Mapping peaks in ChIP-seq data

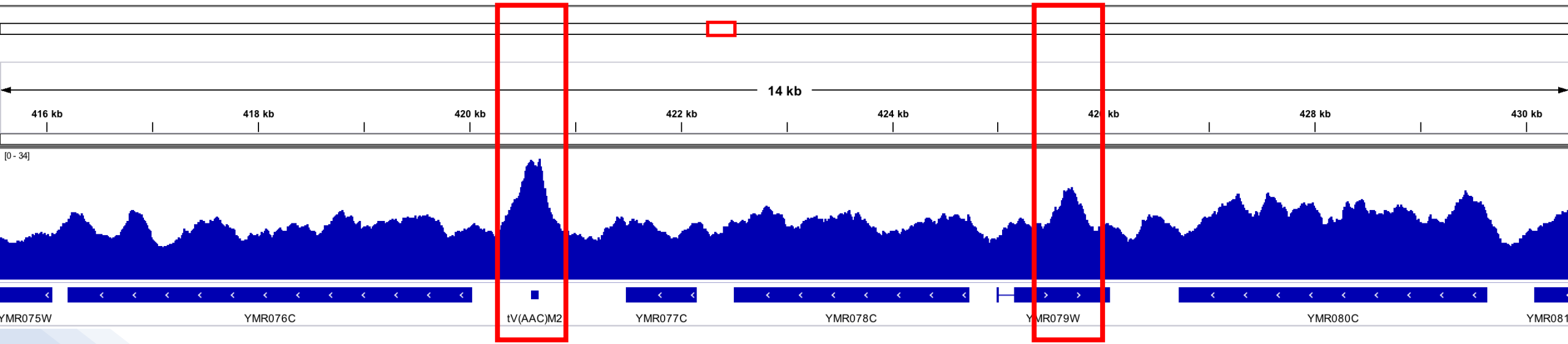
NGS analysis for gene regulation and epigenomics

Physalia 2021

Jacques Serizay

What are “peaks”?

