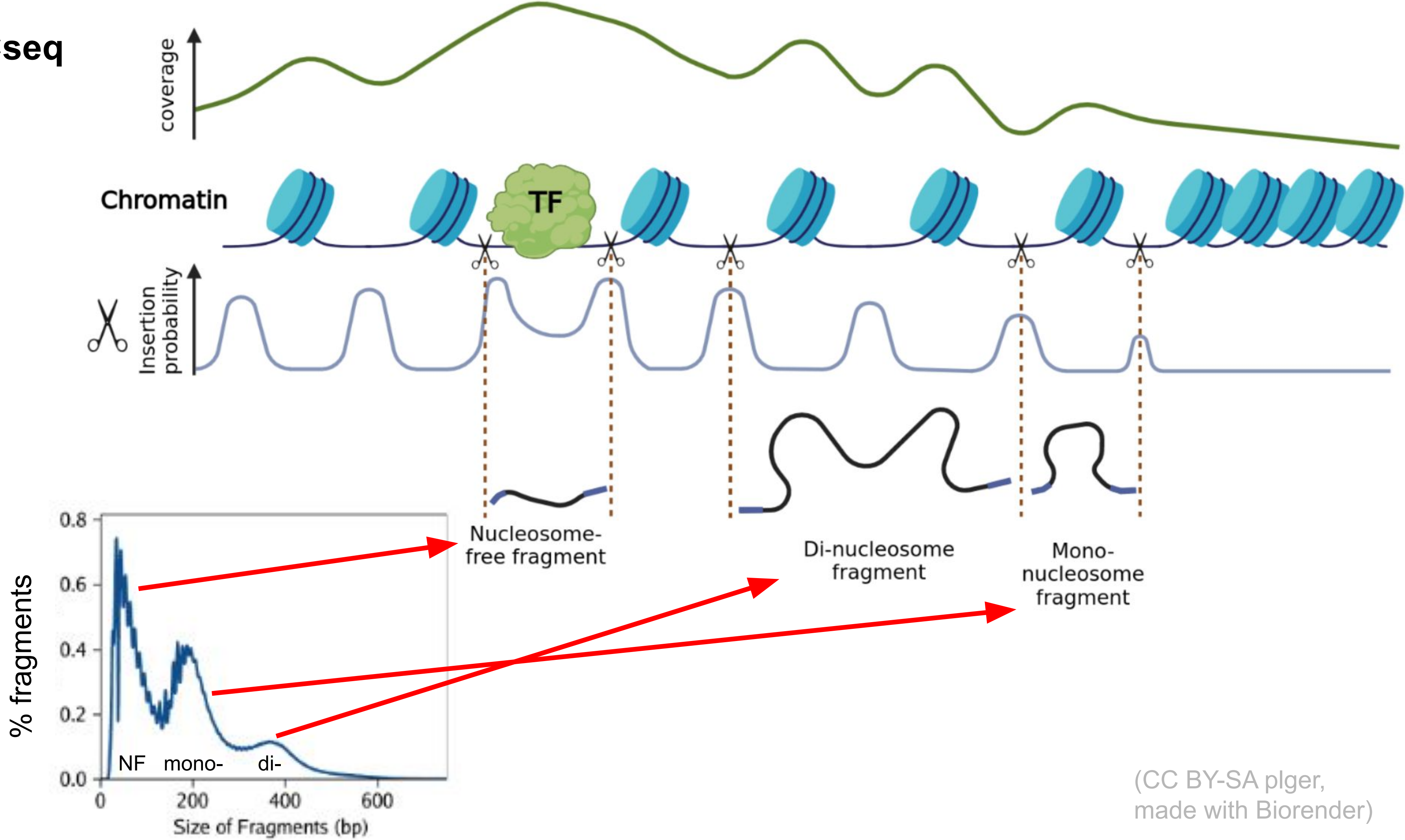
(Hsu et al., Epigenetics in Human Disease, 2018)

