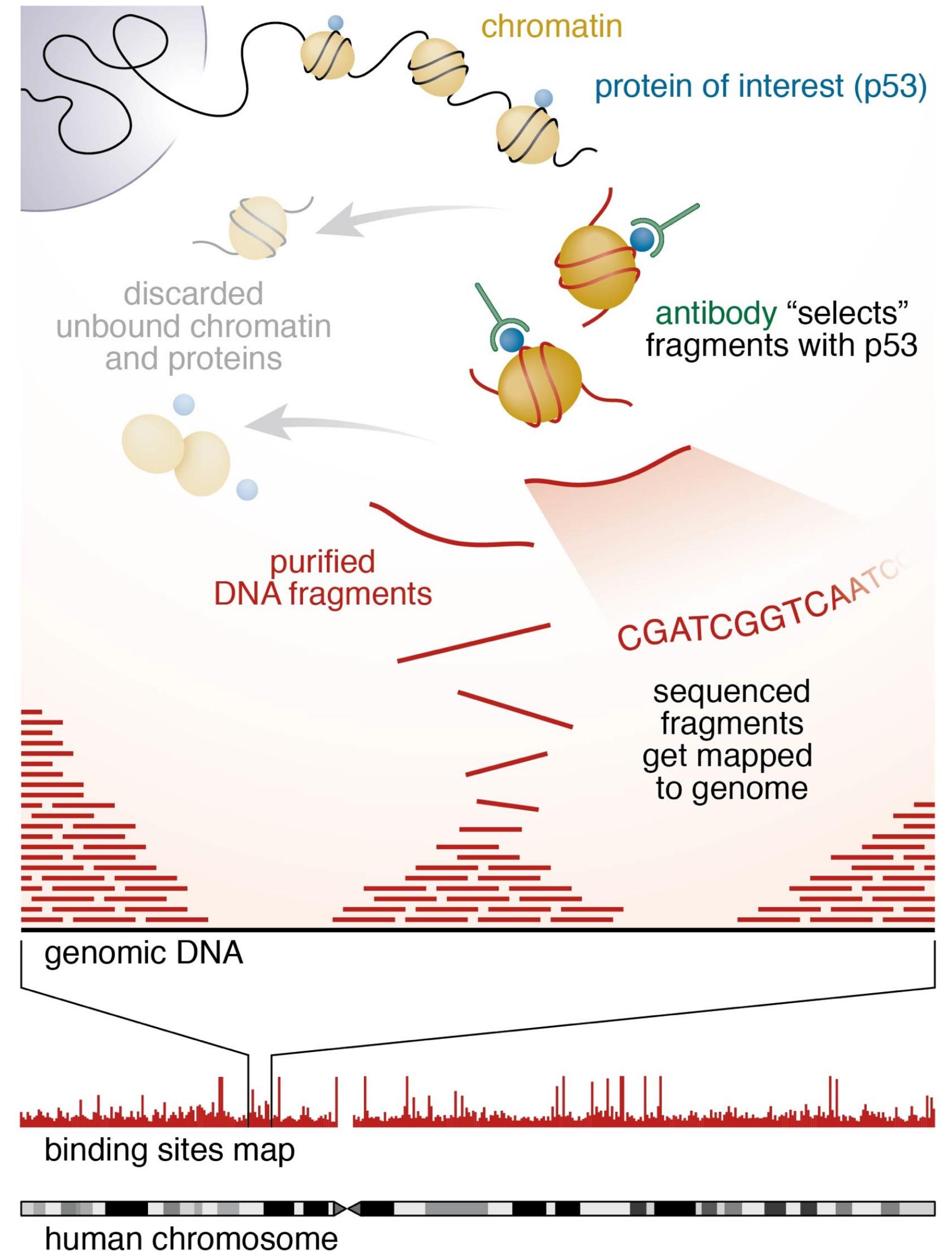
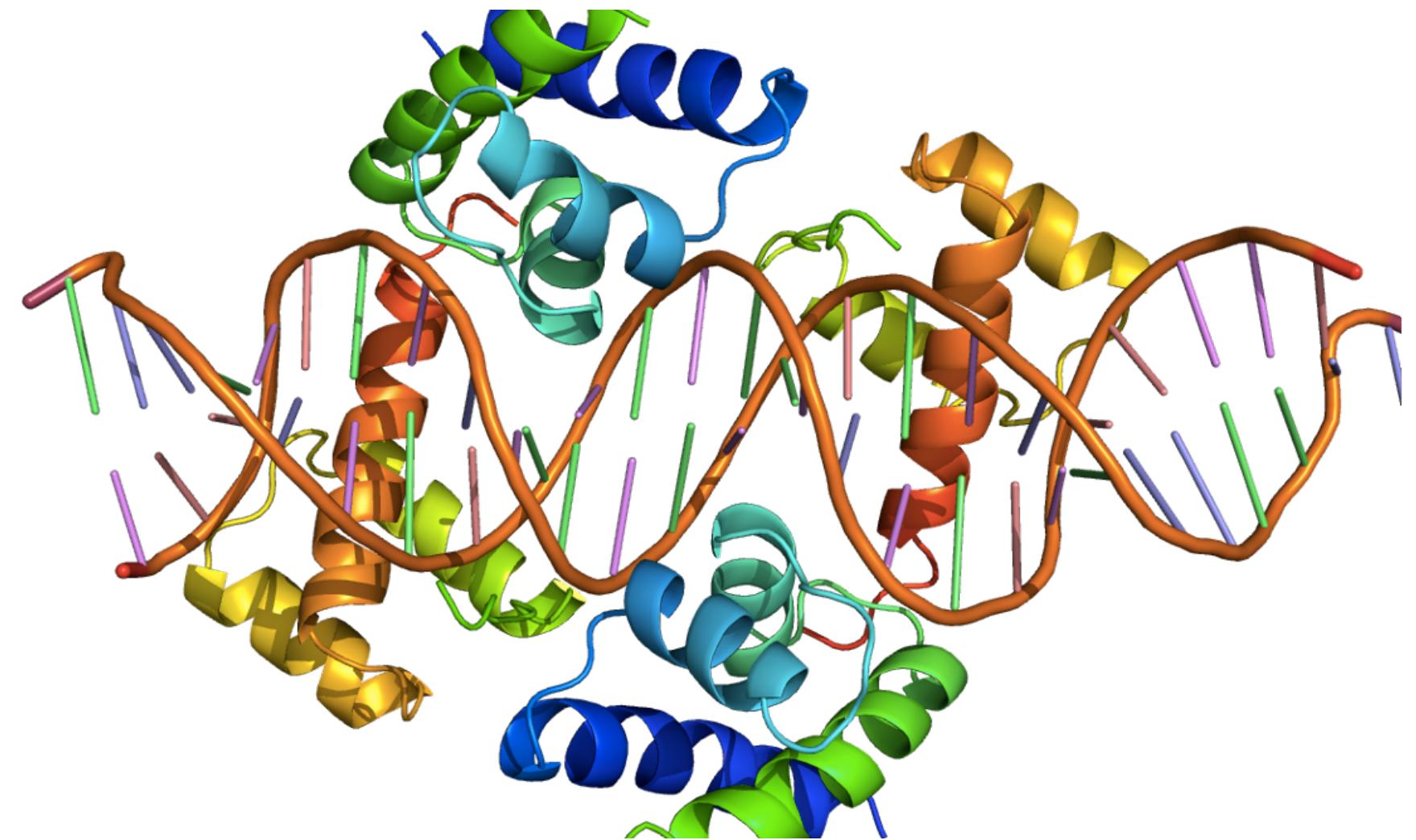


# Introduction to ChIP-seq

HSPH Bioinformatics Core

Shannan Ho Sui

October 3, 2018

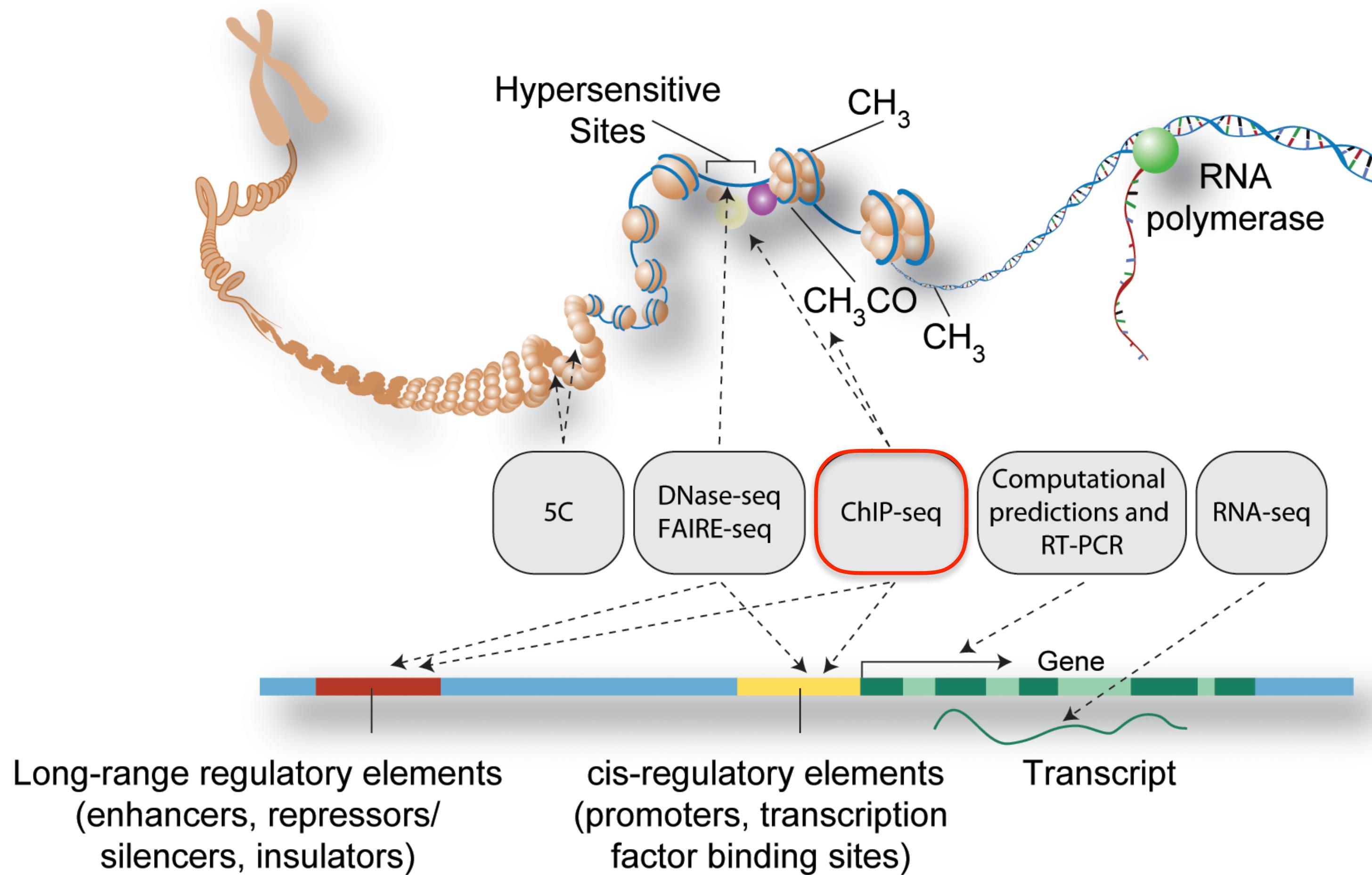


# What is ChIP-seq

---

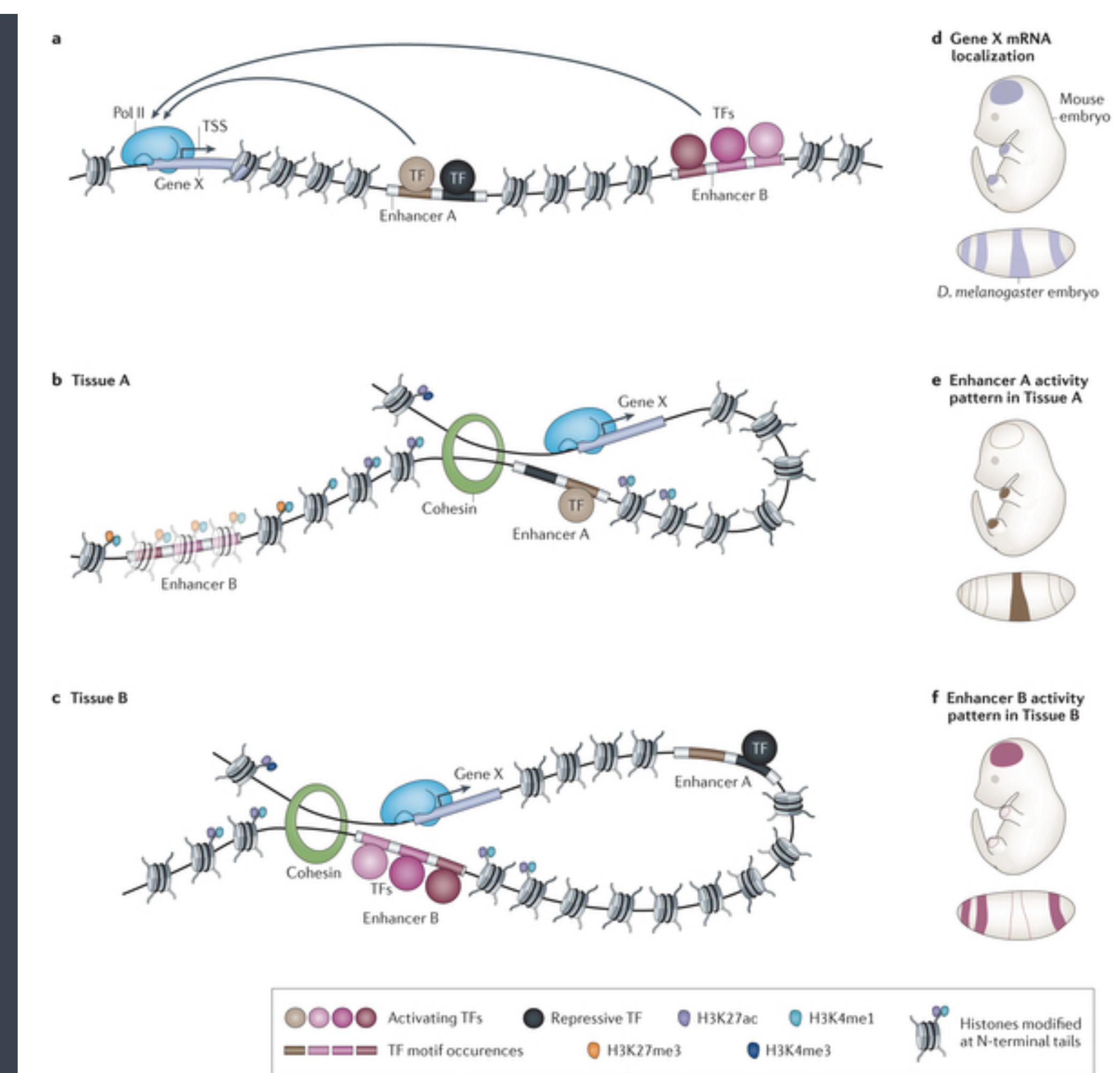
- Assay genome wide binding of protein to DNA
- Uses a combination of chromatin immunoprecipitation and sequencing
- Identifies how transcription factors and histone modifiers interact with DNA *in vivo*
- Complements DNA accessibility studies and gene expression profiling
- Gain an understanding of gene regulation

