

Introduction to ChIP-seq

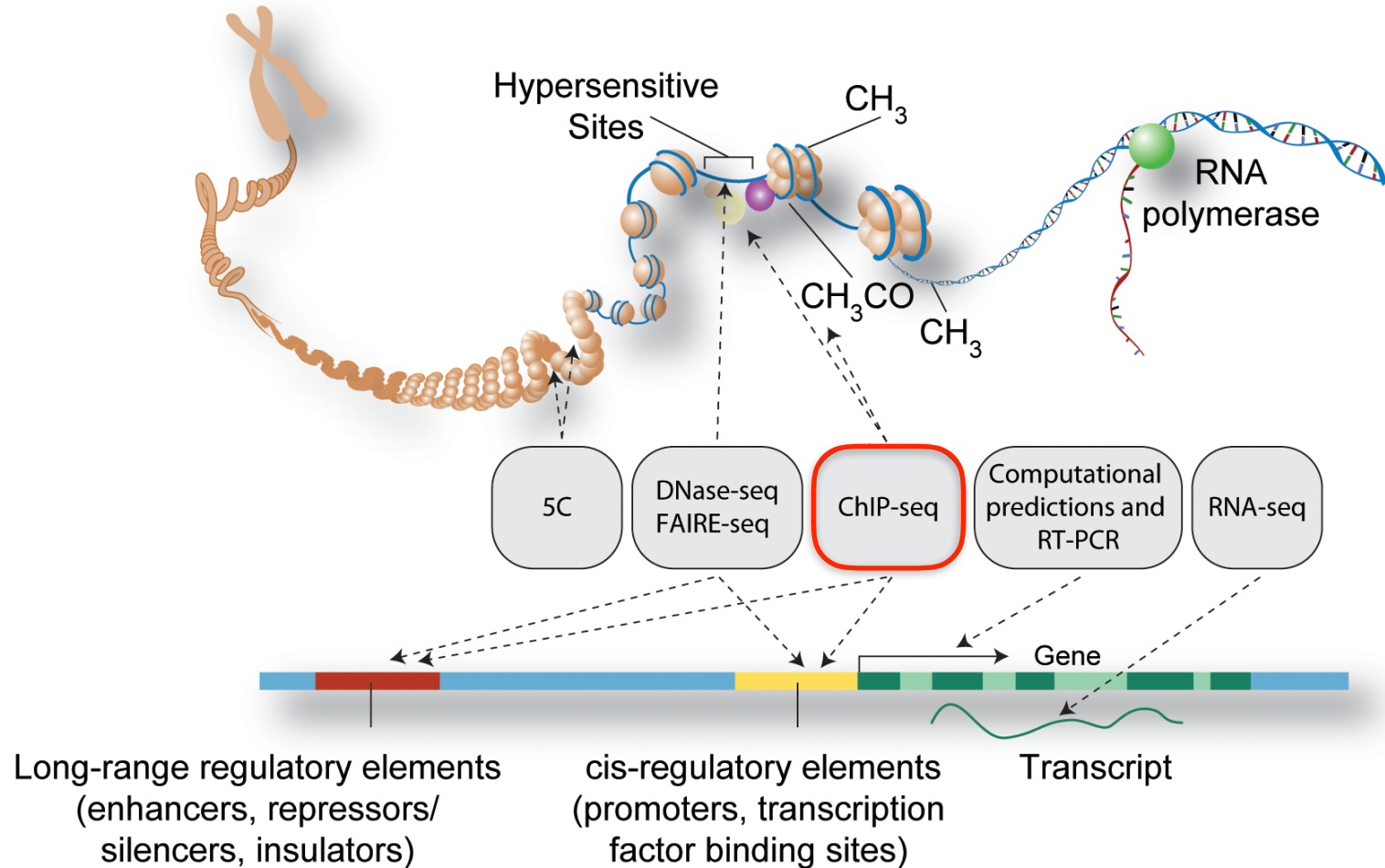
Outline

- Basic concepts of ChIP
- The importance of controls and QC
- Experimental design considerations
- ChIP-seq analysis workflow

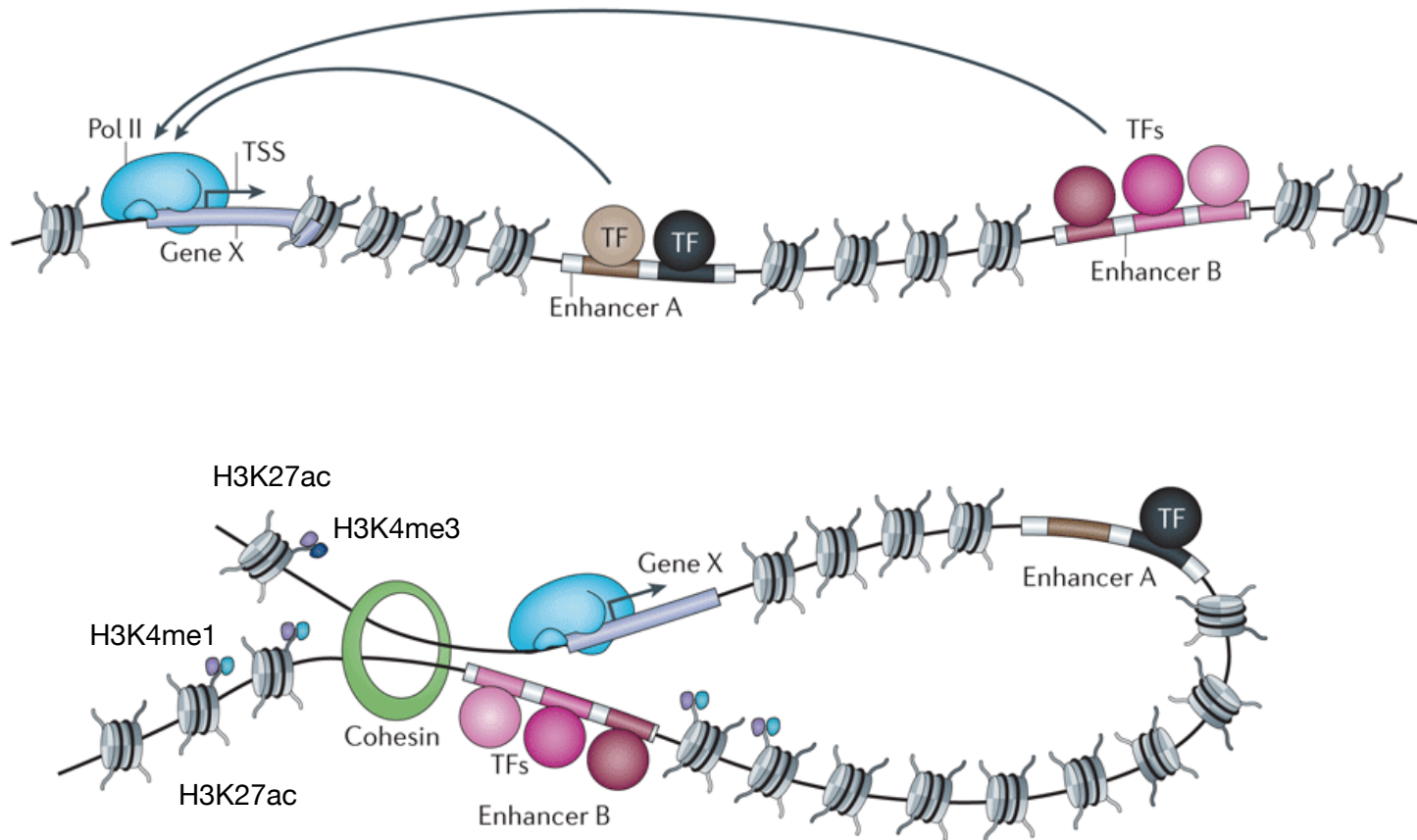
What is ChIP-seq

- Assay genome wide binding of protein to DNA
- Uses a combination of chromatin immunoprecipitation and sequencing
- Identifies how transcription factors and chromatin-associated proteins interact with DNA *in vivo*
- Complements DNA accessibility studies and gene expression profiling
- Gain a more precise picture of gene regulation

