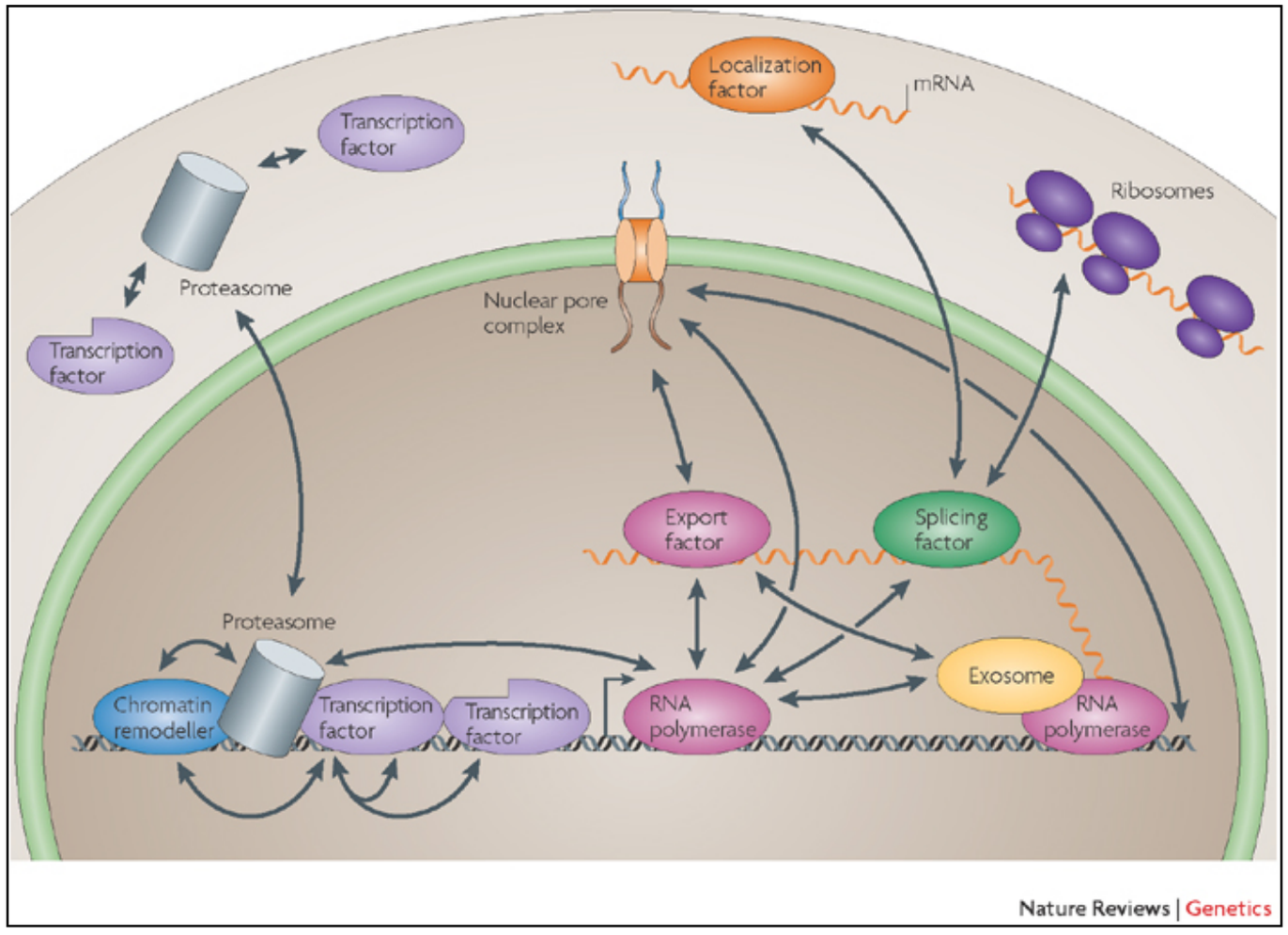
BIO-463
Genomics and bioinformatics

Lecture 11: Transcriptional regulation

Dr Jacques Rougemont

EPFL

Protein-DNA interactions

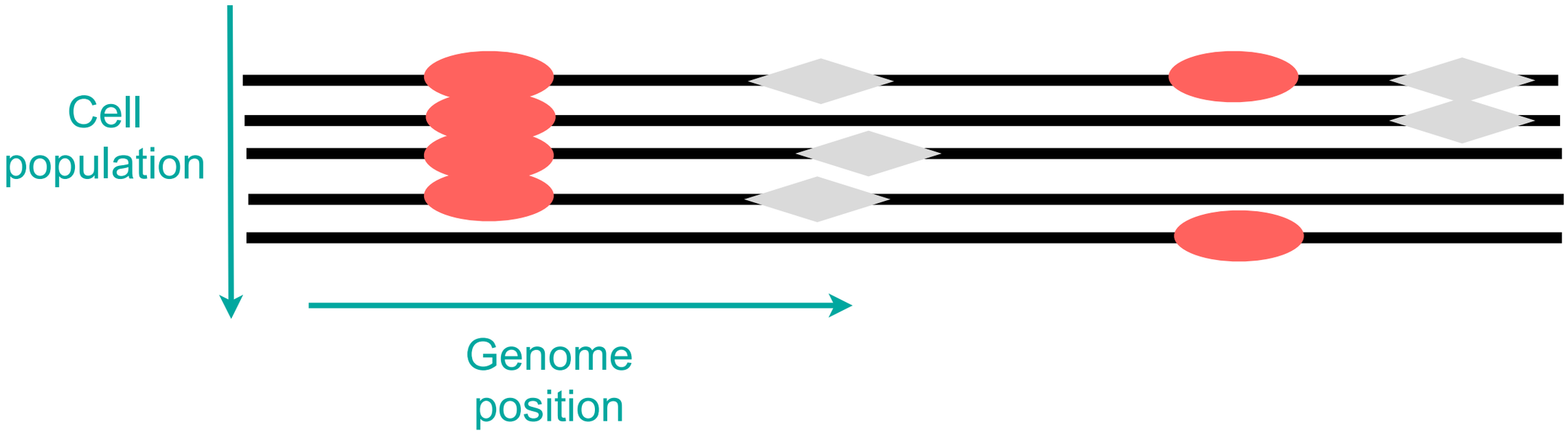


Komili & Silver 2008

- Gene regulation occurs via interaction of DNA with protein complexes
- There is specific binding (transcription factors), indirect binding (co-factors), unspecific binding (Polymerase, histones)
- These can be studied with high-throughput genomic techniques

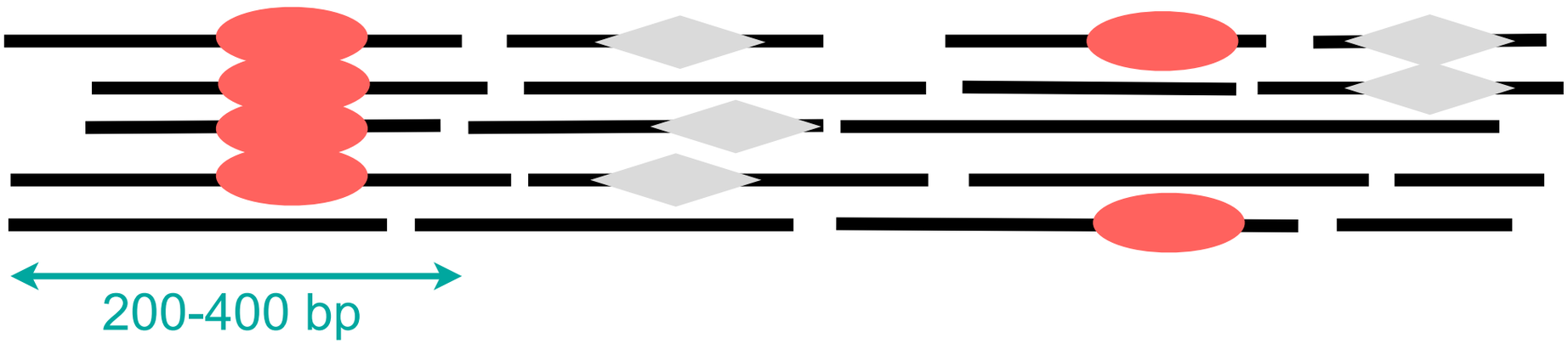
ChIP-Seq: method

1) Cross-link Proteins+DNA



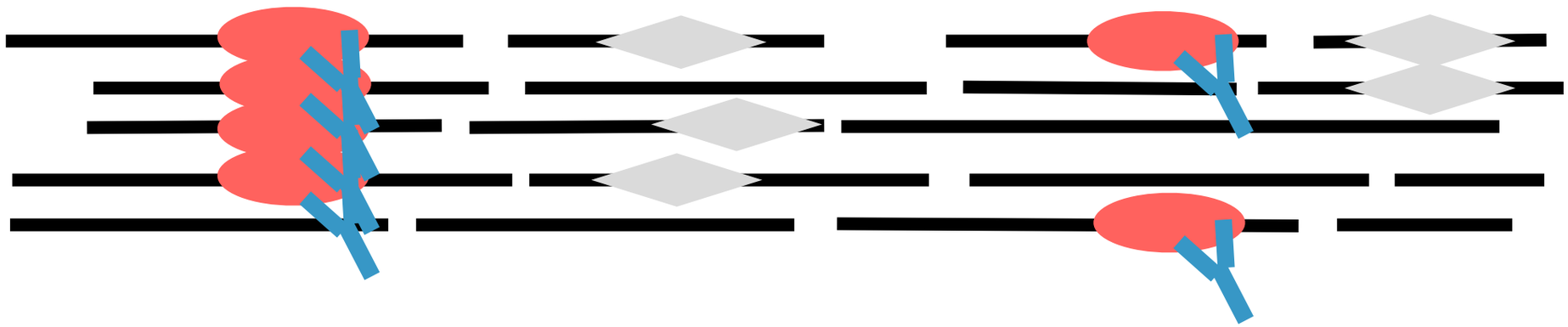
ChIP-Seq: method

2) Sonicate (or digest)



