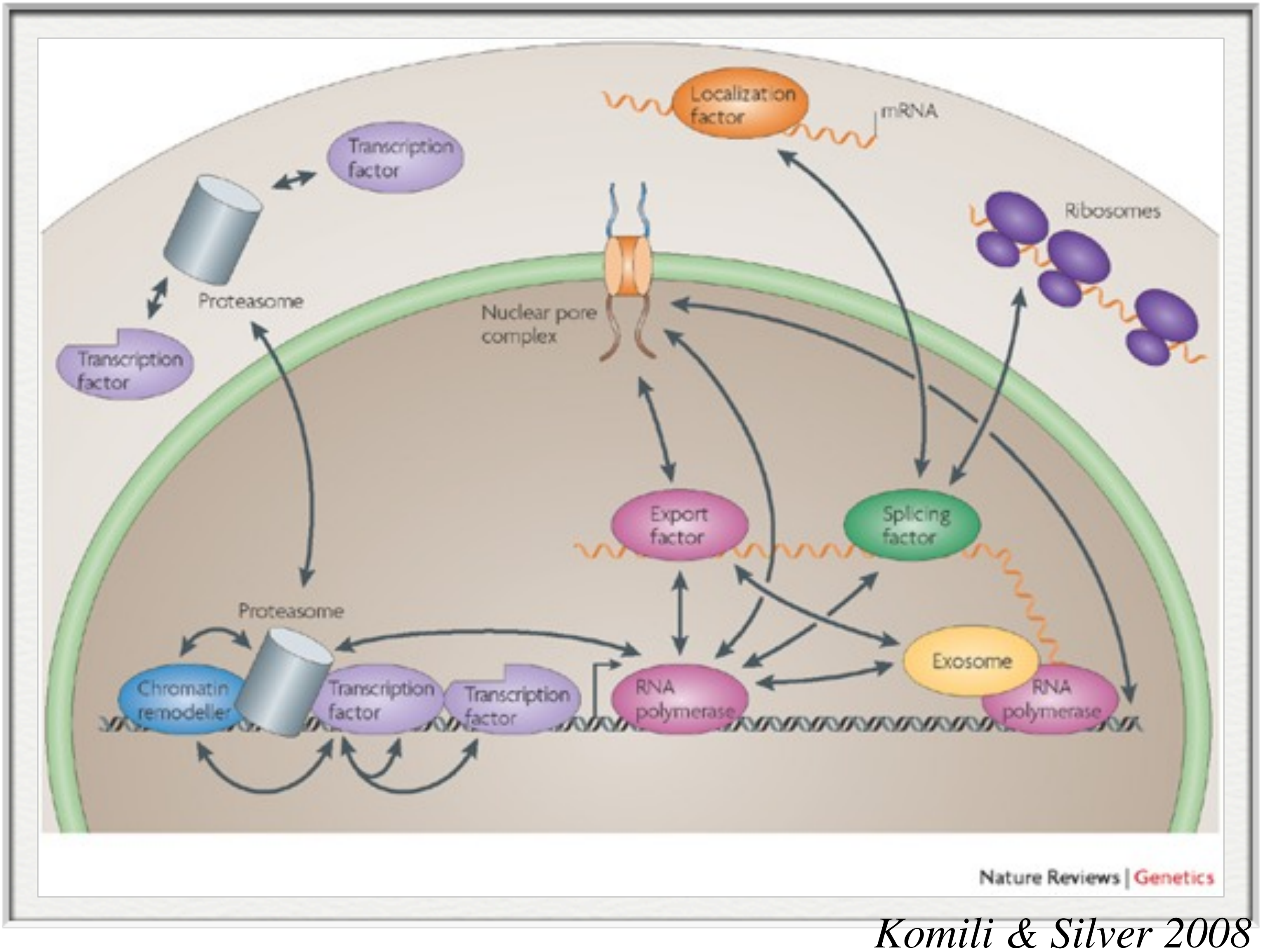
Protein-DNA interactions

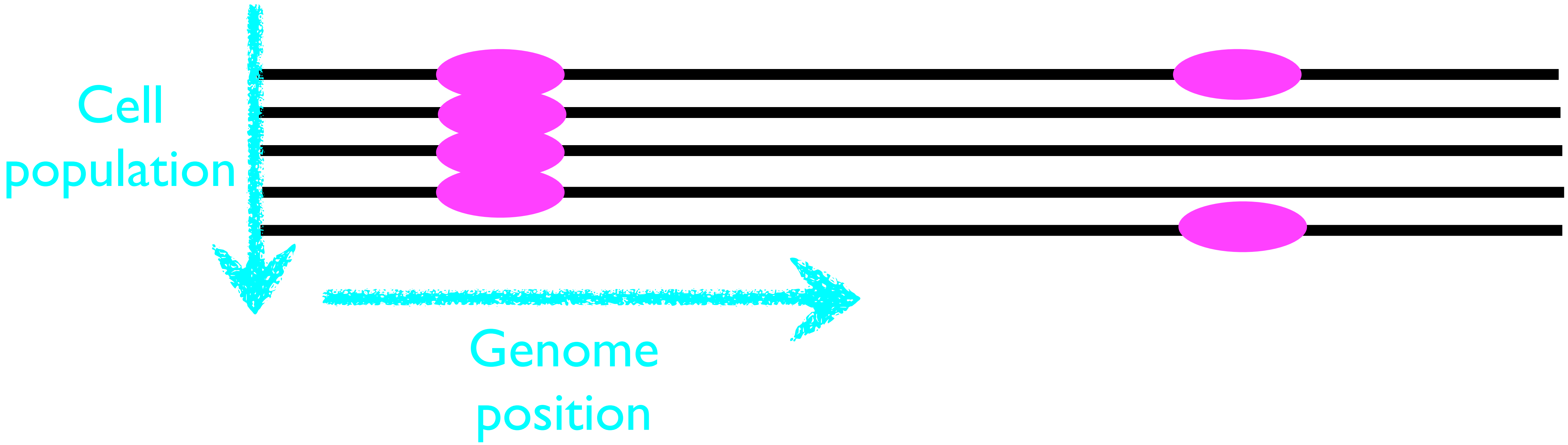
- Gene regulation occurs via interaction of DNA with protein complexes
- There is specific binding (transcription factors), indirect binding (co-factors), unspecific binding (Polymerase, histones)
- All of those can be studied via chromatin-immunoprecipitation (ChIP)



Komili & Silver 2008

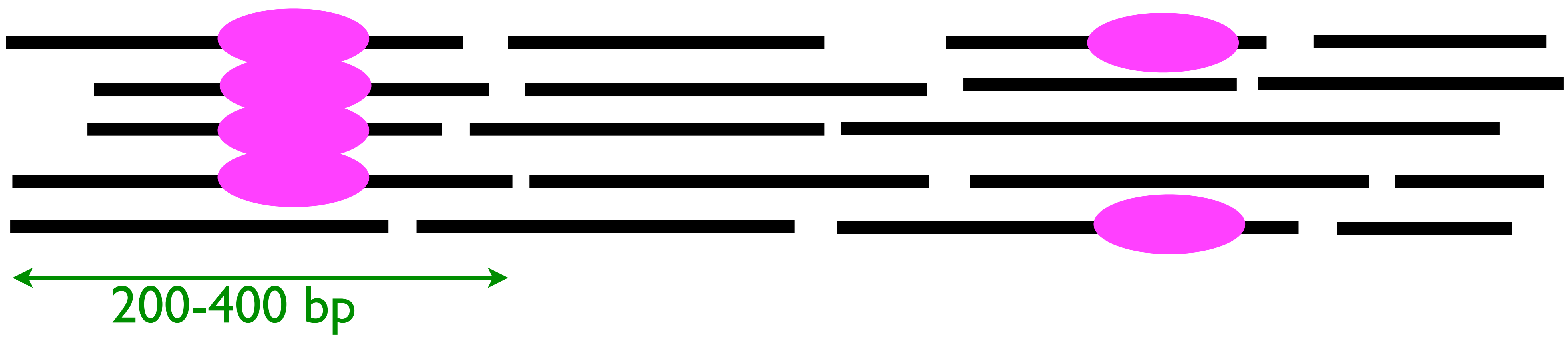
ChIP-Seq: method

1) Cross-link Proteins+DNA



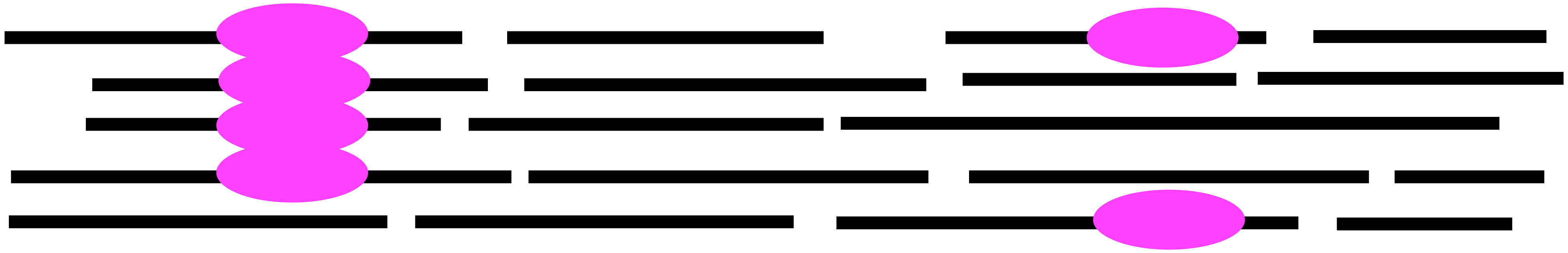
ChIP-Seq: method

2) Sonicate (or digest)



