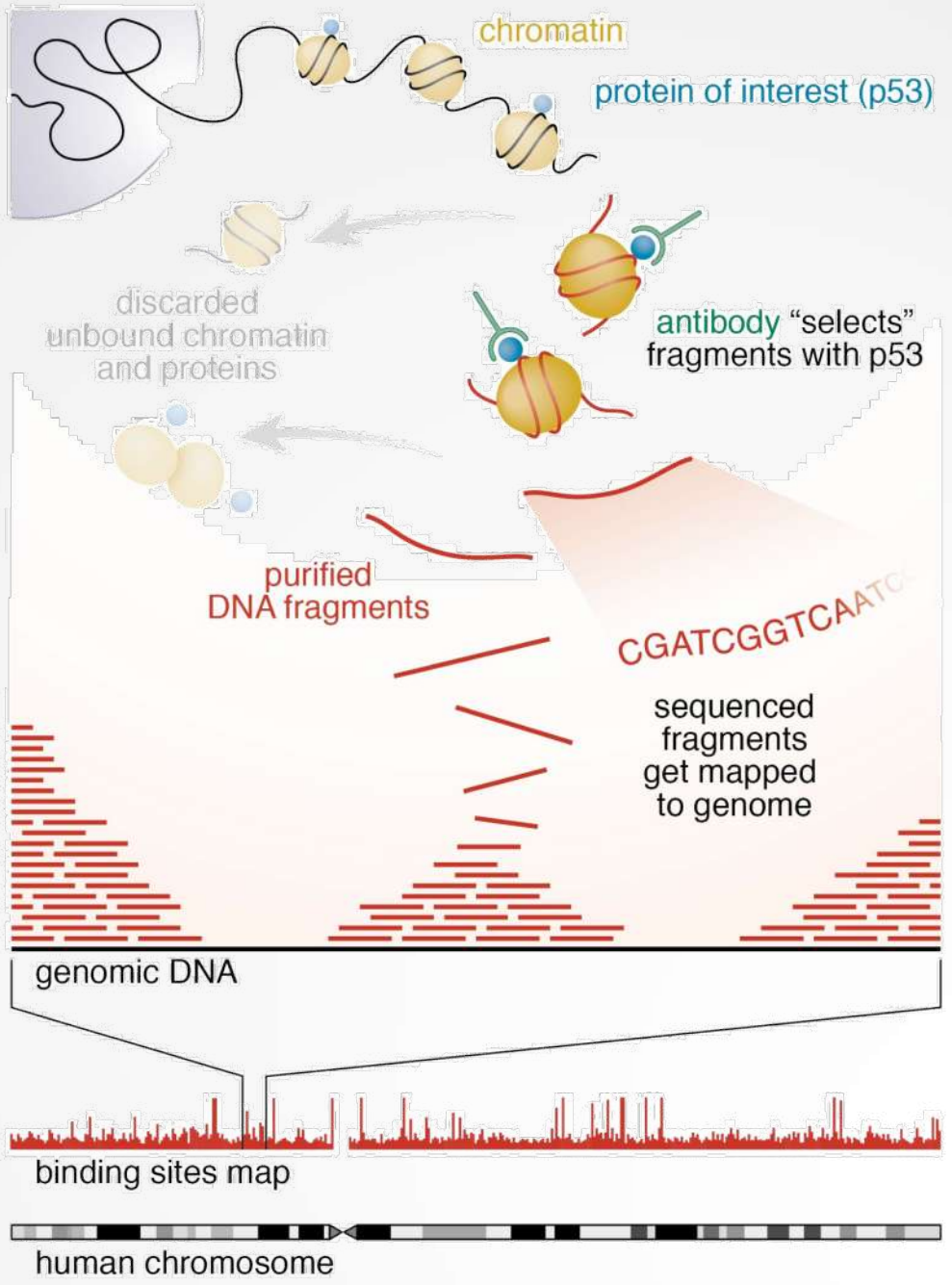
ChIP-sequencing: Library preparation and experimental design

