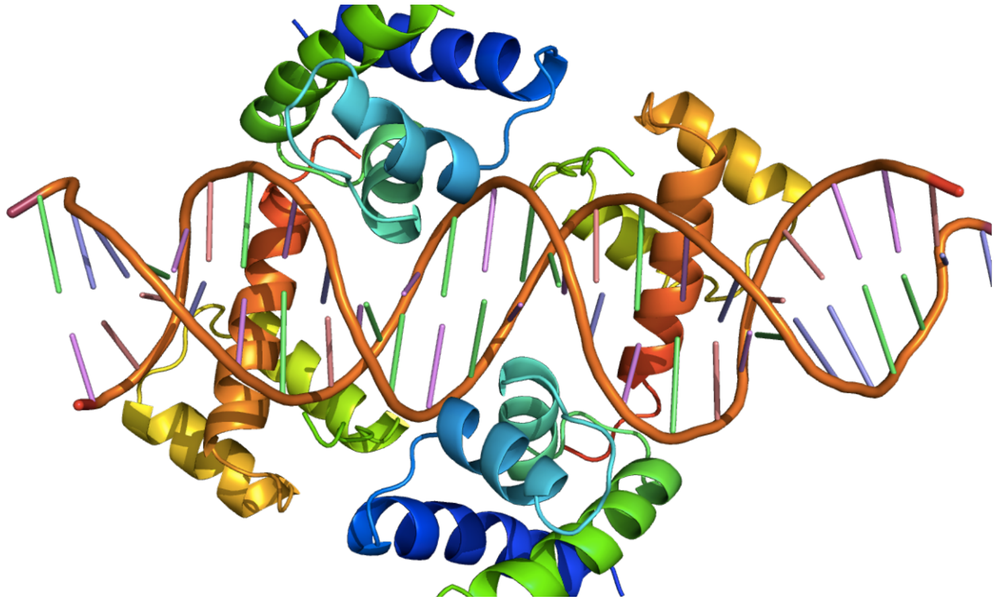
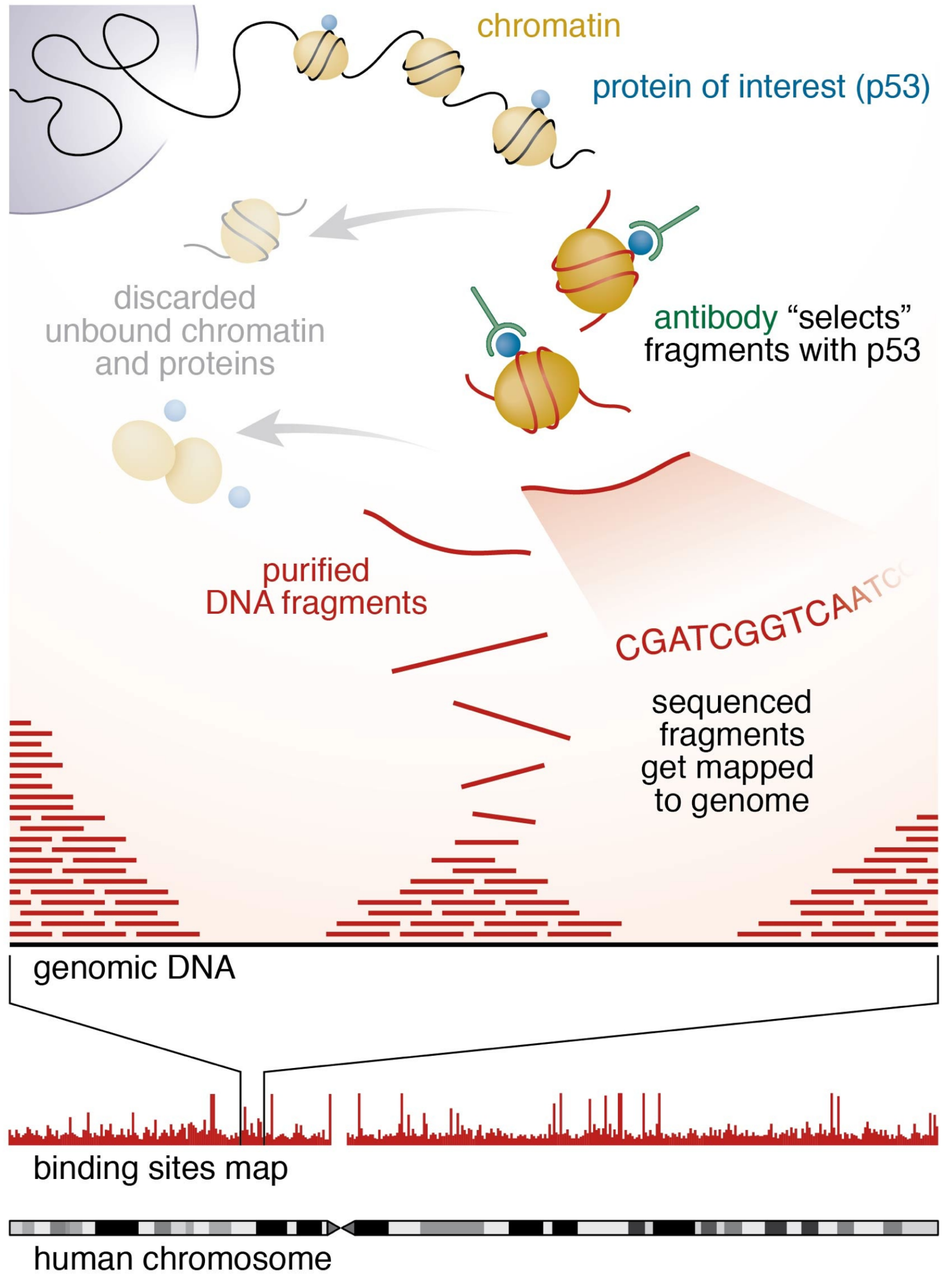


ChIP-seq: Mapping DNA-protein interactions

HSPH Bioinformatics Core

Shannan Ho Sui

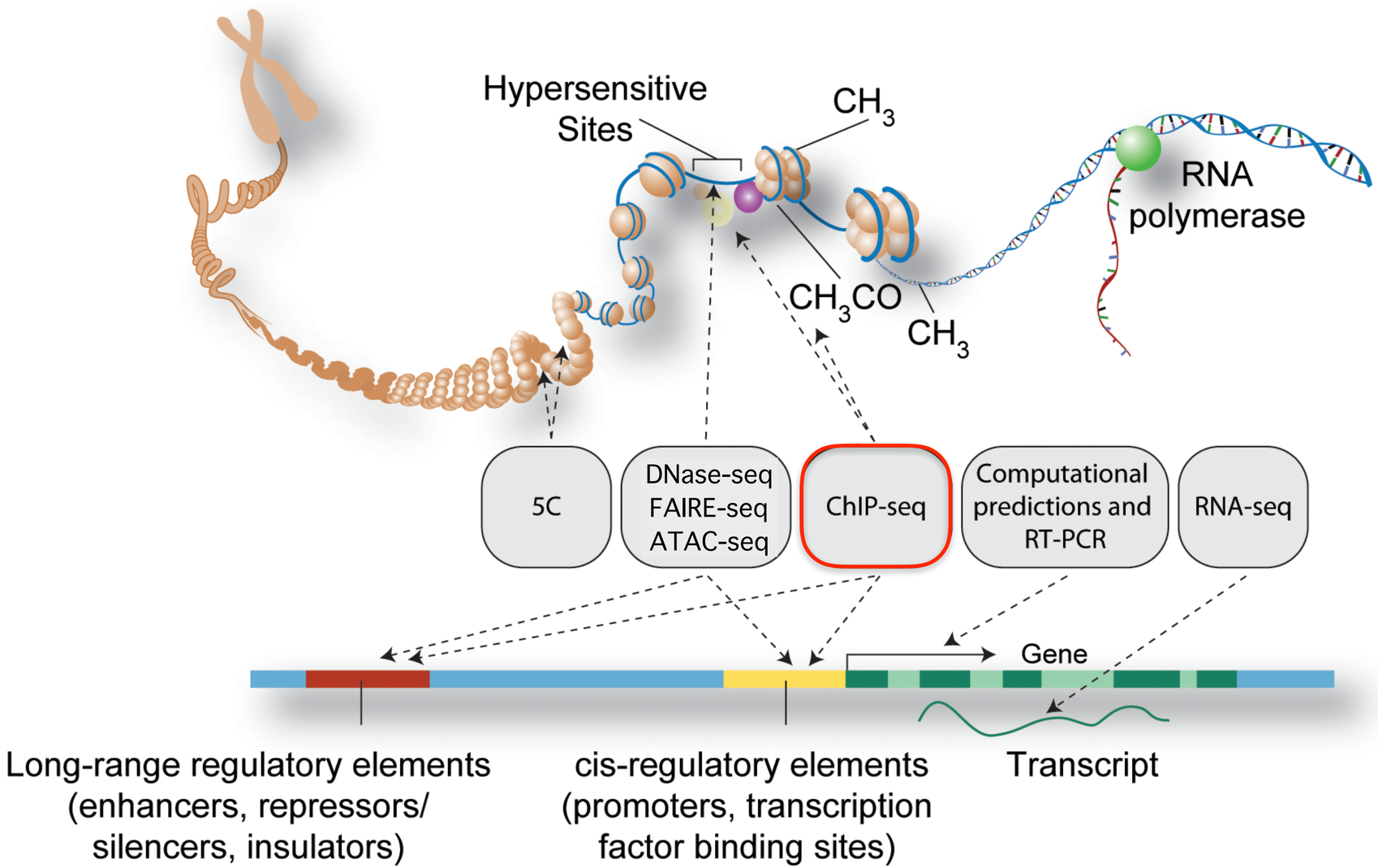
April 29, 2019



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

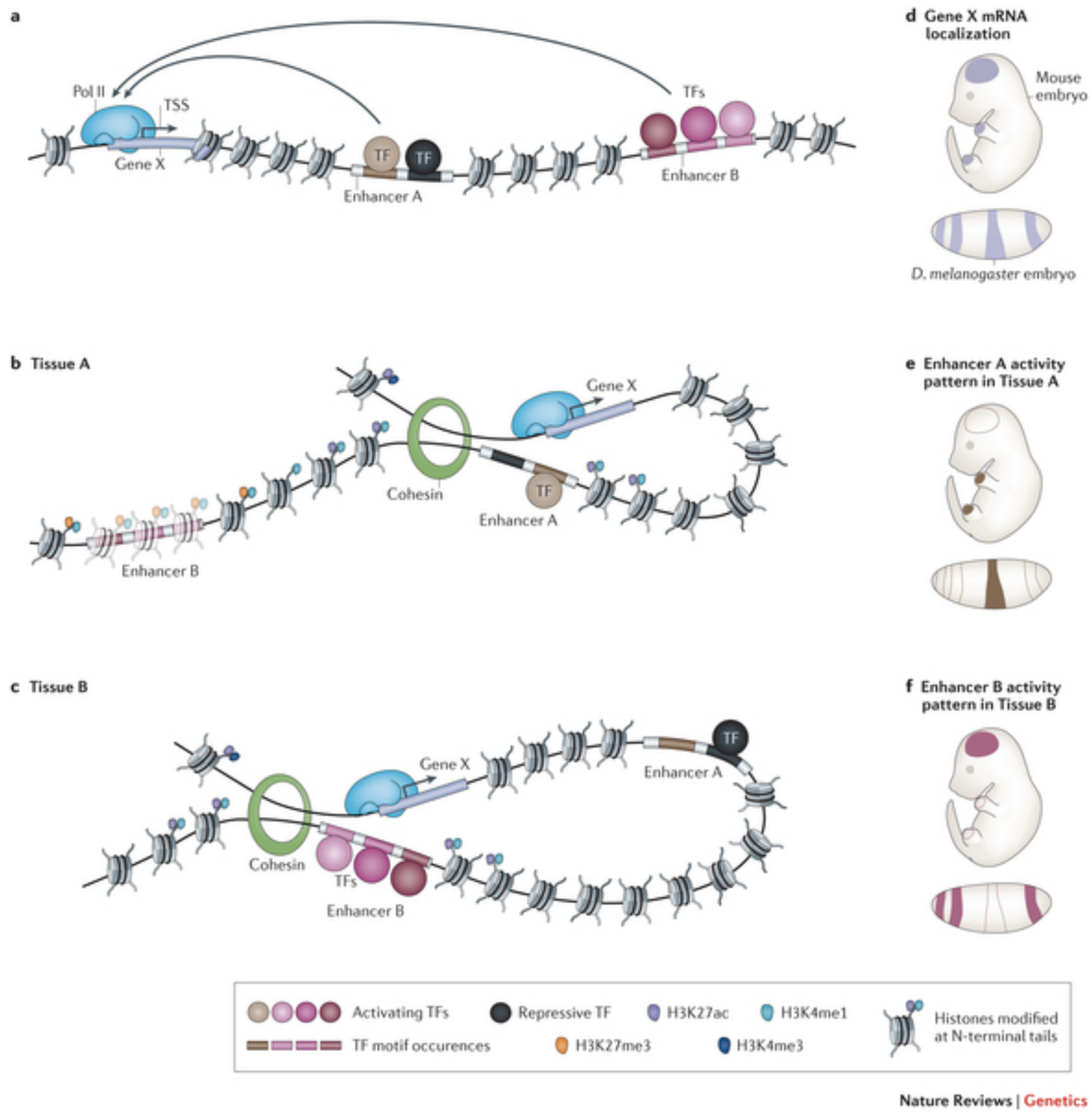
Transcriptional regulation is complex



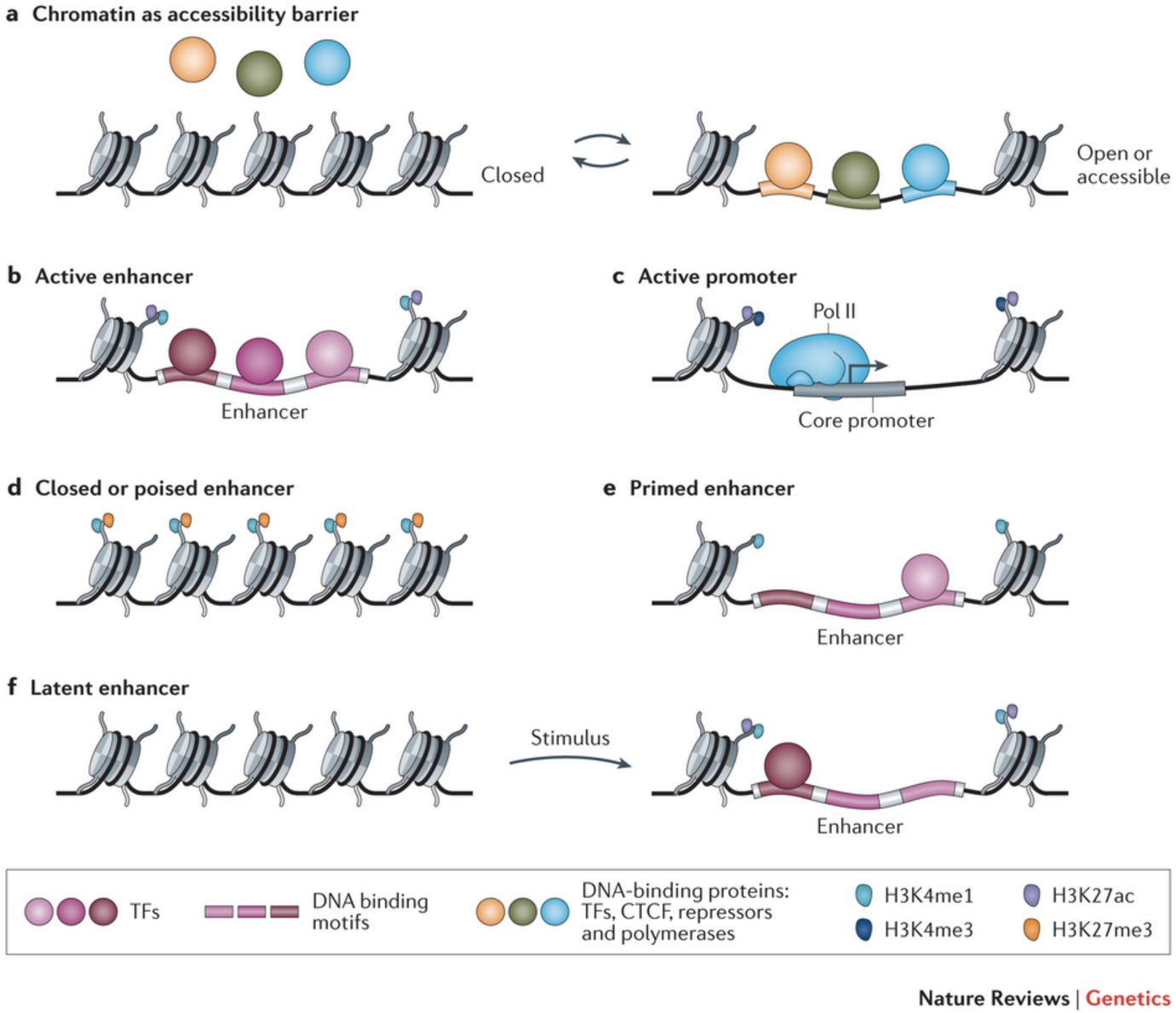
Adapted from The ENCODE Project Consortium (2011). PLOS Biology

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.

