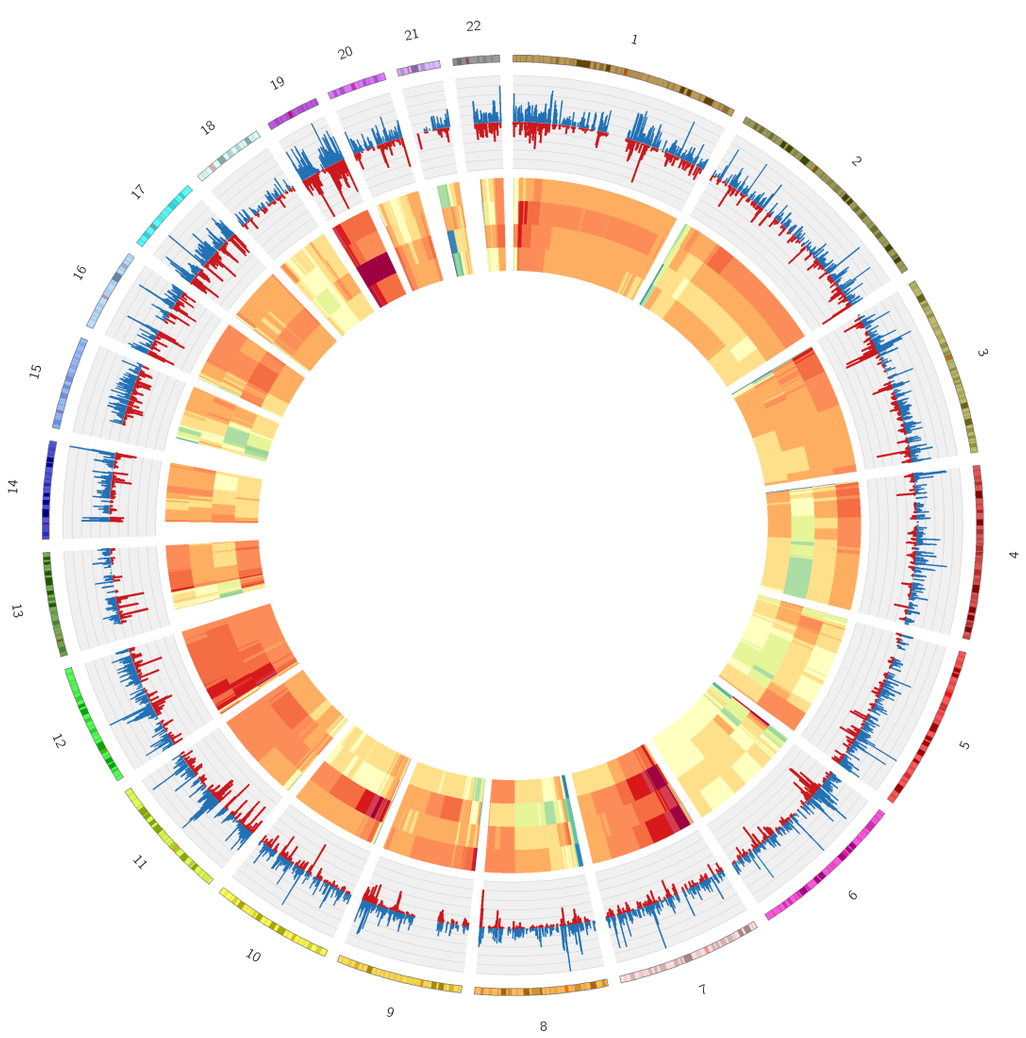


Introduction to ChIP-seq

Joanna Krupka

CRUK Summer School in Bioinformatics

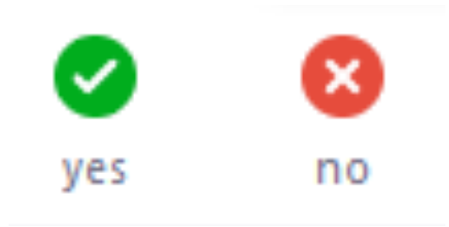
Cambridge, July 2020

CANCER RESEARCH UK

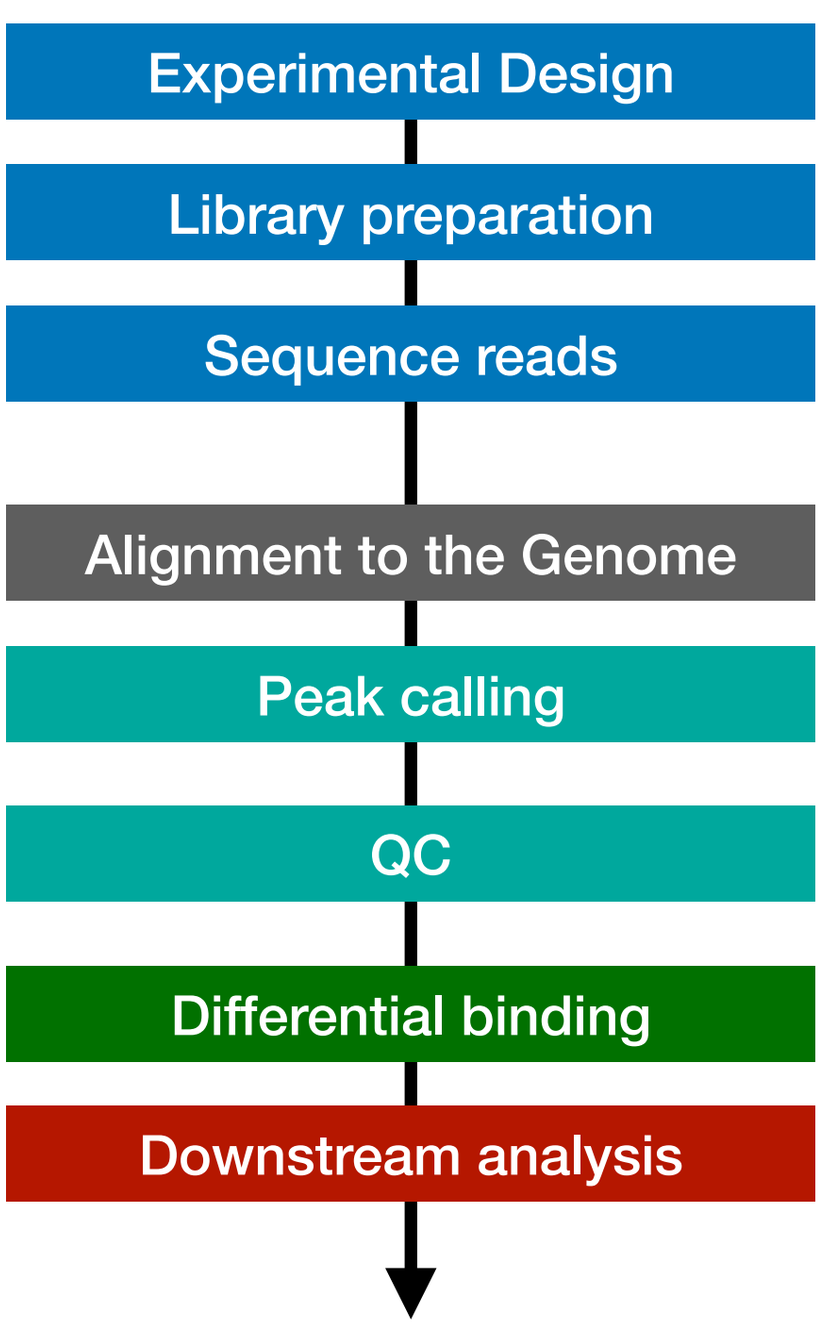
MRC | Cancer Unit

UNIVERSITY OF CAMBRIDGE

How many of you have used ChIP-Seq or think will use it in the future?



