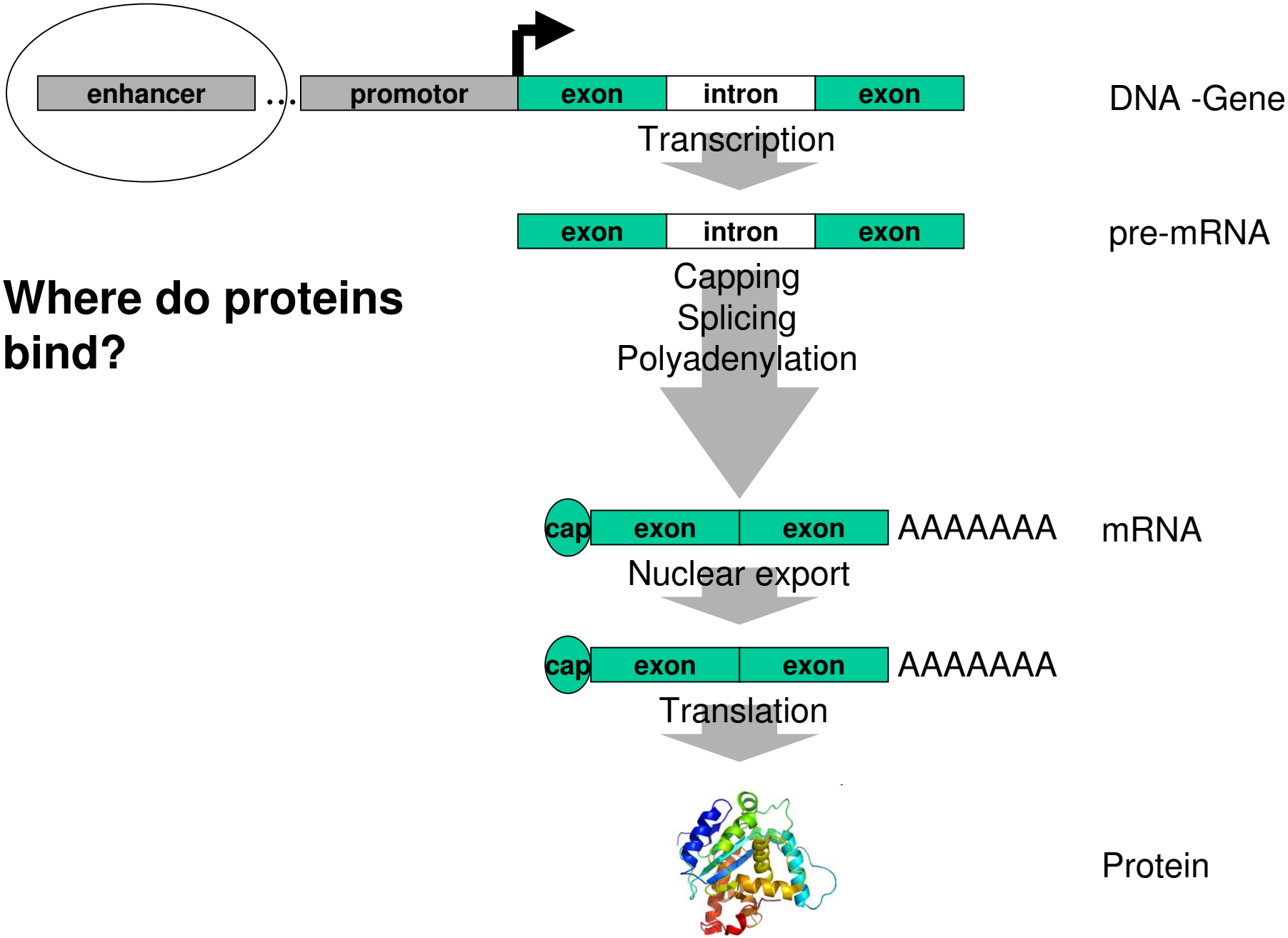
ChIP-seq

Ho-Ryun Chung

Where do proteins bind?



... X-seq

- ChIP-seq – **Ch**romatin **I**mmuno**P**recipitation

  = where are Proteins located in the genome

- DNAse-seq – DNAse I hypersensitive sites

  = *cis*-regulatory regions, enhancer, promoter, insulator, ...

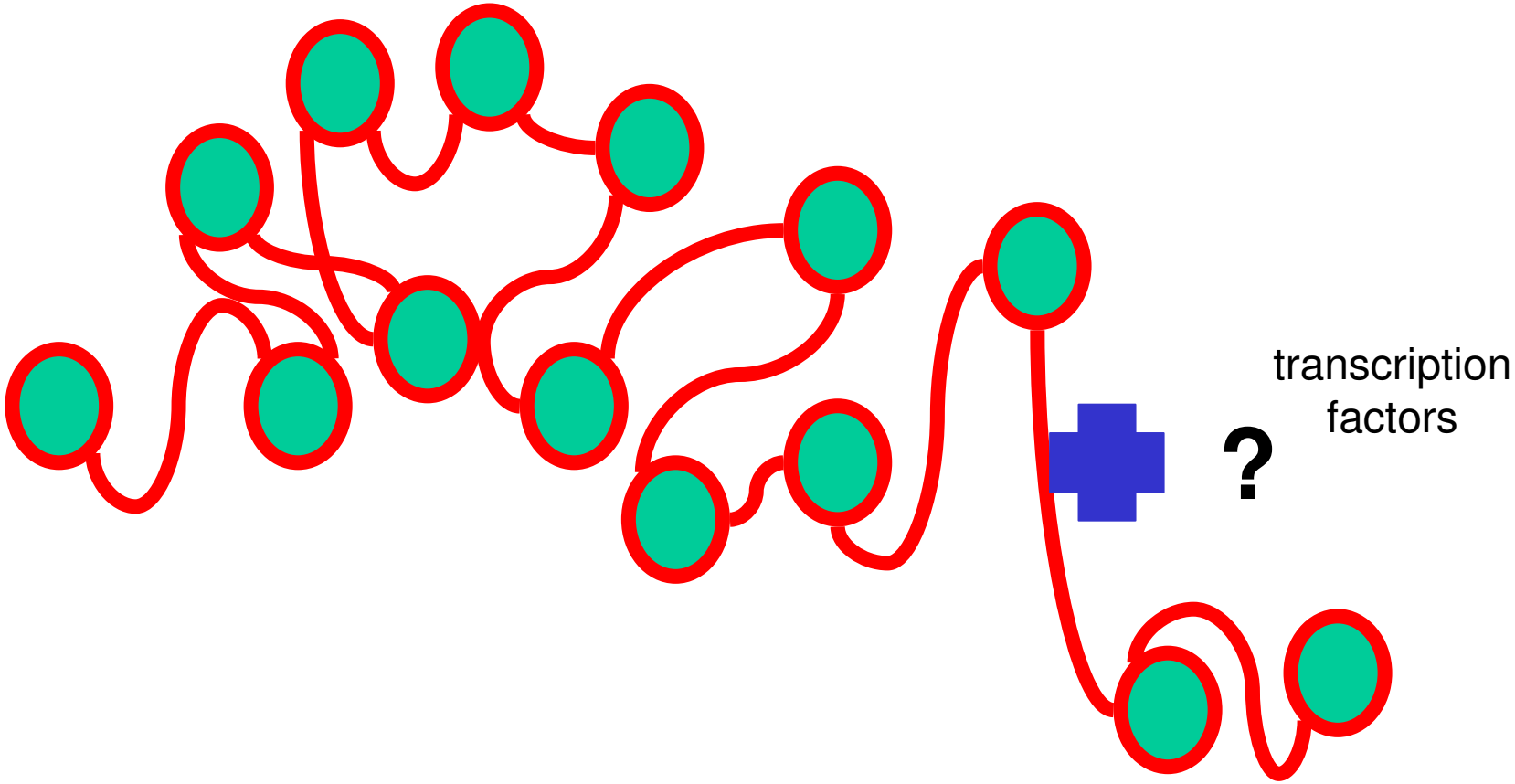
- MNAse-seq – (**M**icrococcal **N**ucle**Ase**)

  = where are nucleosomes

- Cmet-seq – DNA CpG methylation

  = which CpGs are methylated and what is the frequency of methylation

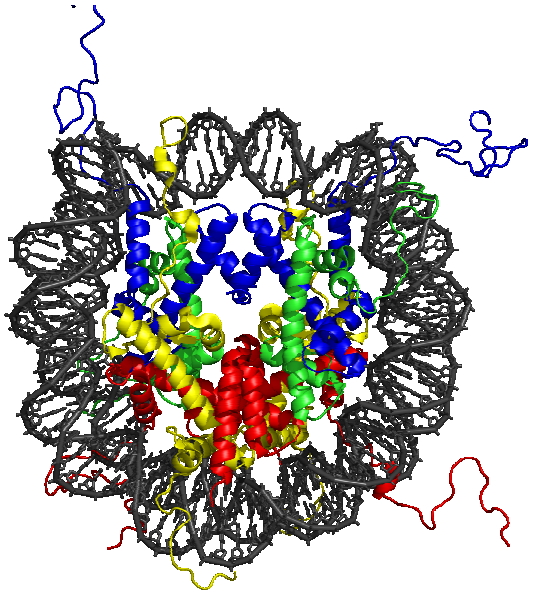
... indentify protein-bound regions

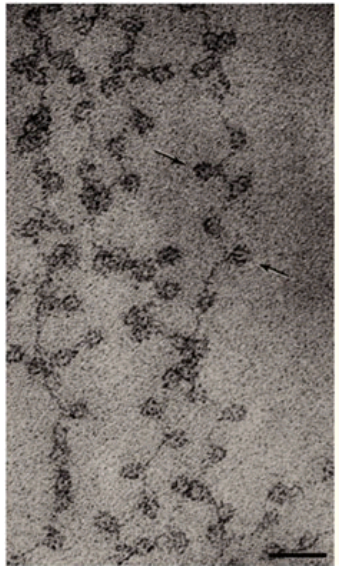


... indentify protein-bound regions

Where are transcription factors located?