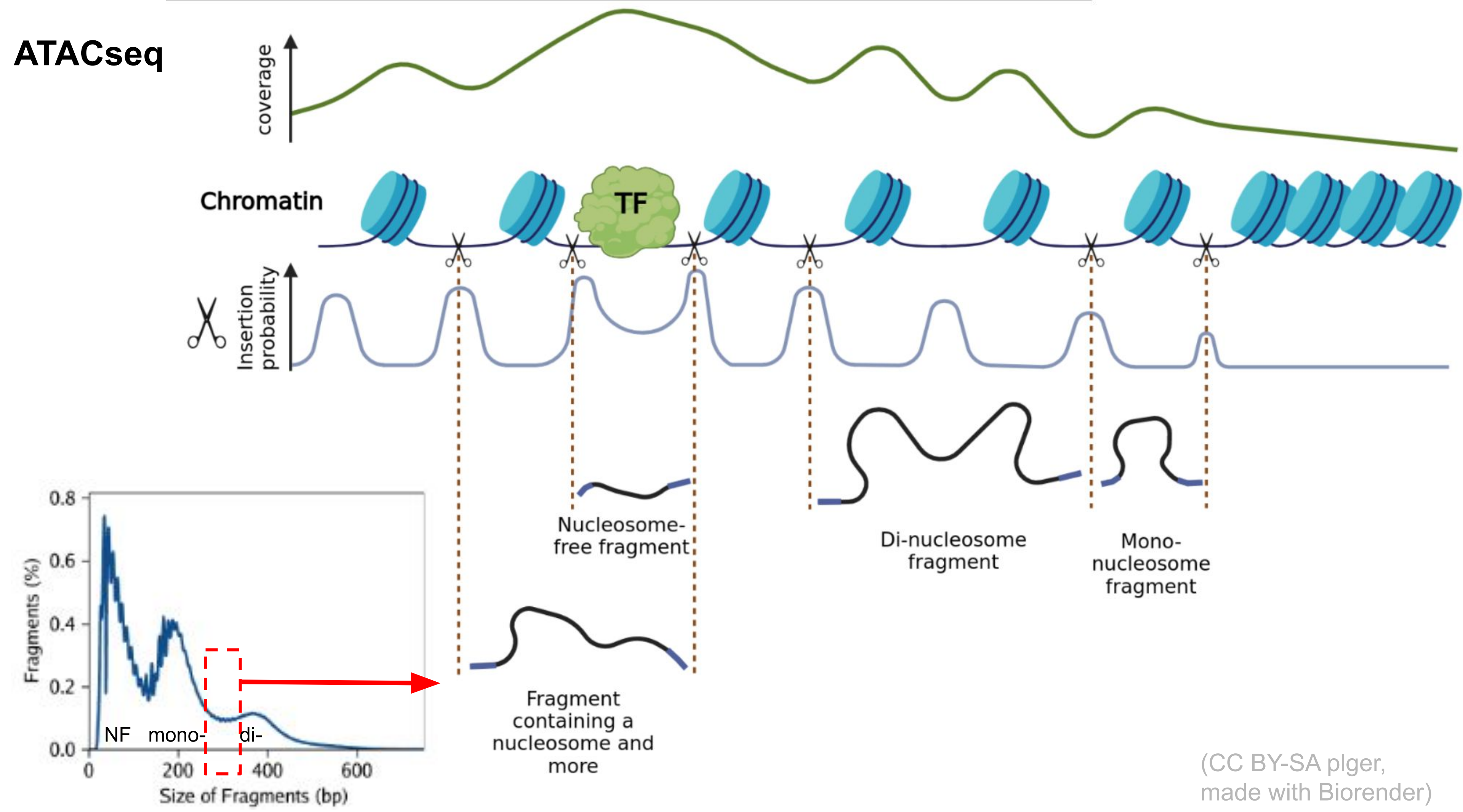


ATAC signal tells us that something is happening, it just doesn't tell us what exactly…

(Adapted from Fox et al., Nat Comm 2020)