# Transcriptional regulation is complex

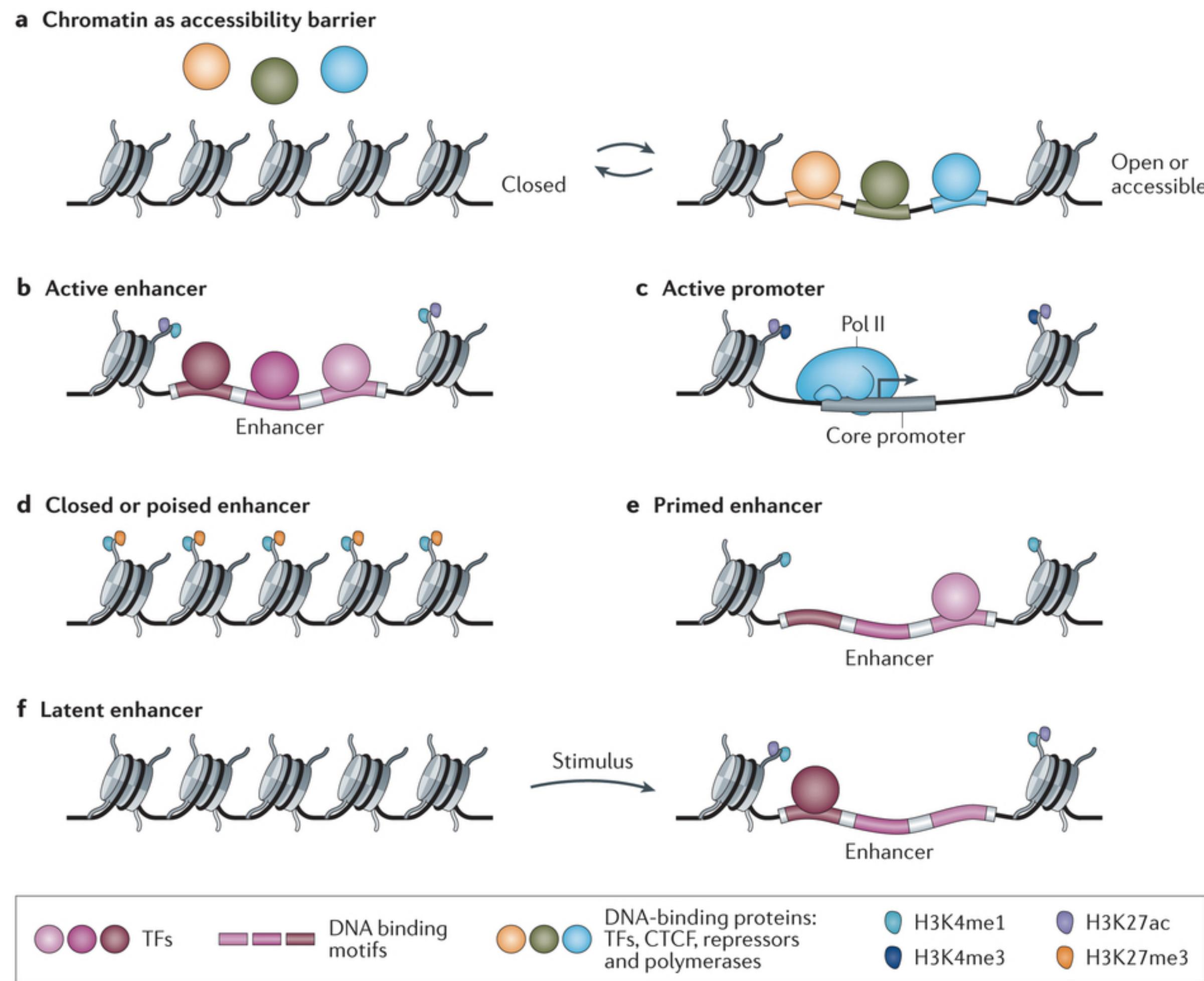


# Complexity in transcriptional regulation

Diverse mechanisms to ensure that genes are expressed at the right time, in appropriate tissues and under specific conditions



Shlyueva, et al (2014). Transcriptional enhancers: from properties to genome-wide predictions.

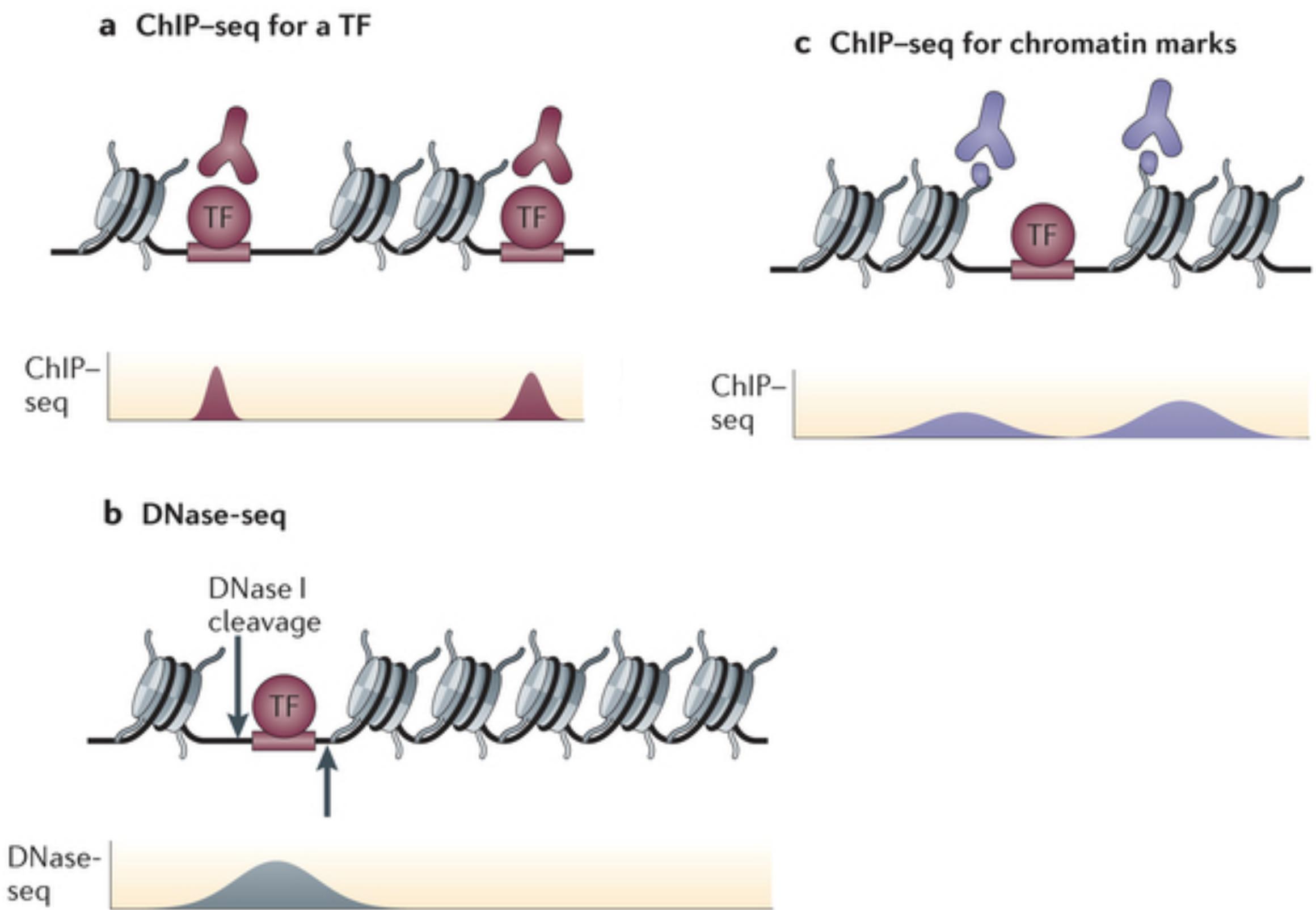


Shlyueva, et al (2014)

Nature Reviews | Genetics

Chromatin structure determines if a gene is expressed or not

# Genomic methods for detecting regulatory elements

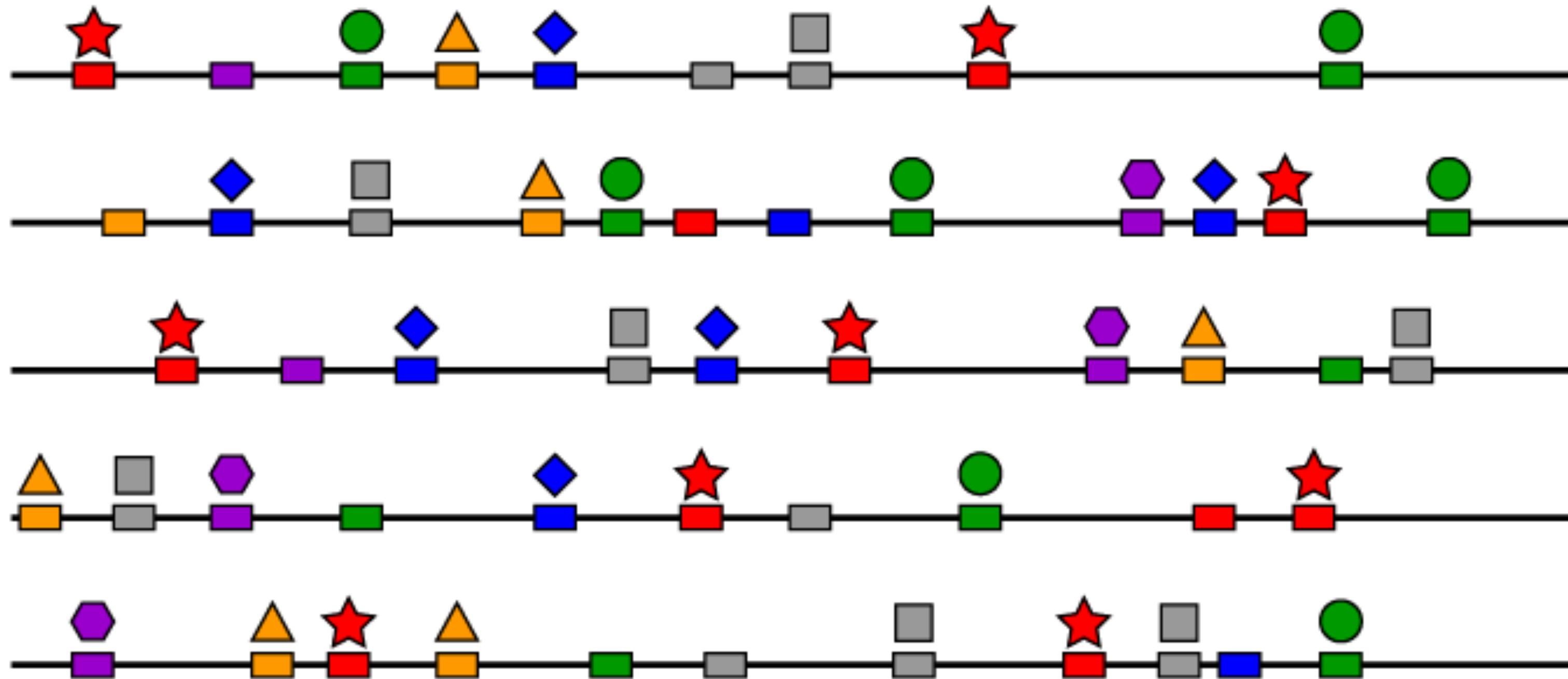


Also ChIA-PET and chromosome conformation capture (3C) based methods to detect to only the contact points but also the pairwise connections between these points

Shlyueva, et al (2014)

# Library Preparation

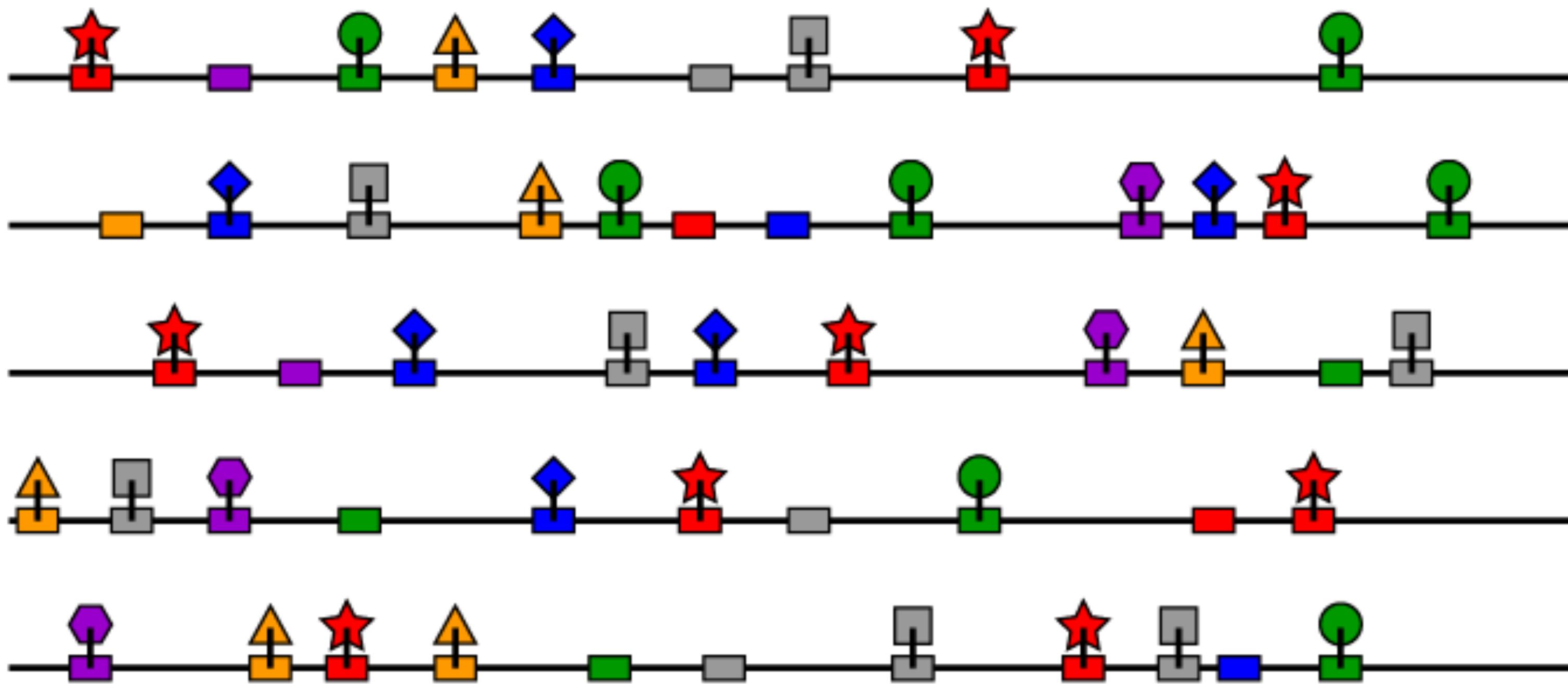
---



- ▶ Need sufficient amount of starting material because the ChIP will enrich for a small proportion
- ▶ Ideally the starting material for one ChIP uses  $10^7$  cells from culture