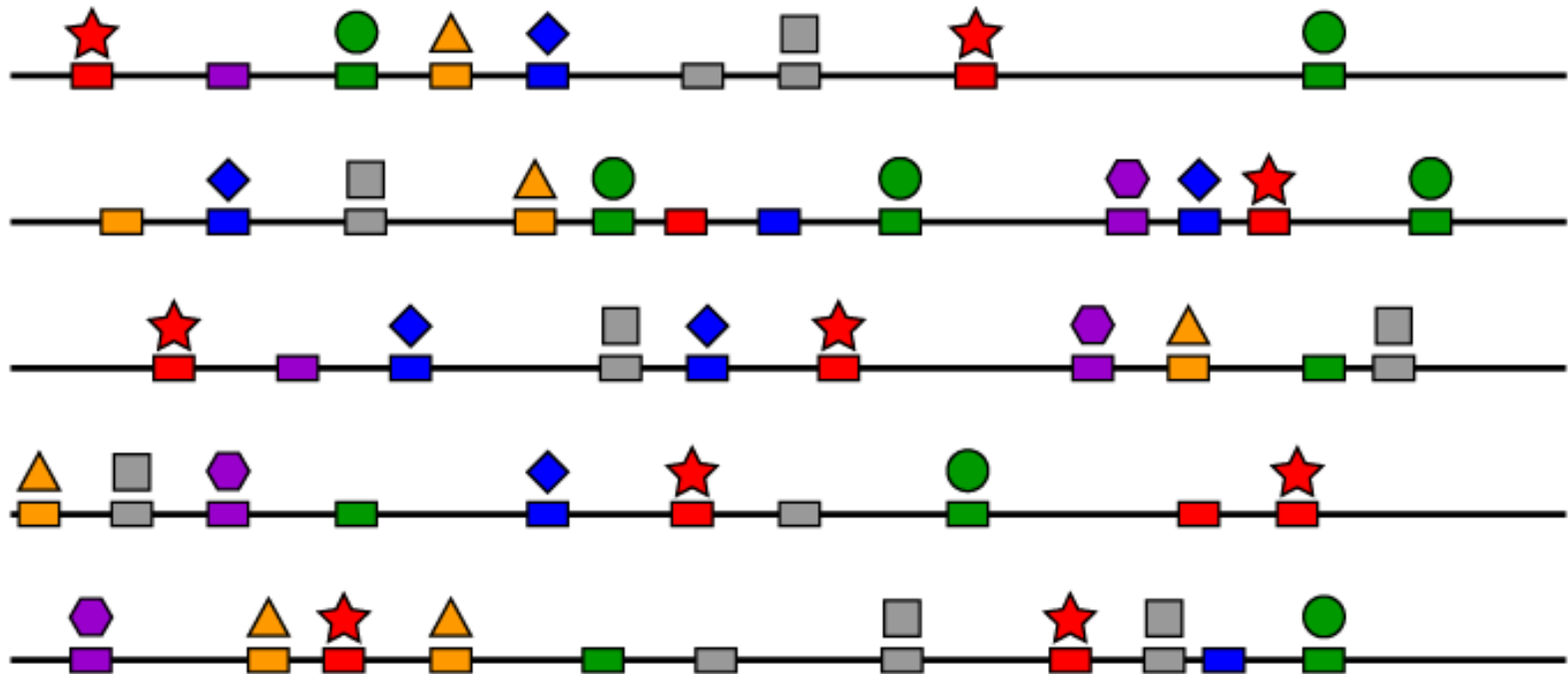
Transcriptional regulation is complex



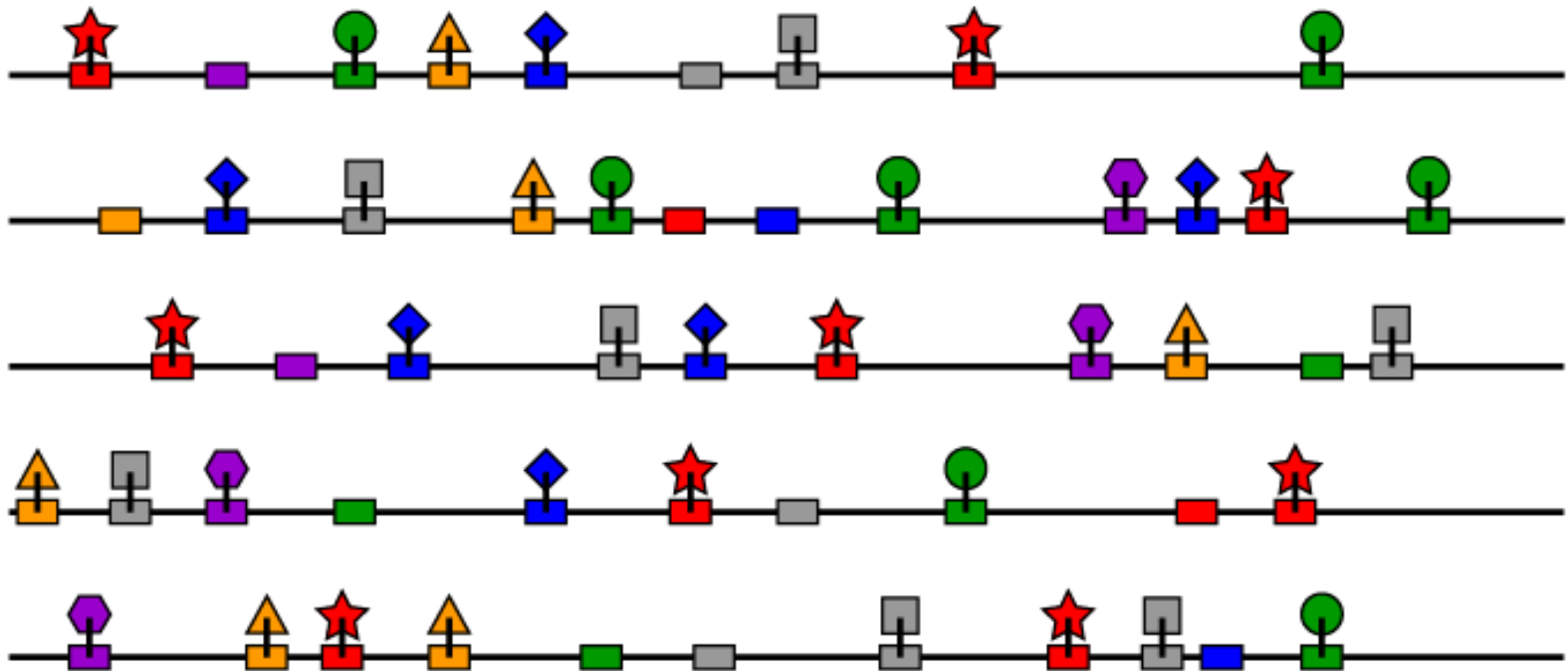
Simplified model of gene regulation



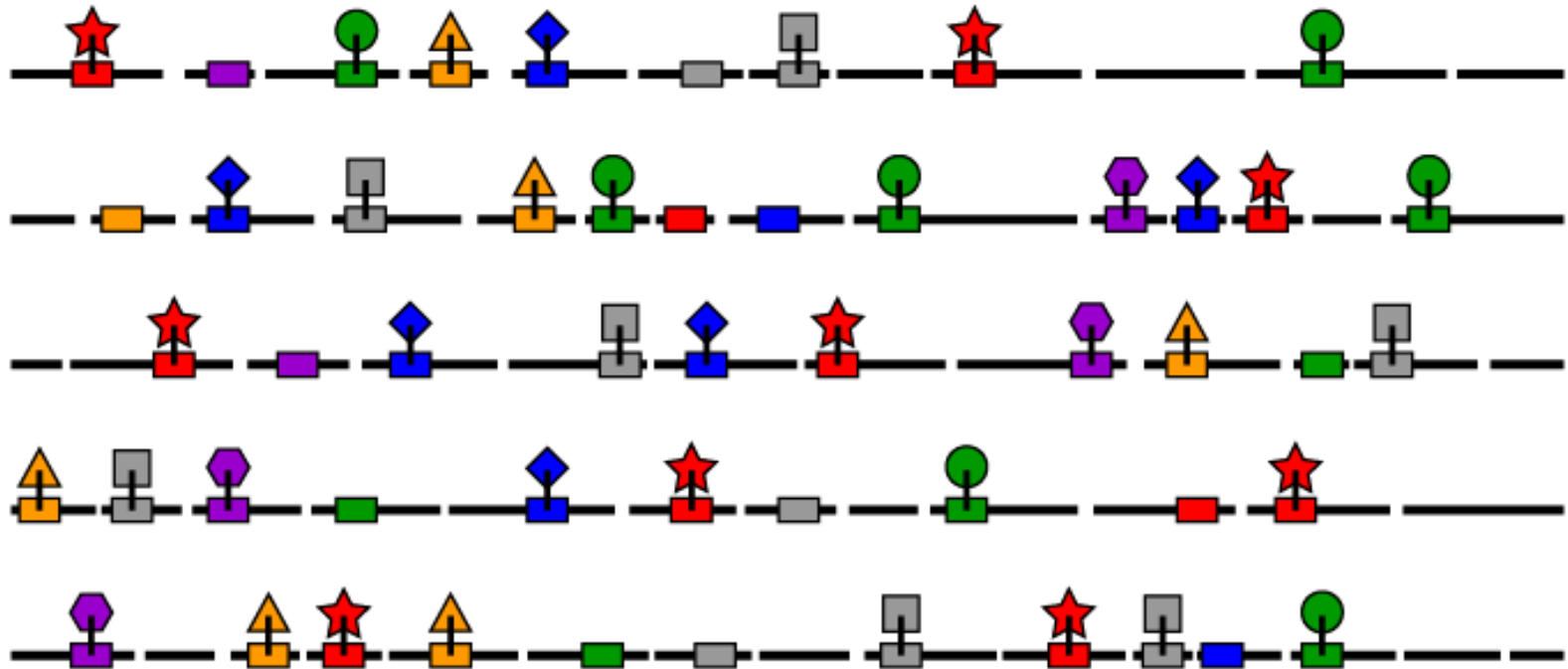
Library Preparation



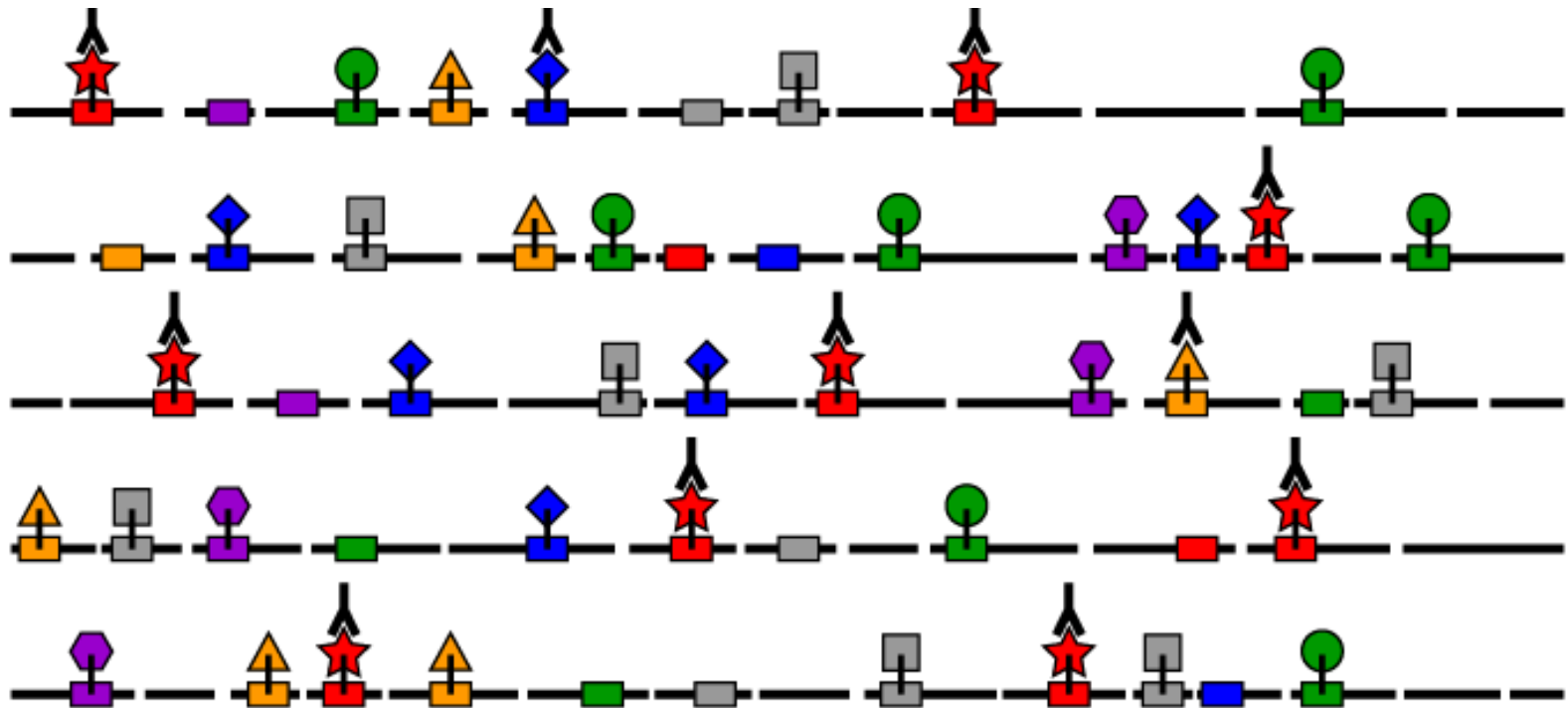
Crosslink proteins to DNA



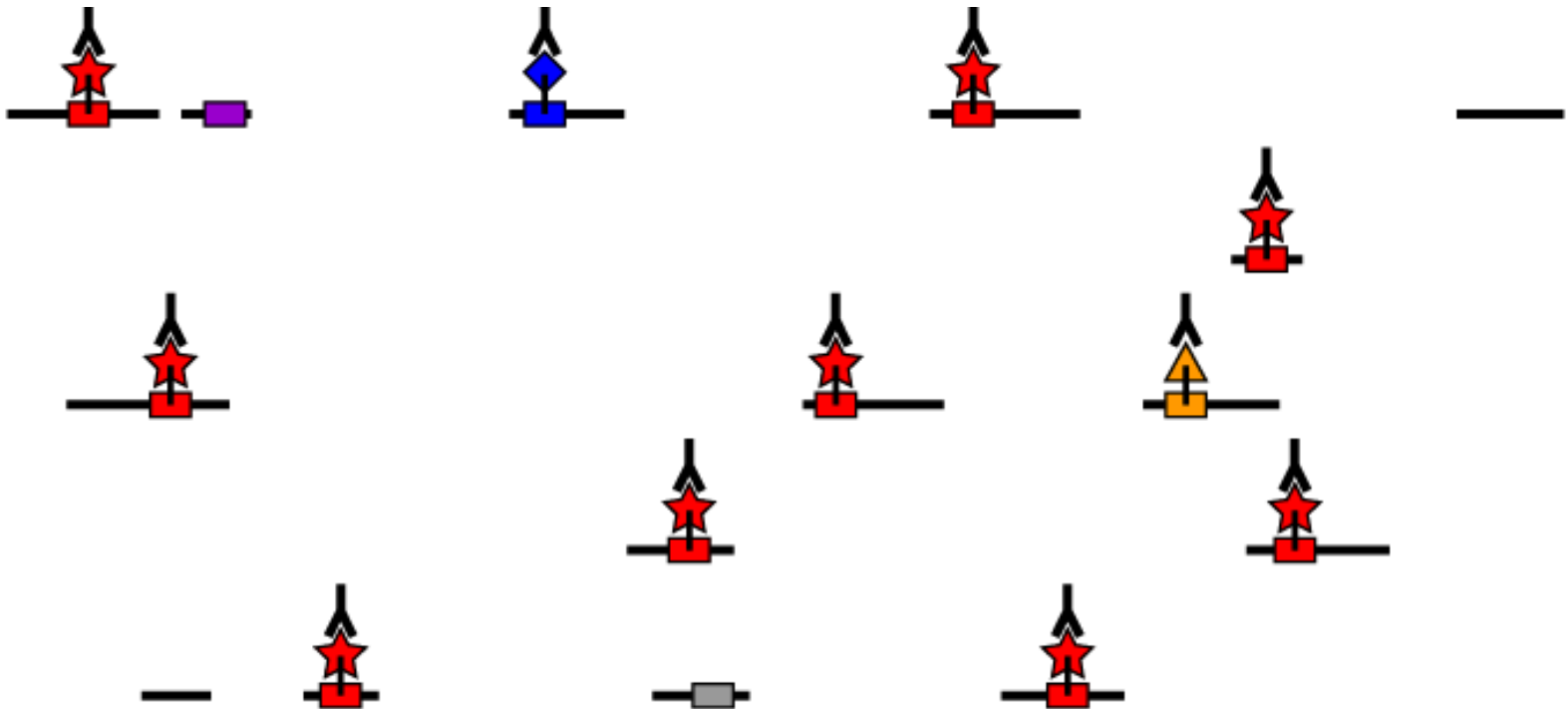