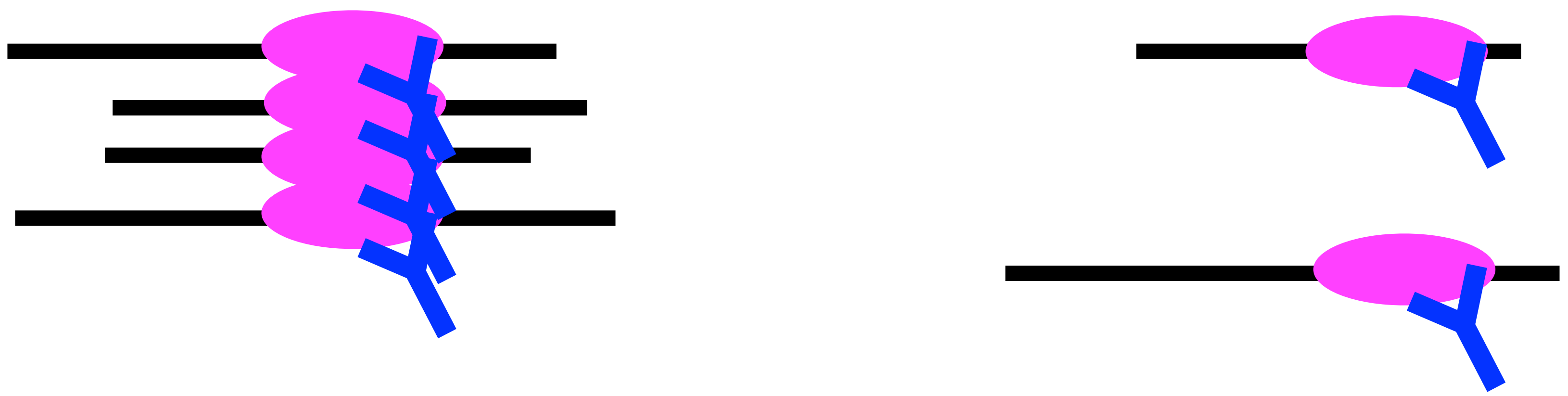
ChIP-Seq: method

3) ImmunoPrecipitate



ChIP-Seq: method

3) ImmunoPrecipitate



ChIP-Seq: method

3) ImmunoPrecipitate

ChIP-Seq: method

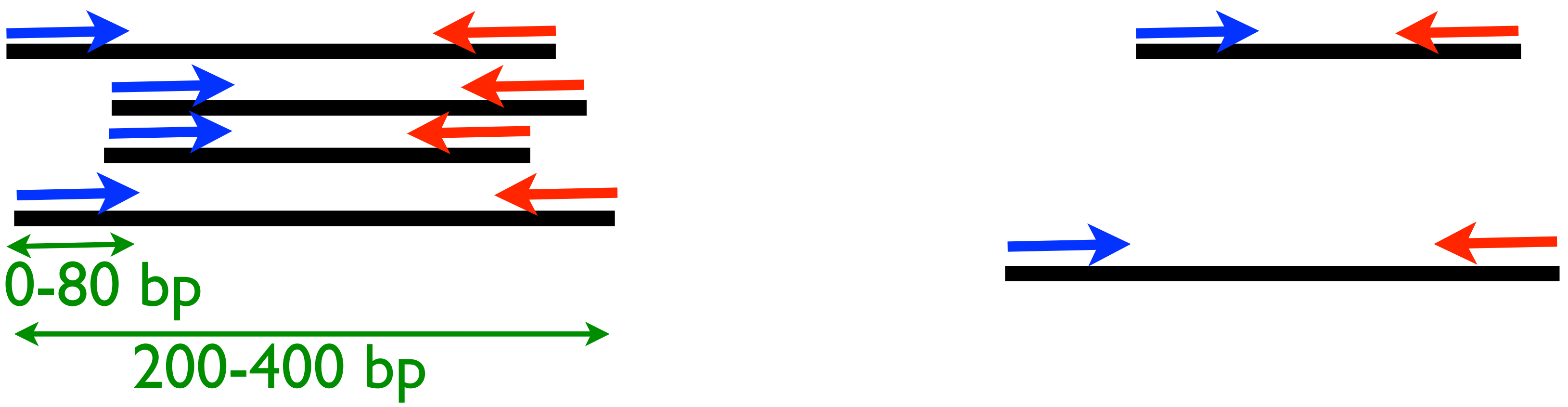
4) Reverse cross-link

ChIP-Seq: method

4) Reverse cross-link

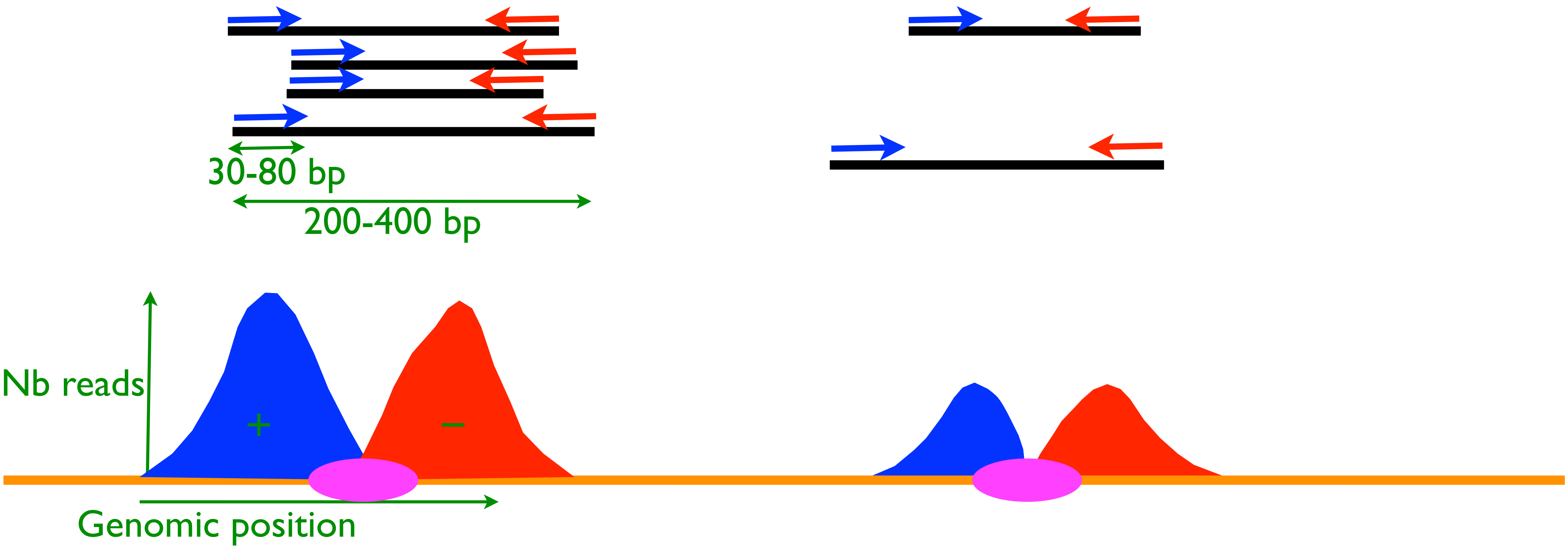
ChIP-Seq: method

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



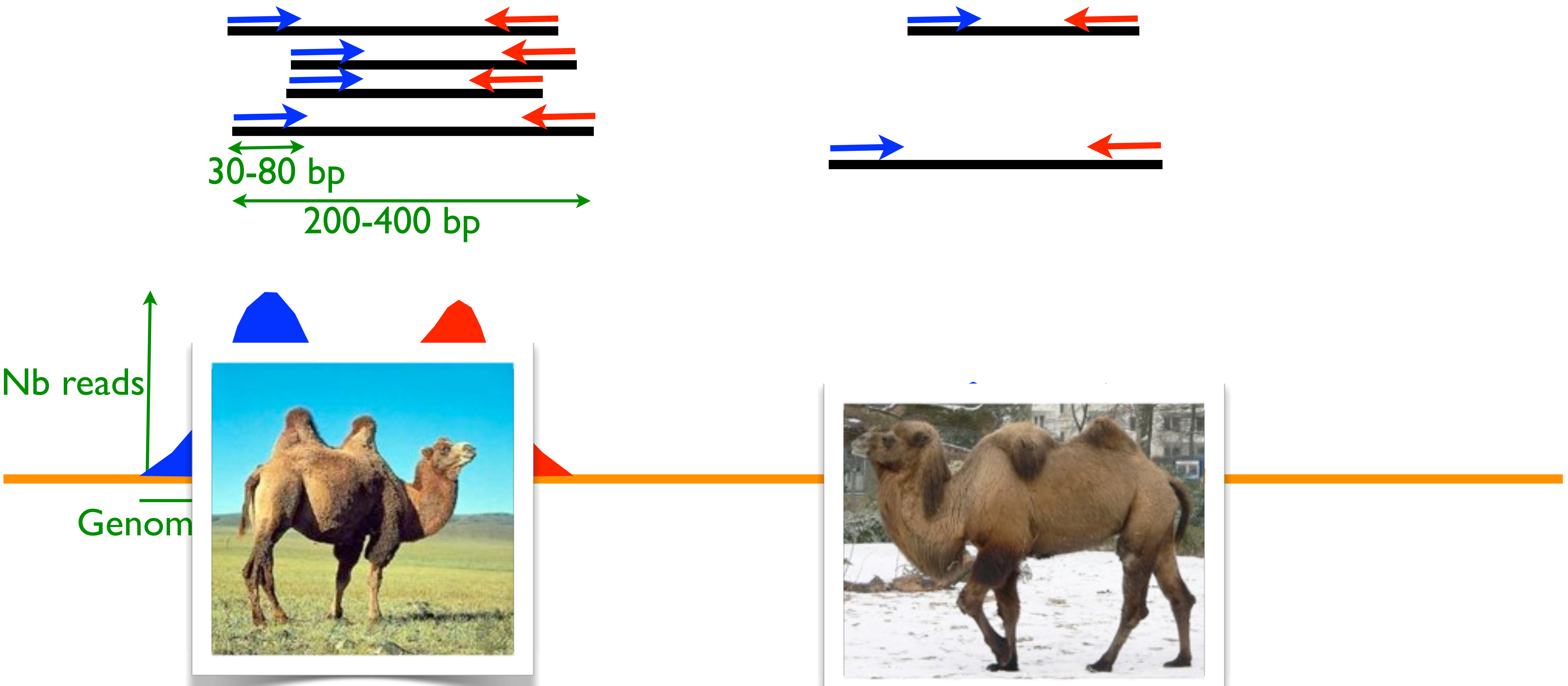
ChIP-Seq: method

6) Map reads to reference sequence



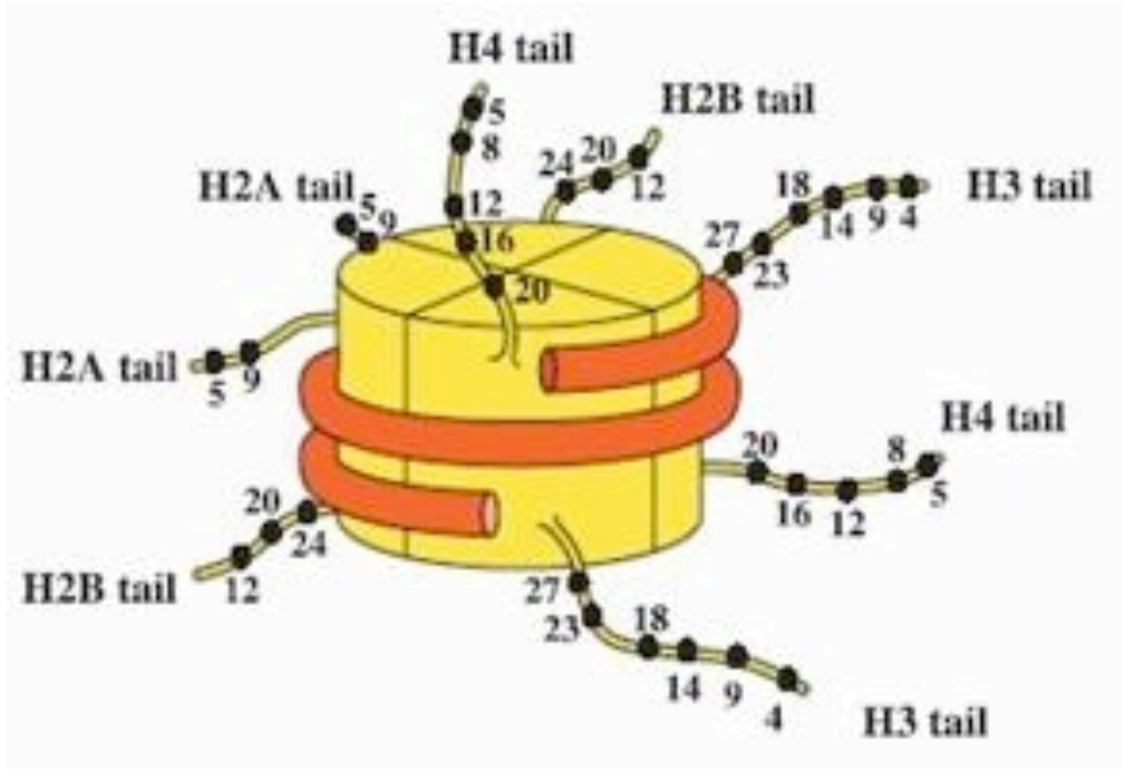
ChIP-Seq: method

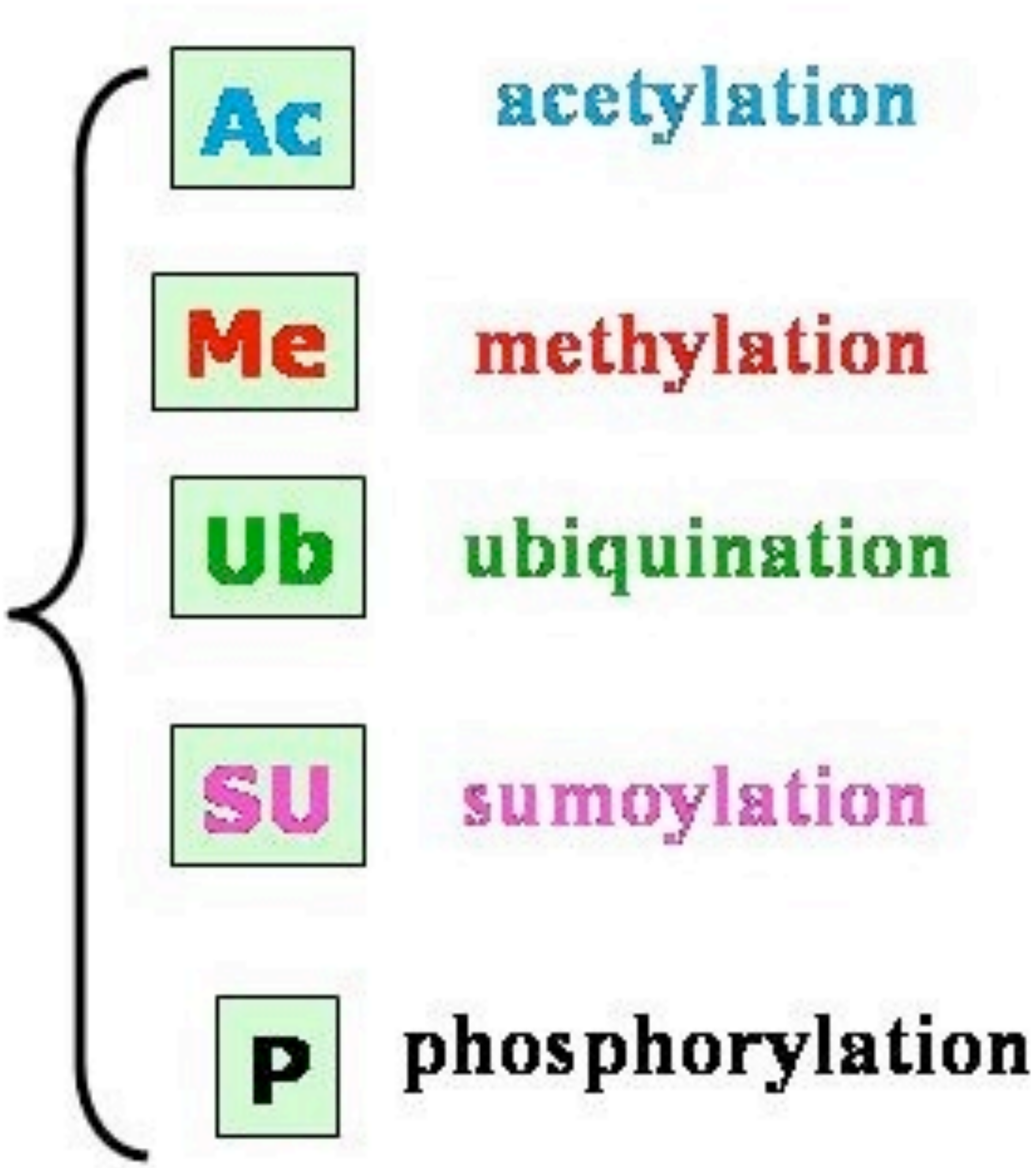
6) Map reads to reference sequence



Histone modifications

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

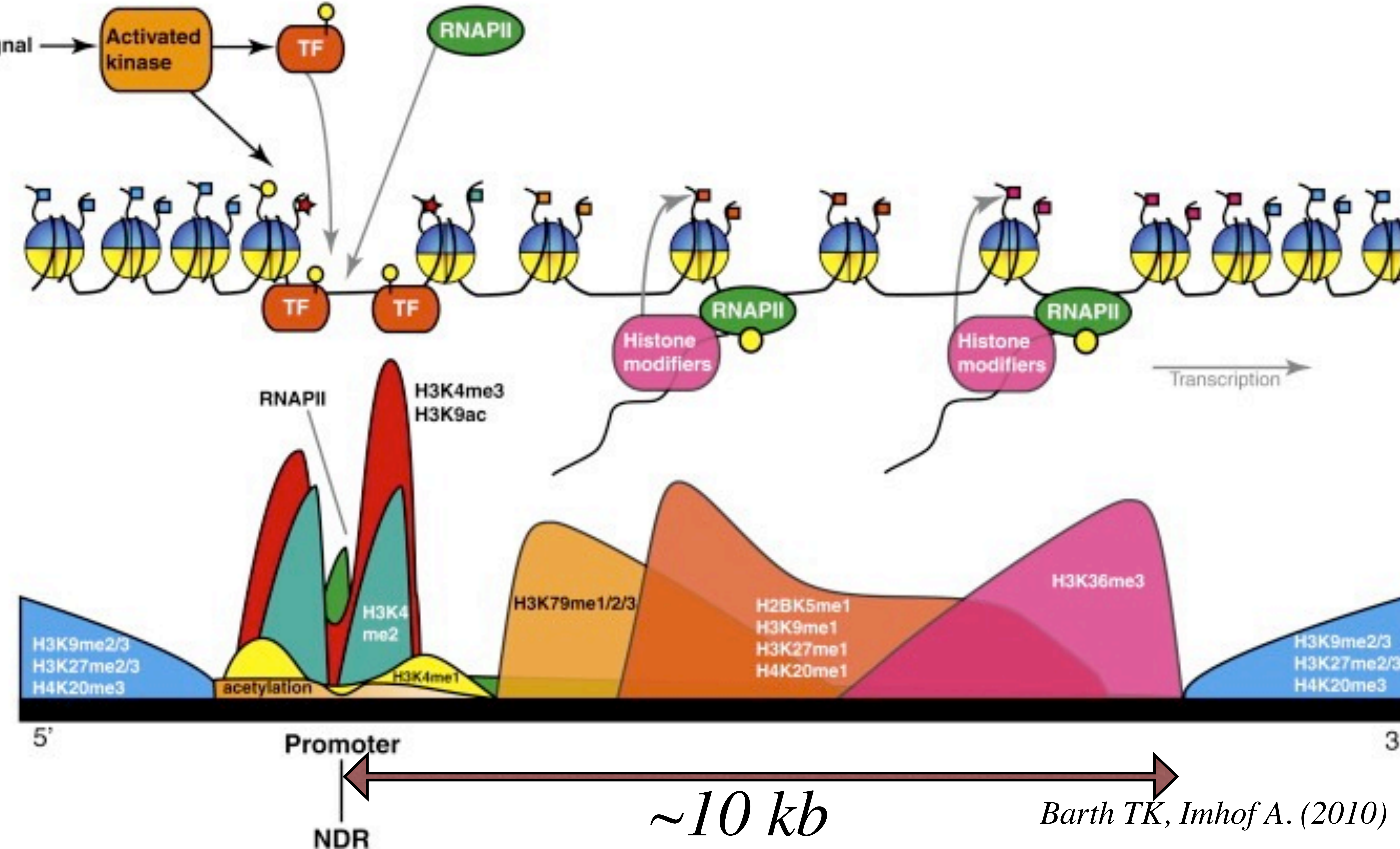




Ac	acetylation