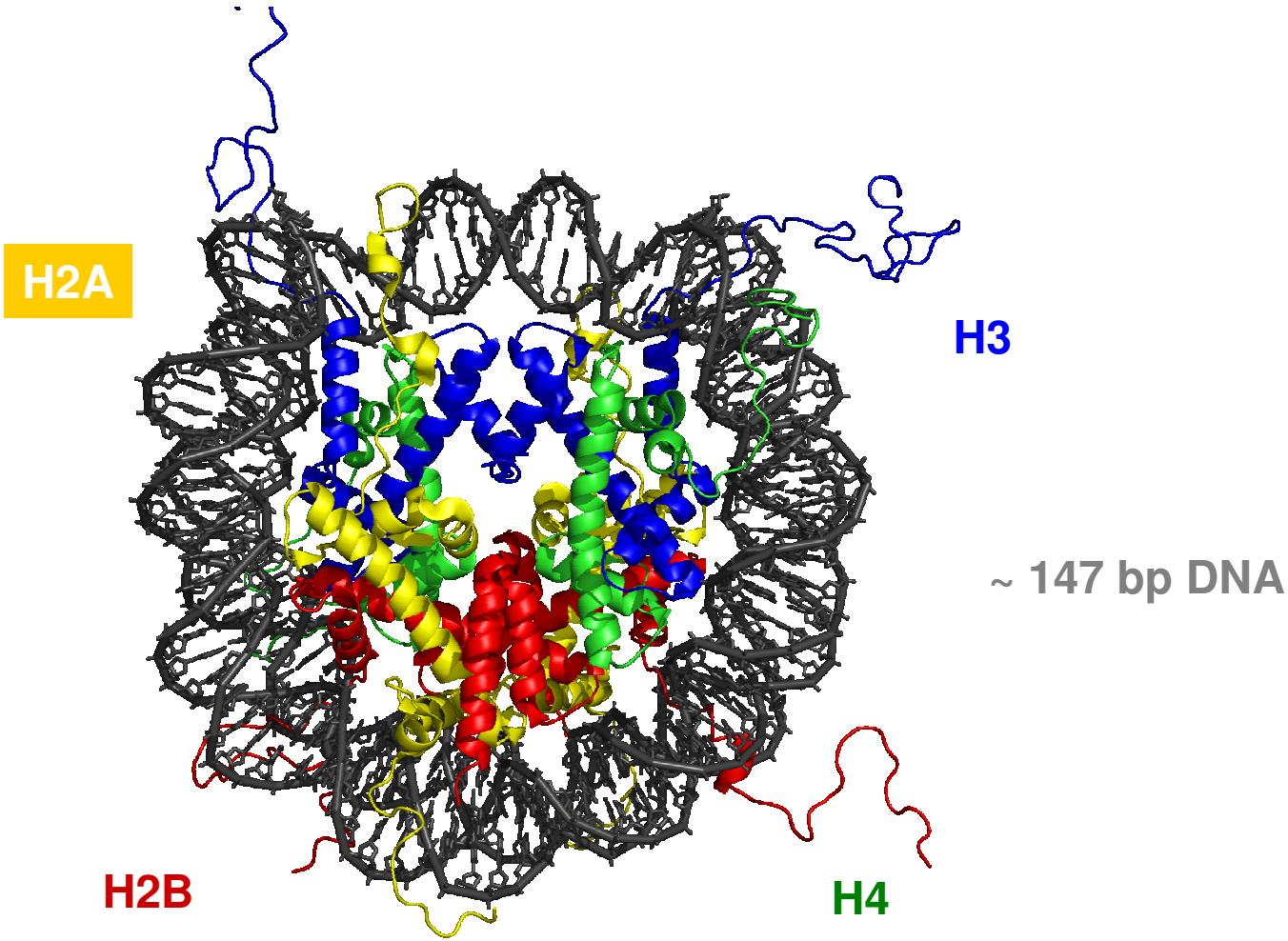
... chromatin



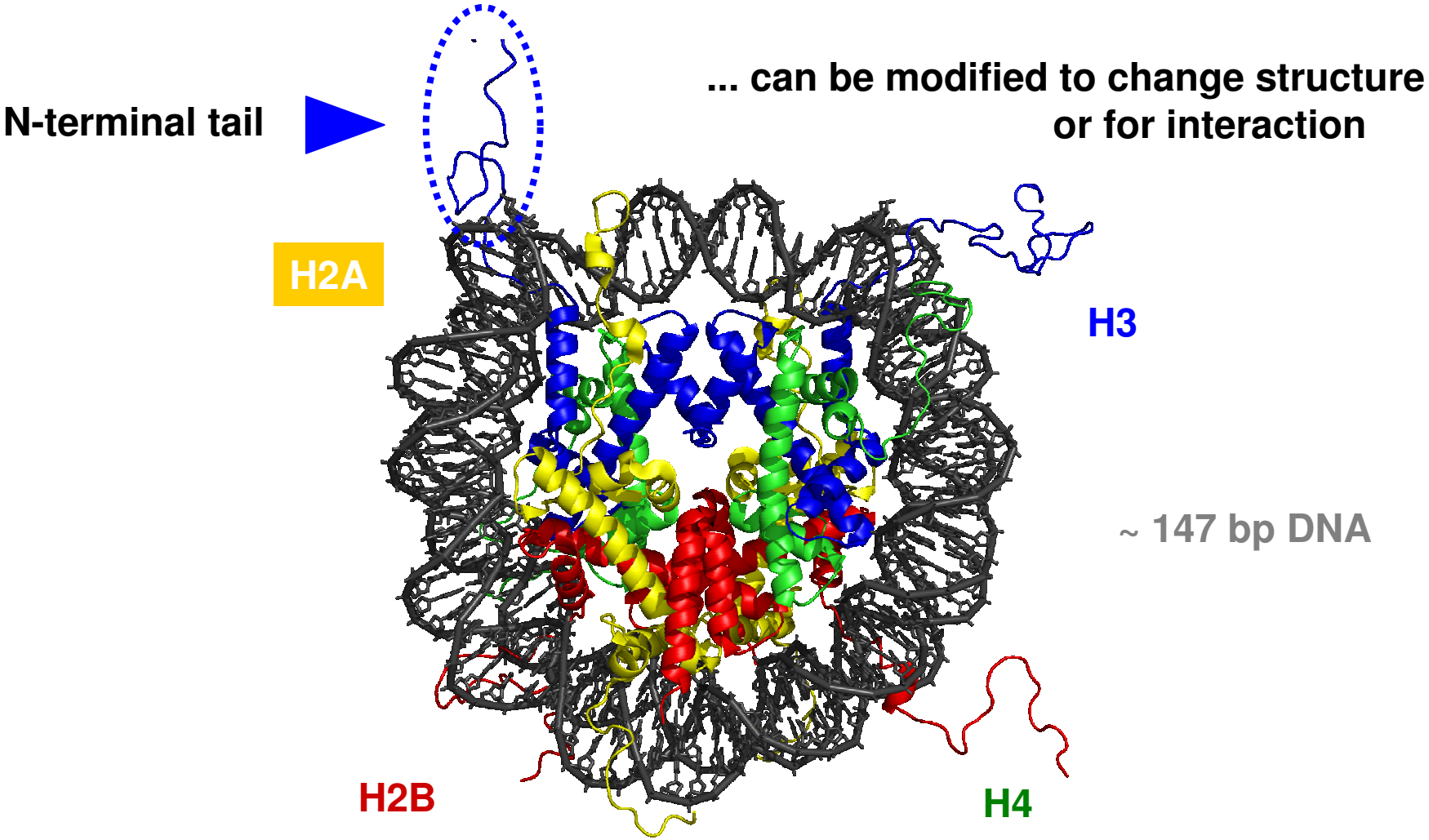


nucleosome **beads on the string** **30 nm fiber**

... nucleosome



... nucleosome and histone modifications



... histone modifications

+ phosphorylation
+ ubiquitination
+ ...

position in the sequence

amino acid

name of the histone — **H3 K 27 me3** — modification

modification	
acetylation	ac
mono-	me1
di-	me2
tri-methylation	me3
mono-ubiquitination	ub1
phosphorylation	P

... histone modifications

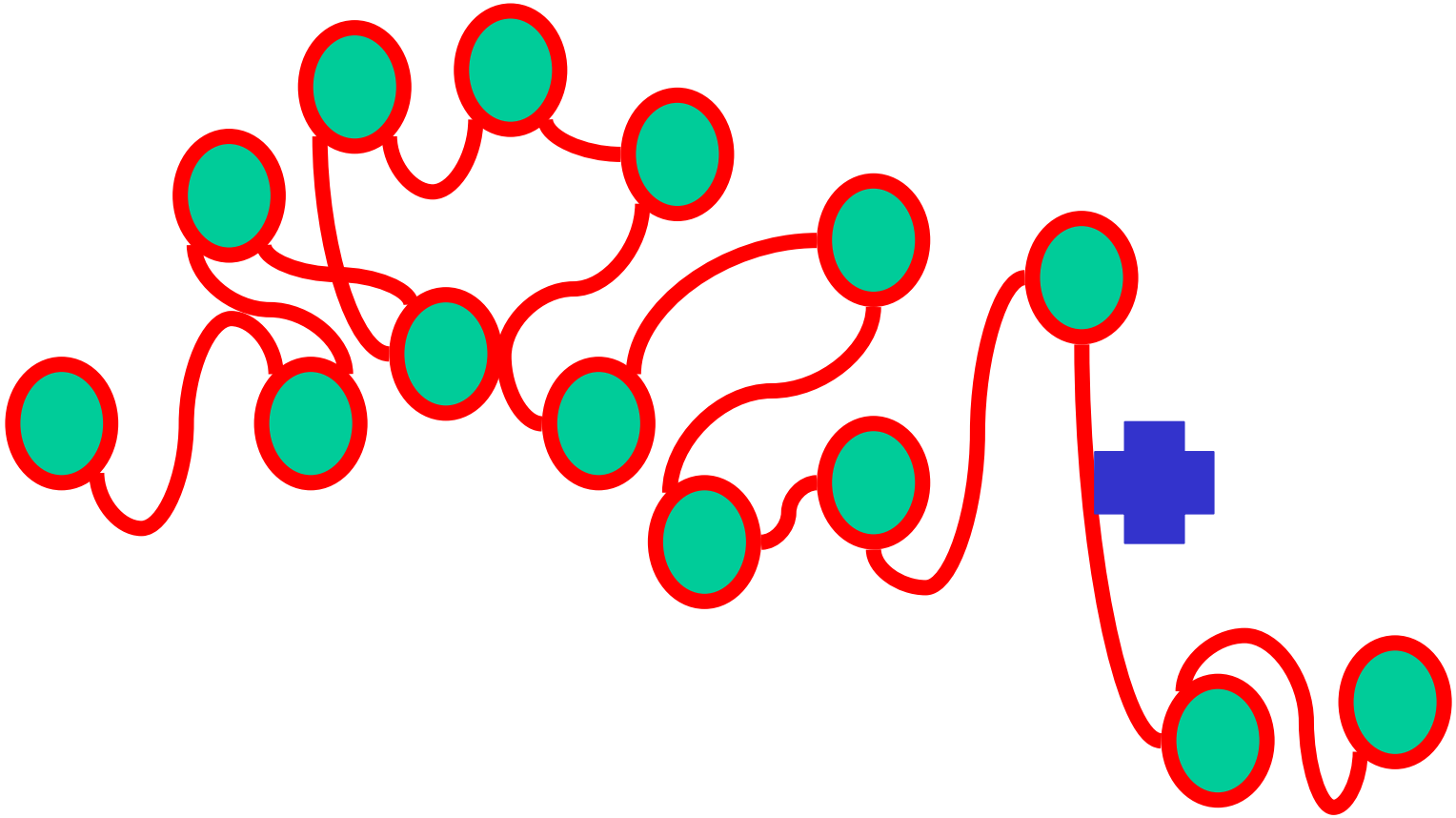
+ phosphorylation
+ ubiquitination
+ ...