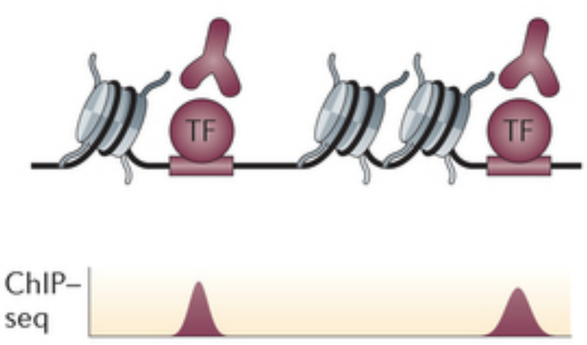


Shlyueva, et al (2014)

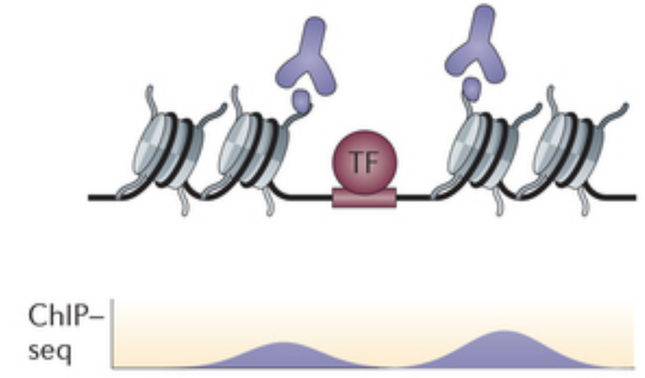
Chromatin structure determines if a gene is expressed or not

Genomic methods for detecting regulatory elements

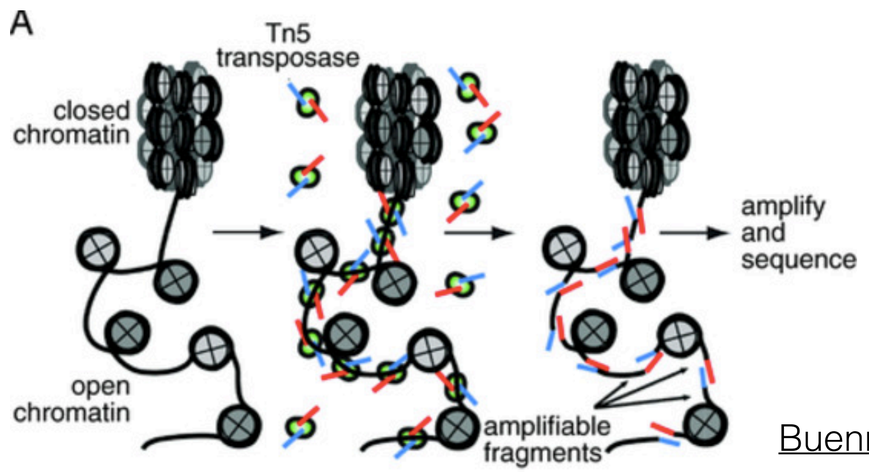
ChIP-seq for a TF



ChIP-seq for chromatin marks

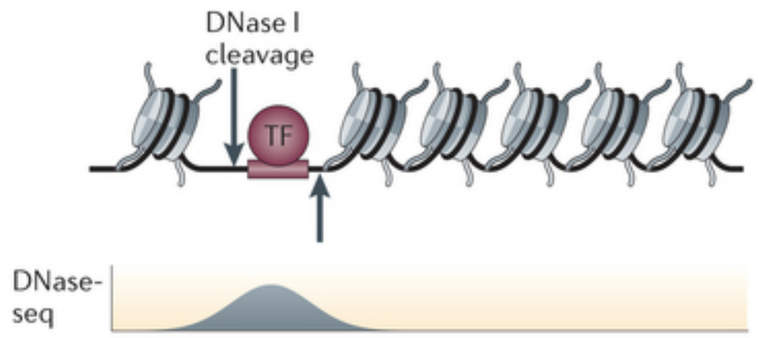


ATAC-seq



Buenrostro et al., 2015

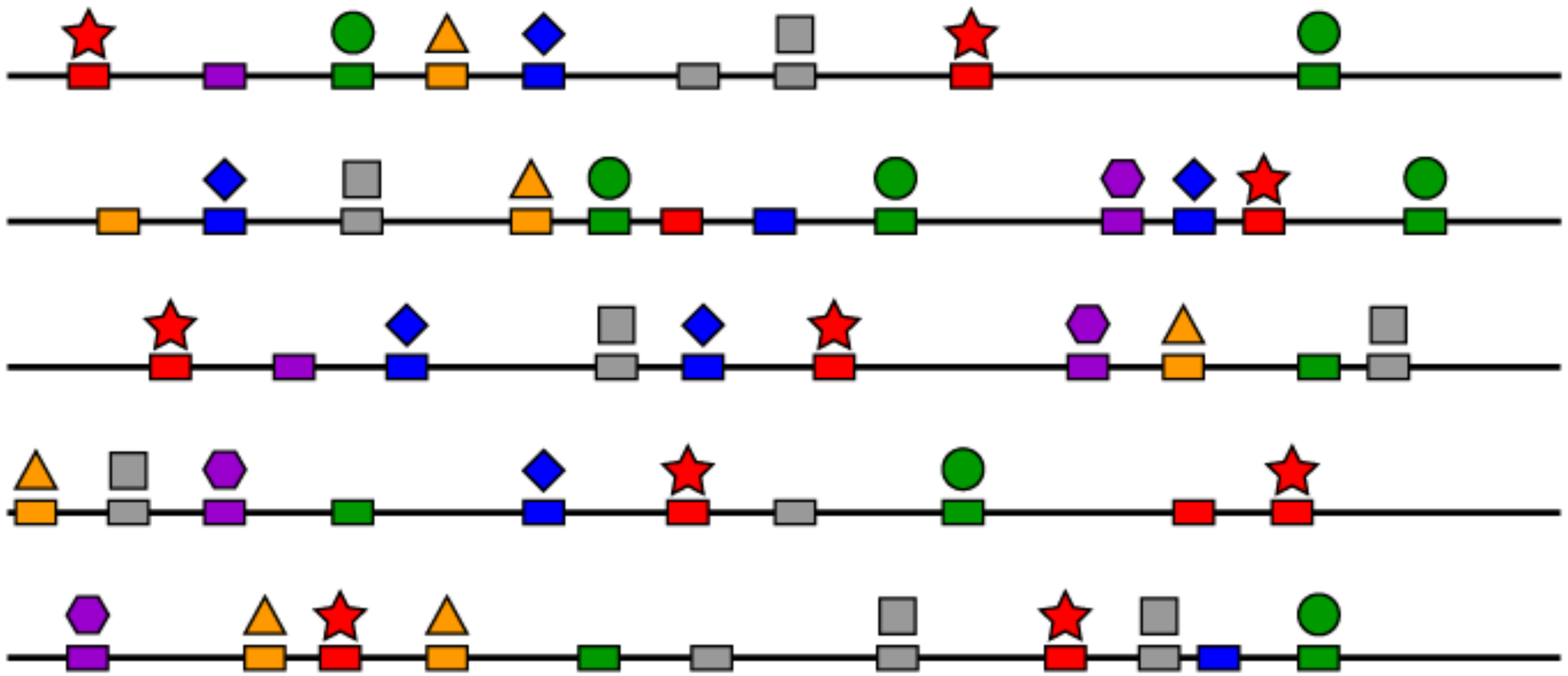
DNase-seq



Shlyueva, et al (2014)

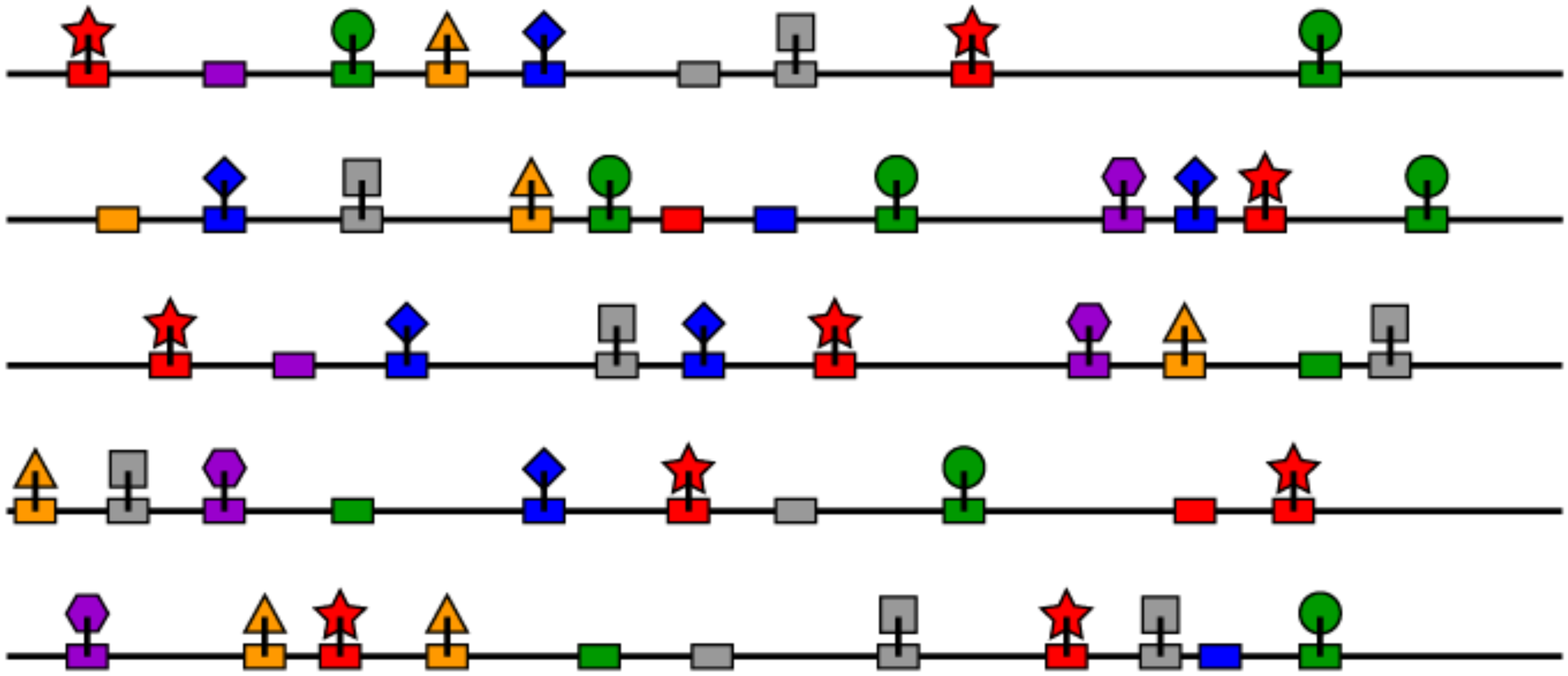
Also ChIA-PET and chromosome conformation capture (3C) based methods to detect not only the contact points but also the pairwise connections between these points

Library Preparation

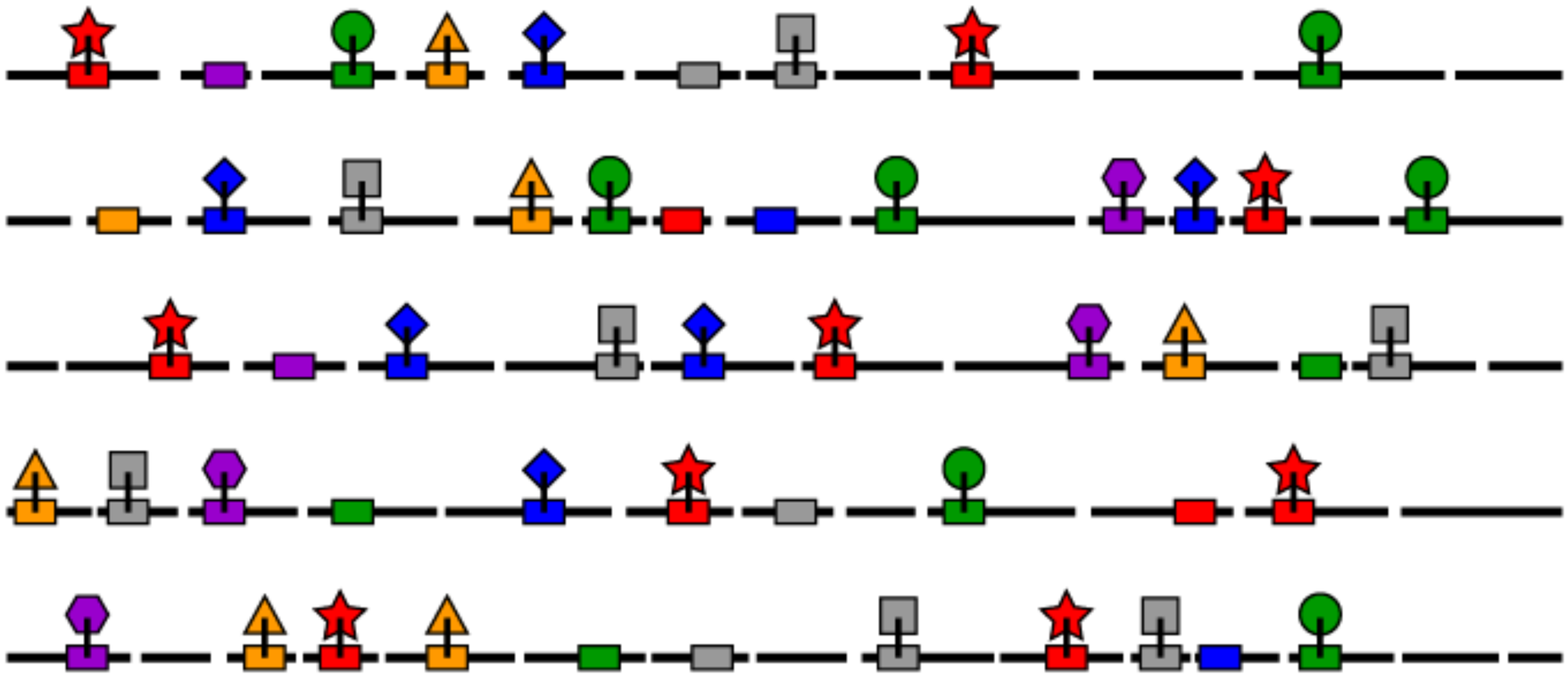


- Need sufficient amount of starting material because the ChIP will enrich for a small proportion
- Ideally the starting material for one ChIP uses 10^7 cells from culture

