Interaktionen und Modifikation von RNA und Proteinen

Prof. Sonja Prohaska

Computational EvoDevo
Universität Leipzig

SS 2018

ChIP-seq

Chromatin **I**mmuno-**P**recipitation followed by **seq**uencing is a **next-generation sequencing (NGS)** technology that retrieves genomic positions of protein - DNA interaction.

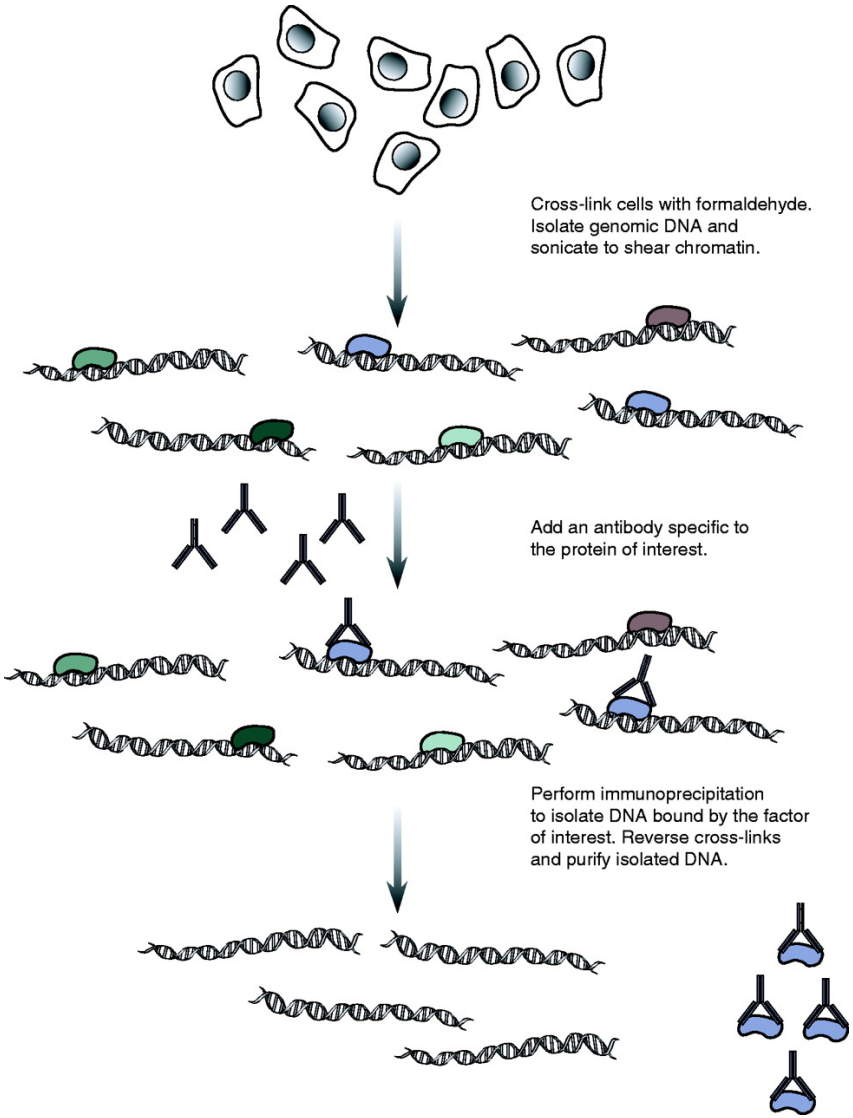
- **Transcription factors**
  - transcription factors
    e.g. estrogen receptor, homeodomain proteins, etc.
  - other protein-based DNA and transcription regulators
    e.g. CTCF, nuclear membrane attachment proteins, etc.
- **Histone modifications**
  - modified nucleosomes/histones
    e.g. acetylation, methylation, etc. on specific sites of histones such as H3K4me3, H3K27ac, etc.

Different protocols during experiment and data analysis.