Workflow for today



9:30-10:30 Introduction to ChIP seq

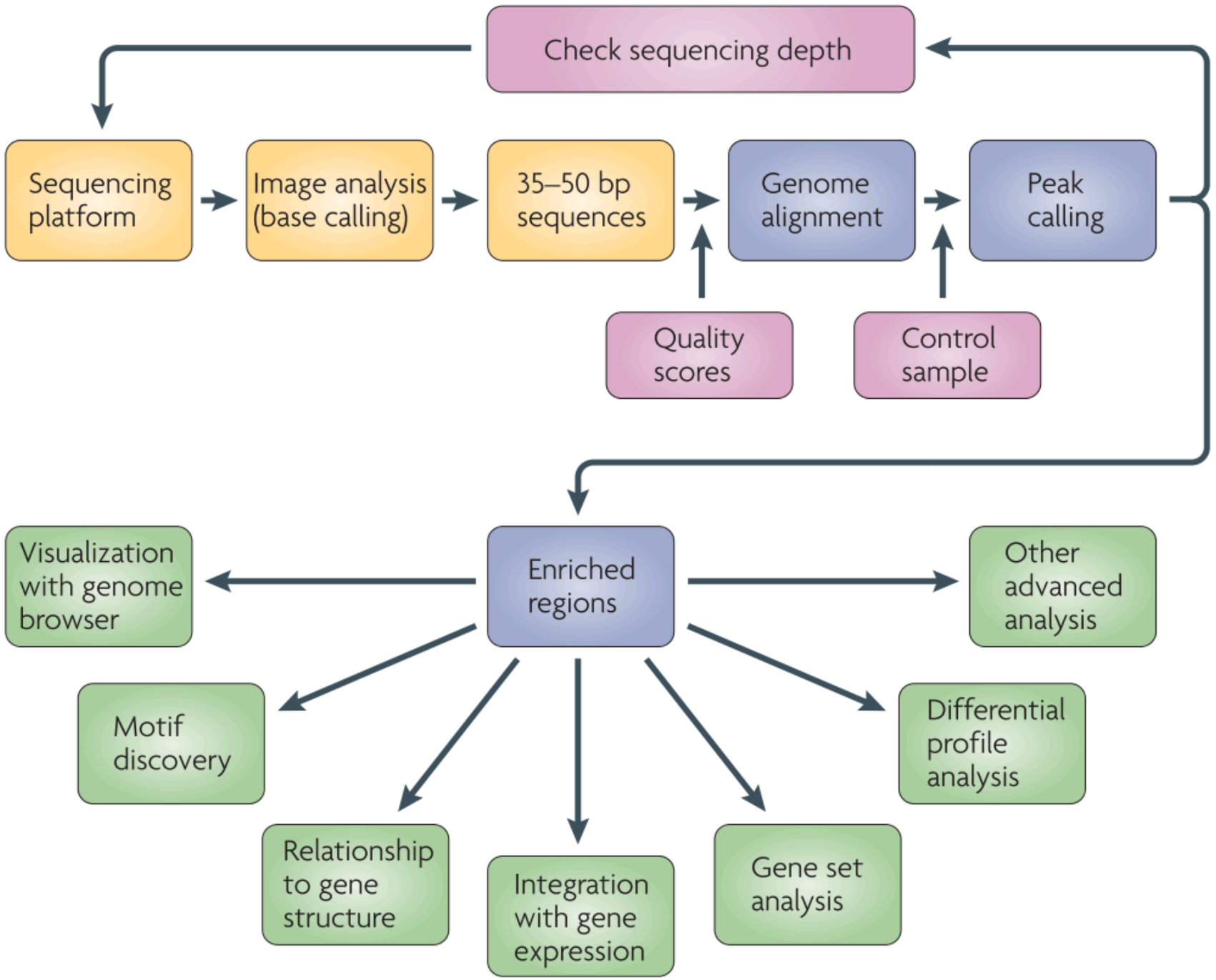
10:40-11:00 Peak calling

10:40-13:50 Evaluation of ChIP seq Data

14:30-15:30 Differential Binding

15:40-17:00 Downstream analysis

ChIP-Seq workflow



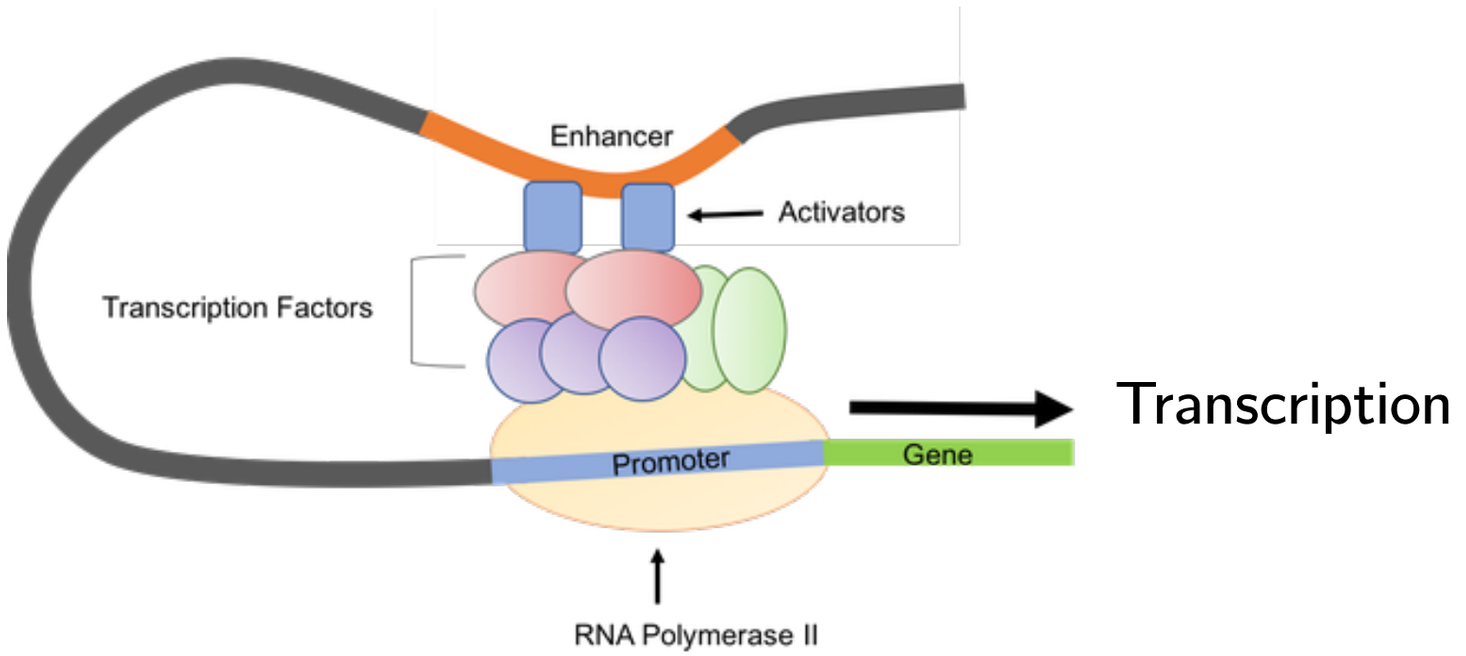
Furey, T. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* **13,** 840–852 (2012).

Gene expression regulation is complex

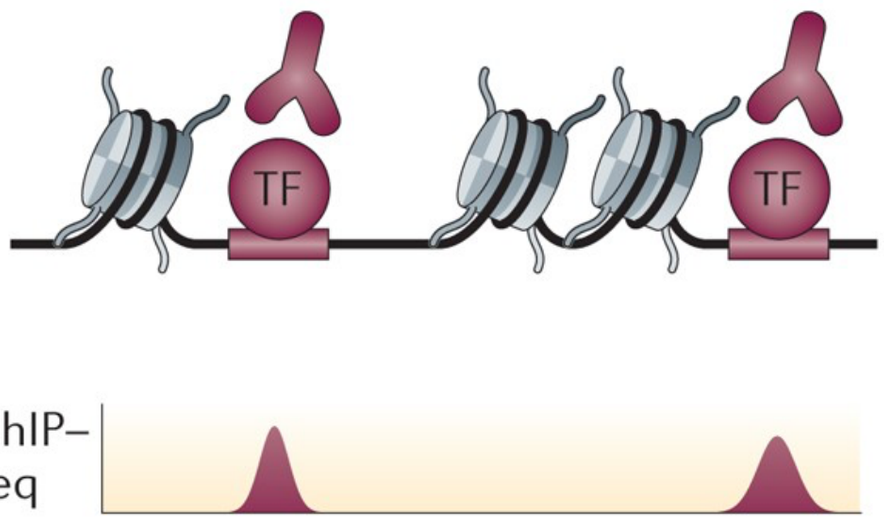
Transcription factor expressed?