Crosslink proteins to DNA

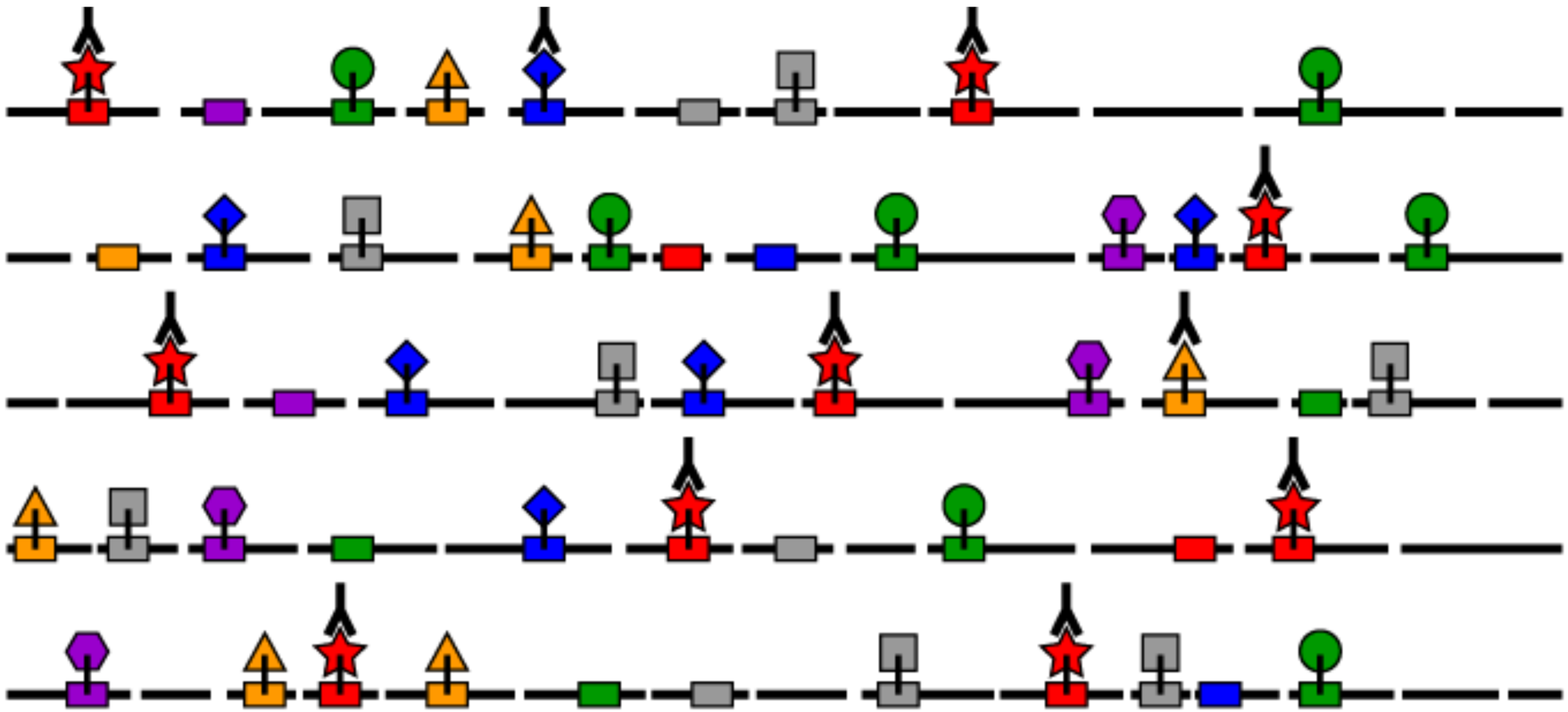


Fragment



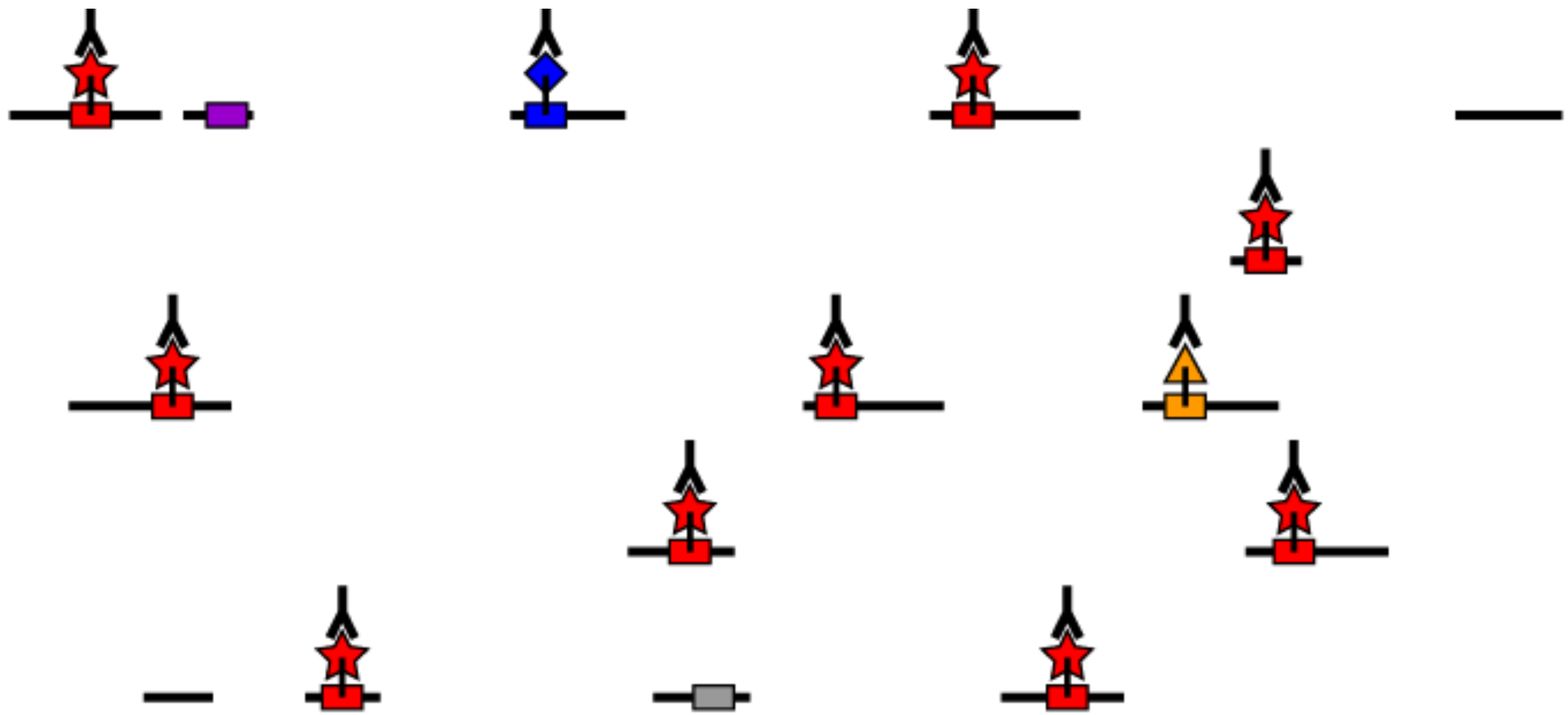
- The DNA is sheared into small fragments - usually 200-500 bp in length
- Check by running on a gel

Protein specific antibody



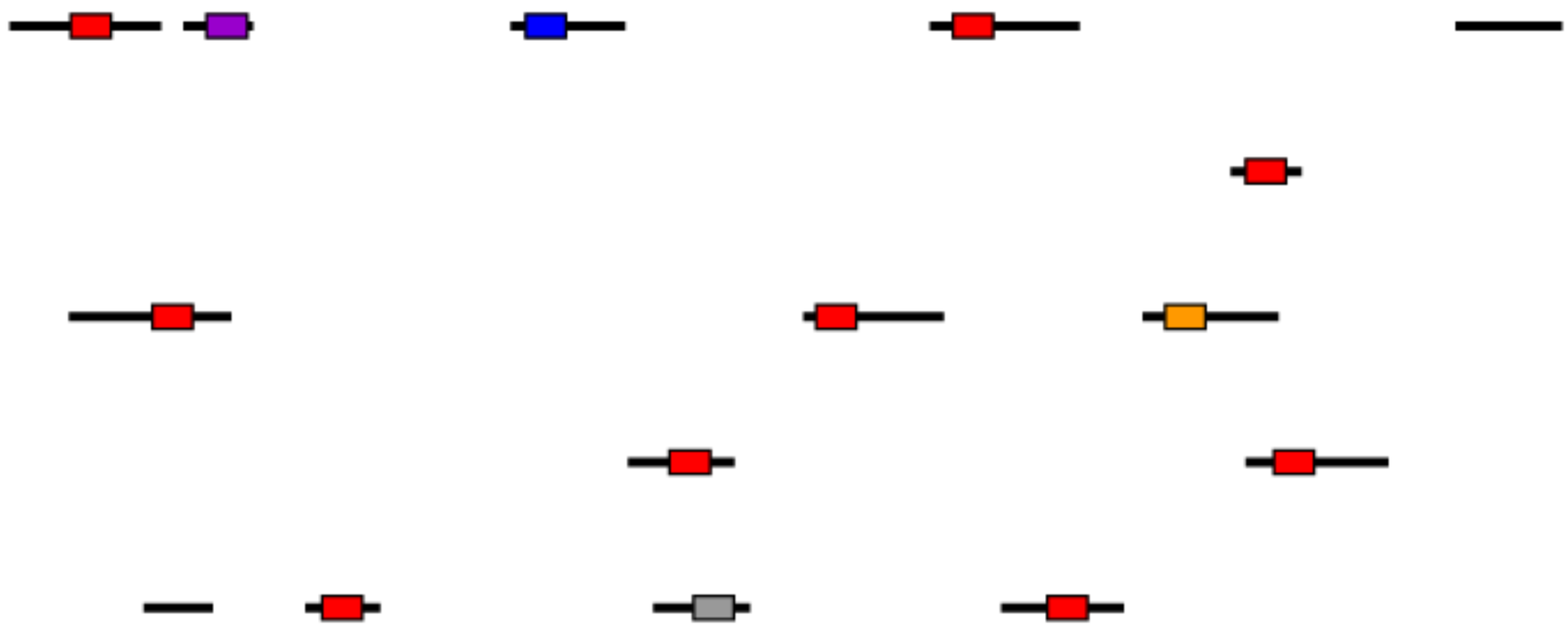
- The sheared protein-bound DNA is immunoprecipitated using a specific antibody

Immunoprecipitate

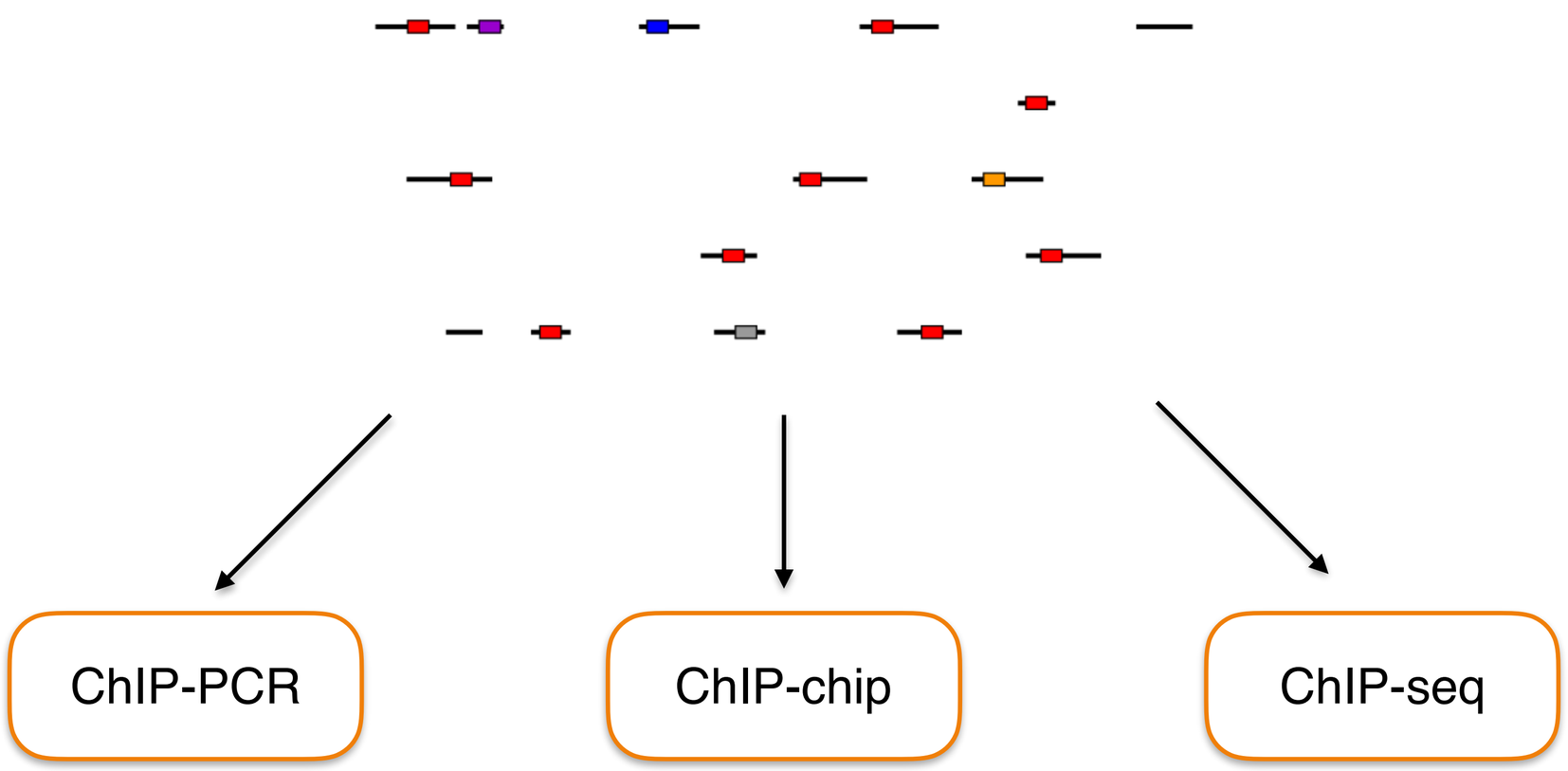


- The antibody binds primarily to the protein of interest but there may be cross reactivity with other proteins with similar epitopes

Reverse crosslink and purify DNA

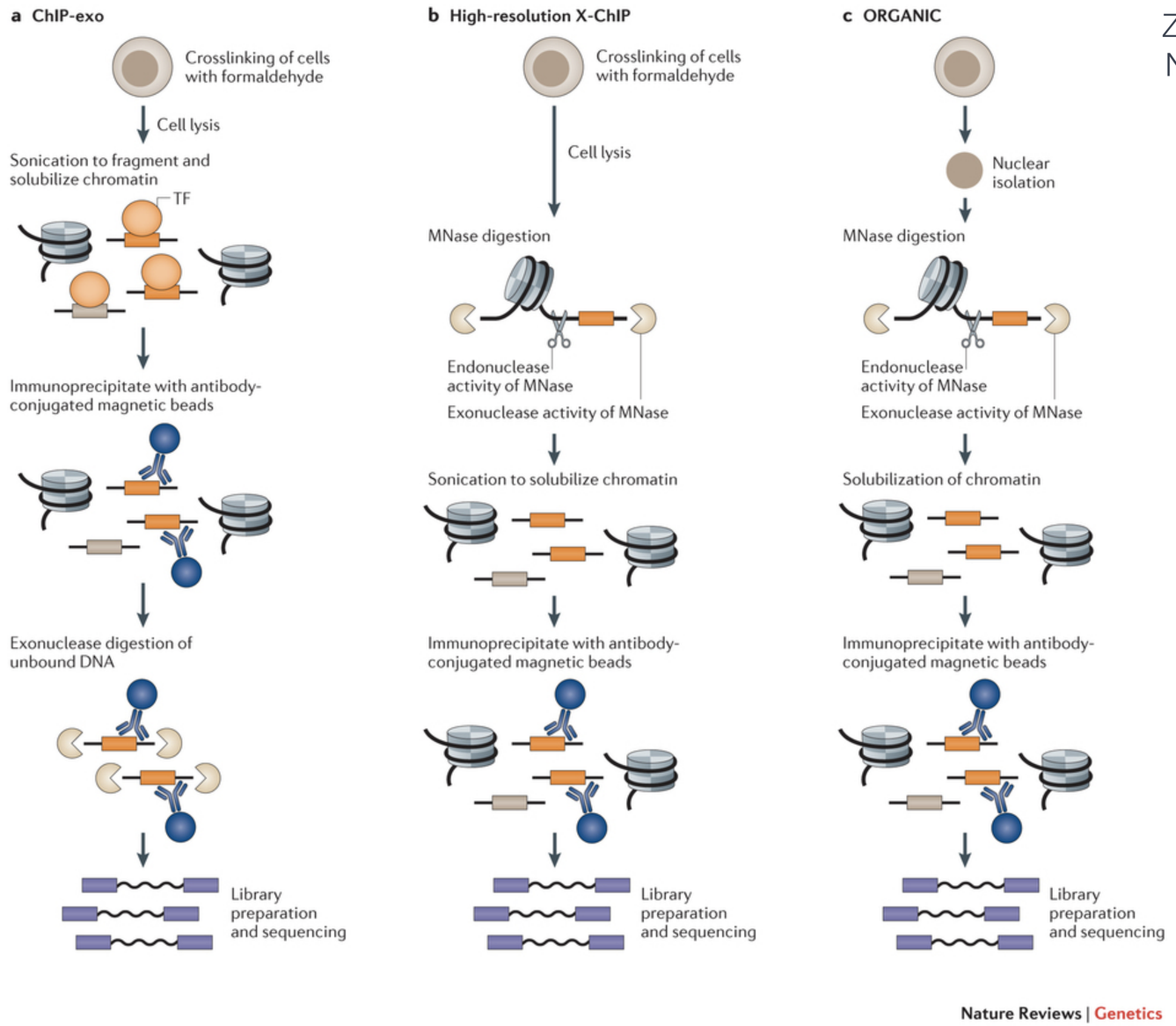


Identify bound regions



~10 ng of ChIP DNA

Zentner & Henikoff (2014).
Nature Reviews Genetics.

