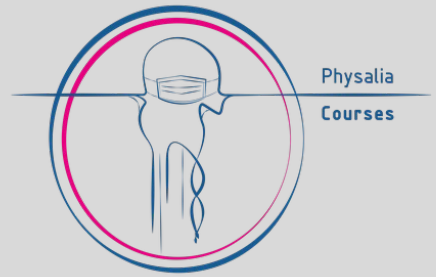
ChIP-seq analysis

Epigenomics Data Analysis

Jacques Serizay

Physalia 2025



Chromatin immunoprecipitation: an old tool rejuvenated by high-throughput sequencing

Chromatin IP is not a new approach. It has been around for the past three decades

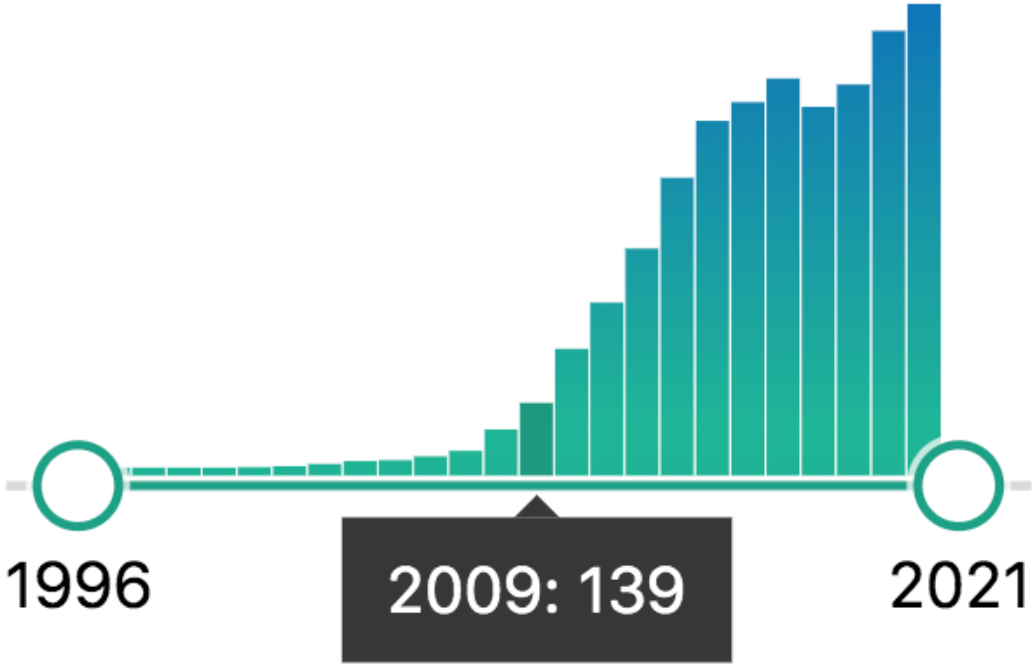
Cell

Summary

We have used formaldehyde-mediated proteinDNA crosslinking within intact cells to examine the in vivo chromatin structure of the D. melanogaster heat shock protein 70 (*hsp70*) genes. In agreement with previous in vitro studies, we find that the heat shockmediated transcriptional induction of the *hsp70* genes perturbs their chromatin structure, resulting in fewer proteinDNA contacts crosslinkable in vivo by formaldehyde. However, contrary to earlier in vitro evidence that histones may be absent from actively transcribed genes, we show directly, by immunoprecipitation of in vivocrosslinked chromatin fragments, that at least histone H4 remains bound to *hsp70* DNA in vivo, irrespective of its rate of transcription. The formaldehyde-based in vivo mapping techniques described in this work are generally applicable, and can be used both to probe proteinDNA interactions within specific genes and to determine the genomic location of specific chromosomal proteins.

Chromatin immunoprecipitation: an old tool rejuvenated by high-throughput sequencing