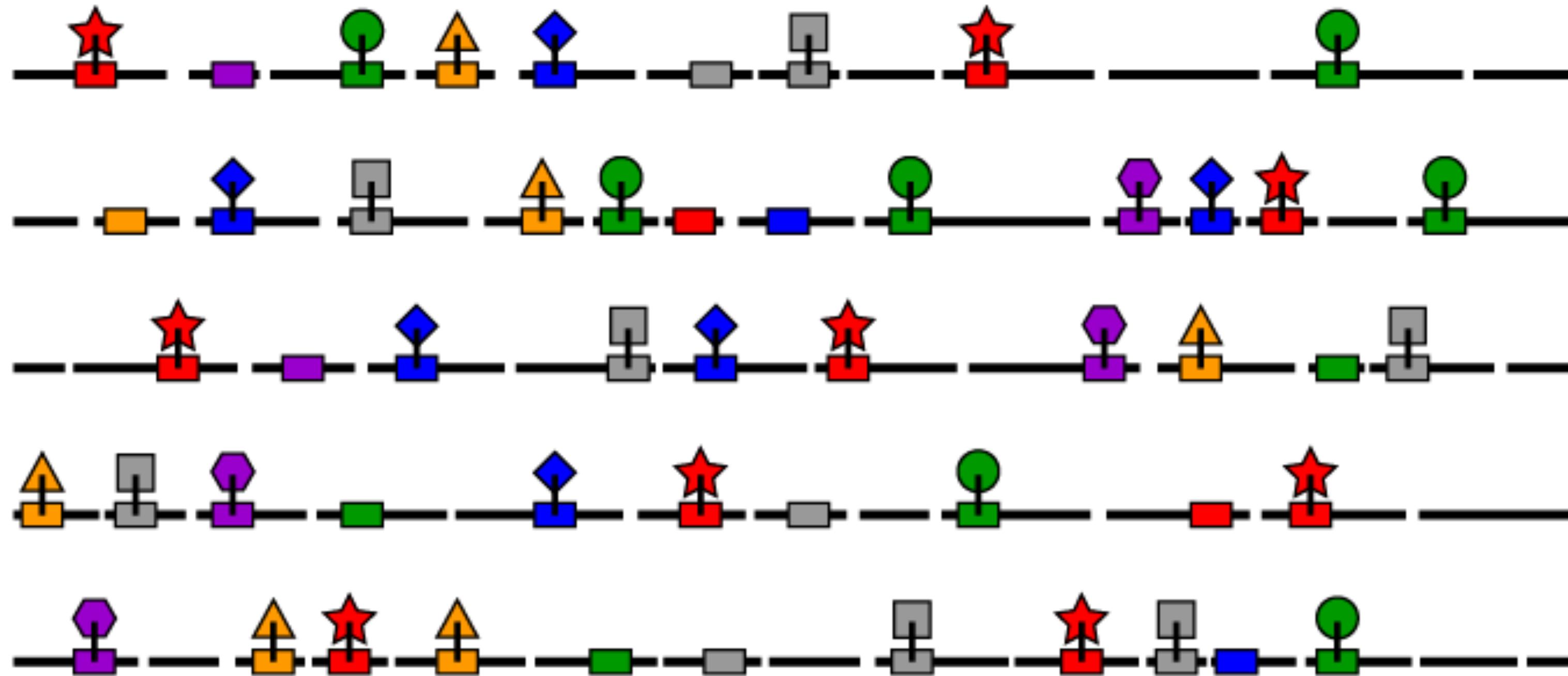
# Crosslink proteins to DNA

---



# Fragment

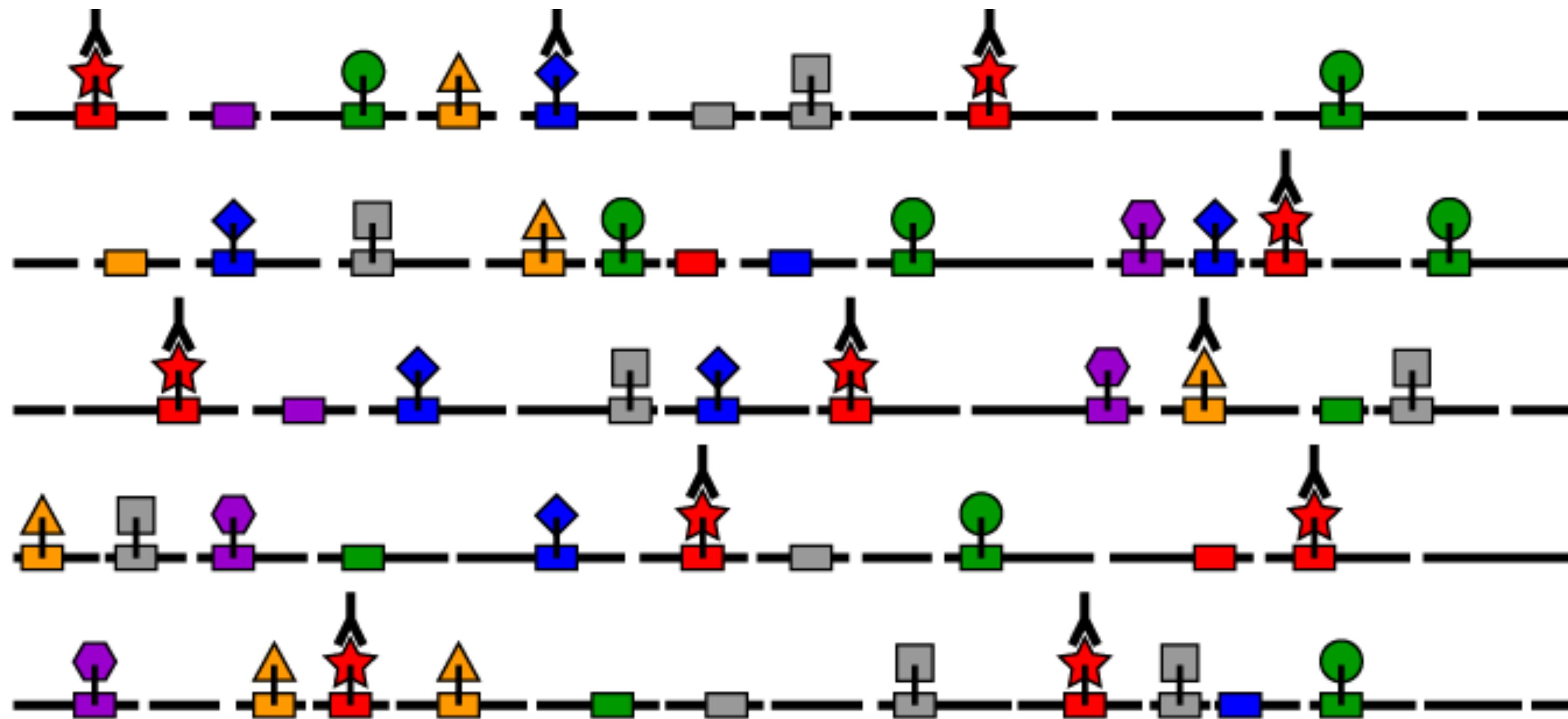
---



- ▶ The DNA is sheared into small fragments - usually 200-500 bp in length
- ▶ Check by running on a gel

# Protein specific antibody

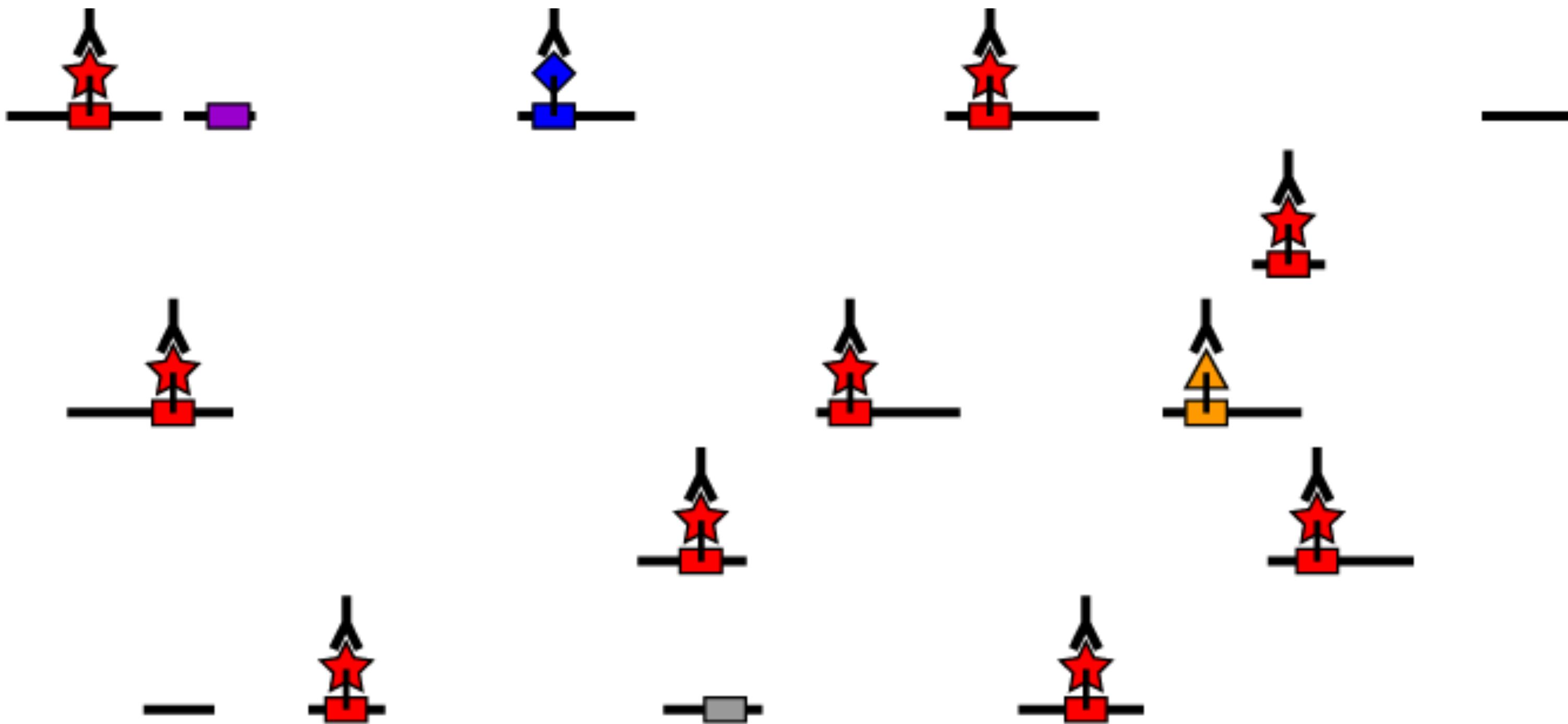
---



- ▶ The sheared protein-bound DNA is immunoprecipitated using a specific antibody

# Immunoprecipitate

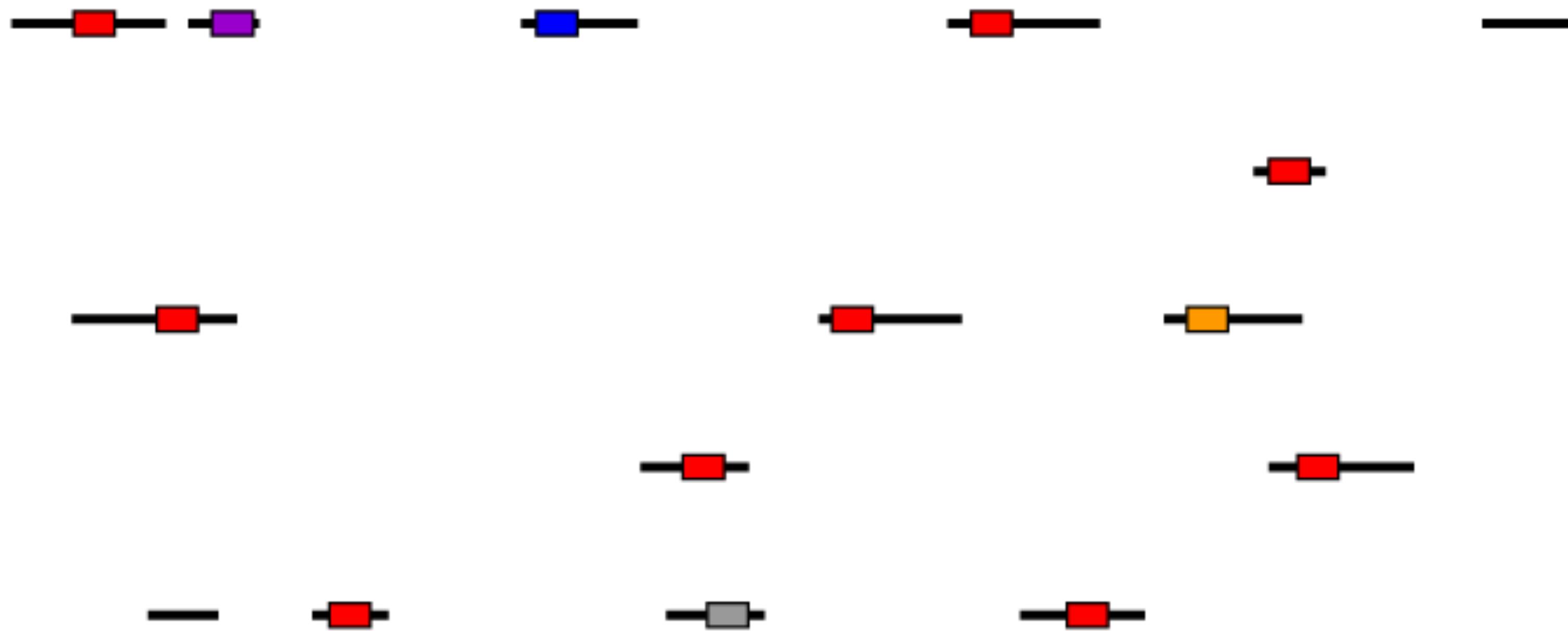
---



- The antibody binds primarily to the protein of interest but there may be cross reactivity with other proteins with similar epitopes

