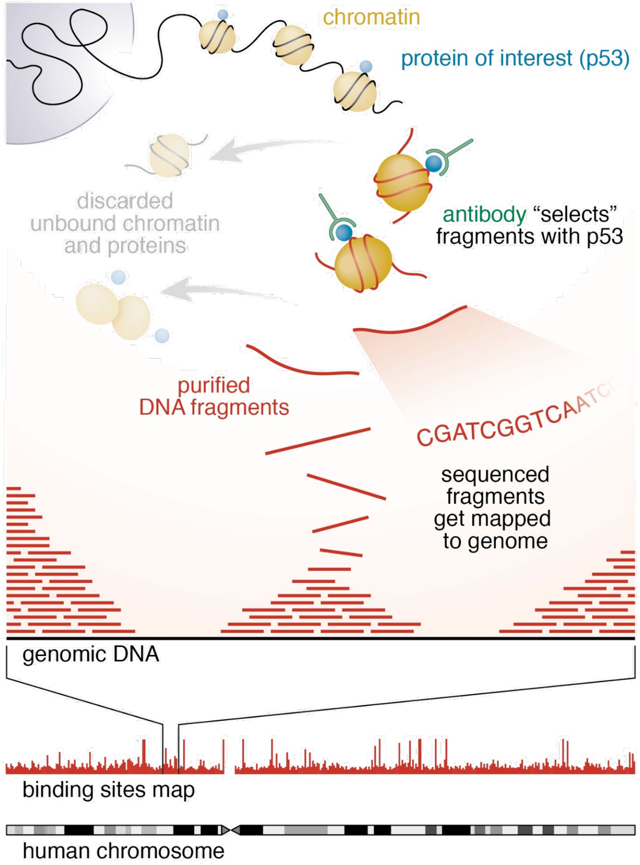
ChIP-sequencing: Library preparation and experimental design

Stéphanie Le Gras
(slegras@igbmc.fr)

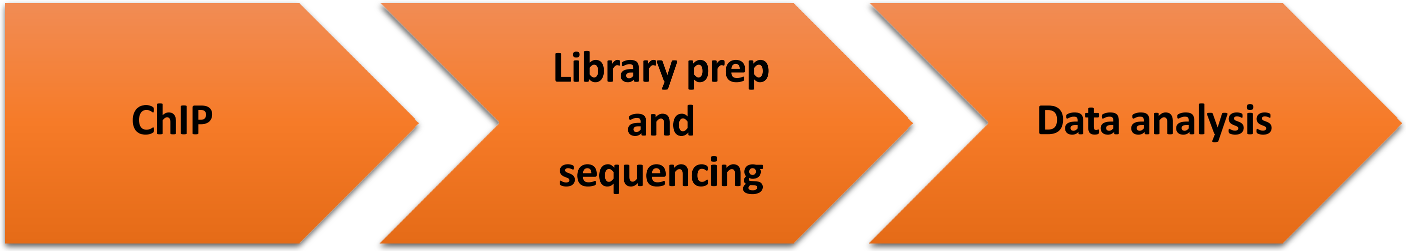
ChIP-seq

- Johnson et al, 2007
- Alternative to ChIP-on-chip (hybridization technique)
- ChIP-seq combines chromatin immunoprecipitation and sequencing to analyze protein interactions with DNA
- It can be used to detect and analyze
  - Binding sites of various proteins bound to DNA such as transcription factors (TF ChIP-seq)
  - Position of histone post translational modifications (Histone ChIP-seq)
  - Nucleotide modification such as methylation (MeDIP-seq)
- Expected results of a ChIP-seq experiment are genomic regions with significant sequenced read enrichments (also called peaks)


chromatin
protein of interest (p53)
discarded unbound chromatin and proteins
antibody "selects" fragments with p53
purified DNA fragments
CGATCGGTCAATC
sequenced fragments get mapped to genome
genomic DNA
binding sites map
human chromosome

https://hbctraining.github.io/Intro-to-ChIPseq/lessons/01_Intro_chipseq_data_organization.html