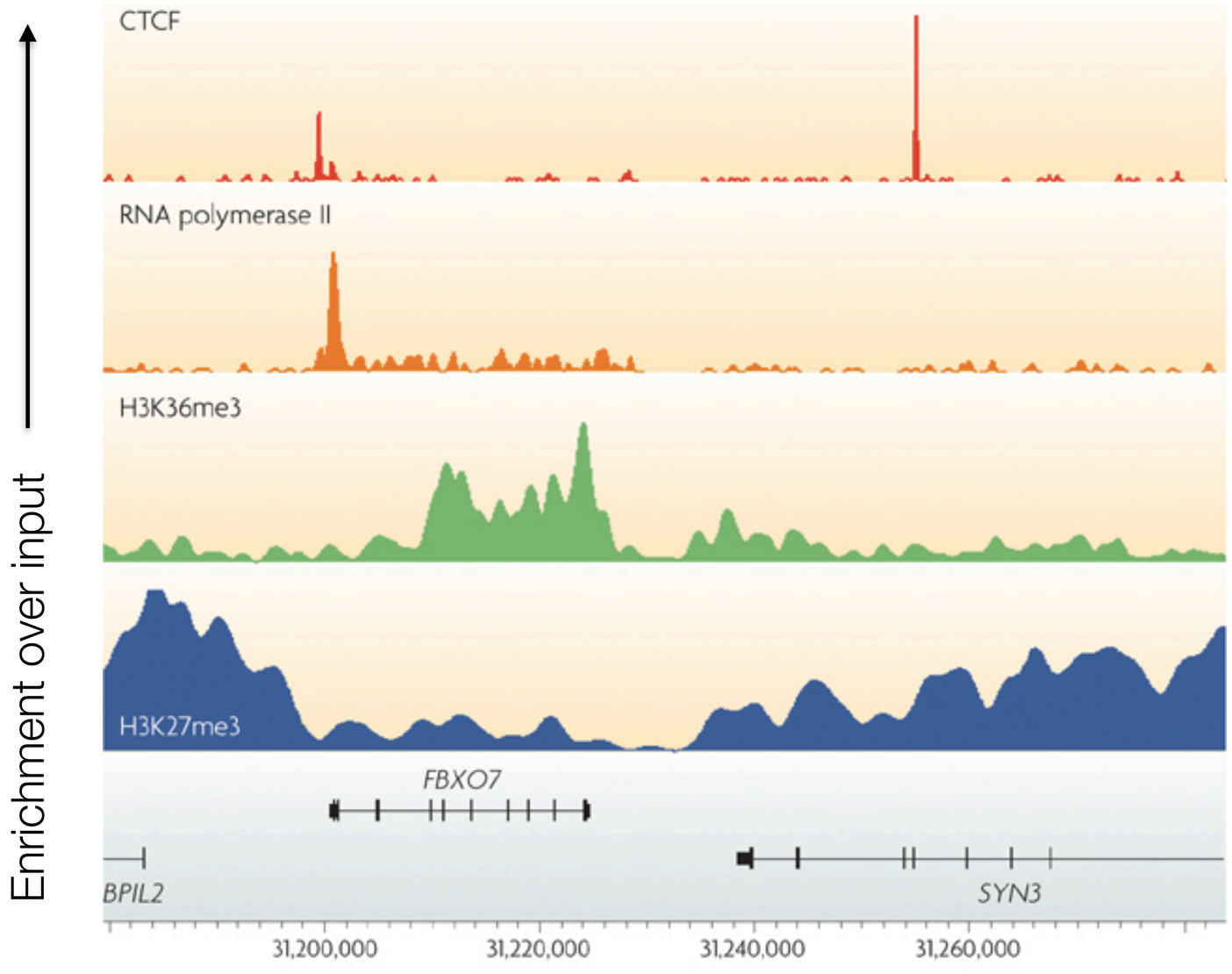


High resolution variations of ChIP-seq

Types of signals

Transcription factors

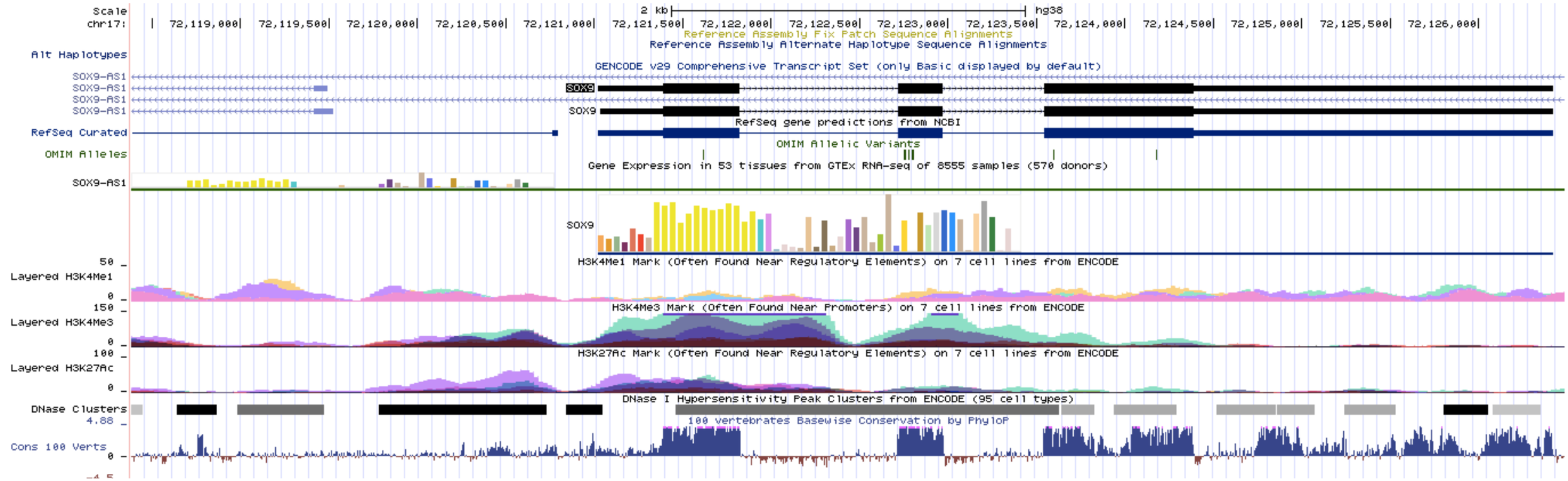
Histone modifications



Adapted from Park (2009). Nature Reviews Genetics.

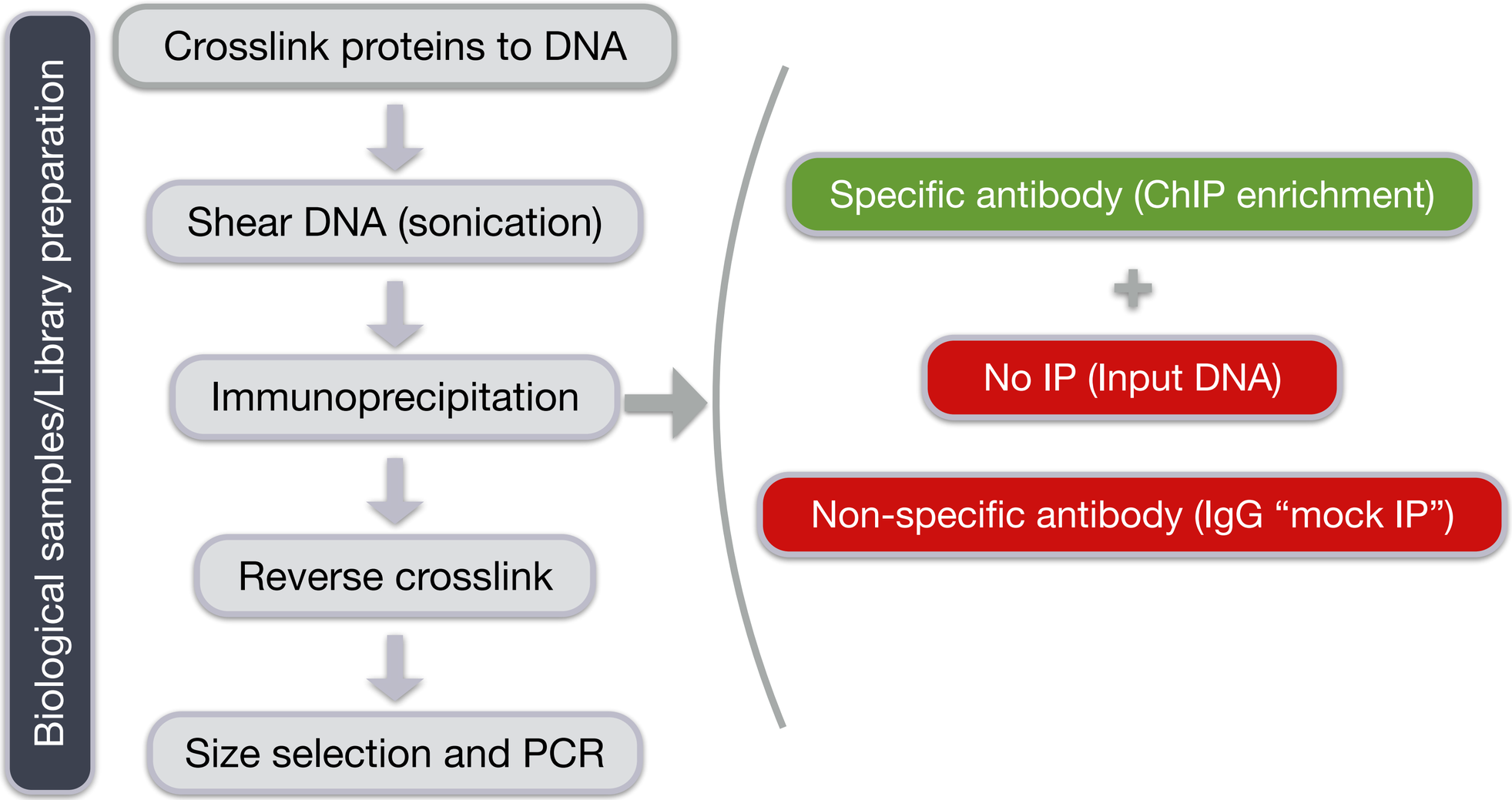
Profiling histone modifications

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

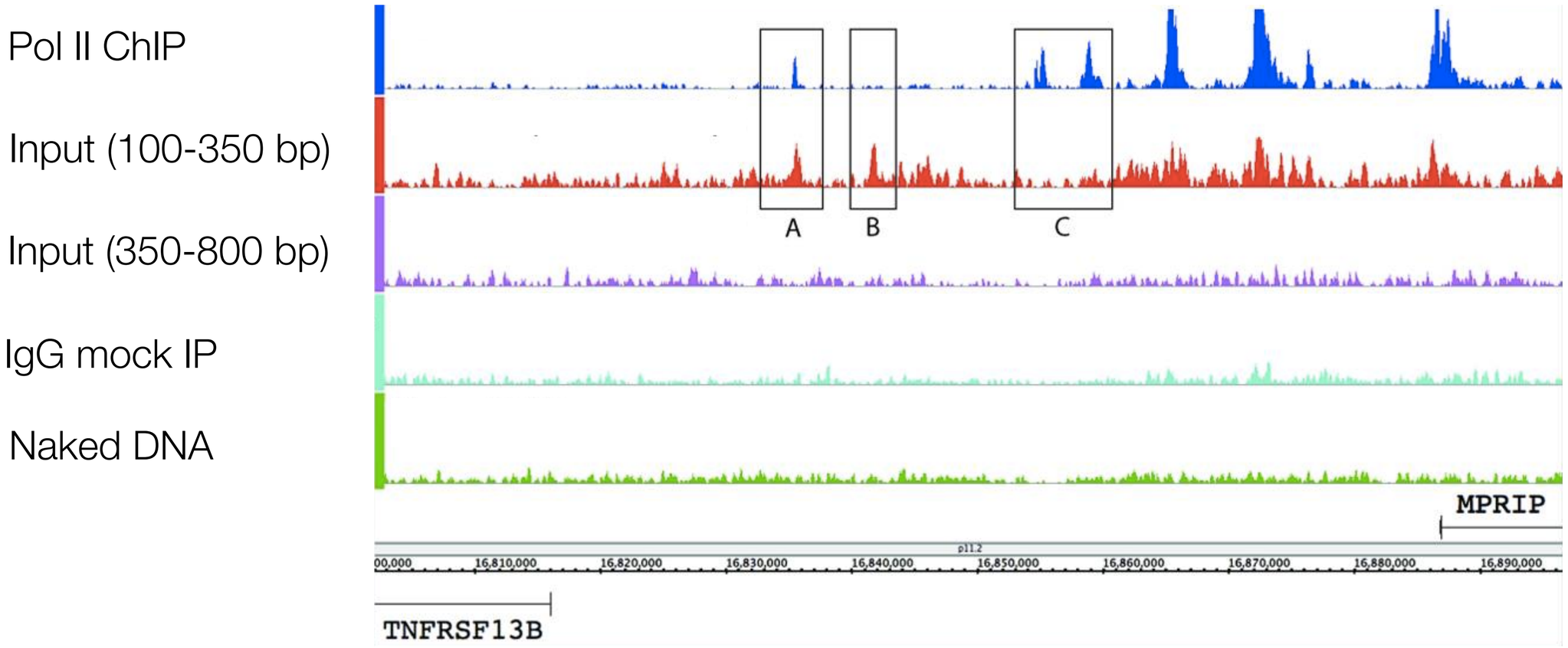


Why are controls necessary?

- 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”



ChIP-Seq Controls



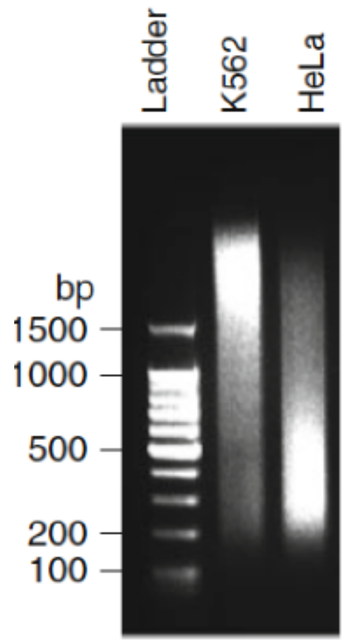
Map of ChIP-seq versus control signals

Parameters for a successful ChIP

- Efficient and specific antibody
- Amount of starting material
  - ChIP DNA yield depends on various factors
    - Cell type in question
    - Abundance of the mark or protein (histones have high binding coverage than TFs)
    - Antibody quality
- “For an IP for histones using 20ug of chromatin DNA from T cells as starting material I have got between 15-50ng DNA in total. For TFs I usually got 5-25ng from 25 million cells (200ug chromatin).” - *Subhash Tripathi,* ResearchGate

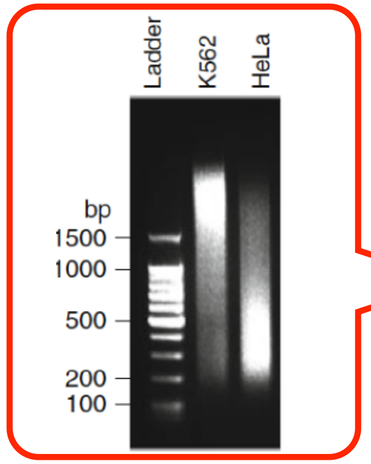
Parameters for a successful ChIP

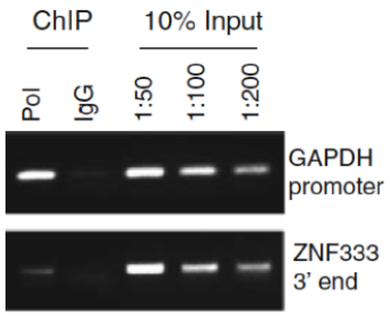
- Chromatin fragmentation
  - Size matters (not too big and not too small)
  - Can vary between cell types
- Stringency of washes



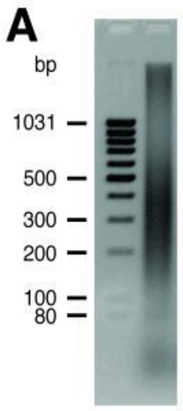
Fragments too big:
Reduced signal to noise ratio in ChIP-seq

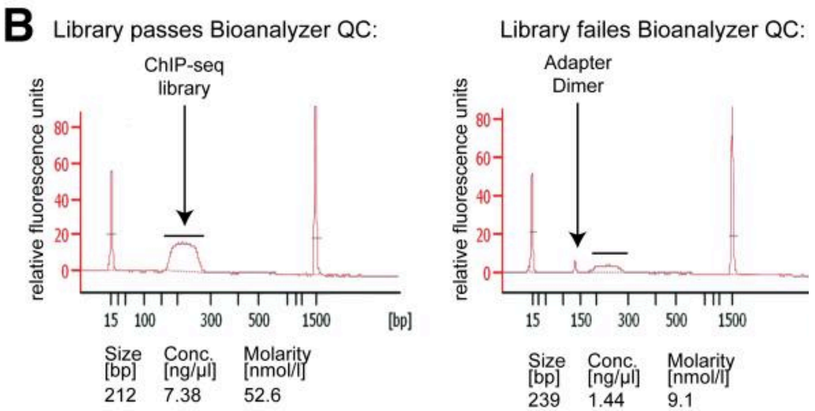
Oversonication:
Fragmentation biased towards promoter regions causes ChIP-seq enrichments at promoters in both, ChIP AND control (input) sample

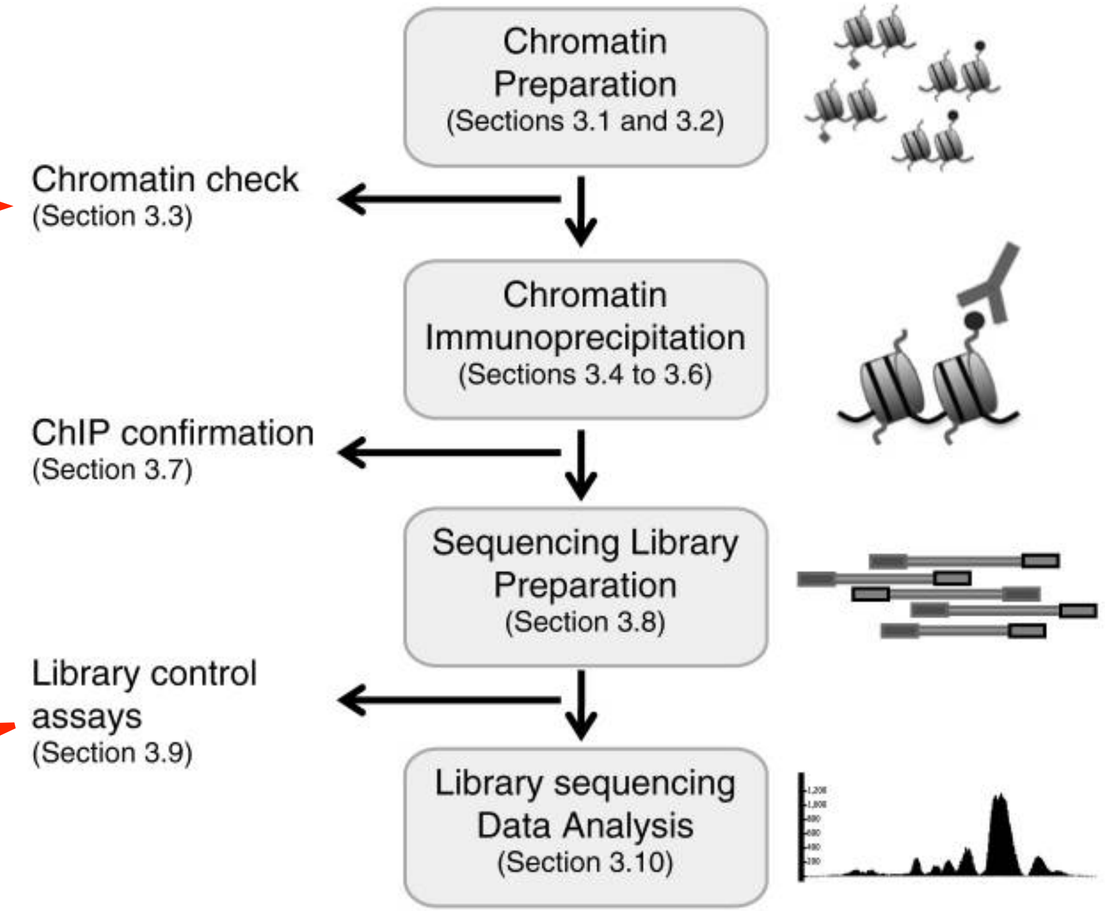




Also quantitate with Qubit (ChIP) / NanoDrop (WCE)