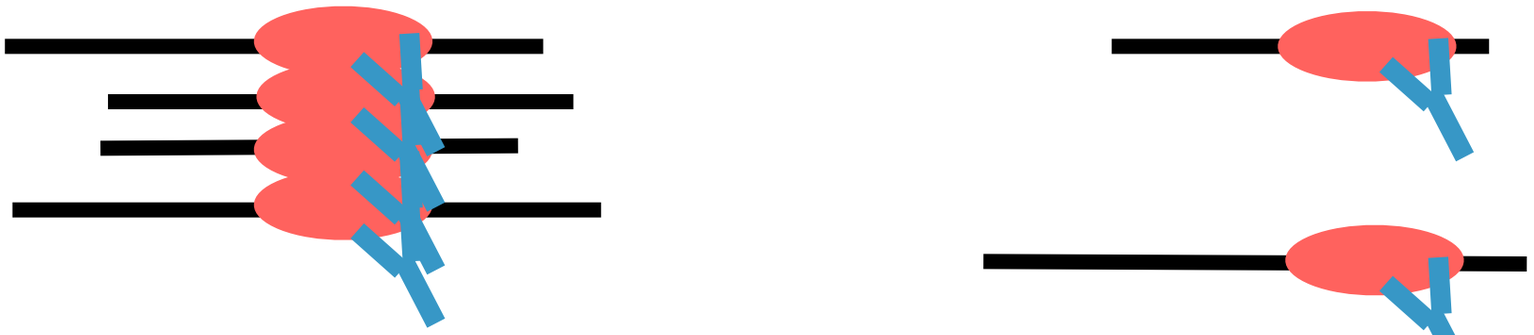
ChIP-Seq: method

3) ImmunoPrecipitate



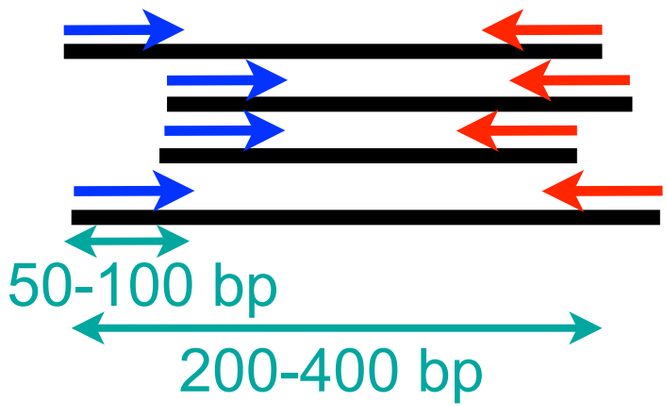
ChIP-Seq: method

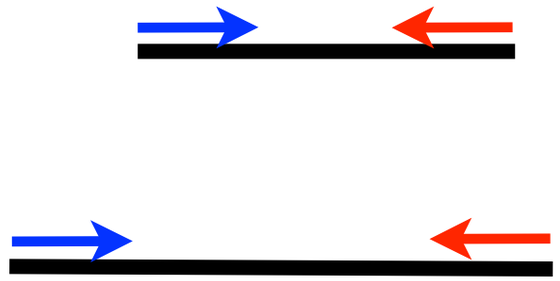
4) Reverse cross-links



ChIP-Seq: method

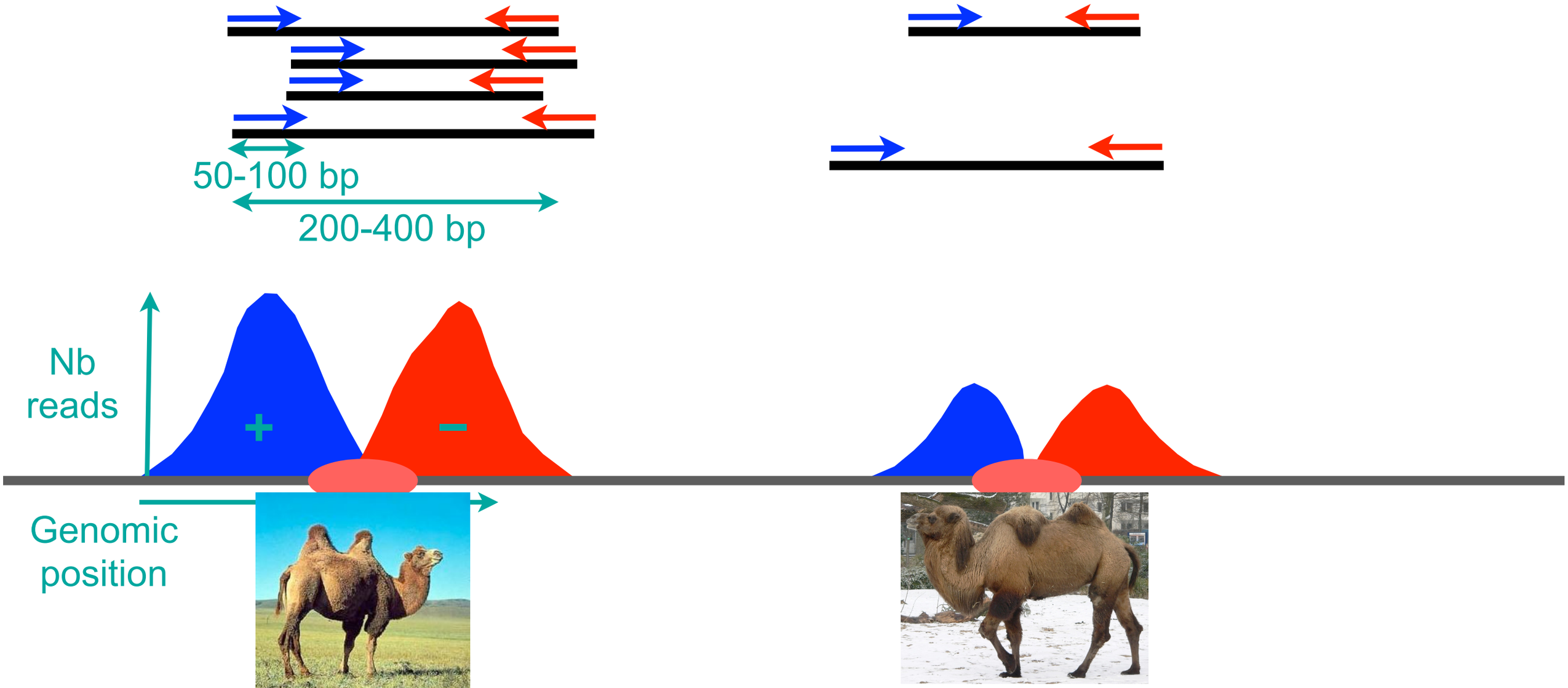
5) Sequence dsDNA (short read 5’ of either strand)





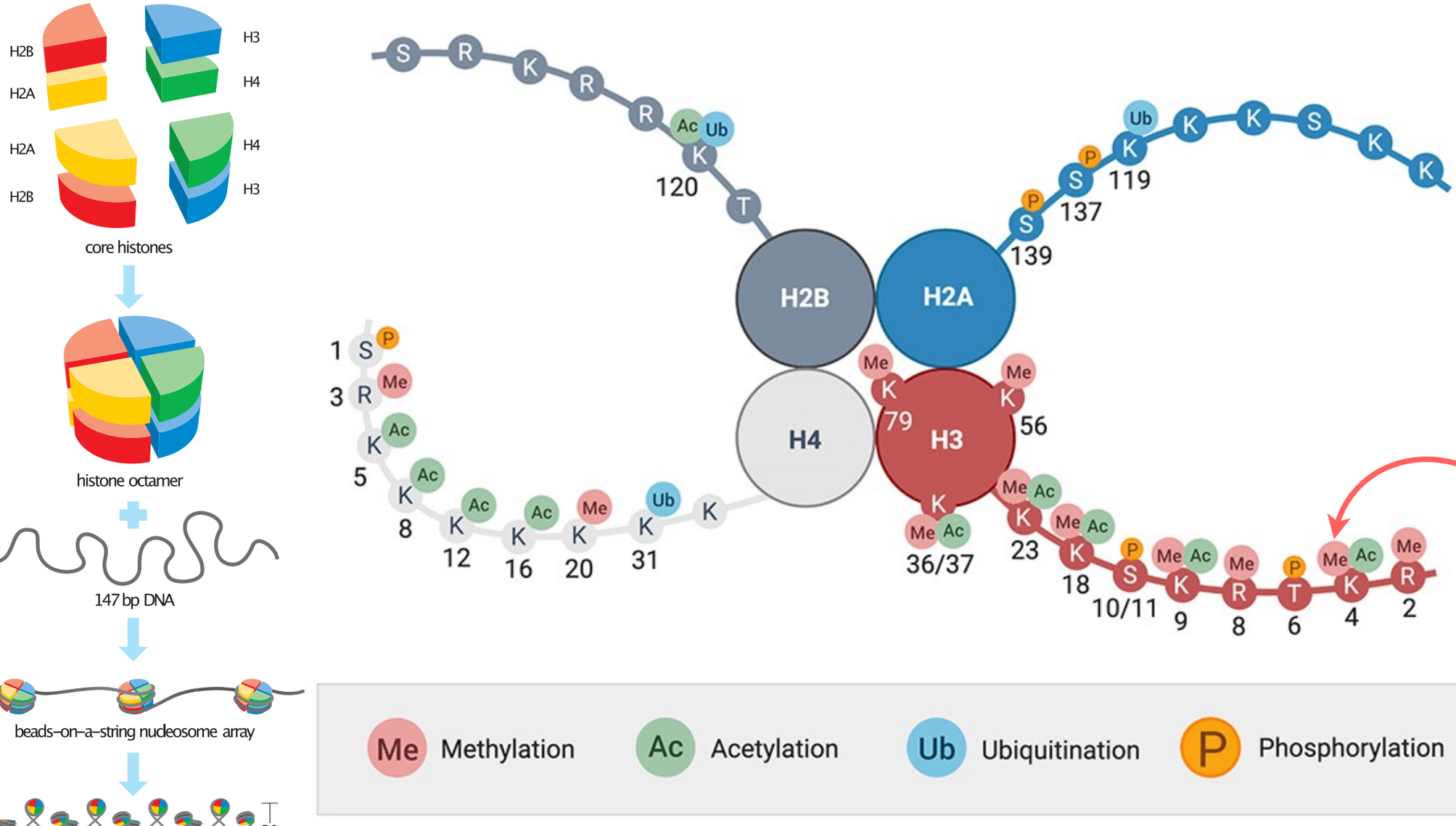
ChIP-Seq: method

6) Map reads to reference sequence



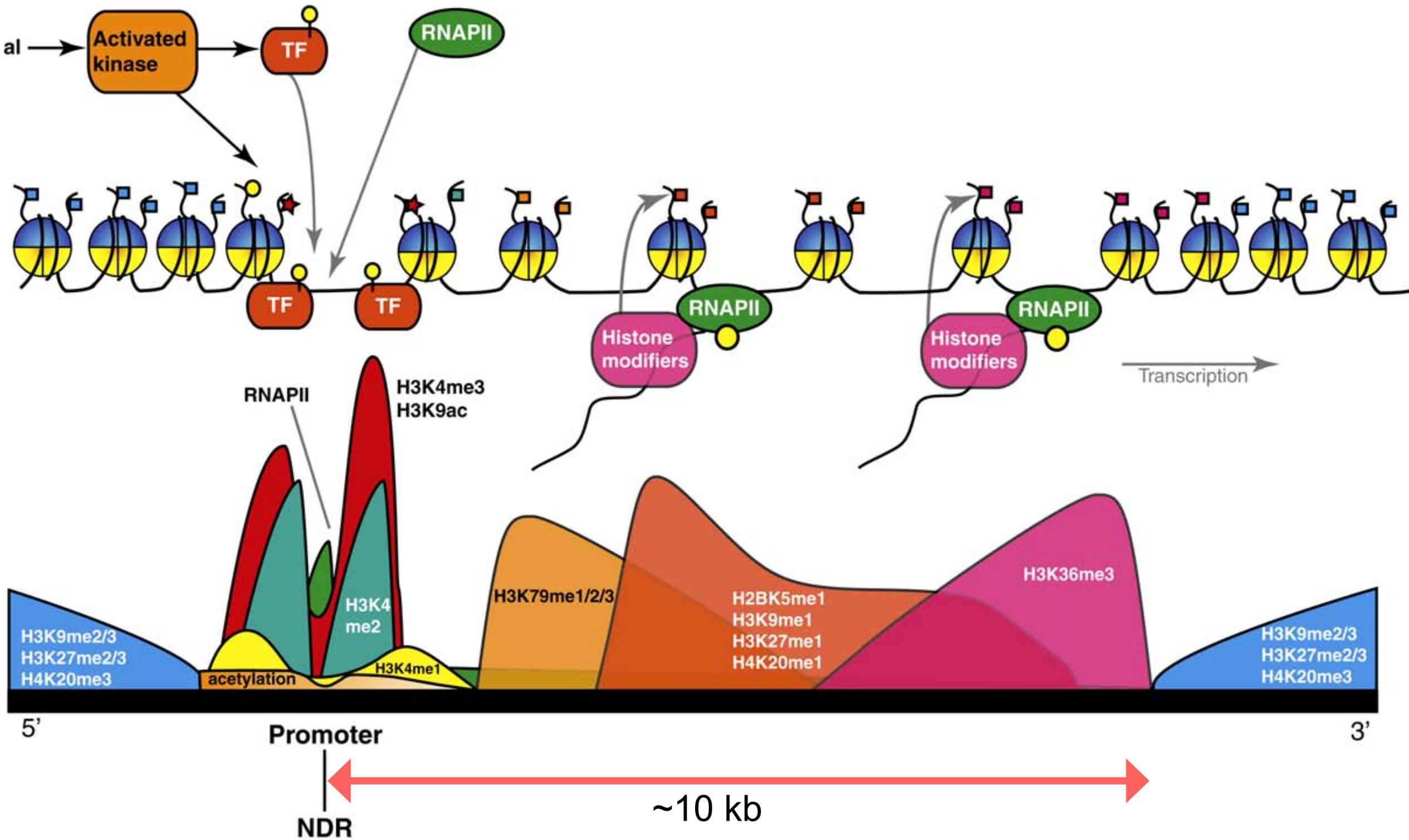
Histone modifications

Chromatin state reflects transcriptional history, modification-specific antibodies can be used



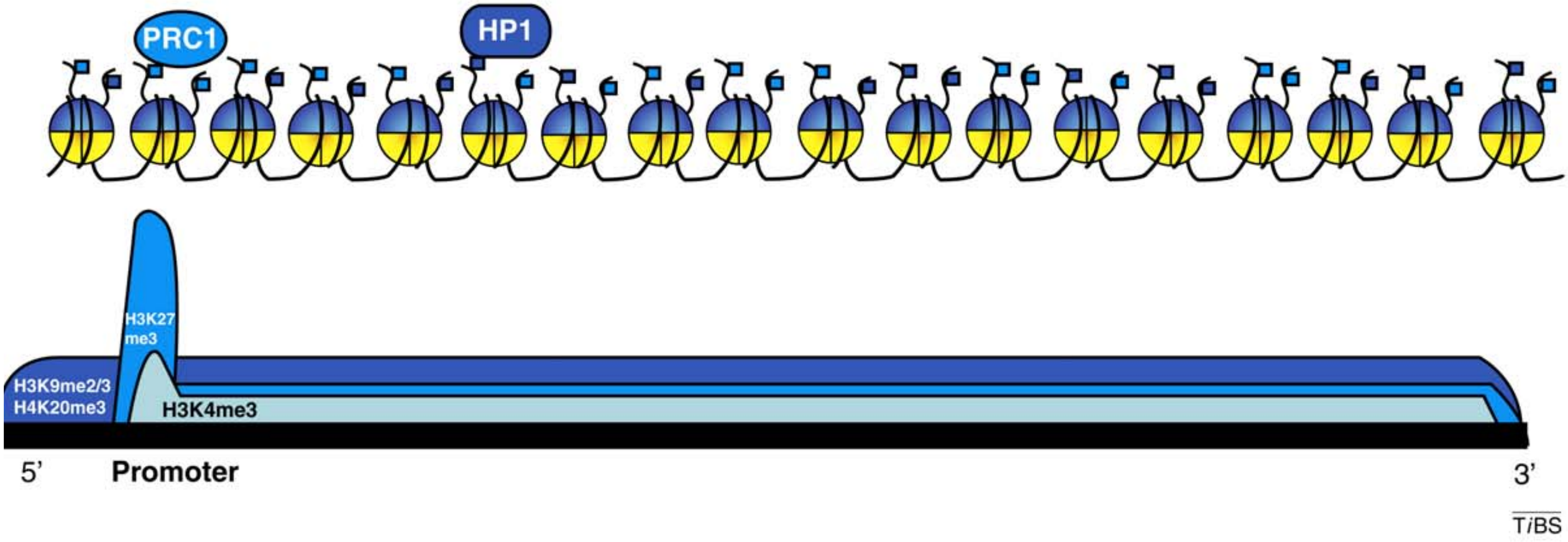
ex.: **H3K4me3** - Lysine (K) at pos. 4 of Histone H3 is 3-methylated