Me	methylation
Ub	ubiquination
SU	sumoylation
P	phosphorylation

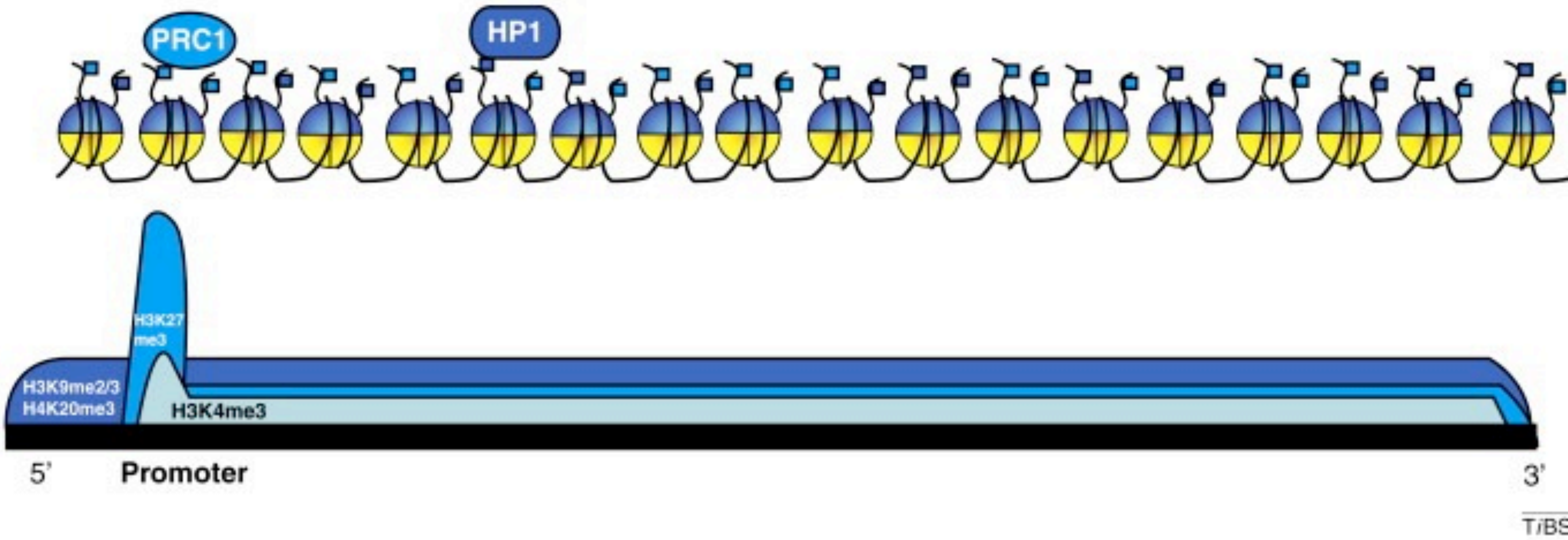
ex.: **H3K4me3** - Lysine (K) at pos. 4 of Histone H3 is 3X methylated

ChIP profiles (active gene)



Barth TK, Imhof A. (2010)

ChIP profiles (inactive gene)



Barth TK, Imhof A. (2010)

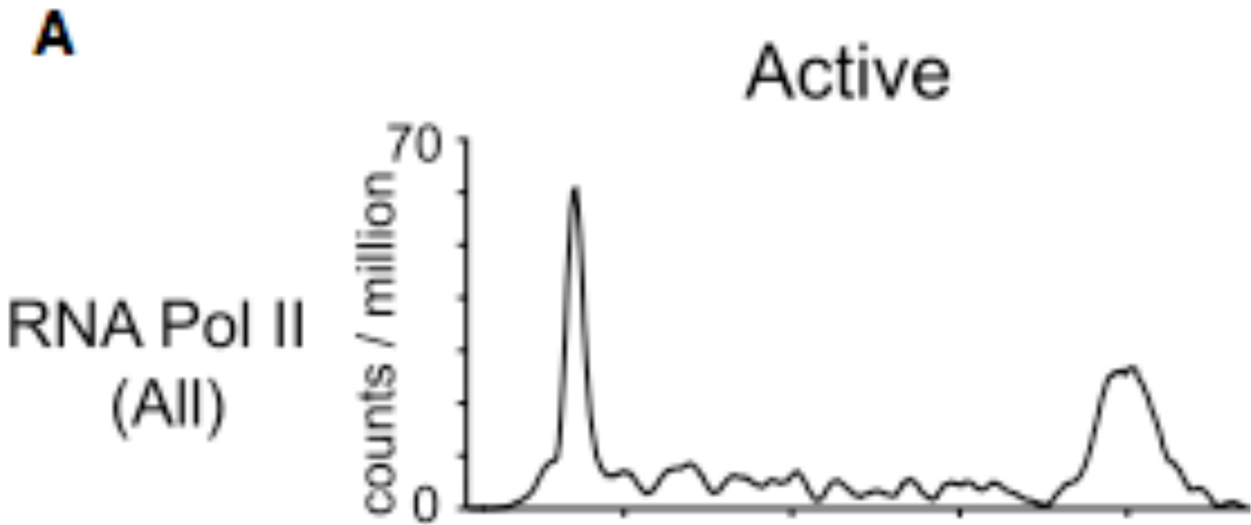
Interpretation of profiles

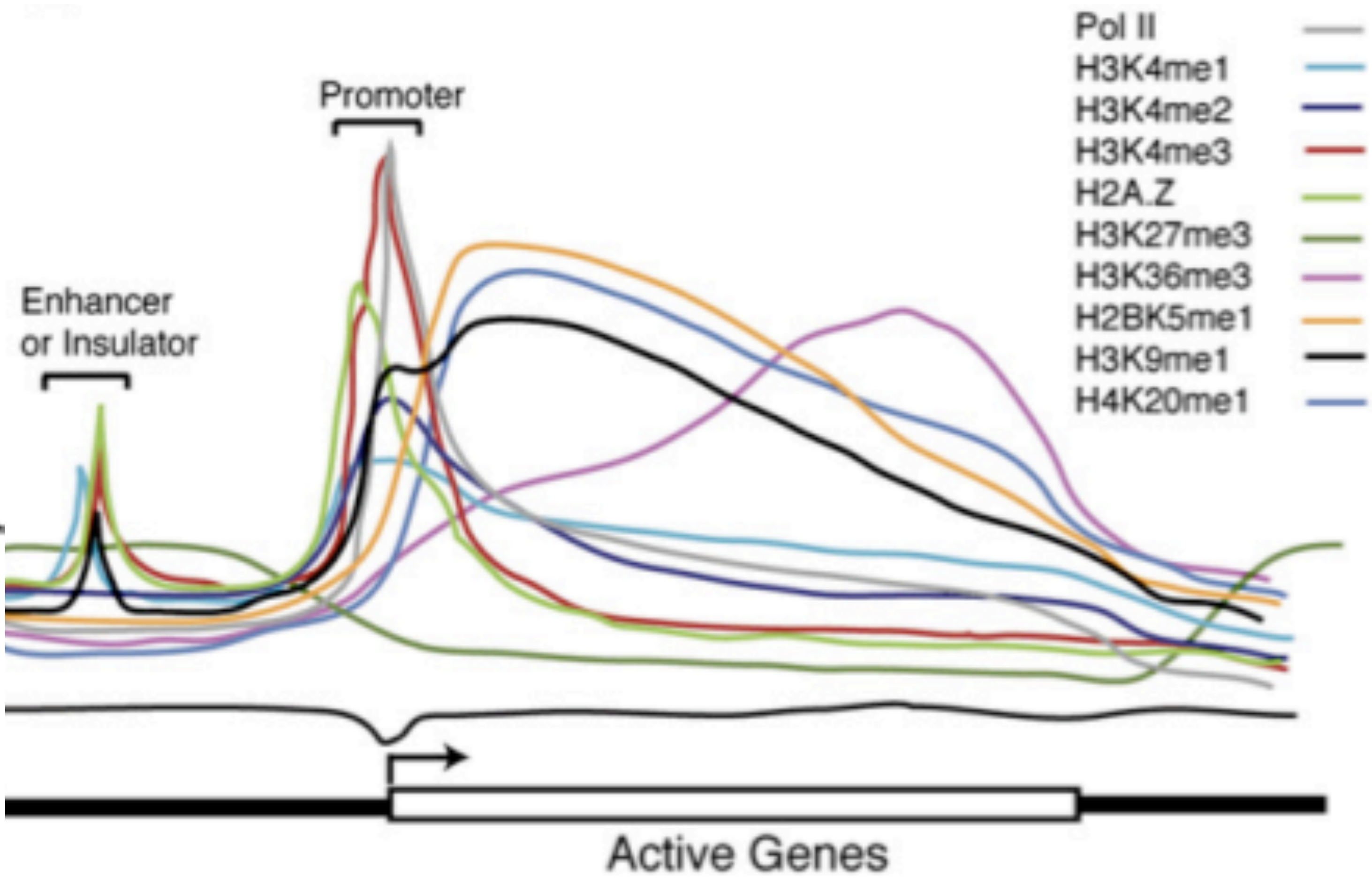
- In general: signal at a genomic position is proportional to **fraction of cells** having the protein bound at this position
    - for histone modifications, this means there is a histone AND it has been modified
    - absence of signal implies either the histone is absent OR it is unmodified
- For travelling proteins (e.g. PolII) this is proportional to **residency times** (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**

Examples

Rahl PB, et al. (2010)

Barski A, et al. (2007)





DNA fragments distribution

Genome size: $3 \cdot 10^9$, fragment size: $3 \cdot 10^2$, hence number of fragments is 10^7.
Typical transcription factors is bound at 1000 sites,
A good antibody will have an enrichment ratio of 100
(bound fragment is 100 times more likely to be selected than unbound fragment).
Therefore the ChIP sample consists of

$$\begin{aligned} 10^3 \cdot 10^2 &= 10^5 \text{ protein-bound fragments, and} \\ 10^7 - 10^3 &\approx 10^7 \text{ background (unbound) fragments .} \end{aligned}$$

$\Longrightarrow$ 99% false positives

Starting material is $\approx 10^7$ cells, typical sequencing throughput is $10^7 - 10^8$ DNA sequences.