It gained a lot of traction when high-throughput sequencing emerged



Current ChIP-seq approaches

ChIP-seq

Low input ChIP-seq

Native ChIP-seq

Indirect ChIP-seq with DamID

ChIP-seq with chemical-based fragmentation

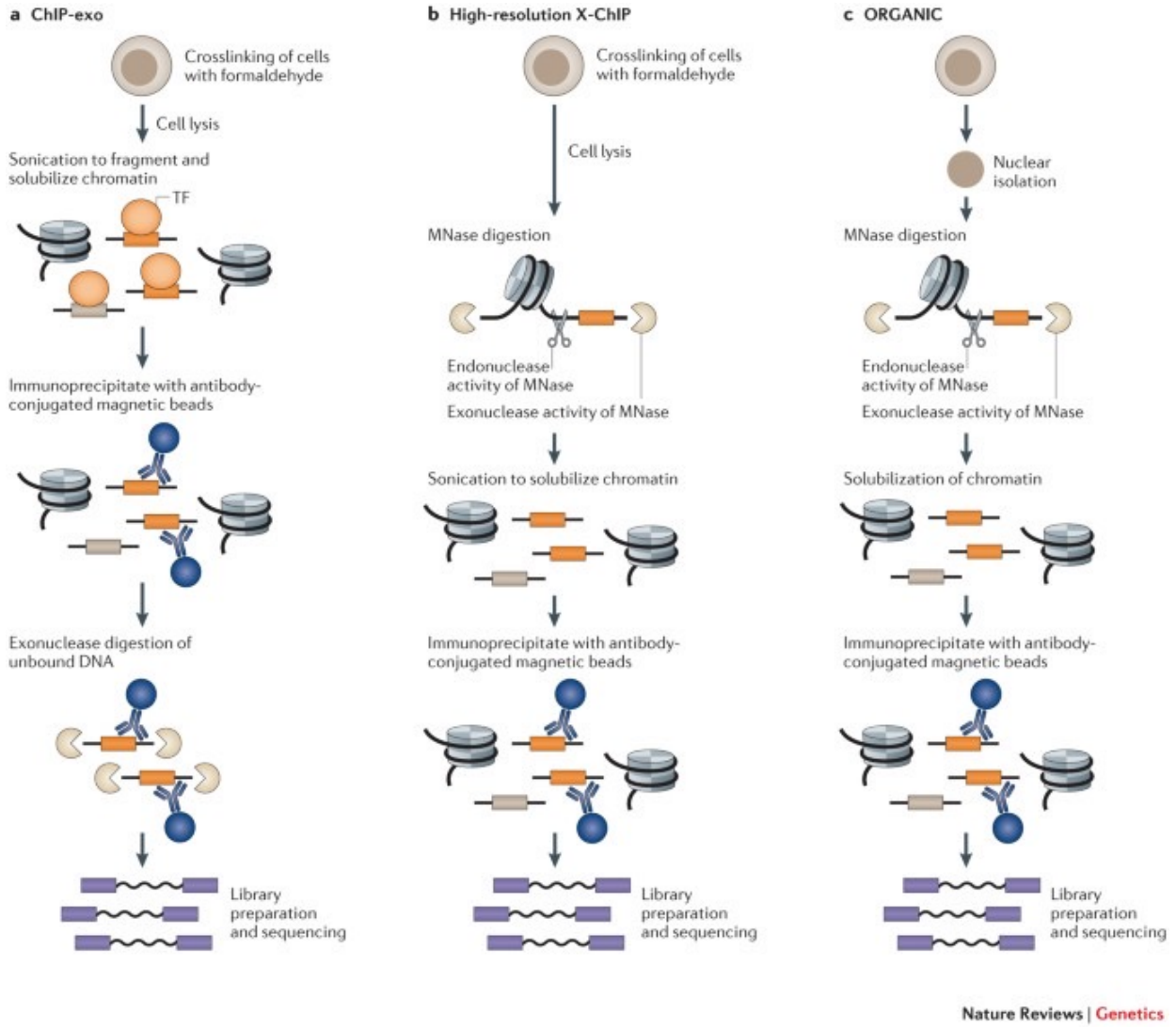
Cut&Run, Cut&Tag

Single-cell ChIP-seq

…

Current ChIP-seq approaches

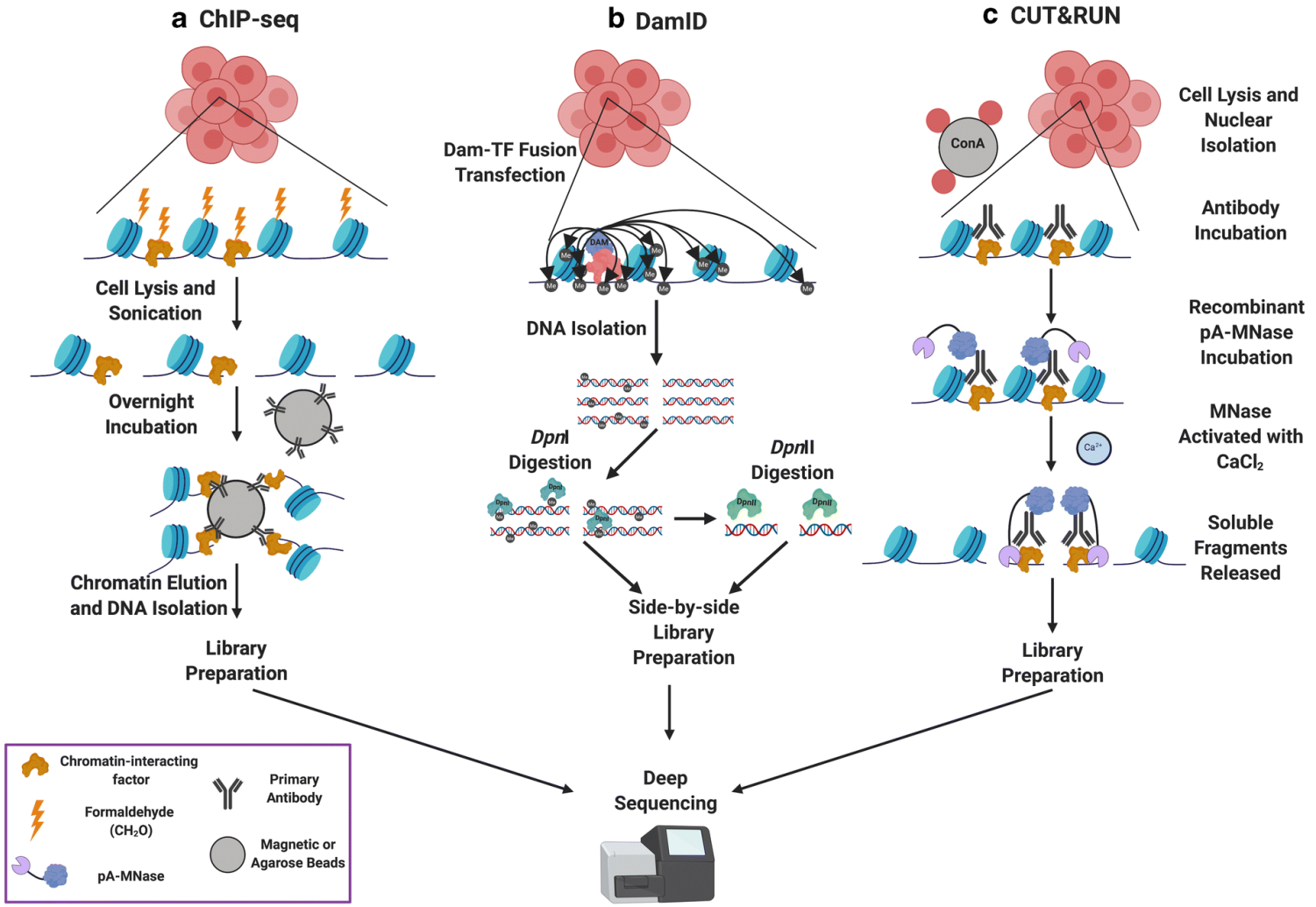
Direct approaches



Zentner & Henikoff, Nat. Rev. Genet. 2014

Current ChIP-seq approaches

Indirect approaches



