

Introduction to Chromatin IP – sequencing (ChIP-seq) data analysis

Workshop on **ChIP-seq data analysis**

Stockholm, 7 November 2018

Agata Smialowska

NBIS, SciLifeLab, Stockholm University



Chromatin state and gene expression



PEV
Position effect
variegation
in *Drosophila* eye
(nature.com)

First observed by
H. Muller
1930

Juxtaposition of eye colour genes with heterochromatin results in the “mottled” eye colouration (red and white).

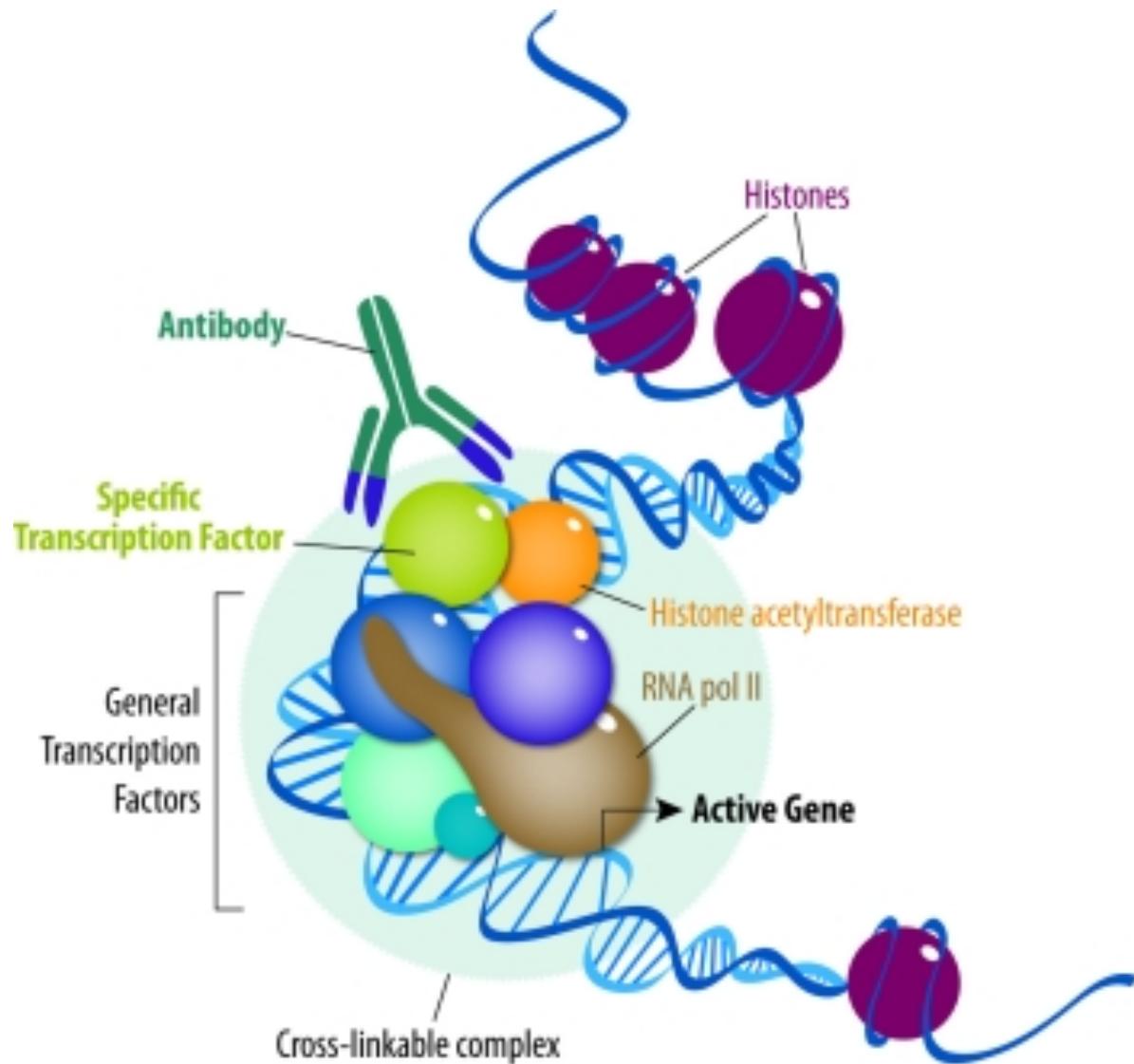
Proteins, which bind heterochromatin, act to “spread” the silencing signal by providing a forward feedback loop.