ChIP-seq Conceptual Lab Workflow

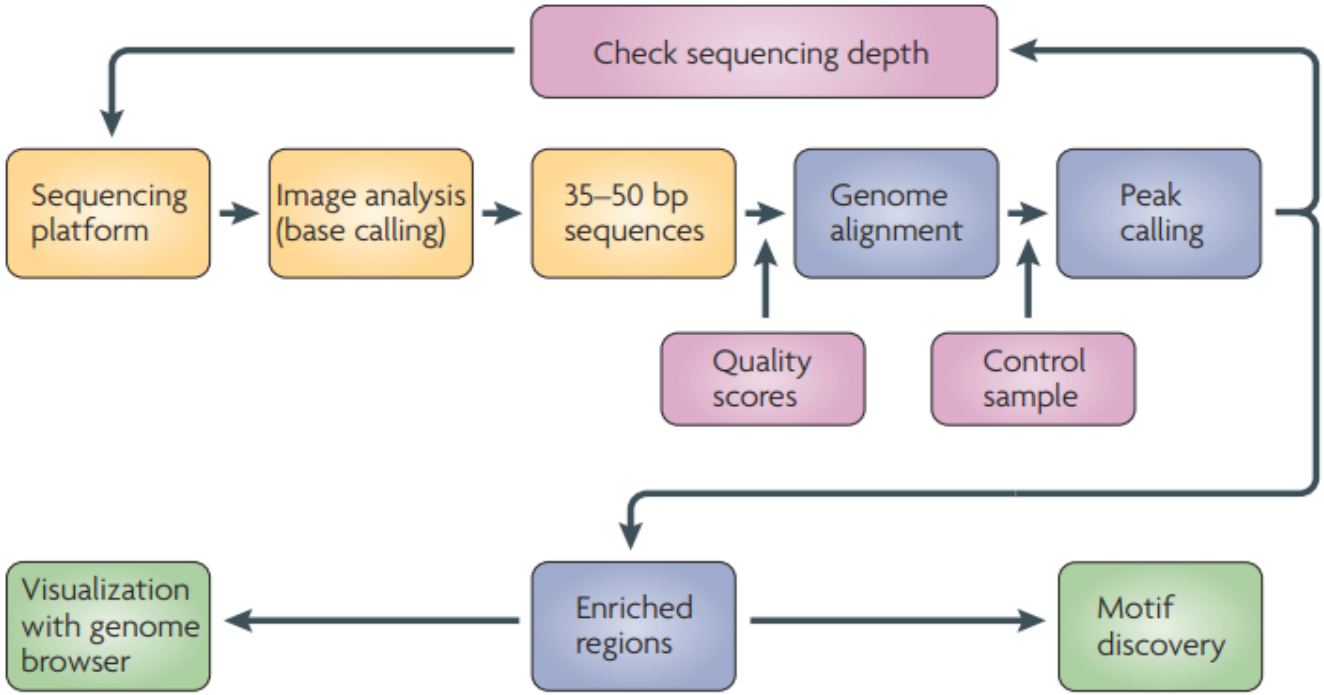
- ChIPseq (**Ch**romatin **I**mmuno-**P**recipitation followed by **seq**uencing)
- proteins are cross-liked to DNA in vivo
- chromatin is isolated
- sonicate to obtain chromatin/DNA fragments in length range 200-300nt
- **immunoprecipitate** with specific antibodies for target protein (modification)
- purify immunocomplexes (remove chromatin fragments not bound by the antibody)
- reverse cross-linking
- purify DNA from chromatin fragments
- get fragments sequenced

ChIP-seq Conceptual Lab Workflow



ChIP-seq Analysis Workflow

Shown are the steps at which bioinformatics is needed.



ChIP-seq Analysis (general)

- **Sequencing**
  steps shown here are based on sequencing with Illumina
- **Read mapping**
- aligning reads against the genome ("genome alignment")
- **Peak calling**
  calculate read density
  compare to/normalize with control/background
  commonly 'input DNA' or 'IgG'
  derive preaks, i.e. regions significantly enriched in reads
- compile list of enriched regions
- **data visualization** e.g. as genome browser track

ChIP-seq Analysis for Transcription Factors (TFs)