Fragment



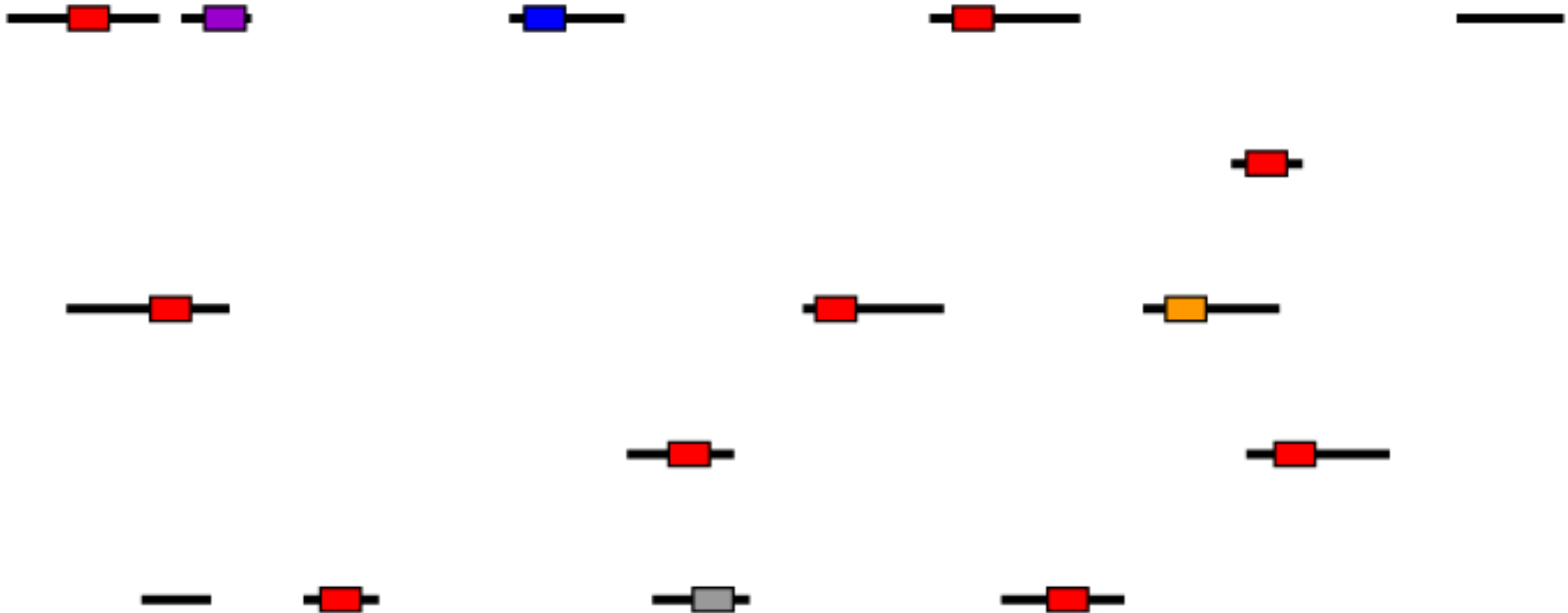
Protein specific antibody



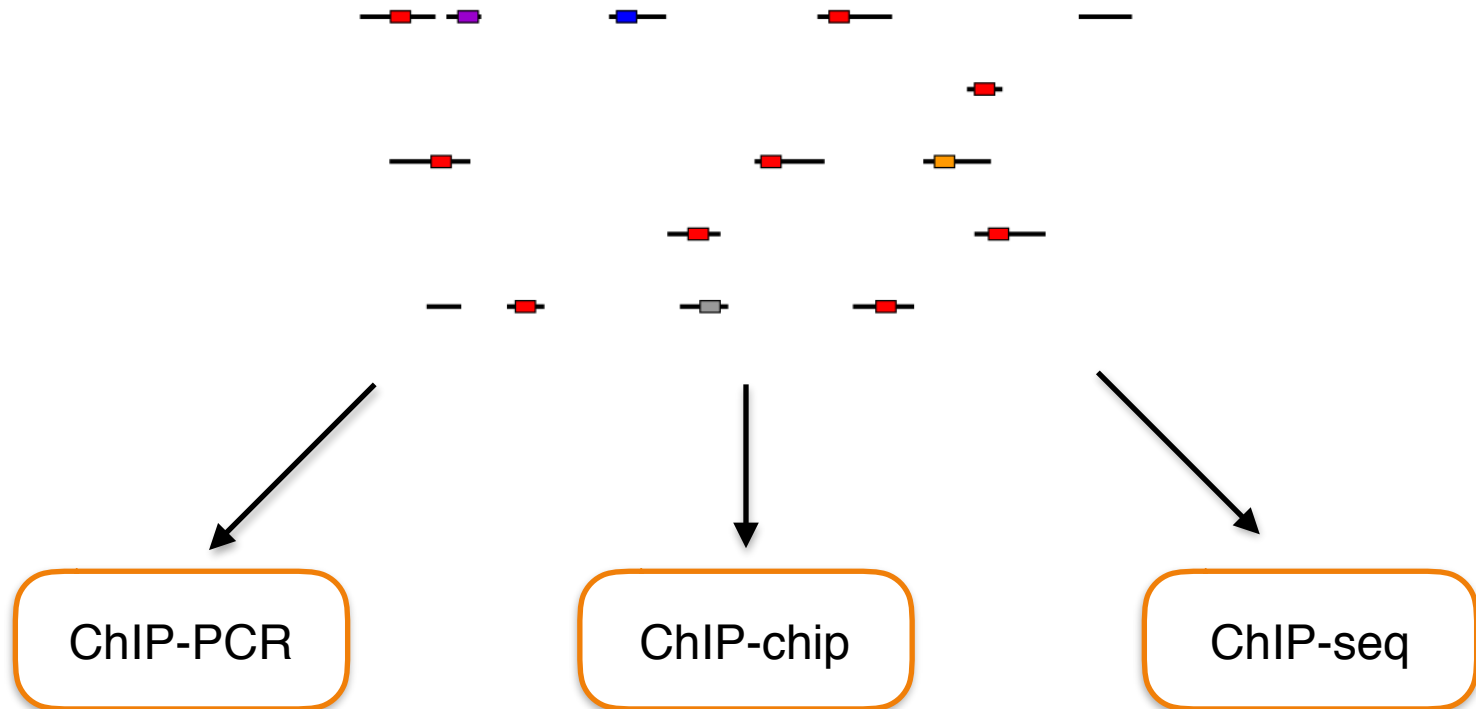
Immunoprecipitate



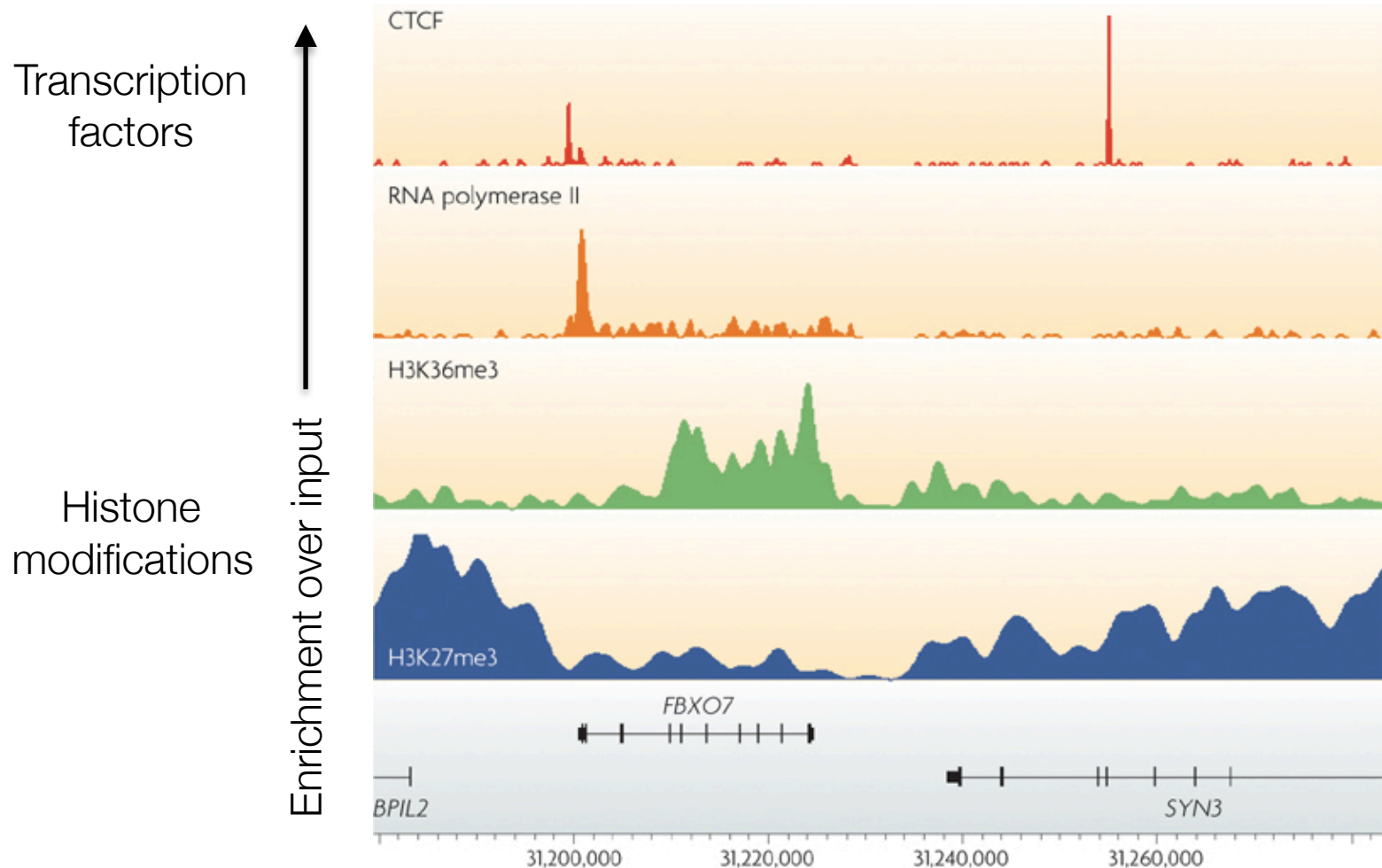
Reverse crosslink and purify DNA



Identify bound regions



Types of signals



Adapted from Park (2009). Nature Reviews Genetics.