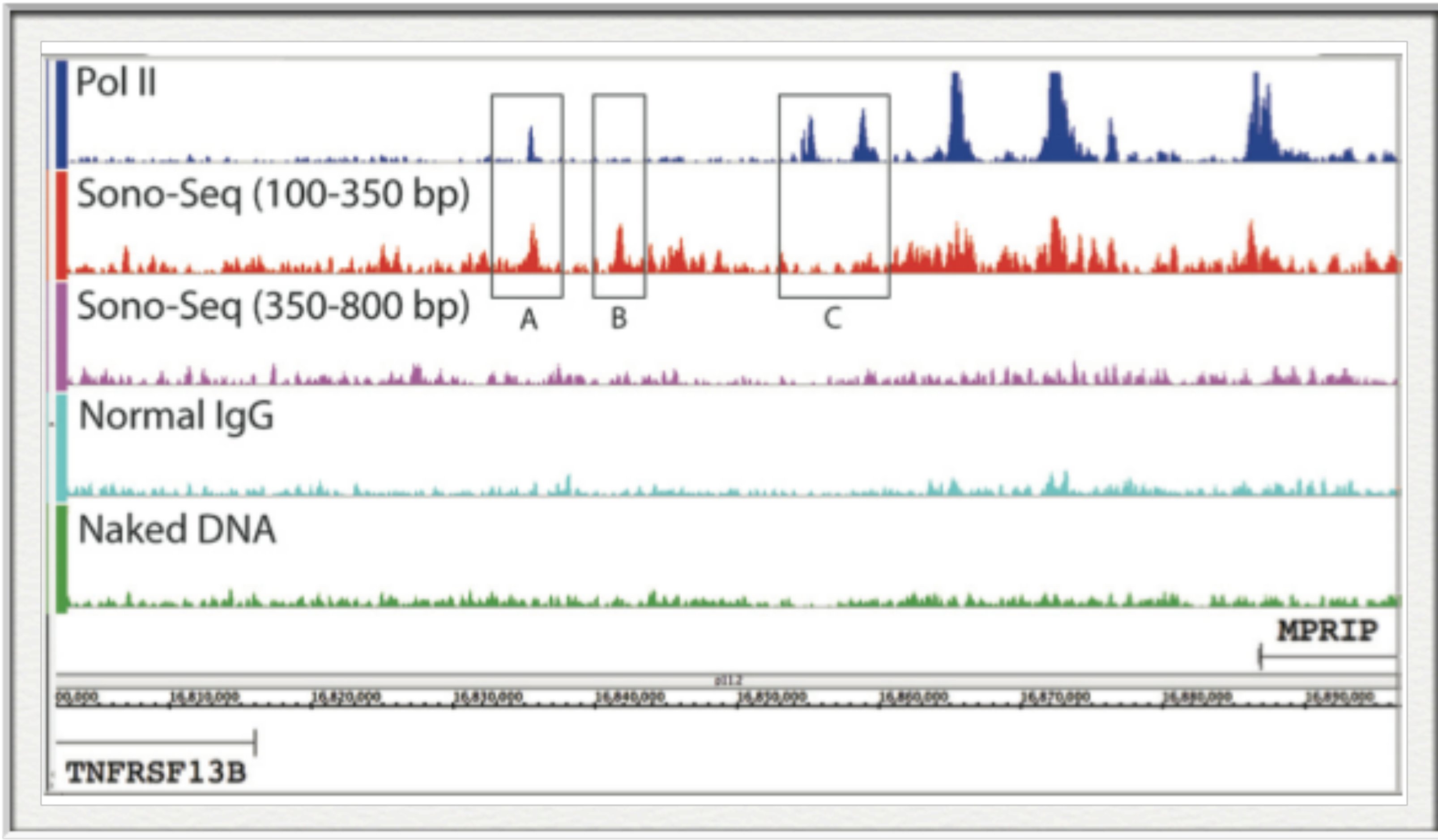
Each protein-bound fragment comes from a different cell

Controls

To detect false positives, several techniques are routinely used:

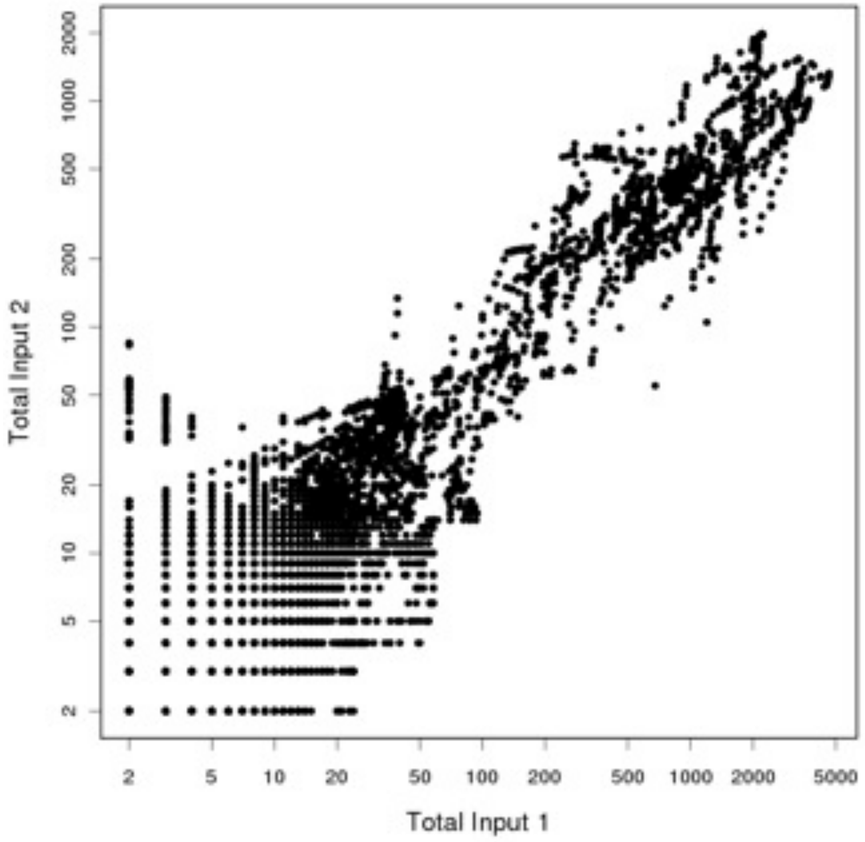
- Naked DNA
- Input (cross-linked) DNA
- Mock IP
- IP on TFΔ KO

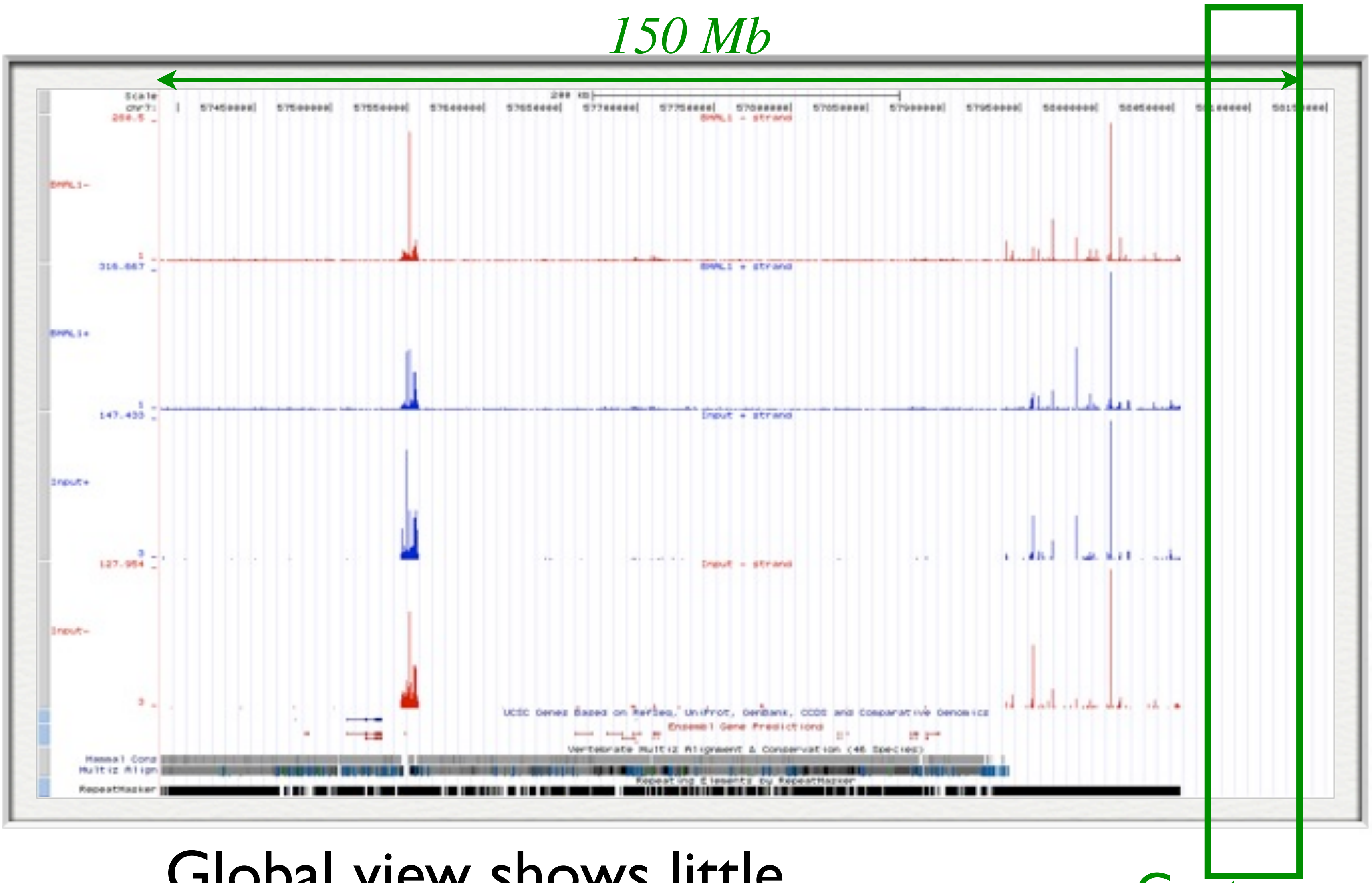
Auerbach et al. (2009)