Where are histone modifications located?

name of the histone — **H3 K 27 me3** — modification

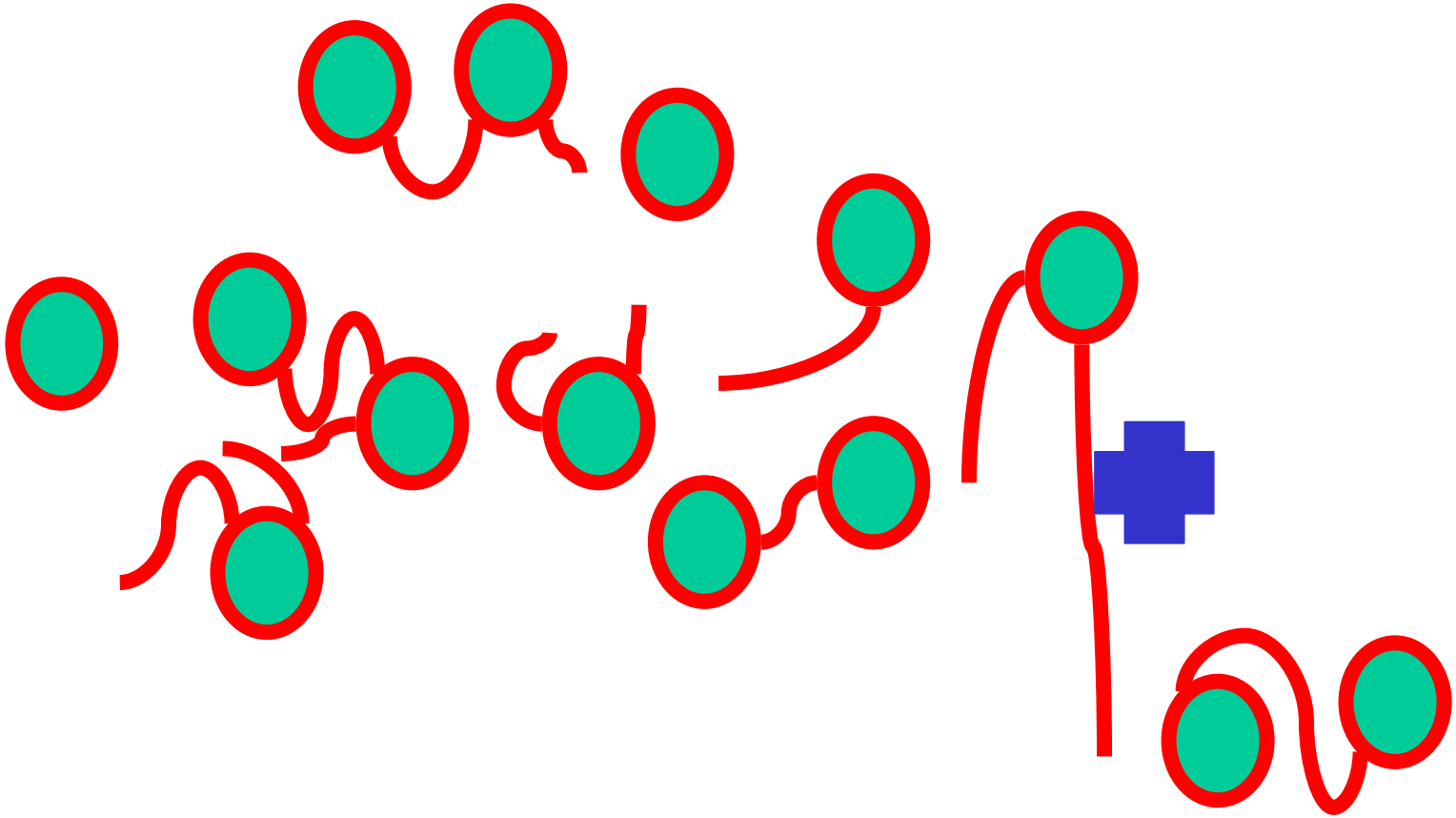
modification	
acetylation	ac
mono-	me1
di-	me2
tri-methylation	me3
mono-ubiquitination	ub1
phosphorylation	P

... 1 make sure the protein sticks to the DNA



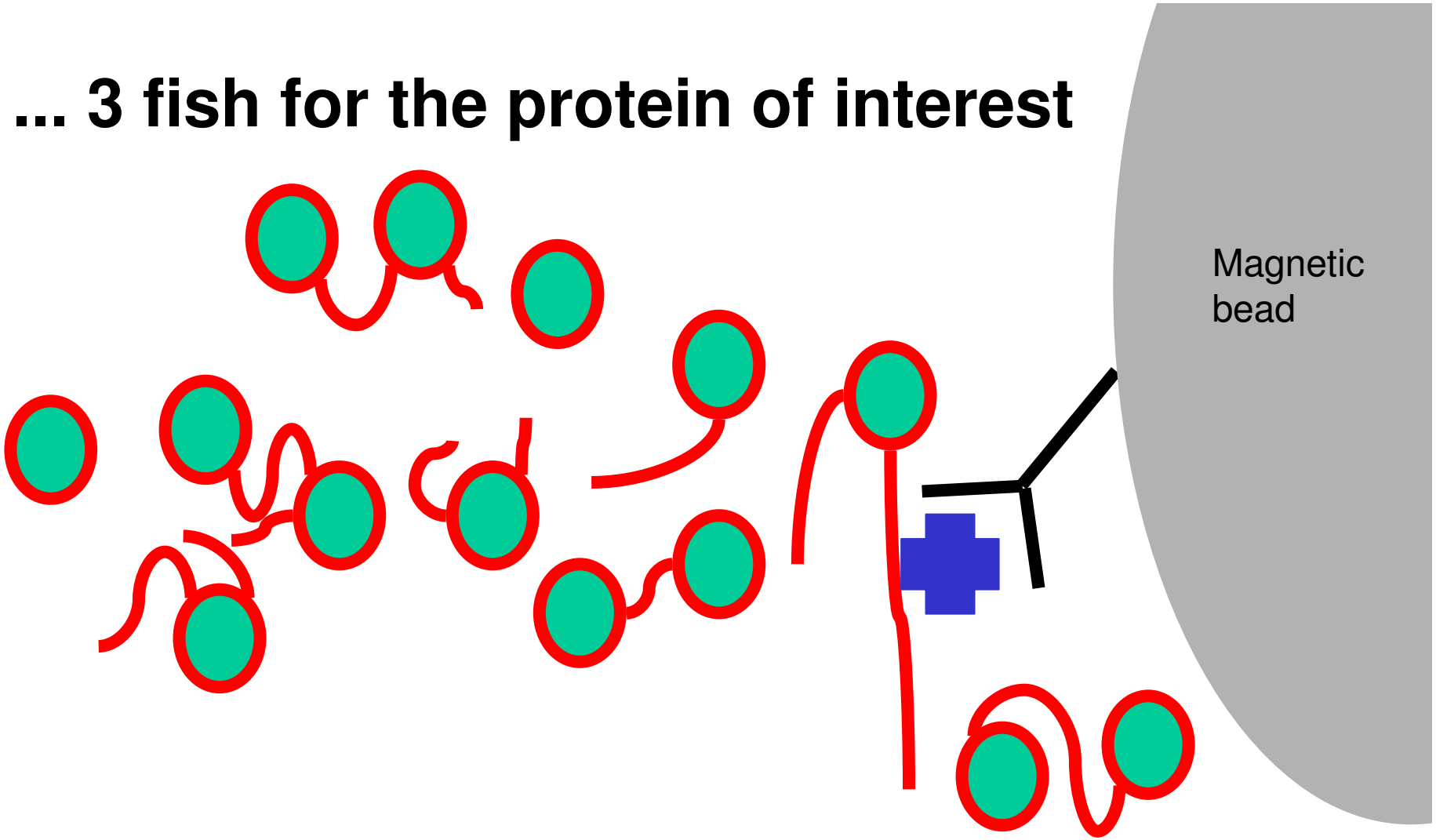
crosslinking – use formaldehyde

... 2 break chromatin into small pieces



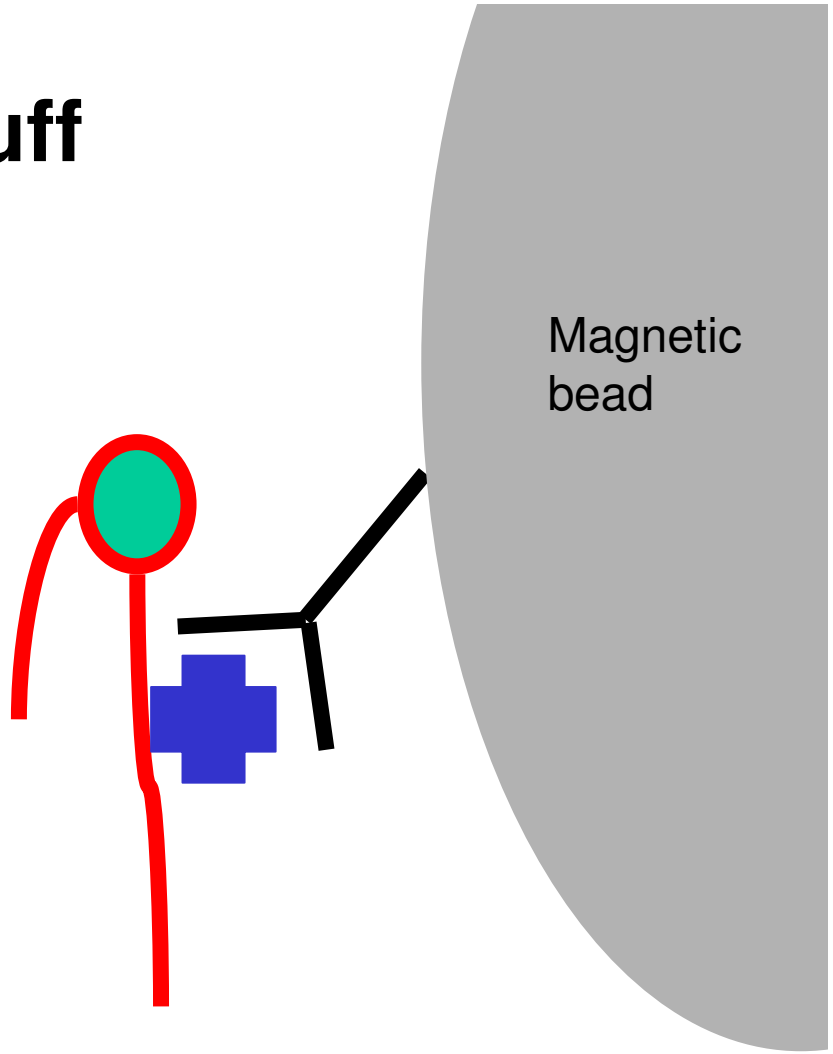
fragmentation – use sound
= sonification