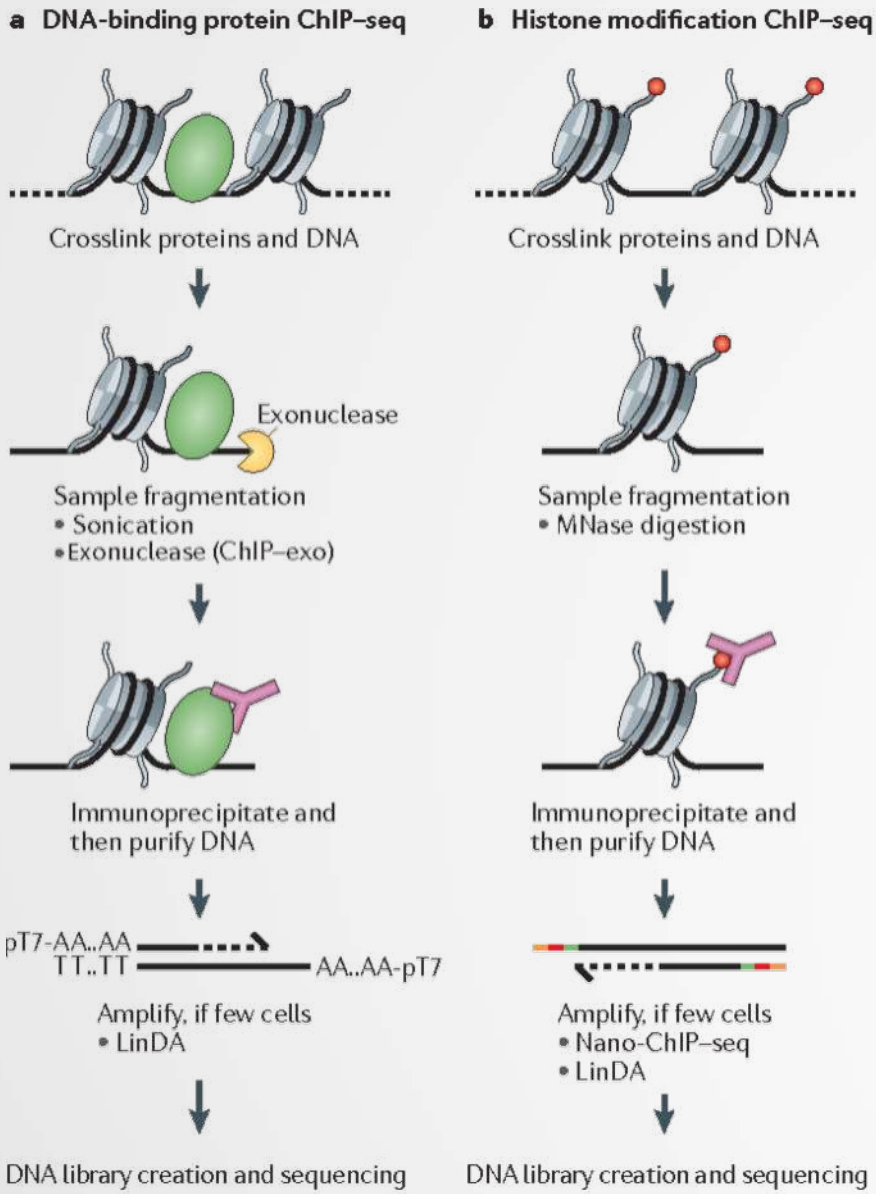
Stéphanie Le Gras
(slegras@igbmc.fr)

ChIP-seq


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

ChIP and library prep considerations

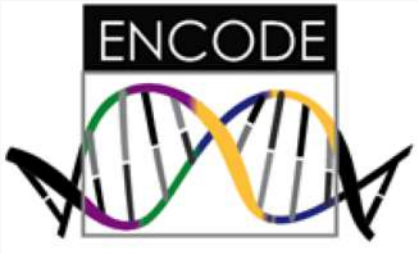
Chromatin ImmunoPrecipitation



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

ENCODE

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



Landt SG, Marinov GK, Kundaje A *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813–1831.