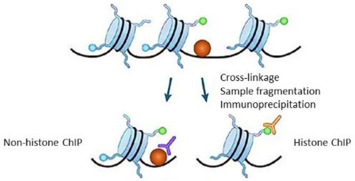
ChIP-seq process

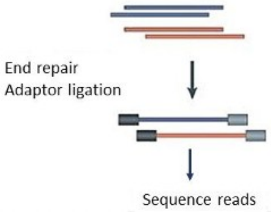


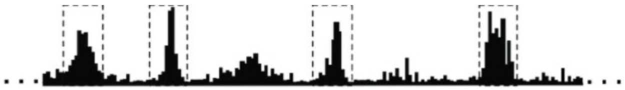
ChIP

Library prep and sequencing

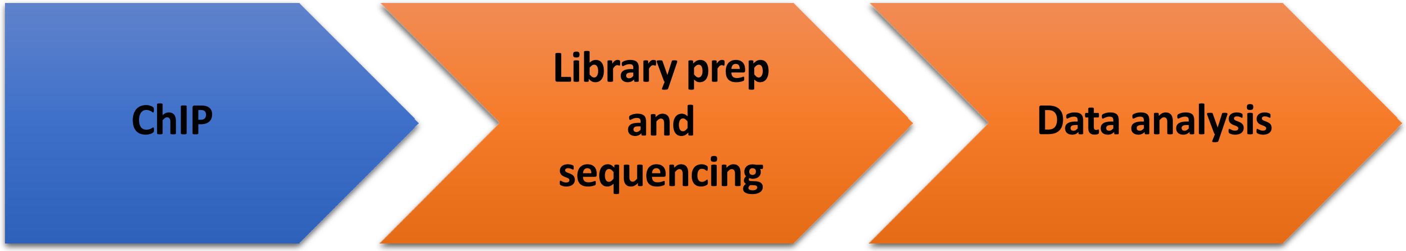
Data analysis







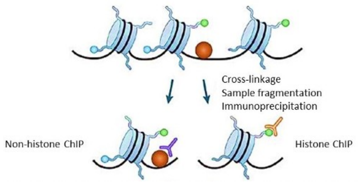
Chromatin ImmunoPrecipitation

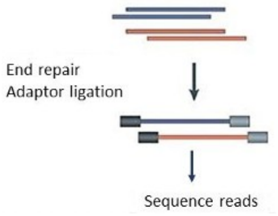


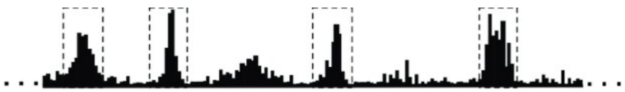
ChIP

Library prep and sequencing

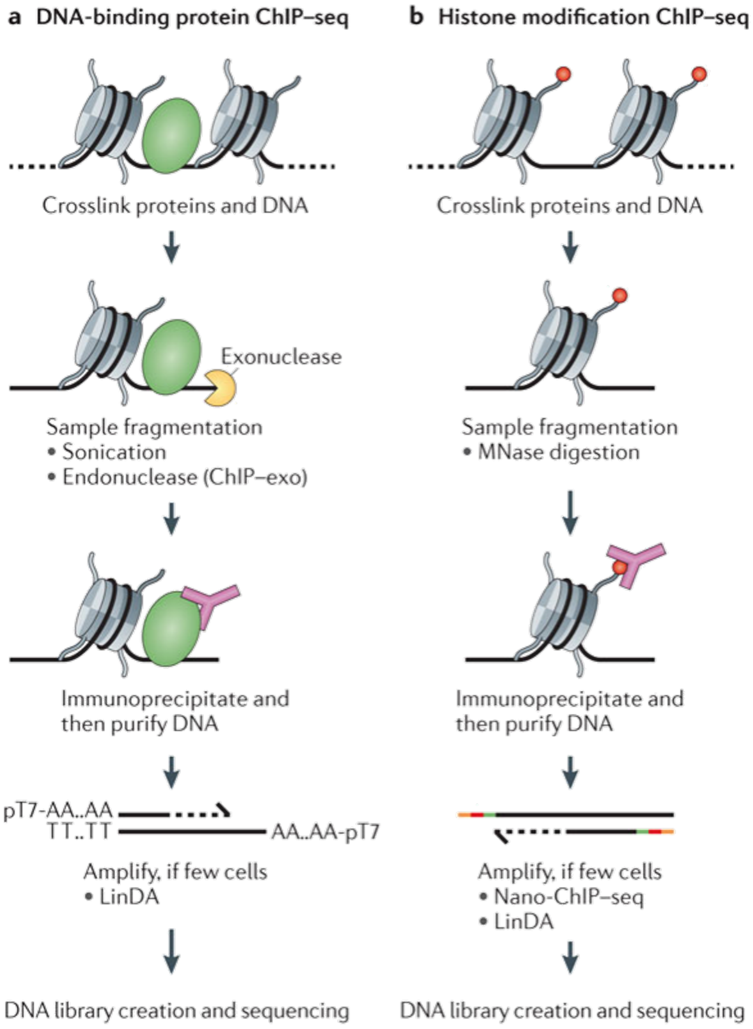
Data analysis







Chromatin ImmunoPrecipitation


a DNA-binding protein ChIP–seq
Crosslink proteins and DNA
Exonuclease
Sample fragmentation
• Sonication
• Endonuclease (ChIP–exo)
Immunoprecipitate and then purify DNA
pT7-AA..AA
TT..TT
AA..AA-pT7
Amplify, if few cells
• LinDA
DNA library creation and sequencing
b Histone modification ChIP–seq