... 3 fish for the protein of interest



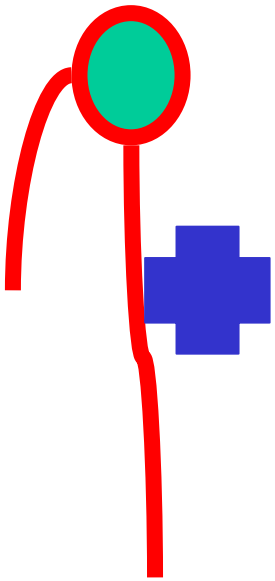
immunoprecipitation – use antibody

... 4 remove unbound stuff

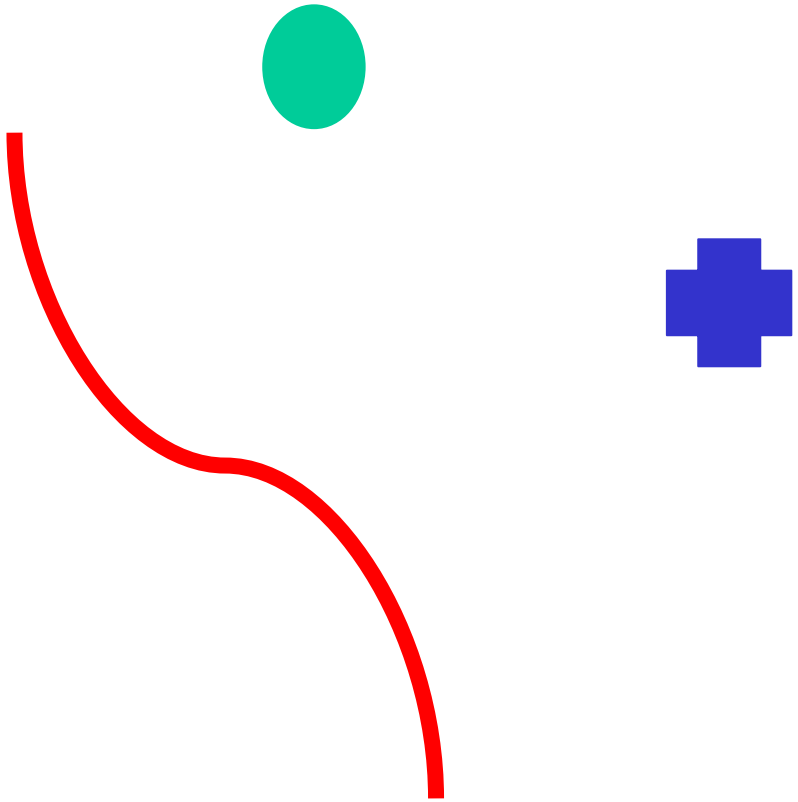


washing

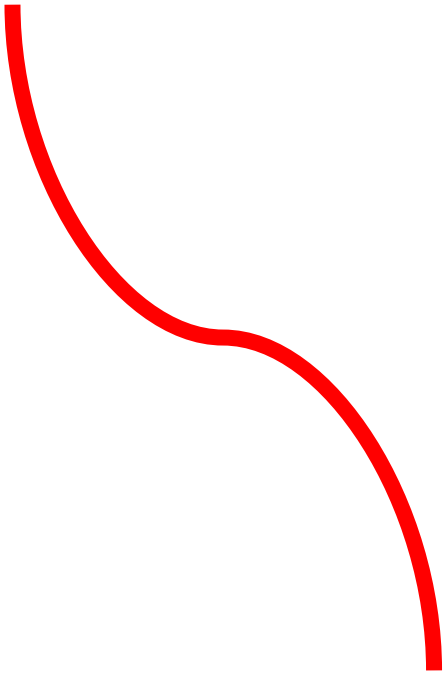
... 5 get unbound stuff



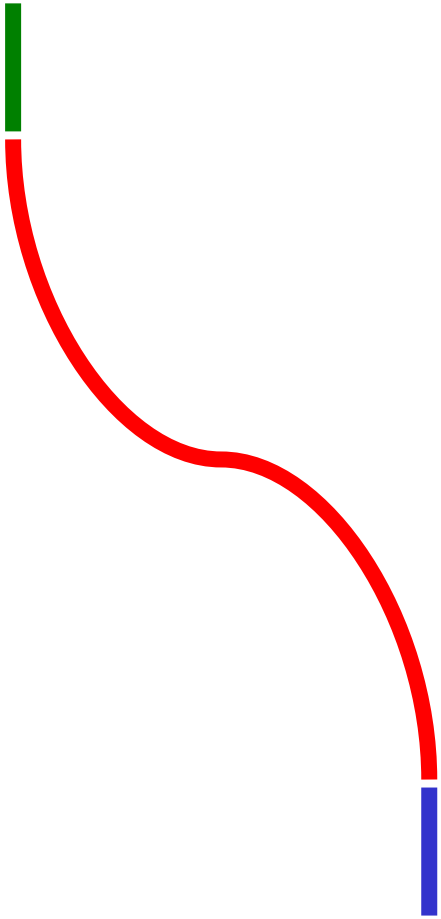
... 6 reverse crosslinks



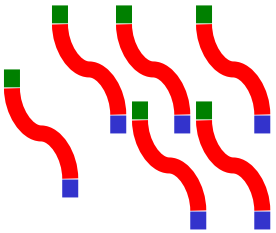
... 7 isolate DNA



... 8 ligate sequencing adaptors



... 9 sequence

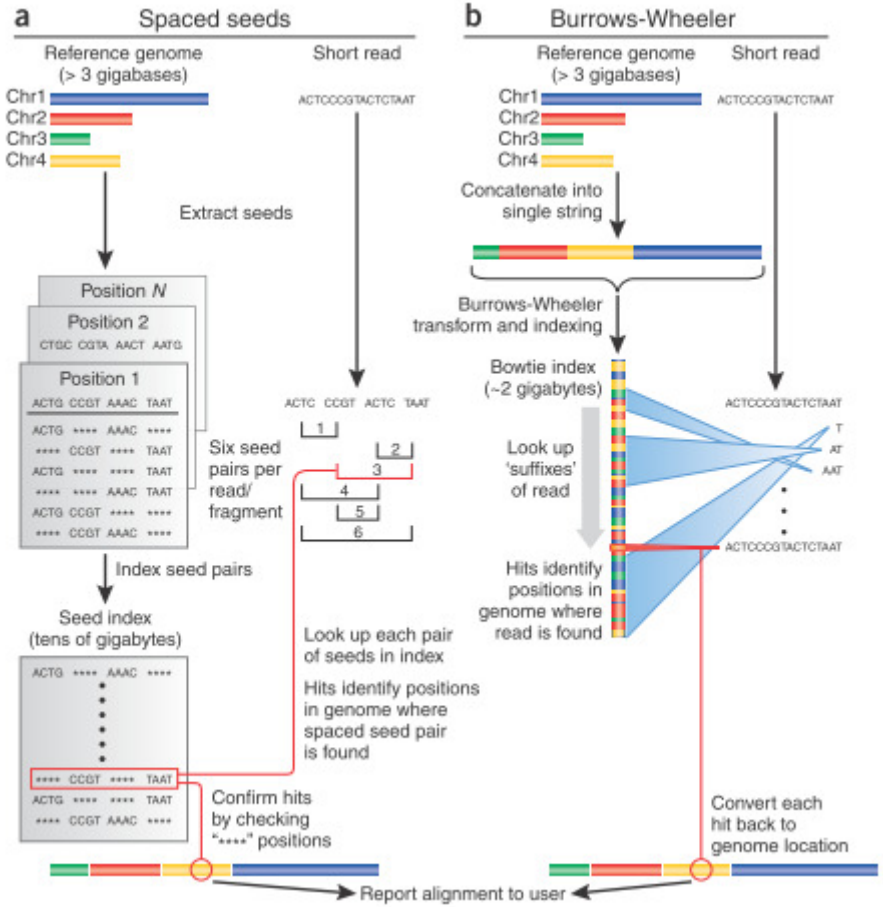


... 10 map tags against genome

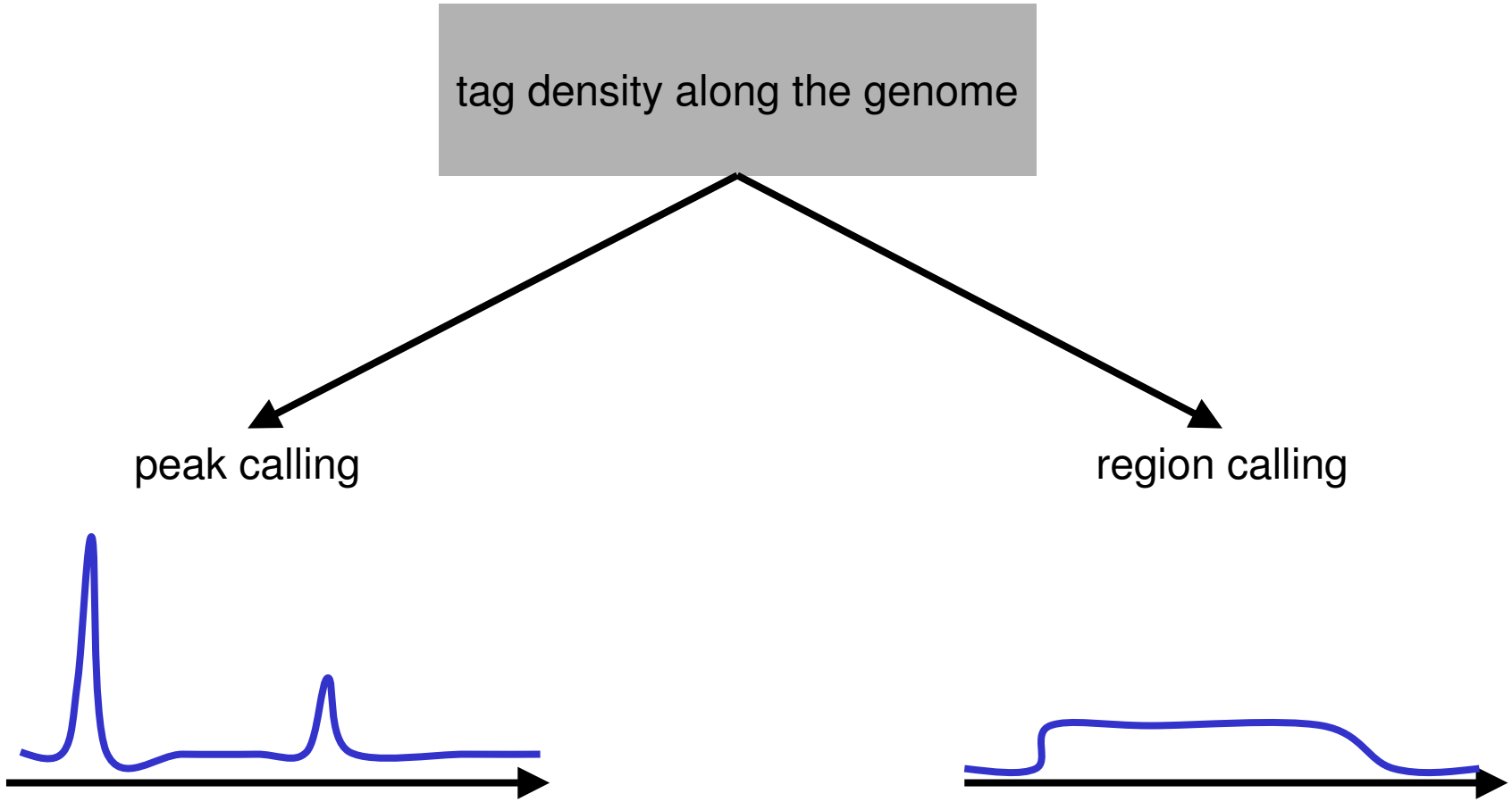
- bowtie (Langmead et al. 2009)
- BWA (Li & Durbin 2009)
- ....

low quality reads?