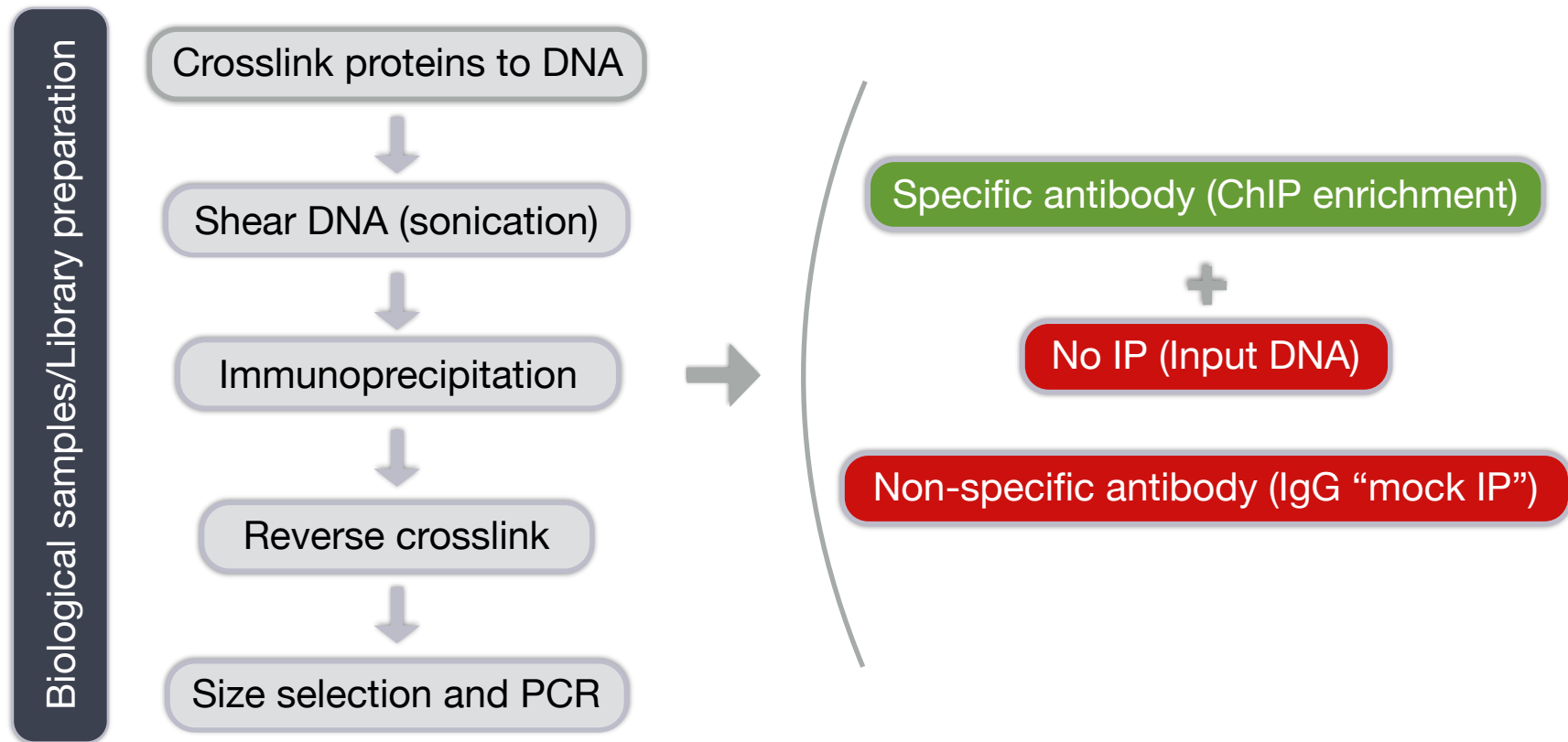
Profiling histone modifications

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

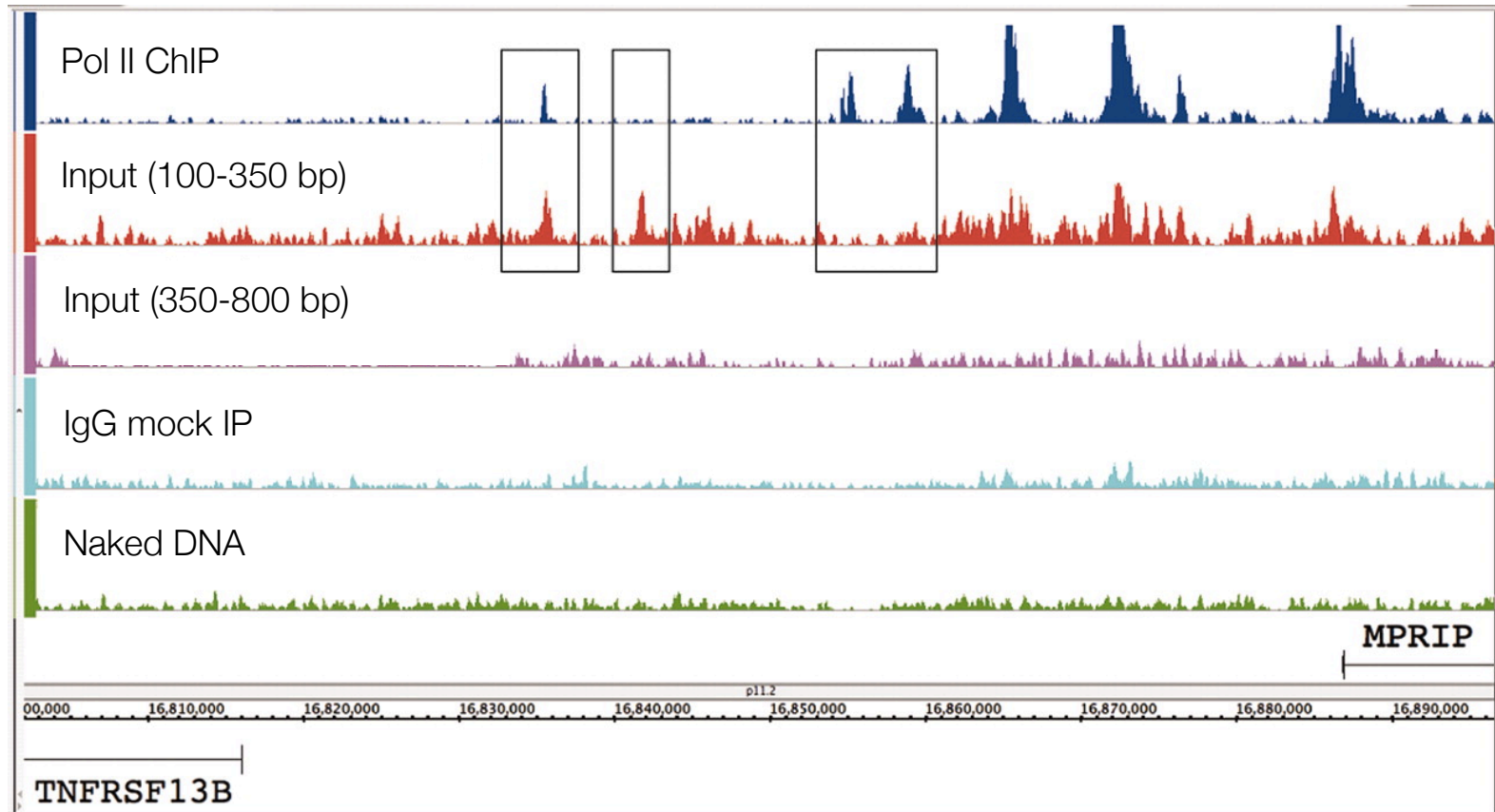
Why controls are necessary

- Allows us to compare with the same region in a matched control
- Artefacts can generate false positive peaks:
 - Open chromatin regions fragment more easily
 - Repetitive sequences might seem to be enriched
 - Uneven distribution of sequence tags across the genome
 - Hyper-ChIPable regions

ChIP-seq controls



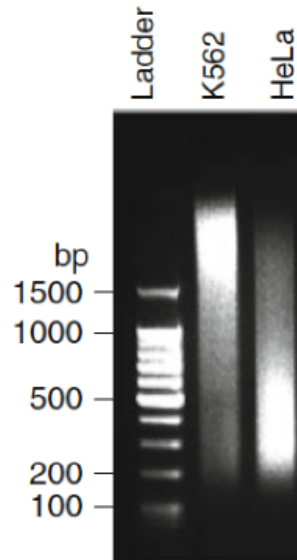
Control vs. ChIP signals



Adapted from Auerbach et al. (2009). PNAS.

Parameters for a successful ChIP-seq

- Efficient and specific antibody!
- Amount of starting material
- Chromatin fragmentation
- Stringency of washes
- Controls



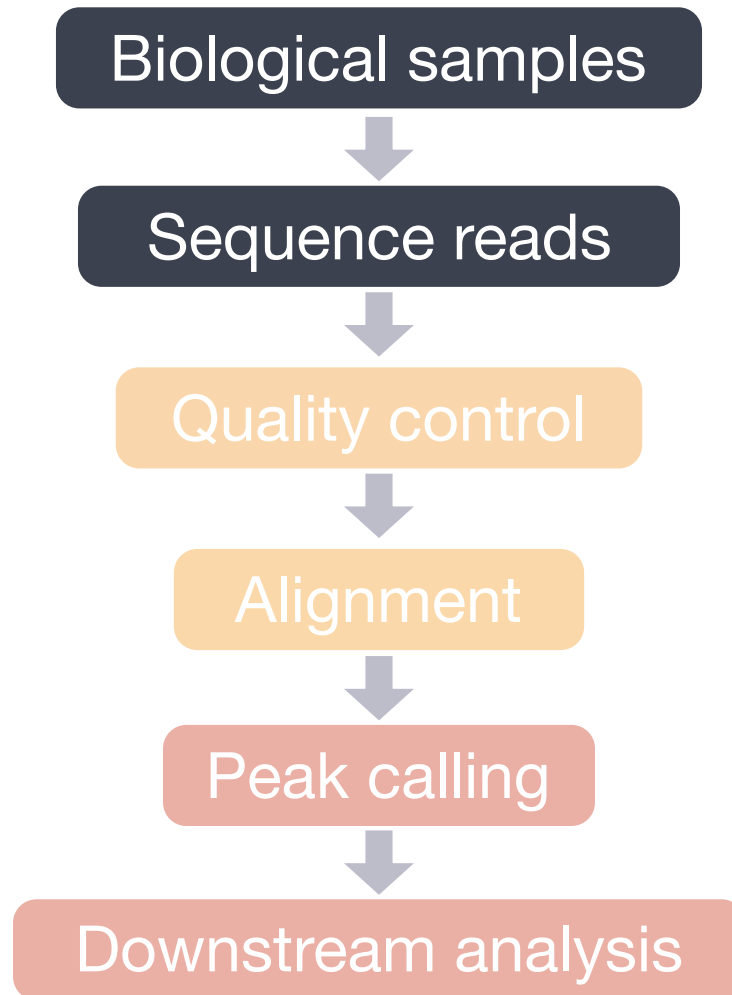
Fragments too big:

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

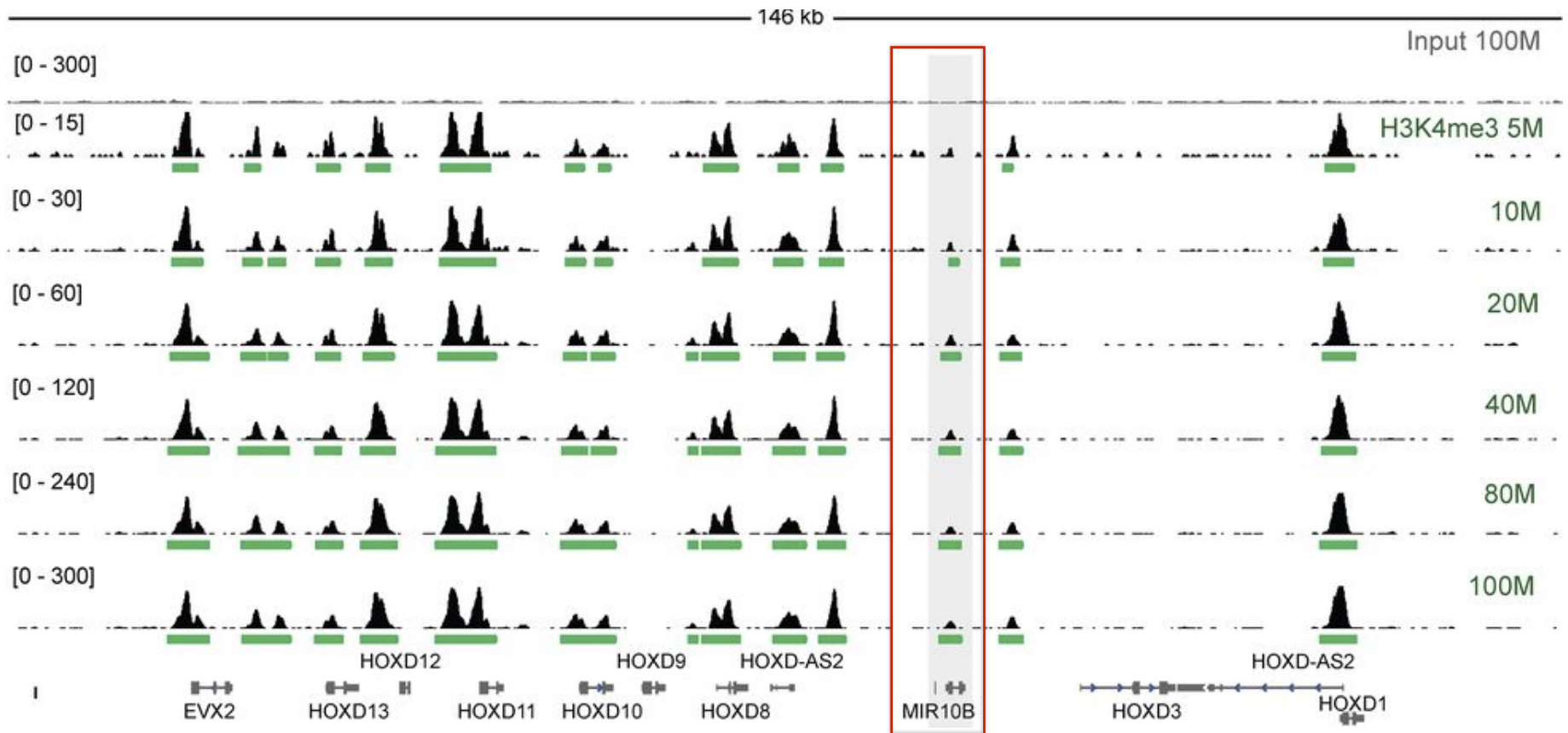


Experimental design considerations

- Read length (25- 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 evenly over batches
- Sequencing depth (5-10M minimum; 20-40M as standard for TFs; higher for broad profiles)
- Input controls should be sequenced to equal or greater depths than IP samples

Impact of sequencing depth

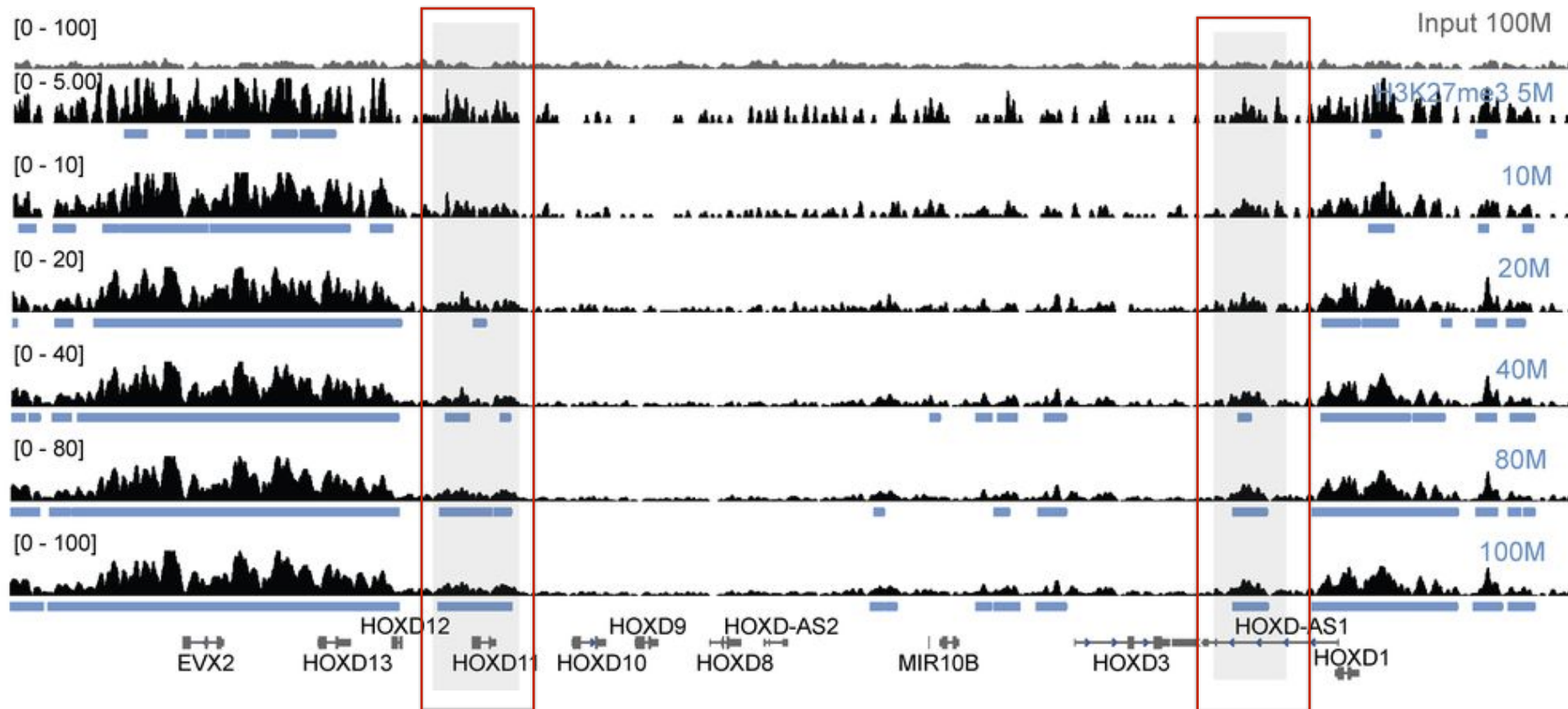
H3K4me3



Adapted from Jung et al (2014). NAR.

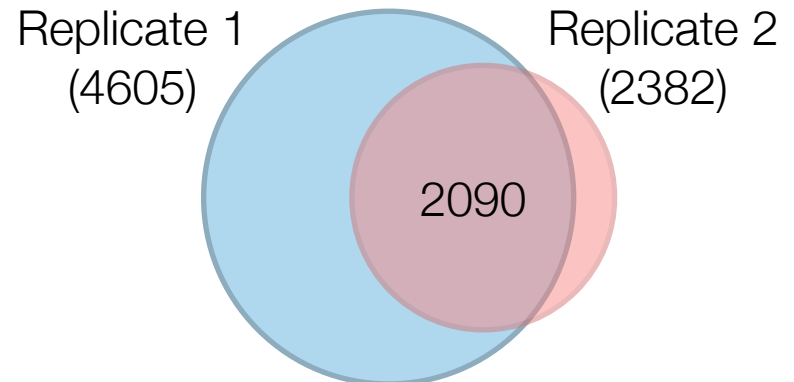
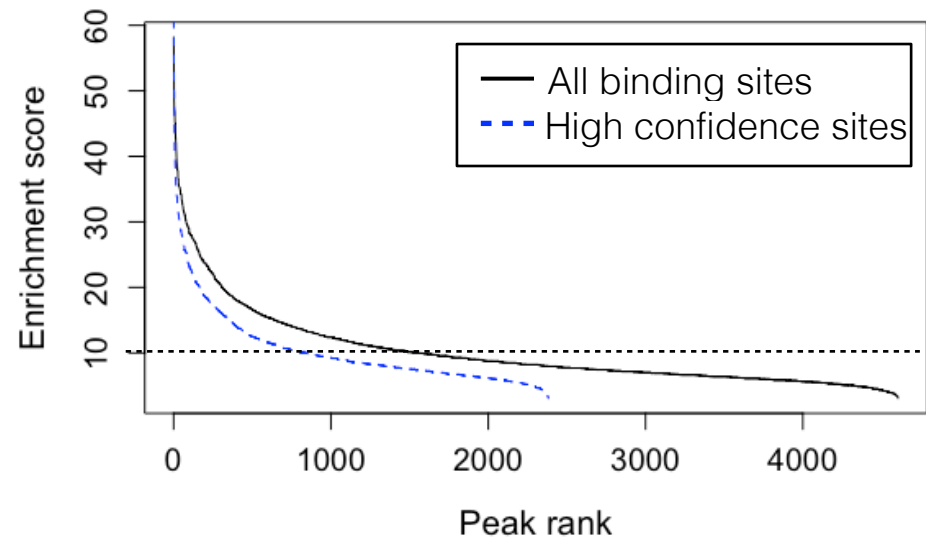
Impact of sequencing depth