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

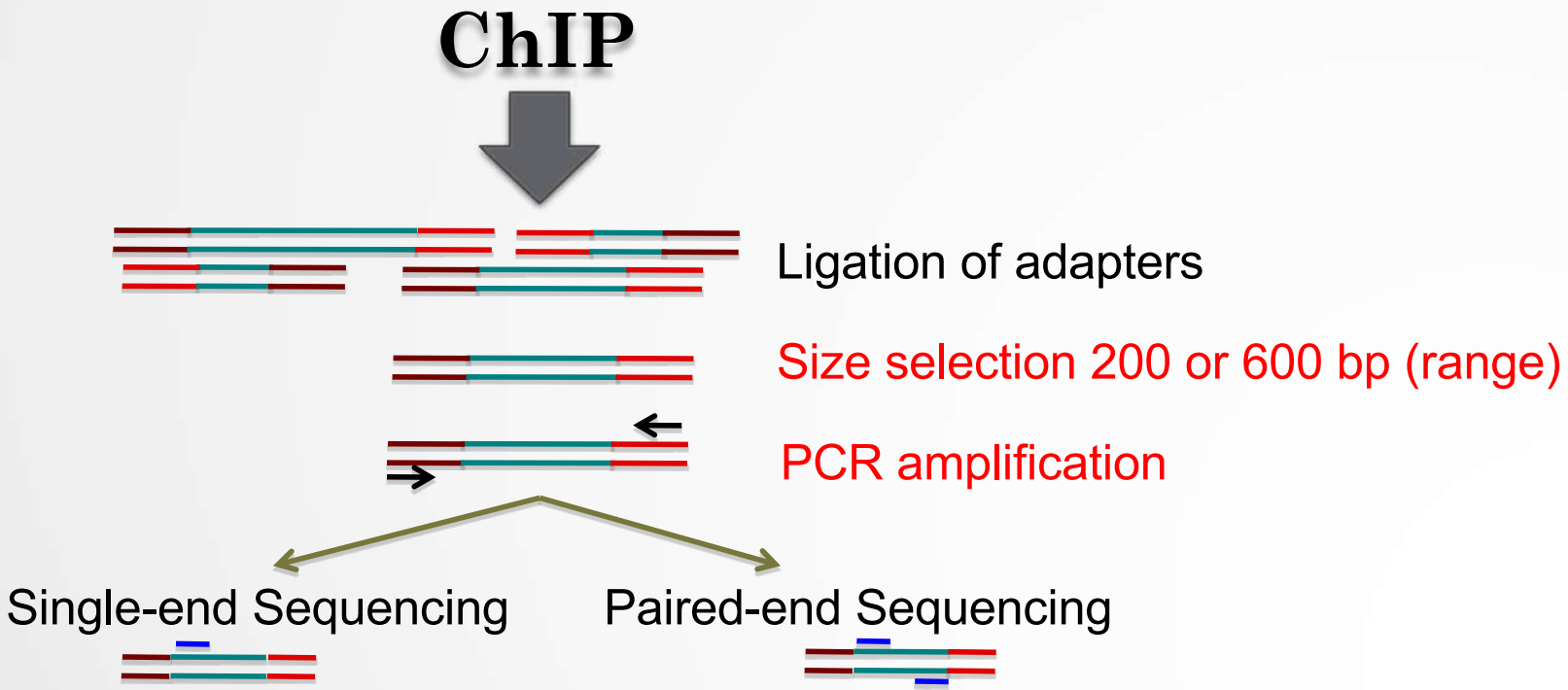
Considerations on chIP

- 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)
- Shearing of DNA (Mnase I, sonication, Covaris): trying to narrow down the size distribution of DNA fragments

⟶ **Complexity in DNA fragments**

Library prep

- Step between chIP and sequencing
- The goal is to prepare DNA for the sequencing
- Starting material: ChIP sample (1-10ng of sheared DNA)