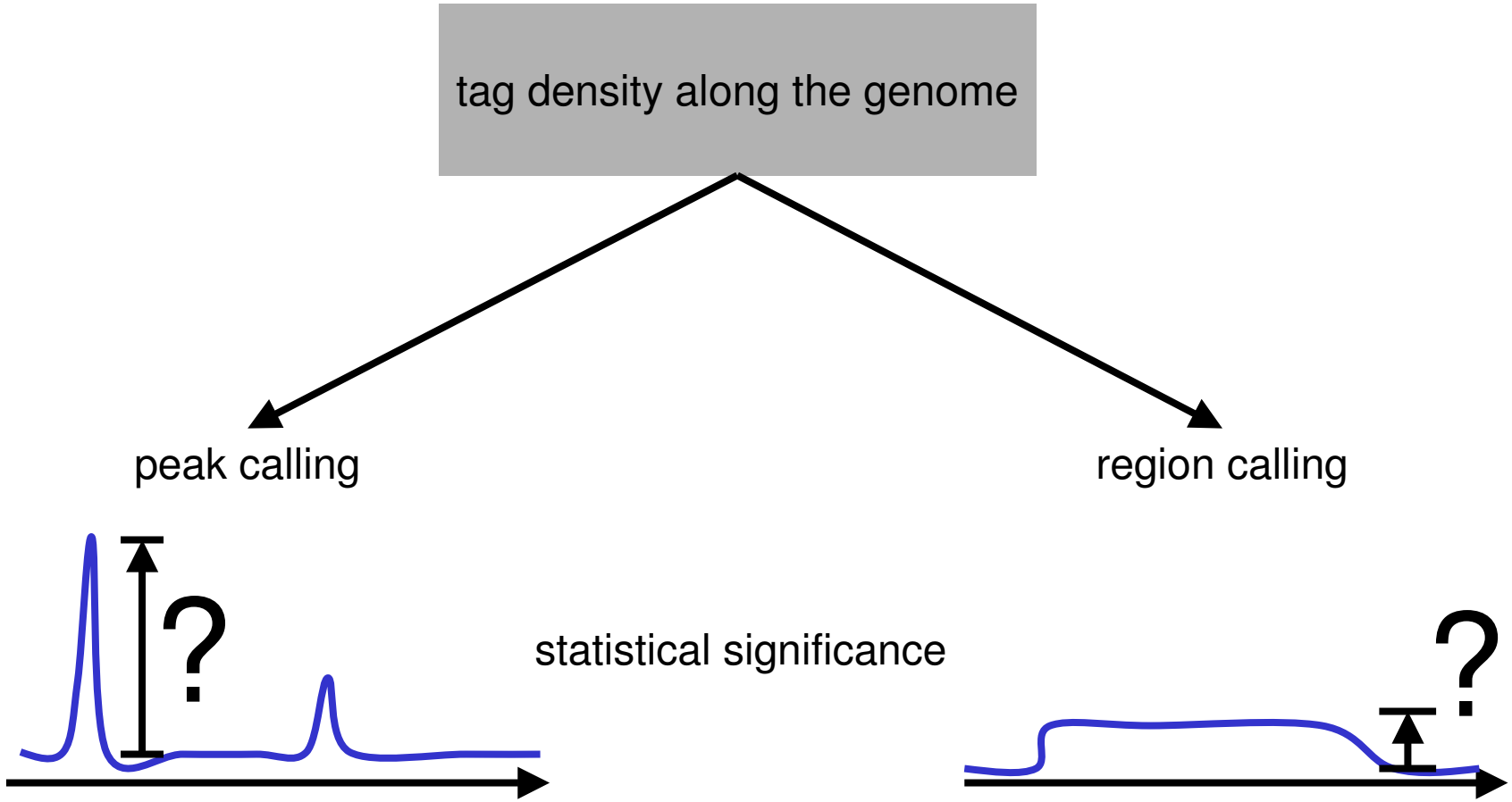
non-unique matches?



... 11 analyze data



... 11 analyze data



... the multiple testing problem

Suppose you carry out 10 hypothesis tests at the 5% level
(assume independance)

- the probability of declaring a particular test significant under its null hypothesis is 0.05
- but the probability of declaring at least 1 of the 10 tests significant is $\boxed{0.401}$ → $1 - 0.95^{10}$
- if you perform 20 hypothesis tests, this probability increases to $\boxed{0.642}$... → $1 - 0.95^{20}$

genome-wide we are testing millions of hypothesis! ($1 - 0.95^{1,000,000} \approx 1$)

... the multiple testing problem

Testing m null hypothesis H_j ($j = 1,\ldots,M$)
M_0 true and M_1 false null hypothesis;
R: # H_0 rejected (false positives)

unobservable random variables

	# H_0 not rejected	# H_0 rejected	
# true H_0	U	V	M_0
# false H_0	T	S	M_1
	M - R	R	M

observable quantity

known quantity

... Benjamini-Hochberg

$$\mathrm{FDR} = \begin{cases} E\left(\dfrac{V}{R}\right) & R > 1 \\ 0 & R = 0 \end{cases}$$

unobservable random variables

	# H_0 not rejected	# H_0 rejected	
# true H_0	U	V	M_0
# false H_0	T	S	M_1
	M - R	R	M

observable quantity

... Benjamini-Hochberg – worked-out

- order the p-values such that $p_1 \leq p_2 \leq \cdots \leq p_M$
- calculate

$$\hat{k} = \arg\max_{1 \leq k \leq M} \left\{ k : p_k \leq \alpha \frac{k}{M} \right\}$$

- then rejecting the null hypotheses corresponding to $p_1, \ldots, p_k$ provides FDR = $M_0/M \cdot \alpha \leq \alpha$.

	# H_0 not rejected	# H_0 rejected	
# true H_0	U	V	M_0
# false H_0	T	S	M_1
	M - R	R	M

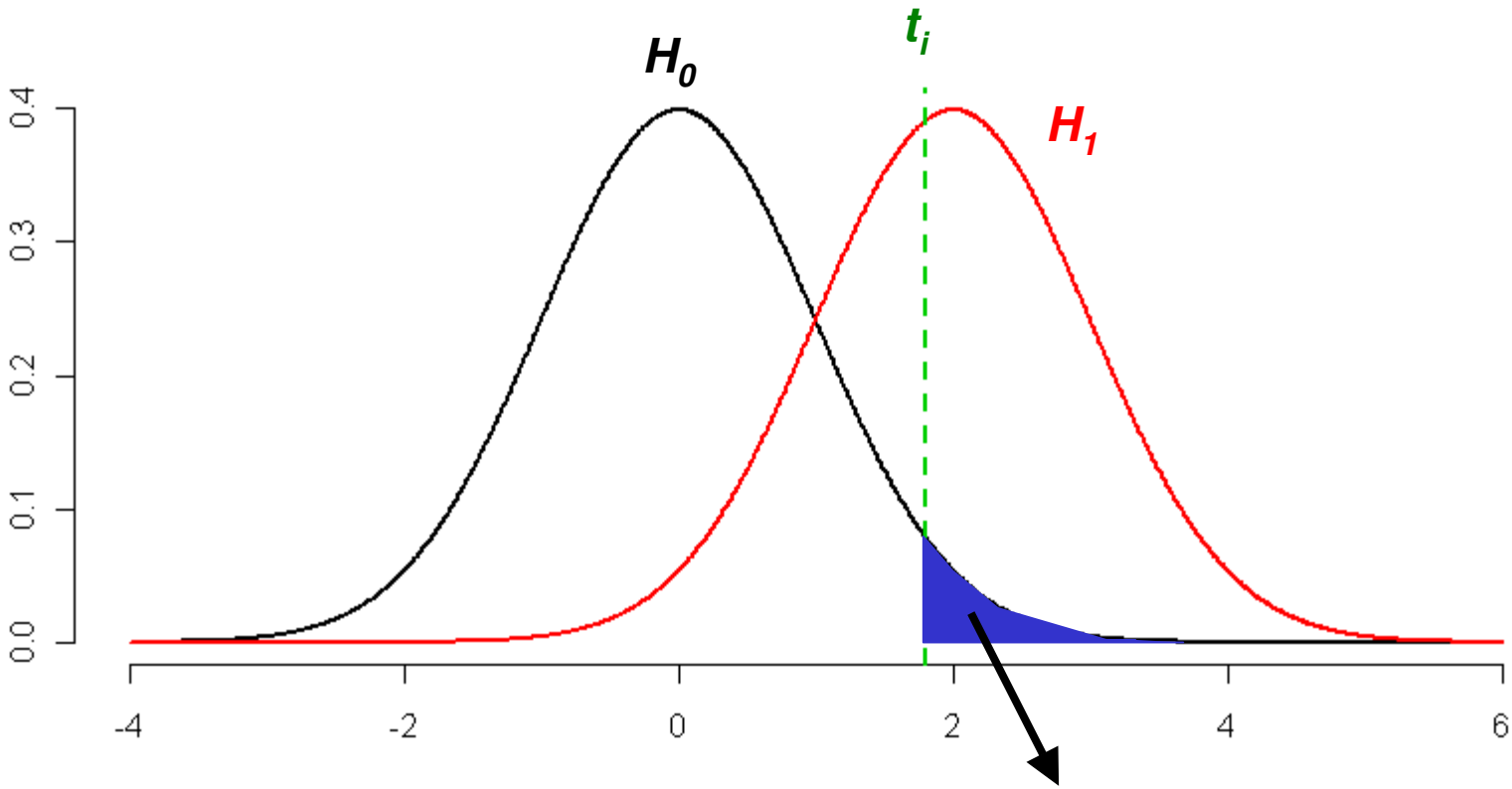
... the q-value

THEOREM 1. *Suppose m identical hypothesis tests are performed with the statistics $T_1, \ldots, T_m$ and significance region Γ. Assume that (T_i, H_i) are i.i.d. random variables, $T_i | H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$ for some null distribution F_0 and alternative distribution F_1, and $H_i \sim \text{Bernoulli}(\pi_1)$ for $i = 1, \ldots, m$. Then*

$$\begin{aligned}
\text{pFDR}(\Gamma) &= E\left(\frac{V(\Gamma)}{R(\Gamma)} \middle| R(\Gamma) > 0\right) \\
&= P(H = 0 | T \in \Gamma) \\
&= \frac{\pi_0 P(T \in \Gamma \mid H = 0)}{\pi_0 P(T \in \Gamma \mid H = 0) + \pi_1 P(T \in \Gamma \mid H = 1)}
\end{aligned}$$

$$\begin{aligned}
q - \text{value}(t) &= \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \text{pFDR}(\Gamma_\alpha) \\
&= \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} P(H = 0 \mid T \in \Gamma_\alpha)
\end{aligned}$$

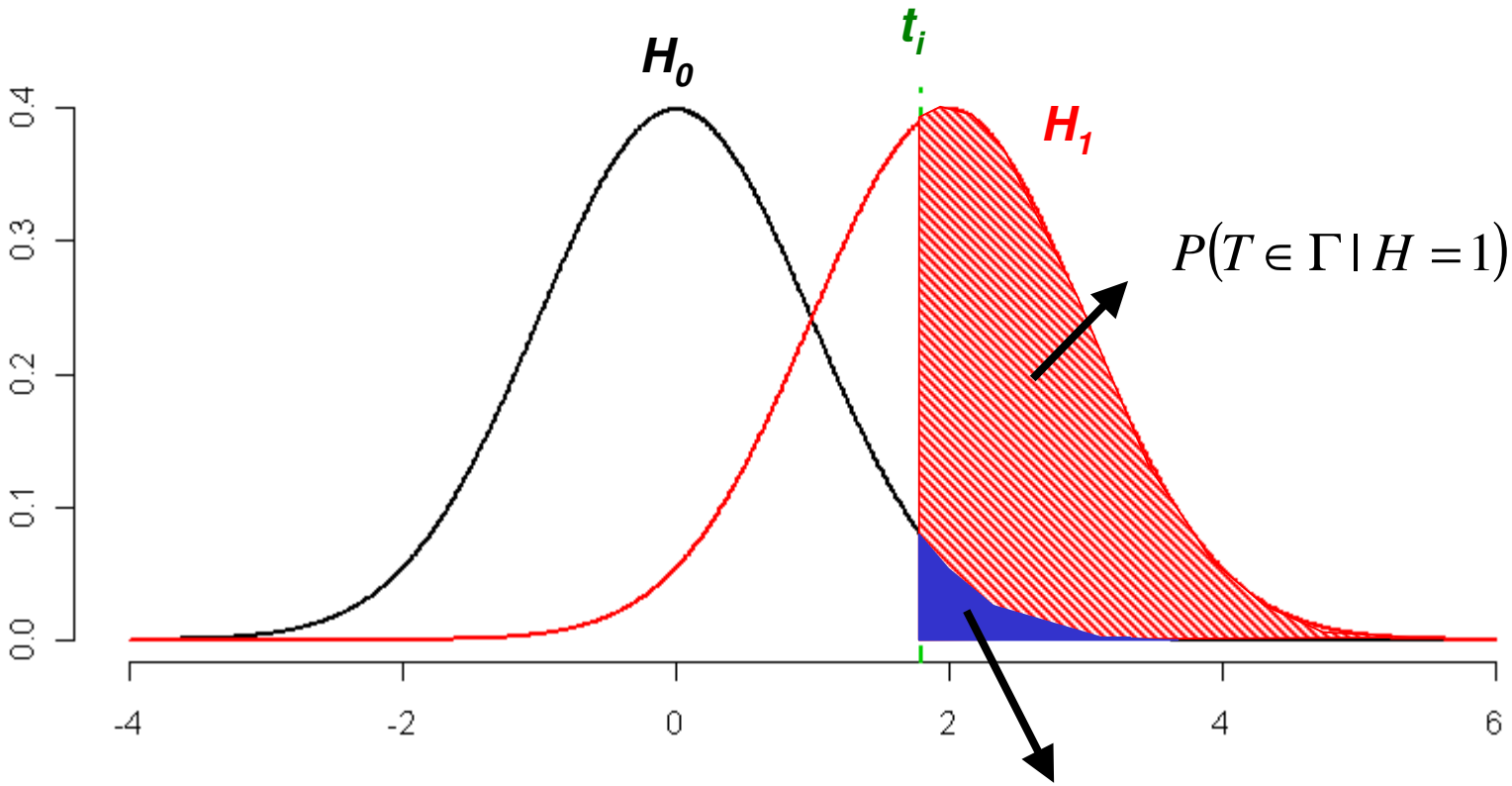
... the p-value



p-value
or type I error
$P(T \in \Gamma \mid H = 0)$

... the q-value

$$\frac{\pi_0 P(T \in \Gamma \mid H = 0)}{\pi_0 P(T \in \Gamma \mid H = 0) + \pi_1 P(T \in \Gamma \mid H = 1)}$$



$P(T \in \Gamma \mid H = 1)$

p-value
or type I error
$P(T \in \Gamma \mid H = 0)$

... what determines tag density variation in one ChIP sample?