H3K27me3

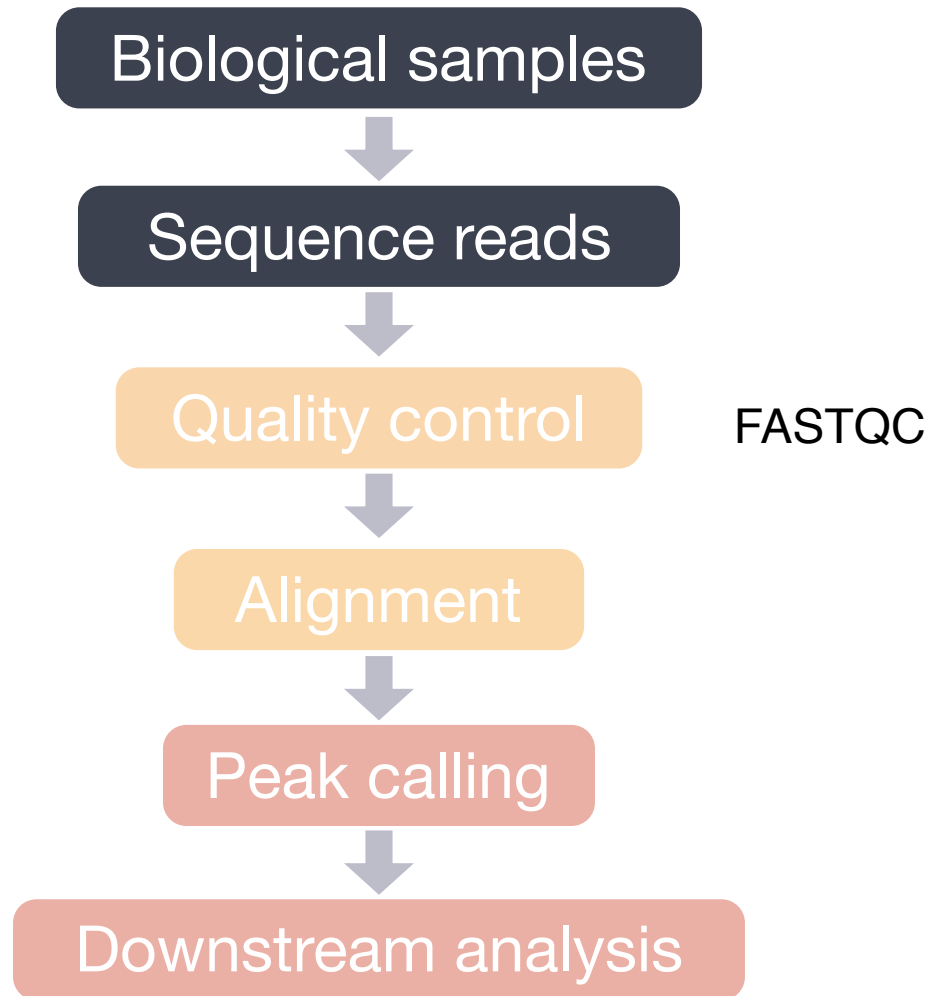


Replicates and reproducibility

- 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



ChIP-seq workflow



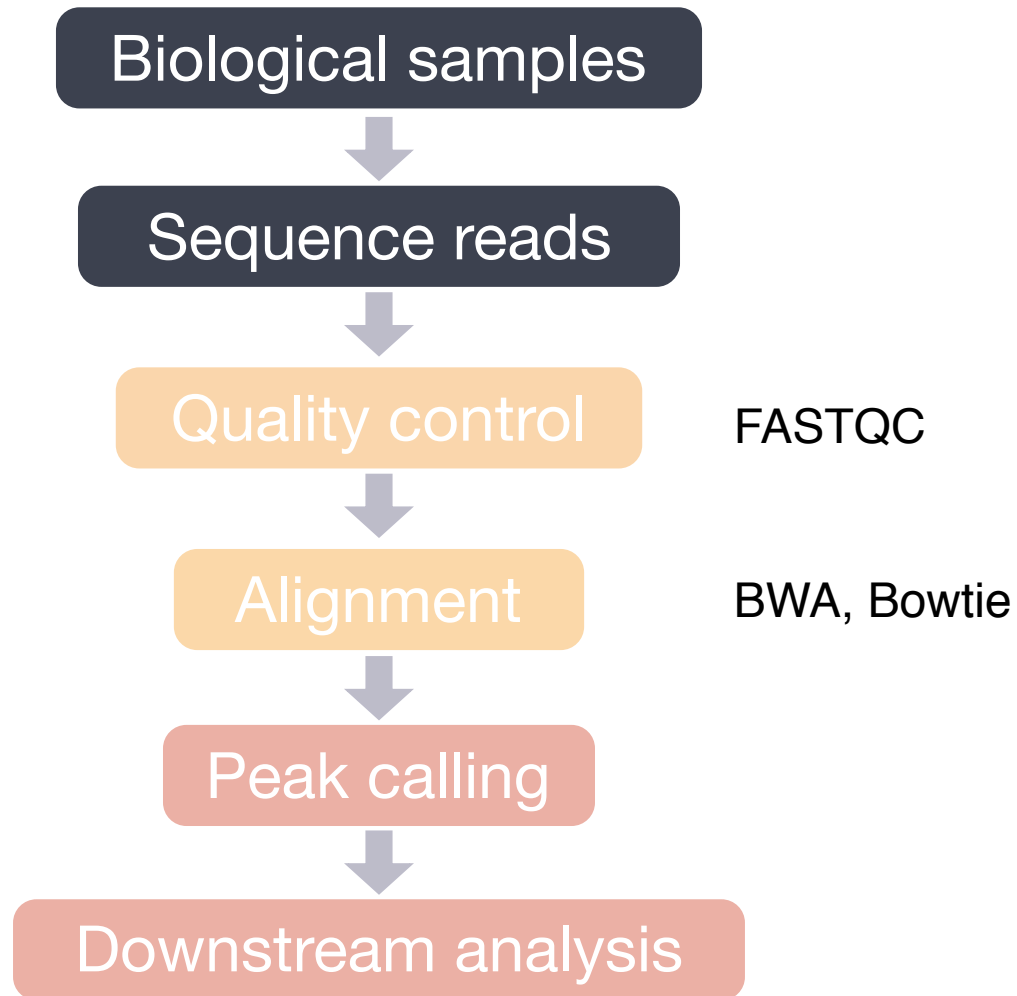
Quality Checks: Raw Data

All NGS analyses require that the **quality of the raw data** is assessed prior to any downstream analysis.

The quality checks at this stage in the workflow include:

1. Checking the quality of the base calls to ensure that there were no issues during sequencing
2. Examining the reads to ensure their quality metrics adhere to our expectations for our experiment
3. Exploring reads for contamination