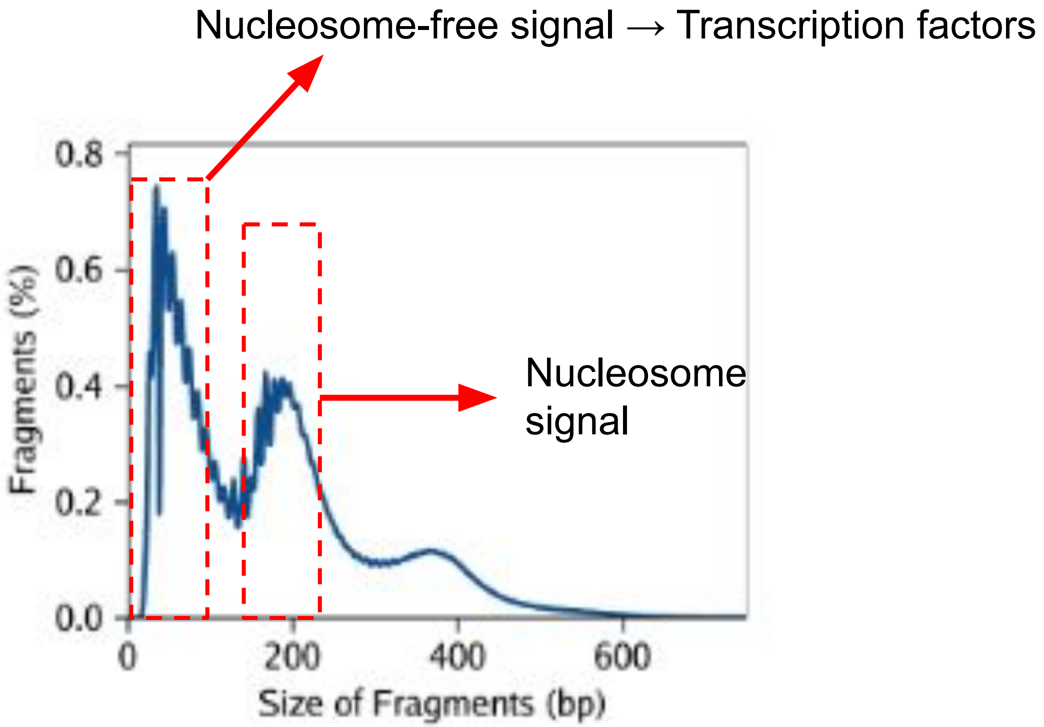
ATACseq


coverage
Chromatin
TF
Insertion probability
Nucleosome-free fragment
Di-nucleosome fragment
Mono-nucleosome fragment
% fragments
0.8
0.6
0.4
0.2
0.0
0
200
400
600
Size of Fragments (bp)
NF
mono-
di-



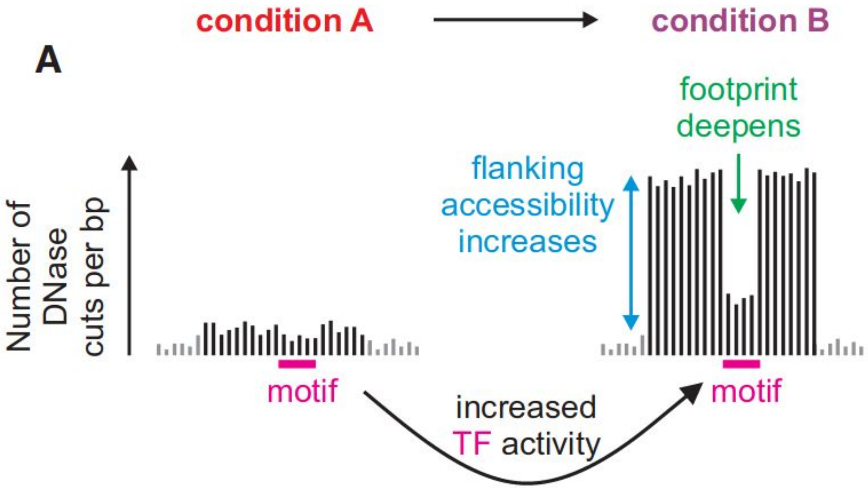
ATACseq
coverage
Chromatin
TF
Insertion probability
Nucleosome-free fragment
Di-nucleosome fragment
Mono-nucleosome fragment
Fragment containing a nucleosome and more
0.8
0.6
0.4
0.2
0.0
Fragments (%)
NF
mono-
di-
0
200
400
600
Size of Fragments (bp)