Heterochromatin Protein 1; Histone methyltransferase Su(var)3-9; H3K9 methylation

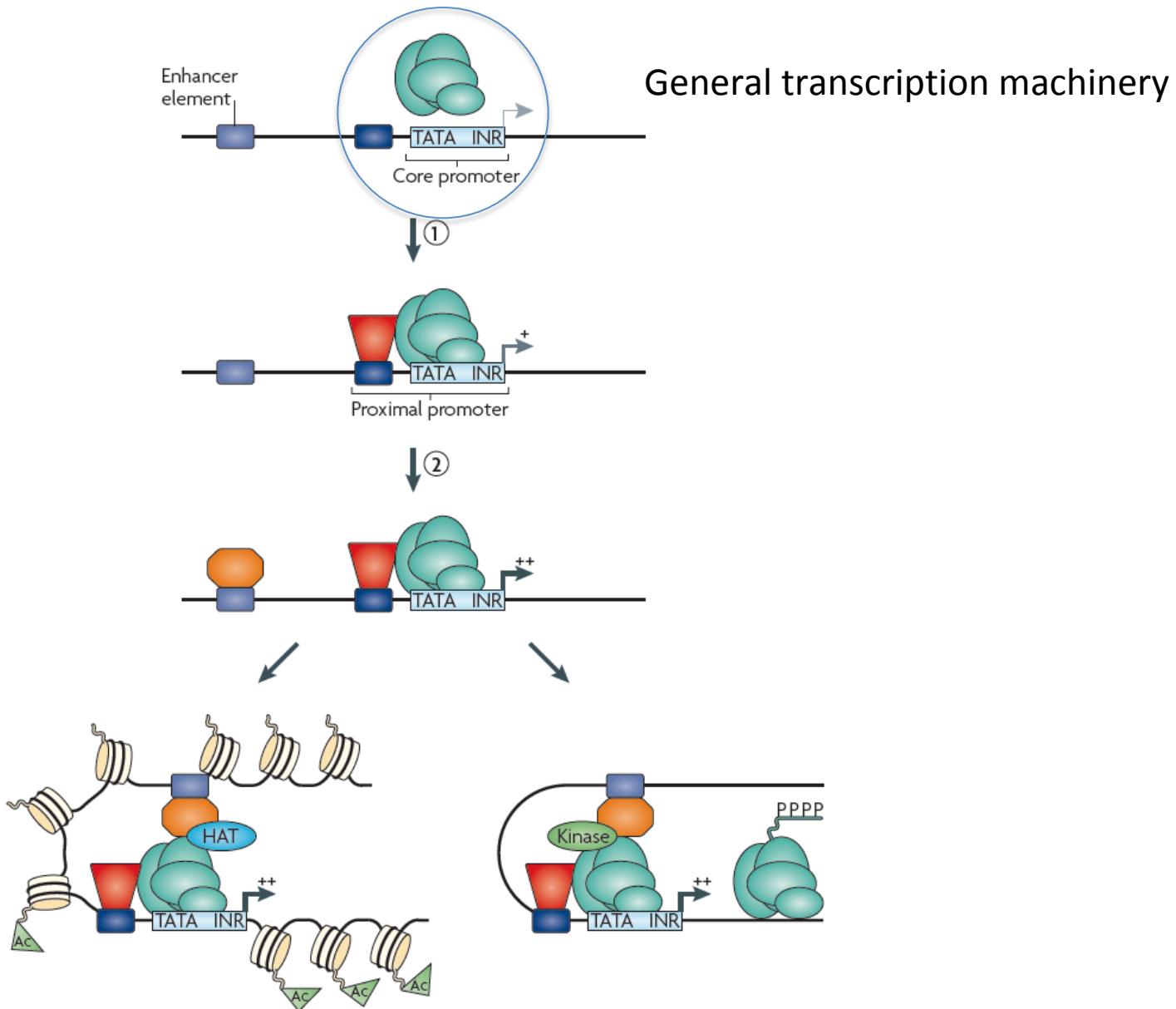
www.pollev.com/AGATASMIALOW506

Give three keywords related to ChIP-seq