Sequencing

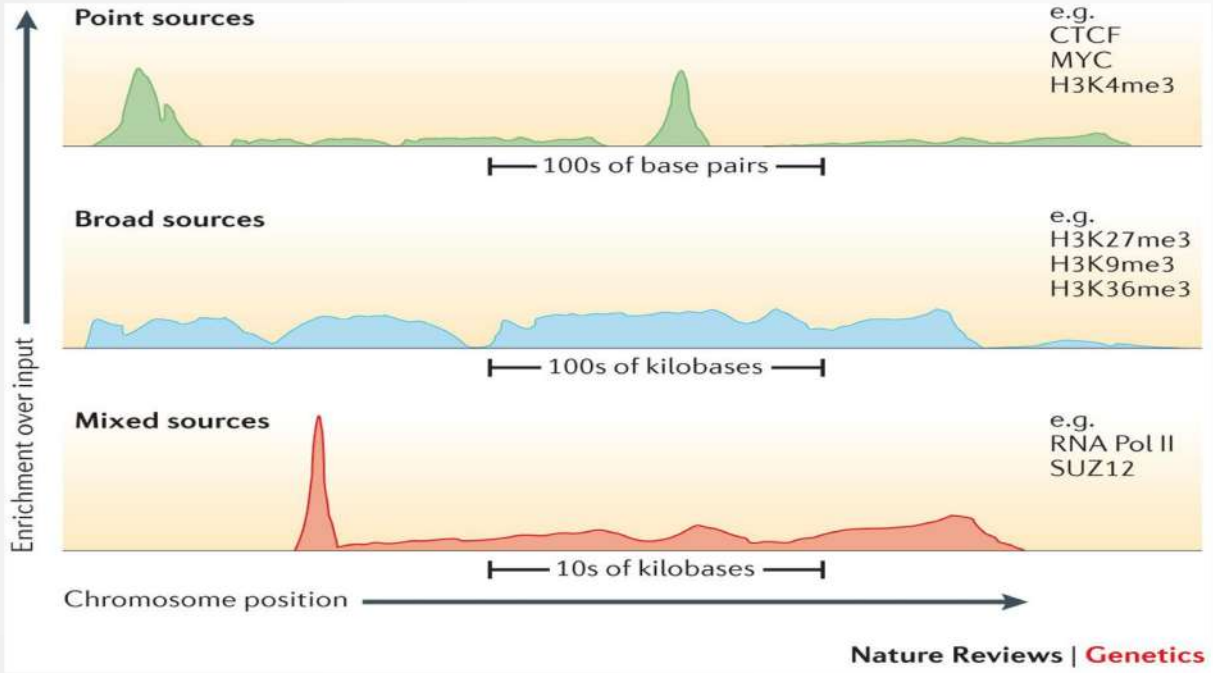
- Sequencer : Illumina HiSeq 4000
- No. of reads per run, per sample :
  - 1st run on the GAIIx : 10-20 millions of reads per lane
  - (HiSeq 2500) 4 samples per lane :~41 millions per sample
  - (HiSeq 4000) 8 samples per lane :~43 millions per sample
- Length of DNA fragment : ~200-600bp
- No. of cycle 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!