This means that once we have the data, we can split the fragments according to size in order to obtain specific information about different kinds of chromatin signals



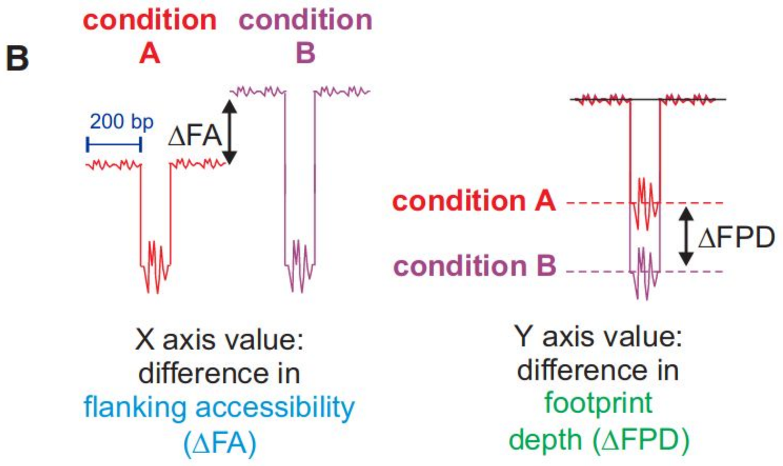
Practical

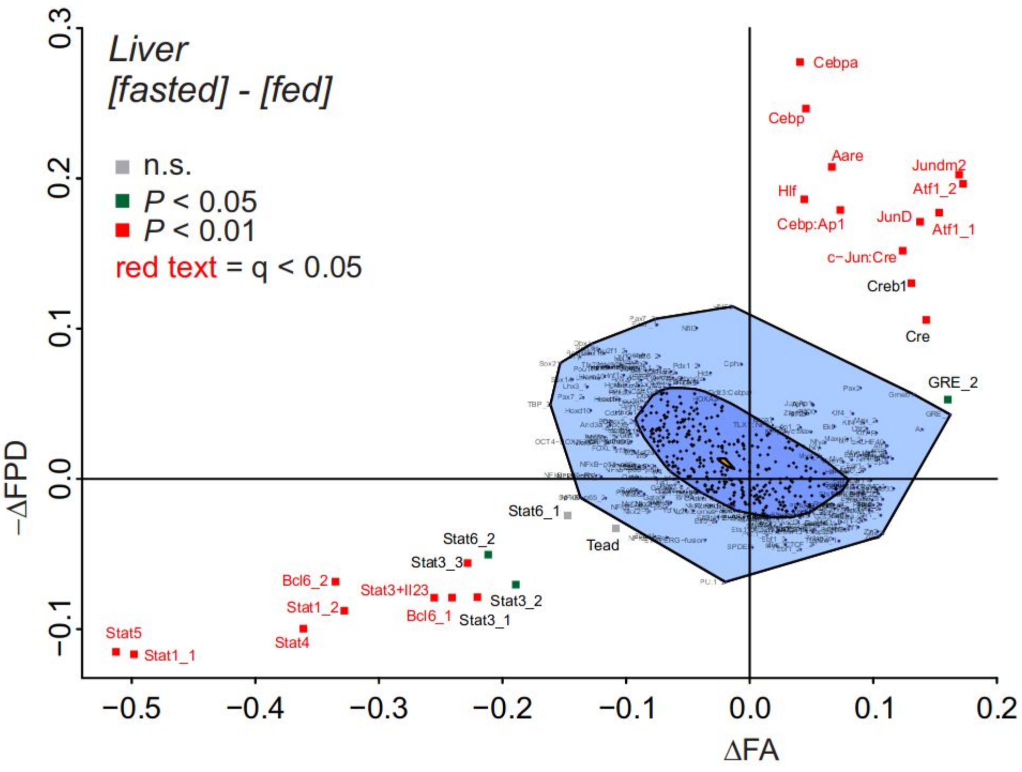
Estimating TF activity from accessibility and footprints