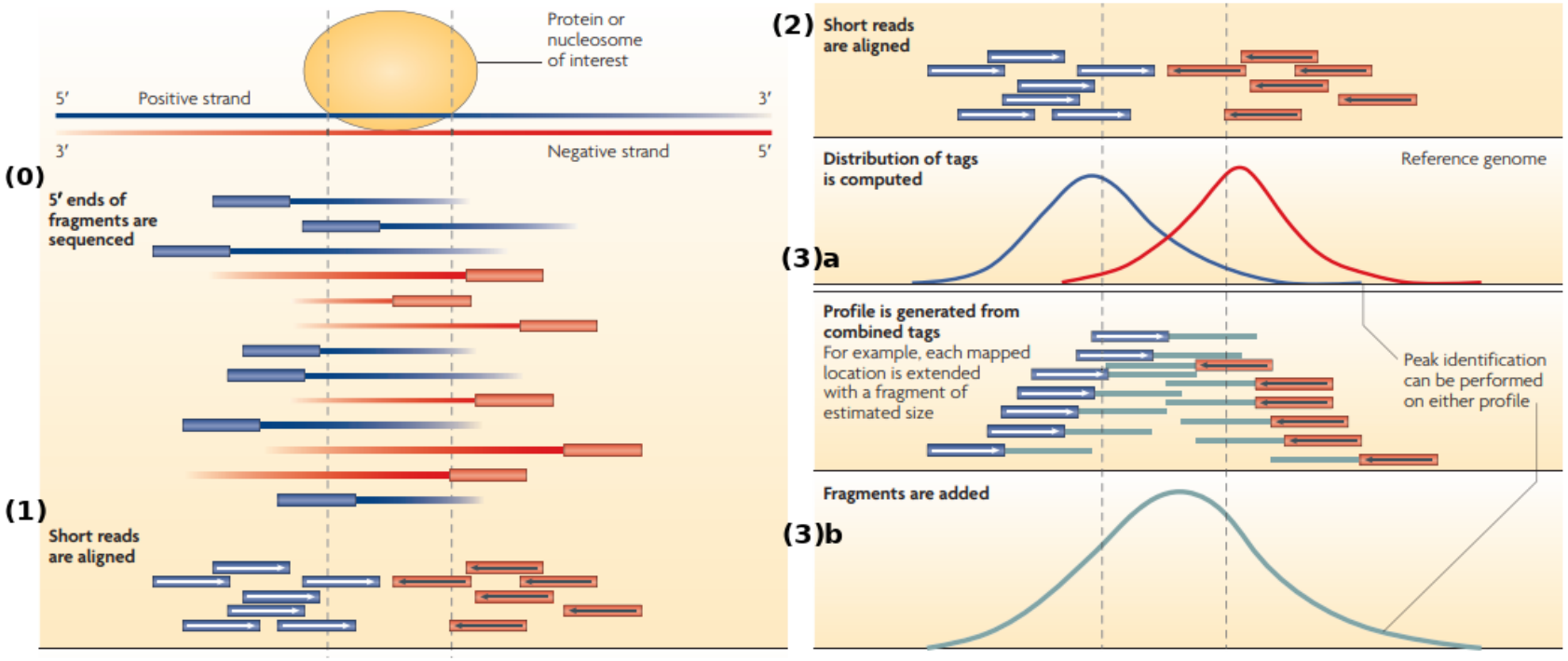


Figure 5 | **Strand-specific profiles at enriched sites.**

ChIP-seq Analysis for Transcription Factors (TFs)

(0) a ChIP-seq experiment returns sequence fragments (200-300nt)

(1) **Sequencing**: with Illumina 5'-ends (35-50nt) are sequenced $\rightarrow$ **reads**

(2) **read mapping** (alignment of reads to the reference genome)
5'-ends of plus (blue boxes) and minus (red boxes) strands are mapped
and read distributions are calculated

(3) **peak calling** identification of peaks (against a background)

a) the peak on the plus strand and the peak on the minus strand is called $\rightarrow$ the peaks flank the traget binding region

b) after conceptual extention of the reads a single peak is called $\rightarrow$ the peaks covers the target binding region

- target binding region (100-200nt) vs. target binding site (4-12nt)
- apply a **motif discovery** tool to derive the binding motif
from the set of binding regions

References

Peter J. Park. *ChIP-seq: advantages and challenges of a maturing technology.* Nature Reviews Genetics, 10(10), p.669.

Rakesh Kaundal (2016). *NGS data analysis with R/Bioconductor: ChIP-Seq workflow.*
`http://biocluster.ucr.edu/~rkaundal/workshops/R_feb2016/ChIPseq/ChIPseq.html`