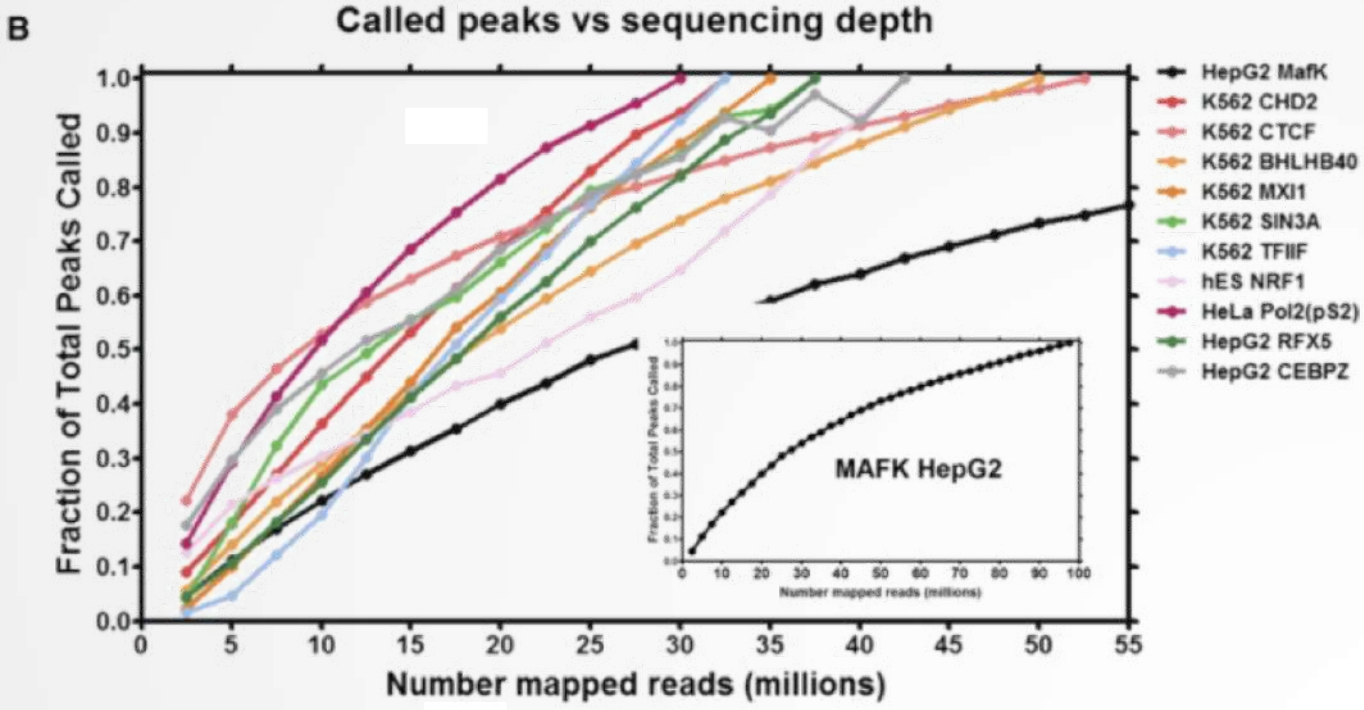
Sequencing depth

- Consider the depth needed depending on:
  - Chipped protein



Sequencing depth

- Consider the depth needed depending on:
  - Chipped protein



Landt et al, 2012

Sequencing depth

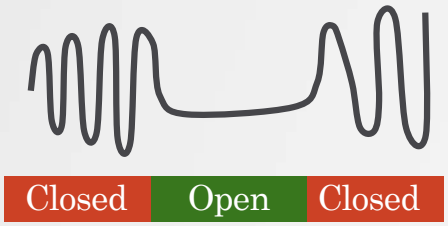
- Consider the depth needed depending on:
  - Chipped protein
  - Number of expected binding sites
  - Size of the genome of interest
- Ex:
  - For human genomes
    - 20 million uniquely mapped read sequences for point-source peaks,
    - 40 million for broad-source peaks.
  - For fly genome: 8 million reads
  - For worm genome: 10 million reads

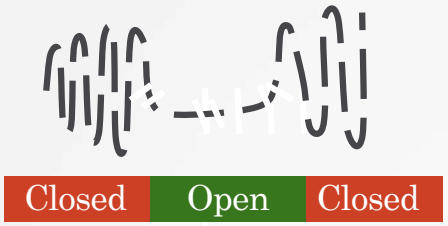
Controls

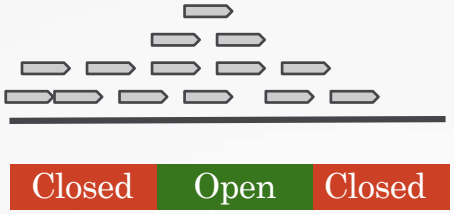
- Used mostly to filter out false positives (high level of noise)
  - Idea: potential false positive will be enriched in both treatment and control.
- A control will fail to filter out false positives if its enrichment profile is very different from the enrichment profile of false positive regions in the treatment sample
- Most commonly used control: Input DNA (a portion of DNA sample removed prior to IP)
- Choice of control is extremely important
- It is recommended to cover the control in a higher extent than the IPs (= the control should cost more money !)

Why an Input is required ?