Crosslink proteins and DNA
Sample fragmentation
• MNase digestion
Immunoprecipitate and then purify DNA
Amplify, if few cells
• Nano-ChIP–seq
• LinDA
DNA library creation and sequencing

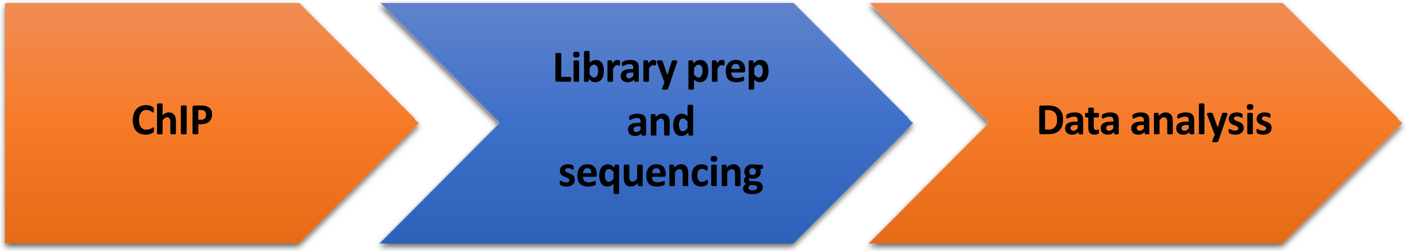
Nature Reviews Genetics 13, 840-852 (December 2012) doi:10.1038/nrg3306

Considerations on chIP

Complexity in DNA fragments ++

- Antibody
  - Antibody quality varies, even between independently prepared lots of the same antibody (Egelhofer, T. A. *et al*. 2011)
- Number of cells
  - large number of cells are required for a ChIP experiment (limitation for small organisms)
    - Nano-ChIP–seq (Adli et al, 2011)
    - LinDA (Shankaranarayanan et al, 2011)
  - Cut&tag, Cut&run
- Shearing of DNA (Mnase I, sonication, Covaris): trying to narrow down the size distribution of DNA fragments