Chromatin structure
(open/close)

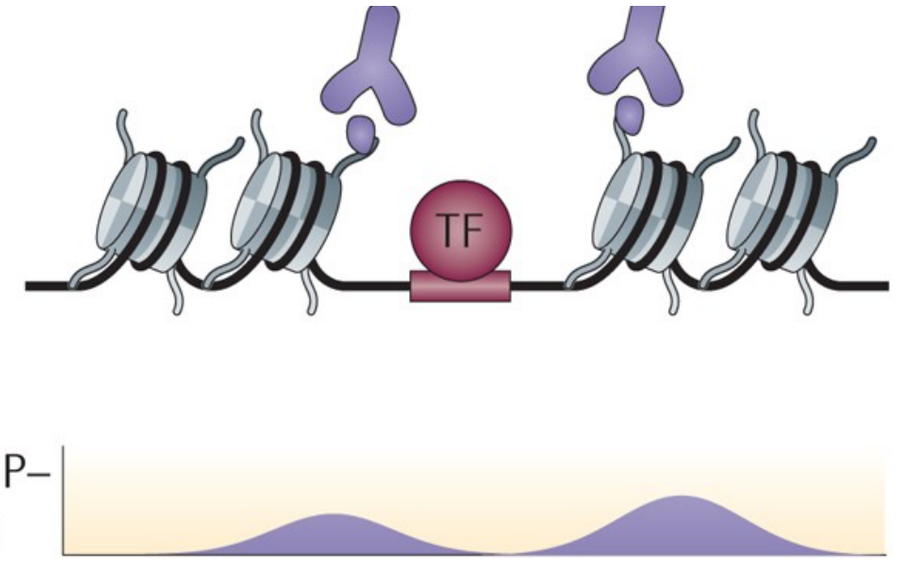
Transcriptional machinery



ChIP-Seq for TF



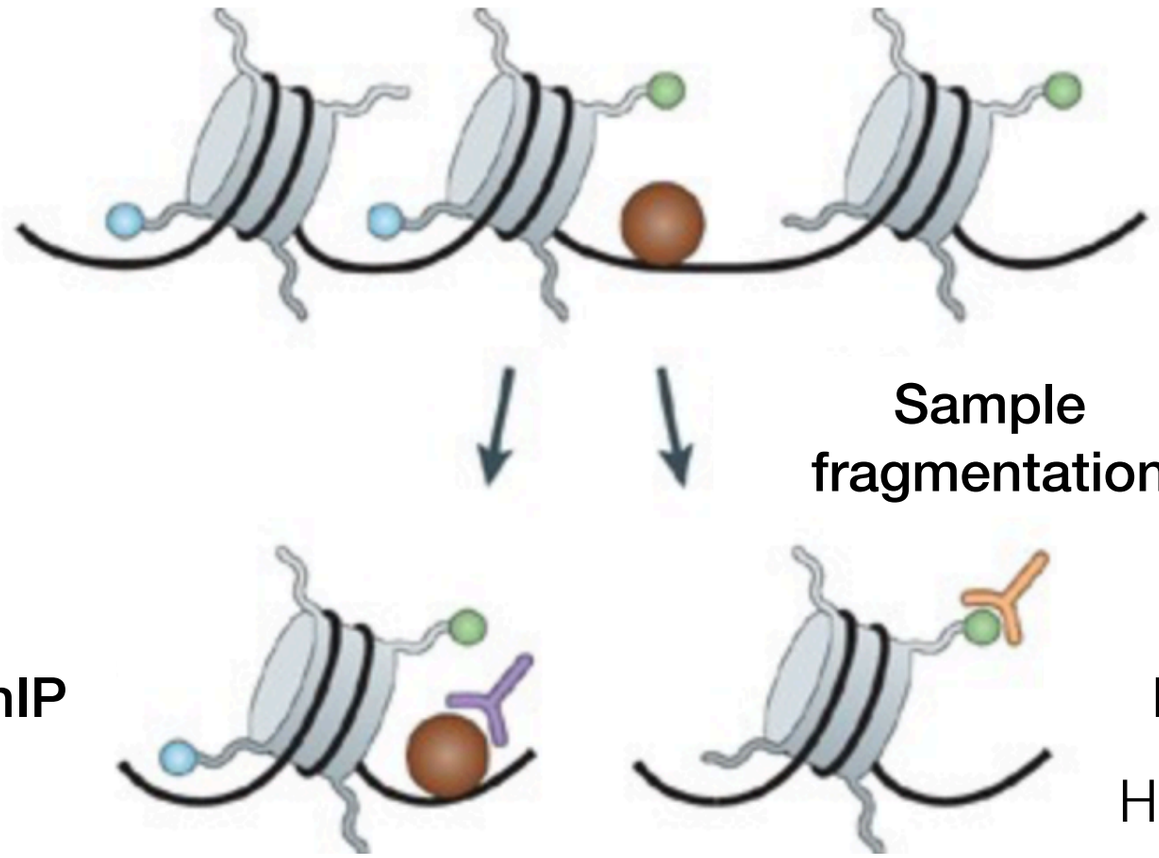
ChIP-Seq for Chromatin marks



Furey, T. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* **13,** 840–852 (2012).

Chromatin immunoprecipitation + NGS

Aim: identify binding sites of DNA-binding proteins or the location of modified histones *in vivo* on a **genome scale**



Transcription factors

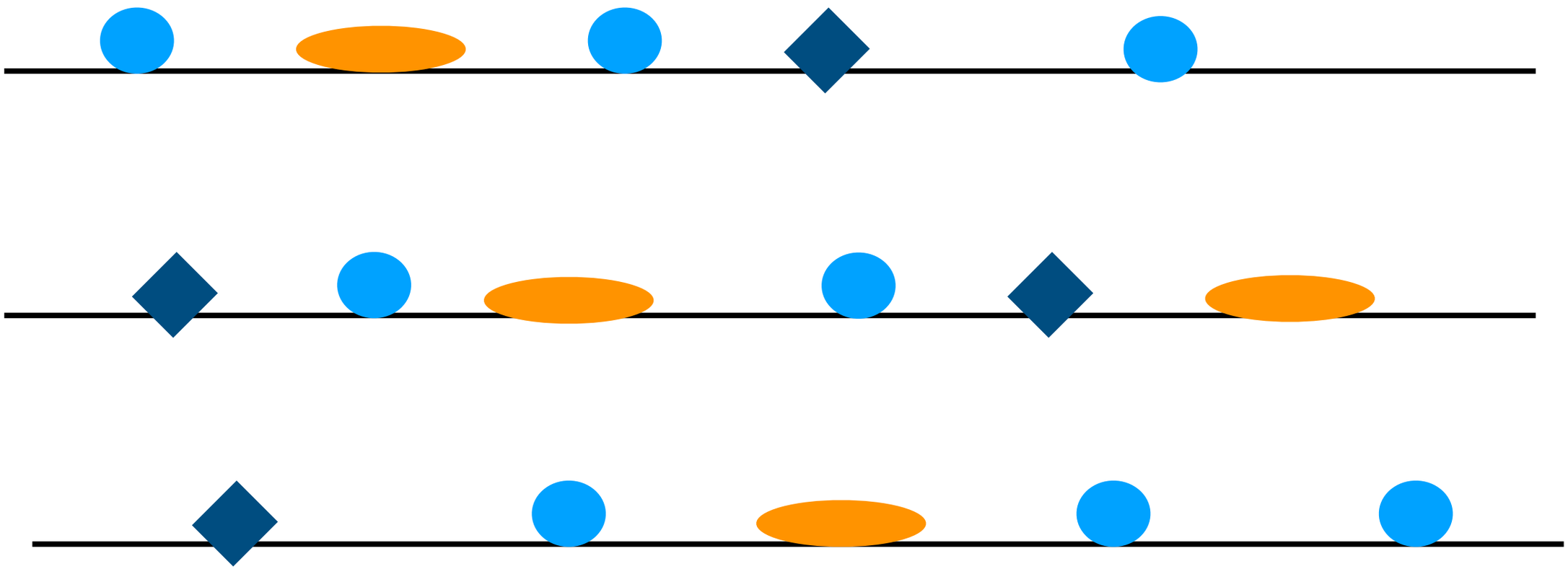
DNA binding proteins (HP1, Lamins, HMGA etc.)

RNA Pol-II occupancy

Histone modification marks

Furey, T. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* **13,** 840–852 (2012).

There are some proteins bound to DNA...



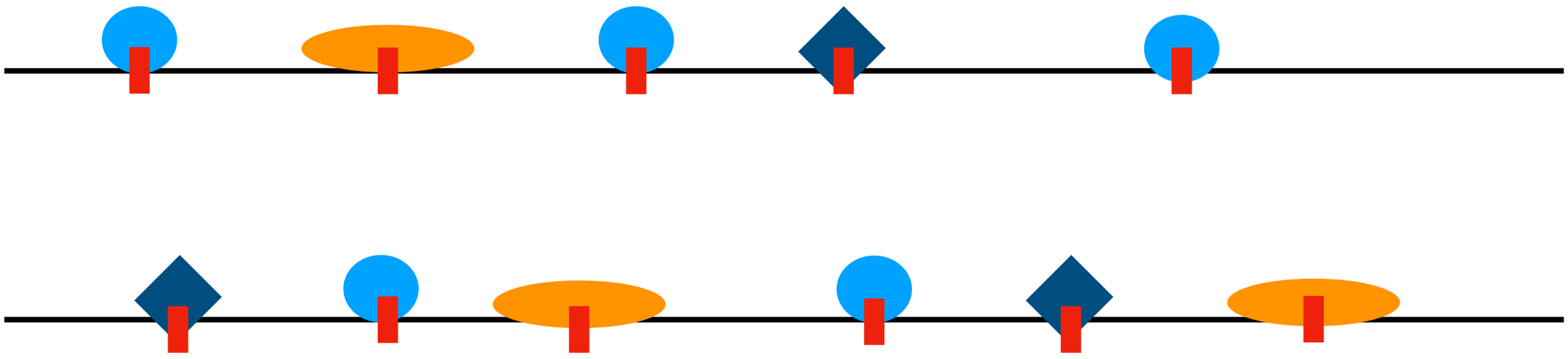
Transcription factors

DNA binding proteins (HP1, Lamins, HMGA etc.)