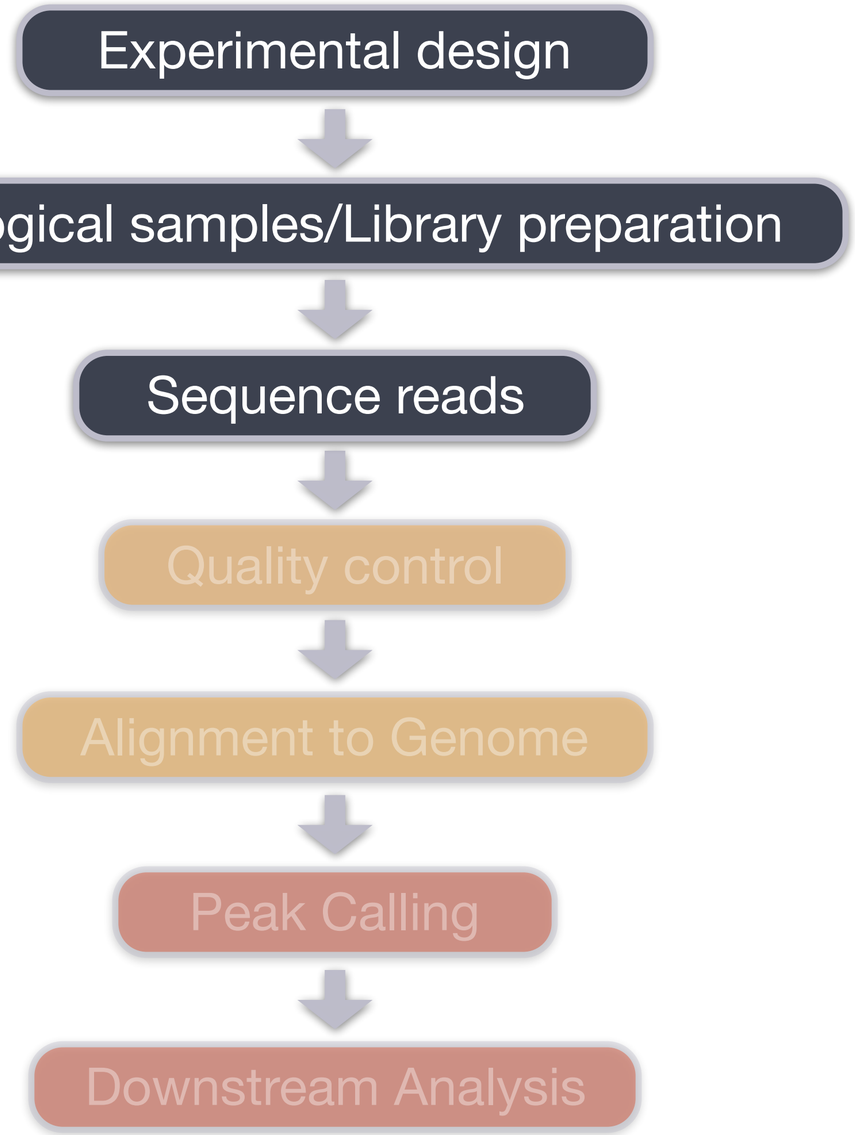






Quality check points

O'Geen et al (2011), Methods Mol Biol: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4151291/
Schmidt et al (2009), Methods;48(3):240-248. doi:10.1016/j.ymeth.2009.03.001.



ChIP-seq workflow

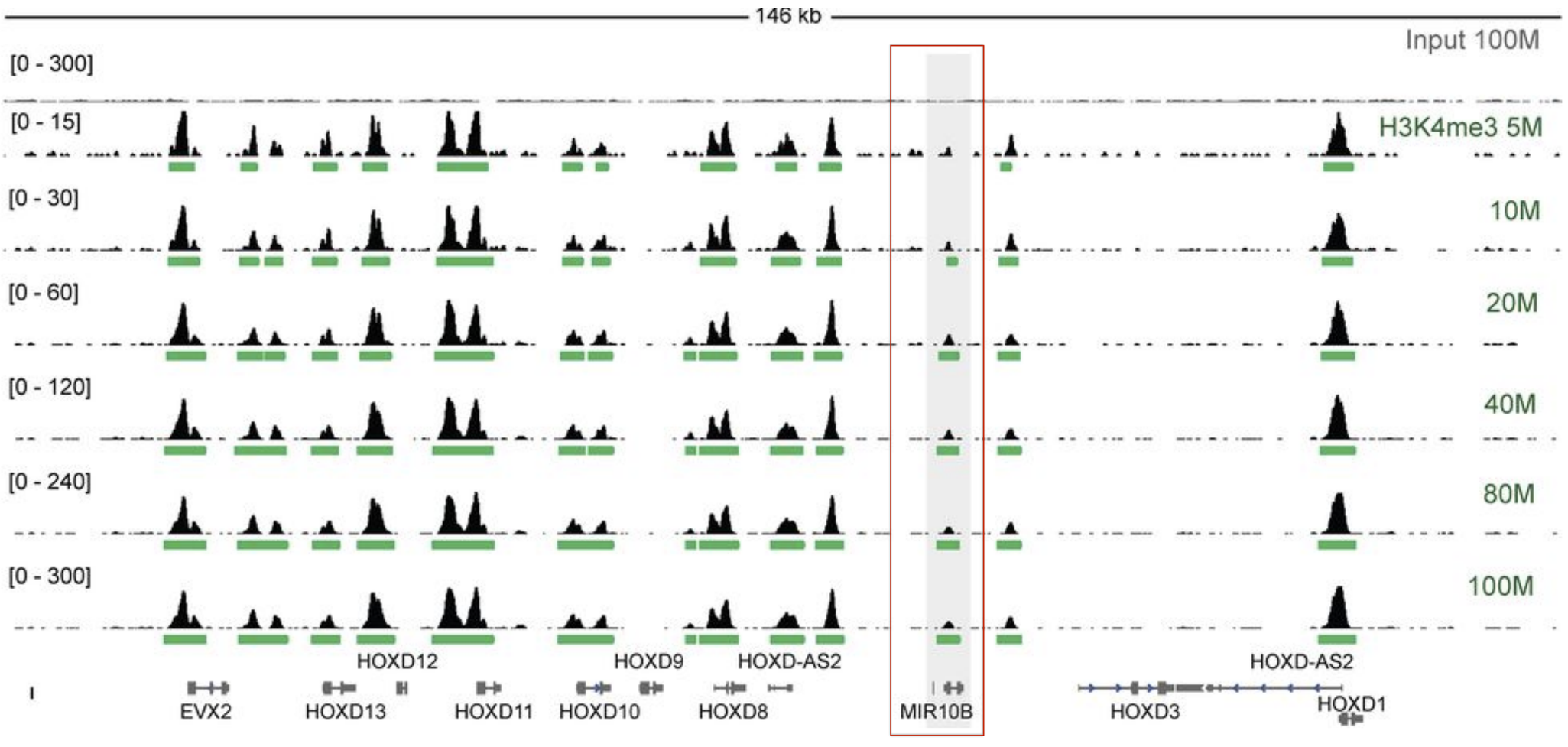
Sequencing considerations

- 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
- Avoid batches or distribute 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

Meyer & Liu, Nature Reviews Genetics, 2014

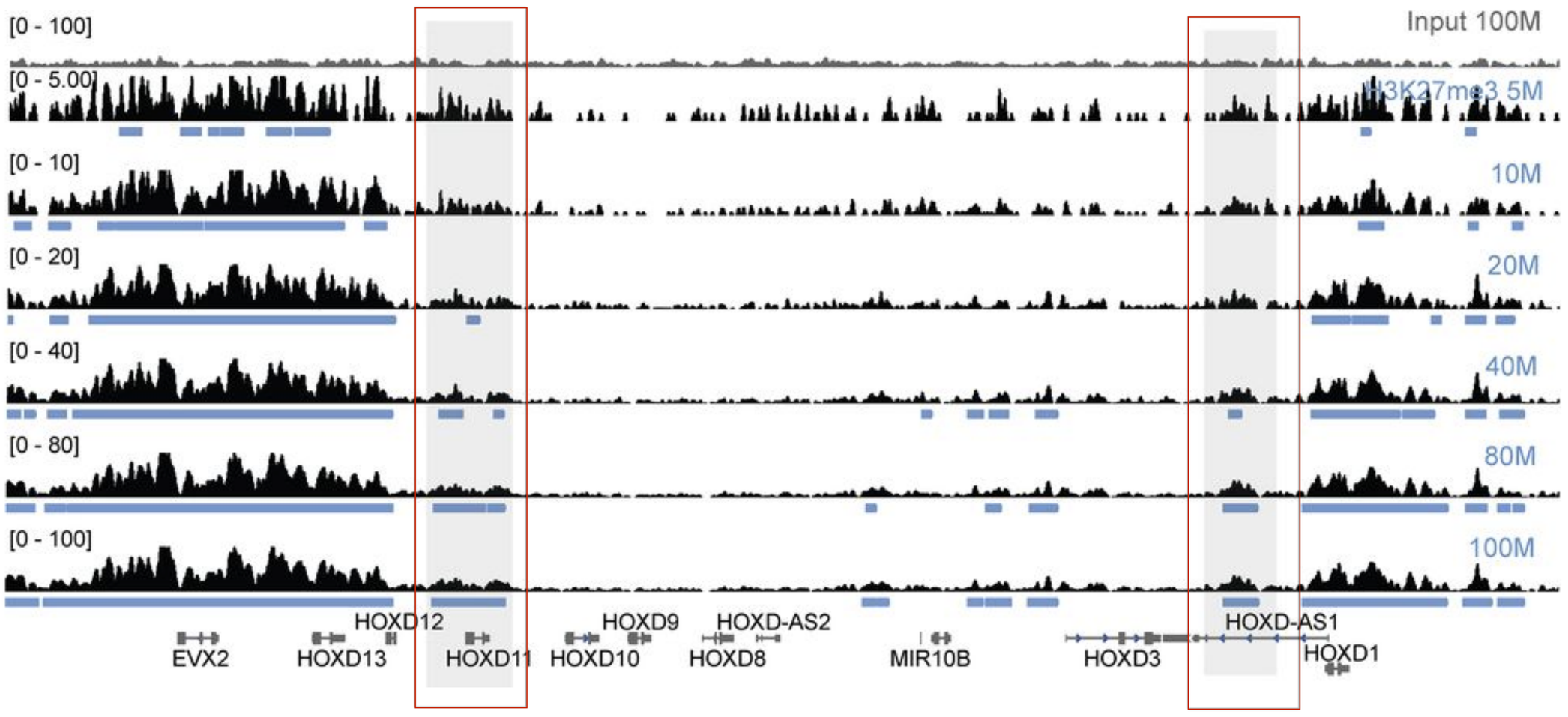
Impact of sequencing depth

H3K4me3



Adapted from Jung et al (2014). NAR.

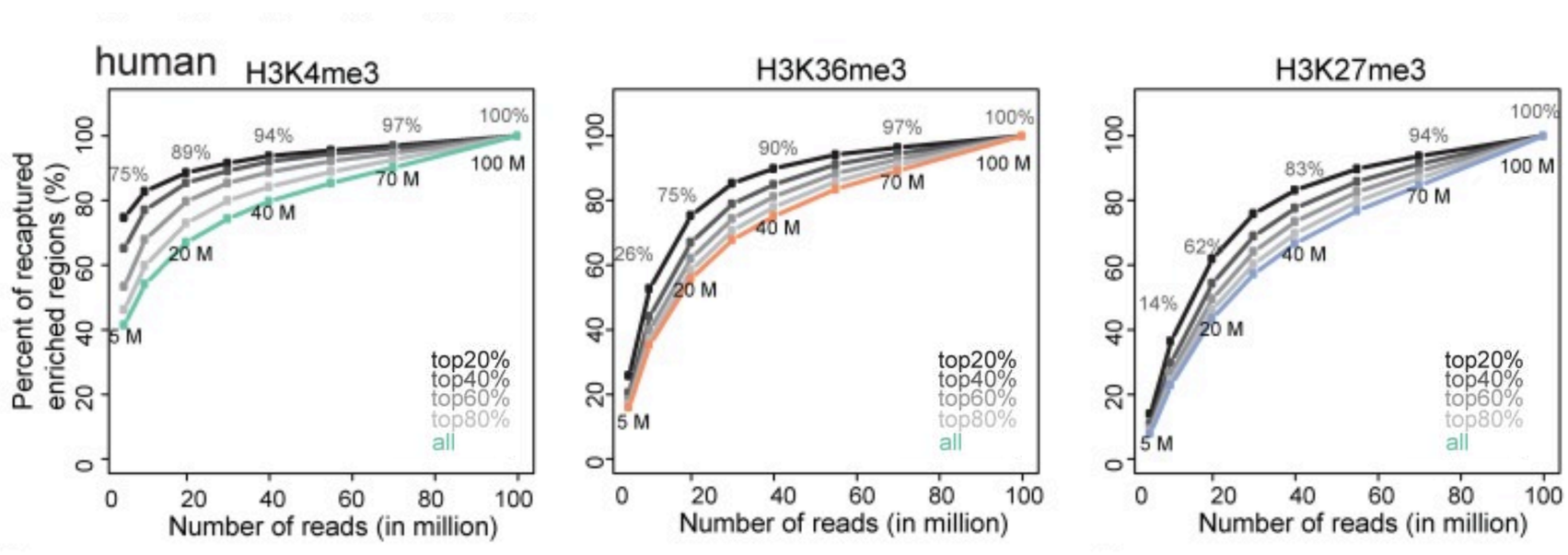
Impact of sequencing depth



Adapted from Jung et al (2014). NAR.

Impact of sequencing depth

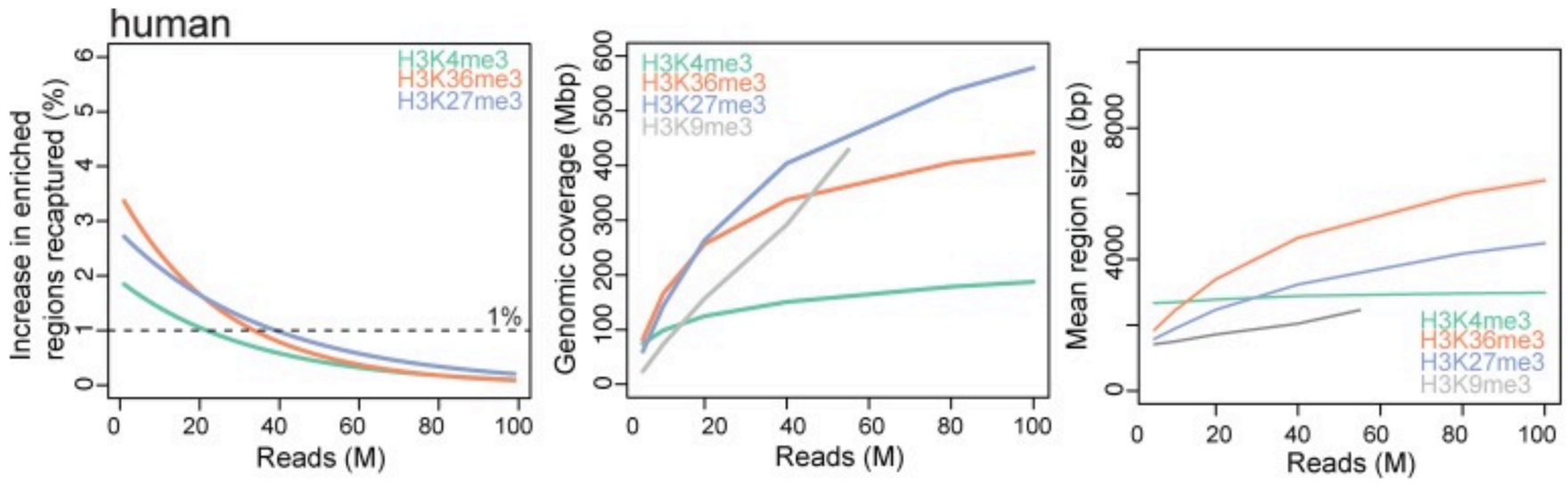
Percentage of significantly enriched regions from the full data recovered in each subsample for H3K4me3, H3K36me3 and H3K27me3



Adapted from Jung et al (2014). NAR.