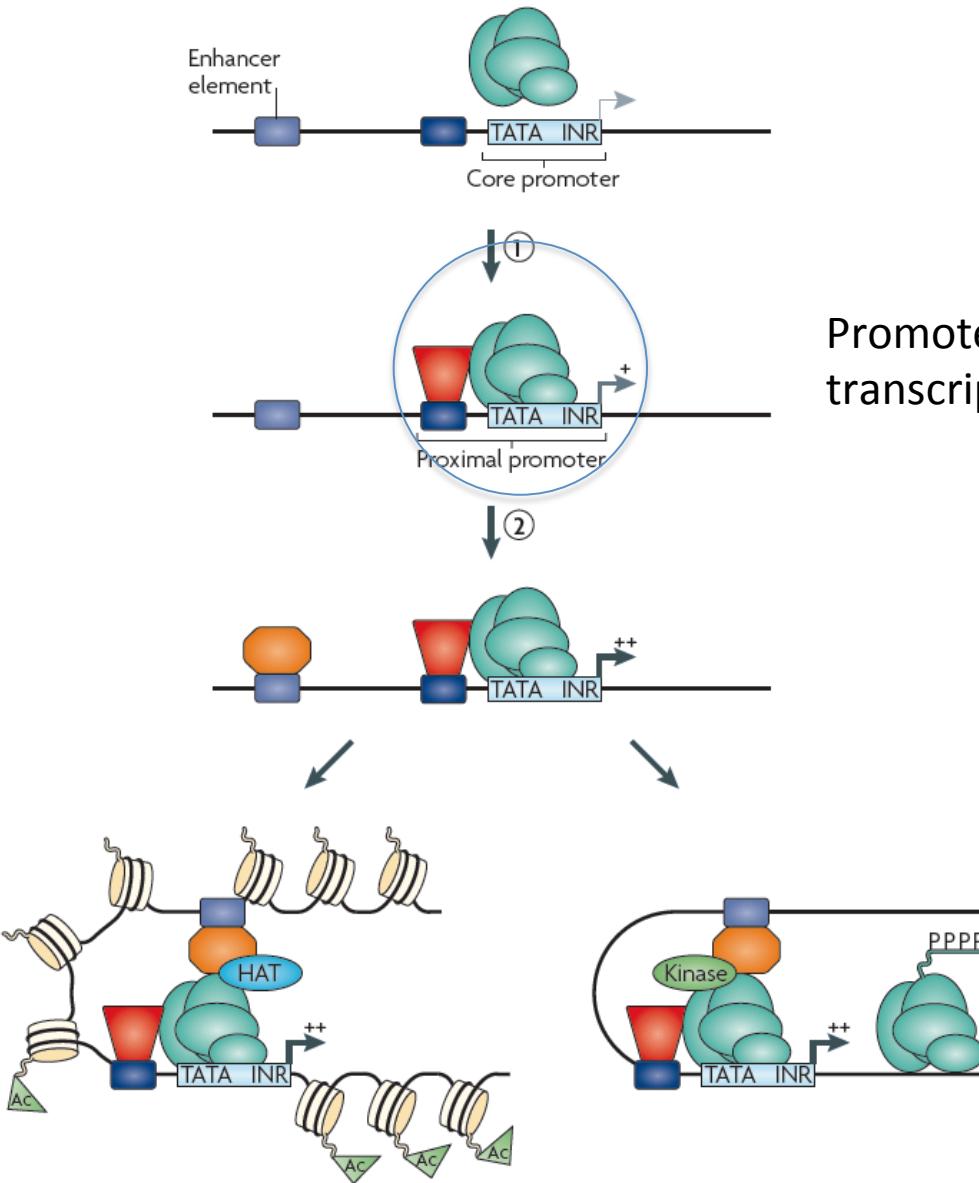
Chromatin immunoprecipitation



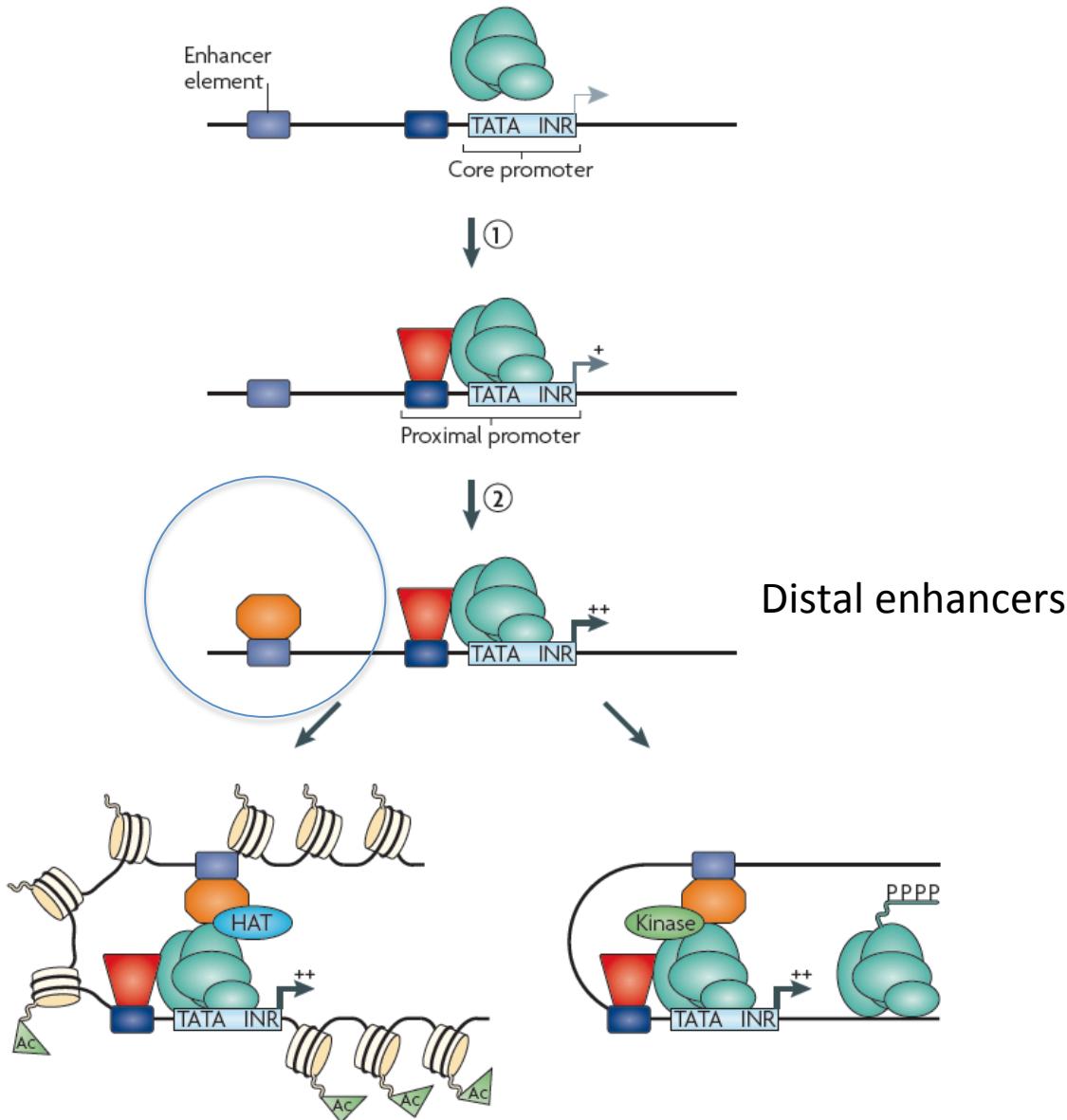
Applications



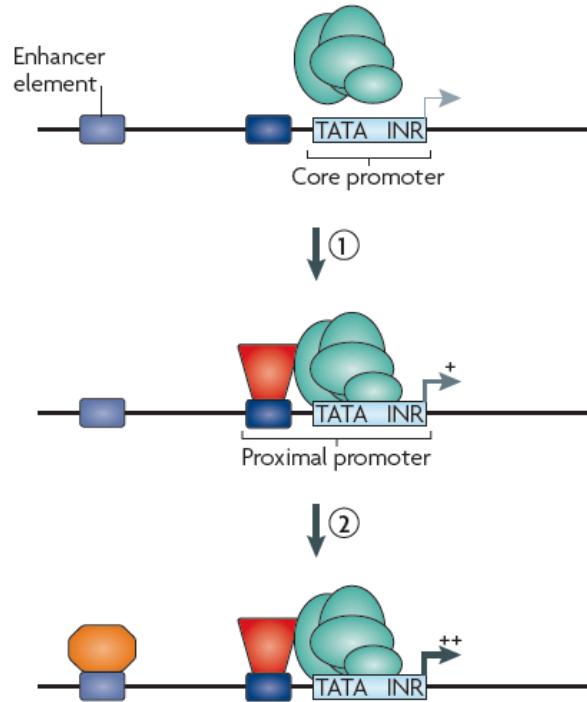
Applications



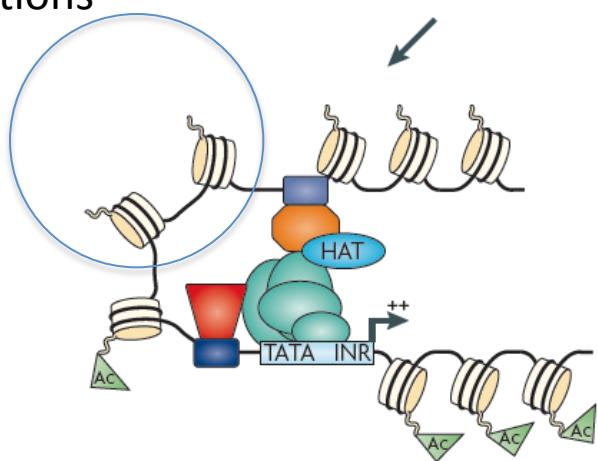