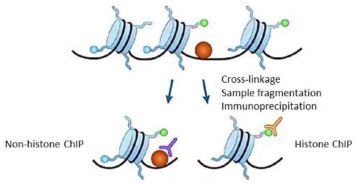
Library prep and sequencing

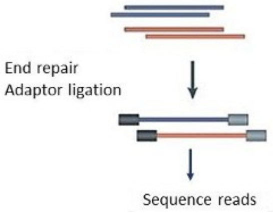


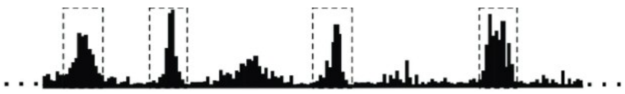
ChIP

Library prep and sequencing

Data analysis

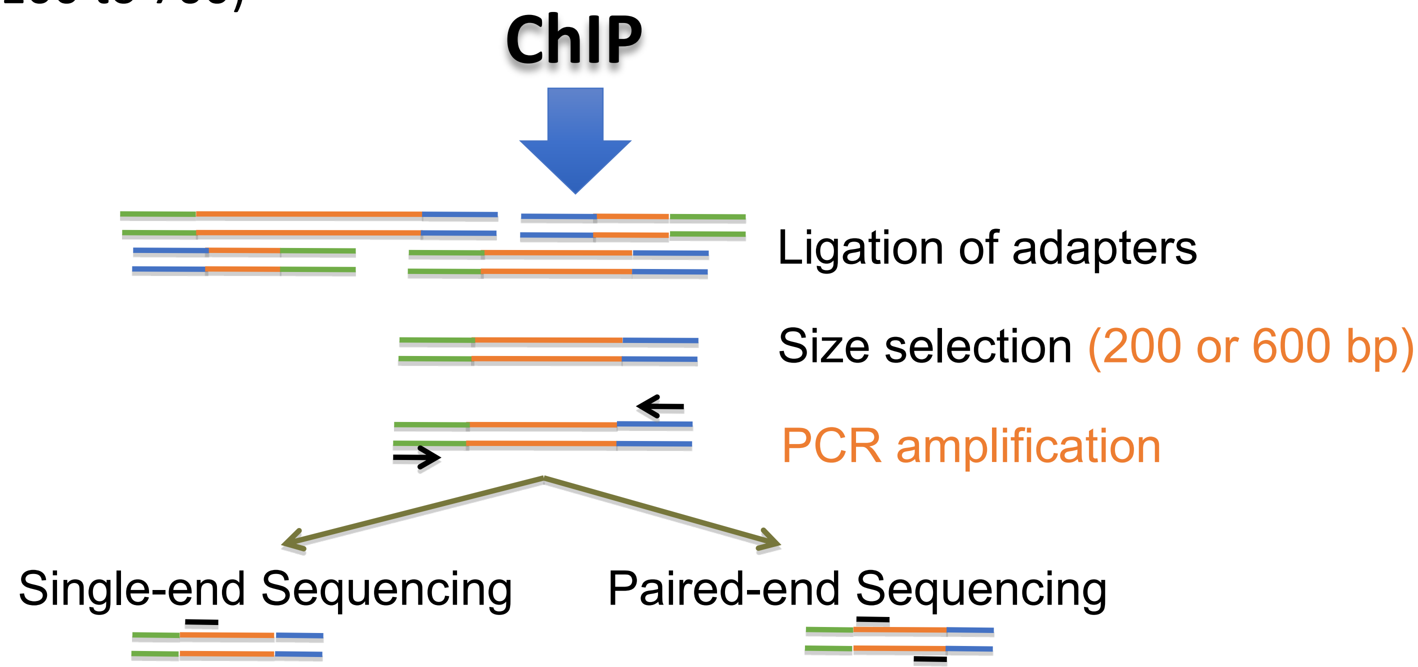






Library prep

- Step between chIP and sequencing
- The goal is to prepare DNA for the sequencing
- Starting material:
    - 2-10ng of sheared DNA (Purification using Agencourt AMPure XP - Beckman Coulter)
    - The mean DNA fragment size should be below 500bp (Smear in a range of 100 to 700)