- probability that a region is bound to the target
- mappability
  - repeat regions => no unique hits
  - repeat-like regions => false positive hits due to sequencing errors
- efficiency of fragmenting chromatin
- efficiency of sequencing

... controls

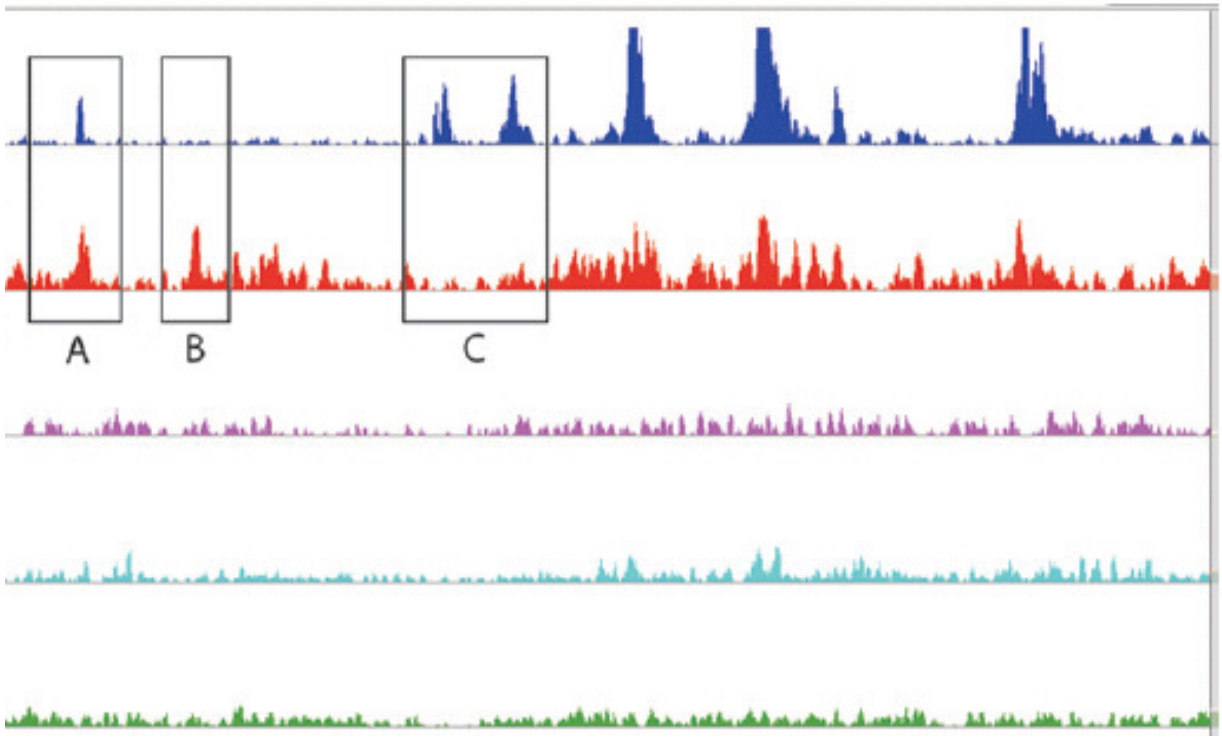
PolII-ChIP

Input (100-350 bps)

Input (350-800 bps)

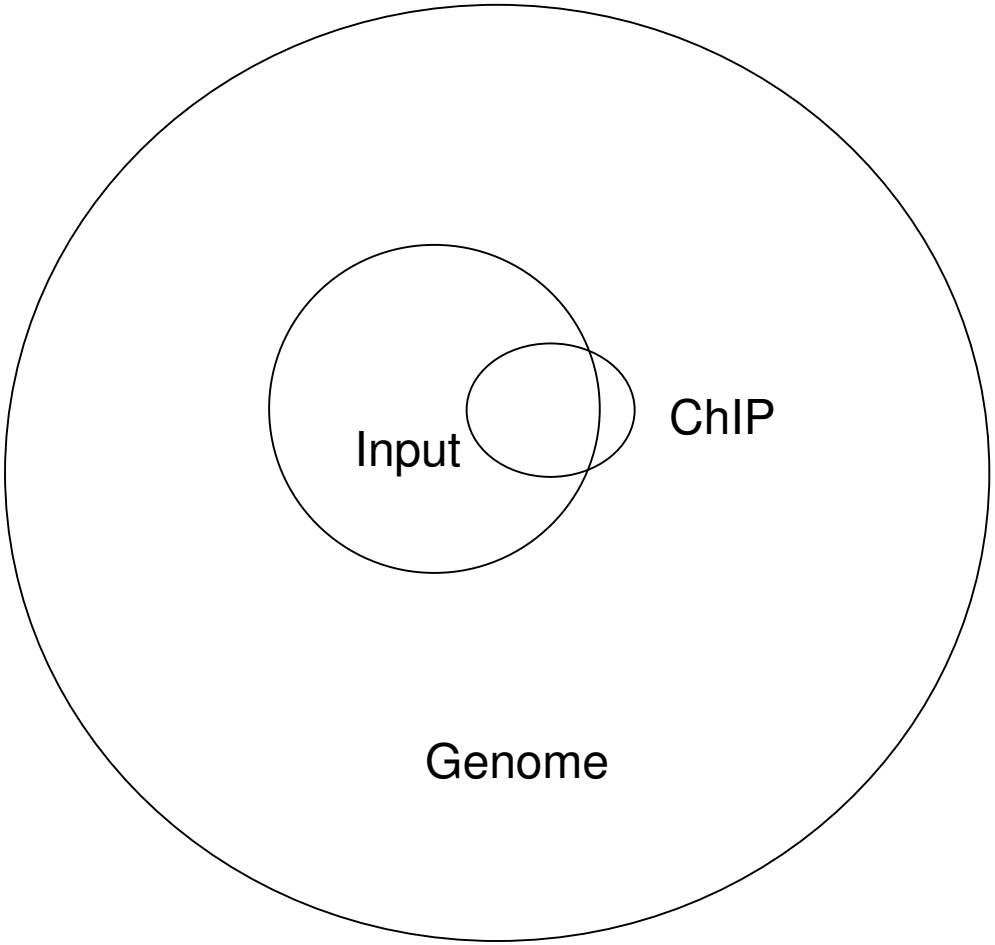
Mock IP

genomic resequencing

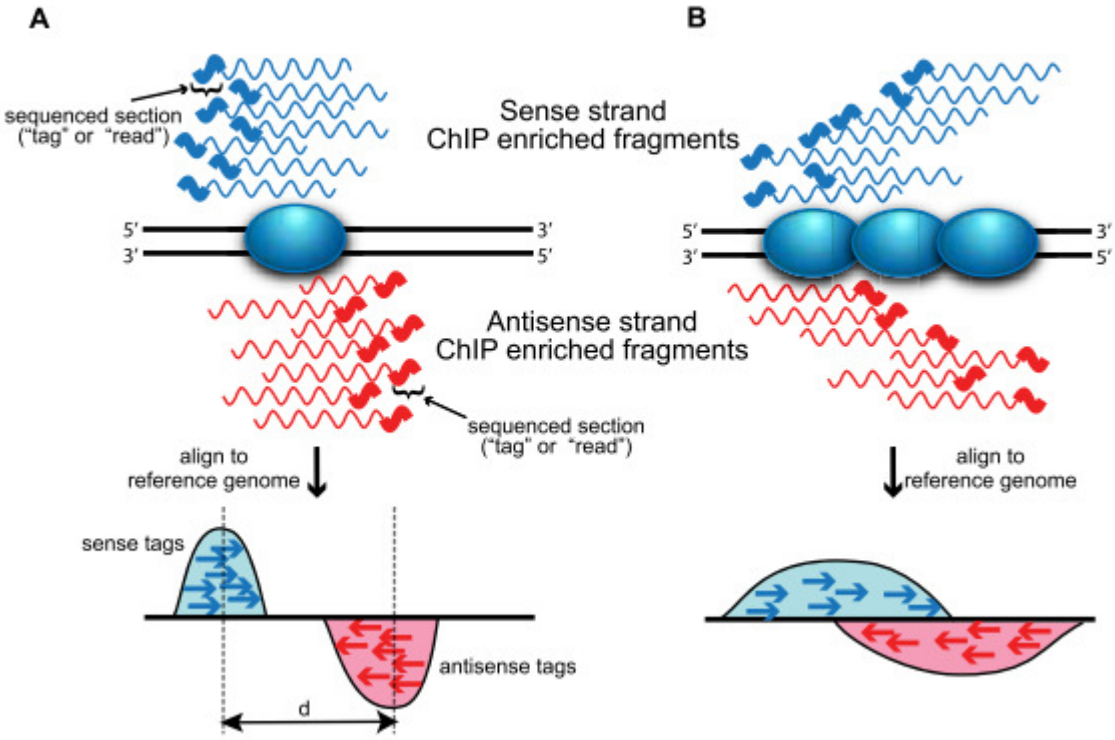


Auerbach et al. (2009)

... is input a good choice?