ChIP profiles (active gene)



Barth TK, Imhof A. (2010)

ChIP profiles (inactive gene)

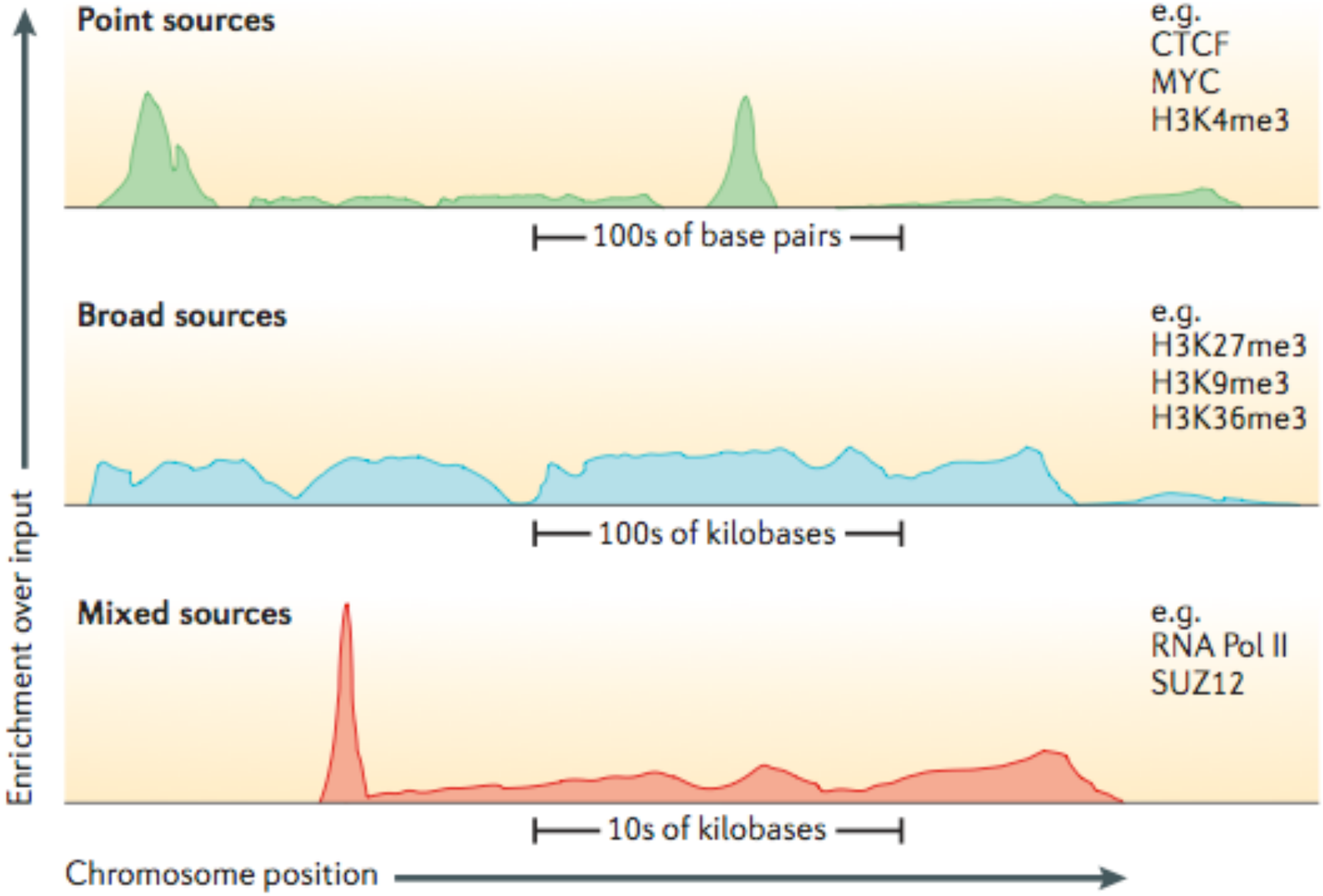


TiBS

Barth TK, Imhof A. (2010)

ChIP profiles (inactive gene)

- In general: signal at a genomic position is proportional to **fraction of cells** having the protein bound at this position
- For travelling proteins (e.g. PolII) this is proportional to **residency time** (inverse of speed): population average is the same as time average
- For sequence-specific binding, this is related in a non-linear way to **binding affinity**

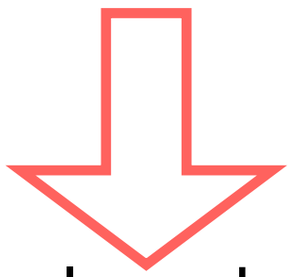


Sims D et al. Nat Rev (2014)

DNA fragment distribution

Genome size: $3 \cdot 10^9$

fragment size: ≈ 300 $\Rightarrow$ nb fragments per cell: $\approx 10^7$



Typical number of (occupied) transcription factor binding sites: $\approx 10^4$.

Antibody “enrichment ratio”: $\approx 100 \times$ (bound fragment is 100 times more likely to be selected than control)

TF-bound fragments are 1/1000 in input, hence 1/10 in IP

$\Rightarrow$ false positives $\approx 90\,\%$.

Starting material is $\approx 10^7$ cells, sequencing throughput is $\approx 10^8$ reads

Consequence: each protein-bound fragment comes from a different cell

But: 0.1×10^8 reads distributed over 10^4 fragments is 10^3 reads per fragment

0.9×10^8 reads distributed over 10^7 fragments is 9 reads per fragment

Controls

Non-specific reads are spread throughout the genome, but not uniformly.

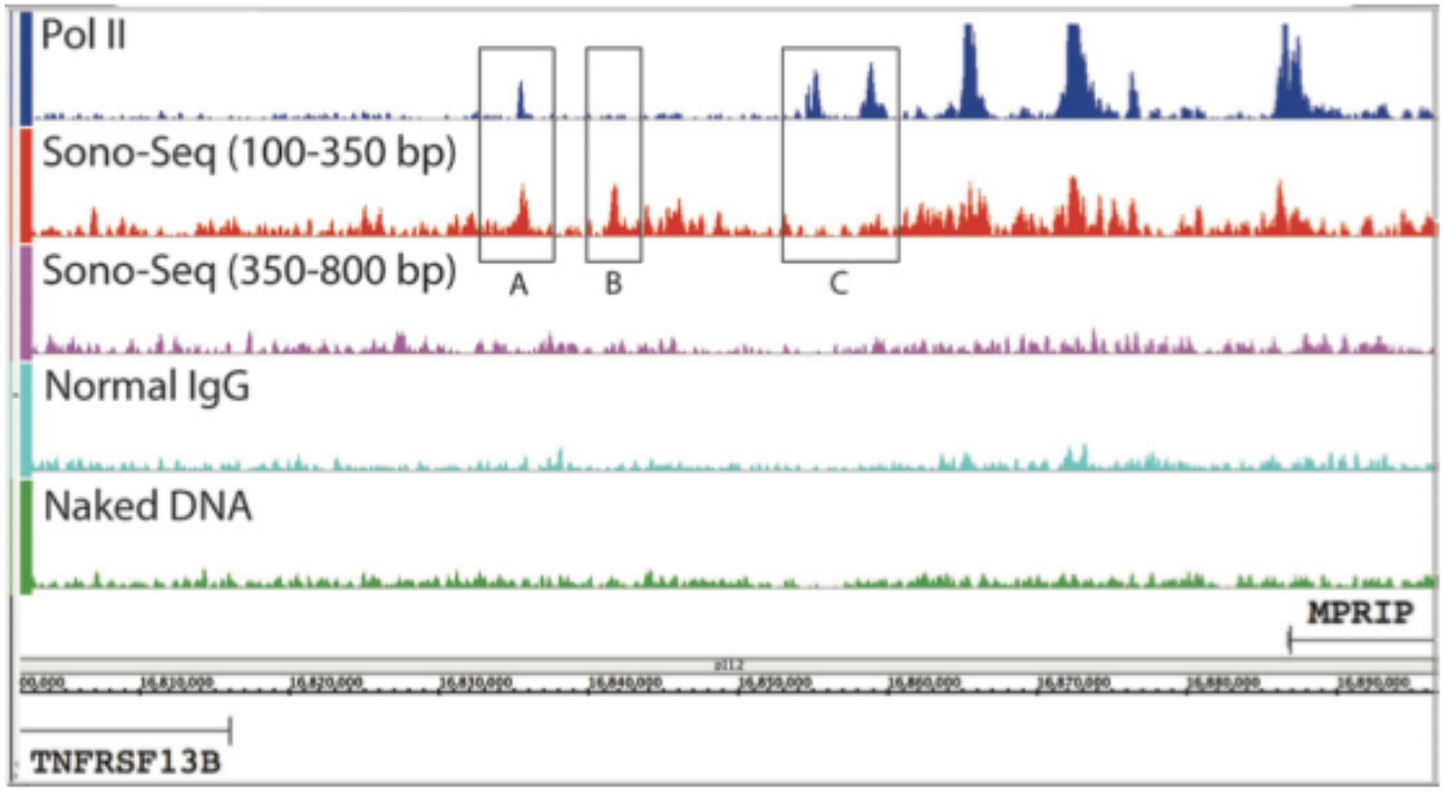
To detect false positives, several techniques are routinely used:

ChIP-seq

Input (cross-linked sonicated DNA)

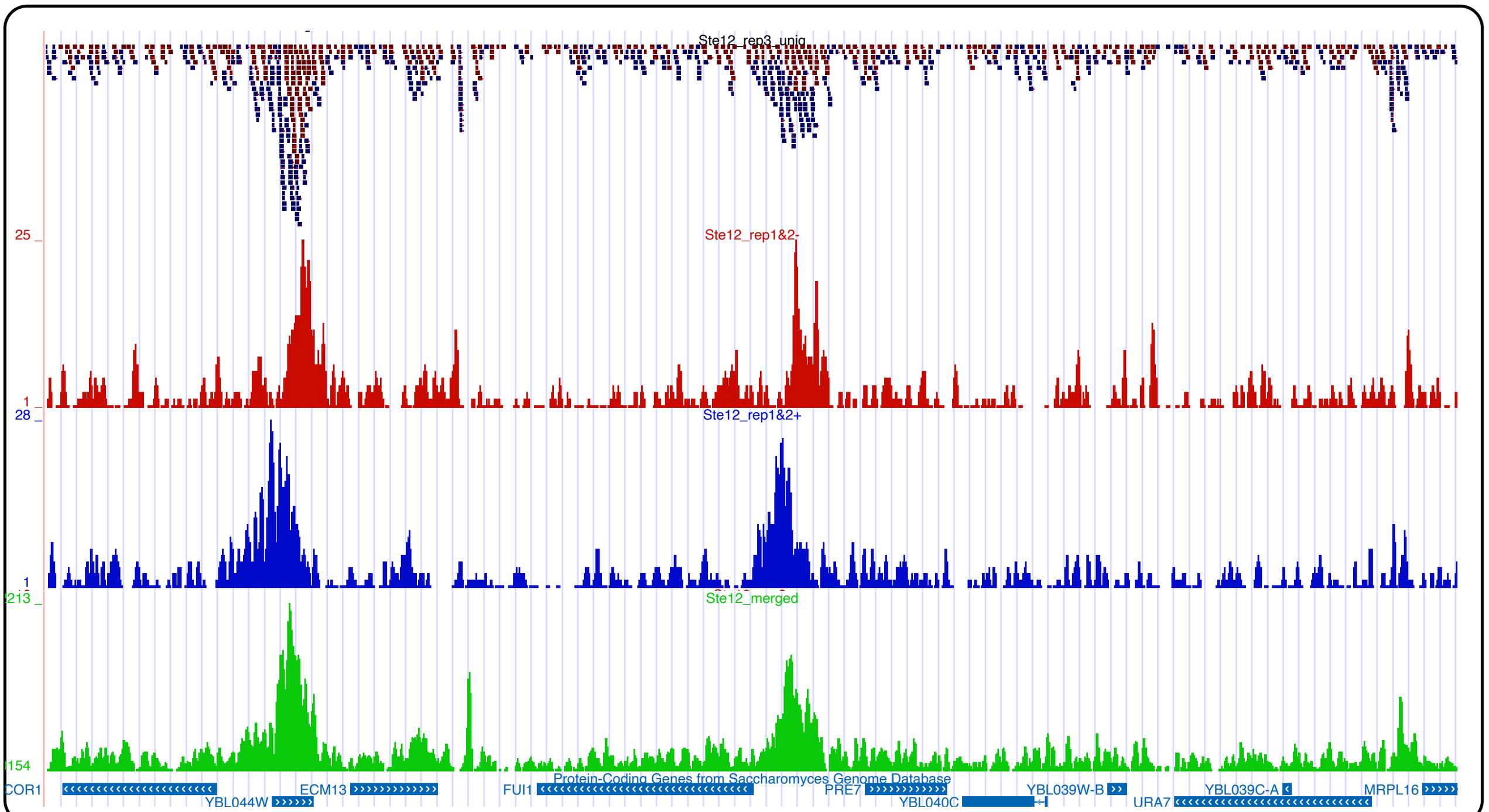
"mock" IP (unspecific antibody)