Impact of sequencing depth

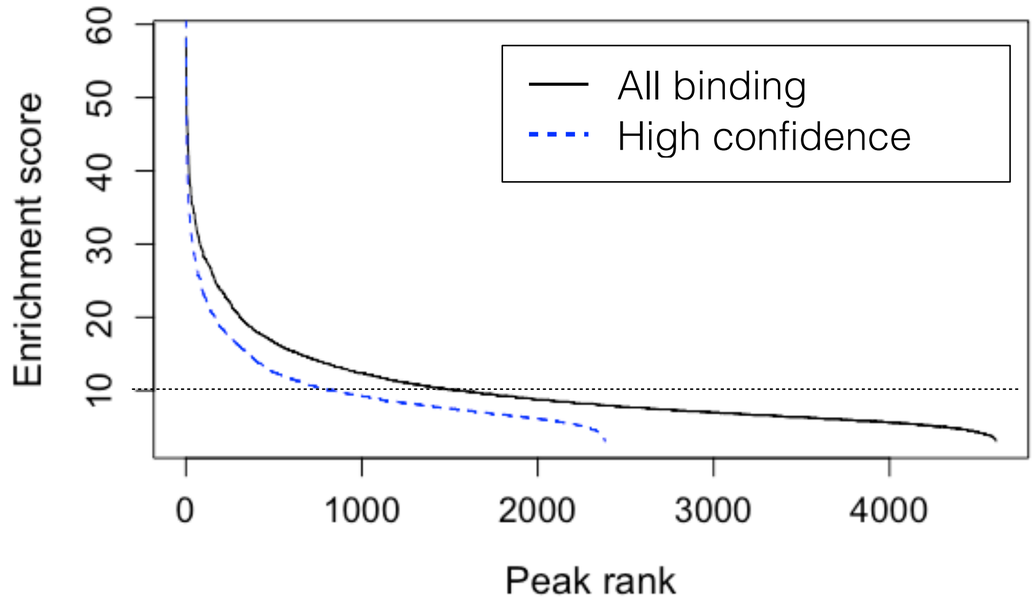
Percentage of increase in enriched regions recaptured when an additional 1 million reads were sequenced

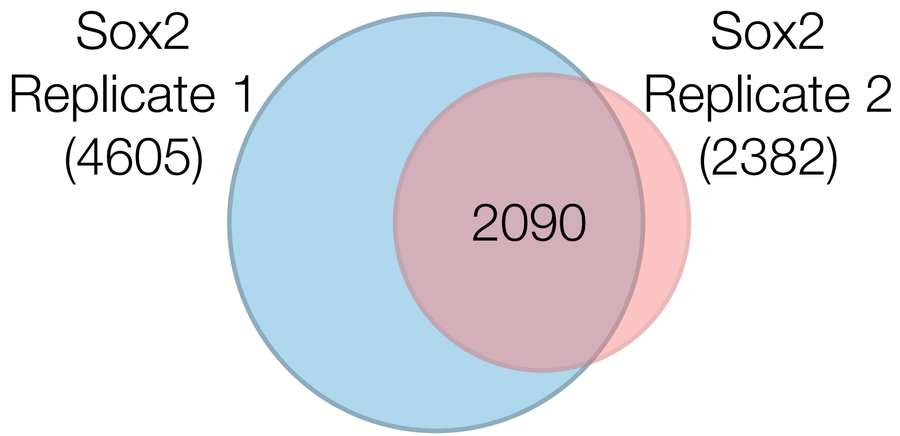


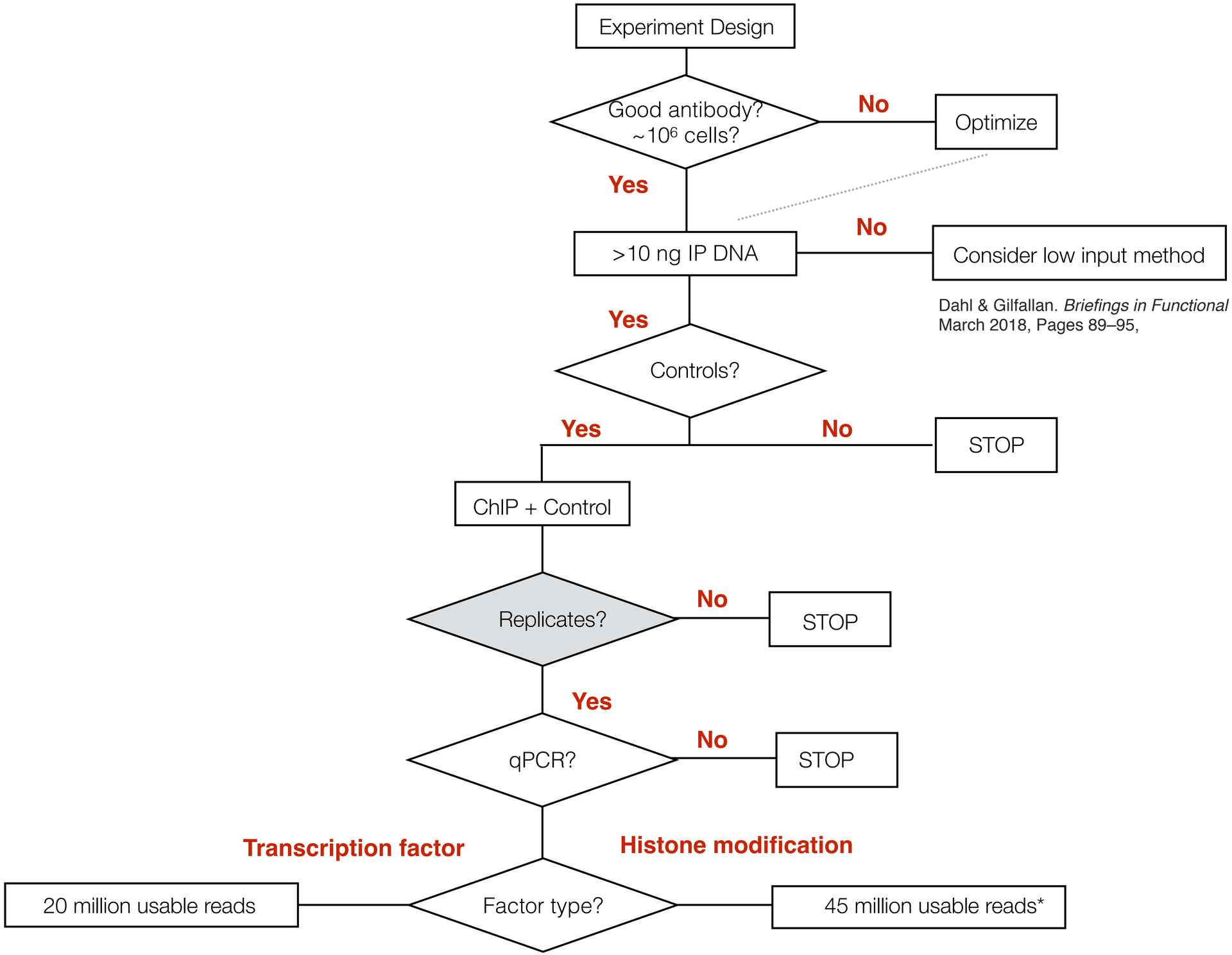
Adapted from Jung et al (2014). NAR.

Replicates and reproducibility

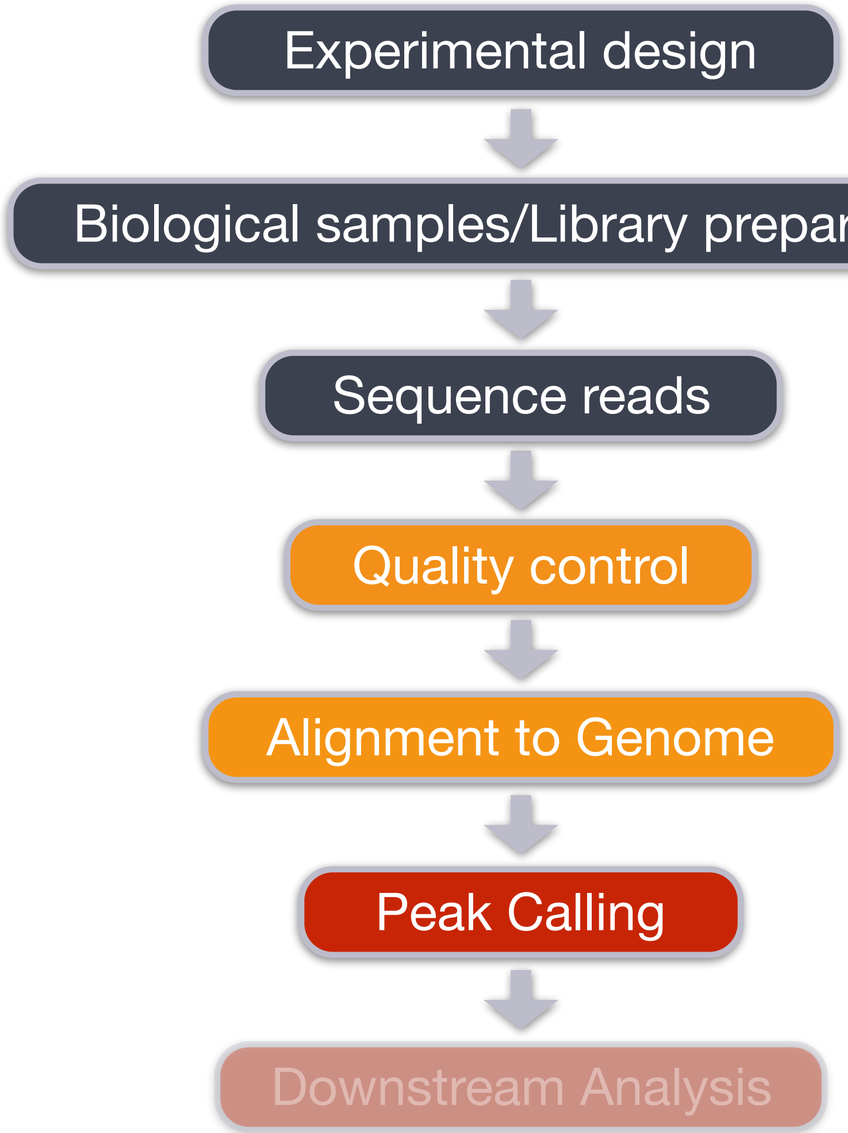
- Biological replicates are essential to understand variation and for differential binding analysis
- More replicates is often preferable to greater depth
- Better to sequence high-quality sample at lower depth than low-quality sample to higher depth







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



FASTQC

BWA, Bowtie2

Filter duplicates, multi mappers, blacklist

ChIP-seq workflow

Quality check and filtering

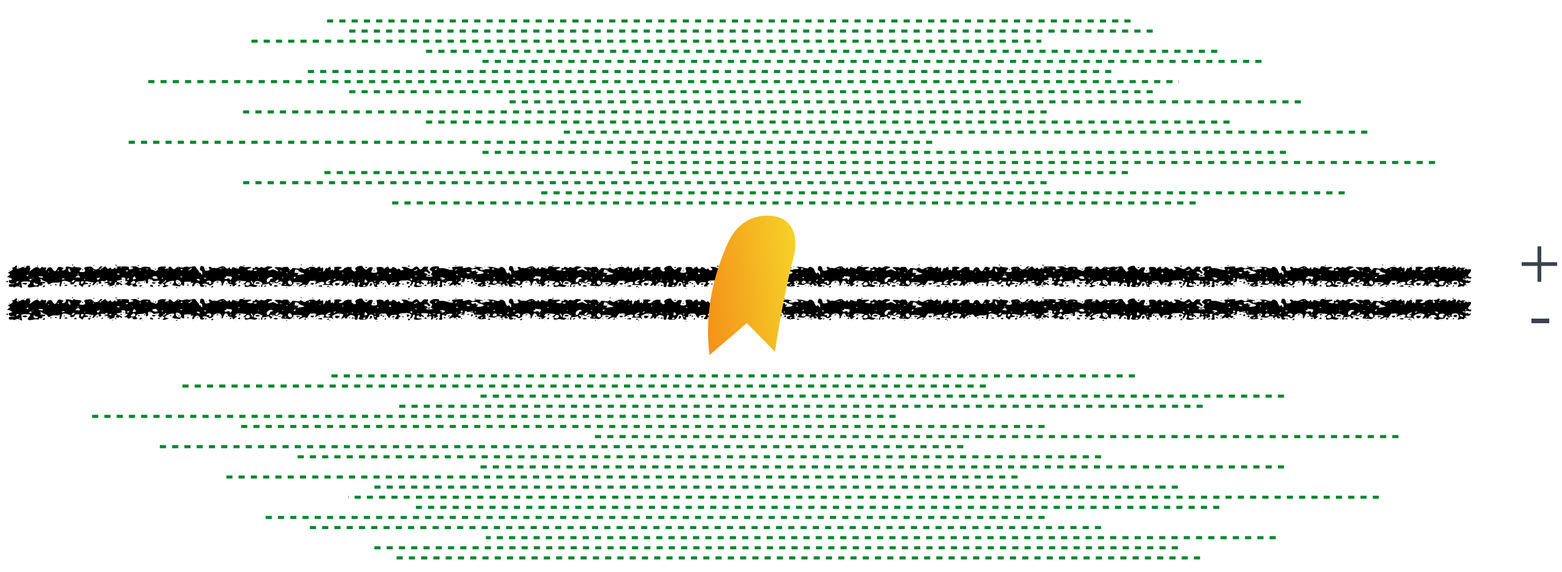
- Raw sequence QC is similar to RNA-seq
- However,
  - Explore duplication rates and possibly remove duplicates
  - Remove blacklisted regions
  - Assess cross correlation scores and Fraction of Reads in Peaks (FRiP)