RNA Pol-II

Histones (H1, H2A, H2B, H3 and H4)

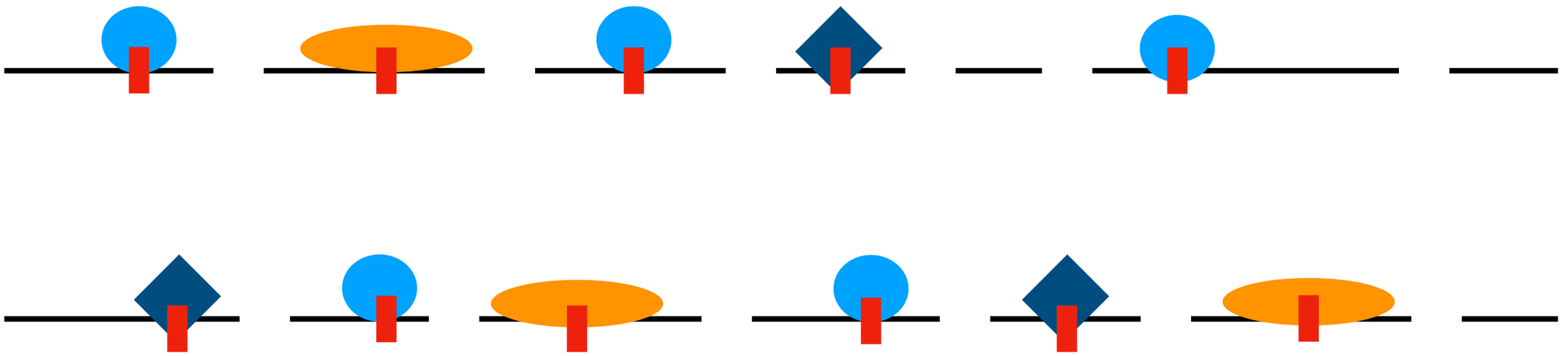
Crosslinking



Usually - formaldehyde crosslinking

There may also be changes in nucleosome positions and histone modifications during the course of the experiment in the absence of crosslinking.
ChIP without X-linking is called: **N-ChIP ("native")** - more effective in some biological models (eg. muscle tissue)

Fragmentation

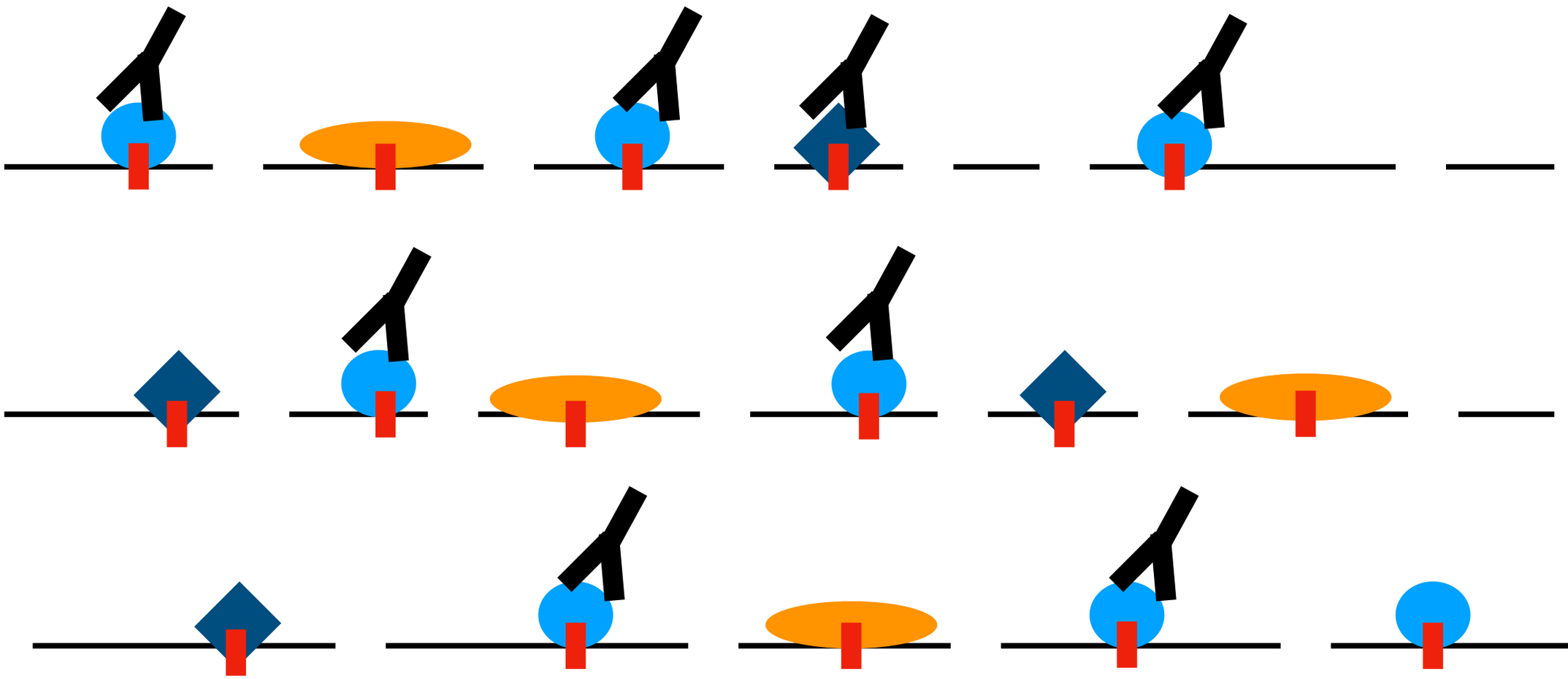


The DNA is sheared into small fragments - usually 200-500 bp in length

But it is not entirely random!

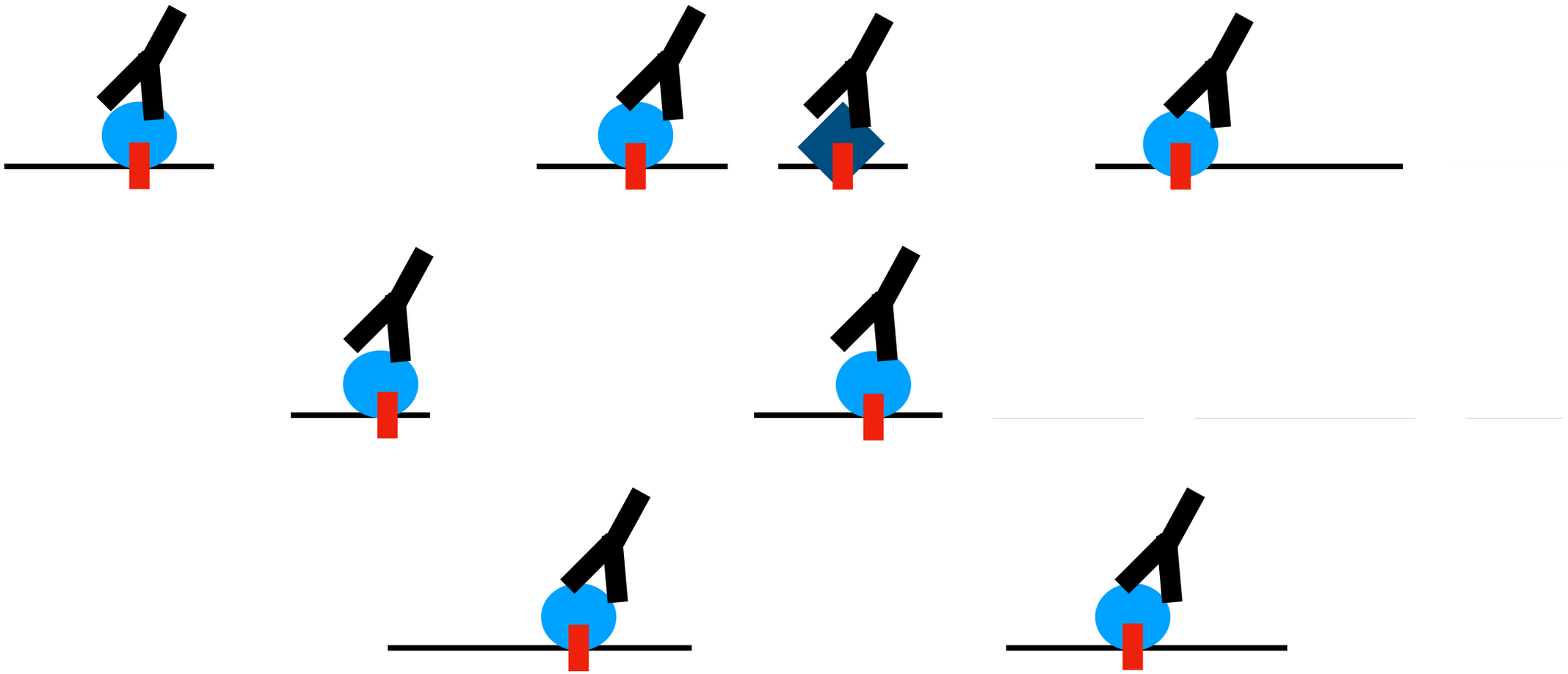
Eg. open chromatin regions tend to be fragmented more easily than closed regions, which creates an uneven distribution of sequence tags across the genome.