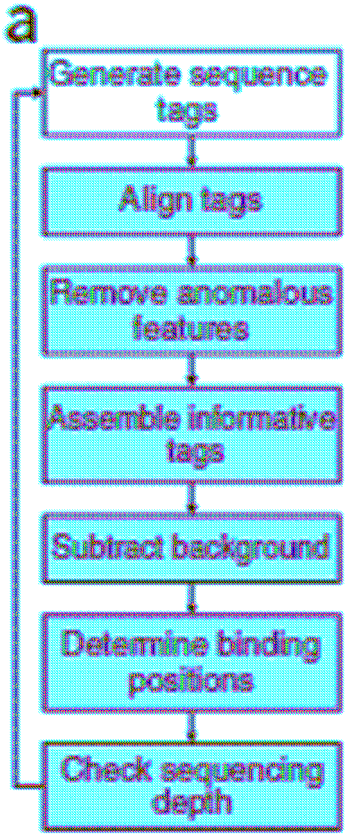


... strand dependent bimodality



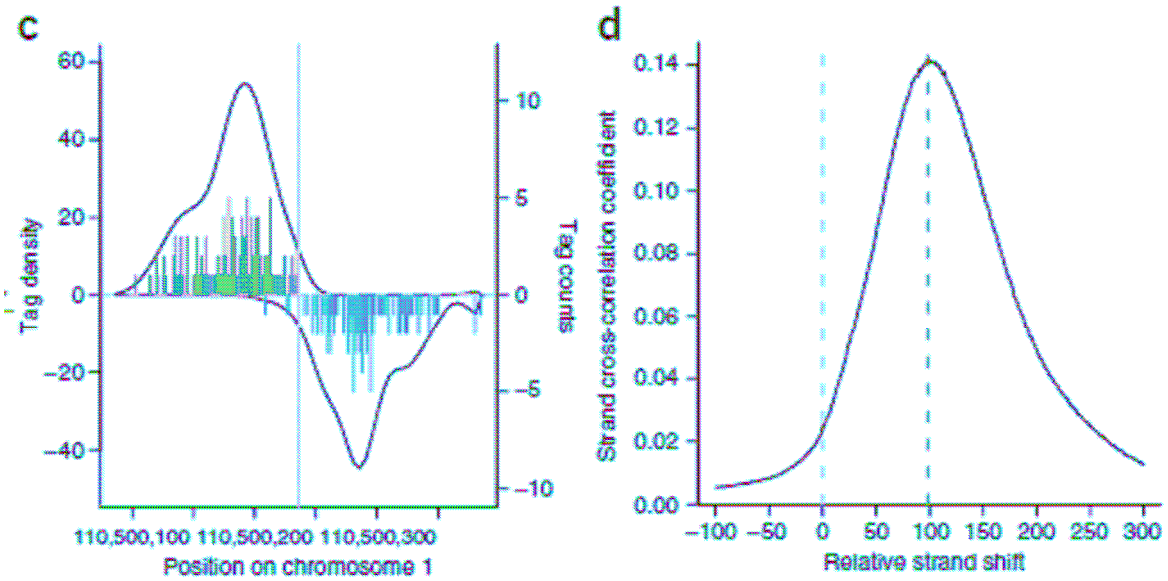
Wilbanks et al. (2010)

... ssp



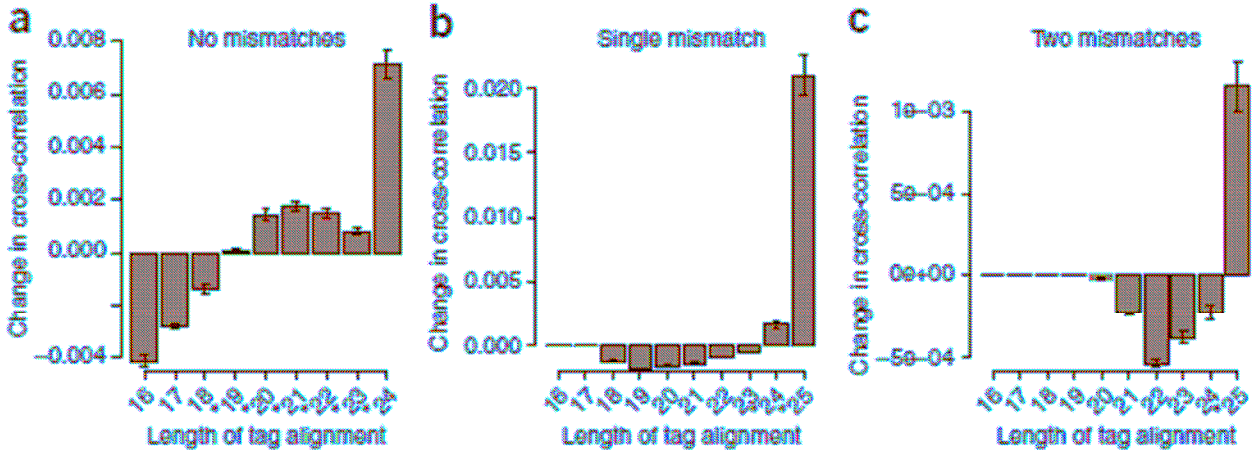
Kharchenko et al. (2008)

... shift the tags



Kharchenko et al. (2008)

... allow gaps and mismatches?

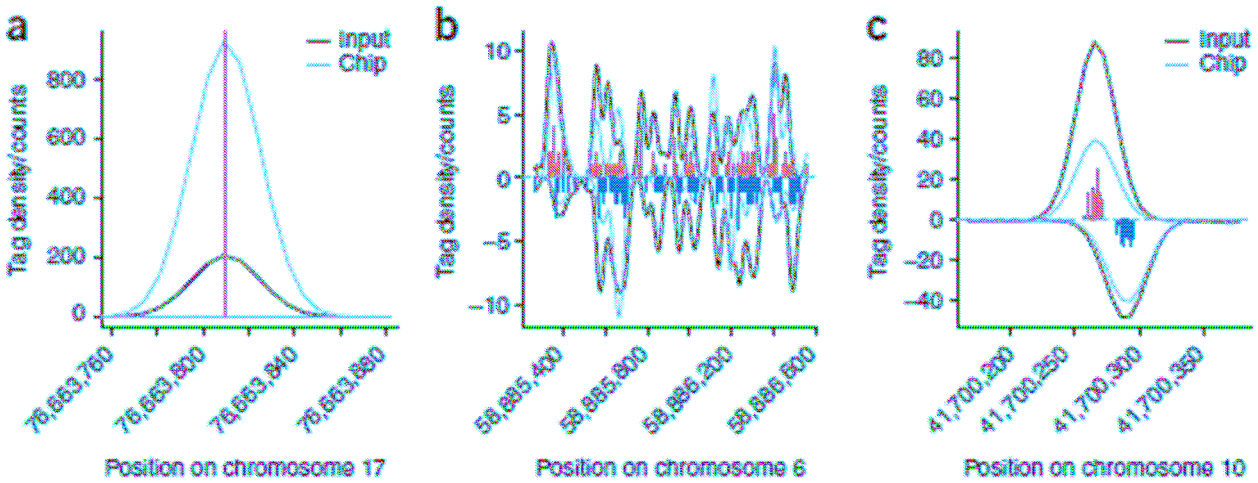


Kharchenko et al. (2008)

... background anomalies

singular high
tag density
only one strand

Background tag density
patterns resembling
true protein-binding positions



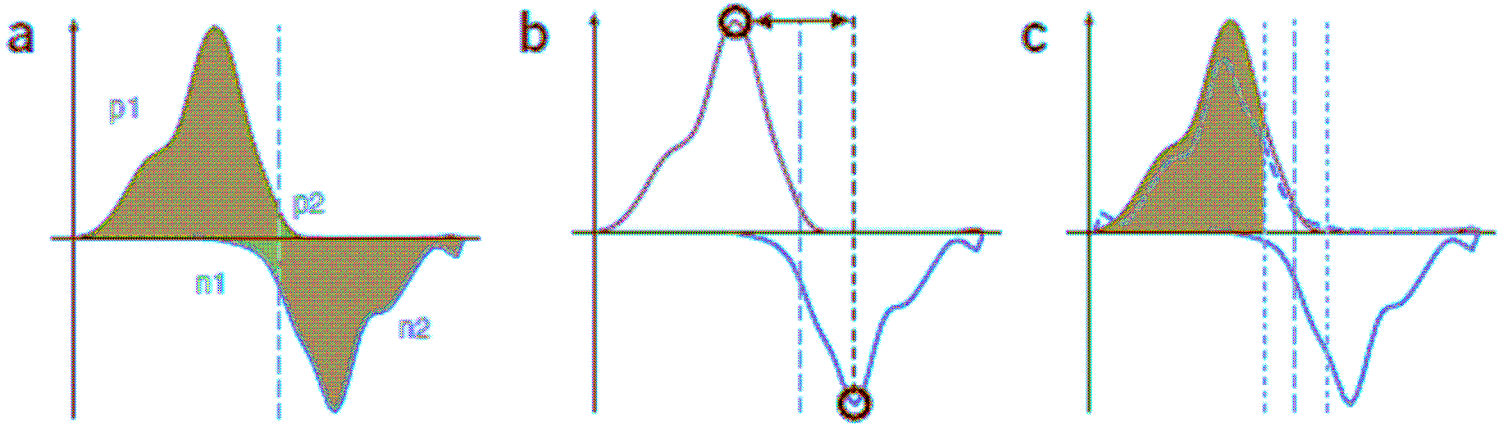
Larger, nonuniform
regions of increased
background tag density

Kharchenko et al. (2008)

... peak calling

Window tag density
(WTD)

Mirror tag correlation
(MTC)



(p1 + n2) / (n1 + p2)

Matching strand peaks
(MSP)

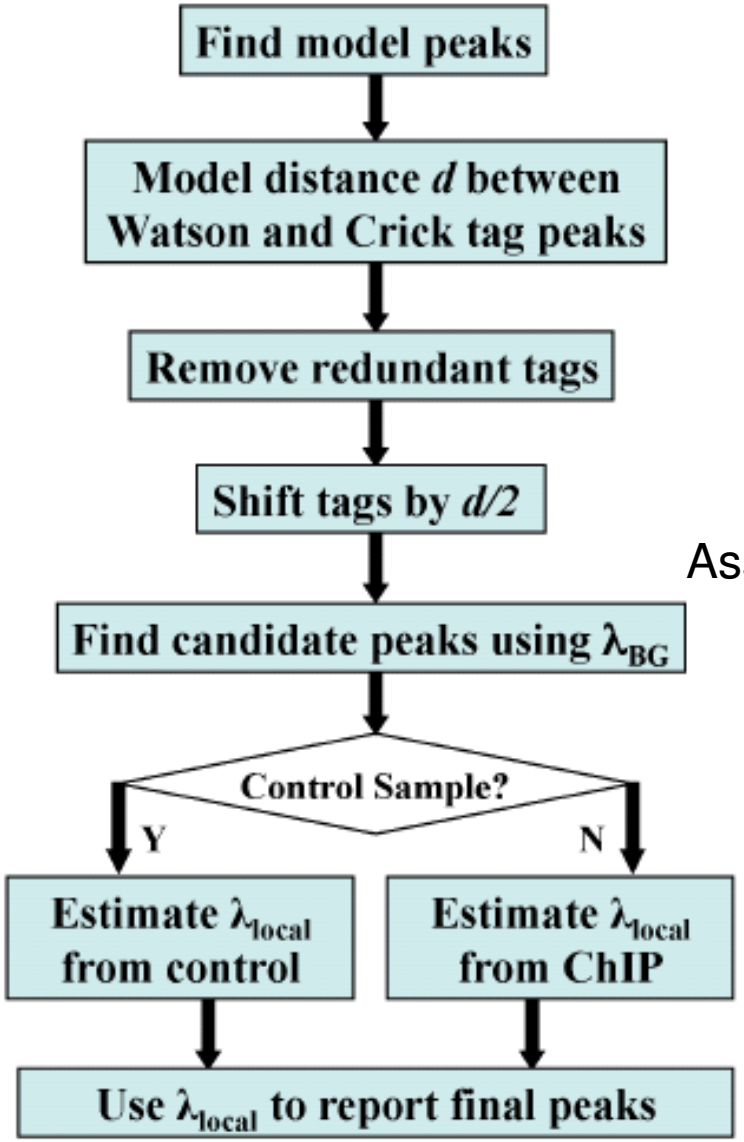
Kharchenko et al. (2008)

... FDR?

$$\mathrm{FDR}(s) = \frac{N_r(s) + 0.5}{N_c(s) + 0.5}$$

Kharchenko et al. (2008)