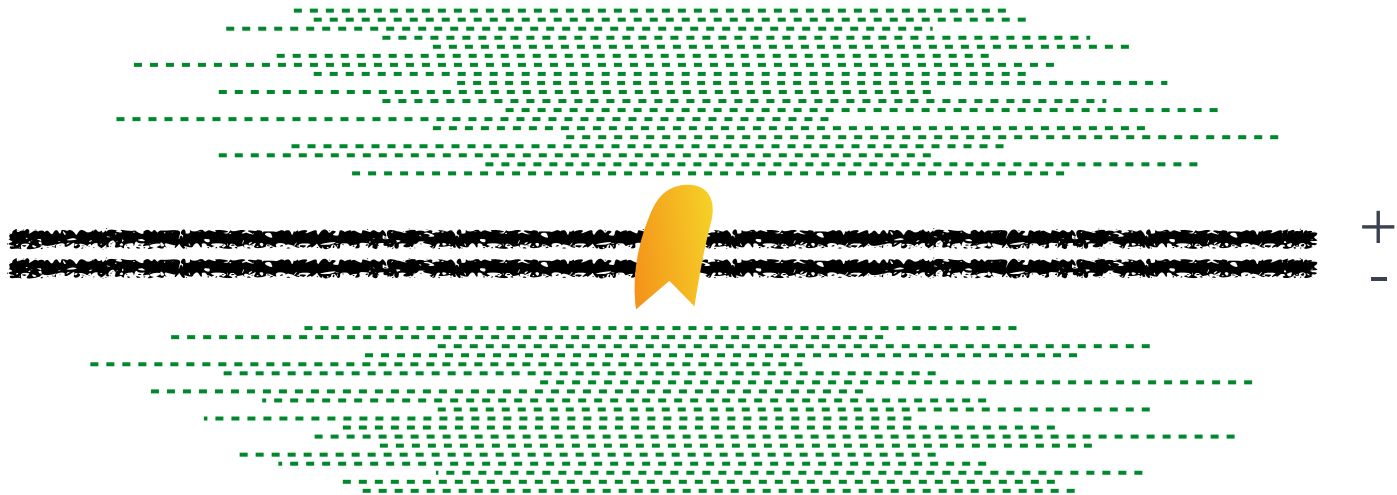
ChIP-seq workflow



Understanding strand cross-correlation

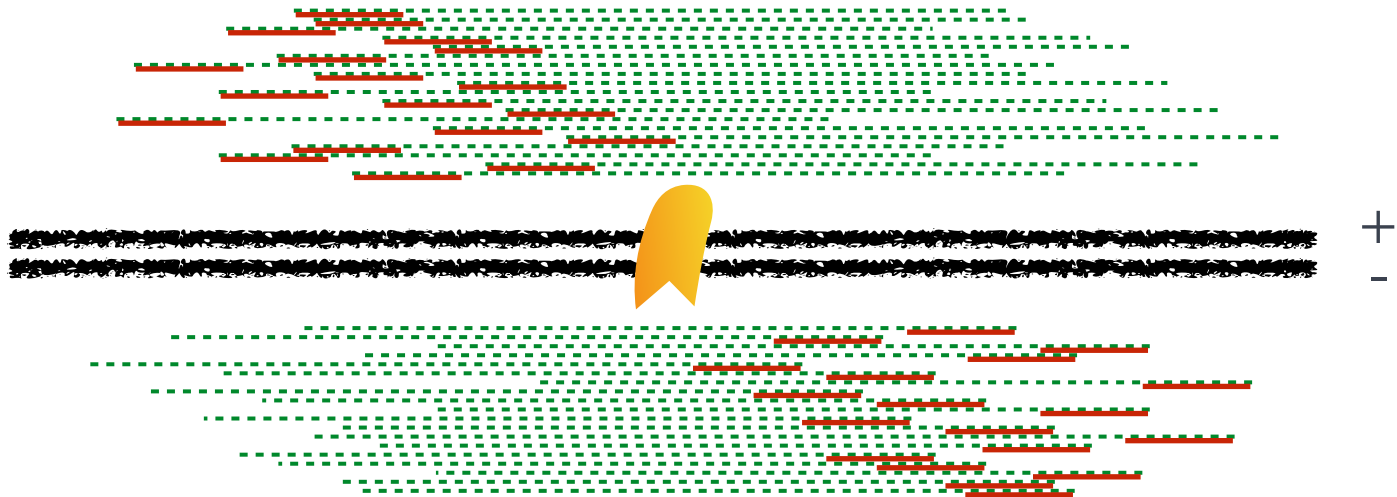
 = binding site

- - - = size selected DNA fragment



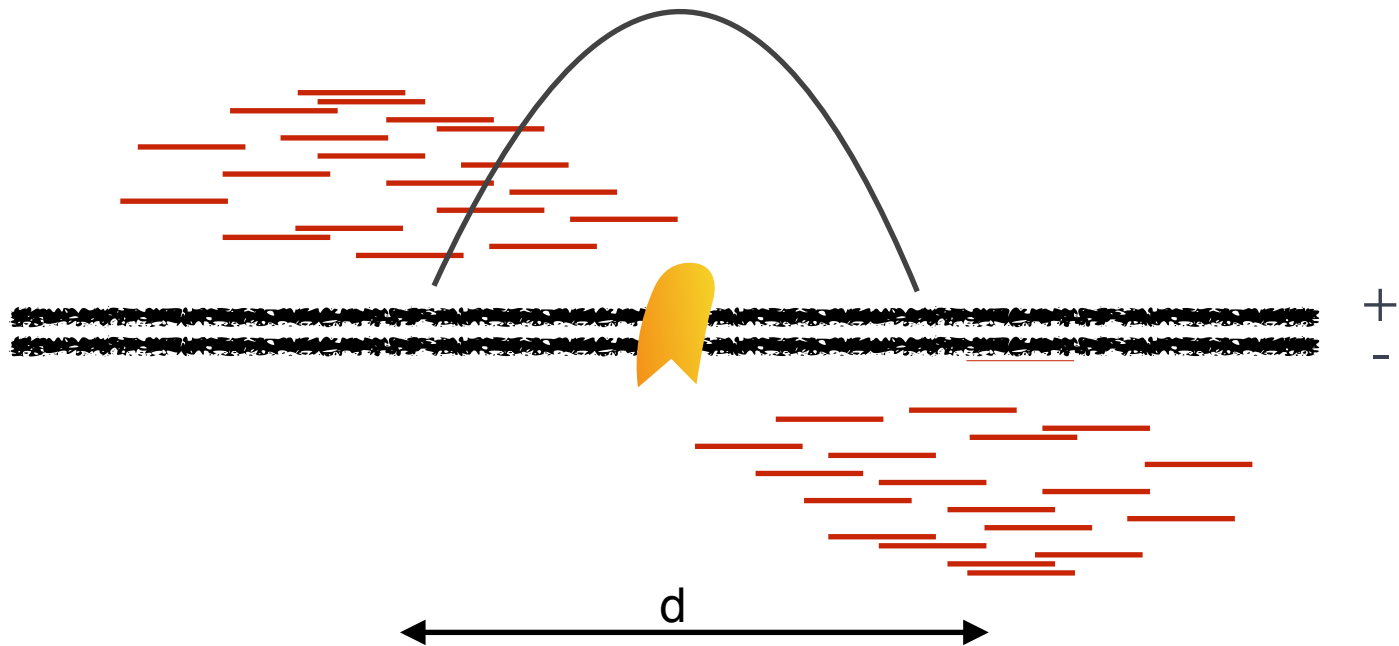
Understanding strand cross-correlation

ChIP-seq fragments are sequenced from the 5' end



Understanding strand cross-correlation

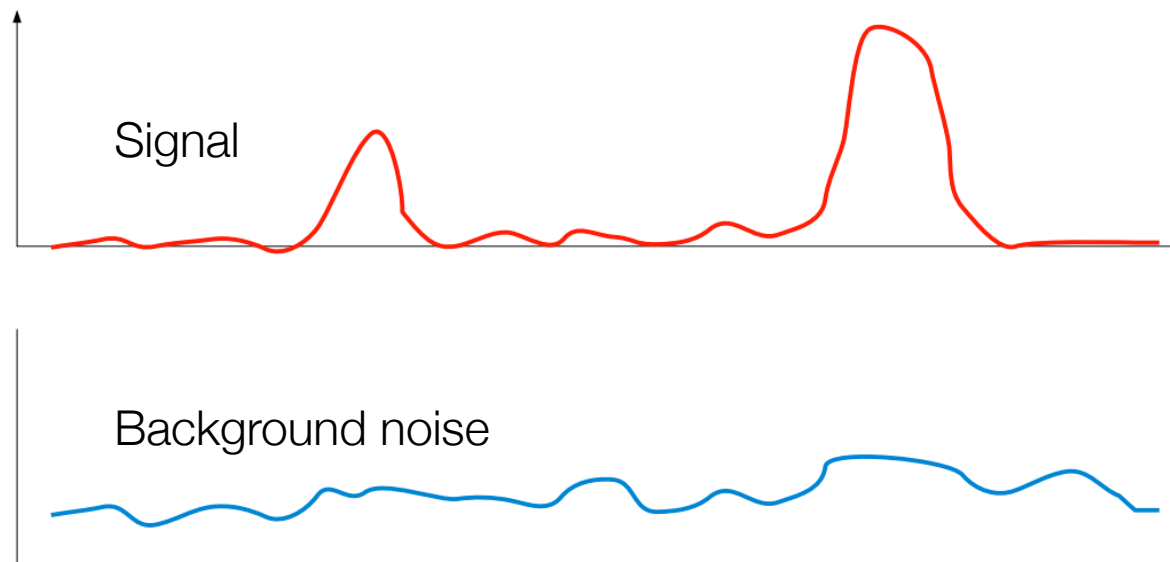
Alignment generates a **bimodal pattern** on the plus and minus strands around binding sites



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

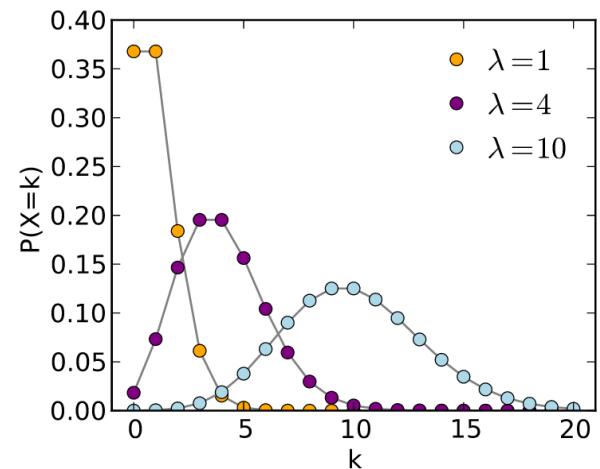
Modeling noise to detect real peaks

- How much signal we have is dependent on:
 - the number of active binding sites
 - the number of starting genomes (or cells)
 - the efficiency of the IP



Peak detection

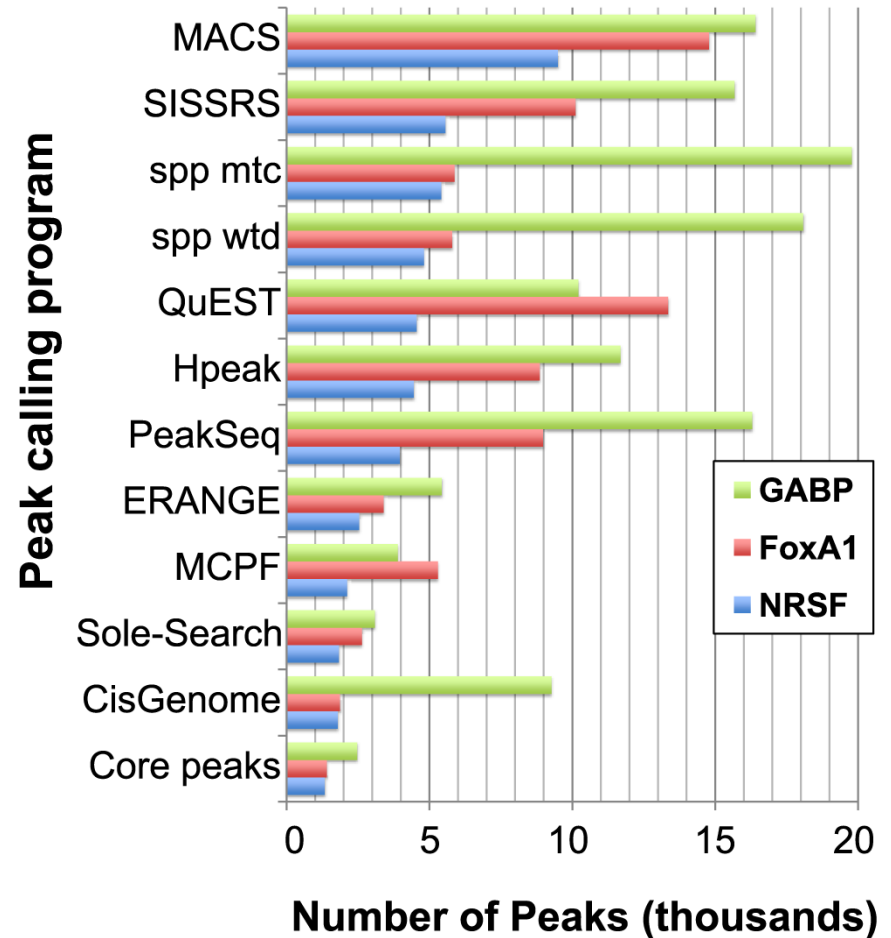
- Most algorithms model the number of reads from a genomic region/window using a Poisson distribution
- Often more variance in real data than assumed by the Poisson (overdispersion)
- MACS (model-based analysis of ChIP-Seq) uses multiple Poisson distributions to model the local background noise within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



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)

Quality Checks

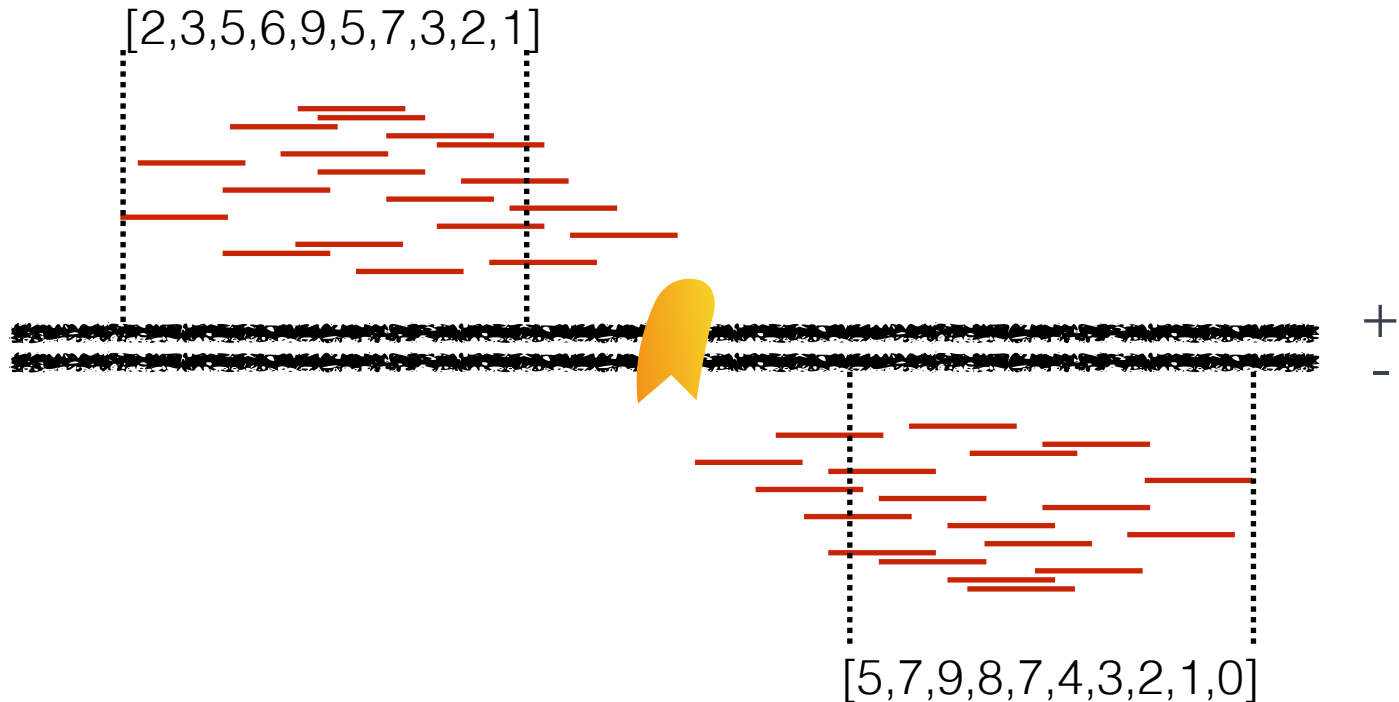
Evaluating the **quality of the aligned data and peak calls** can give important information about the quality of the library. The quality checks at this stage in the workflow include:

1. Checking the percent of reads aligning to the genome
2. Removing blacklisted regions
3. Exploring duplication rates, cross correlation scores and fraction of reads in peaks (FRiP)

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

Understanding strand cross-correlation

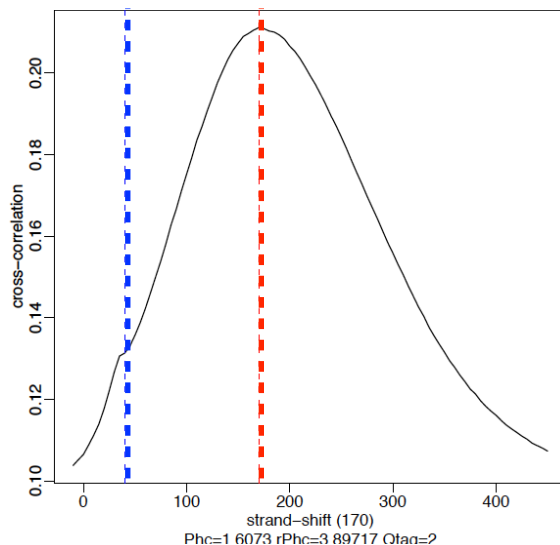
At strand shift of zero, the Pearson **correlation** between the two vectors is **0.539**