ChIP

Ligation of adapters

Size selection (200 or 600 bp)

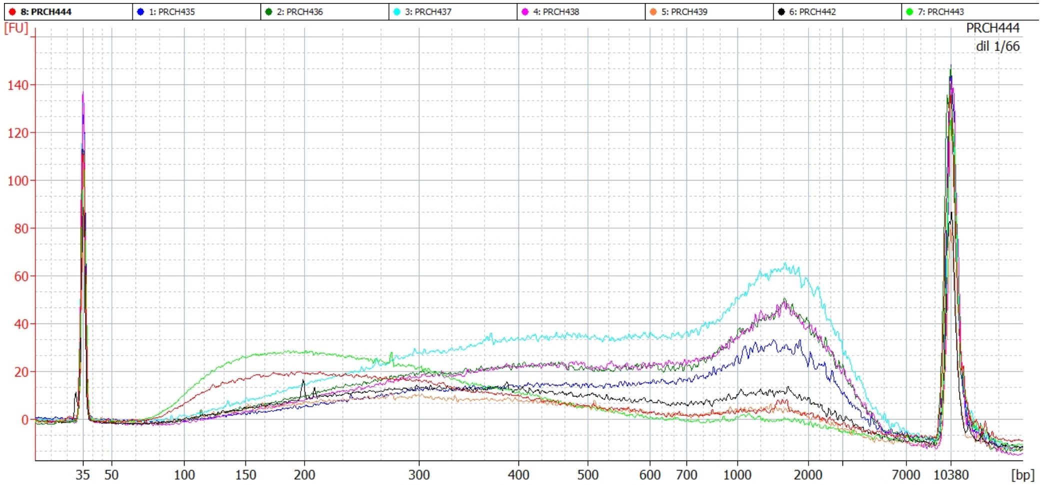
PCR amplification

Single-end Sequencing

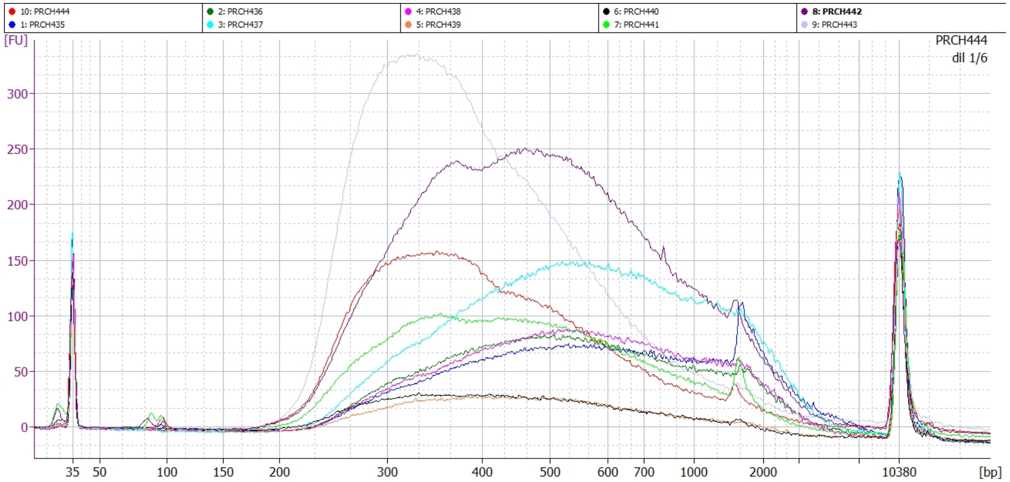
Paired-end Sequencing

DNA shearing

Before library prep

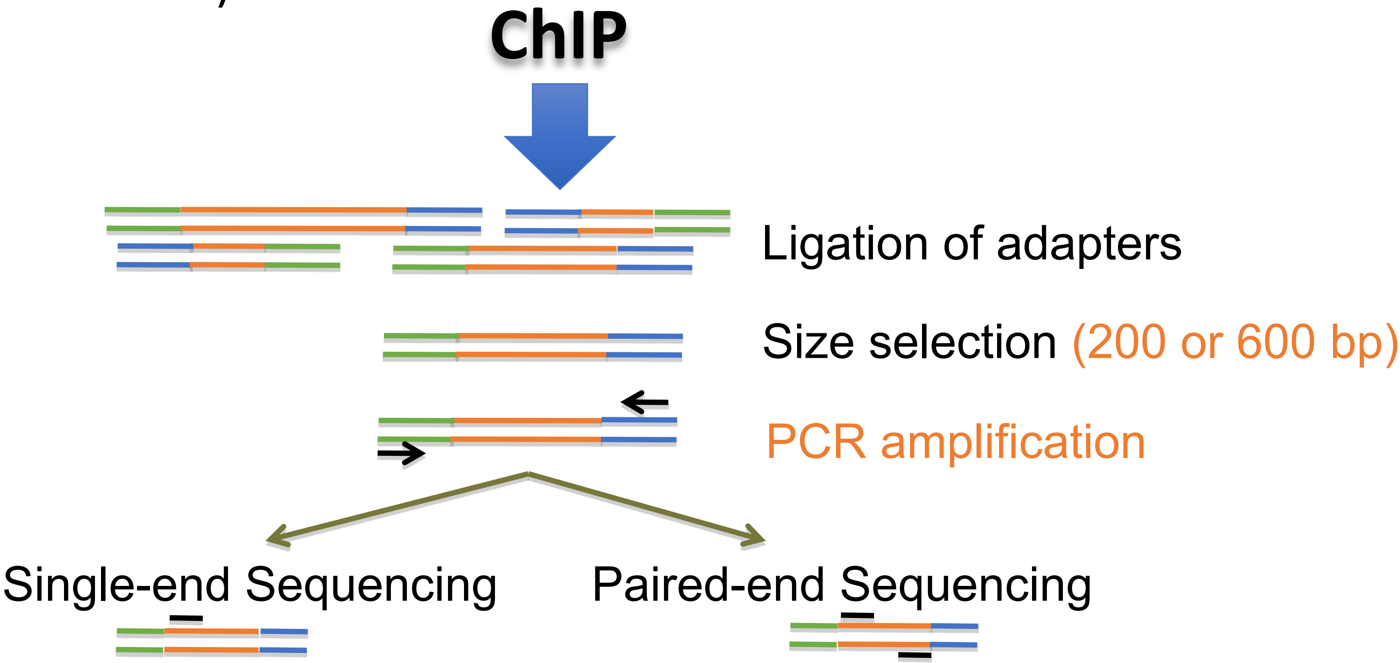


After library prep



Library prep

- Step between chIP and sequencing
- The goal is to prepare DNA for the sequencing
- Starting material:
  - 2-10ng of sheared DNA (Purification using Agencourt AMPure XP - Beckman Coulter)
  - The mean DNA fragment size should be below 500bp (Smear in a range of 100 to 700)



ChIP

Ligation of adapters

Size selection (200 or 600 bp)

PCR amplification

Single-end Sequencing

Paired-end Sequencing

Sequencing

- Sequencer : Illumina HiSeq 4000
- No. of reads per run, per sample :
  - HiSeq 4000: 300-350 millions reads per lane
  - Multiplexing 8 samples per lane :~43 millions per sample
- Length of DNA fragment : ~200-600bp
- No. of cycle per run : 50

Single end or paired end?