(Baek, Goldstein and Hager, Cell Reports 2017)

Estimating TF activity from accessibility and footprints





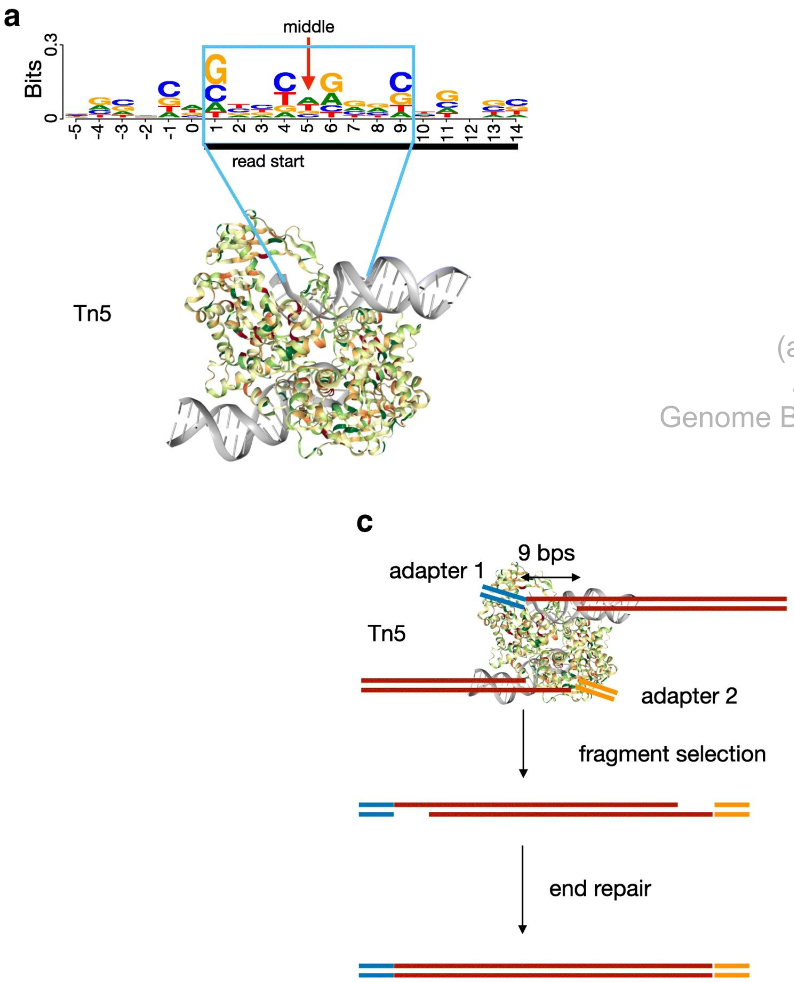
(Baek, Goldstein and Hager, Cell Reports 2017)

“Shifting” ATAC-seq alignments

From a given ATAC-seq insertion site, the exact region that is accessible is a few nucleotides from the start of the read