Enriched regions are rare

Inputs from different cell types are highly correlated

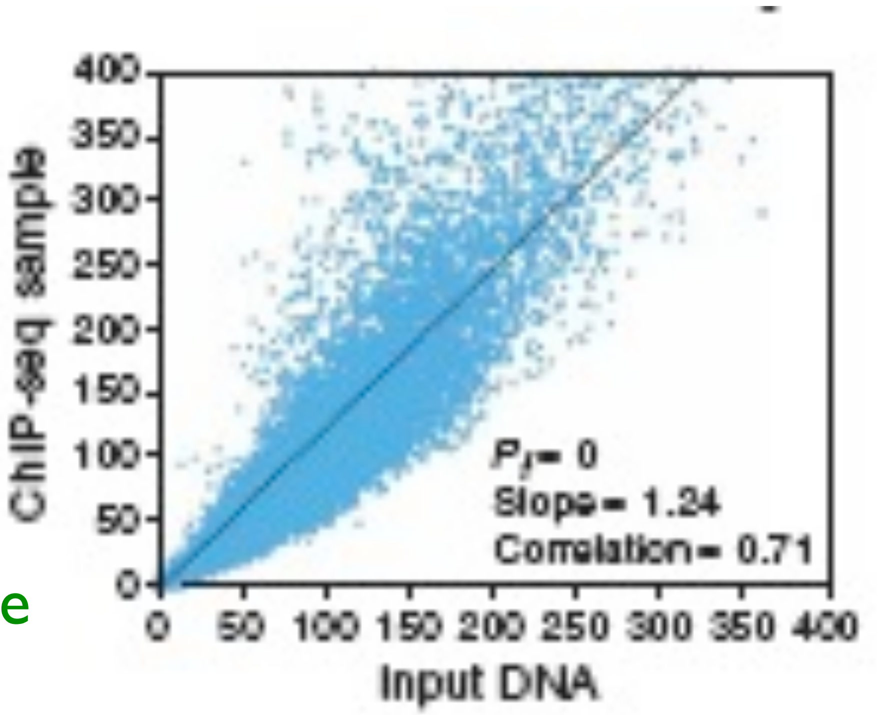




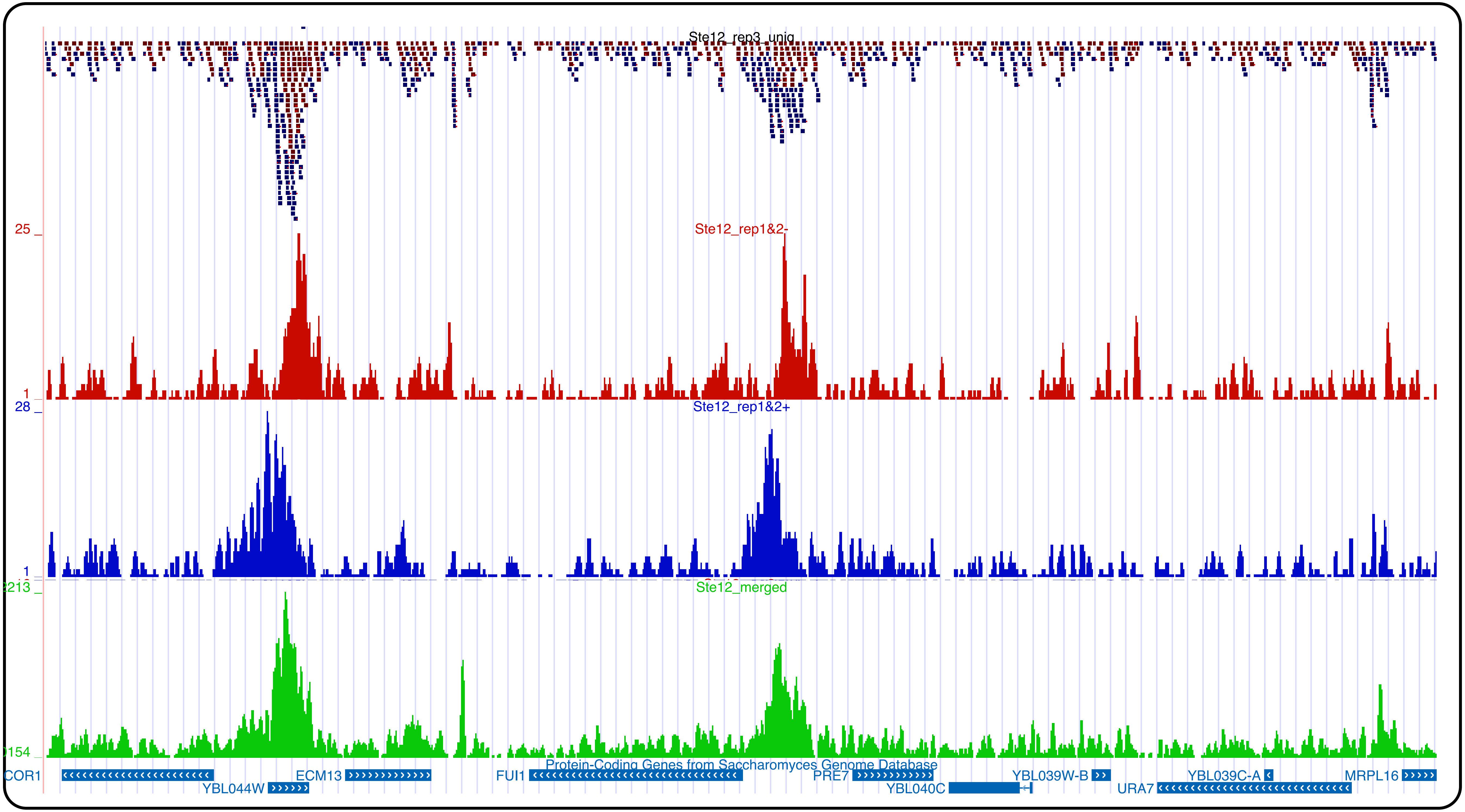
150 Mb
BMAL1 - strand
BMAL1 + strand
Input + strand
Input - strand
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Ensembl Gene Predictions
Vertebrate Multiz Alignment & Conservation (46 Species)
Repeating Elements by RepeatMasker