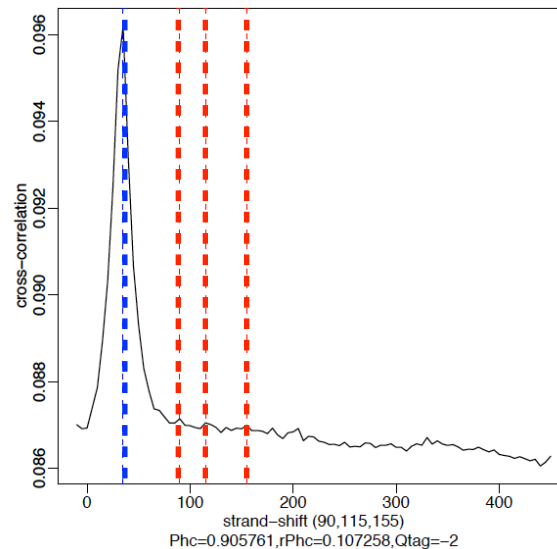
Cross correlation profiles

- Red vertical line shows the dominant peak at the true peak shift
- Blue vertical line is at read-length

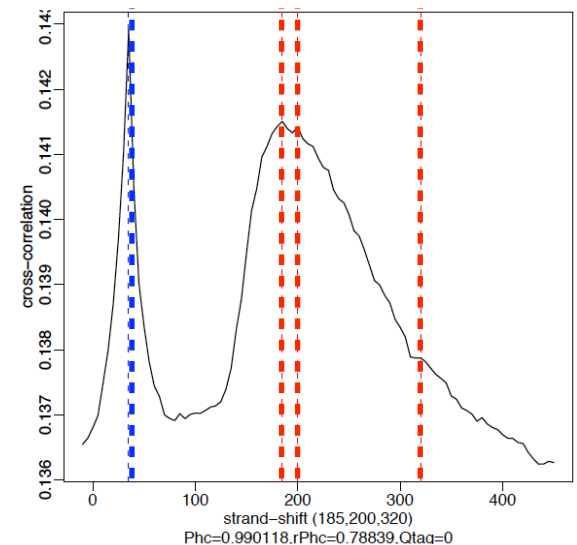
Strong signal



Input DNA



Marginal signal



Metrics based on cross correlation

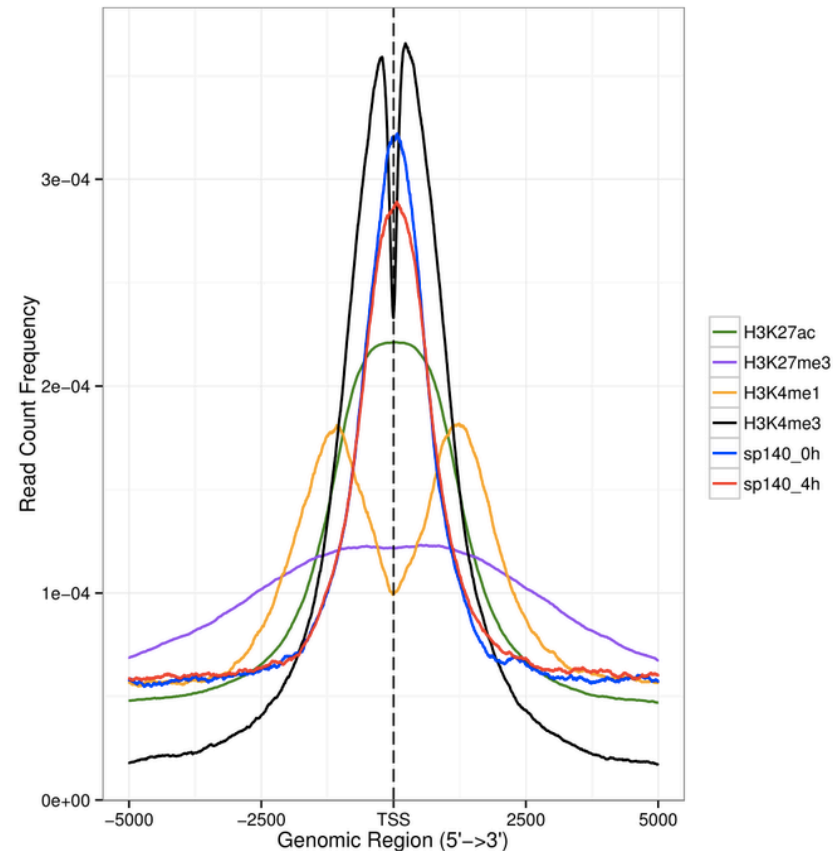
- **Normalized strand cross-correlation coefficient (NSC):** CC at fragment length peak / minimum CC value. Higher values indicate more enrichment.
 - Minimum value: 1
 - Critical threshold: 1.1
- **Relative strand cross-correlation coefficient (RSC):** (CC at fragment-length peak - min CC) / (CC at read-length peak - min CC).
 - Minimum value: 0
 - Critical threshold: 1
- Low scores indicate low signal to noise
 - Failed ChIP, poor sequence quality (leading to mismapping), inadequate sequencing depth
 - OR factor only binds a few sites

Downstream analysis

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

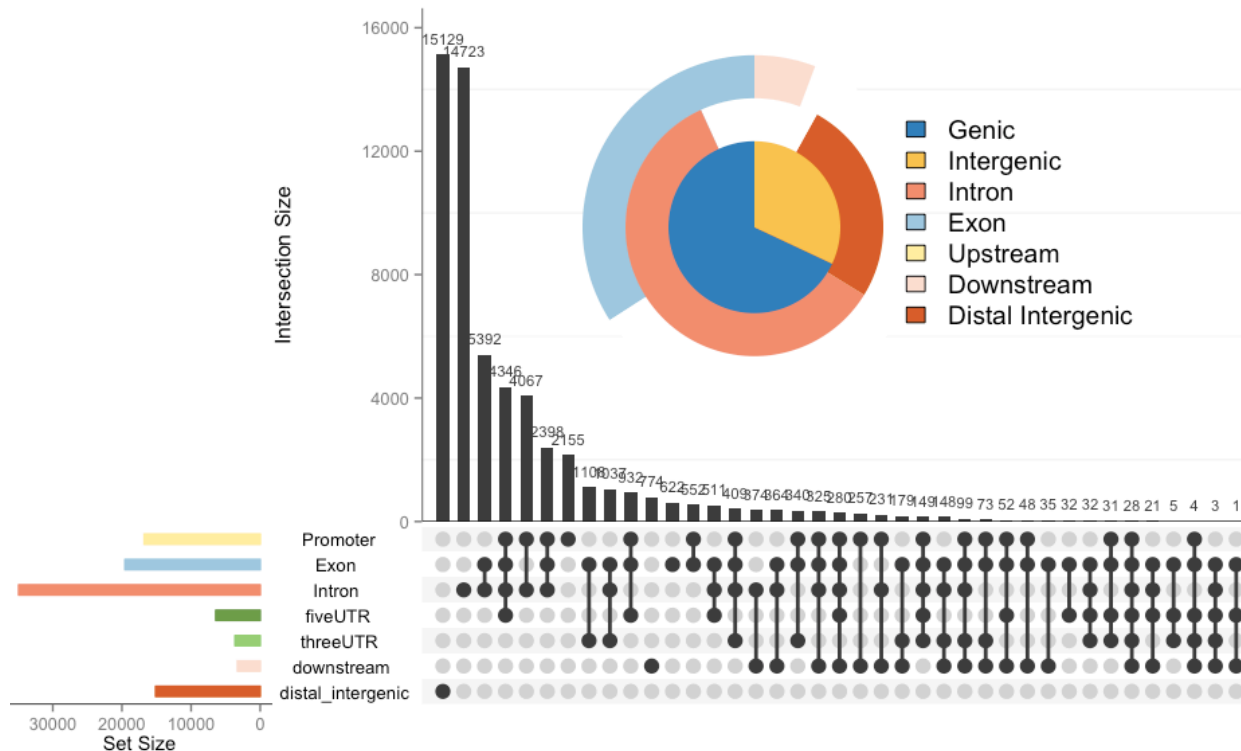
Downstream analysis

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



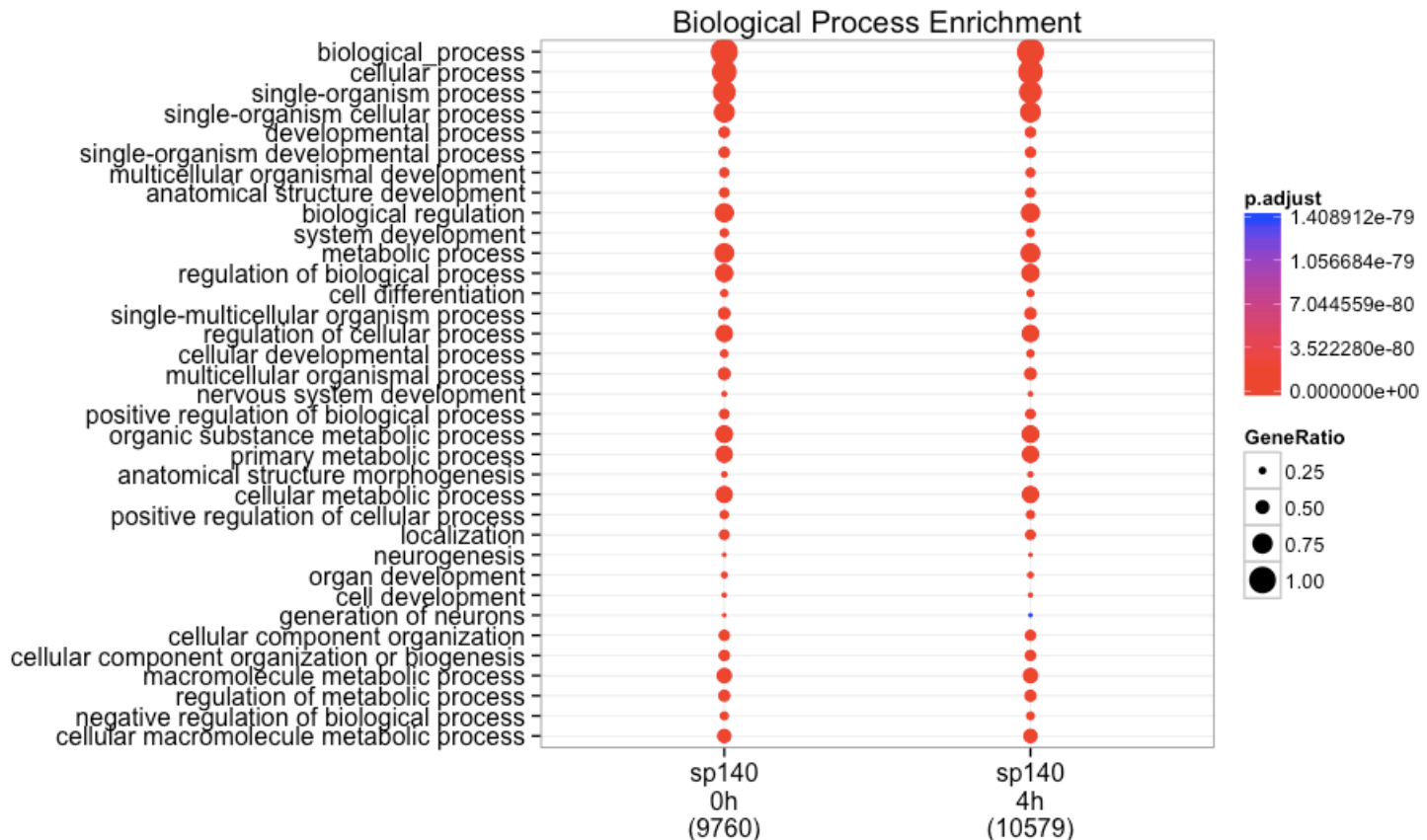
Downstream analysis

- 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 ?	Unersased 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	↓	...→

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