Protein-specific antibody



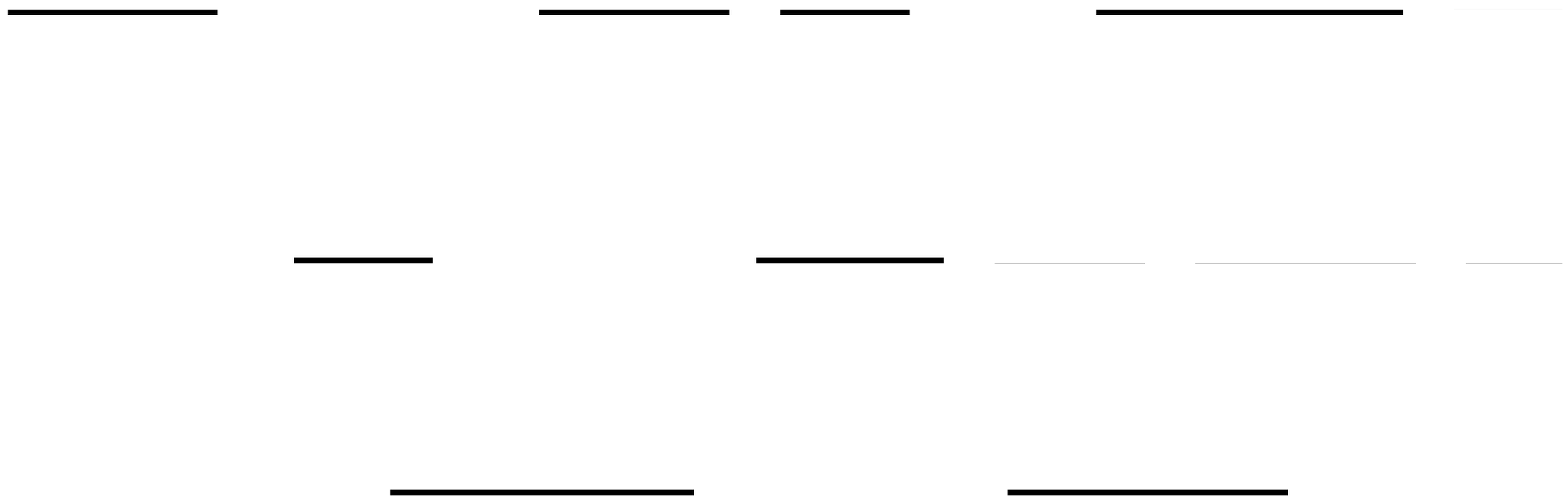
The sheared protein-bound DNA is immunoprecipitated using a specific antibody

Immunoprecipitation



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

Reverse cross-reaction, purify DNA, sequence



Sequencing ~10 ng of ChIP DNA recommended

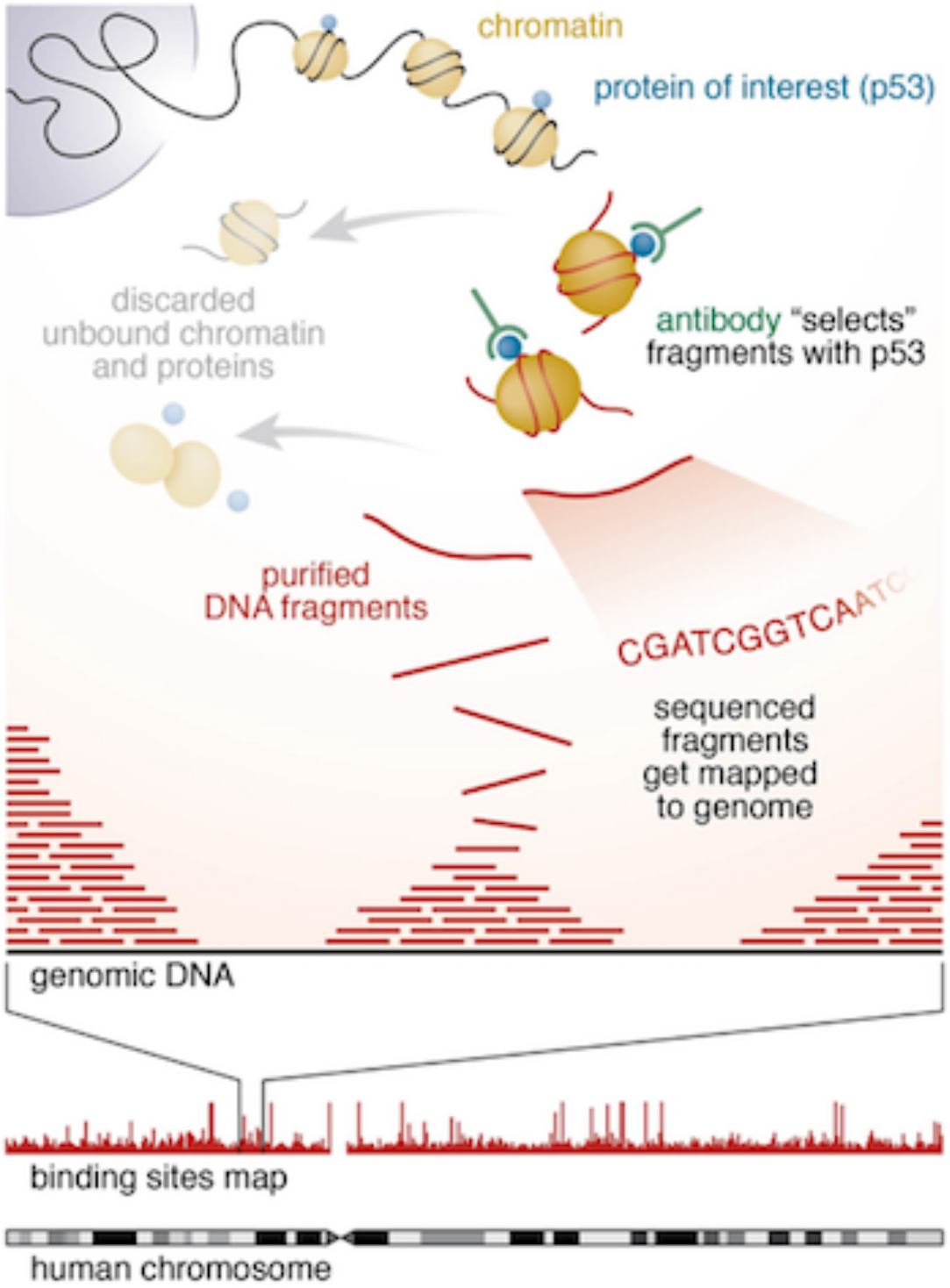
NOTE: beware of amplification bias - fewer cycles, the better!

Main experimental steps in the ChIP-Seq protocol

The typical ChIP assay
usually take 4–5 days, and require
approx. 10^6 - 10^7 cells.

Recipe for successful experiment:

- **Good Experimental Design** (enough replicates!)
- Optimized Conditions (Cells, Antibodies ...)
- Good biological question that can be answered with this technique
- Efficient and specific antibody
- Sufficient amount of starting material (ChIP DNA depends on cell type, abundance of the mark or protein, quality of antibody)



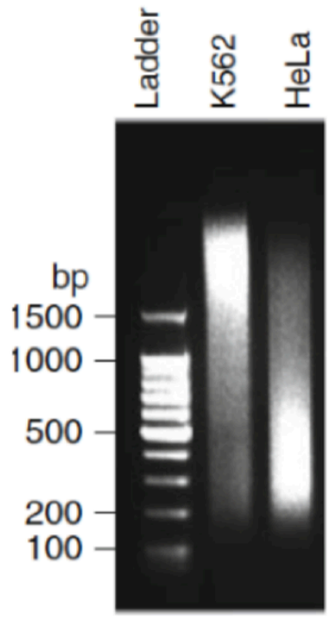
Pitfalls during ChIP-Seq protocol

1. 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

2. Gel size selection

The most variable step

Differences between investigators!

3. Specificity of the antibody

Variability between different lot numbers of the same antibody!

Time-consuming, but rewarding validation: ~1/4 of the tested histone antibodies failed specifici
criteria by dot blot or western blot

Histone modifications:

- the reactivity of the antibody with unmodified histones or non-histone proteins should be checked by western blotting.
- cross-reactivity with similar histone modifications (validated using eg. siRNAs against enzymes that are predicted to add the modifying group)

O'Geen et al (2011), Methods Mol Biol, Schmidt et al (2009), Methods;48(3):240-248.