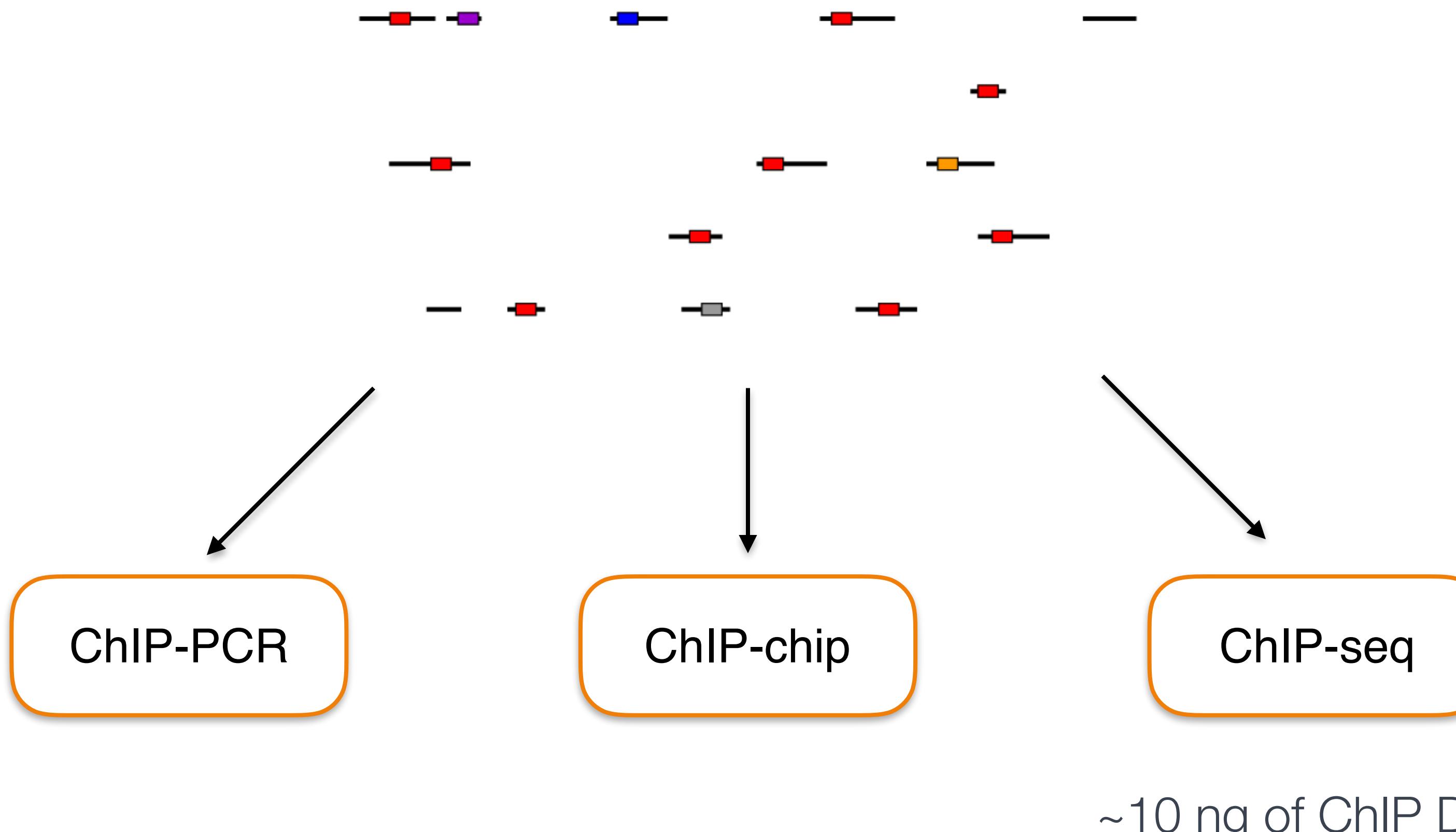
# Reverse crosslink and purify DNA

---

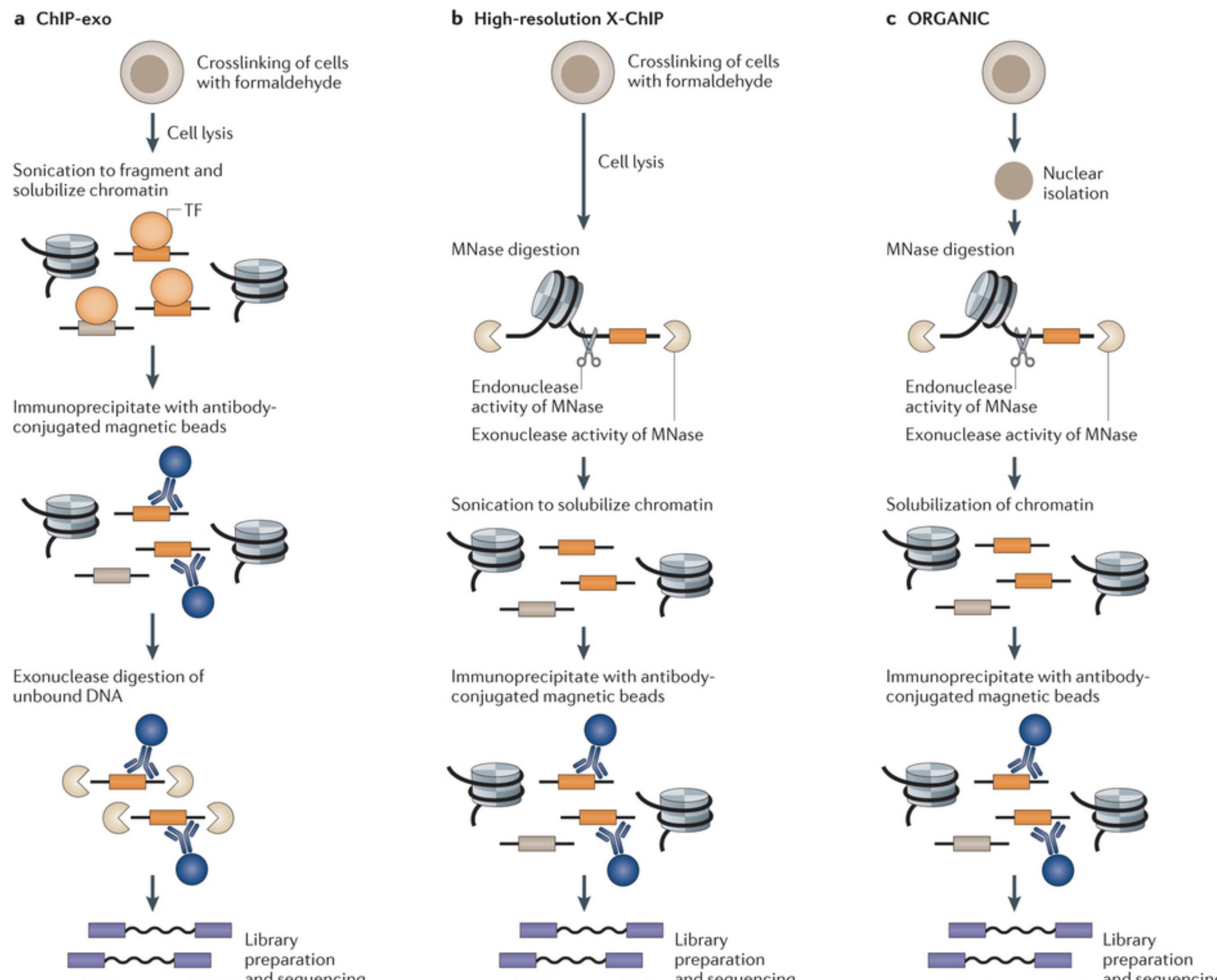


# Identify bound regions

---



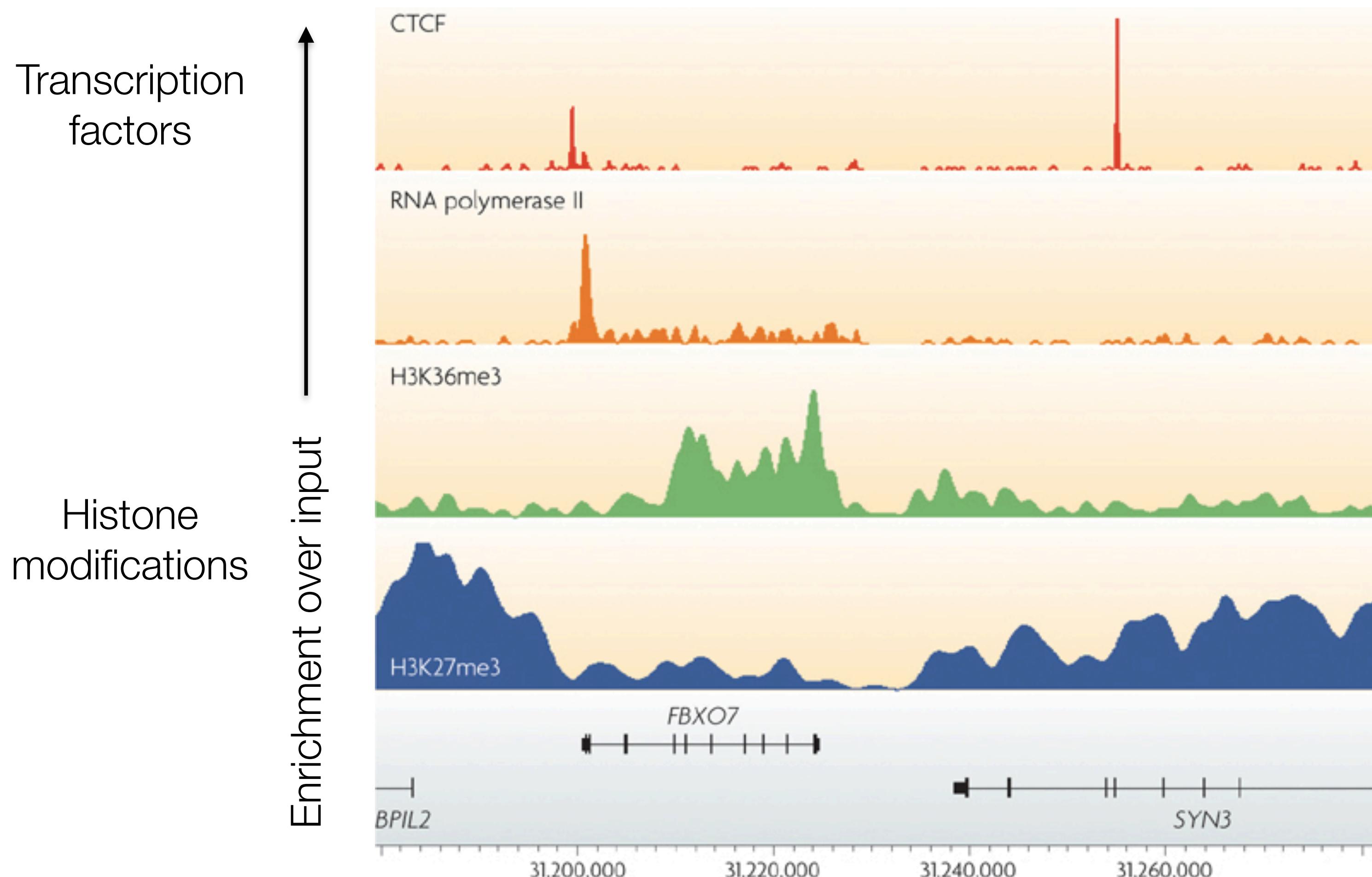
Zentner & Henikoff (2014).  
Nature Reviews Genetics.



Nature Reviews | Genetics

# High resolution variations of ChIP-seq

# Types of signals



Adapted from Park (2009). Nature Reviews Genetics.

# Profiling histone modifications

---

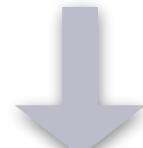
- Active promoters: H3K4me3, H3K9Ac
- Active enhancers: H3K27Ac, H3K4me1
- Repressors: H3K9me3, H3K27me3
- Transcribed gene bodies: H3K36me3

# Why are controls necessary?

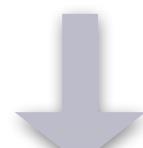
- Signal depends on # active binding sites, the number of starting genomes, IP efficiency
- Open chromatin regions fragment more easily than closed regions
- Repetitive sequences might seem to be enriched
- Uneven distribution of sequence tags across the genome
- Hyper-ChIPable regions
- Allows us to compare with the same region in a matched control
- ENCODE also provides a “Black List”

Biological samples/Library preparation

Crosslink proteins to DNA



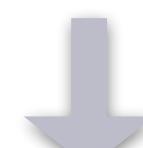
Shear DNA (sonication)



Immunoprecipitation



Reverse crosslink



Size selection and PCR

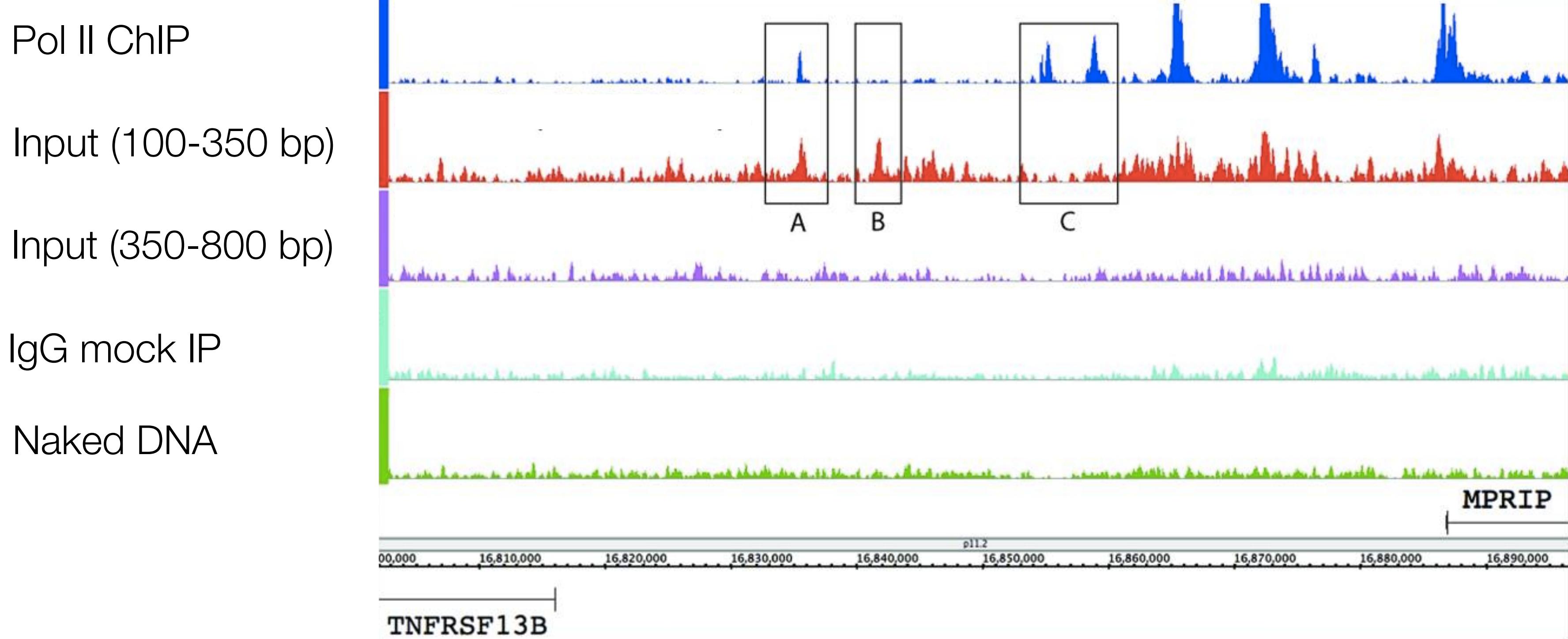
Specific antibody (ChIP enrichment)



No IP (Input DNA)

Non-specific antibody (IgG “mock IP”)