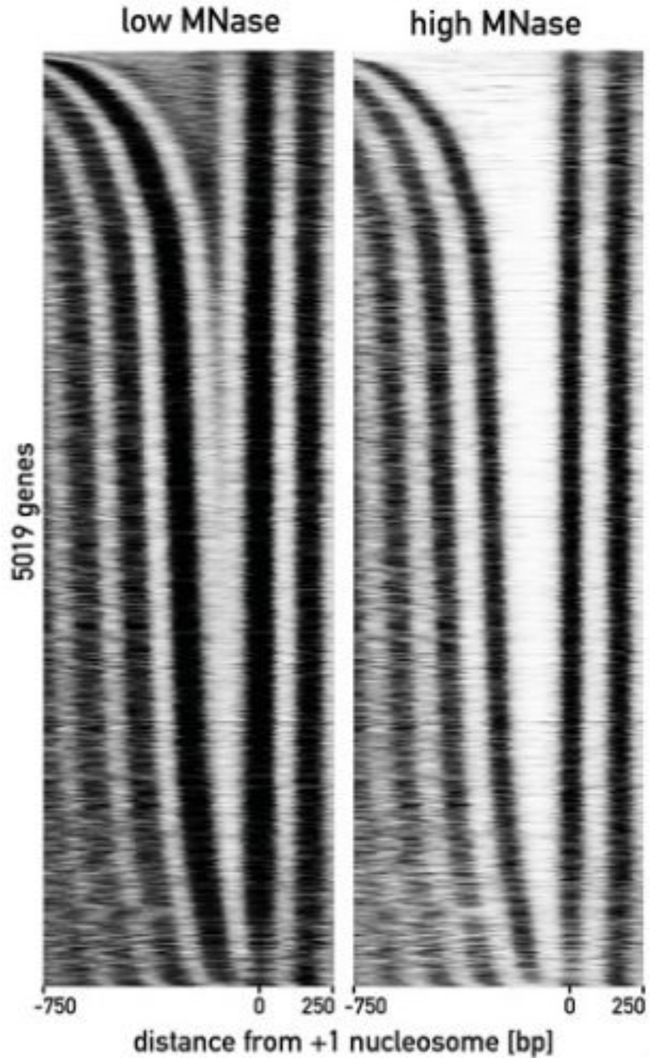
When doing high-resolution things like footprinting, one therefore typically shifts the cut sites by +4/-5nt, so that it is placed in the middle of where the Tn5 was binding

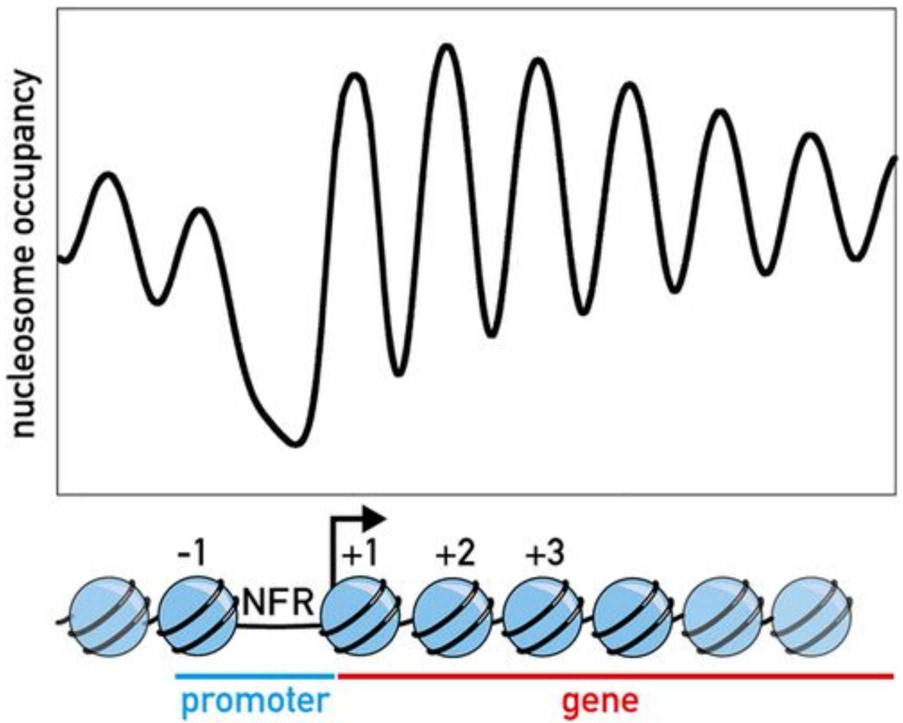
(For most other purposes, this is too fine-grained to make a difference)