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

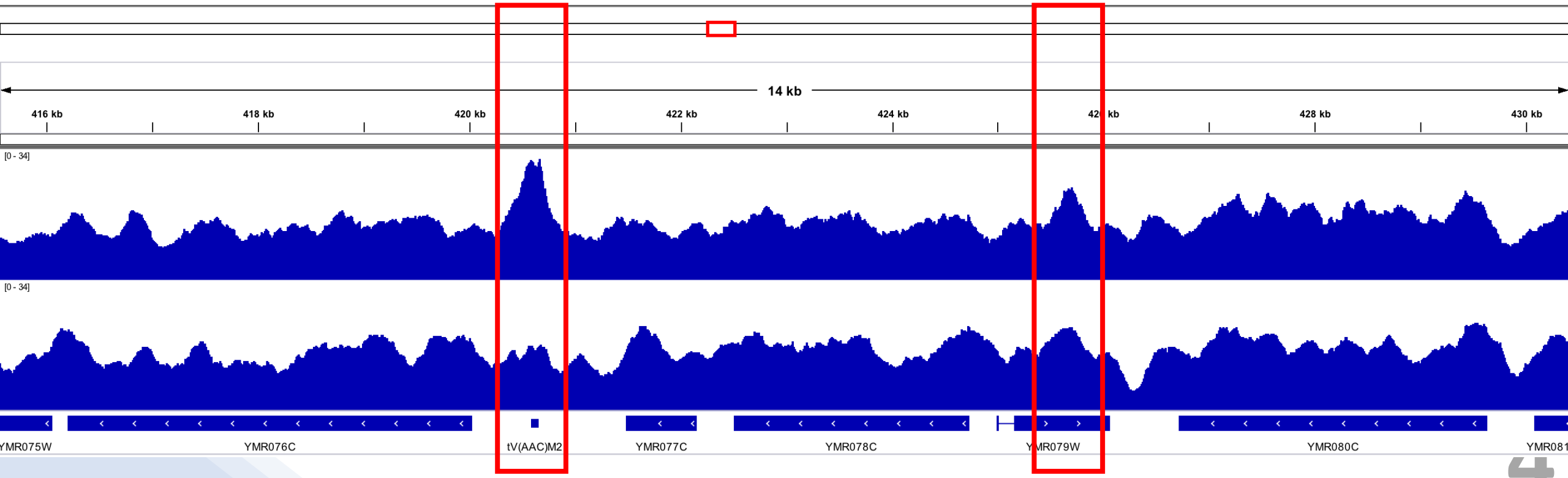


Inherent ChIP-seq artefacts

- 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

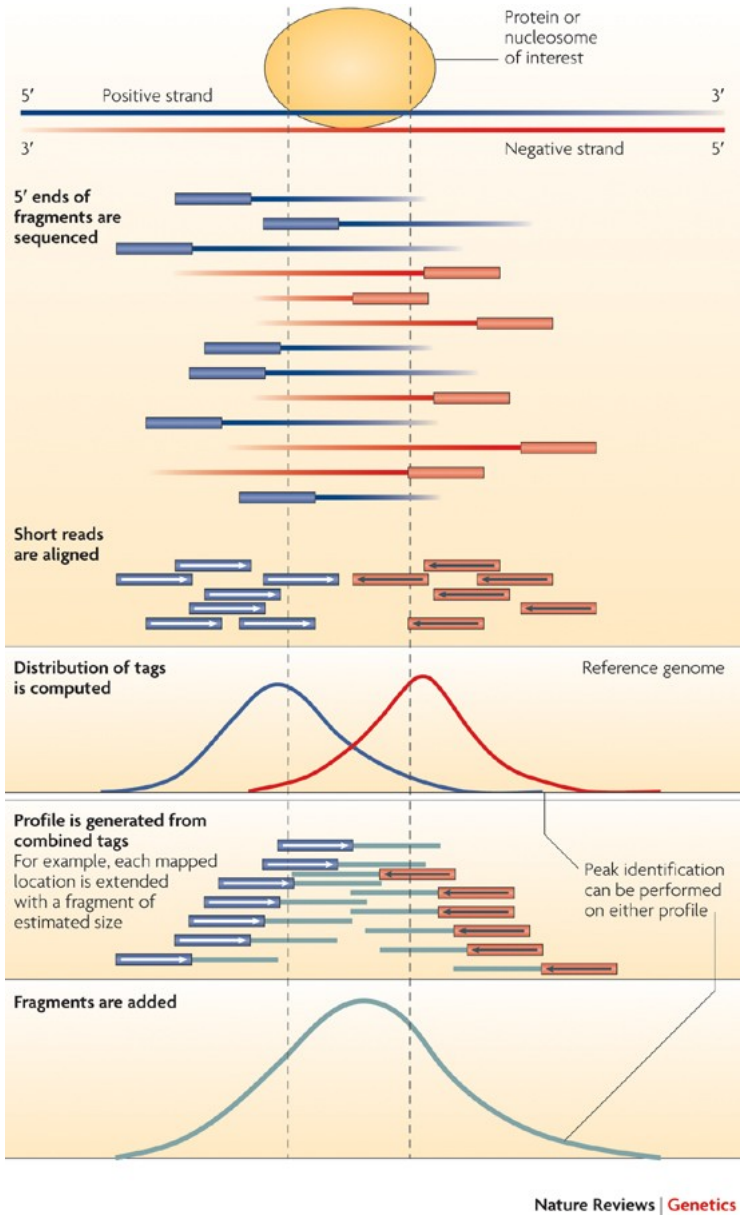
Dealing with ChIP-seq artefacts: using an “input” sample