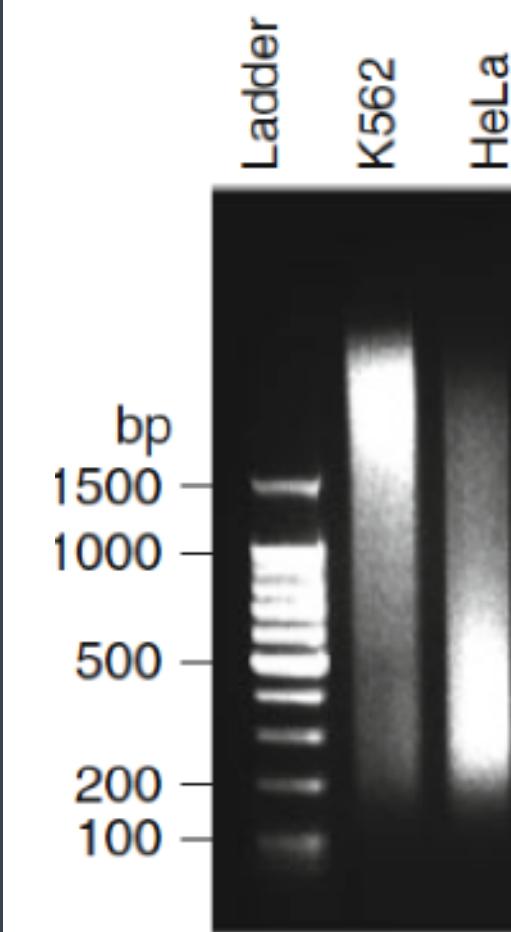
ChIP-Seq Controls



Map of ChIP-seq versus control signals

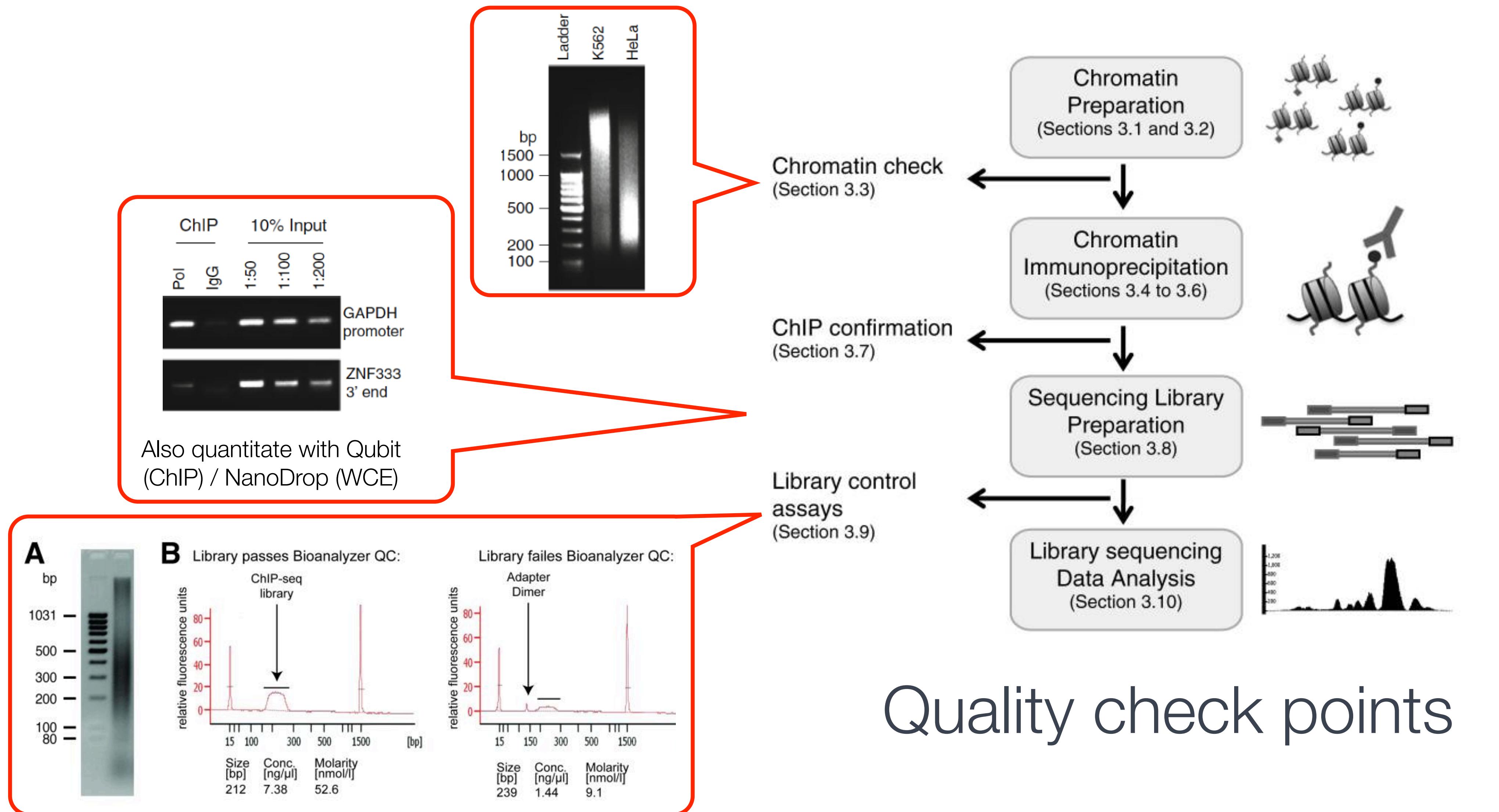
# Parameters for a successful ChIP

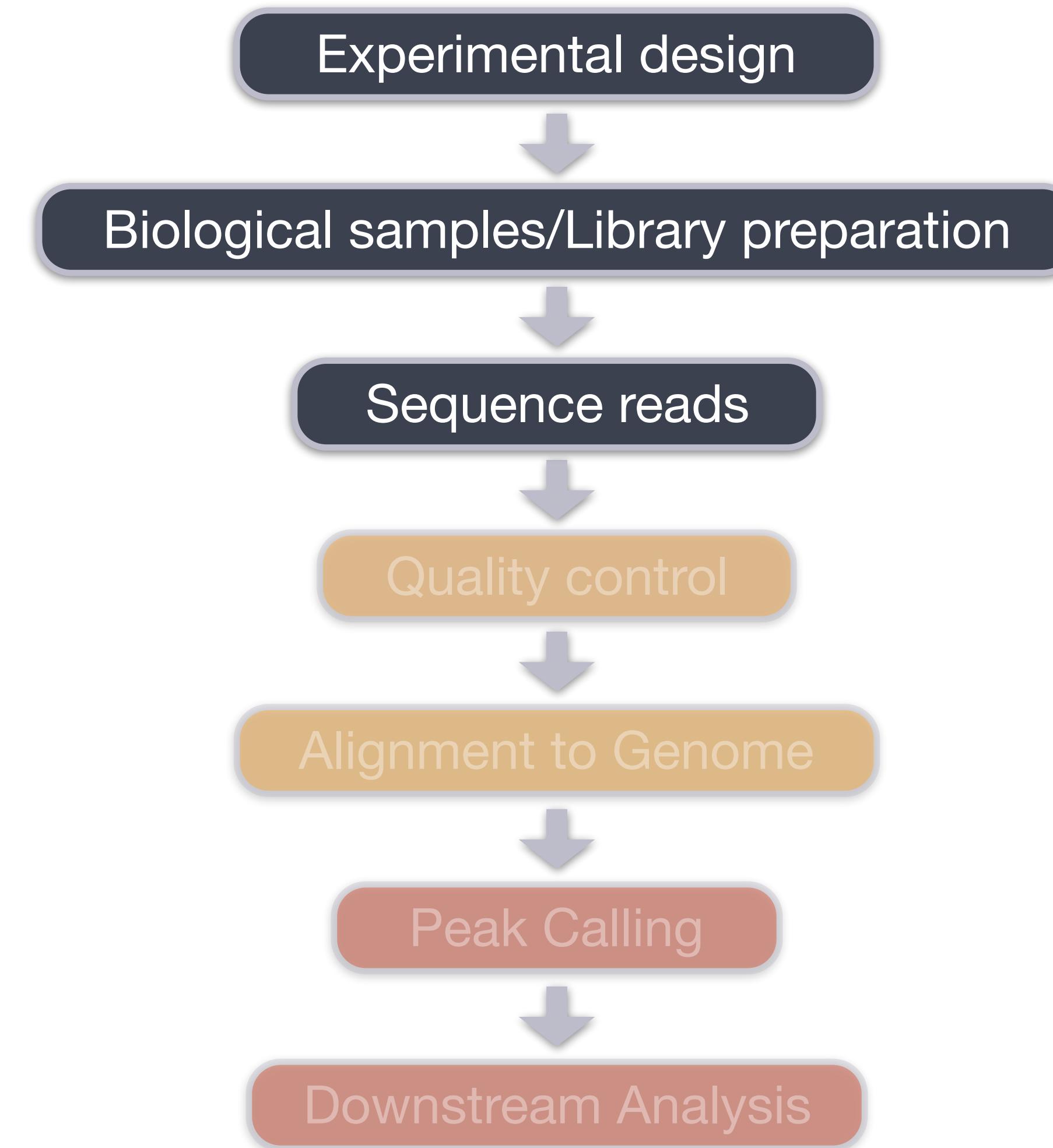
- ▶ Efficient and specific antibody
- ▶ Amount of starting material
- ▶ Chromatin fragmentation
  - ▶ Size matters (not too big and not too small)
  - ▶ Can vary between cell types
  - ▶ Stringency of washes



**Fragments too big:**  
Reduced signal to noise ratio  
in ChIP-seq

**Oversonication:**  
Fragmentation biased towards  
promoter regions causes  
ChIP-seq enrichments at  
promoters in both, ChIP AND  
control (input) sample





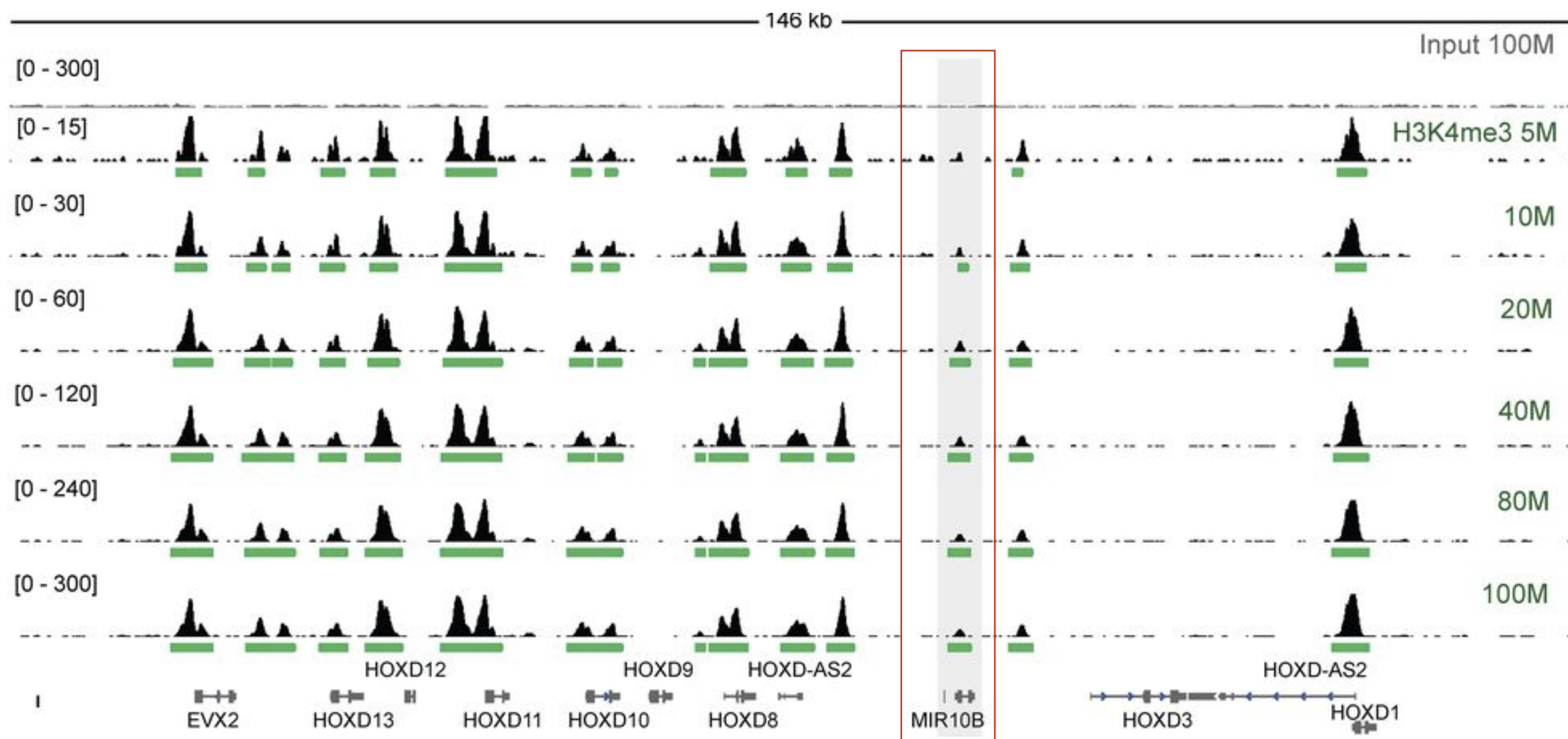
ChIP-seq workflow

# Sequencing considerations

- ▶ Read length (50- to 150-bp)
  - > Longer reads and paired-end reads improve mappability
  - > Only necessary for allele-specific chromatin events, investigations of transposable elements)
  - > Balance cost with value of more informative reads
- ▶ Avoiding batches or distributing samples evenly over batches
- ▶ Sequencing depth (5-10M min; 20-40M as standard for TFs; higher for broad profiles)
- ▶ Sequence input controls to equal or higher depth than IP samples

# Impact of sequencing depth

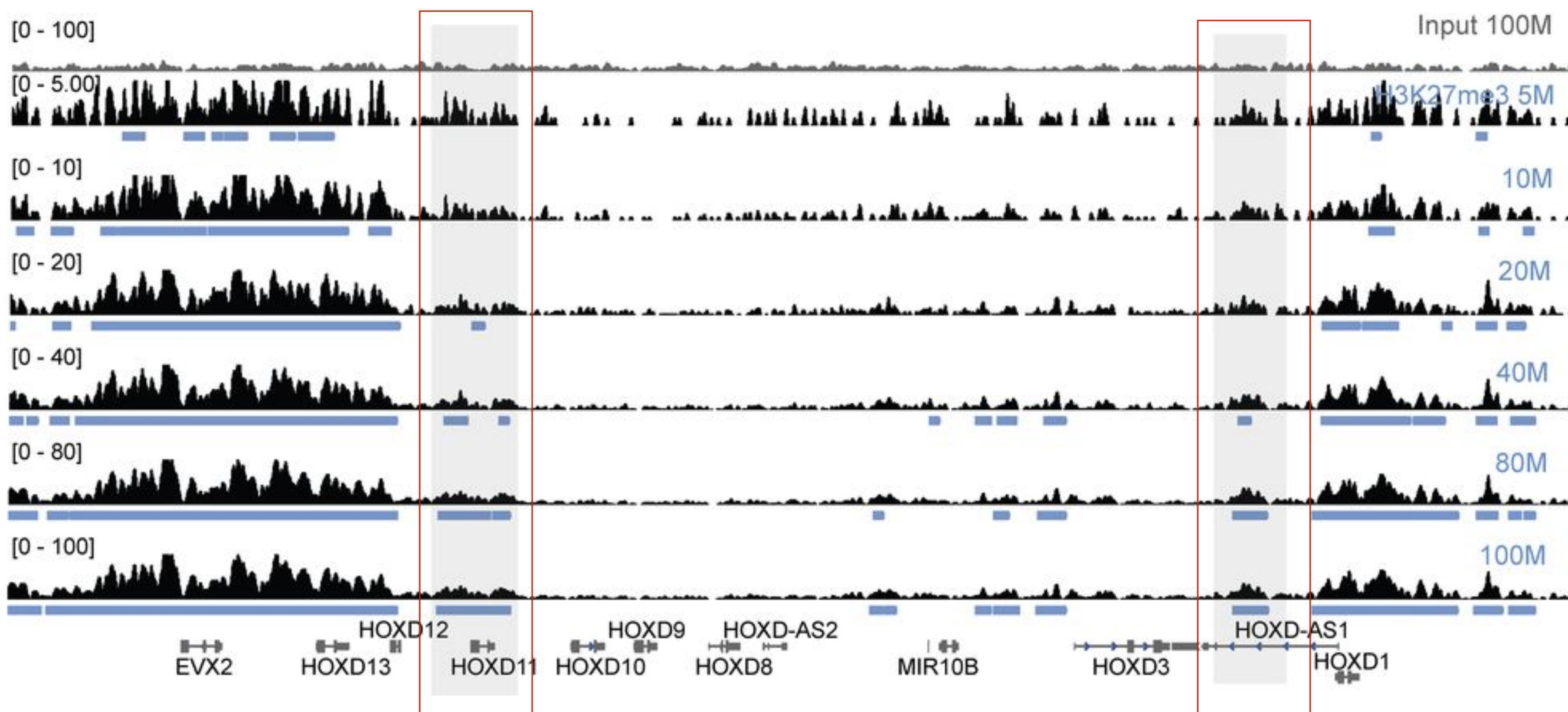
H3K4me3



Adapted from Jung et al (2014). NAR.

# Impact of sequencing depth

H3K27me3

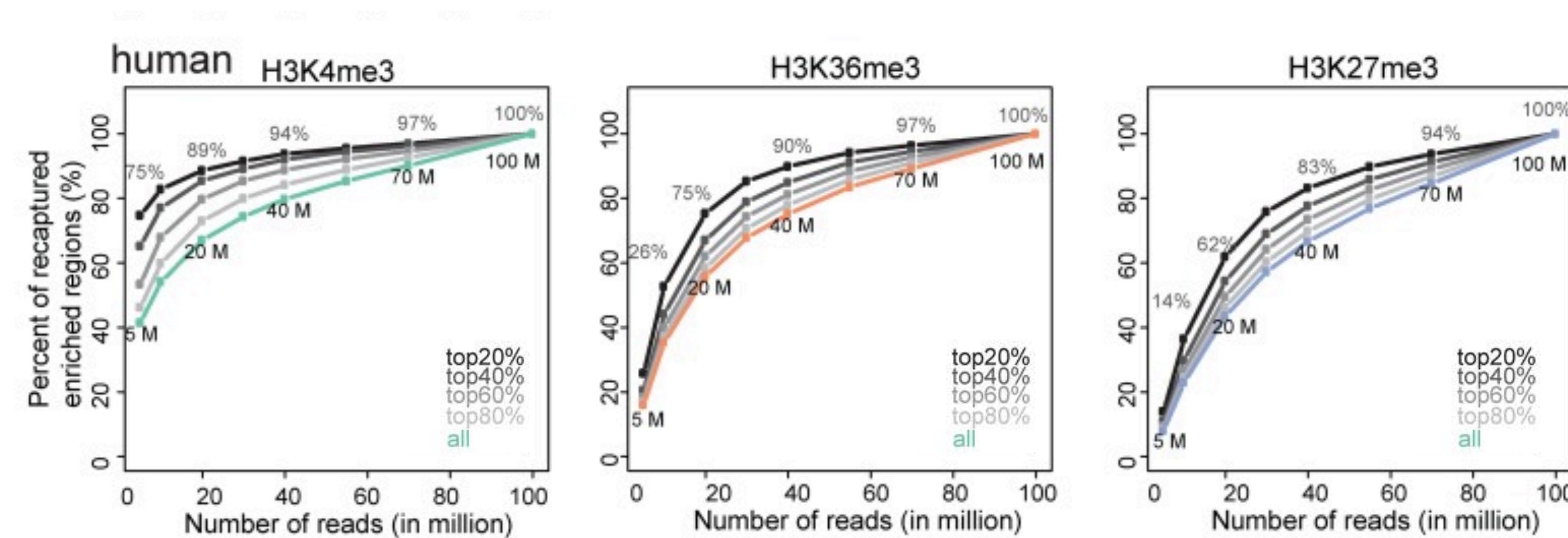


Adapted from Jung et al (2014). NAR.

