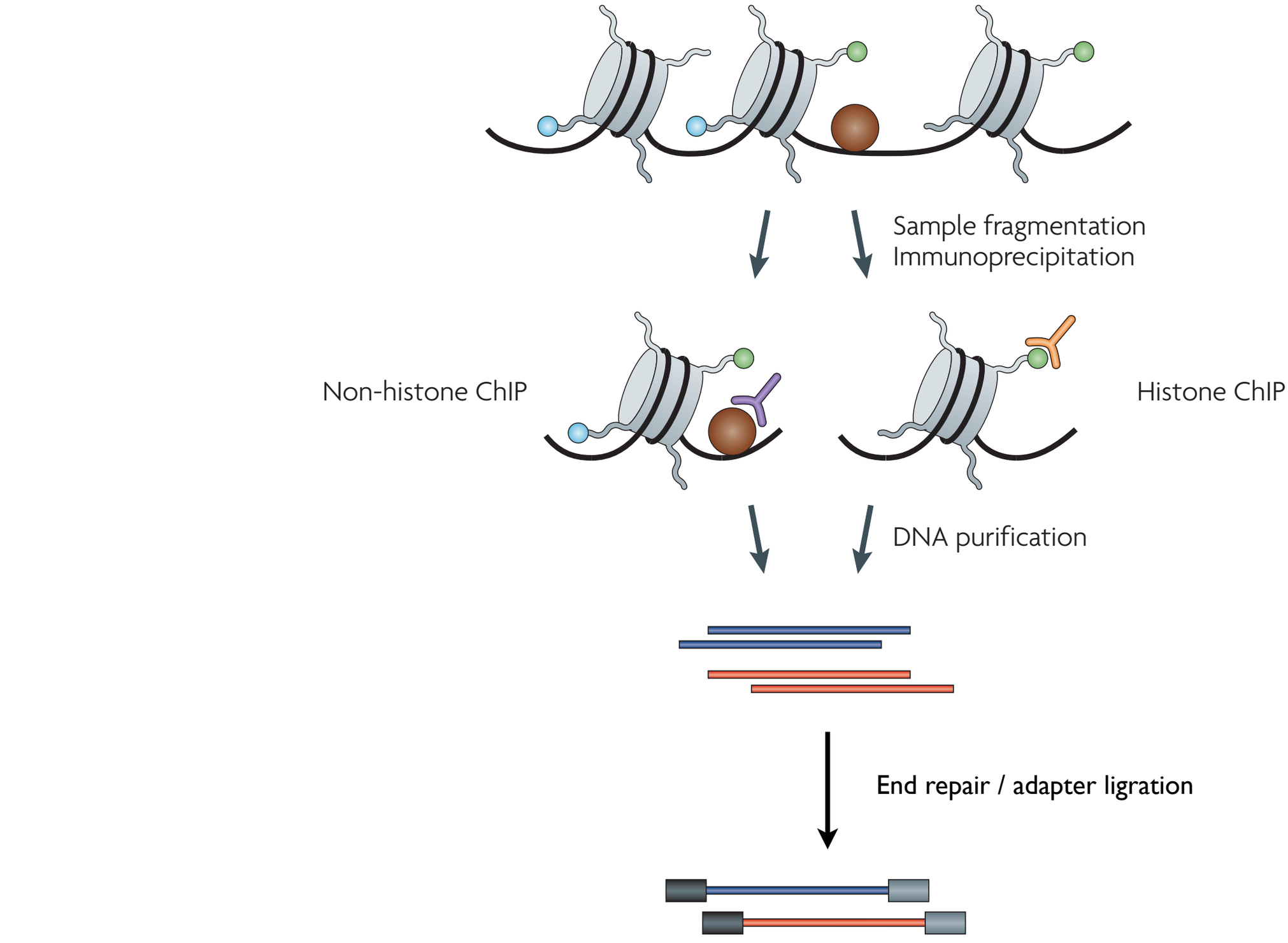
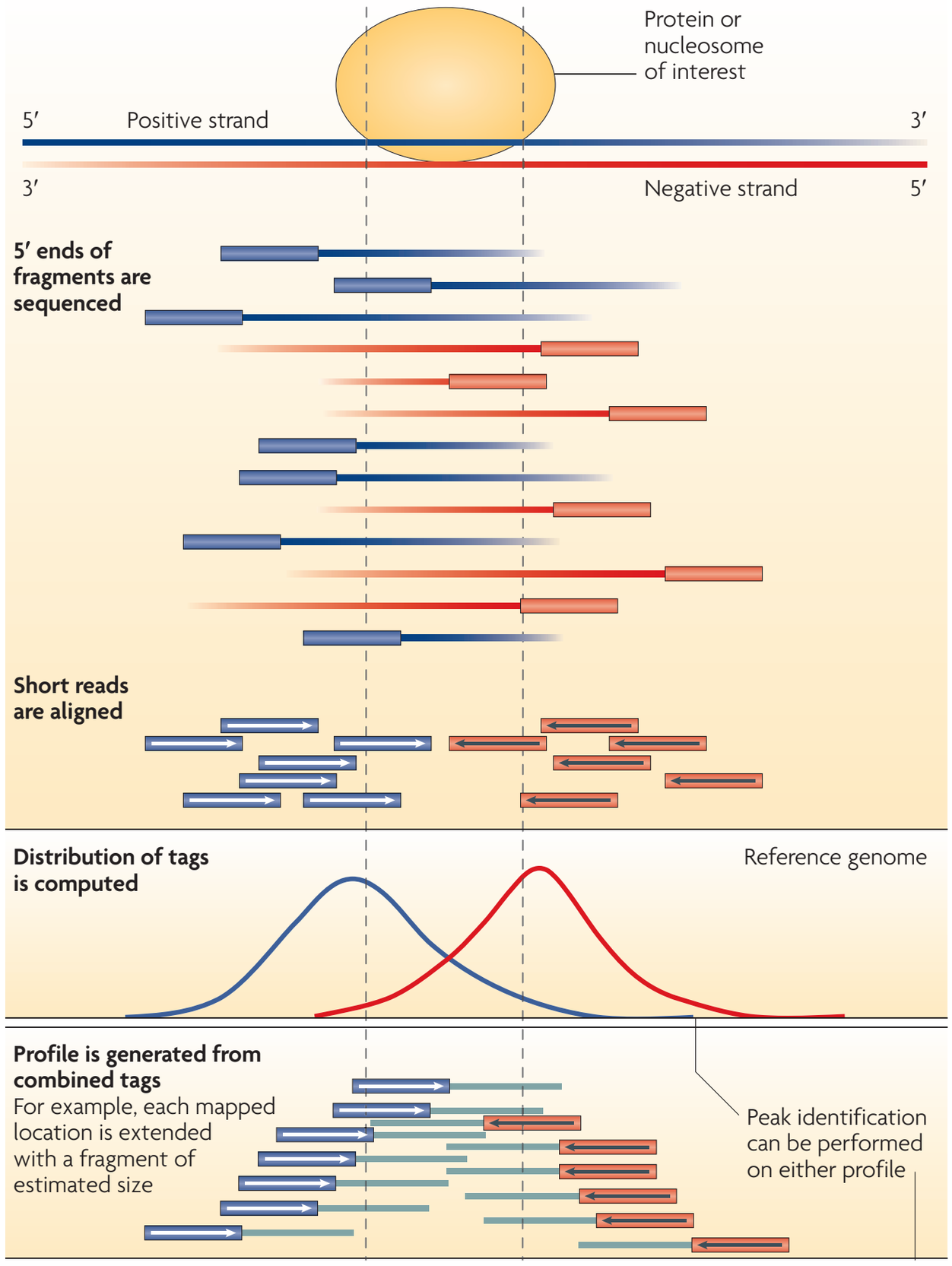
Galaxy