Software: ChIPQC, Homer, ChiLin, DiffBind

Understanding strand cross-correlation

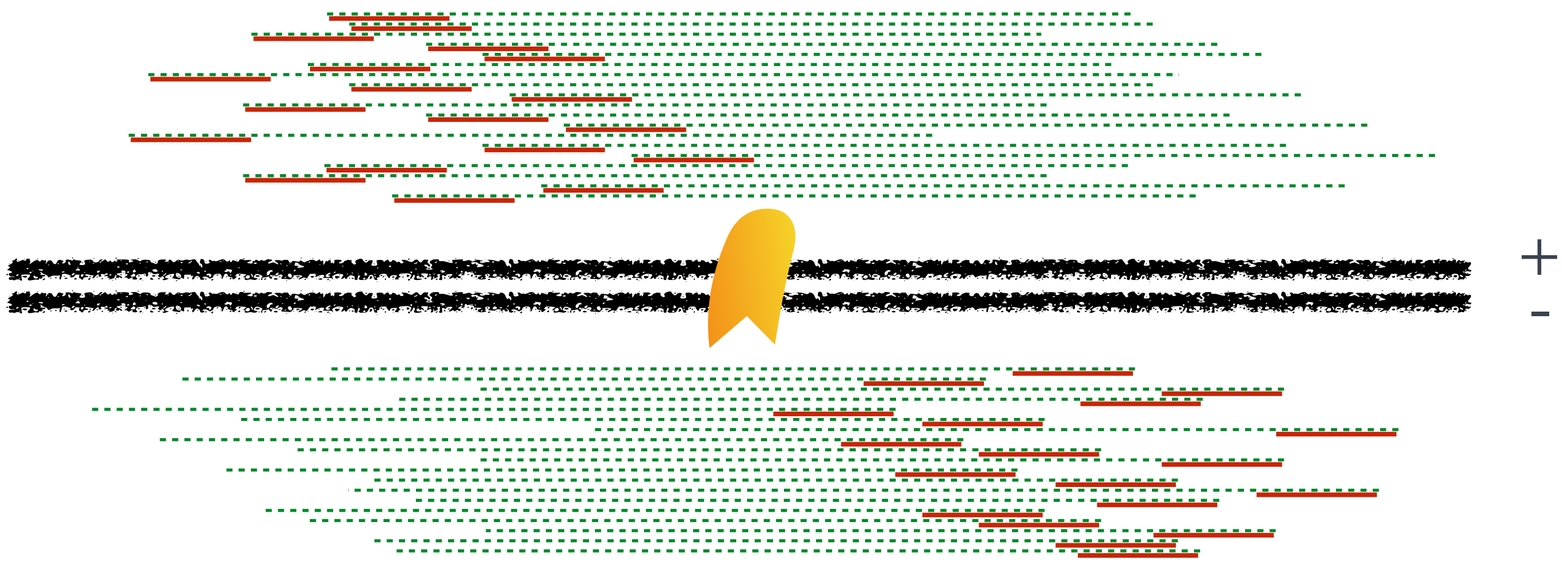
= binding site

---- = size selected DNA fragment



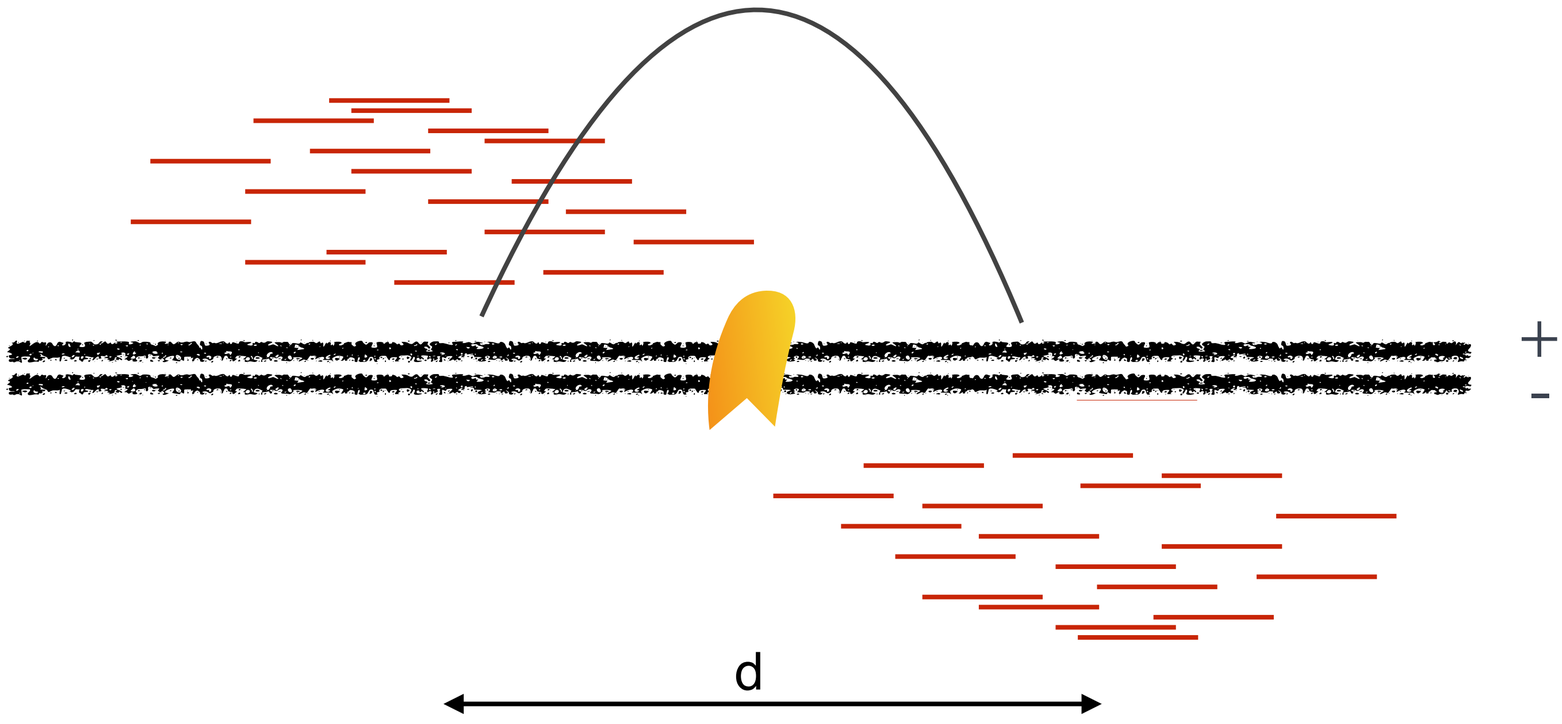
Understanding strand cross-correlation

ChIP-seq fragments are sequenced from the 5’ end



Understanding strand cross-correlation

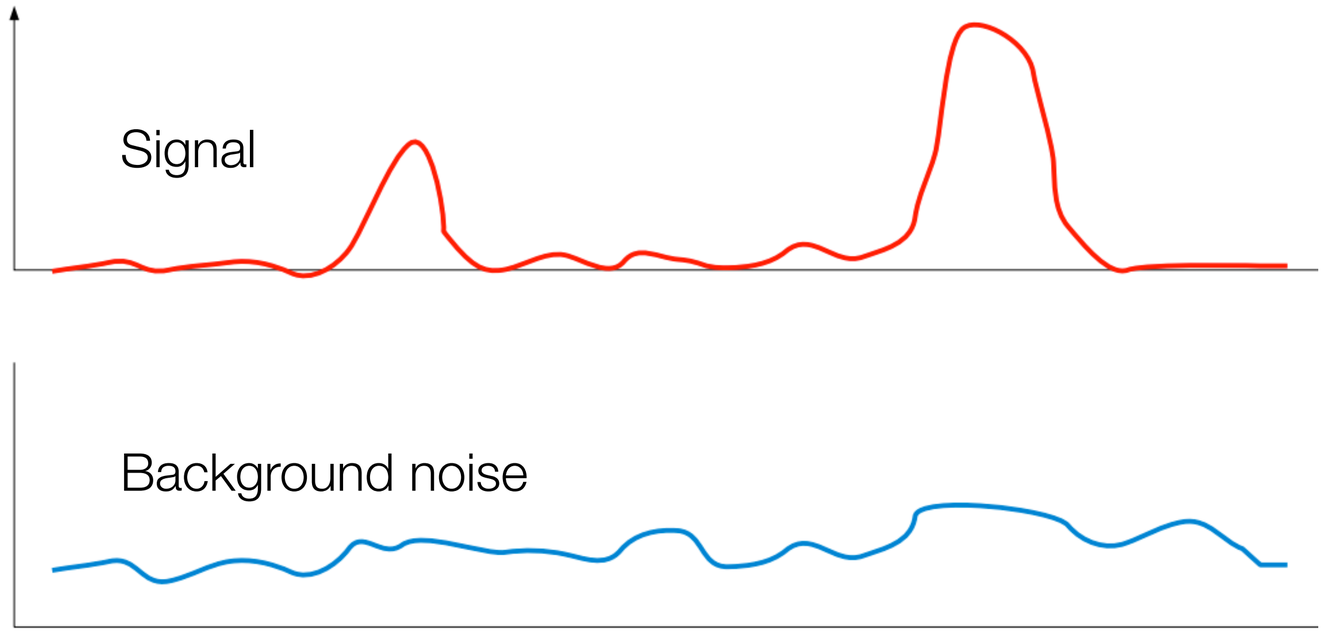
Alignment generates a **bimodal pattern** on the plus and minus strands around binding sites



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)



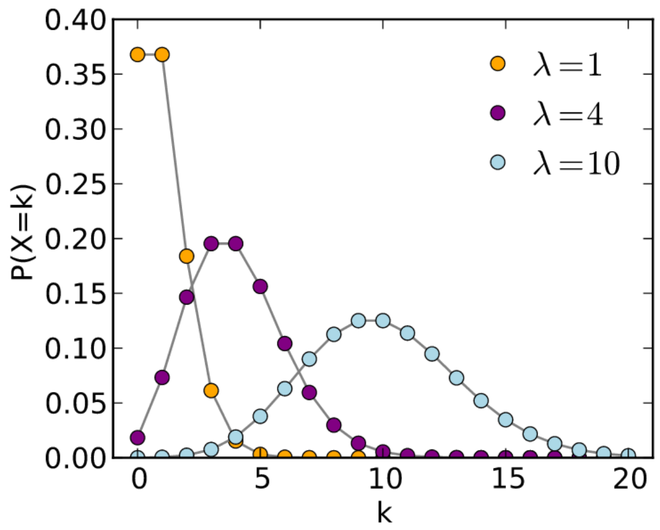
Peak detection

- 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

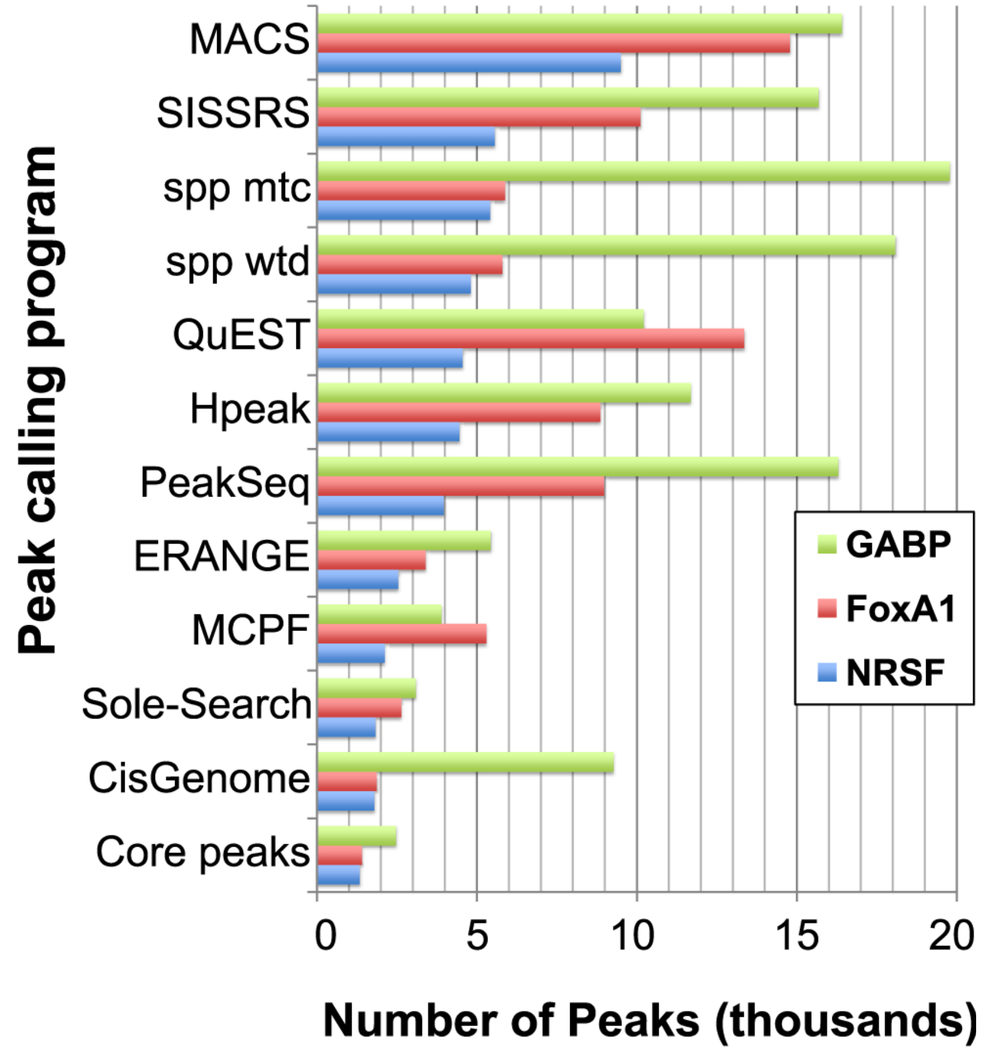
- λ is the average number of events per interval
- e is the number 2.71828... (Euler's number) the base of the natural logarithms
- k takes values 0, 1, 2, …
- $k! = k \times (k-1) \times (k-2) \times \ldots \times 2 \times 1$ is the factorial of k.



http://en.wikipedia.org/wiki/Poisson_distribution

Peak callers

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



Wllbanks & Facciotti (2010). PLoS ONE.

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)

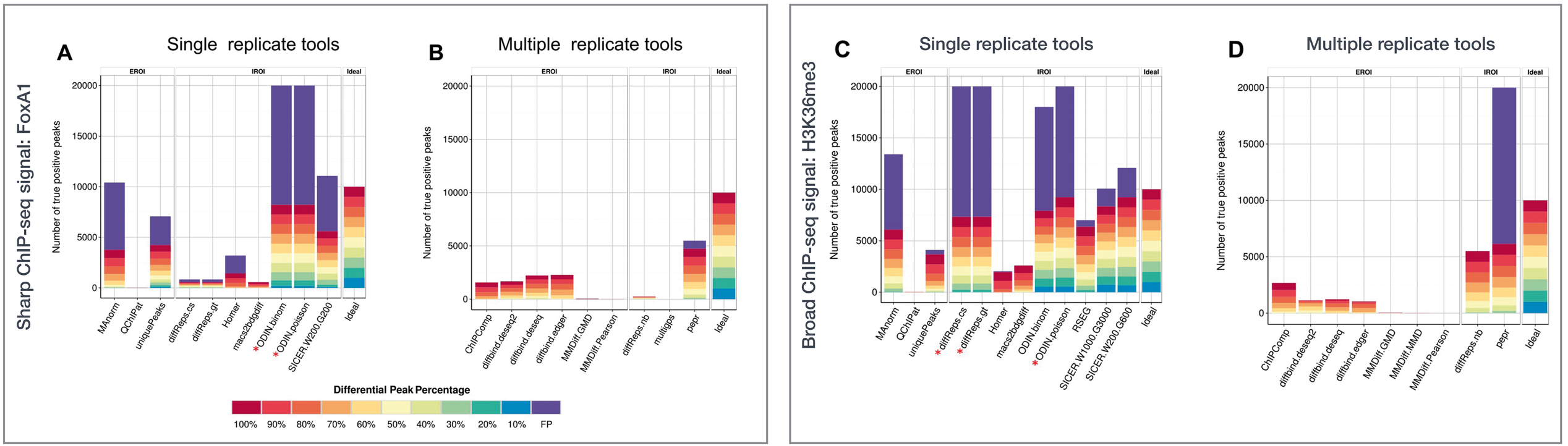


Figure 4. Proportion of true and false positives for each tool on the simulated FoxA1 data set (A, B) and H3K36me3 data (C, D)

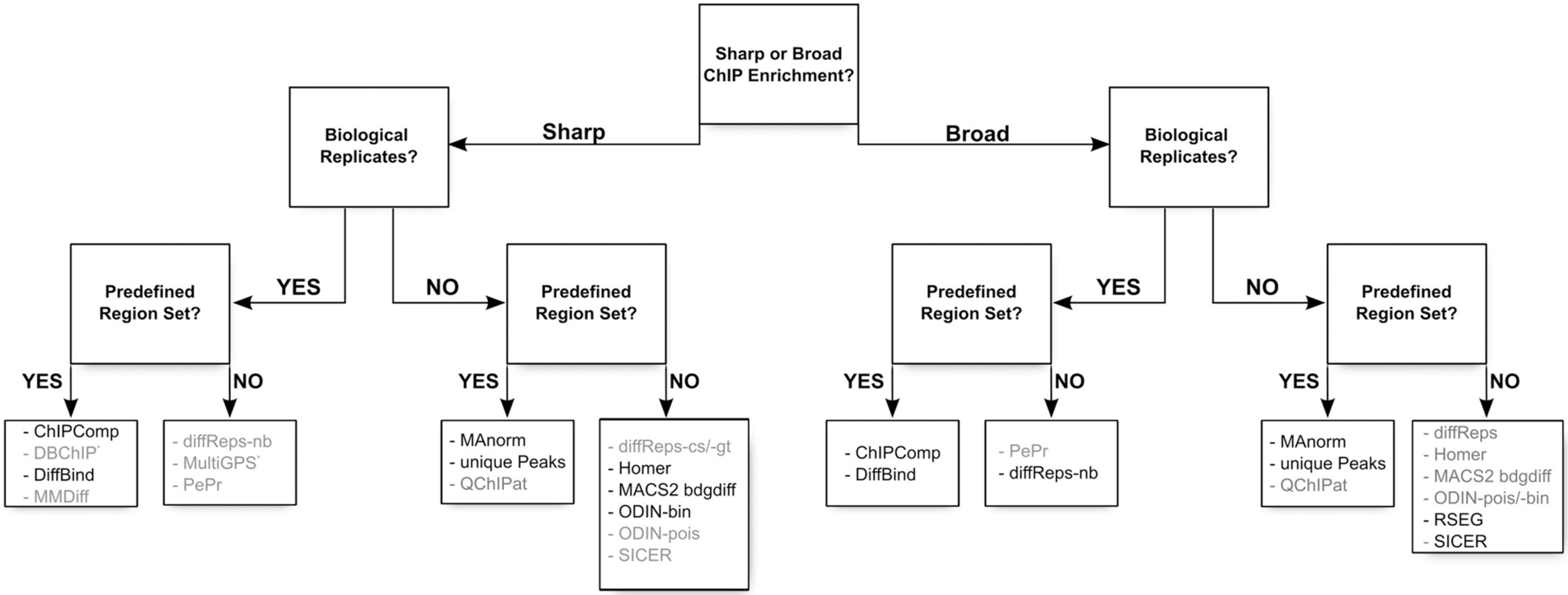
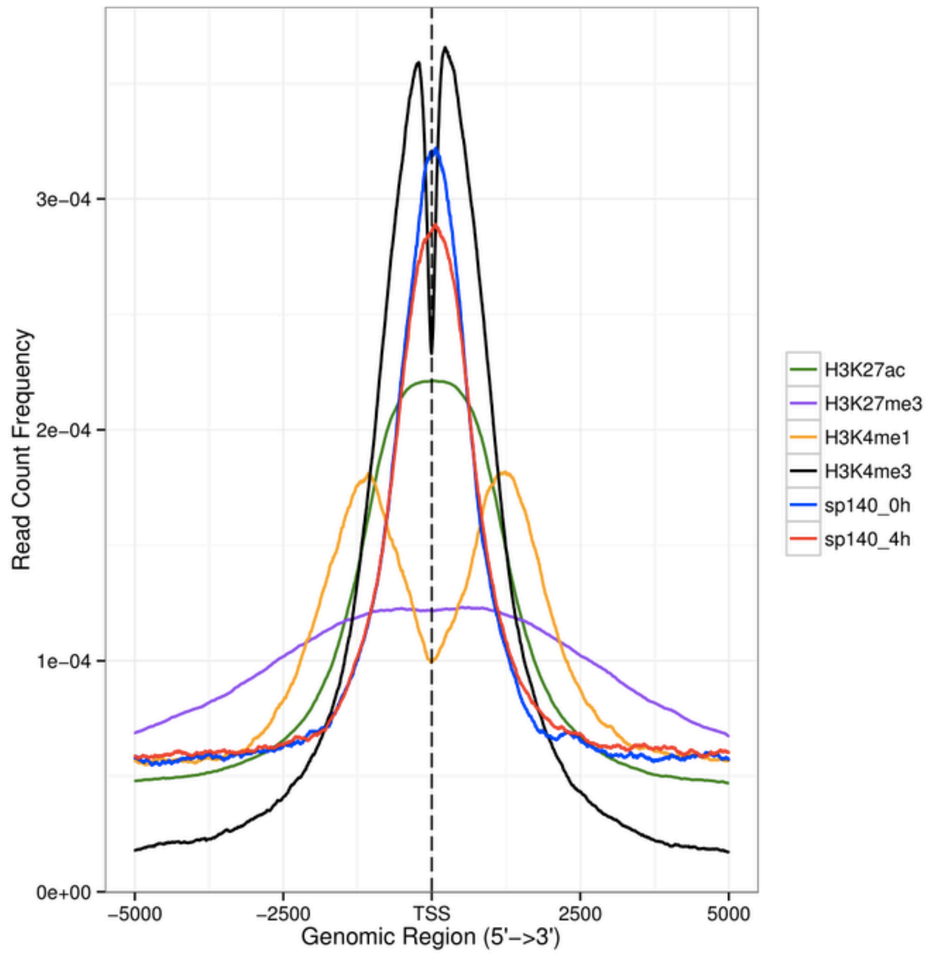


Figure 7. Decision tree indicating the proper choice of tool depending on the data set: shape of the signal (sharp peaks or broad enrichments), presence of replicates and presence of an external set of regions of interest [*Steinhauser, et al, 2016*].