Klein & Hainer, Chromosome Research 2020

ChIP-seq downstream analysis

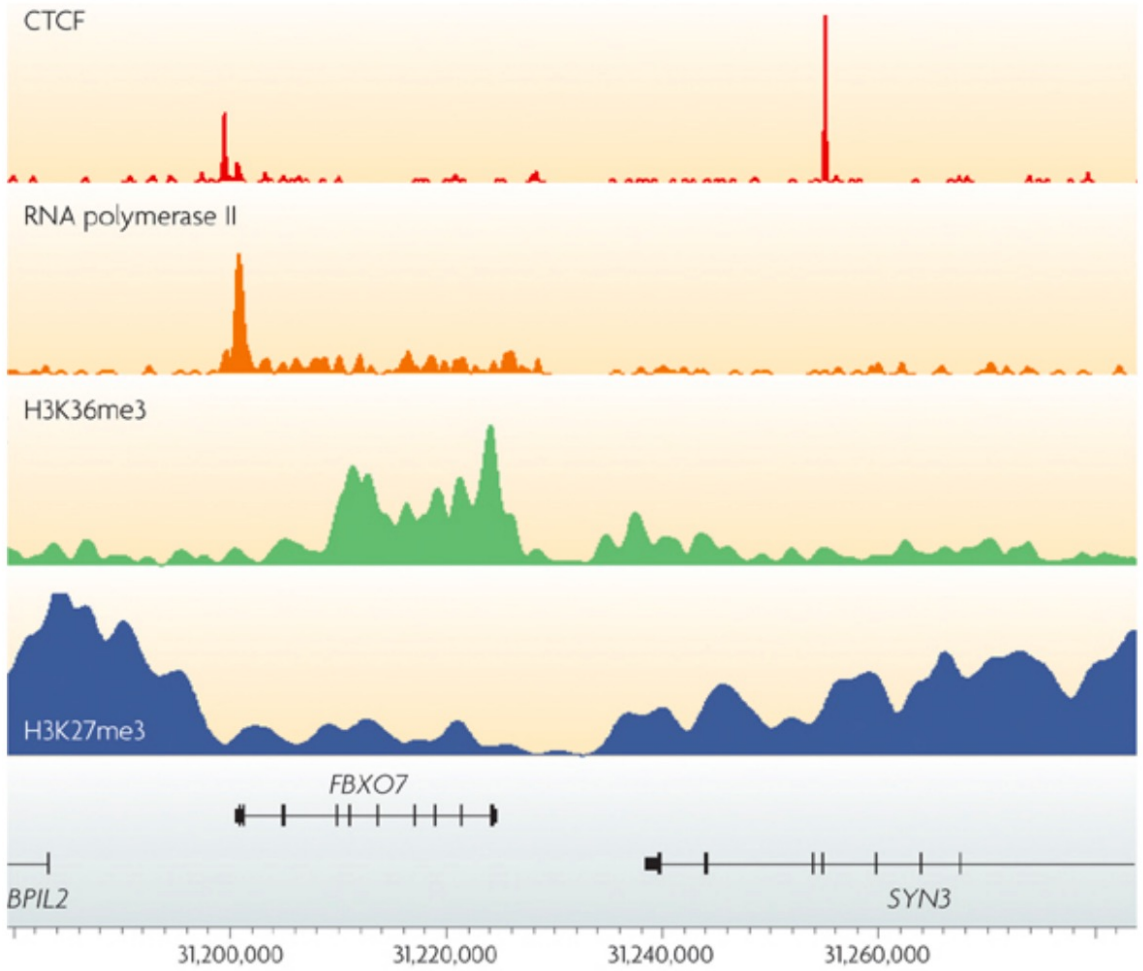
- ✓ Get .bcl files 
- ✓ Create fastq files  Or **`bcl2fastq`**
- ✓ QC: remove/trim low quality reads — E.g. **`cutadapt`**
- ✓ Align fastq to BAM — E.g. **`bowtie2`**
- ✓ Filter duplicates, artifacts, … — E.g. **`samtools`**
- ✓ Generate tracks — E.g. **`deepTools`**
- ● Assay-specific downstream analysis

Peak calling
Differential peak analysis
Motif finding
Peak co-occurrence analysis