- 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

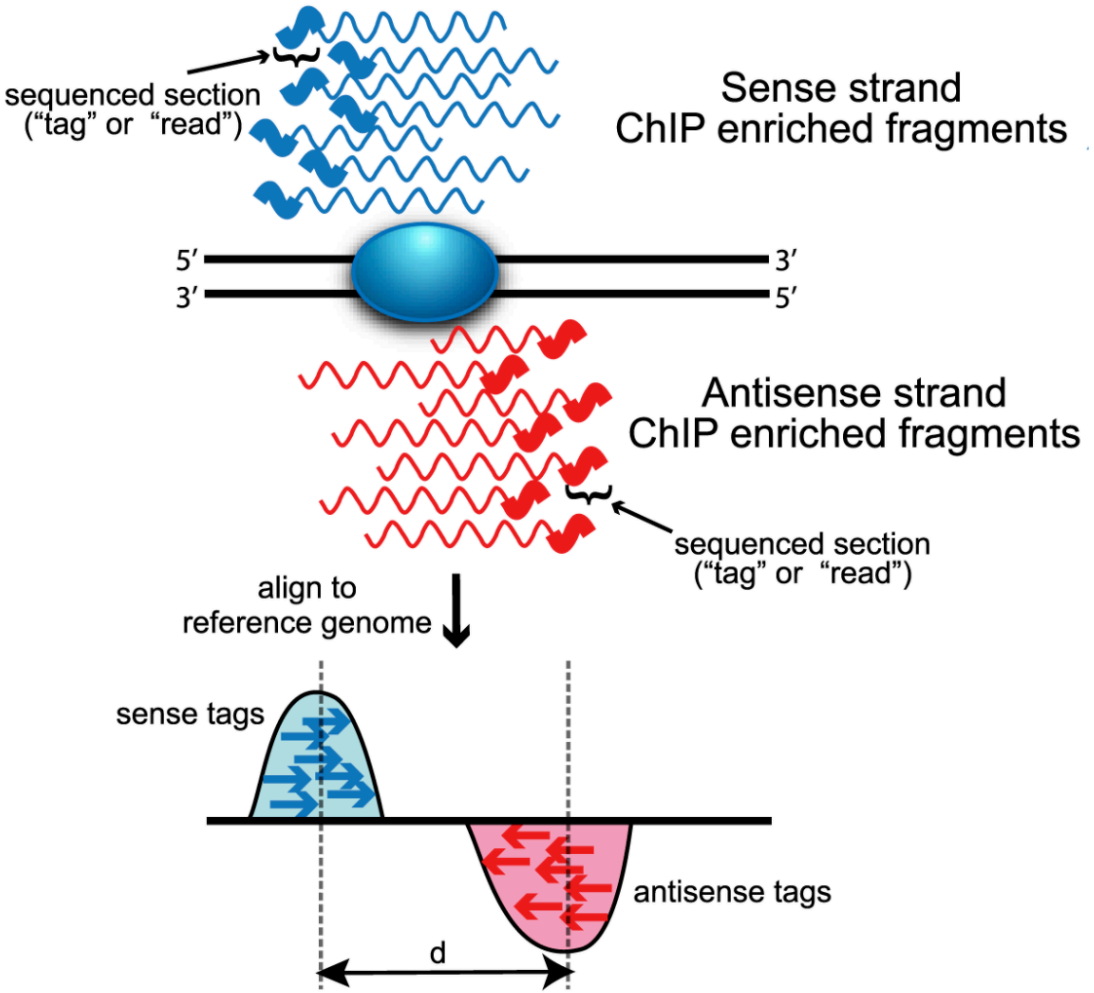


Park, Nat. Rev. Genetics 2009

Finding peaks in ChIP-seq (2)