# Impact of sequencing depth

---

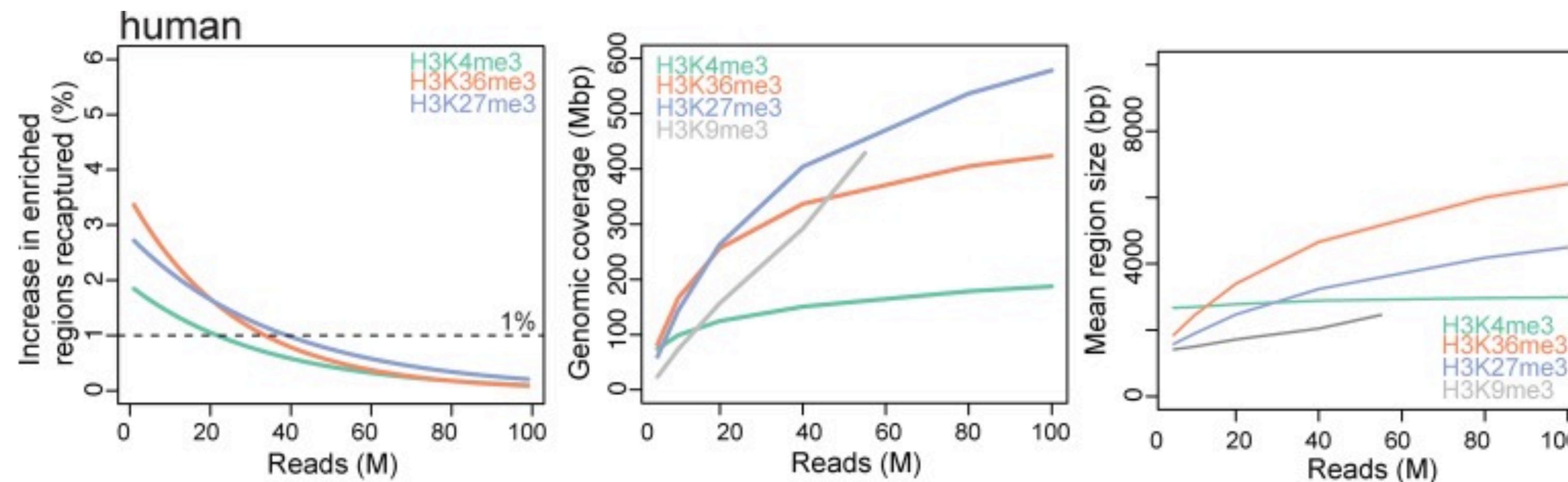
Percentage of significantly enriched regions from the full data recovered in each subsample for H3K4me3, H3K36me3 and H3K27me3



# Impact of sequencing depth

---

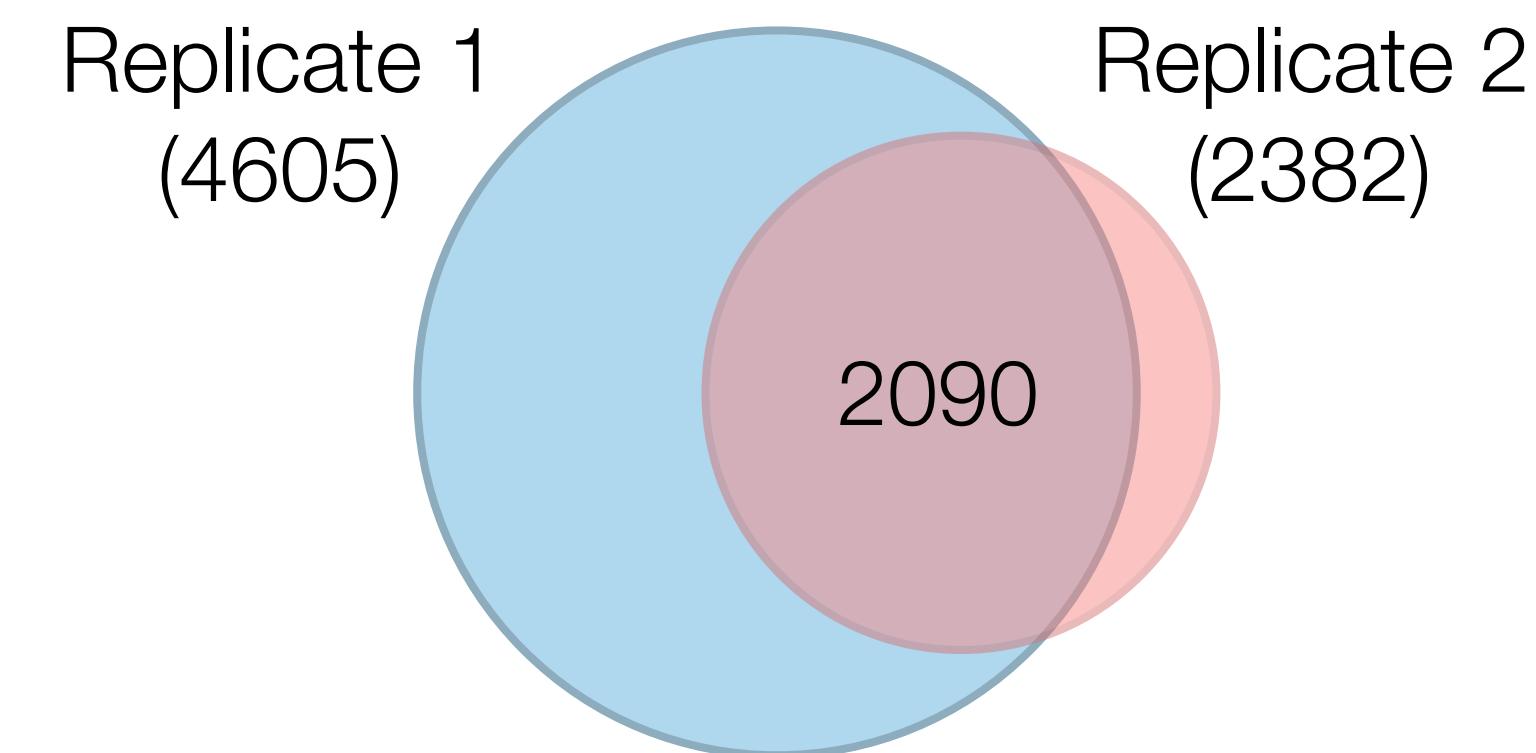
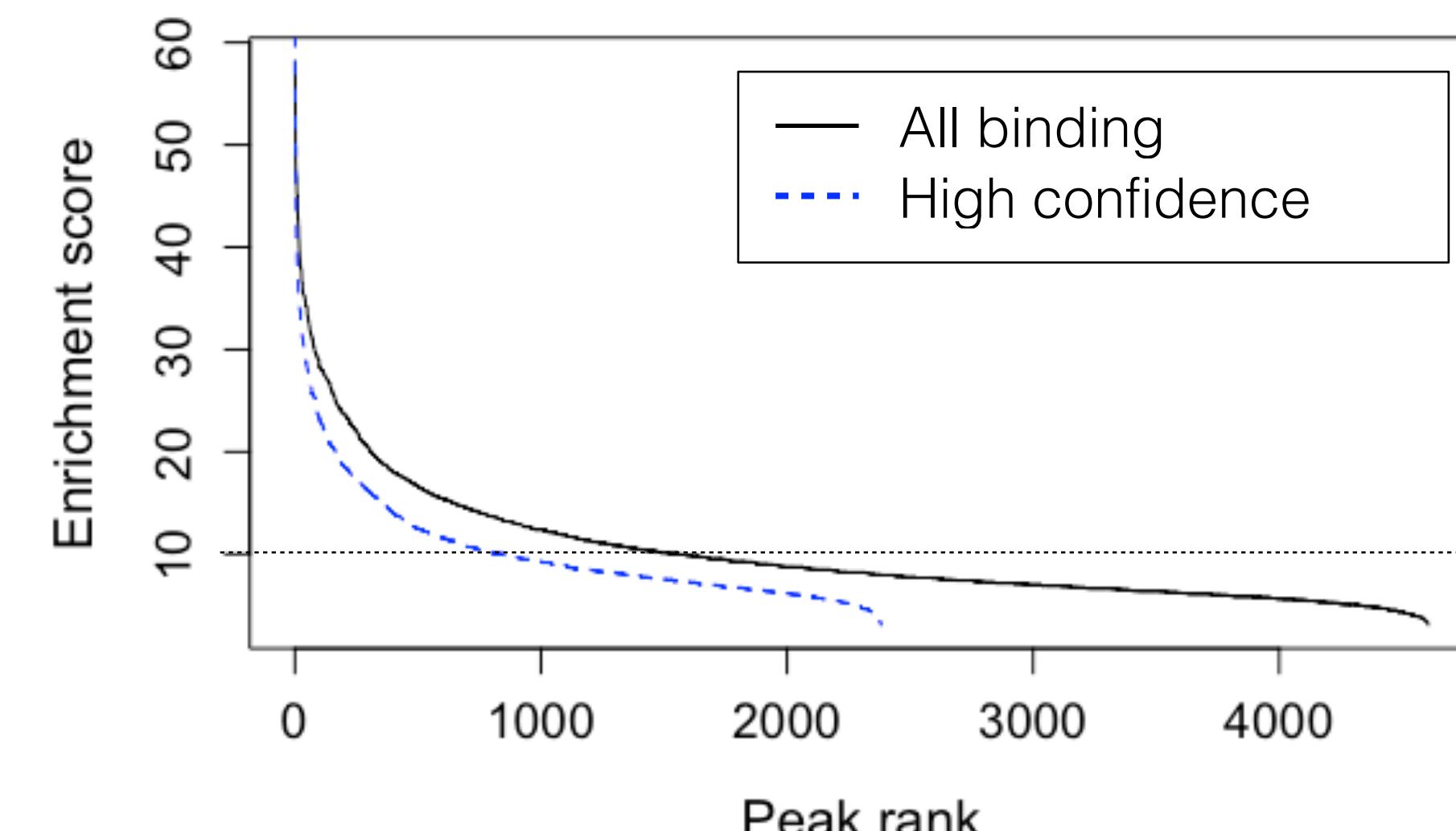
Percentage of increase in enriched regions recaptured when an additional 1 million reads were sequenced

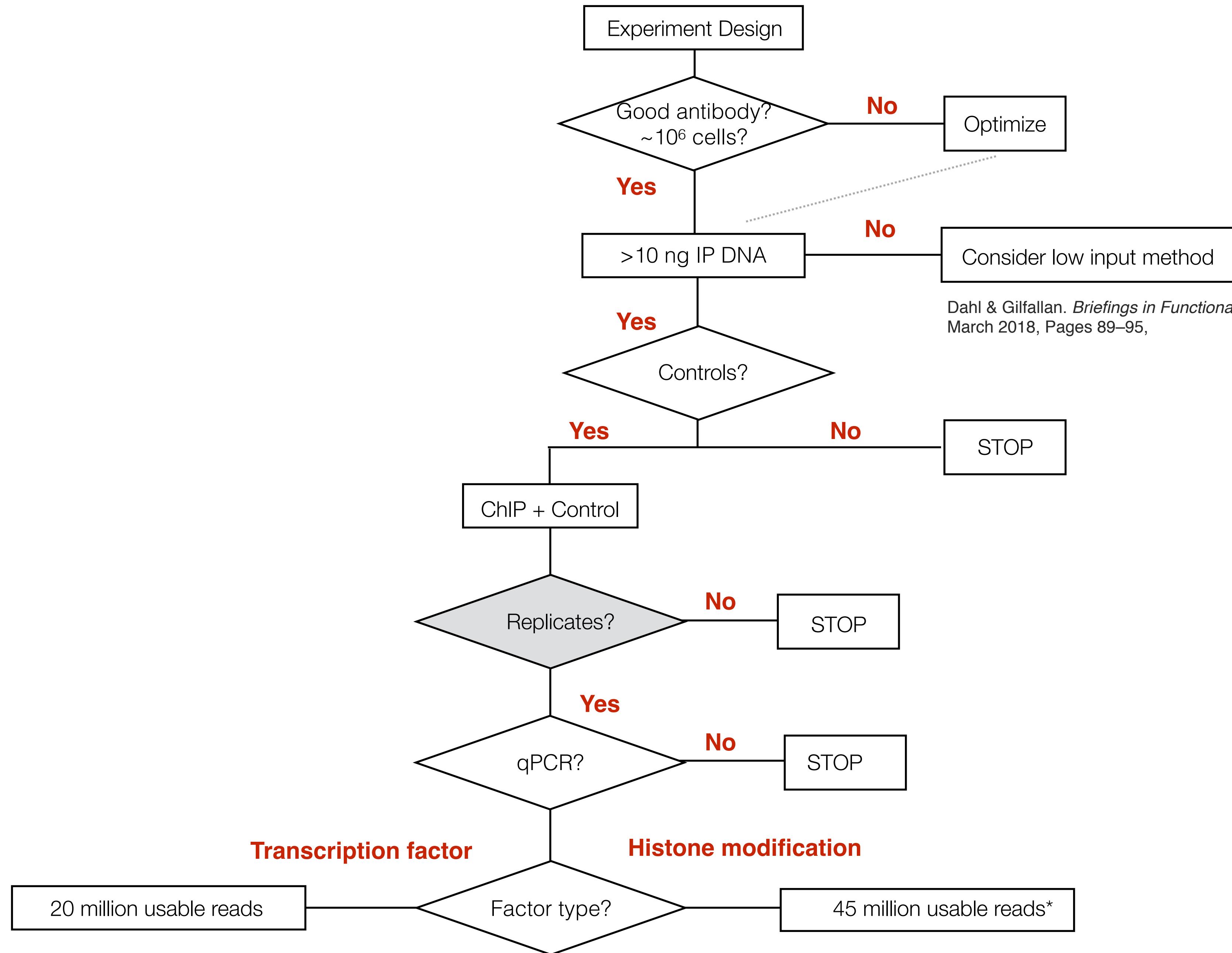


# Replicates and reproducibility

---

- Biological replicates are essential to understand variation and for differential binding analysis
- More replicates is often preferable to greater depth
- Better to sequence high-quality sample at lower depth than low-quality sample to higher depth



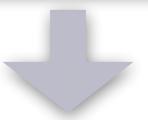


Dahl & Gilfallan. *Briefings in Functional Genomics*, Volume 17, Issue 2, 1 March 2018, Pages 89–95,

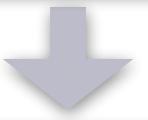
## Experimental design



## Biological samples/Library preparation



## Sequence reads



## Quality control

FASTQC



## Alignment to Genome

BWA, Bowtie2



## Peak Calling

Filter duplicates, multi mappers, blacklist



## Downstream Analysis

ChIP-seq workflow

# Quality check and filtering

---

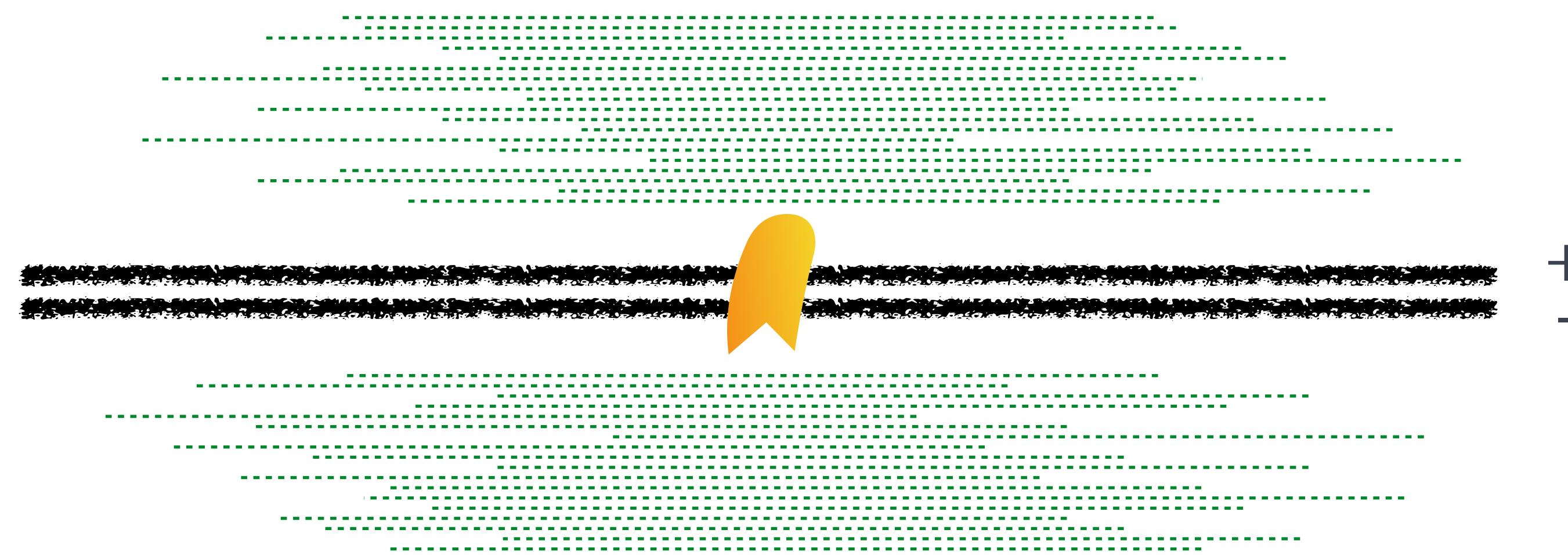
- Raw sequence QC is similar to RNA-seq
- However,
  - Explore duplication rates and possibly remove duplicates
  - Remove blacklisted regions
  - Assess cross correlation scores and Fraction of Reads in Peaks (FRiP)