ENCODE guidelines for antibody and immunoprecipitation characterisation

Primary mode of characterization

- immunoblot of immunofluorescence
- demonstrate that the protein of interest can be efficiently immunoprecipitated from a nuclear extract.

Secondary mode of characterization

- Knock-d
- Immunoprecipitation followed by mass-spectrometry
- Immuoprecipitation with multiple antibodies against different parts of the target protein or members of the same complex to demonstrate specificity of the antibody

Full guideline:

Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. 2012;22(9):1813-1831. doi:10.1101/gr.136184.111

What generates ChIP-Seq signal?

ChIP-Seq signal depends on:

- The number of active binding sites
- The number of starting genomes (number of cells)
- IP efficiency (antibody quality, biological model used)
- GC rich content (bias in fragment selection, during amplification)

Globaly

- Open chromatin regions fragment more easily than closed regions (open region will generate more reads than closed one due to non-random fragmentation)
- Differential mappability of short reads to repeat-rich genomic regions (Teytelman et al., 2009, Aird et al., 2011)
- Hyper-ChIPable regions

Localy

What generates ChIP-Seq signal?

ChIP-Seq signal depends on:

Globaly

- The number of active binding sites
- The number of starting genomes (number of cells)
-

A peak in the ChIP–seq profile must be compared with the same region in a matched control sample to determine its significance.

Localy

- differential mappability of short reads to repeat-rich genomic regions (Teytelman et al., 2009, Aird et al., 2011)
- Hyper-ChIPable regions

We DO need controls

2 types of controls:

Check of preferential enrichment step:

test for different types of artefacts

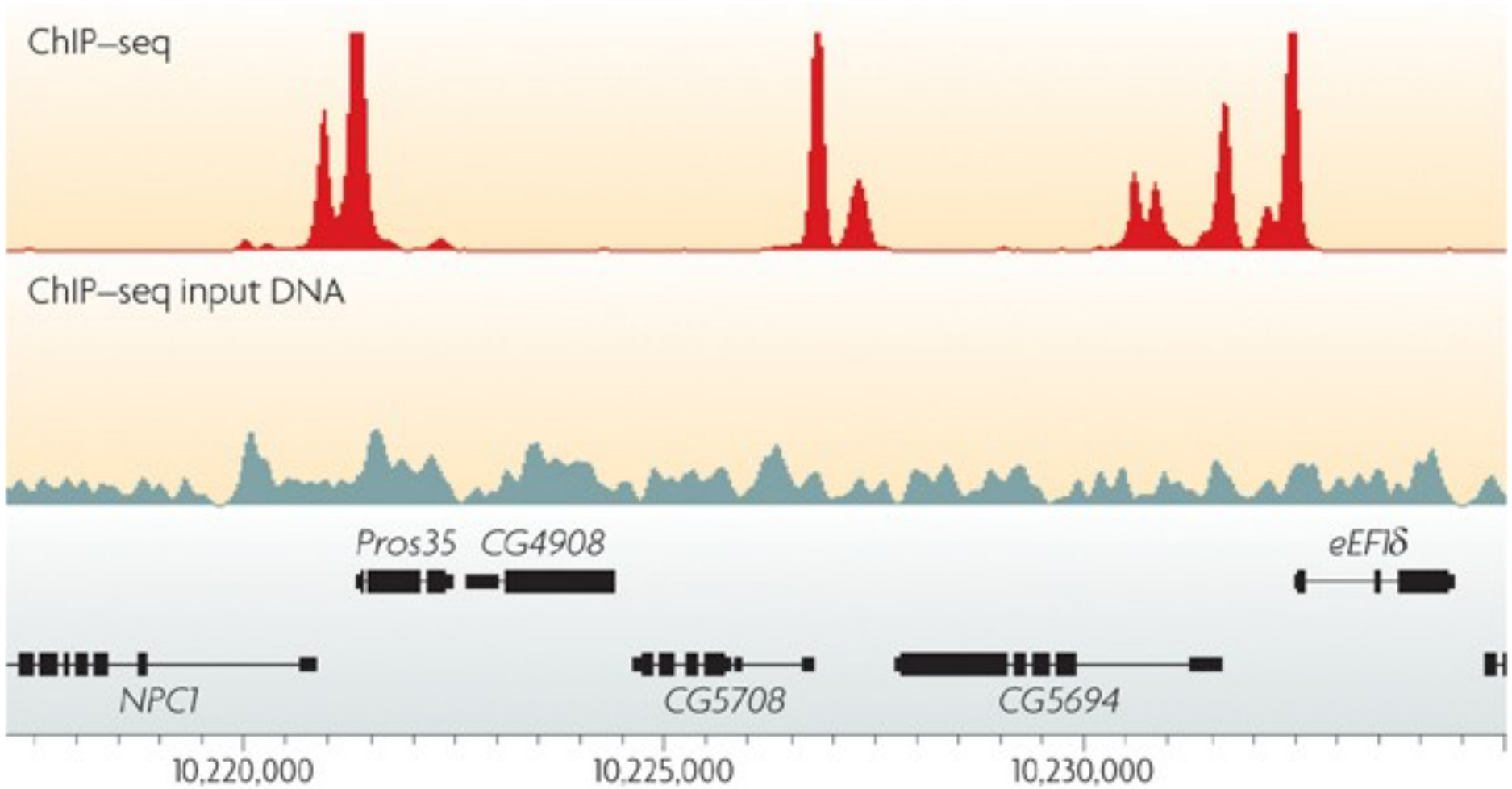
- sonicated DNA before immunoprecipitation (input)
- mock immunoprecipitation with an unrelated antibody (IgG)
- mock IP DNA (DNA obtained from IP without antibodies)

Check of biological specificity of the signal

make biological interpretation easier

- Knock-down/WT sample

Signal-to-noise



Different types of signal

Transcription factors

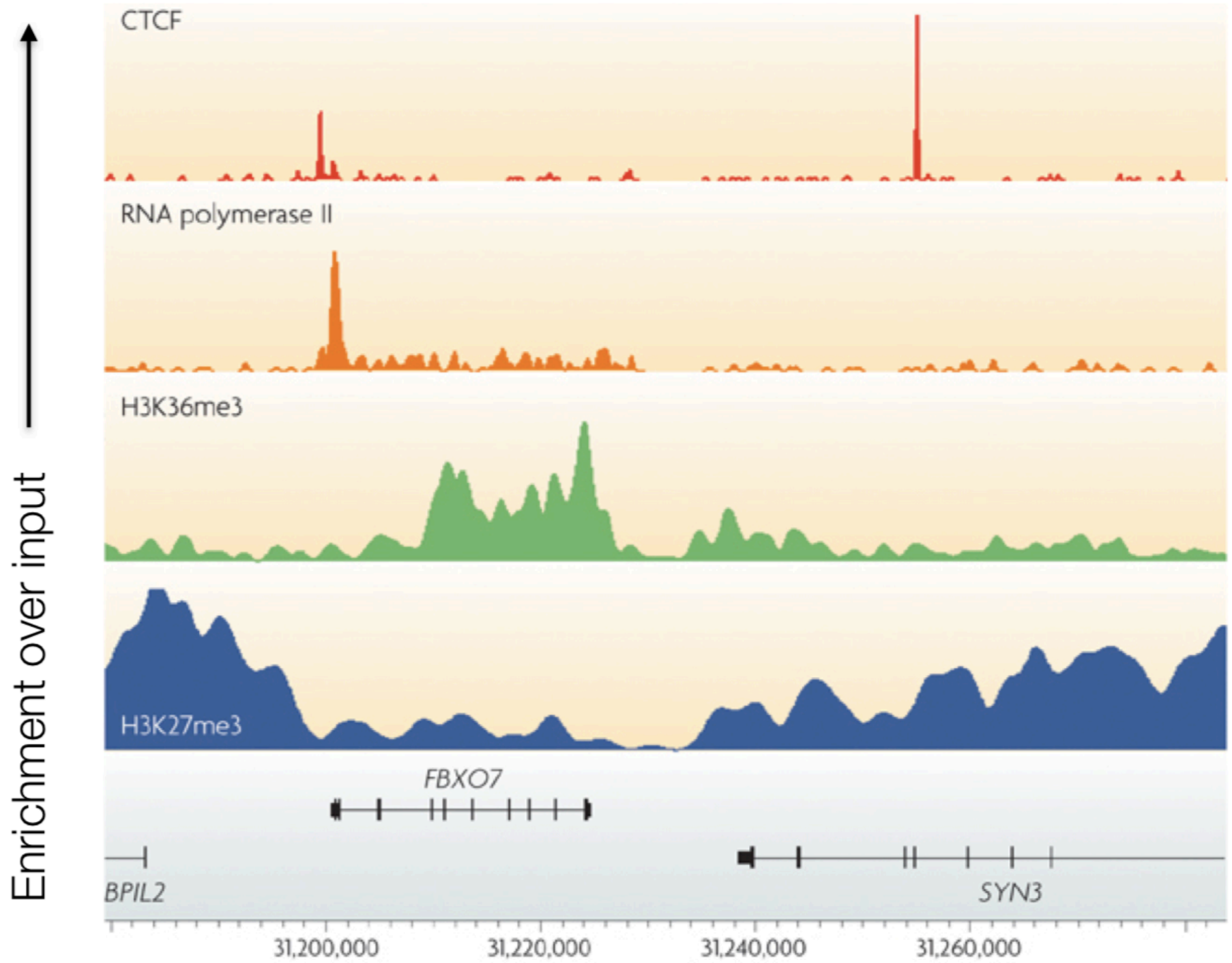
Sharp & localised

Histone modifications

Varying:

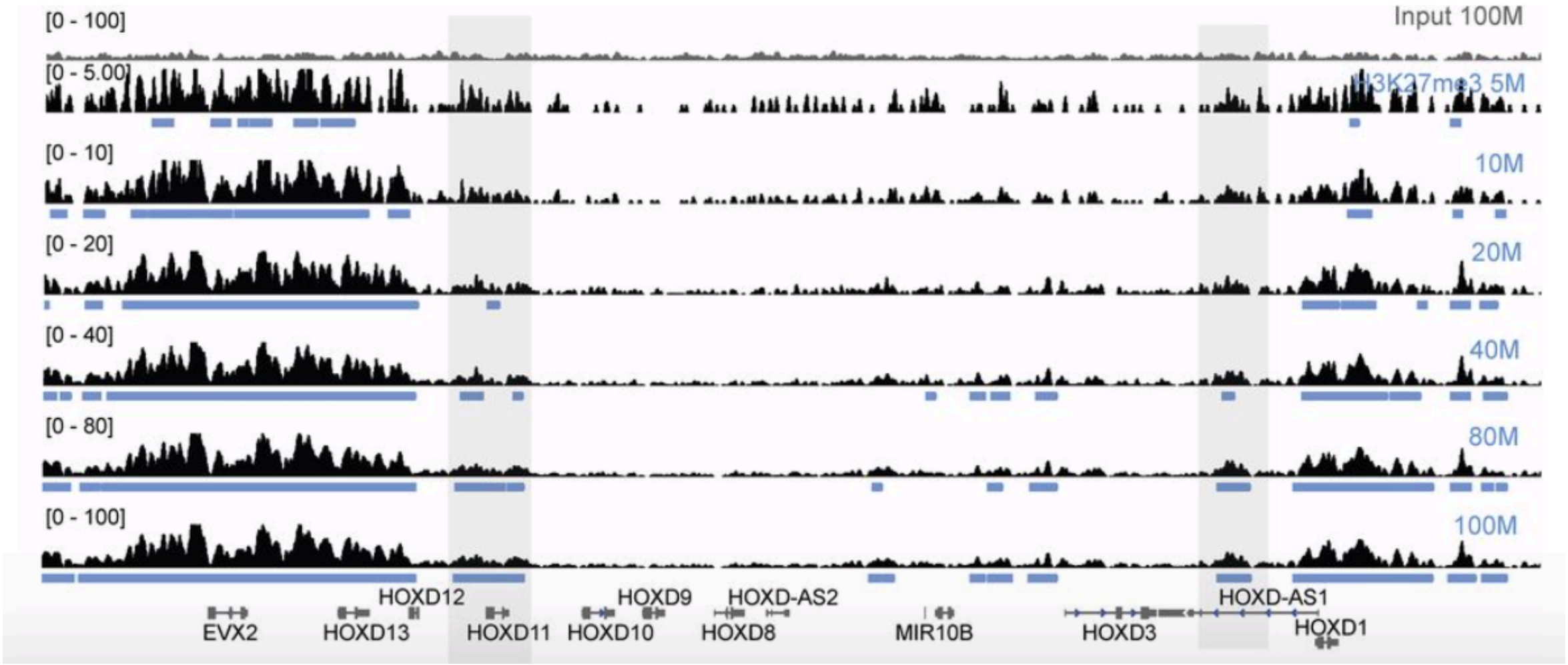
Sharp or broad

Broad peaks: more difficult to find, need deeper sequencing!



Park (2009). Nature Reviews Genetics.

ChIP-Seq signal & sequencing depth



Rule of thumb: More prominent peaks are identified with fewer reads, versus weaker peaks that require greater depth.

Youngsook L. Jung, Lovelace J. Luquette, Joshua W.K. Ho, Francesco Ferrari, Michael Tolstorukov, Aki Minoda, Robbyn Issner, Charles B. Epstein, Gary H. Karpen, Mitzi I. Kuroda, Peter J. Park, Impact of sequencing depth in ChIP-seq experiments, *Nucleic Acids Research*, Volume 42, Issue 9, 14 May 2014, Page e74, https://doi.org/10.1093/nar/gku178

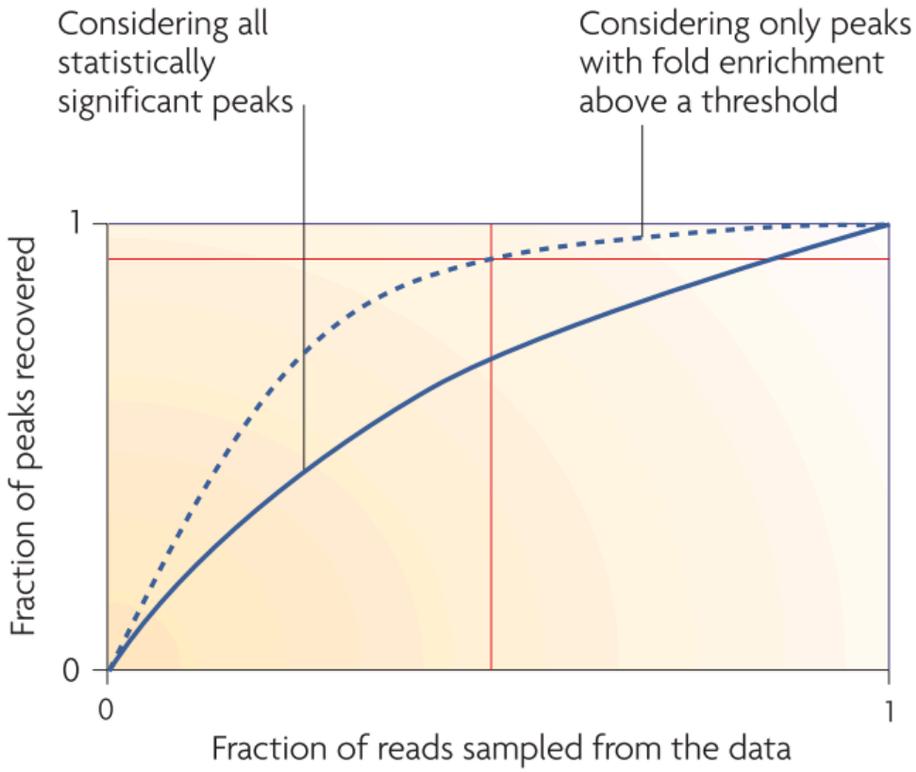
How deep is deep enough?

It's not a simple question!

Saturation:

measure of the fraction of library complexity that was sequenced in a given experiment; depends on library complexity and sequencing depth

Ideally - sequencing should be deep enough to capture all real binding sites (fully saturated the library)

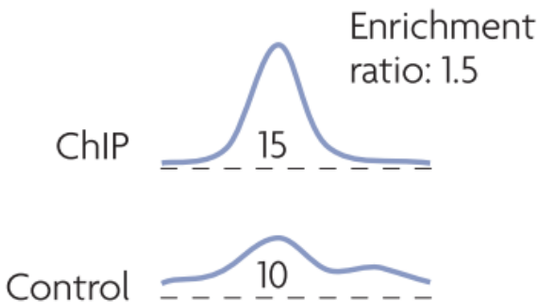


Park P et al. 2009

Youngsook L. Jung, Lovelace J. Luquette, Joshua W.K. Ho, Francesco Ferrari, Michael Tolstorukov, Aki Minoda, Robbyn Issner, Charles B. Epstein, Gary H. Karpen, Mitzi I. Kuroda, Peter J. Park, Impact of sequencing depth in ChIP-seq experiments, *Nucleic Acids Research*, Volume 42, Issue 9, 14 May 2014, Page e74, https://doi.org/10.1093/nar/gku178

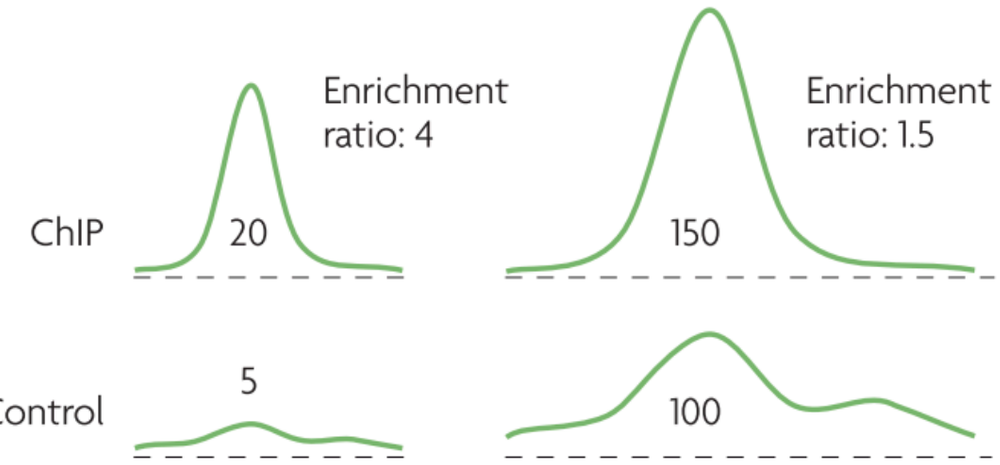
Significant or not?