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

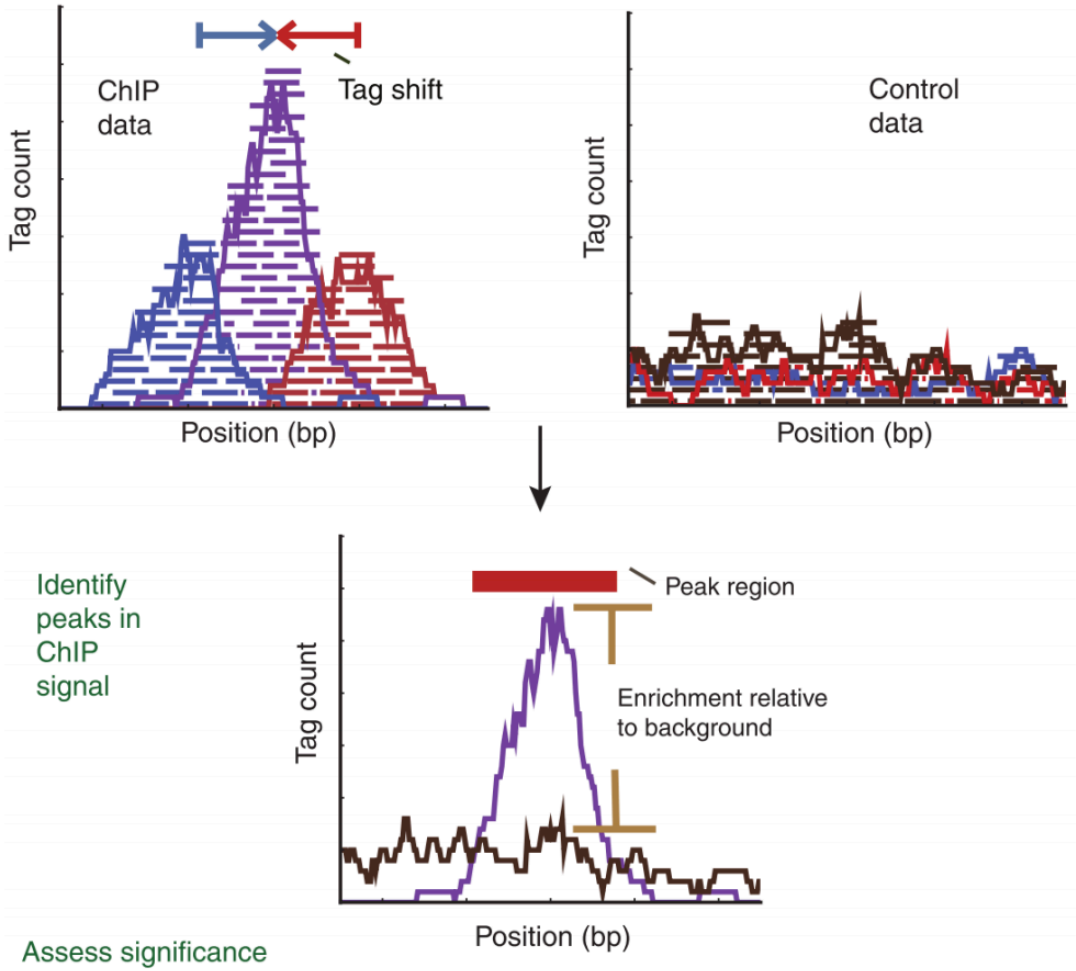
Note: this step is specific to single-end libraries, as paired-end libraries do not have this bias.



Wilbanks and Facciotti, PLoS One 2010

Finding peaks in ChIP-seq (3)

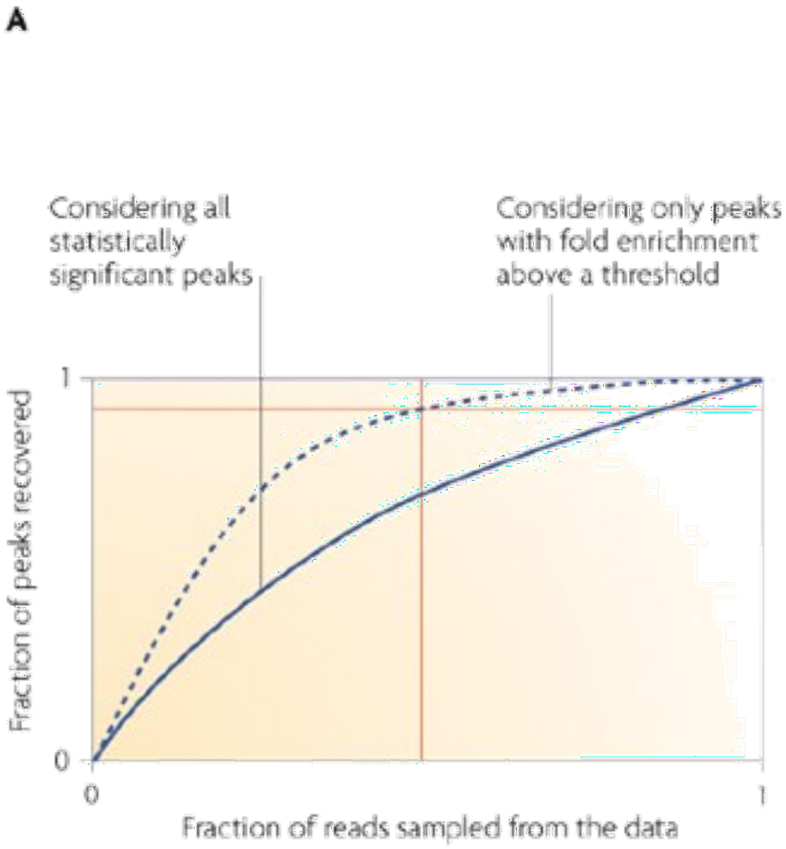
- Then MACS2 scans the genome again using a window size which is twice the fragment length.
- For each peak, MACS2 calculates a p-value using a dynamic Poisson distribution to capture local biases in read background levels.
- If a control sample is available, it is used to calculate the local background.