Software: [ChIPQC](#), Homer, ChiLin, DiffBind

# Understanding strand cross-correlation

---

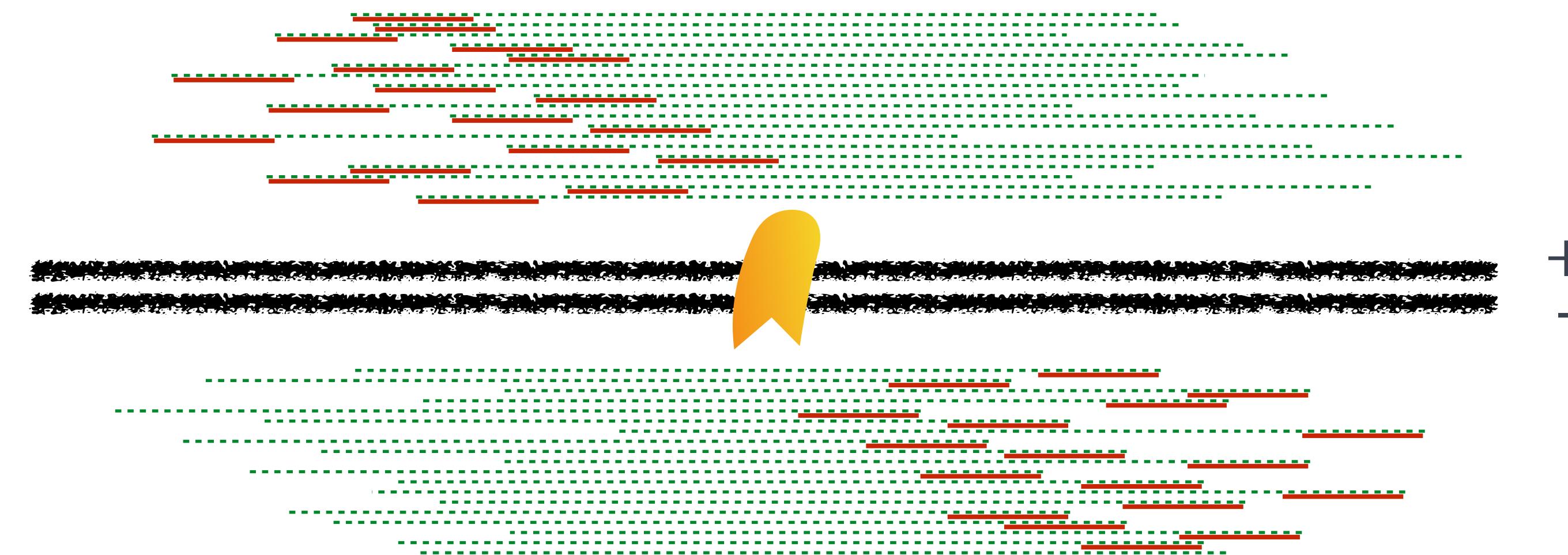
Yellow arrowhead = binding site  
Dashed green line = size selected DNA fragment



# Understanding strand cross-correlation

---

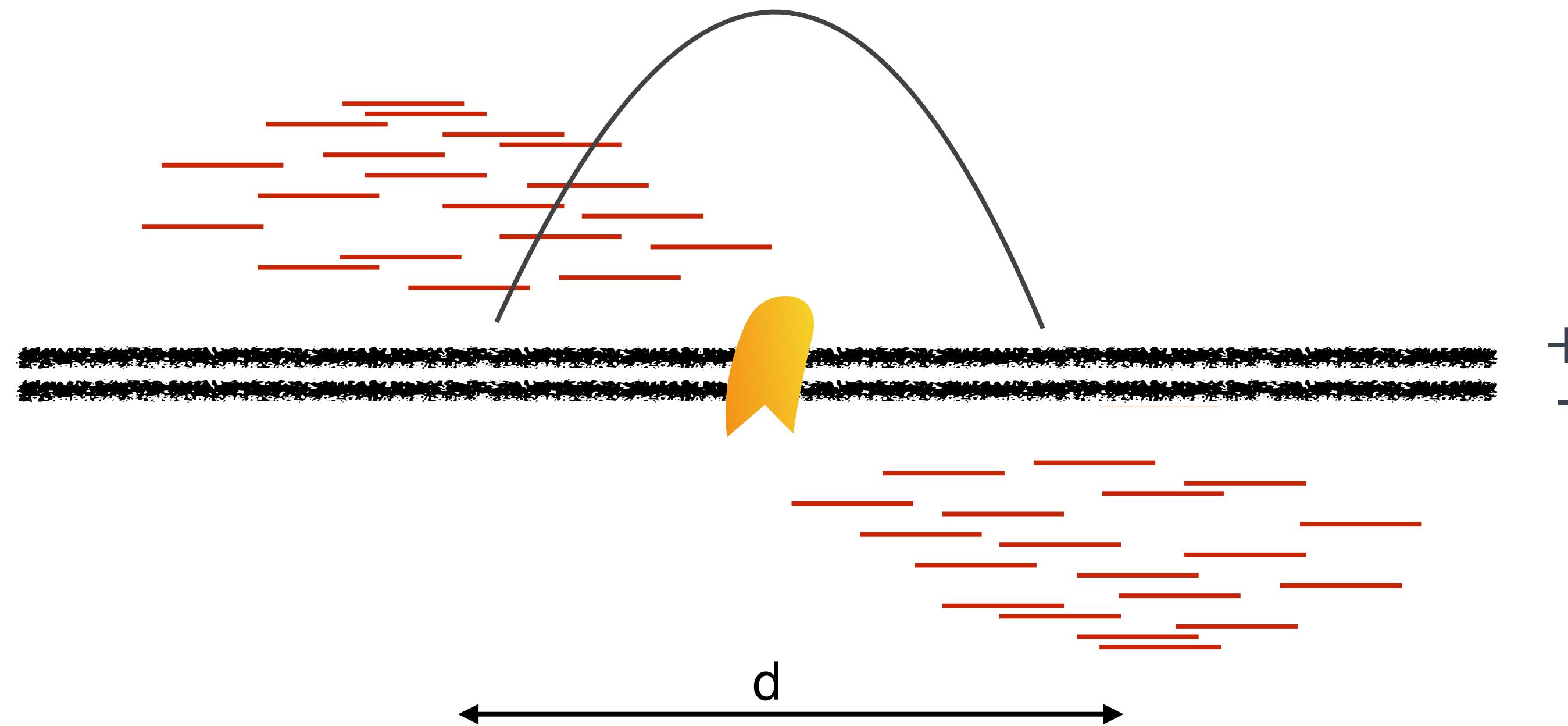
ChIP-seq fragments are sequenced from the 5' end



# Understanding strand cross-correlation

---

Alignment generates a **bimodal pattern** on the plus and minus strands around binding sites

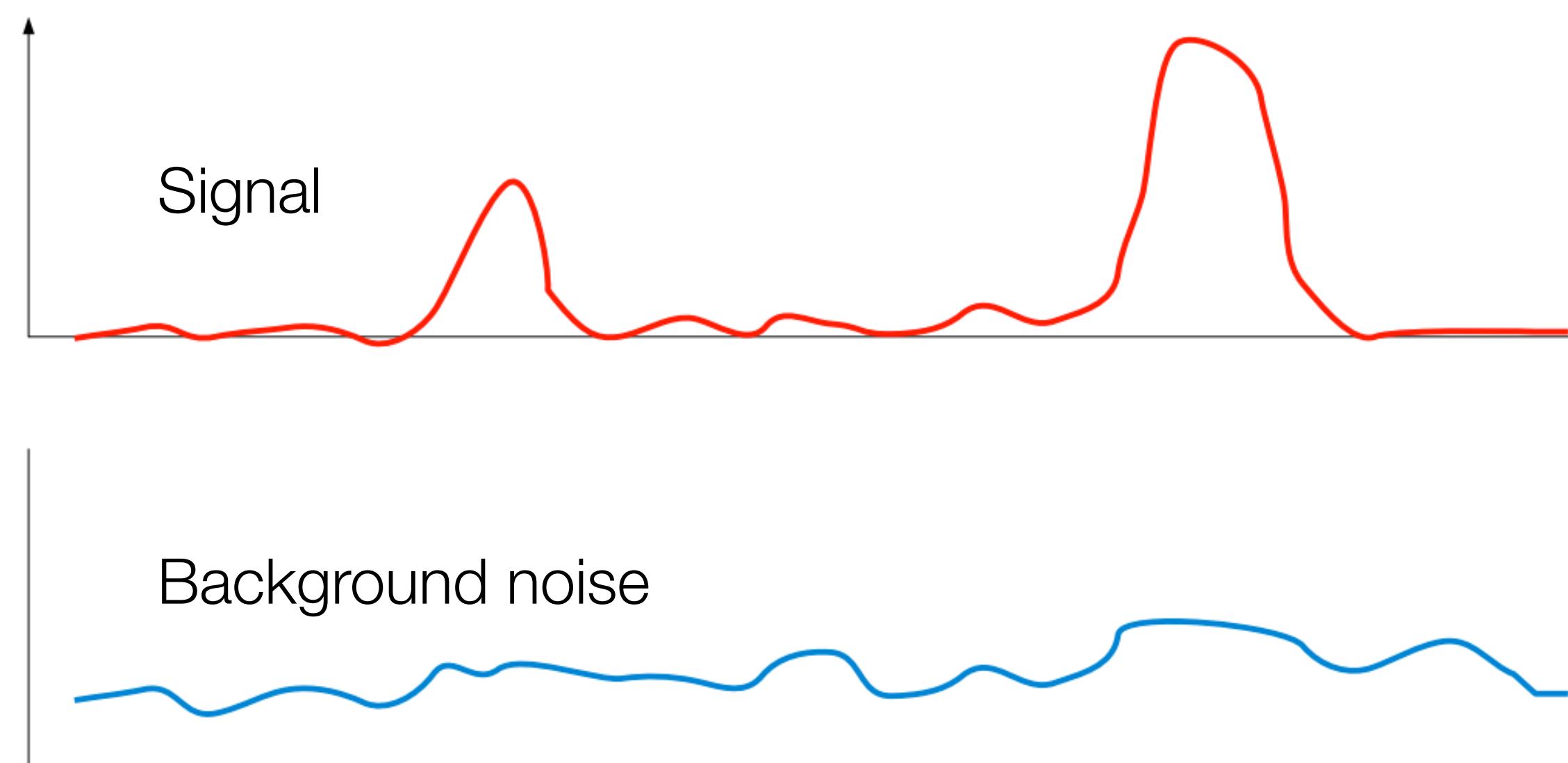


Peak calling algorithms use this pattern to estimate the relative strand shift

# Modeling noise to detect real peaks

---

- Noise is not uniform (chromatin conformation, local biases, mappability)
- Input data is mandatory for a reliable estimation of noise (even though some tools don't require it)



# Peak detection

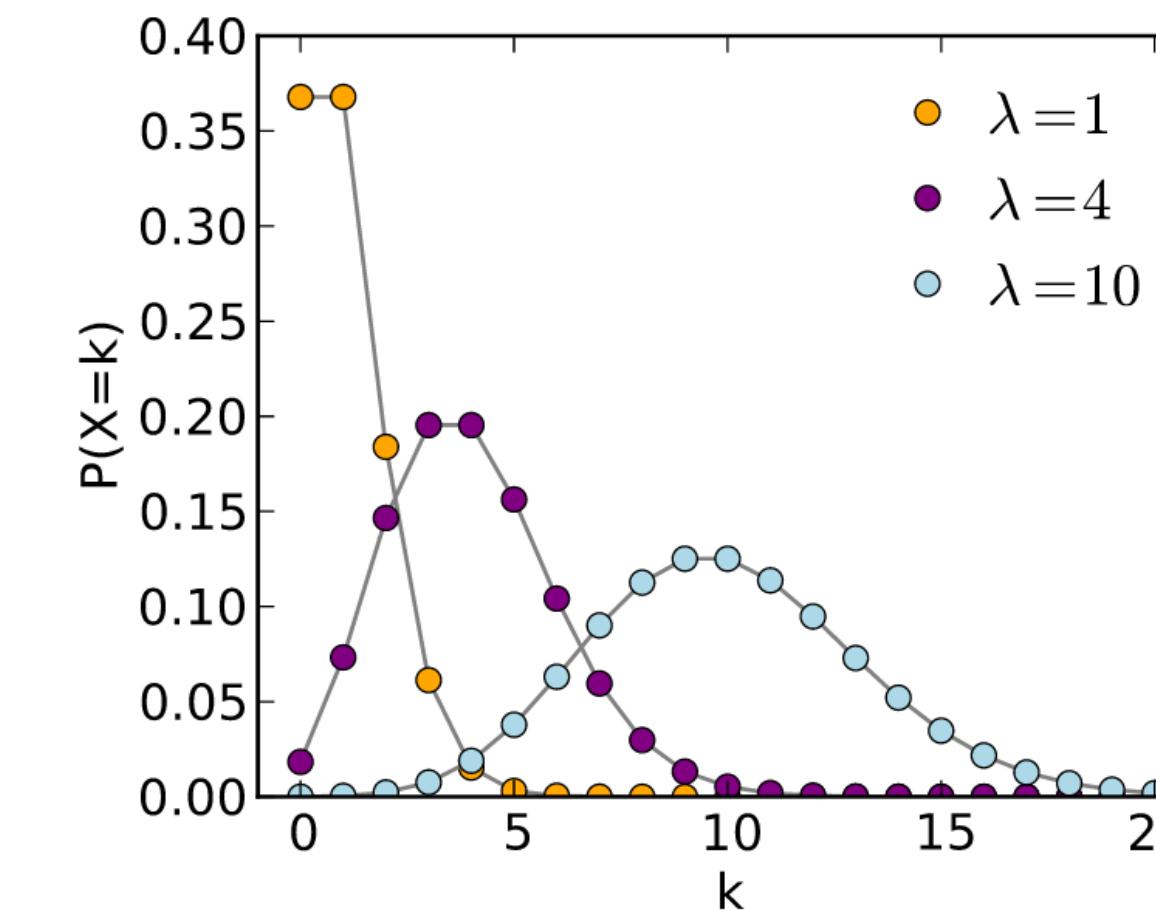
---

- Most algorithms model the number of reads from a genomic region/window using a Poisson distribution
- One parameter model for estimating the expected number of reads in the window
- Often more variance in real data than assumed by the Poisson (overdispersion)
- MACS (model-based analysis of ChIP-Seq) uses multiple Poisson distributions to model the local background noise within each region from the input data

$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

where

- $\lambda$  is the average number of events per interval
- $e$  is the number 2.71828... (Euler's number) the base of the natural logarithms
- $k$  takes values 0, 1, 2, ...
- $k! = k \times (k - 1) \times (k - 2) \times \dots \times 2 \times 1$  is the factorial of  $k$ .