ChIP-seq outputs

Genome-wide coverage tracks

- Can be directly viewed in a genome browser

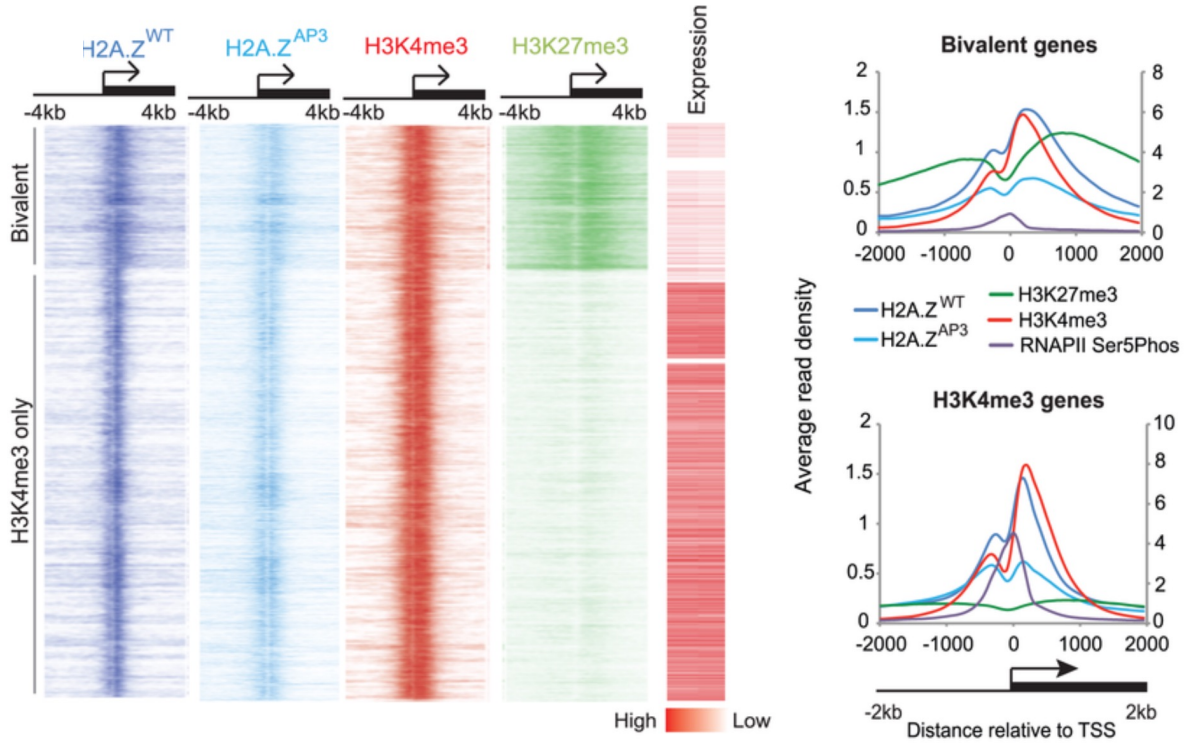


Park et al., Nat. Rev. Genet. 2009

ChIP-seq outputs

Genome-wide coverage tracks

- Can be directly viewed in a genome browser
- Can also be aligned at genomic features of interest (e.g. TSSs)



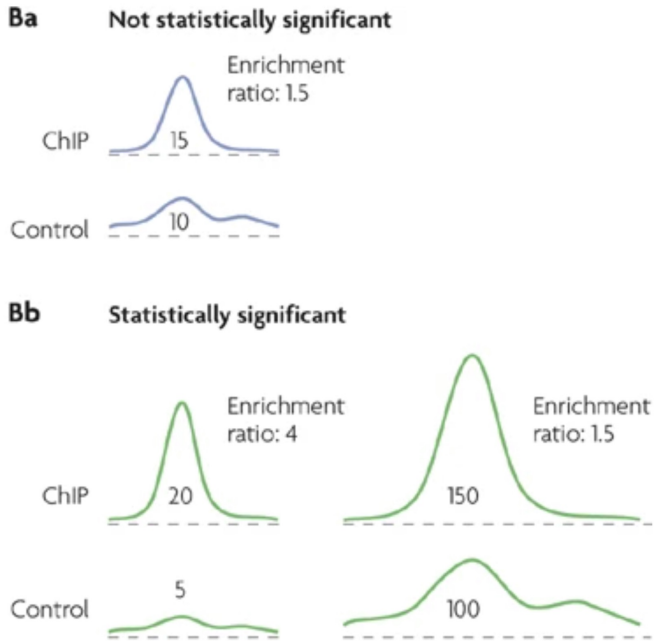
Subramanian et al., PLoS Genetics 2013

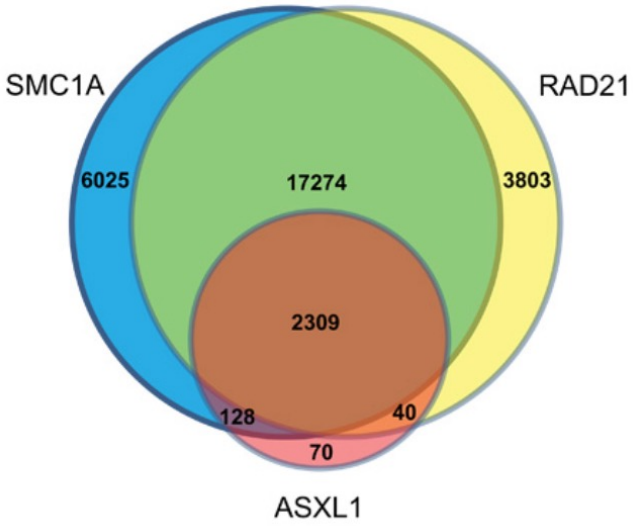
ChIP-seq outputs

Genome-wide coverage tracks

- Can be directly viewed in a genome browser
- Can also be aligned at genomic features of interest (e.g. TSSs)