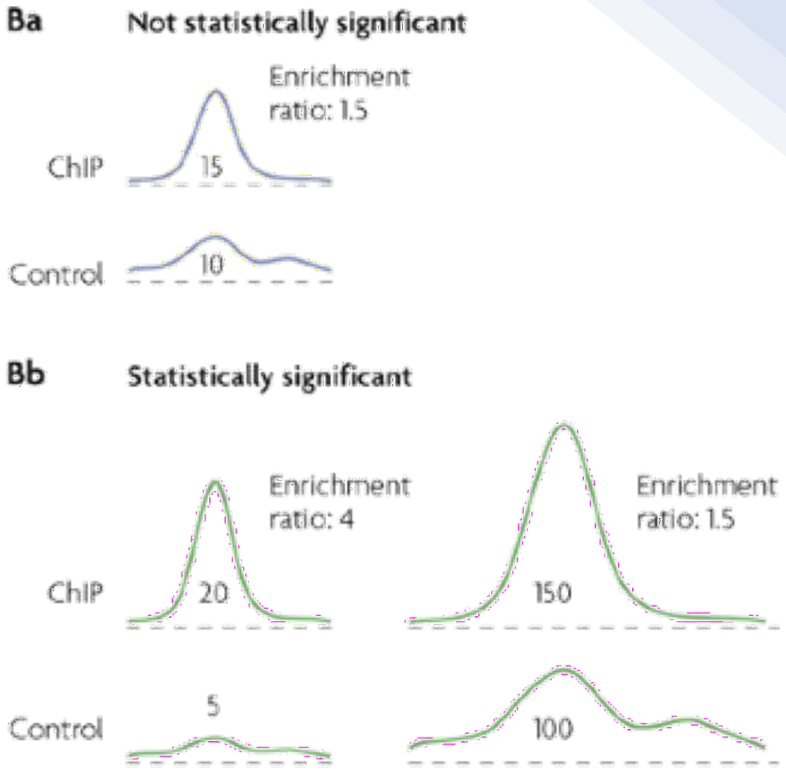


Zhang et al., Genome Biol. 2008

Can we find “all” the statistically significant peaks in a ChIP-seq dataset?

- With greater sequencing depth, more peaks become statistically significant





Nature Reviews | Genetics

Park, Nat. Rev. Genetics 2009

Using replicates to peaks

- Usually, by taking overlapping peak calls across replicates
- Additionally, there are more complex methods that employ statistical testing and evaluate the reproducibility between replicates (e.g. IDR)

More on IDR here: https://hbctraining.github.io/Intro-to-ChIPseq/lessons/07_handling-replicates-idr.html

Park, Nat. Rev. Genetics 2009

Downstream analysis on peak sets

- Differential binding analysis
- Functional enrichment analysis: annotate the closest gene for each peak
- Find DNA sequences enriched

Downstream analysis on peak sets

- Differential binding analysis
- ChIP-seq datasets comparison
- Functional enrichment analysis: annotate the closest gene for each peak
- Find enriched DNA sequences

Ex. 03-2

Downstream analysis on peak sets

- Differential binding analysis
- ChIP-seq datasets comparison **Ex. 03-3**
- Functional enrichment analysis: annotate the closest gene for each peak
- Find DNA sequences enriched

Downstream analysis on peak sets

- Differential binding analysis

Ex. 04-2

- ChIP-seq datasets comparison
- Functional enrichment analysis: annotate the closest gene for each peak
- Find DNA sequences enriched

Downstream analysis on peak sets

- Differential binding analysis
- ChIP-seq datasets comparison
- Functional enrichment analysis: annotate the closest gene for each peak

Ex. 05-1

- Find DNA sequences enriched