Global view shows little enrichment in ChIP vs Input

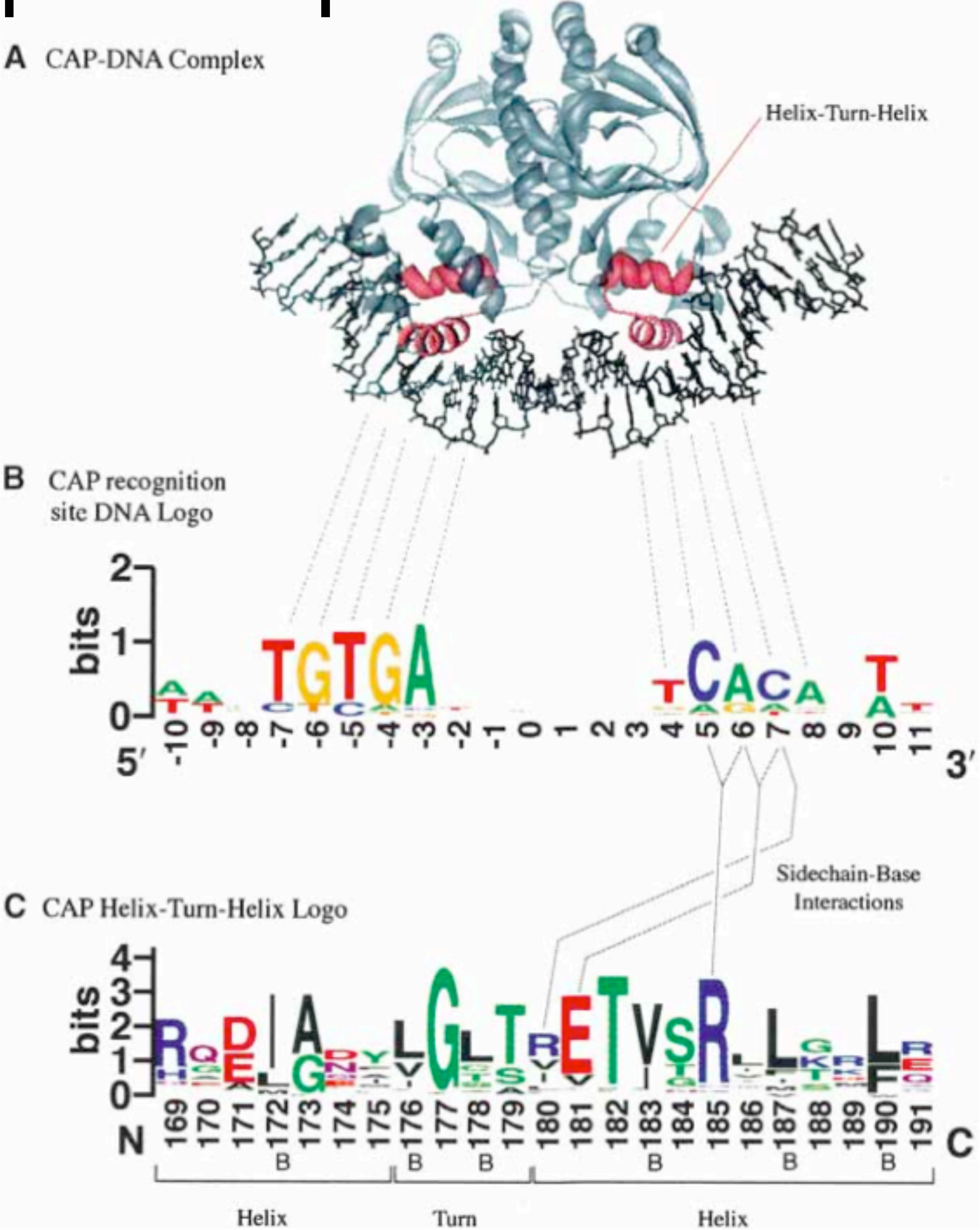
Centromere



Binding regions have characteristic peak shape



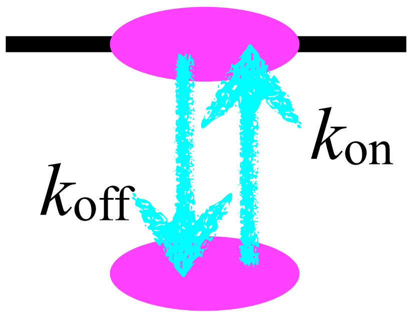
Sequence-specific DNA binding



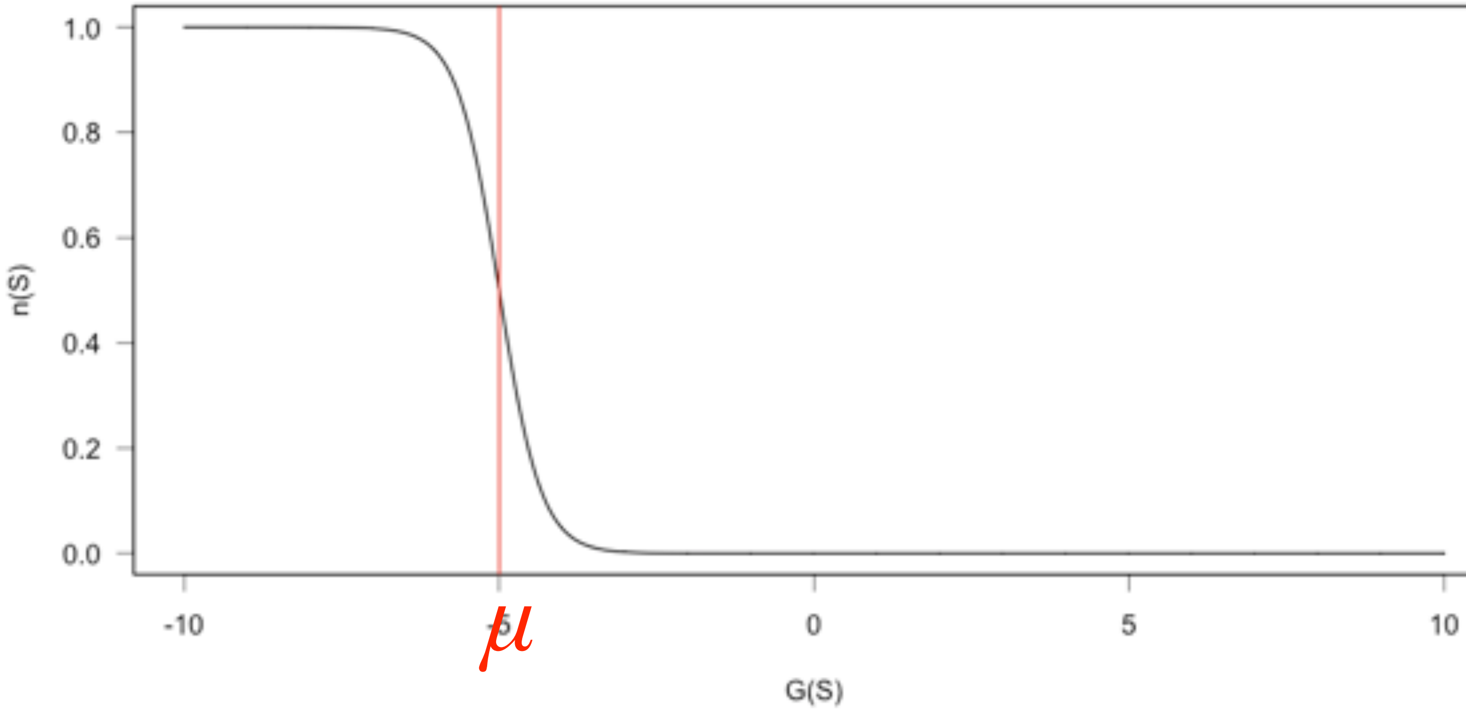
Sequence-specific occupancy

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

$$
\begin{aligned}
K_d^{-1}(S) &= \frac{k_{\text{on}}}{k_{\text{off}}} = \frac{[P \cdot S]}{[P][S]} = e^{-\beta G(S)} \,, \\
n(S) &= \frac{[P \cdot S]}{[P \cdot S] + [S]} = \frac{1}{1 + [S]/[P \cdot S]} \\
&= \frac{1}{1 + K_d(S)/[P]} = \frac{1}{1 + e^{\beta(G(s) - \mu)}} \,,
\end{aligned}
$$



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



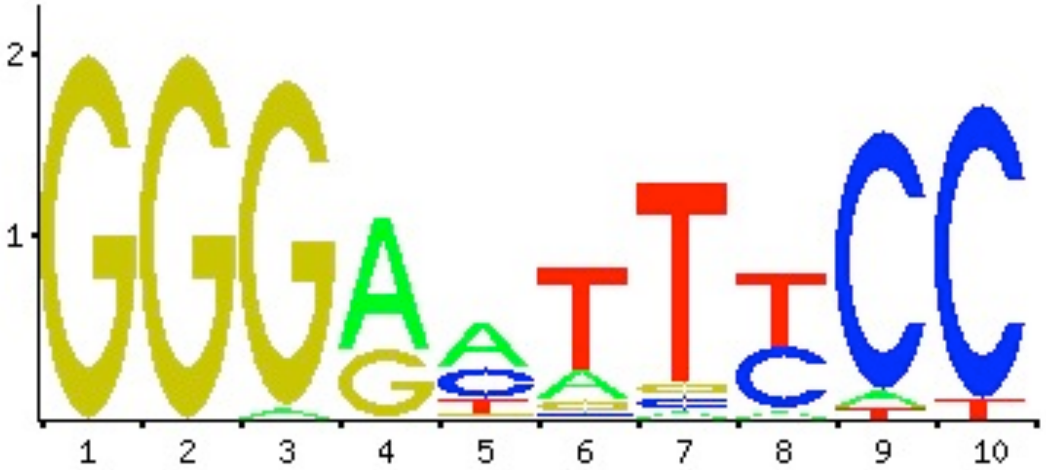
Sequence-specific affinity

Binding energy is well approximated by Position-Weight Matrices (PWM)

We assume binding via L consecutive bases, where each bond contributes an independent additive weight (log of prob.):

$$G(S) \quad = \quad \sum_{k=1}^{L} W(S_k, k) \ ,$$

$$e^W \quad = \quad \begin{array}{cc} \left(\begin{array}{cccc} 0.000000 & 0.000000 & 1.000000 & 0.000000 \\ 0.000000 & 0.000000 & 1.000000 & 0.000000 \\ 0.026316 & 0.000000 & 0.973684 & 0.000000 \\ 0.657895 & 0.000000 & 0.342105 & 0.000000 \\ 0.500000 & 0.342105 & 0.026316 & 0.131579 \\ 0.184211 & 0.026316 & 0.078947 & 0.710526 \\ 0.026316 & 0.052632 & 0.052632 & 0.868421 \\ 0.052632 & 0.447368 & 0.000000 & 0.500000 \\ 0.052632 & 0.921053 & 0.000000 & 0.026316 \\ 0.000000 & 0.947368 & 0.000000 & 0.052632 \end{array}\right) & \begin{array}{c} 1 \\ 2 \\ \\ \\ \dots \\ \\ \\ \\ \\ L{=}10 \end{array} \\ \begin{array}{cccc} \quad\text{A} \quad & \quad\text{C} \quad & \quad\text{G} \quad & \quad\text{T} \quad \end{array} & \end{array}$$

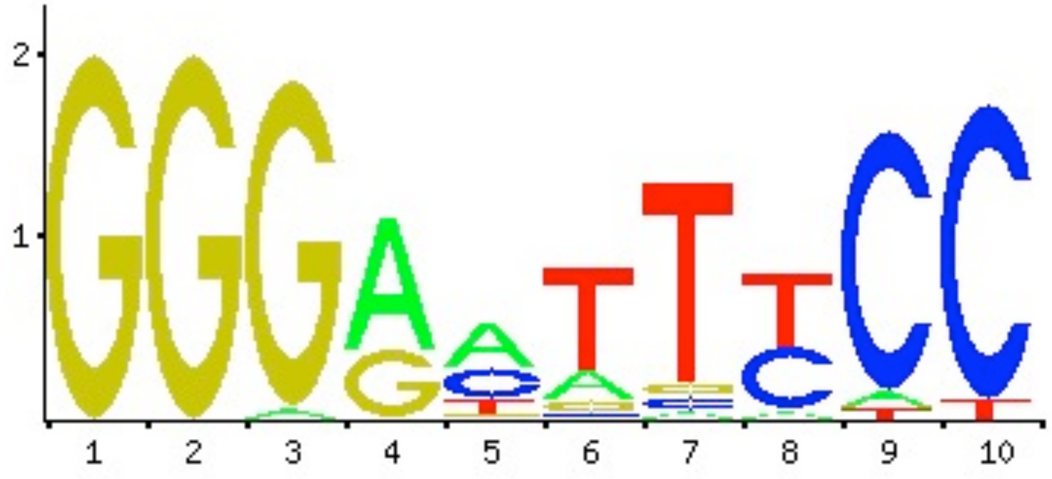


Sequence motifs and logos

Sequence logos are a way of representing graphically the PWM

- In each column (each position in the sequence) each letter is represented with a size proportional to probability (exp of weight)
- Total size is scaled to information content *I(i)*

$$I(i) \; = \sum_{\alpha \in \{A,C,G,T\}} e^{W(\alpha,i)} \log_2(e^{W(\alpha,i)}/f_\alpha)$$