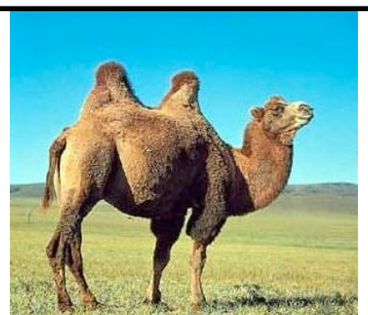
genome sequencing



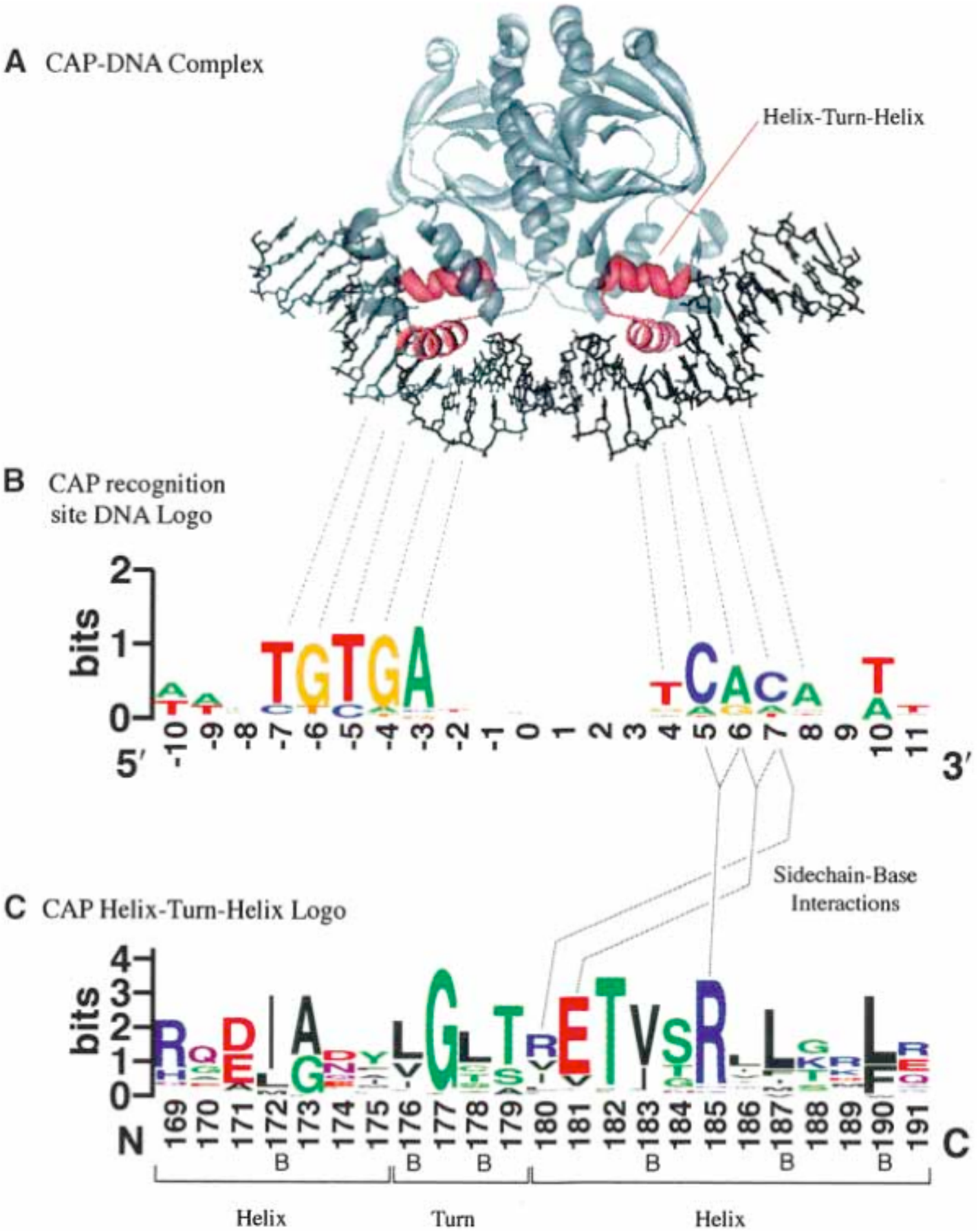
Auerbach et al. (2009)

Binding regions have characteristic peak shape





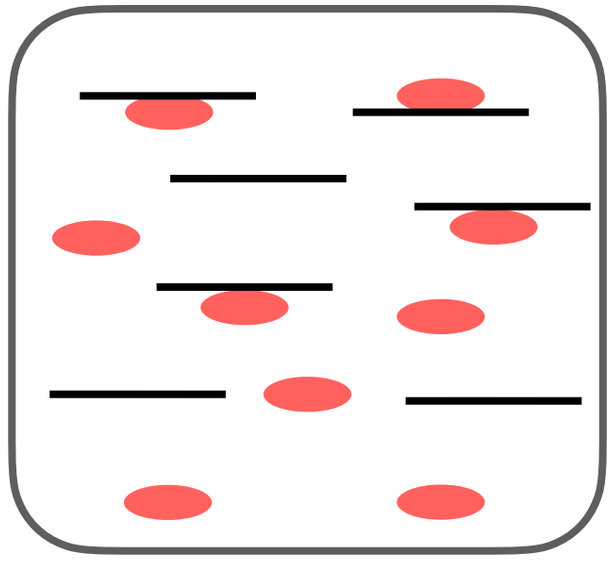
Sequence-specific DNA binding

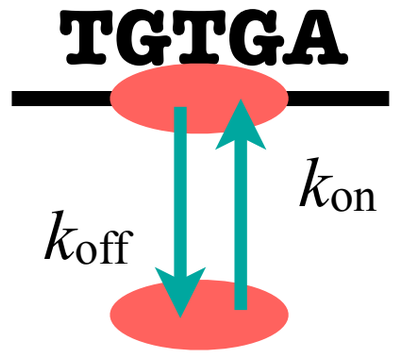


Sequence-specific occupancy

DNA binding proteins have a sequence-dependent binding energy $G(S)$:

$$\frac{[P \cdot S]}{[P][S]} = e^{-\beta G(S)}$$



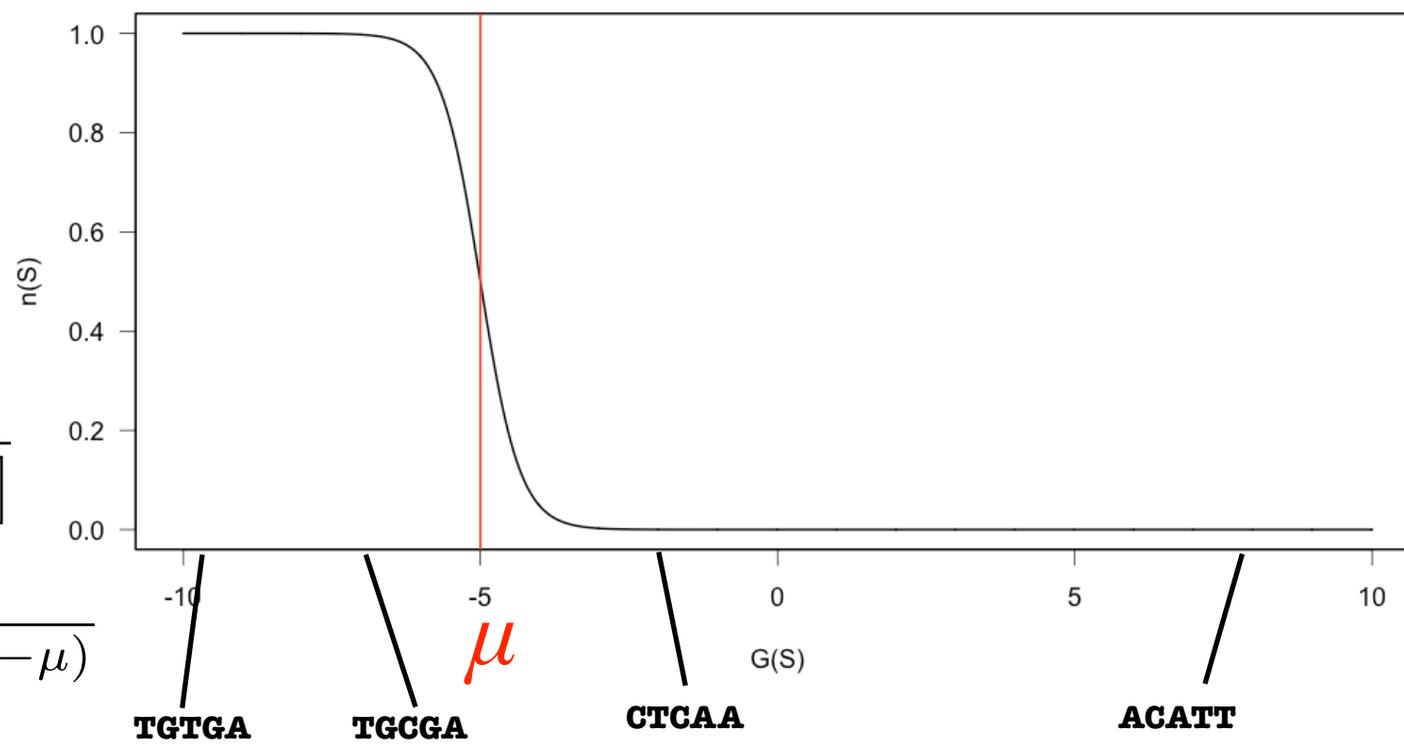


$$\frac{k_{\text{on}}}{k_{\text{off}}} = \frac{[P \cdot S]}{[P][S]}$$

Occupancy $n(S)$ is a non-linear (monotone) function of energy and protein concentration

$$n(S) = \frac{[P \cdot S]}{[P \cdot S] + [S]} = \frac{1}{1 + [S]/[P \cdot S]}$$

$$= \frac{1}{1 + e^{\beta G(S)}/[P]} = \frac{1}{1 + e^{\beta(G(S) - \mu)}}$$



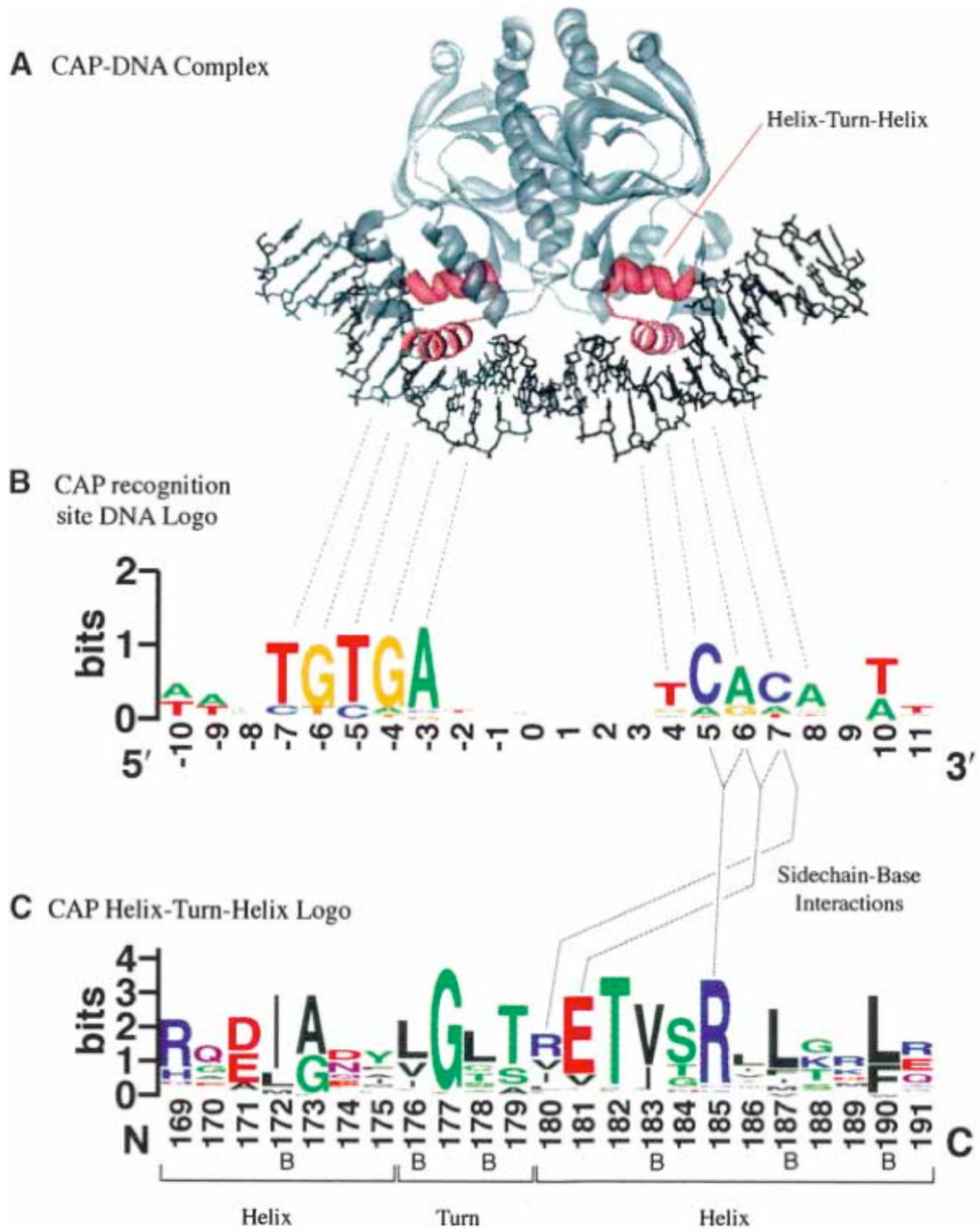
Sequence-specific affinity

We assume binding via L consecutive bases, each bond contributes an independent additive weight:

$$G(S) = G_0 - \sum_{k=1}^{L} g(k, S_k)$$

$$e^{-\beta G(S)} = e^{-\beta G_0} \prod_{k=1}^{L} e^{\beta g(k,S_k)}$$

$$= Z_0 \prod_{k=1}^{L} W(k, S_k)$$

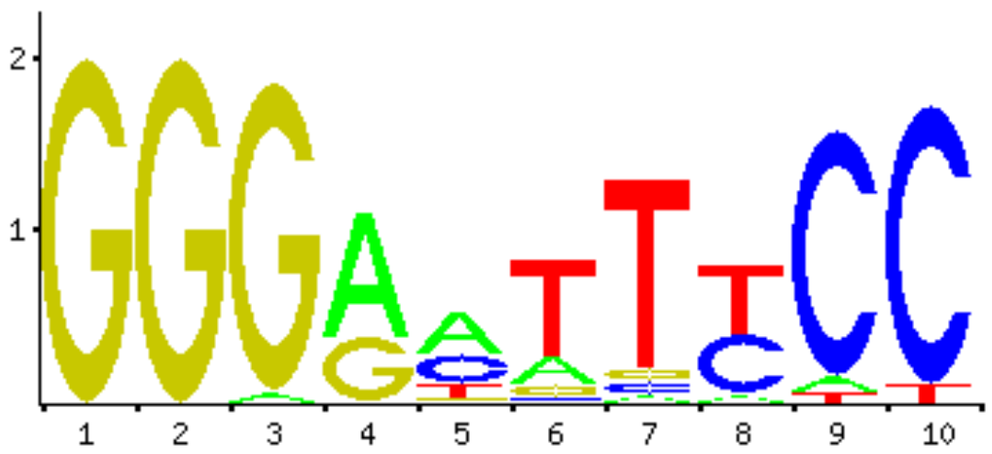


Sequence-specific affinity

In this approximation, the binding affinity is represented by Position-Weight Matrices (PWM):

$$W_{k\alpha} = \begin{pmatrix} 0.000000 & 0.000000 & 1.000000 & 0.000000 \\ 0.000000 & 0.000000 & 1.000000 & 0.000000 \\ 0.026316 & 0.000000 & 0.973684 & 0.000000 \\ 0.657895 & 0.000000 & 0.342105 & 0.000000 \\ 0.500000 & 0.342105 & 0.026316 & 0.131579 \\ 0.184211 & 0.026316 & 0.078947 & 0.710526 \\ 0.026316 & 0.052632 & 0.052632 & 0.868421 \\ 0.052632 & 0.447368 & 0.000000 & 0.500000 \\ 0.052632 & 0.921053 & 0.000000 & 0.026316 \\ 0.000000 & 0.947368 & 0.000000 & 0.052632 \end{pmatrix}$$

k (1, 2, ⋮, L=10) — α (A, C, G, T)



consensus: GGGAATTTCC

Relation between energy, frequency and score

There are 2 kinds of PWM:

- Position-Probability Matrix (PPM in units of probability/frequency: W)
- Position-Specific Scoring Matrix (PSSM in units of relative energy: g)

	A	C	G	T	
PPM (absolute prob)	0.184211	0.000000	0.105263	0.710526	← sum=1
PPM (relative to consensus)	0.259260	0.000000	0.148148	1.000000	← max=1
PSSM (log of PPM)	-2.440569	-∞	-3.247930	-0.493041	
PSSM - constant	-1.947528	-∞	-2.754889	0.000000	

A strong binding site has:

- low k_{off}
- low (negative) G
- high (close to 1) PPM
- high PSSM ($g = -G$)

Motif score (protein affinity for a given sequence S) is calculated as

$$S(s_1 s_2 \ldots s_9 s_{10}) = \log_2 \left(\prod_{j=1}^{10} \text{PPM}(j, s_j) \right)$$

$$= \sum_{j=1}^{10} \text{PSSM}(j, s_j)$$

Sequence-specific affinity

Finding the matrix by maximum likelihood: data S is a set of protein-bound sequences

Sequence scoring is relative to a specific set of background frequencies $f(\alpha), \alpha = \mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}$

S $\quad k = 1,\ldots,L \rightarrow$

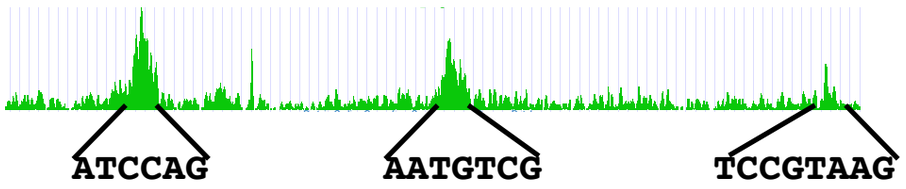
$n = 1,\ldots,N \downarrow$

```
GGGAATTTCC
GGGAATTCCC
GGGAATTCCC
GGGGATGTCC
GGGGCTTTCC
GGGACATTCC
GGGGAATTCC
GGAAATATCC
GGGAATTTCC
GGGAATCTCC
GGGAATCTCC
GGGAATTTCC
GGGAAATTCC
```

$$P(W|S) = \frac{P(S|W)P(W)}{P(S)}$$

$$\log\left(\frac{P(S|W)}{P(S)}\right) = \sum_n \sum_k \Big(g(k, S_{nk}) - \log f(S_{nk}) \Big)$$

EM algorithm



ATCCAG
AATGTCG
TCCGTAAG

Set of ChIP-seq enriched sites

Suppose each site contains one instance of a $L = 3$ motif

	A	C	G	T
1	0.27	0.27	0.15	0.31
2	0.30	0.30	0.23	0.17
3	0.25	0.25	0.30	0.20

PPM

score

find high scoring triplets

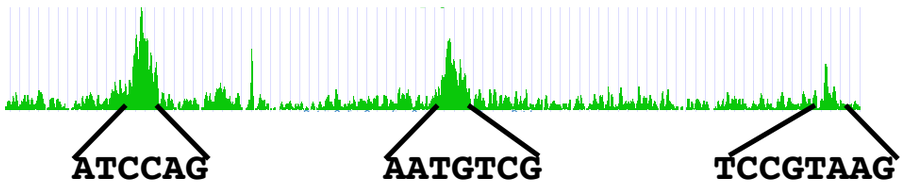
0.024
ATCCAG

0.028
AATGTCG

0.024 0.024
TCCGTAAG

meme-suite.org

EM algorithm



ATCCAG
AATGTCG
TCCGTAAG

Set of ChIP-seq enriched sites

Suppose each site contains one instance of a $L = 3$ motif

	A	C	G	T
1	0.5	1.5	0	1
2	1.5	1.5	0	0
3	0	0	3	0

PPM * 3

score

update

Average ties

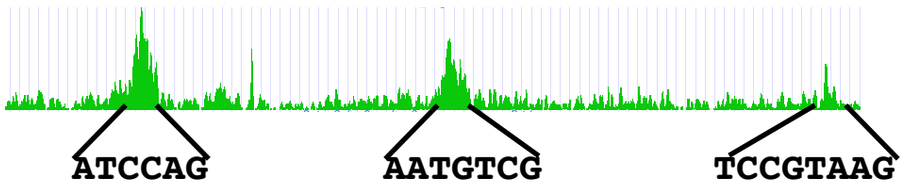
0.024
ATCCAG

0.011
AATGTCG

0.024 0.024
TCCGTAAG

meme-suite.org

EM algorithm



ATCCAG
AATGTCG
TCCGTAAG

Set of ChIP-seq enriched sites

Suppose each site contains one instance of a $L = 3$ motif

	A	C	G	T
1	0.75	1.75	0.25	1.25
2	1.75	1.75	0.25	0.25
3	0.25	0.25	3.25	0.25

PPM * 4

score

update

ATCCAG

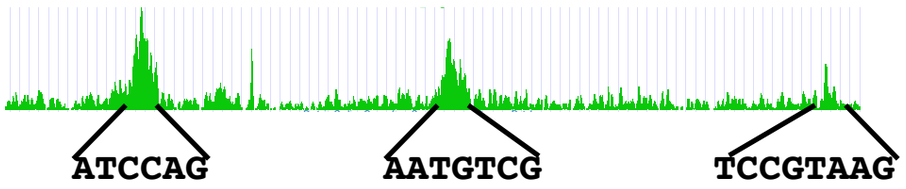
AATGTCG

TCCGTAAG

pseudo-counts: add .25 to each row to avoid 0
then divide by 4 = nb sites + 1

meme-suite.org

EM algorithm



ATCCAG
AATGTCG
TCCGTAAG

Set of ChIP-seq enriched sites

Suppose each site contains one instance of a $L = 3$ motif

	A	C	G	T
1	0.25	2.25	0.25	1.25
2	1.25	2.25	0.25	0.25
3	0.25	0.25	3.25	0.25

PPM * 4

iterate

New scores

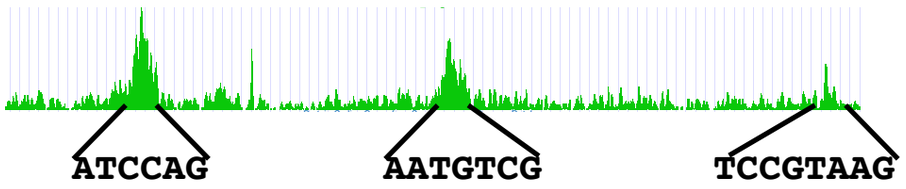
0.155
ATCCAG

0.11
AATGTCG

0.155
TCCGTAAG

meme-suite.org

EM algorithm



ATCCAG
AATGTCG
TCCGTAAG

Set of ChIP-seq enriched sites

Suppose each site contains one instance of a $L = 3$ motif

	A	C	G	T
1	0.06	0.56	0.06	0.32
2	0.32	0.56	0.06	0.06
3	0.06	0.06	0.82	0.06

PPM

iterate

0.155
ATCCAG

0.11
AATGTCG

0.155
TCCGTAAG

Initialize with "flat" matrix

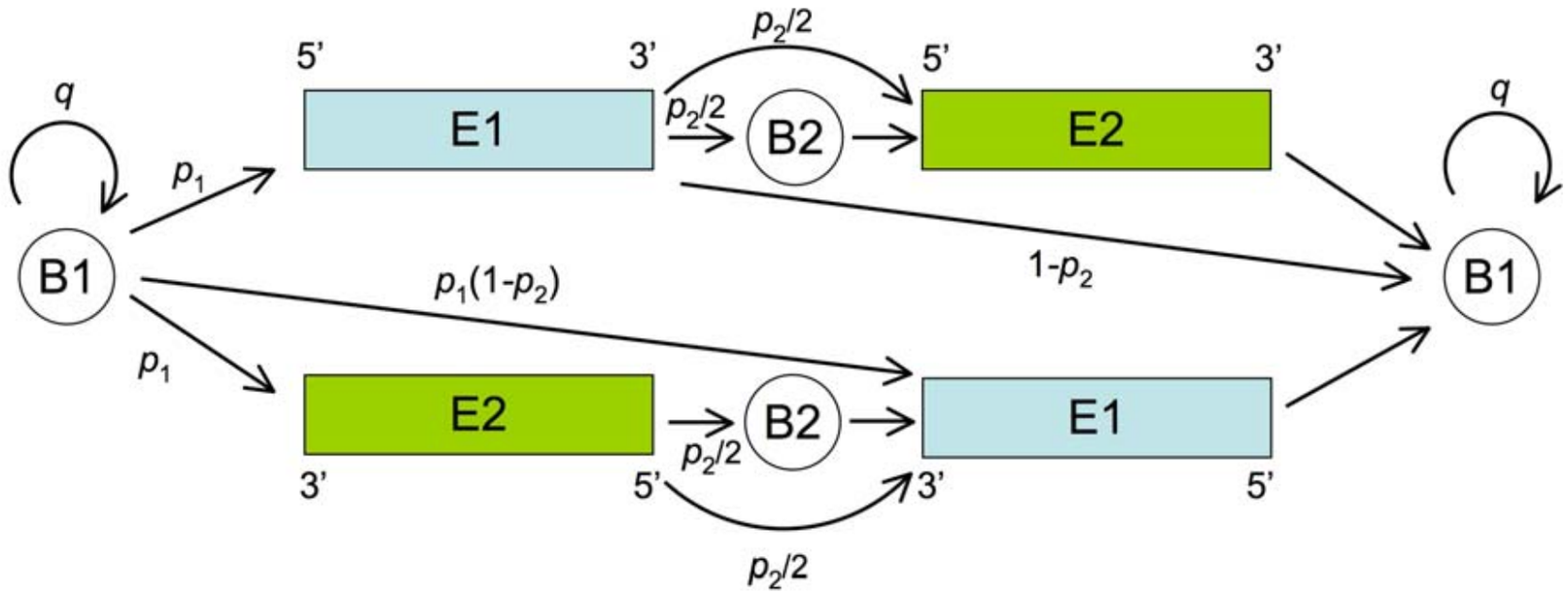
	A	C	G	T
1	0.25	0.25	0.25	0.25
2	0.25	0.25	0.25	0.25
3	0.25	0.25	0.25	0.25