Not significant



Too low enrichment
Too few tags

Significant

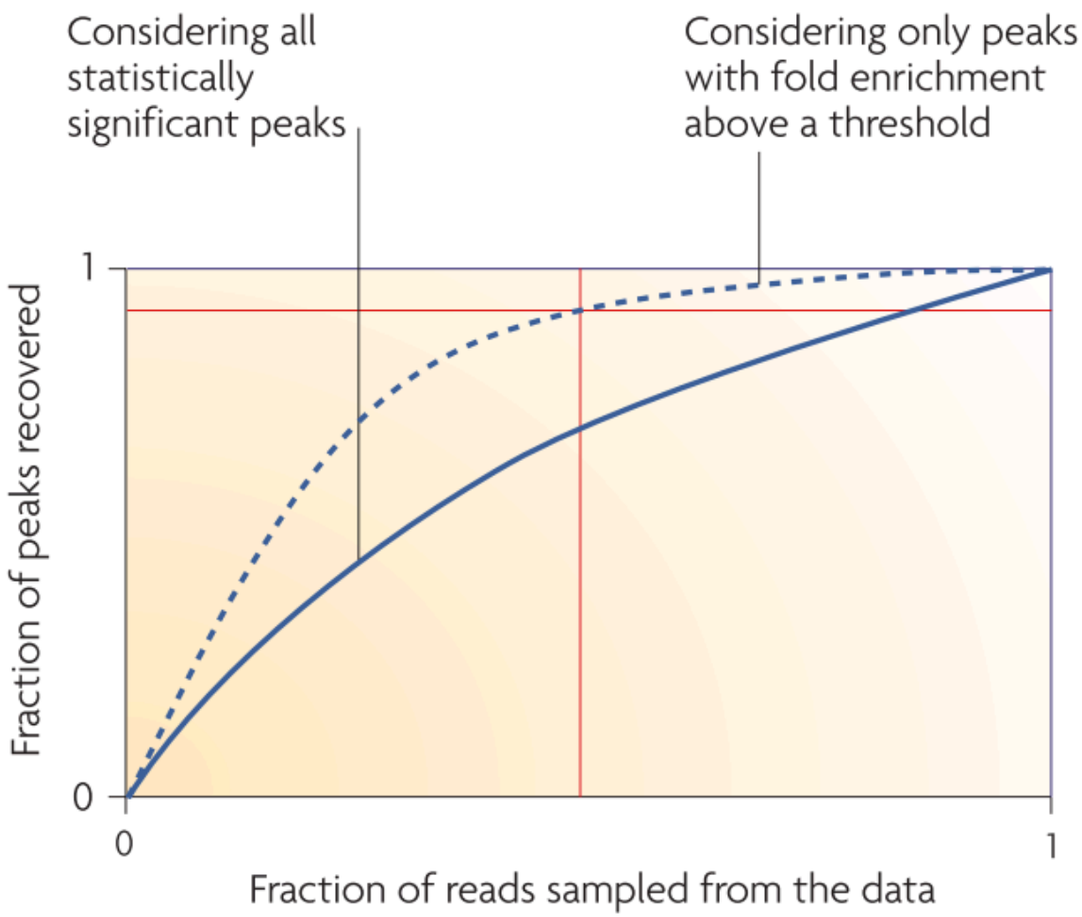


High enrichment
Too few tags

Low enrichment
A lot of tags

How deep is deep enough?

Simulation to characterise the fraction of the peaks that would be recovered if a smaller number of tags had been sequenced



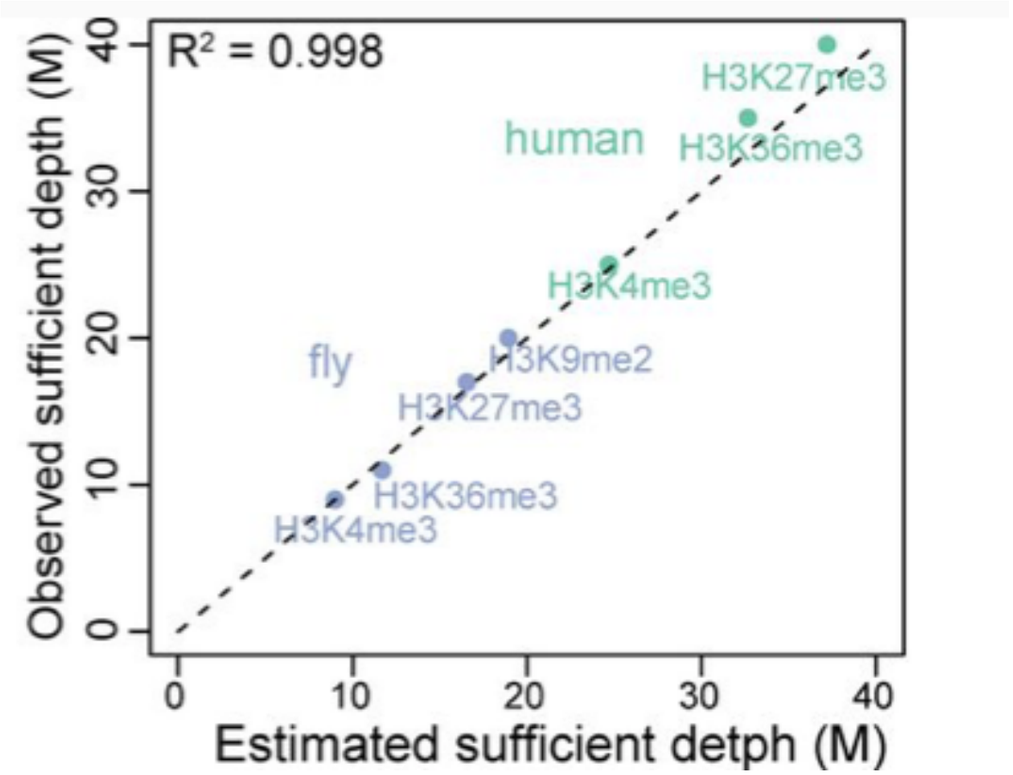
NOTE: Even for transcription factors (sharp, clear peaks), the number of valid peaks increases without saturation as more reads are sequenced <u>if only statistical significance is used</u>.

Even very small peaks become statistically significant when the number of reads at those peaks gets larger.

Park P et al. 2009

Saturation of ChIP-Seq signal

'Sufficient depth': the sequencing depth at which the percent gain in enriched regions per 1 million additional sequence reads falls below 1%



There is no universal "sufficient" sequencing depth

20 mln reads - TFs (ENCODE standard)

25 mln reads - H3K4me3

35 mln reads - H3K36me3

40 mln reads - H3K27me3

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

Youngsook L. Jung, Lovelace J. Luquette, Joshua W.K. Ho, Francesco Ferrari, Michael Tolstorukov, Aki Minoda, Robbyn Issner, Charles B. Epstein, Gary H. Karpen, Mitzi I. Kuroda, Peter J. Park, Impact of sequencing depth in ChIP-seq experiments, *Nucleic Acids Research*, Volume 42, Issue 9, 14 May 2014, Page e74, https://doi.org/10.1093/nar/gku178

Blacklisted regions

Blacklisted regions are genomic regions with anomalous, unstructured, high signal or read counts in NGS experiments, independent of cell type or experiment. Including these regions can lead to false-positive peaks. Often found at repetitive regions (centromeres, telomeres, Satellite repeats)

Problems:

- tend to have a lot of multi-mapping reads
- high variance of mappability
- difficult to remove with simple mappability filters

Once reads have been aligned to the reference genome, "blacklisted regions" are removed from BAM files before peak calling. Alternatively, peaks overlapping blacklisted regions can be removed.

https://www.encodeproject.org/files/ENCFF356LFX/

Take home messages - highlights from ENCODE guidelines

1. Sufficient sequencing depth: varies between different targets
2. Sufficient amount of starting material
3. Optimisation (antibodies, cells)
4. Control libraries!
5. Reproducibility: at least 3x!