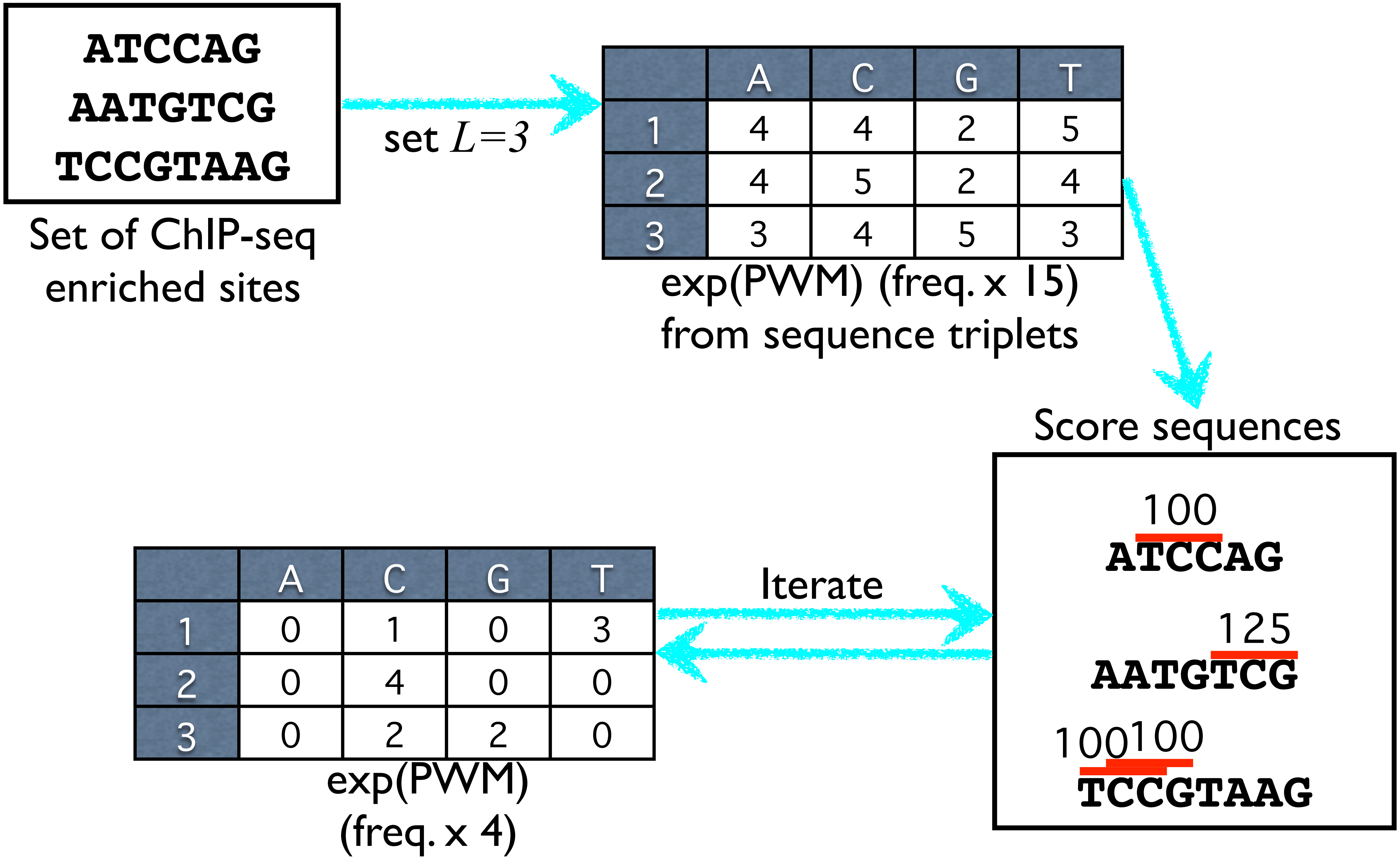
Sequence-specific affinity

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_s

$$
\begin{aligned}
P(W|S) &= \frac{P(S|W)P(W)}{P(S)}\;, \\
\log\left(\frac{P(S|W)}{P(S)}\right) &= \sum_k W(S_k,k) - \log f_{S_k}\,.
\end{aligned}
$$

EM algorithm



ATCCAG
AATGTCG
TCCGTAAG

Set of ChIP-seq enriched sites

set *L=3*

	A	C	G	T
1	4	4	2	5
2	4	5	2	4
3	3	4	5	3

exp(PWM) (freq. x 15) from sequence triplets

Score sequences

100
ATCCAG

125
AATGTCG

100100
TCCGTAAG

Iterate

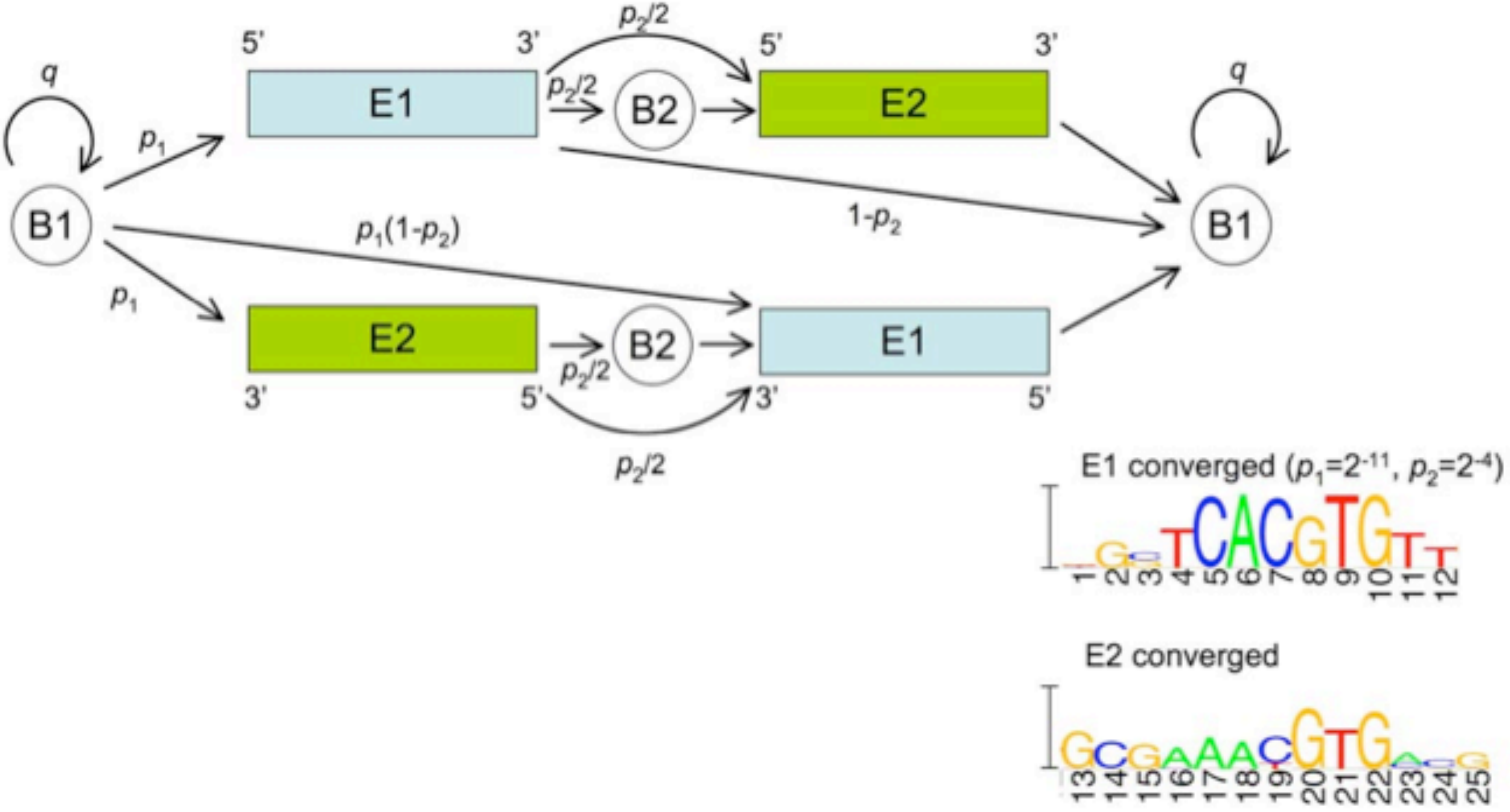
	A	C	G	T
1	0	1	0	3
2	0	4	0	0
3	0	2	2	0

exp(PWM)
(freq. x 4)

http://meme.sdsc.edu/meme/intro.html

HMMs

HMMs are particularly well adapted to modeling multiple binding sites in promoters, example, the double E-box structure of circadian promoters


q
B1
p1
p1
p1(1-p2)
5'
3'
E1
p2/2
p2/2
B2
5'
3'
E2
1-p2
q
B1
E2
3'
5'
B2
p2/2
p2/2
E1
3'
5'
E1 converged (p1=2-11, p2=2-4)
TCACGTGTT
1 2 3 4 5 6 7 8 9 10 11 12
E2 converged
GCGAAACGTG
13 14 15 16 17 18 19 20 21 22 23 24 25