meme-suite.org

HMMs

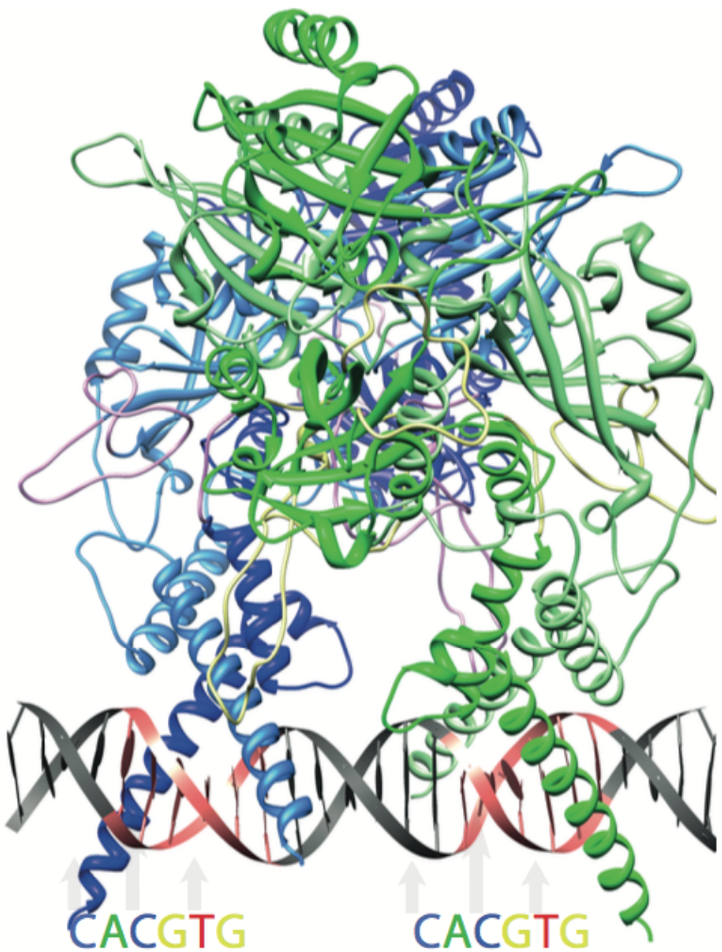
HMMs are particularly well adapted to modeling multiple binding sites in promoters.

Example: the double E-box structure of circadian promoters



E1 converged ($p_1=2^{-11}$, $p_2=2^{-4}$)

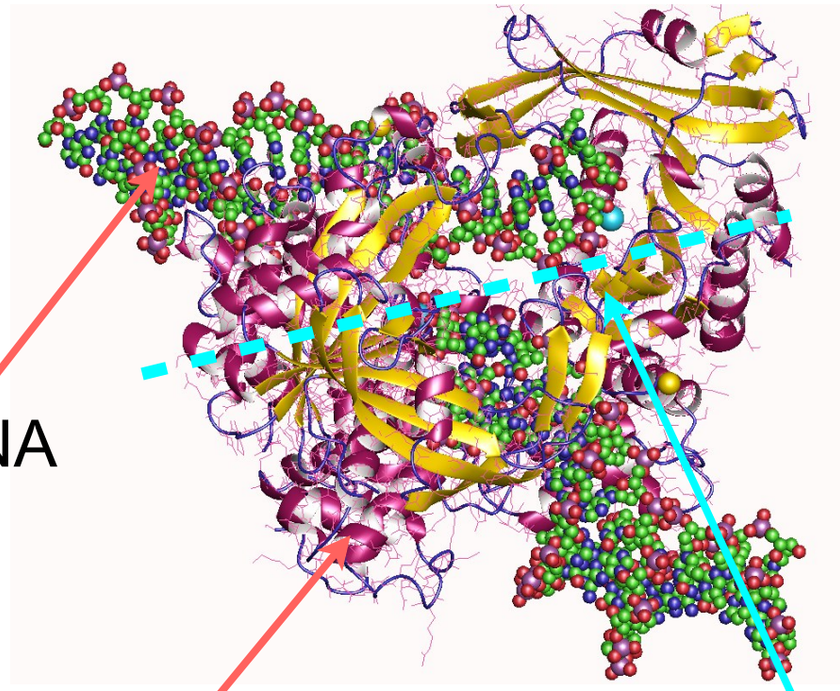
E2 converged



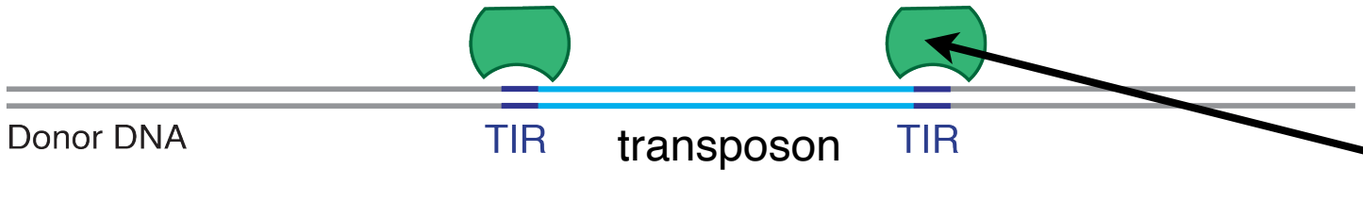
Paquet et al. PLoS Comp Bio (2009)

Tn5 transposons

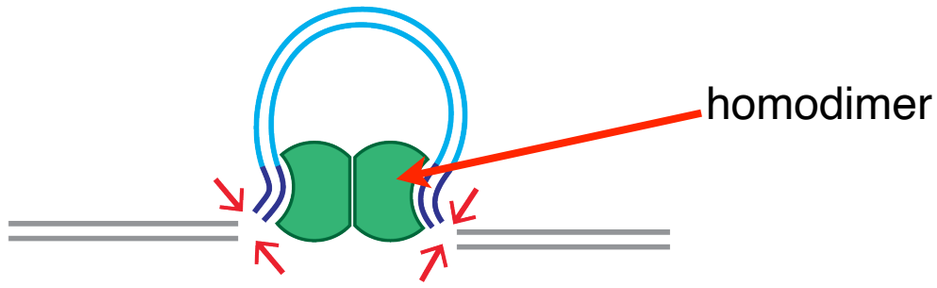
Transposase: a protein that can cleave a segment of DNA and transpose it elsewhere