ChIP-Seq analysis with MACS

www.galaxyproject.org



(Park, *Nature Reviews Genetics*, 2009)



(Park, *Nature Reviews Genetics*, 2009)

Model-based Analysis of ChIP-Seq (MACS)

(Zhang et al. *Genome Biology*, 2008)

- Tag distributions represent the end of the sequenced fragment. A shift is necessary to resolve the center of the bound region (peak)
- MACS empirically models the amount of this shift to better determine the bound region

MACS Fragment Shift

- Given a sonication size (b) and fold enrichment (f), find windows of size 2b that are f-fold enriched
- Sample 1000 of these high-quality windows, separate tags by strands, and find distance between the modes to estimate d
- Shift all tags by d/2 in the 3’ direction

ChIP-Seq Analysis: Get the Data

Shared Data → Data Libraries →

Demonstration Datasets

Select everything in the

Mouse ChIP-seq: G1E CTCF Binding

folder

(We're ignoring quality control,

in practice this would be a good time for FASTQC)

ChIP-Seq Exercise: Mapping with Bowtie

Use Bowtie2 (could also use BWA)

NGS Mapping: → Bowtie2

FASTQ file → G1E_CTCF (chr19)

ChIP-Seq Analysis: Find Peaks

NGS: Peak Calling → MACS

Experiment name → MACS G1E_CTCF

Tag File → G1E_CTCF BAM file

Tag Size → 36

Leave MFOLD → 32

Check Perform the new peak detection method (futuredir)

ChIP-Seq Analysis: Results

Look at the HTML report dataset

Potential ChIP Biases

- Problems:
  - Chromatin accessibility affects fragmentation
  - Amplification bias
  - Repetitive regions
- Solution: Controls
  - Input DNA (after fragmentation but before IP)
  - Non-specific IP

MACS peak detection

- After shift, slide windows of size 2d across genome
- Model tag count for windows as a poisson distribution, and calculate a p-value for each window
- For the λ parameter (~expected number of tags per window), estimate from sample or control if available
    - Estimates for local windows of size 1kb, 5kb, 10kb or the whole genome and uses the max

ChIP-Seq Analysis: Results

Rerun with control:

NGS Mapping: → Bowtie2 on G1E_Input

NGS: Peak Calling → MACS on the resulting mapped reads

Summary

MACS is one tool, available in Galaxy, for analysis of ChIP-seq data

Controls are extremely important for accurately calling ChIP-seq peaks

As for most genomics problems, there are other tools that may be more appropriate depending on the type of data, for example SICER for broad histone modifications