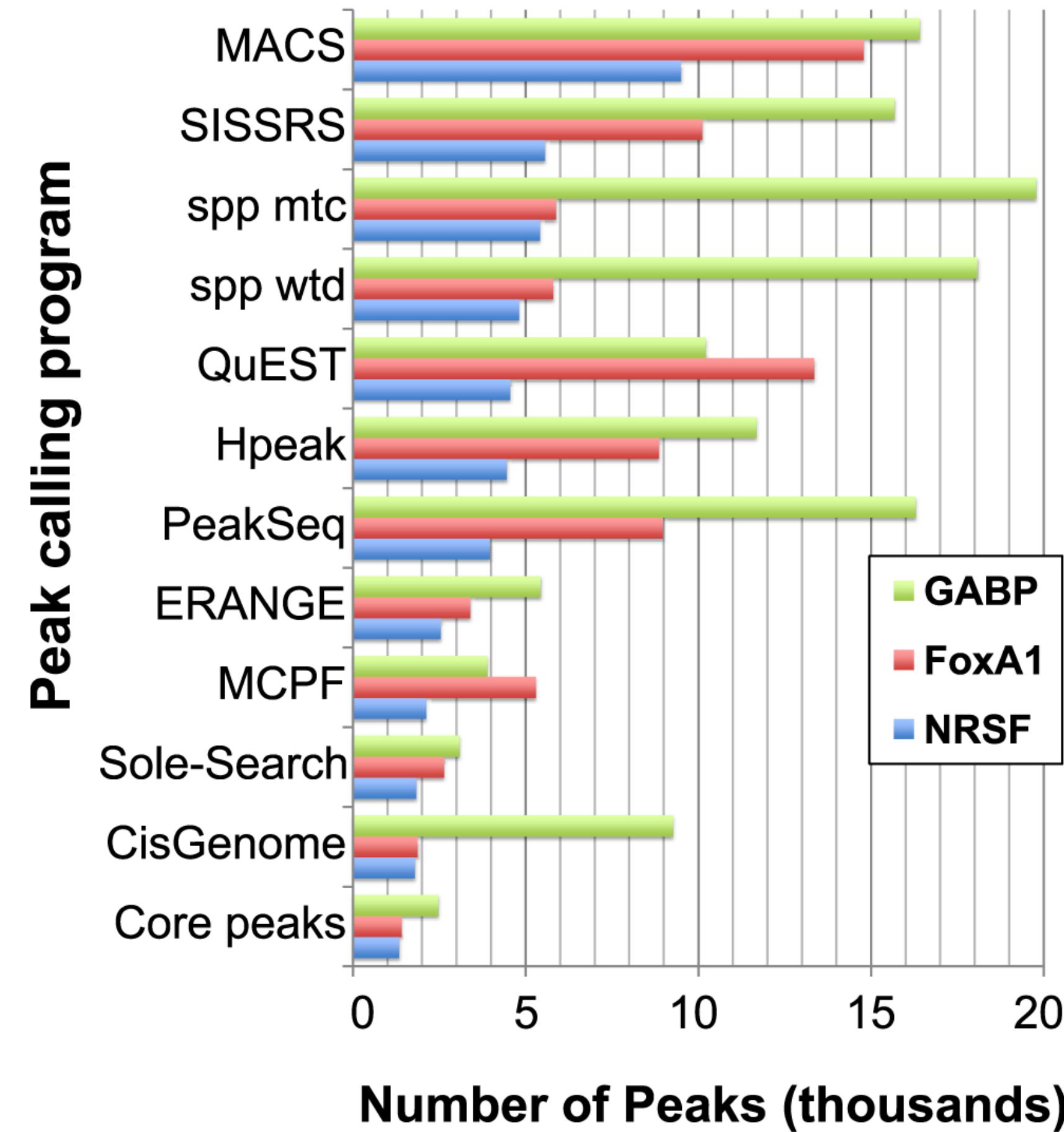


[http://en.wikipedia.org/wiki/Poisson\\_distribution](http://en.wikipedia.org/wiki/Poisson_distribution)

# Peak callers

---

- Variability in number of peaks called
- Tend to agree on the strongest signals



# How to choose one

---

- Widely used
- Actively maintained and updated
- Default settings are a good start but know your parameters for your peak caller
- Be critical! Visually inspect your data (IGV)

# Downstream analysis

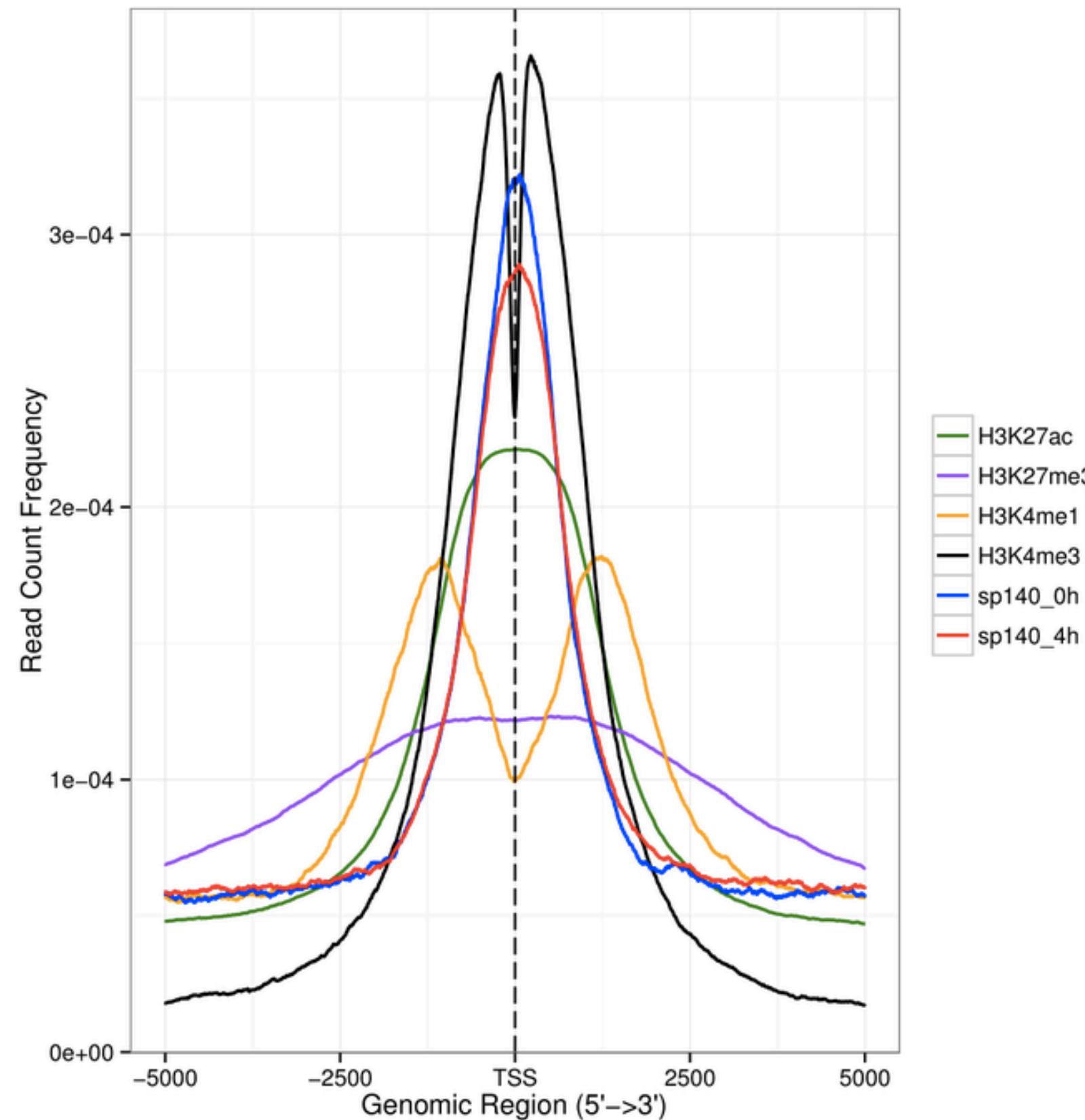
---

- Detecting differential enrichment across samples
  - Steinhauser et al, Brief Bioinform. (2016)

# Downstream analysis

---

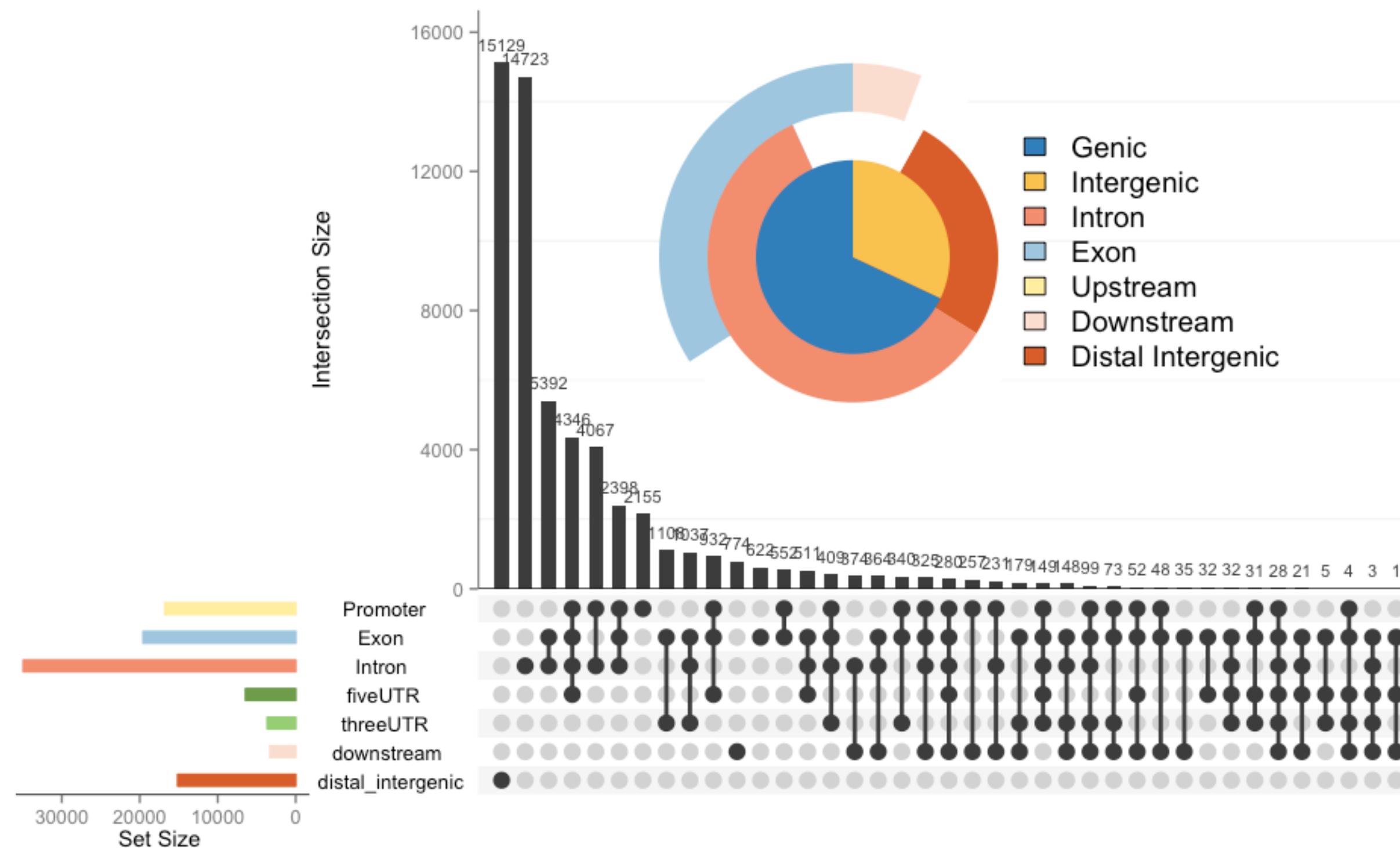
- Annotation of peaks - distance from TSS
  - [ChIPseeker](#), Homer, ChiLin



# Downstream analysis

---

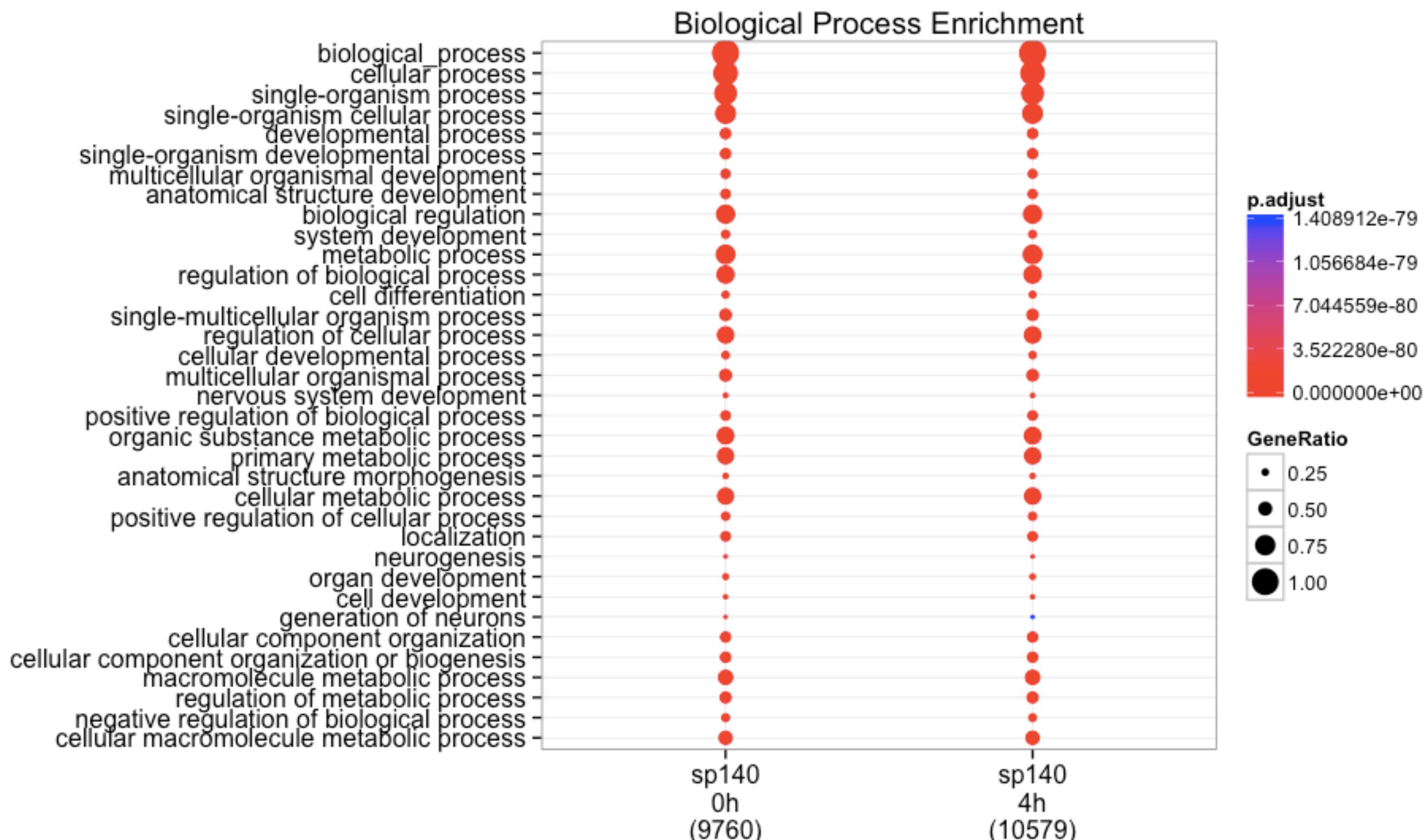
- Annotation of peaks - genomic context
  - [ChIPseeker](#), Homer, ChiLin



# Downstream analysis

---

- Functional enrichment analysis
  - [ChIPseeker](#), GREAT, Homer, ChiLin



# Downstream analysis

---

- Motif discovery
  - MEME suite, ChiLin, Homer



For further information on how to interpret these results or to get a copy of the MEME software please access <http://meme.nbcr.net>.

If you use DREME in your research please cite the following paper:

Timothy L. Bailey, "DREME: Motif discovery in transcription factor ChIP-seq data", *Bioinformatics*, 27(12):1653-1659, 2011. [\[full text\]](#)

[DISCOVERED MOTIFS](#) | [INPUTS & SETTINGS](#) | [PROGRAM INFORMATION](#)

## DISCOVERED MOTIFS

Motif	Logo	RC Logo	E-value	Unerased E-value	More	Submit/Download
1. CYWTTGTB			4.2e-299	4.2e-299	<a href="#">↓</a>	<a href="#">→</a>
2. ATGBWAAT			8.4e-179	1.1e-179	<a href="#">↓</a>	<a href="#">→</a>
3. CCMCDCCC			1.3e-130	1.1e-131	<a href="#">↓</a>	<a href="#">→</a>

# Summary

---

- Basics of the ChIP protocol
- Better understanding of how to design a ChIP experiment
- How to analyze the data
- What to look for in a good ChIP data set

# Ask us questions

[shosui@hsph.harvard.edu](mailto:shosui@hsph.harvard.edu)

[bioinformatics.sph.harvard.edu](mailto:bioinformatics.sph.harvard.edu)