... Model based analysis of ChIP-Seq - MACS

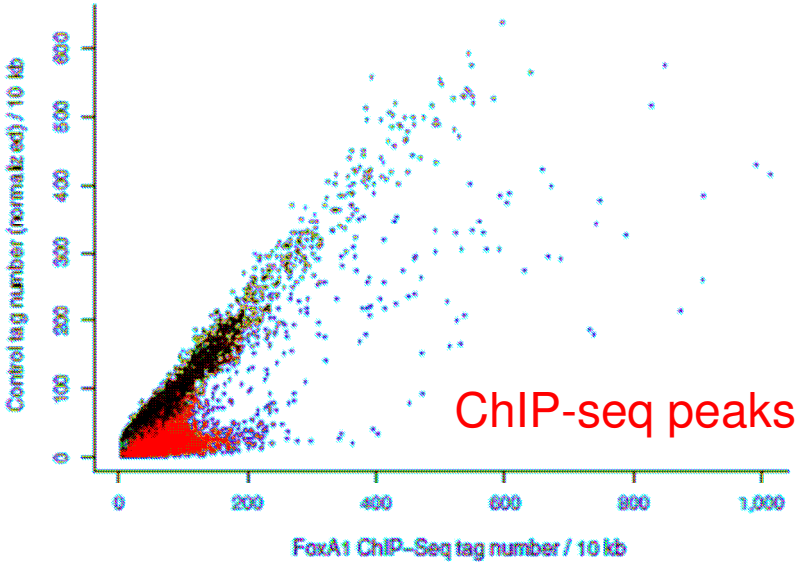


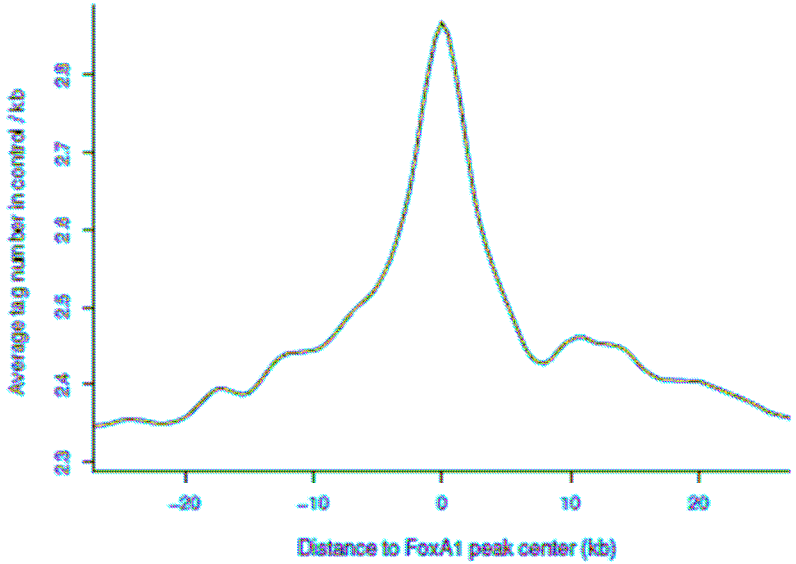
Assume Poisson-distribution

Zhang et al. (2008)

... Model based analysis of ChIP-Seq - MACS

Control = Input peaks

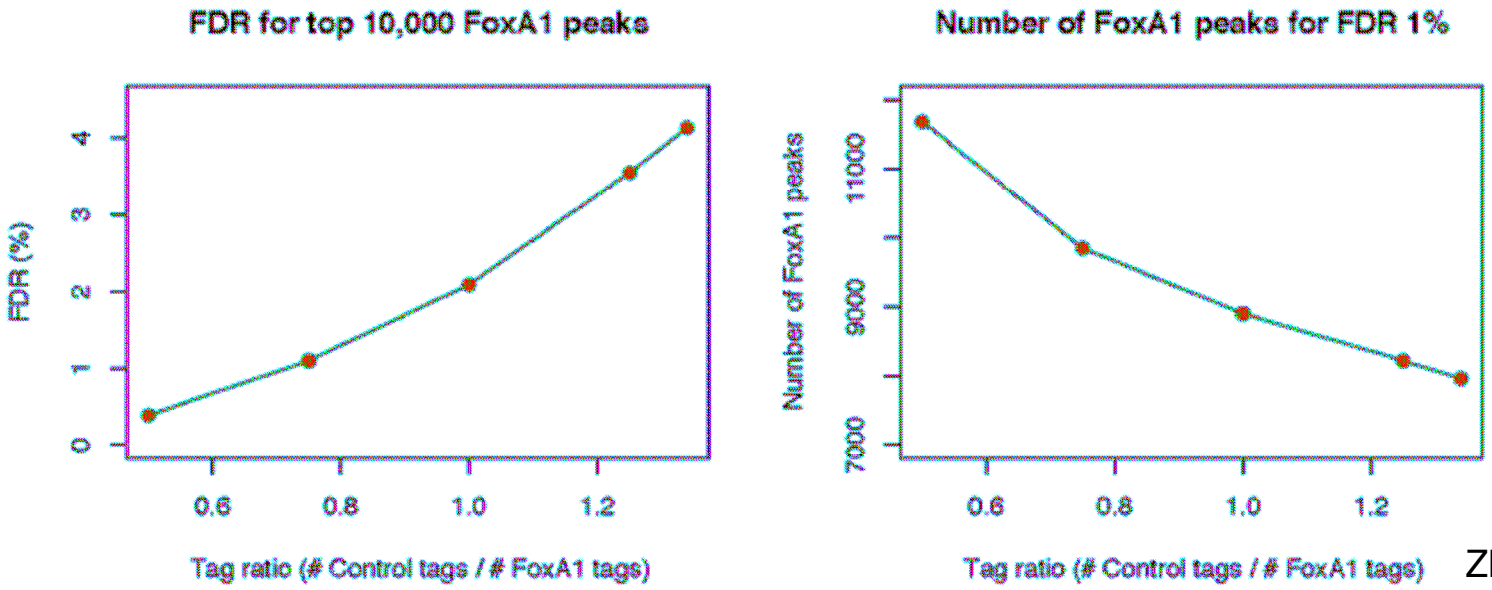




Zhang et al. (2008)

... Model based analysis of ChIP-Seq - MACS FDR?

MACS calculates the FDR based on the number of peaks from control over ChIP that are called at the same p-value cutoff. This FDR estimate is more robust than calculating the FDR from randomizing tags along the genome. However, we notice that when tag counts from ChIP and controls are not balanced, the sample with more tags often gives more peaks even though MACS normalizes the total tag counts between the two samples (Figure S5 in Additional data file 1).



Zhang et al. (2008)

... peak calling - comparison

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

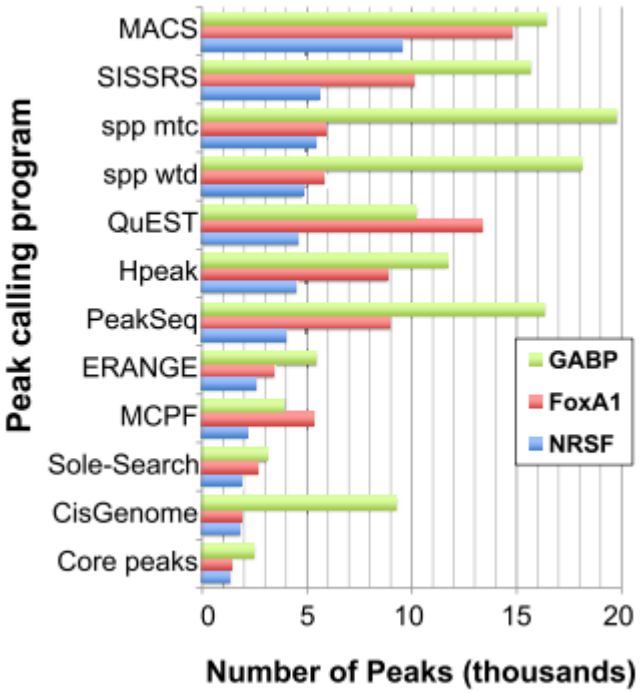
X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

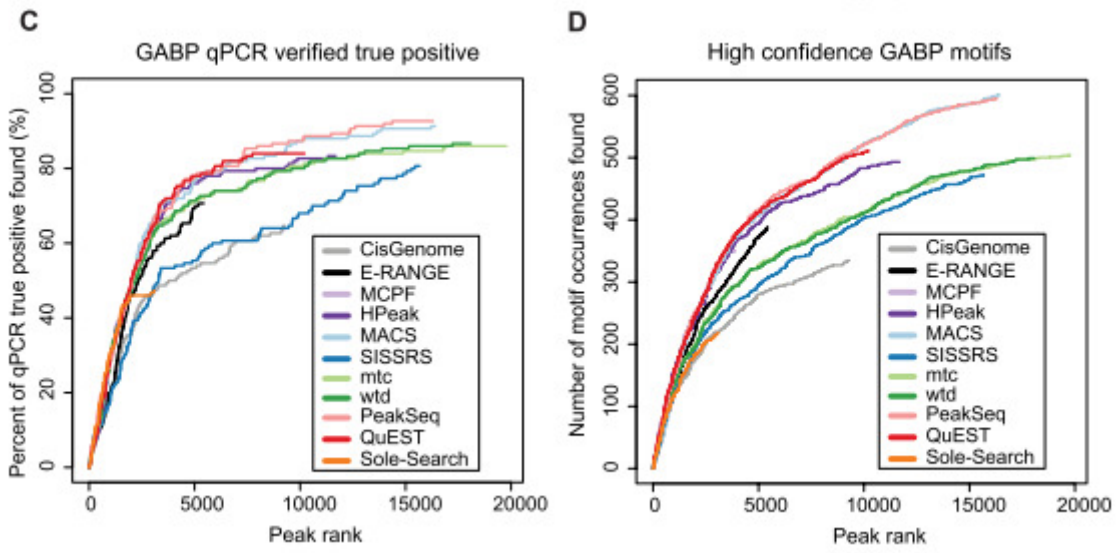
Wilbanks et al. (2010)

... peak calling - comparison



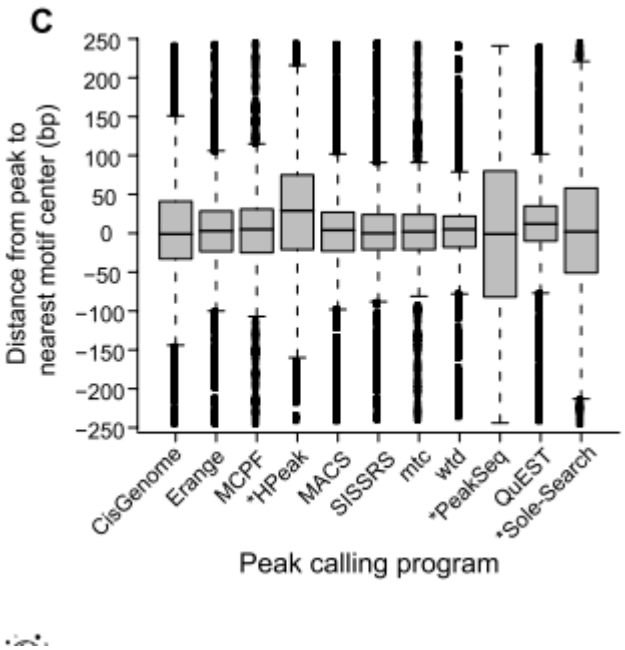
Wilbanks et al. (2010)

... peak calling - comparison



Wilbanks et al. (2010)

... peak calling - comparison



Wilbanks et al. (2010)