Peak sets





Subramanian et al., PLoS Genetics 2013

ChIP-seq outputs

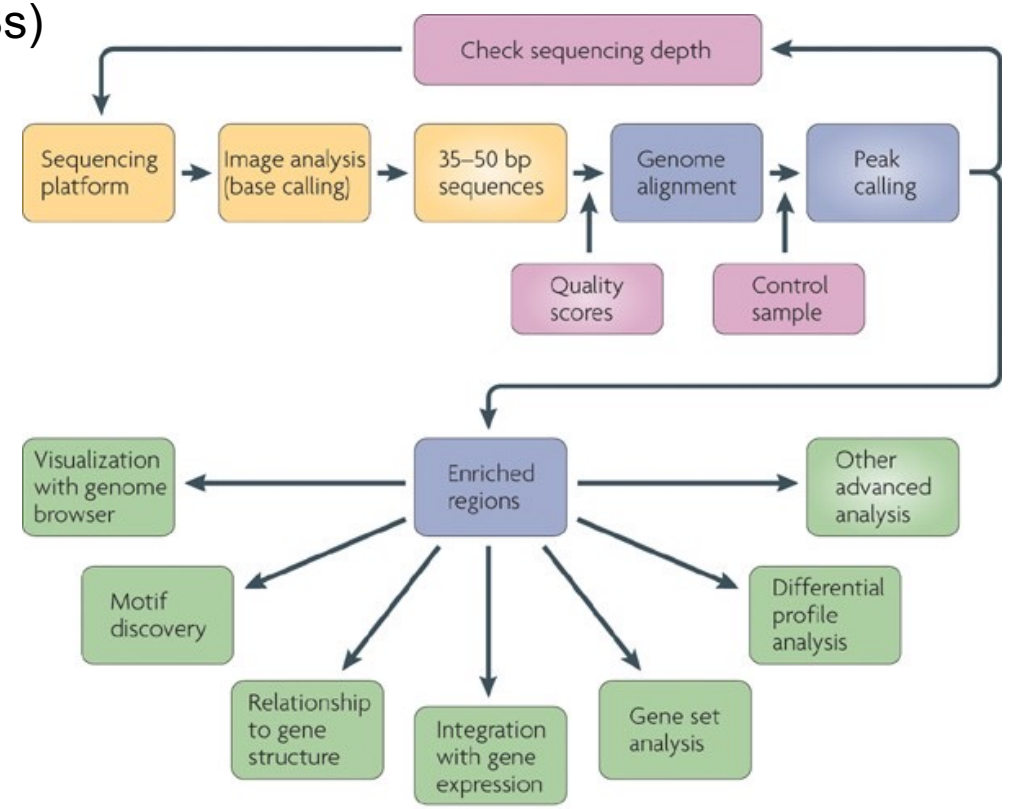
Genome-wide coverage tracks

- Can be directly viewed in a genome browser
- Can also be aligned at genomic features of interest (e.g. TSSs)

Peak sets

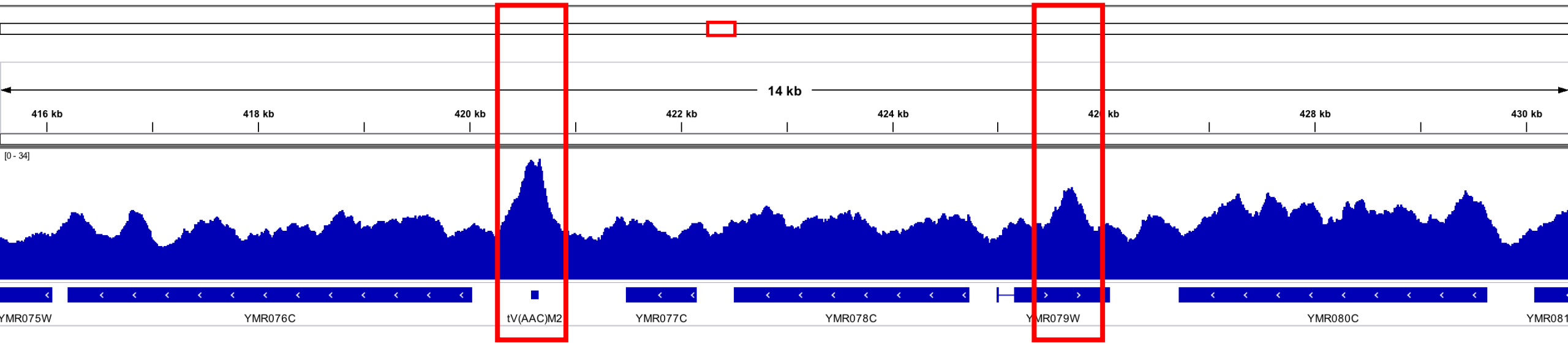
And more from downstream analyses, according to the biological question:

- DNA binding motifs
- Differential binding of a factor
- Cooperation between factors
- Biological functions in which a factor is involved
- Temporal dynamics of factor binding



What are “peaks”?

ChIP-seq libraries show uneven genomic coverage: loci with high local coverage compared to neighboring environment are “peaks”.



Inherent ChIP-seq artefacts

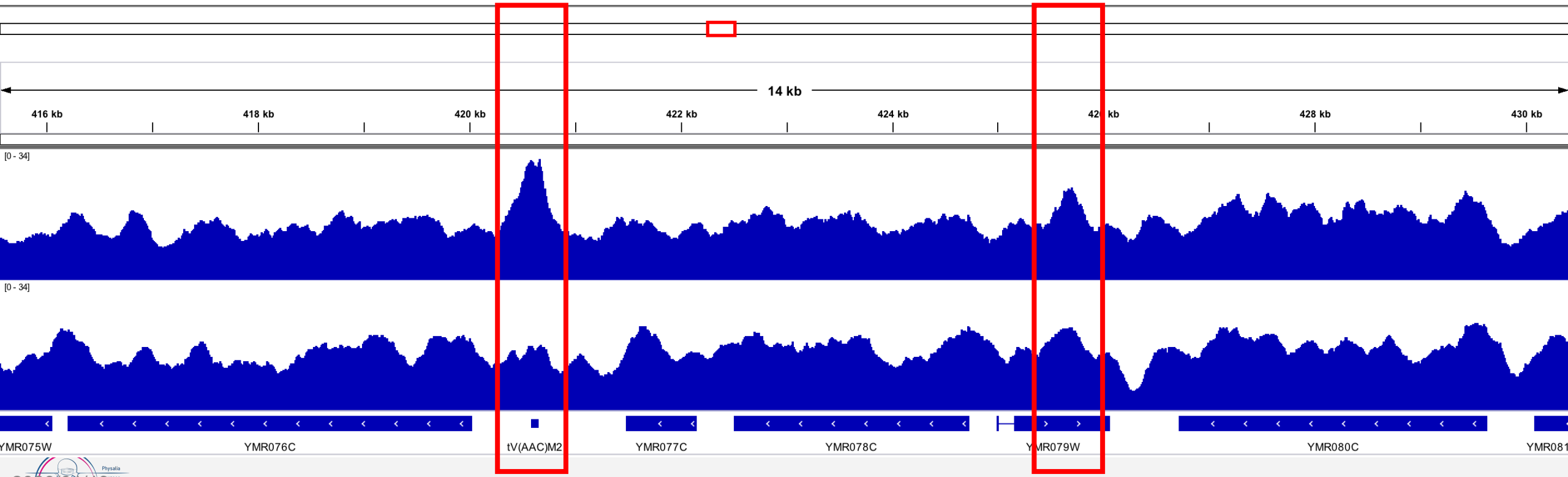
Potential sources of artefacts in ChIP-seq experiments are:

- DNA shearing: not uniform across genome, which results in more reads in open chromatin regions.
- Amplification bias (GC content)
- Repetitive regions might appear enriched due to underestimated repeat copies in the reference genome
- Sequencing depth may be too low, resulting in noisy peaks

This impedes straightforward identification of peaks in ChIP-seq data

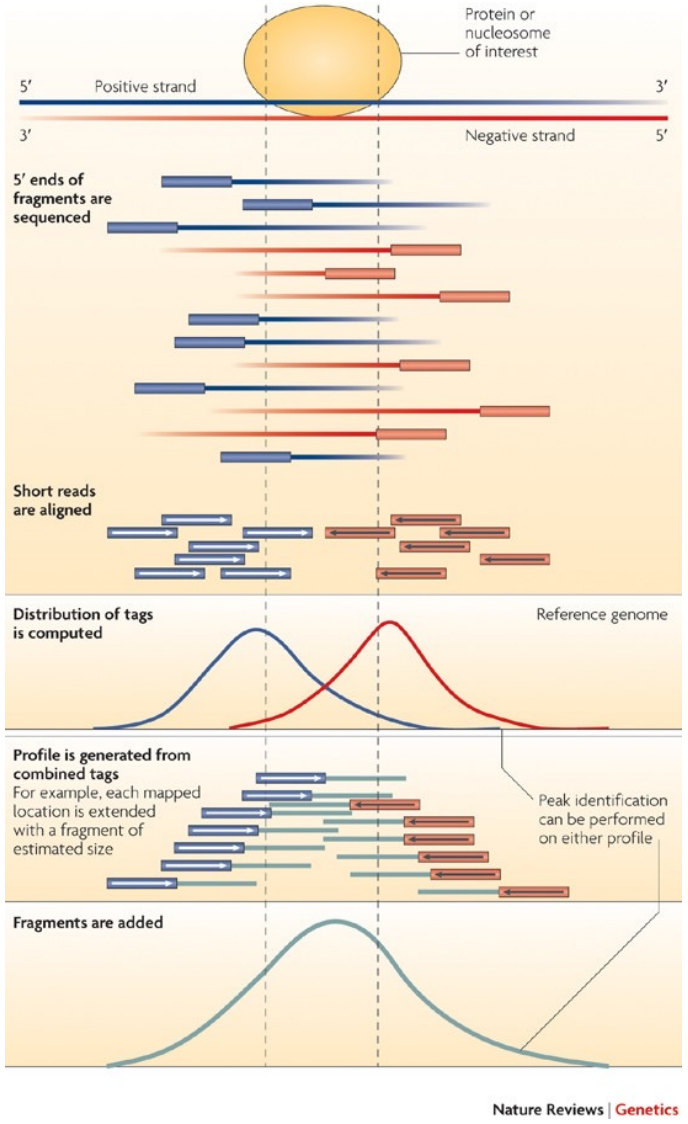
Dealing with ChIP-seq artefacts: using an “input” sample

Input control: DNA is isolated from cells that have been cross-linked and fragmented under the same conditions as the immunoprecipitated DNA



Finding peaks in ChIP-seq (1)