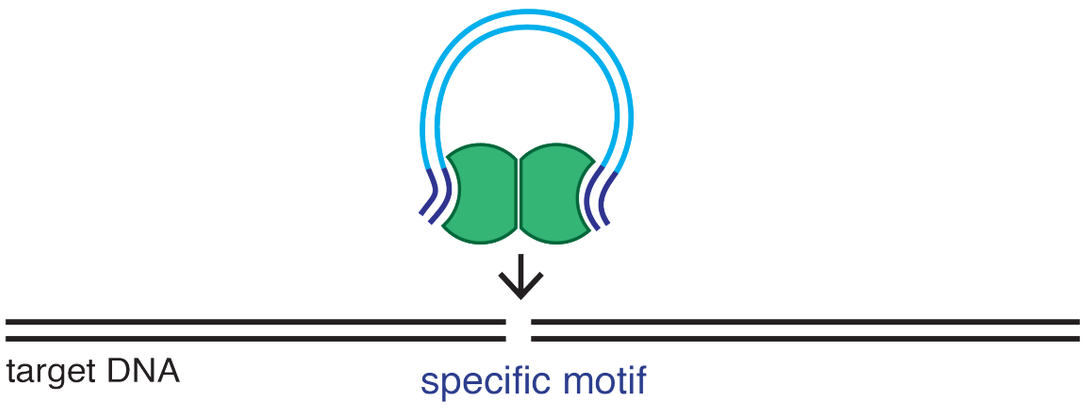
Transposase binding



Cleavage



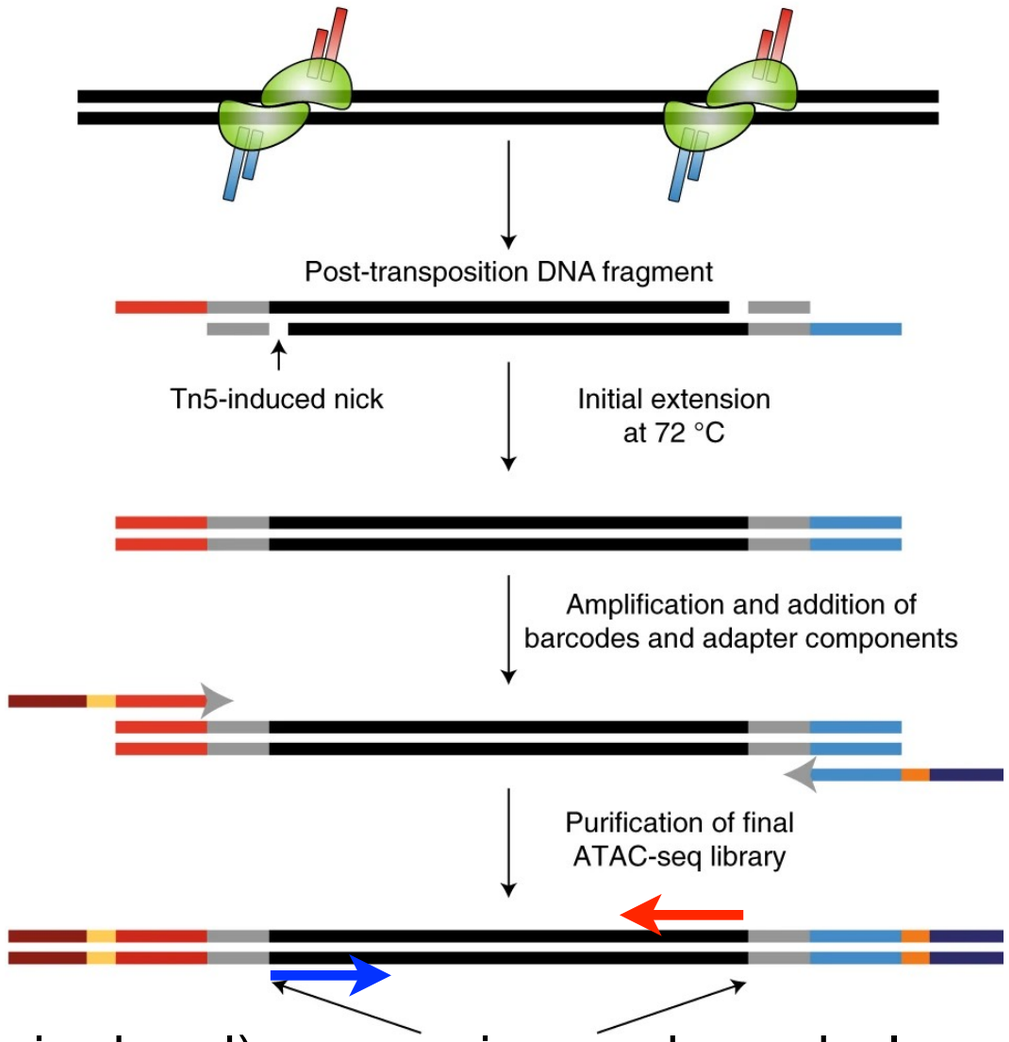
Target capture and strand transfert



insertion into remote location

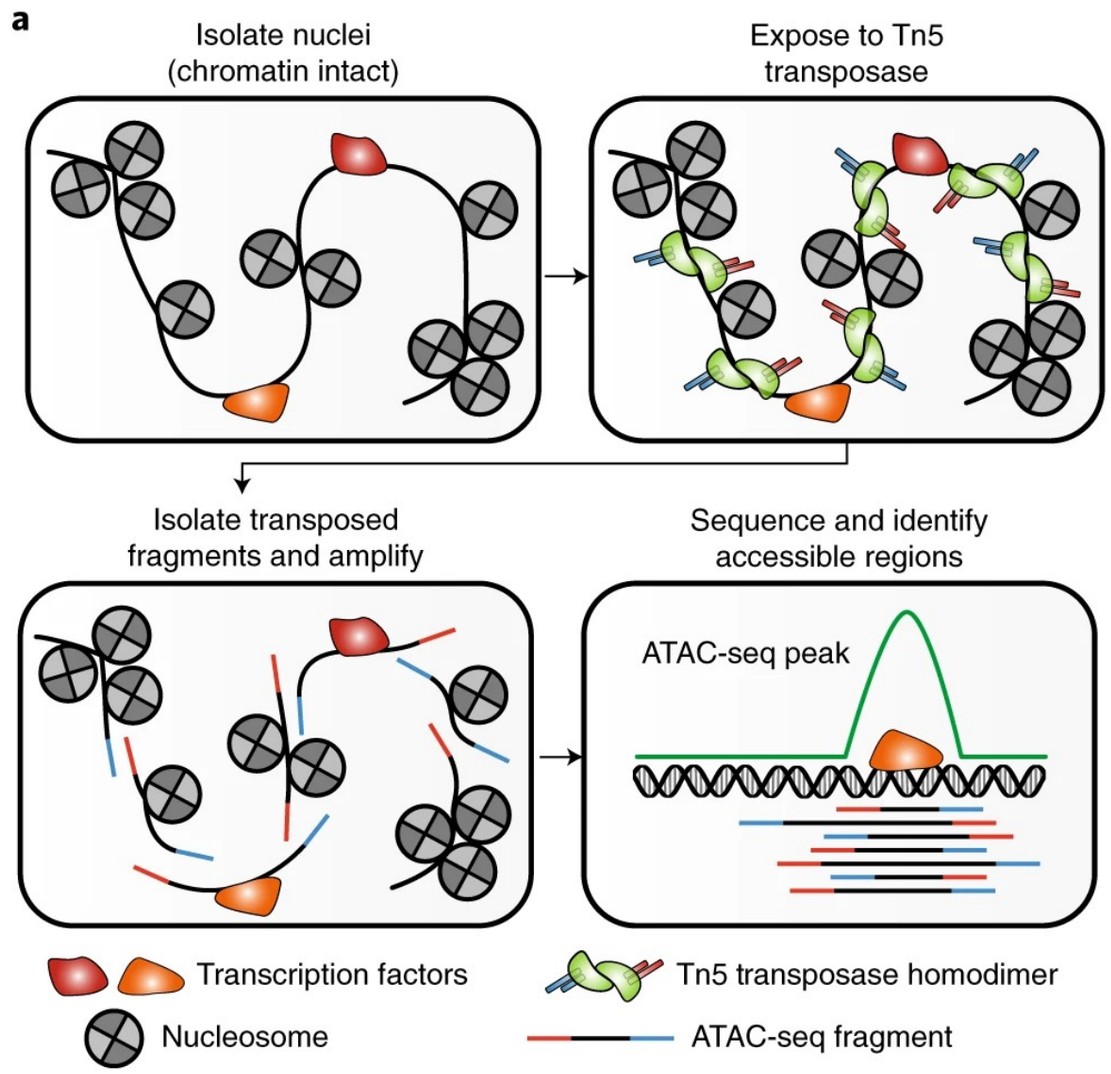
ATAC-seq

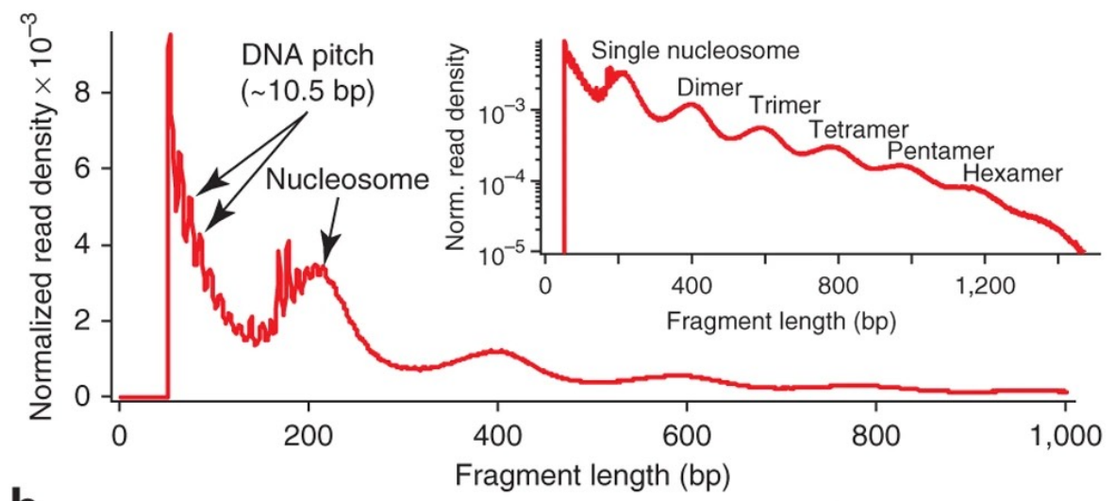
Use Tn5 transposases preloaded with sequencing adapters (+ barcode, UMI, etc.)



(paired-end) sequencing reads mark cleavage site

Fragments enriched near protected regions

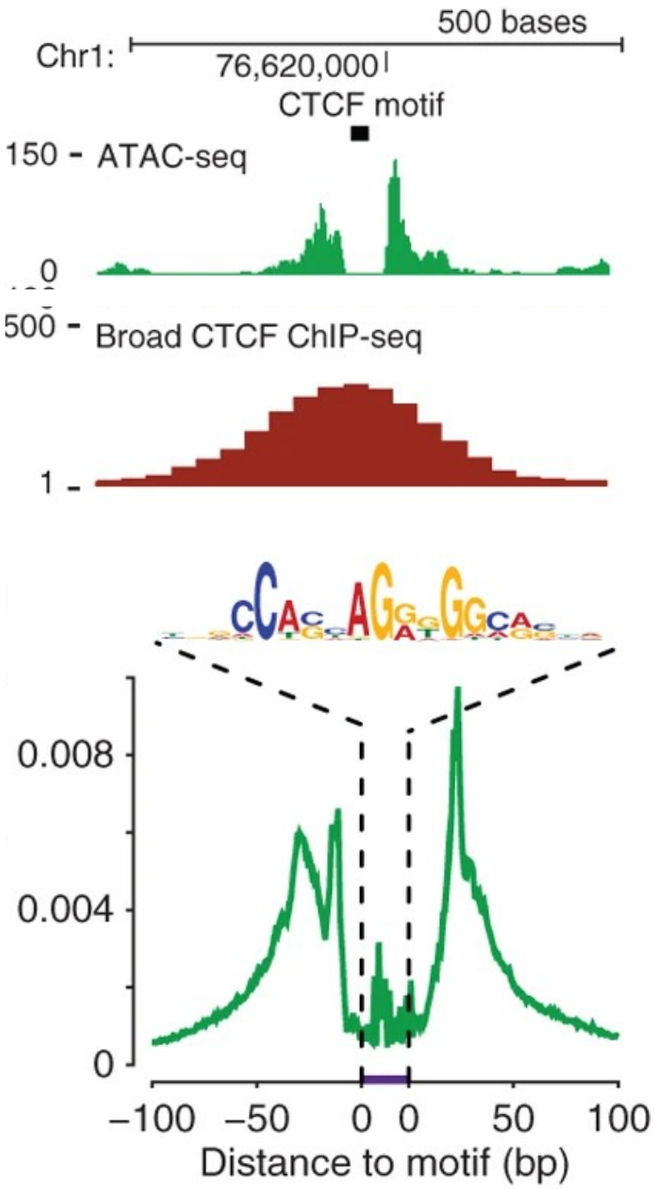


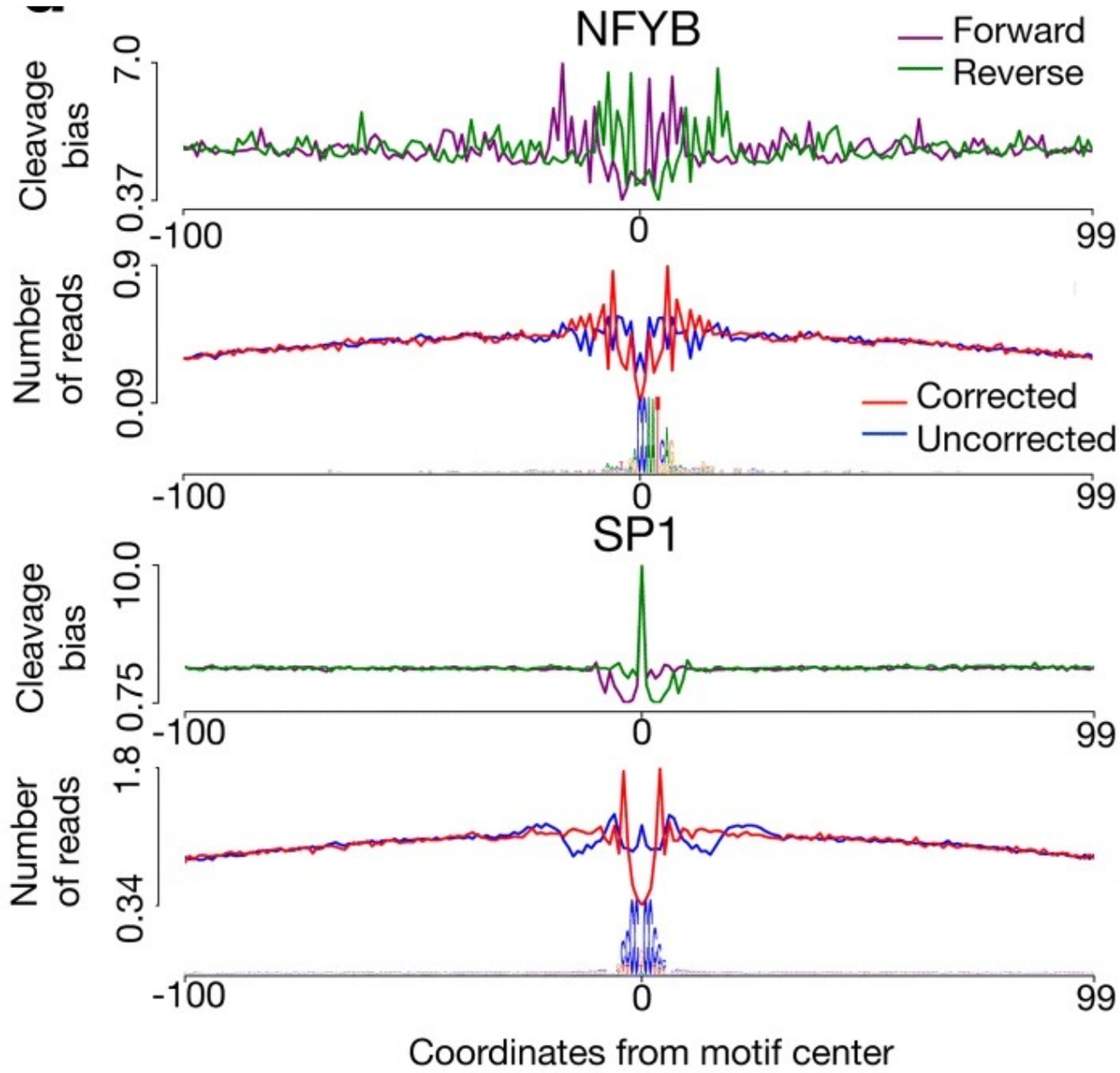


Grandi FC et al. Nat. Prot. (2022)
Buenrostro JD et al. Nat. Meth (2013)

Footprints

Bound transcription factors create a "footprint" inside a cleavage peak



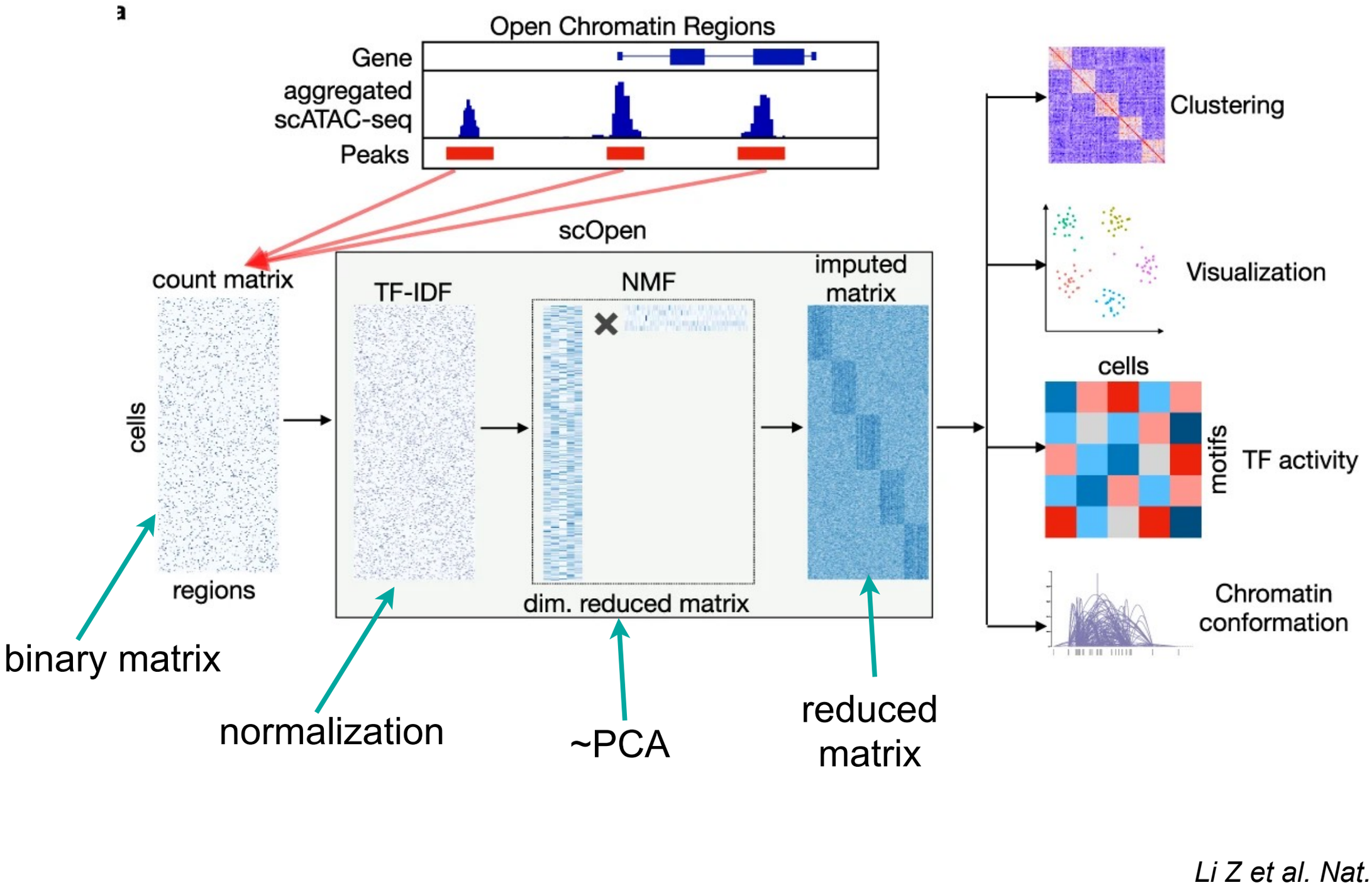


Li Z et al. Genome Biol. (2019)

Single cell ATAC-seq

Technique is sensitive enough to use on single nuclei.