Applications



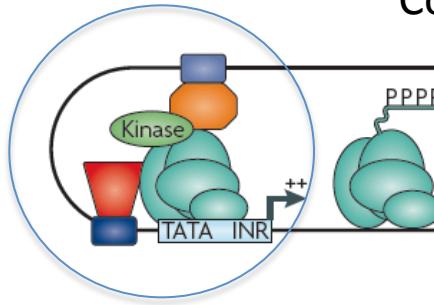
Applications



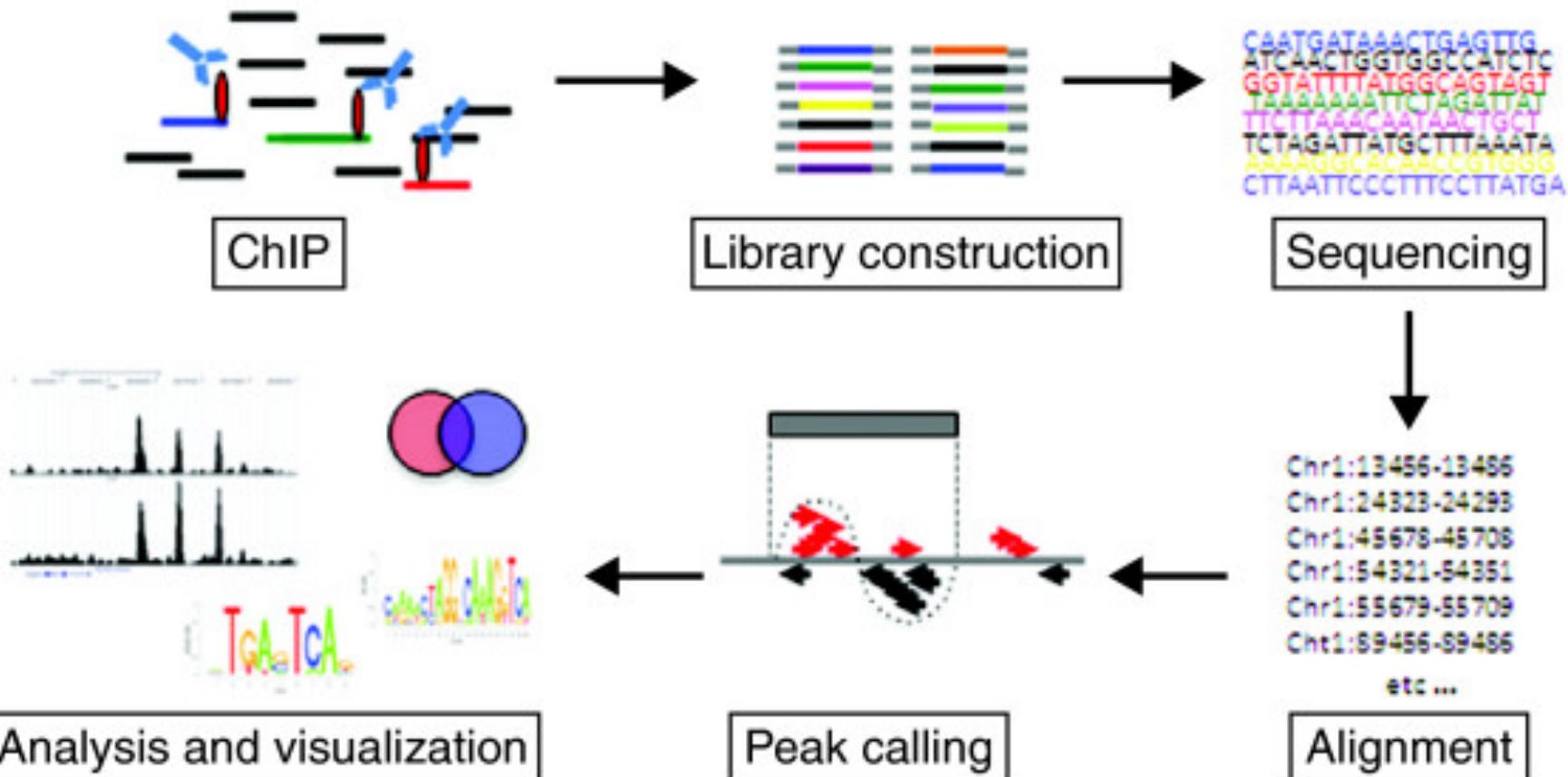
Histone modifications
and variants



Activation states
Co-factors



ChIP-seq workflow



Workflow of a ChIP-seq study

design study

obtain input chromatin

perform precipitation

construct library

sequence library

bioinformatic analysis



Critical factors

- Antibody selection
 - Proper control sample (input chromatin or mock IP)
 - Library cloning and sequencing
 - Algorithm for peak detection
-
- Enough material and biological replicates
 - Reproducibility in chromatin fragmentation
 - Cross-linker choice