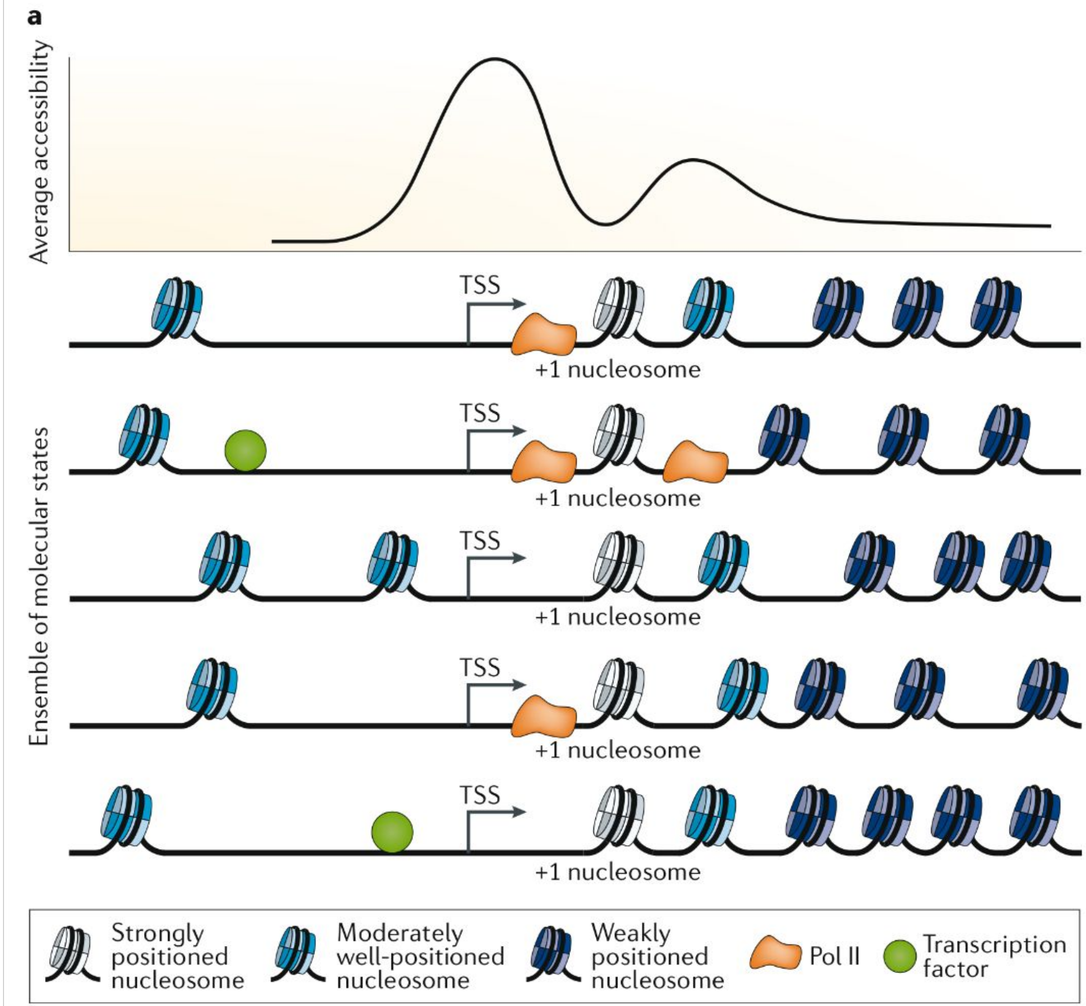
(adapted from Zhijian et al., Genome Biology 2019)

Nucleosome positioning





(Kubik, Bruzzone and Shore, 2017)



(Klemm, Shipony and Greenleaf, 2019)

Due to holidays we're seeing each other next time on the **29th**!

Assignment

In the same dataset of ATAC on chr19, plot the insertion (i.e. 'cuts') profile of, respectively, nucleosome-free and nucleosome-containing fragments, around the high-confidence motifs of two factors.
You can choose your own factors of interest, or for instance use REST and the glucocorticoid receptor (search "GCR")

Expected form of the answer: 2 figures (one for each factor),
each containing the heatmaps of the two signals around the motifs

Don't forget to render your markdown and push it as `assignment.html` !