General workflow relies on comparing local read coverage to the input

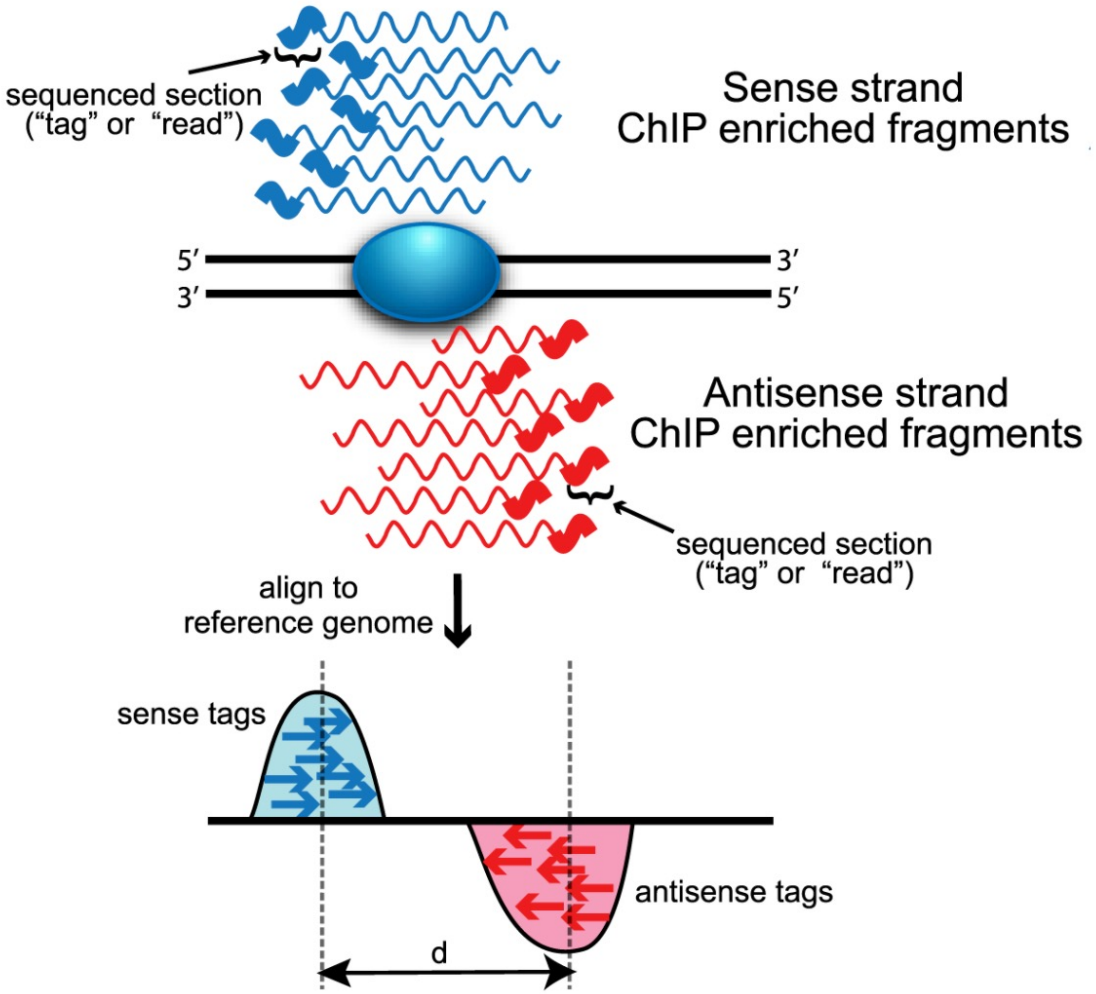


Park, Nat. Rev. Genetics 2009

Finding peaks in ChIP-seq (2)

MACS2 finds a model estimating how to **shift** sense and antisense reads towards a central position

Note: this step is specific to single-end libraries, as paired-end libraries do not have this bias.



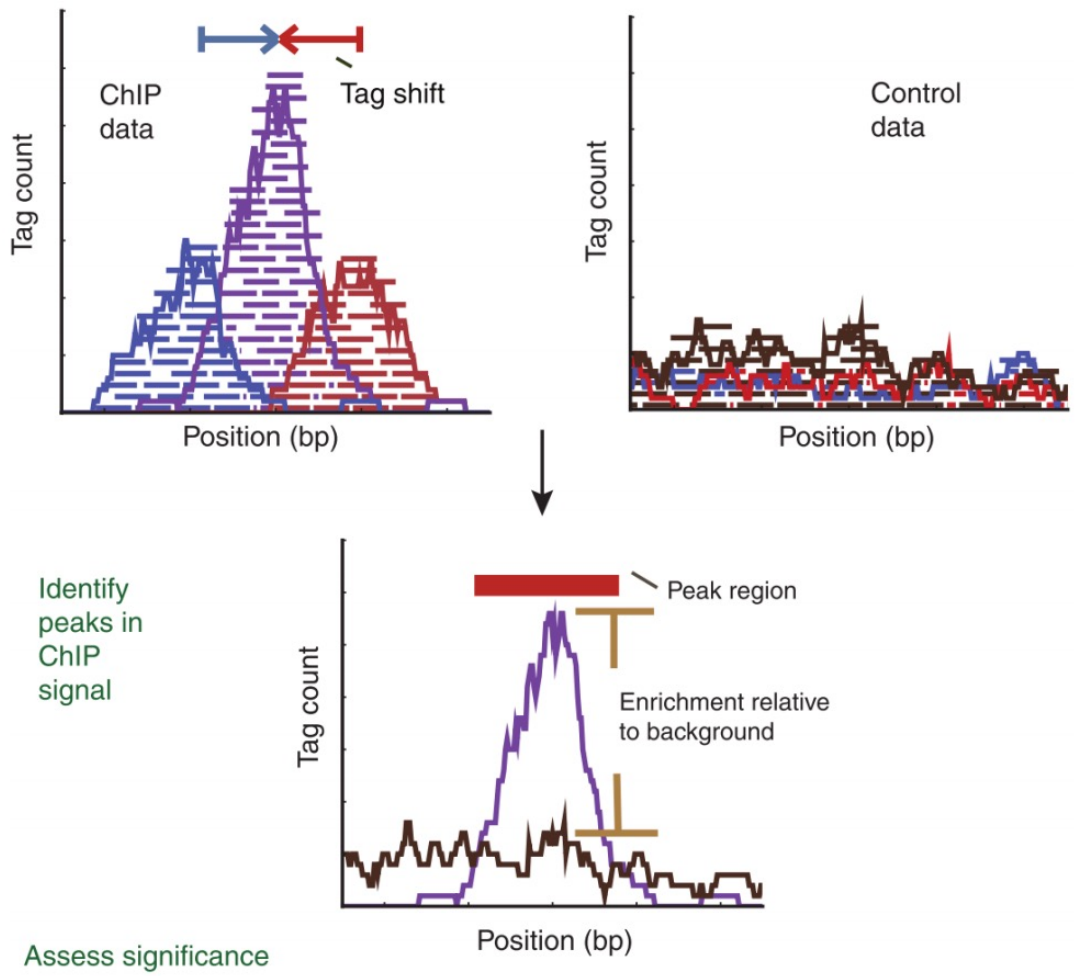
Wilbanks and Facciotti, PLoS One 2010

Finding peaks in ChIP-seq (3)

Then MACS2 scans the genome again using a window size which is twice the fragment length.