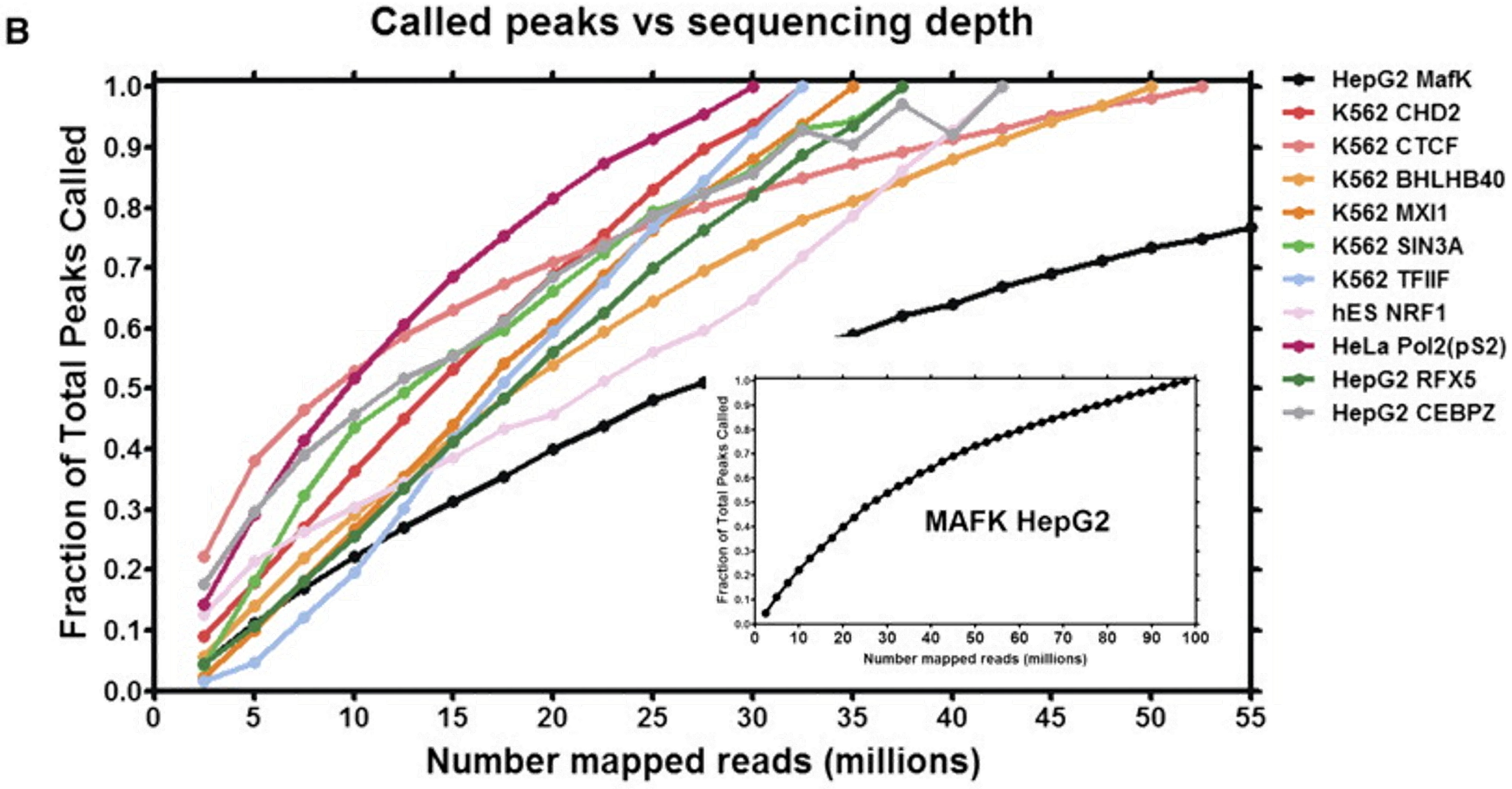
- Single end (most of the time)
- Paired-end sequencing
  - ☺ Improve identification of duplicated reads
  - ☺ Better estimation of the fragment size distribution
  - ☺ Increase the mapping efficiency to **repeat regions**
  - ☹ The price!