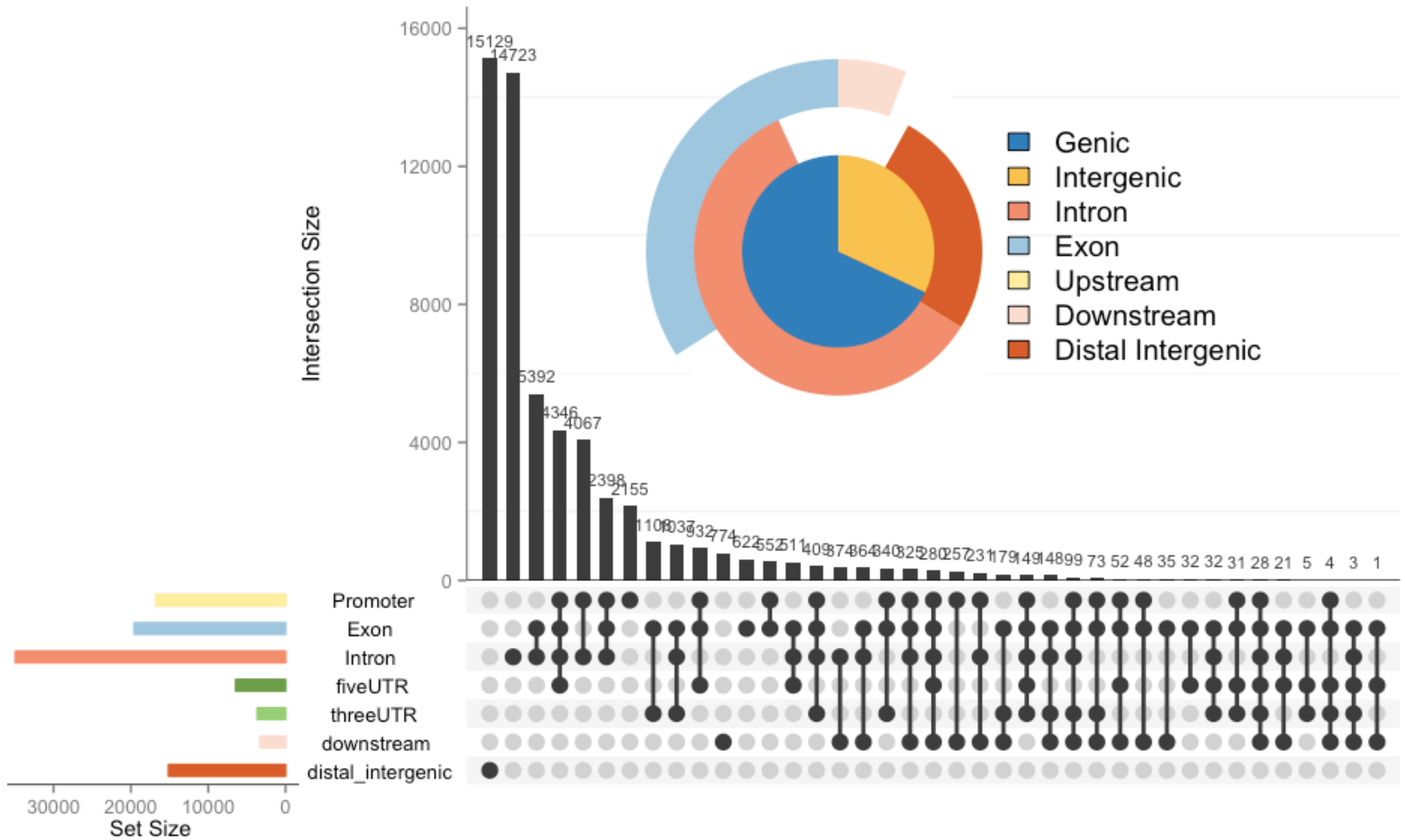
Downstream analysis

- Annotation of peaks - distance from TSS
  - ChIPseeker, Homer, ChiLin



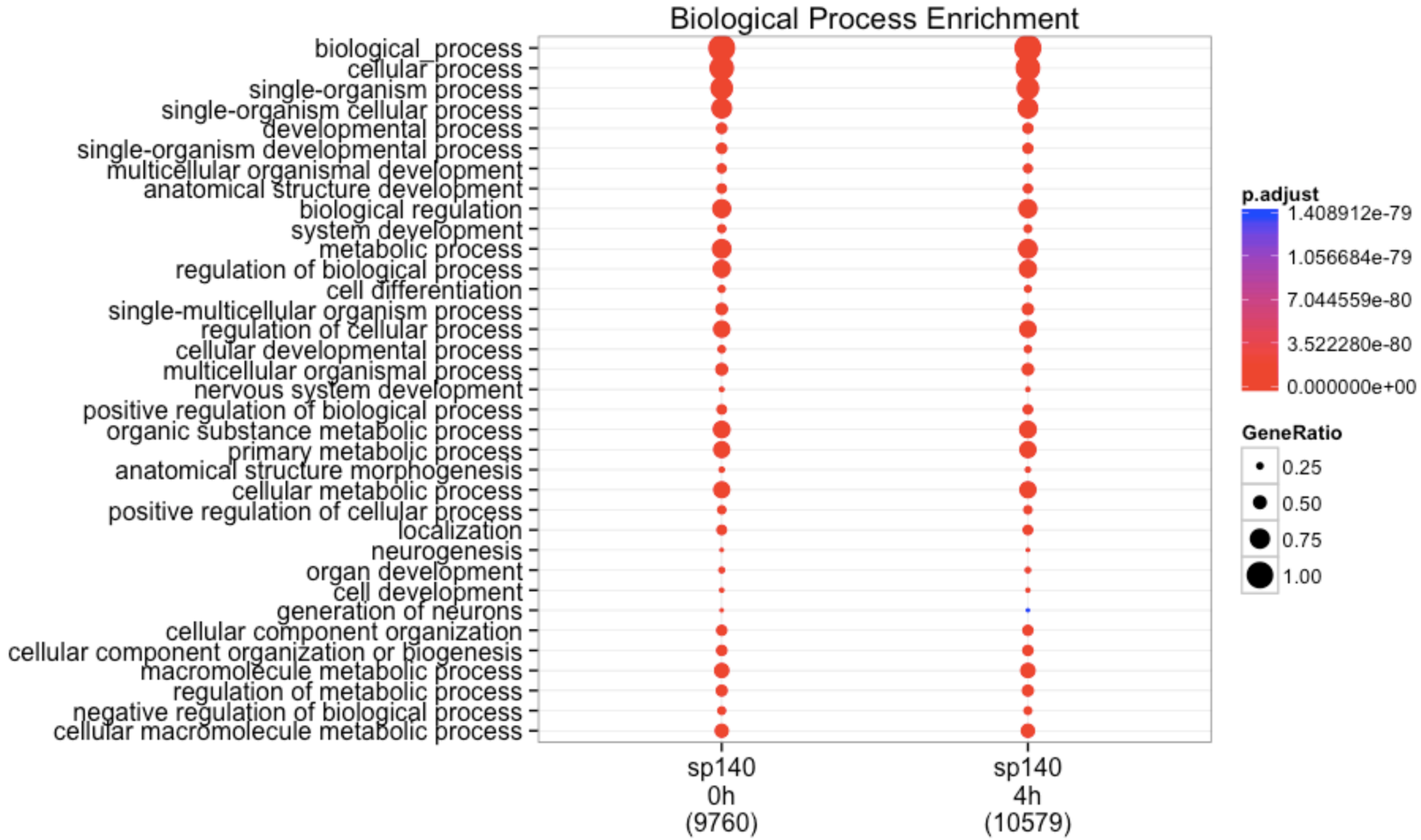
Downstream analysis

- 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]

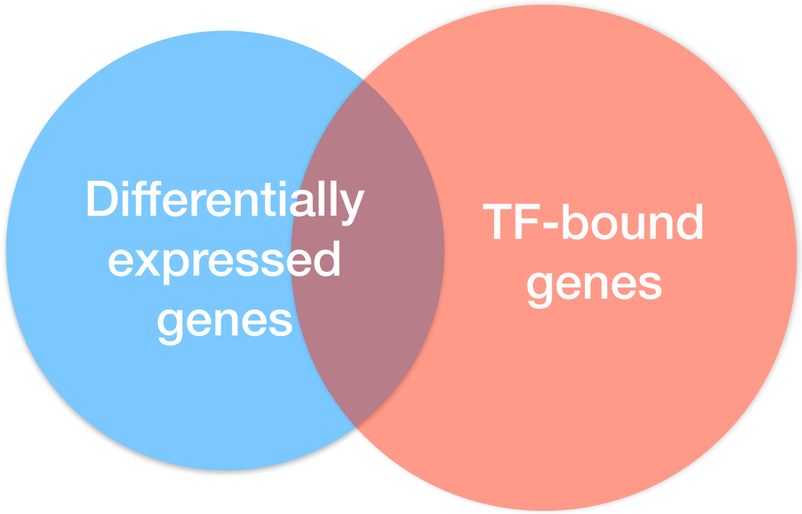
DISCOVERED MOTIFS | INPUTS & SETTINGS | PROGRAM INFORMATION

DISCOVERED MOTIFS

	Motif	Logo	RC Logo	E-value	Unerased E-value	More	Submit/Download
1.	CYWTTGTB			4.2e-299	4.2e-299		
2.	ATGBWAAT			8.4e-179	1.1e-179		
3.	CCMCDCCC			1.3e-130	1.1e-131		

Downstream analysis

- Integrative analysis of RNA-seq and ChIP-seq
  - Which of the regulated genes are direct targets of the TF?
  - Is the TF an activator, repressor, or both?
  - Does the TF have different binding partners depending on the direction of regulation?



Some notes on ATAC-seq

- Main advantage over existing methods is the simplicity of the library preparation protocol: Tn5 insertion followed by two rounds of PCR.
  - requires no sonication or phenol-chloroform extraction like FAIRE-seq
  - no antibodies like ChIP-seq
  - no sensitive enzymatic digestion like MNase-seq or DNase-seq
- Unlike similar methods, which can take up to four days to complete, ATAC-seq preparation can be completed in under three hours.
- Lower starting cell number than other open chromatin assays (500 to 50K cells recommended for human).

Slide by Meeta Mistry

What does it give us?

- Multiple aspects of chromatin architecture simultaneously at high resolution.
  - Maps open chromatin
  - TF occupancy
  - nucleosome occupancy

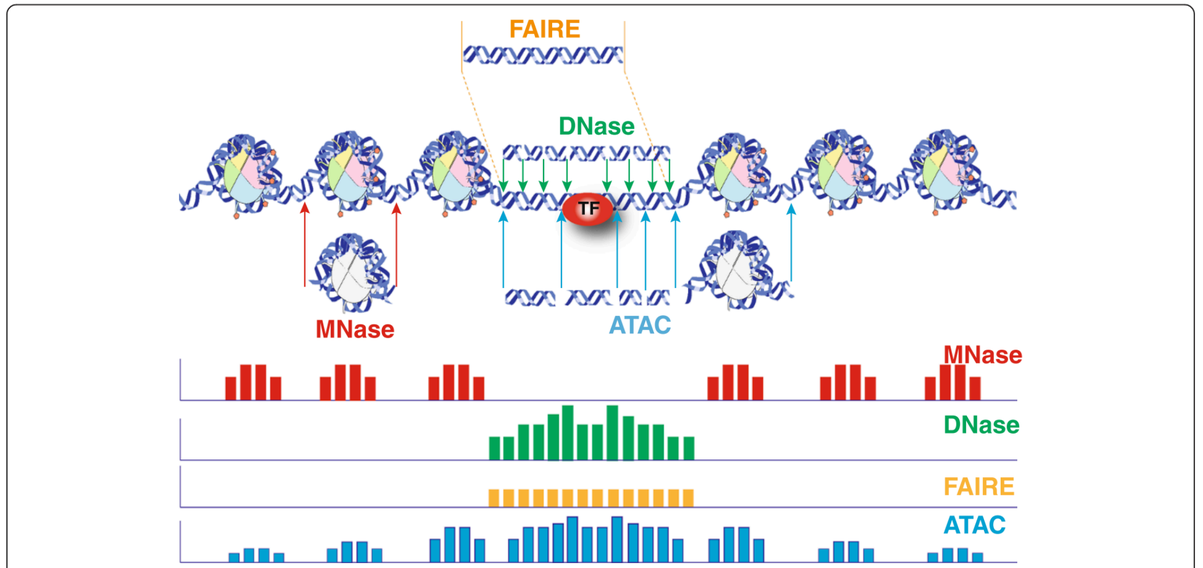


Figure 1 Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each assay across the entire region. The footprint created by a transcription factor (TF) is shown for ATAC-seq and DNase-seq experiments.

Tsompana and Buck, 2014

Slide by Meeta Mistry

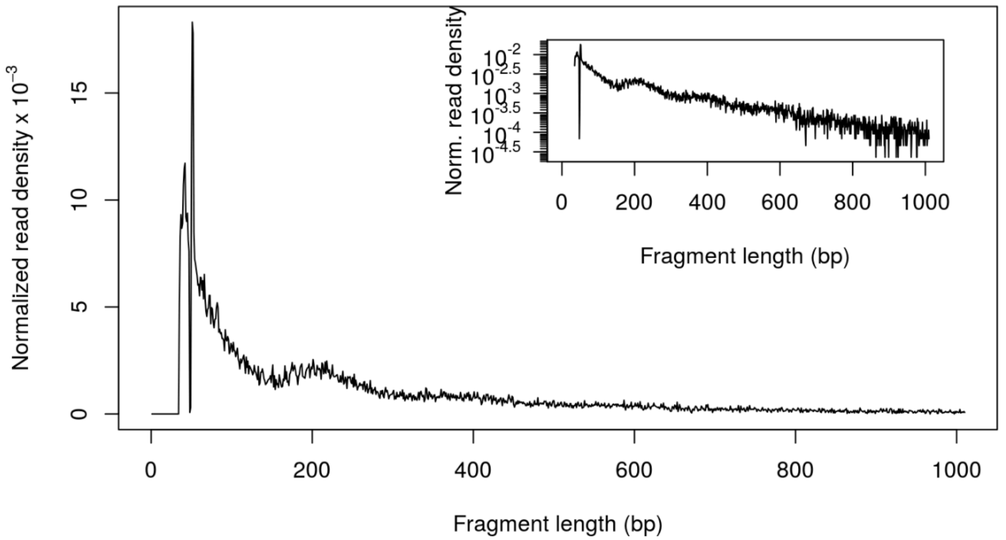
Planning your ATAC-seq experiment

- Replicates: more is better
- Controls: not typically run, but could use deproteinized "naked" genomic DNA
- PCR amplification: as few cycles as possible
- Sequencing depth: varies based on size of reference genome and degree of open chromatin expected
- Sequencing mode: paired-end
- Mitochondria: discarded from computational analyses; option to remove during prep

Adapted from slide by
Meeta Mistry

ATAC-seq data analysis

- Peak calling using MACS2 with PE settings and without model building
- Remove mitochondrial reads
- Shift alignments
- Separate nucleosome free regions (NFR) from nucleosome containing regions



Ou et al, Bioconductor 2018

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

bioinformatics.sph.harvard.edu