1. Identify peaks on combined data
2. Cluster cells + peaks with similar patterns



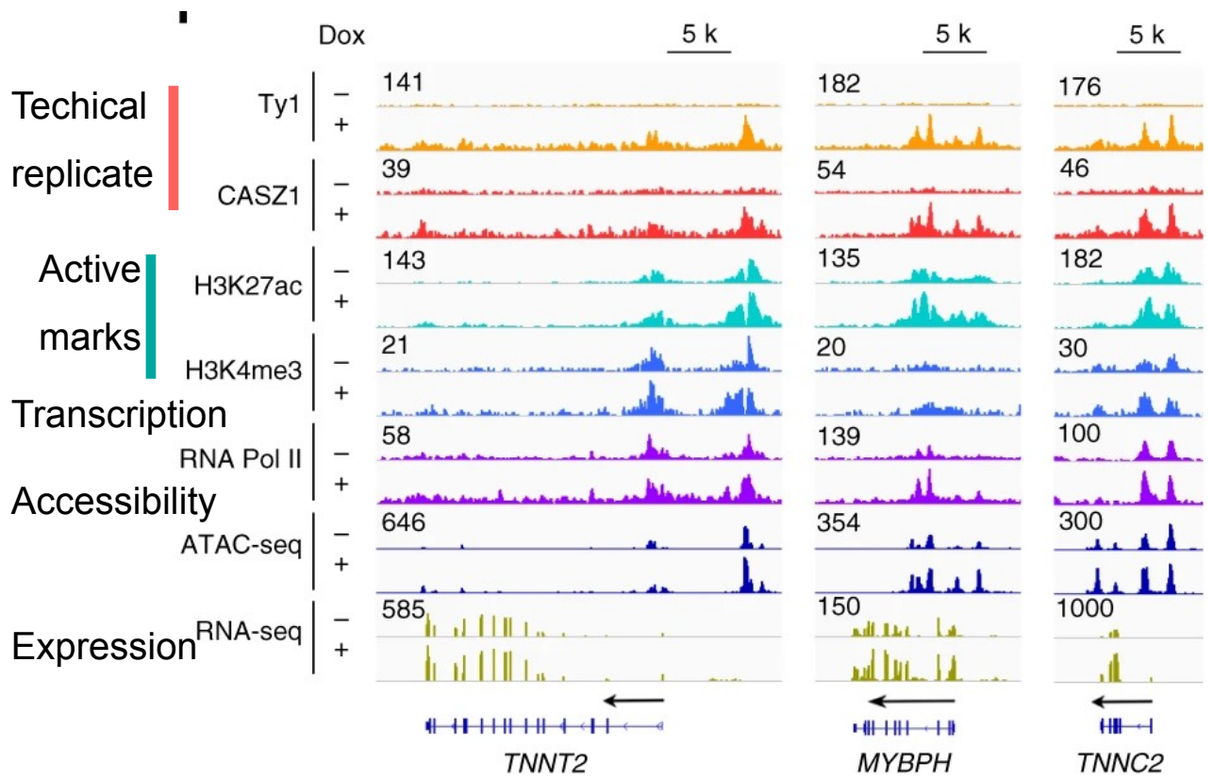
Li Z et al. Nat. Comm. (2021)

ChIP-seq + ATAC-seq + TF motifs

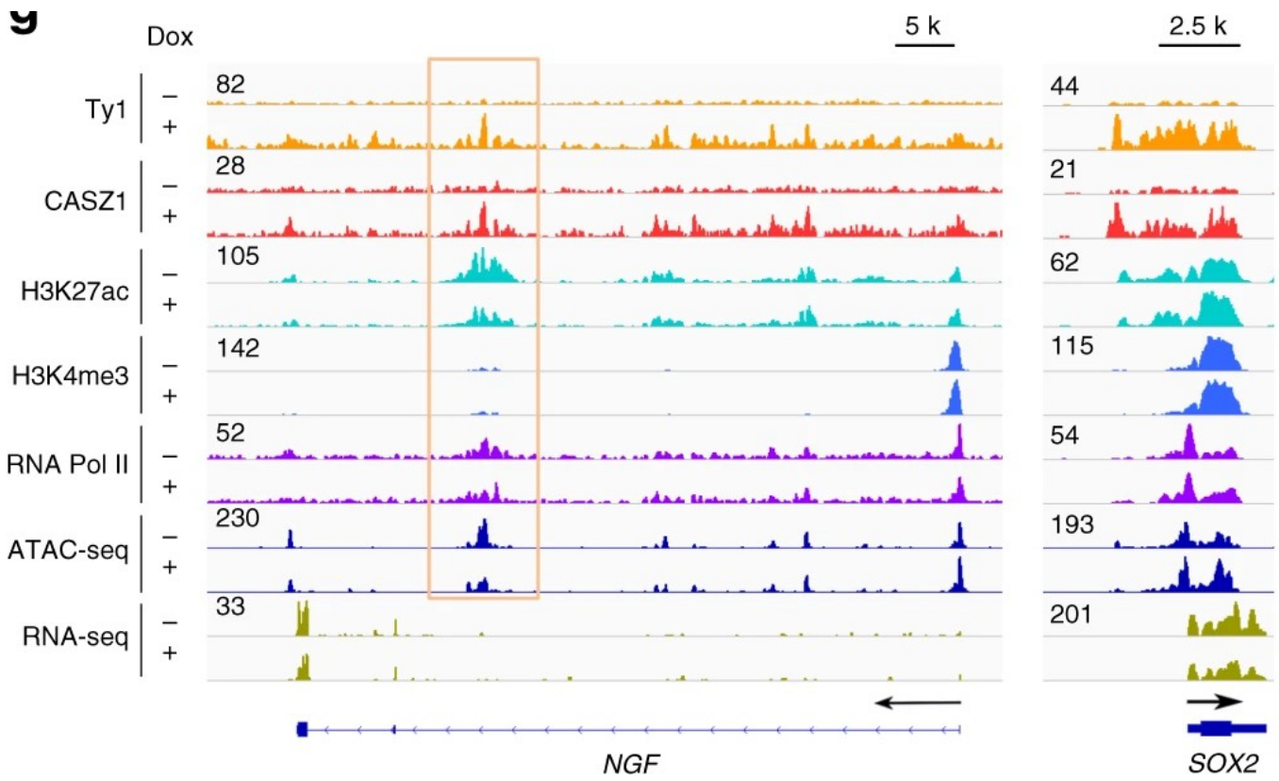
- Induce expression of transcription factor CASZ1 to trigger muscle differentiation
- Cooperates with specific myogenic regulatory factors MYOD, MYOG, MEF2D

Chromatin landscape

Upregulated muscle genes



Downregulated neural genes



Liu Z et al. Nat. Comm. (2020)

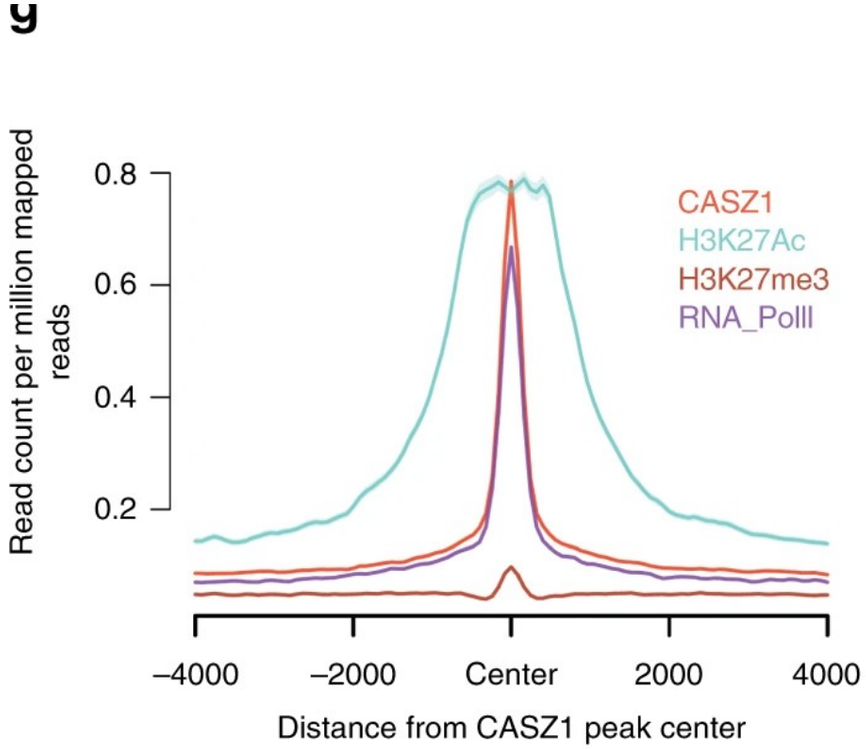
ChIP-seq + ATAC-seq + TF motifs

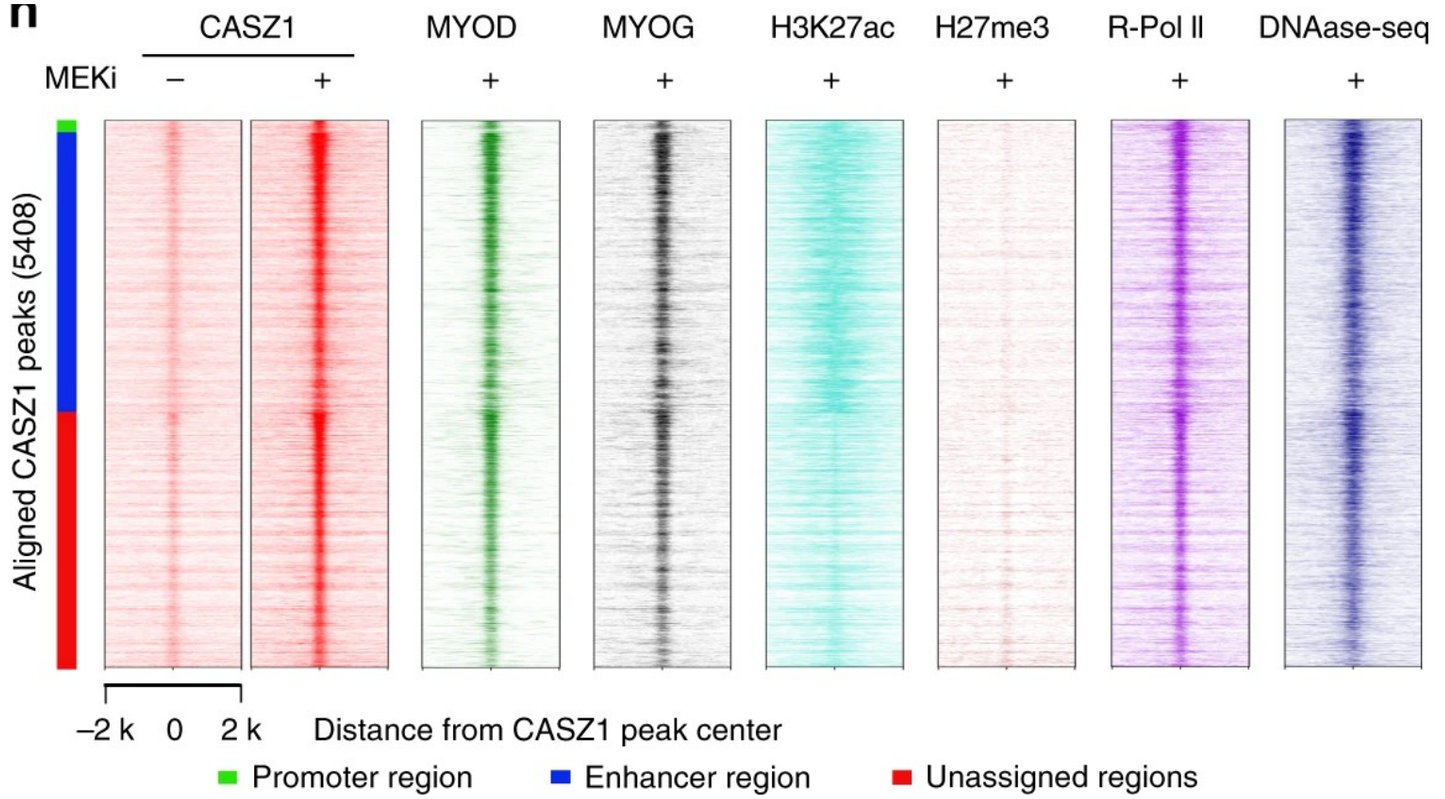
Motif inference from ChIP-seq peaks

CASZ1 HOMER de novo motif			
Rank	Motif	*P*-value	Best match
1		1e-1357	MYOD, MYOG, MYF5, TCF12
2		1e-271	TEAD1, TEAD2, TEAD3, TEAD4

CASZ1 binds to same motif as MYOD, MYOG, TEAD (MEF2)

Averaged / cumulative profiles





Liu Z et al. Nat. Comm. (2020)