Sequencing depth

Consider the depth needed depending on:

- chipped protein,

Sequencing depth



Landt et al, 2012

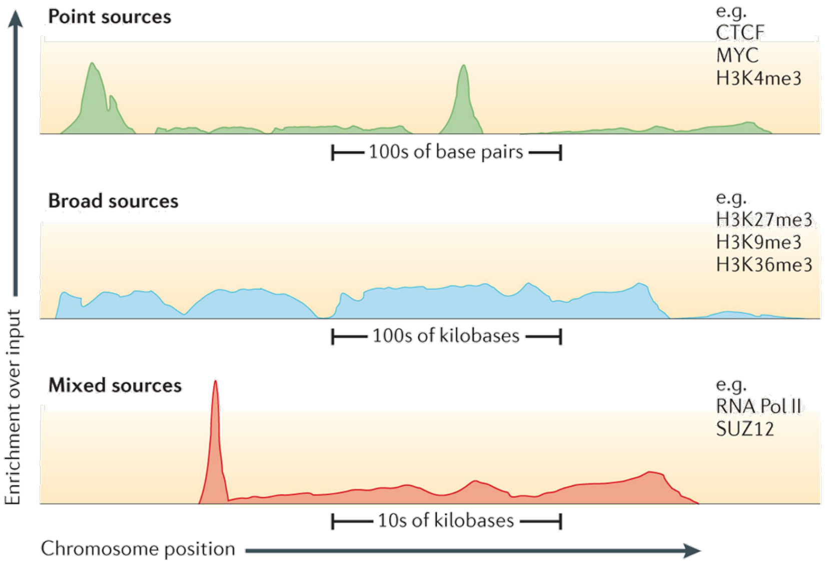
Sequencing depth

Consider the depth needed depending on:

- chipped protein,
- type of expected profile,
- number of expected binding sites,
- size of the genome of interest.

Ex:

- For human genomes, 20 million uniquely mapped read sequences are suggested for point-source peaks, or 40 million for broad-source peaks.
- For fly genome: 8 million reads
- For worm genome: 10 million reads



Nature Reviews | Genetics

Control sample

- **Problem**:
  - Only a fraction of sequenced reads corresponds to true signal (DNA fragments specifically selected during ChIP). Non-specific enrichment is called background.
  - We want to extract regions in the genome that are specifically bound by the protein of interest and not due to some experimental artefacts.
- **Solution**:
  - In addition to chipped samples, control samples are also generated and sequenced. Control samples don’t undergo specific immunoprecipitation thus read enrichments they might exhibit, are non specific. By comparing chipped to control samples, regions showing read enrichment both in control and chipped samples might be false positive enrichments and thus not retained as true peak.
- 3 types of control are commonly used :
  - Input DNA : a portion of DNA sample removed prior to IP
  - DNA from non specific IP : DNA obtained from IP with an antibody not known to be involved in DNA binding or chromatin modification such as IgG
  - Mock IP DNA : DNA obtained from IP without antibodies

Replicates

- A minimum of two replicates should be carried out per experiment.
- Each replicate should be a **biological** rather than a technical replicate; that is, it represents an independent cell culture, embryo pool or tissue sample.

ENCODE

- The Encyclopedia of DNA Elements (ENCODE) Consortium has carried out hundreds of ChIP–seq experiments and has used this experience to develop a set of working standards and guidelines

See: https://www.encodeproject.org/about/experiment-guidelines/

Data used in this course

Sample name	No. of raw reads
MITF	31,334,257
Ctrl	29,433,042
H3K4me3	11,192,622
polII	10,404,820