For each peak, MACS2 calculates a p-value using a dynamic Poisson distribution to capture local biases in read background levels.

If a control sample is available, it is used to calculate the local background.



Zhang et al., Genome Biol. 2008

Finding peaks in ChIP-seq (3)

Then MACS2 scans the genome again using a window size which is twice the fragment length.

For each peak, MACS2 calculates a p-value using a dynamic Poisson distribution to capture local biases in read background levels.

```
macs2 callpeak --treatment ….bam
```

Finding peaks in ChIP-seq (3)

Then MACS2 scans the genome again using a window size which is twice the fragment length.

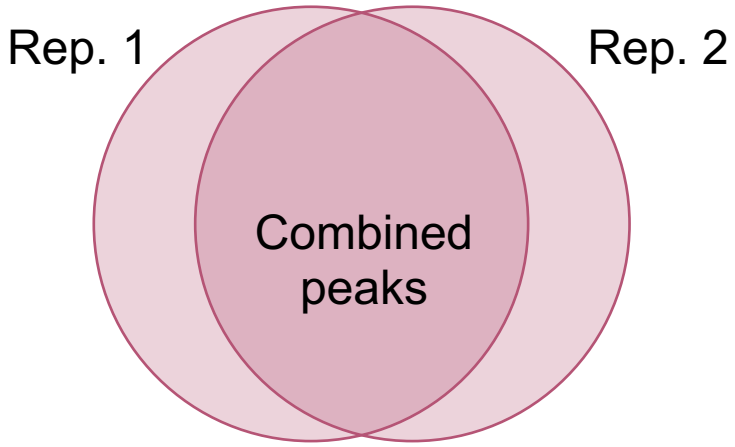
For each peak, MACS2 calculates a p-value using a dynamic Poisson distribution to capture local biases in read background levels.

If a control sample is available, it is used to calculate the local background.

```
macs2 callpeak --treatment ….bam --control ….bam
```

Using replicates to peaks

The easiest approach it to take overlapping peak calls across replicates.

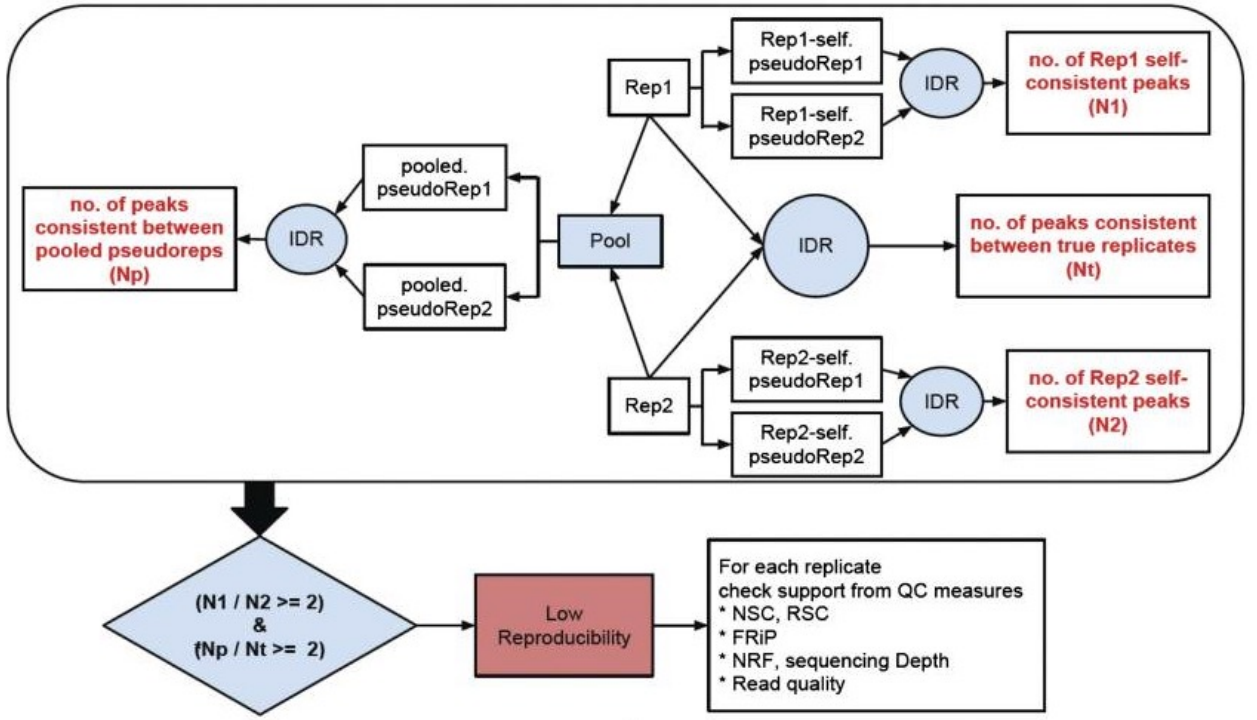


Using replicates to peaks

The easiest approach it to take overlapping peak calls across replicates.

For more advanced users, there are more complex methods that employ statistical testing and evaluate the reproducibility between replicates.

The standard approach to leverage replicates is the IDR (Irreproducibility Discovery Rate) approach.



Qi et al., The Annals of Applied Statistics 2011

Using replicates to peaks

The easiest approach it to take overlapping peak calls across replicates.

For more advanced users, there are more complex methods that employ statistical testing and evaluate the reproducibility between replicates.

The standard approach to leverage replicates is the IDR (Irreproducibility Discovery Rate) approach.

```
idr --samples rep1.narrowPeak rep2.narrowPeak \
  --input-file-type narrowPeak \
  --rank p.value \
  --output-file idr \
  --plot \
  --log-output-file idr.log
```

Qi et al., The Annals of Applied Statistics 2011