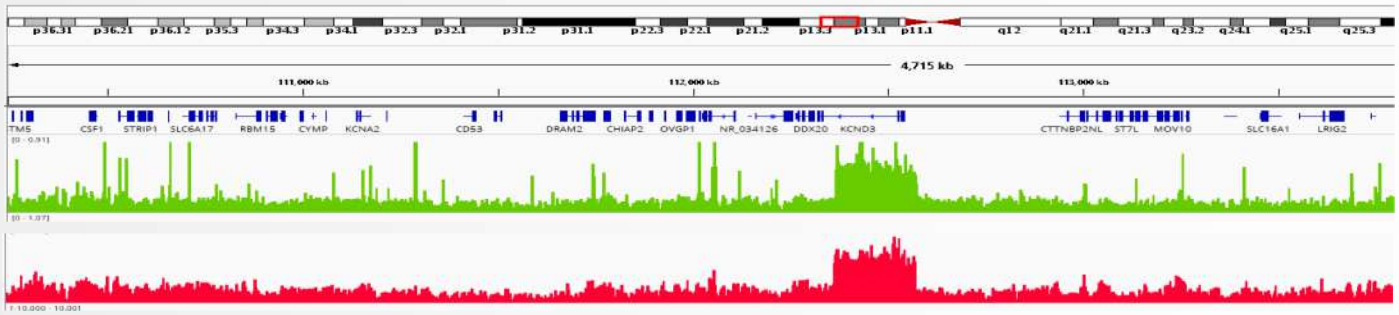
- The input is used to model local noise level
  - Accessible regions are expected to produce **more reads**





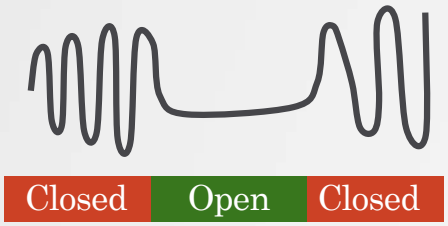


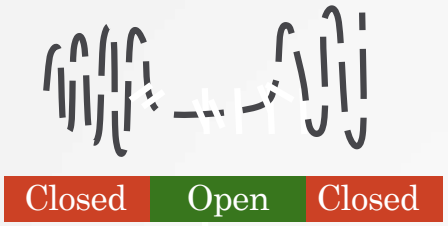
  - **Amplified regions (CNV) are expected to produce more reads**

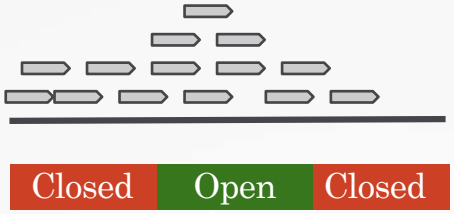


Why an Input is required ?

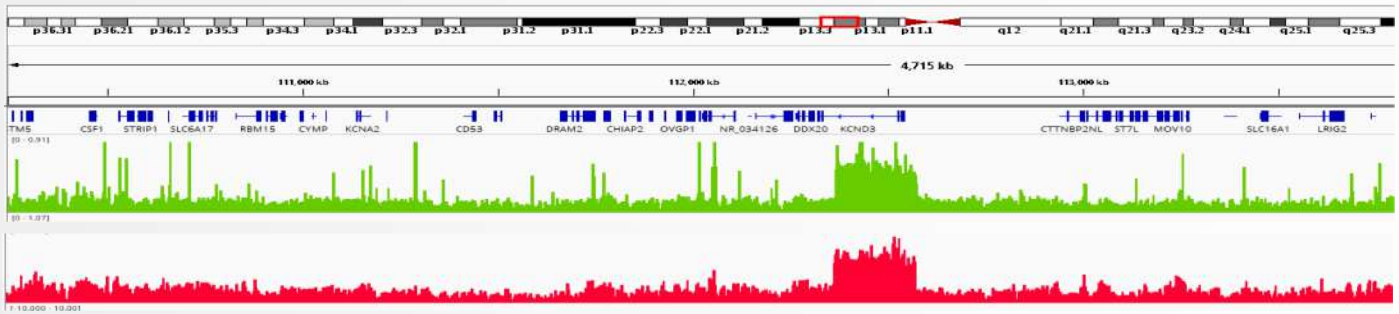
- The input is used to model local noise level
  - Accessible regions are expected to produce **more reads**







  - **Amplified regions (CNV) are expected to produce more reads**



  - **Moreover, most peak callers** are configured **with an input as control**

Other controls

- IgG (mock IP): controls for non-specific IP enrichment
  - Problem : low-complexity library (few reads)
- Histone H3 (for H3 variants)
- Uninduced condition (for inducible TFs)
  - Example : Glucocorticoid Réceptor
  - Induced by Dexamethasone (Dex)
  - Control vehicule = Ethanol (EthOH)
- KO of your protein of interest
- Non flagged cell lines
- ...
-

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 results from an independent cell culture, embryo pool or tissue sample.

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