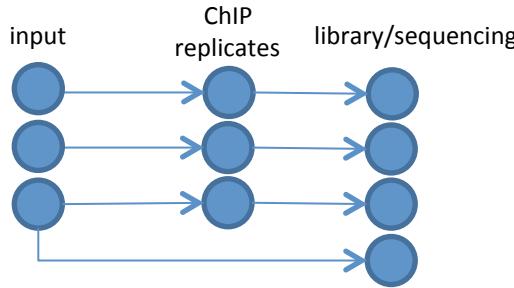
Experiment design

- Sound experimental design: replication, randomisation and blocking (R.A. Fisher, 1935)
- In the absence of a proper design, it is essentially impossible to partition biological variation from technical variation
- Sequencing depth: depends on the structure of the signal; cannot be linearly scaled to genome size
- Single- vs. paired-end reads: PE improves read mapping confidence and gives a direct measure of fragment size, which otherwise has to be modelled or estimated

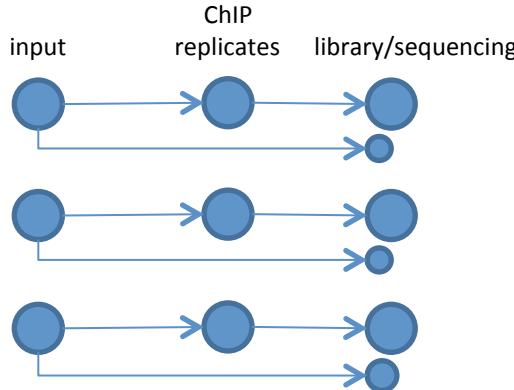
Experiment design

Ideal design:

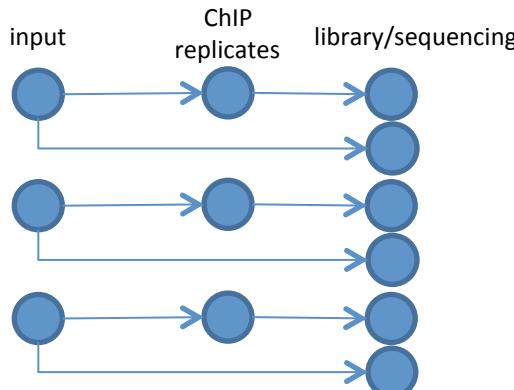
X



X

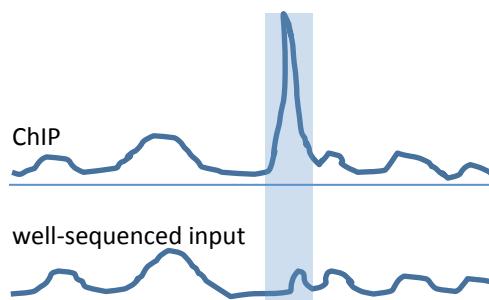
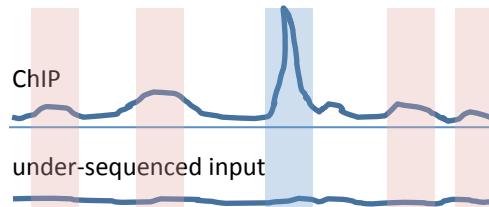


✓



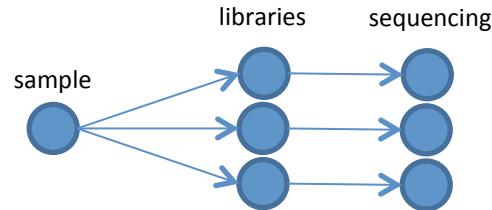
Each sample has a matched input
Input sequenced to a comparable depth
as IP sample

≥2 biological replicates for site identification
≥3 biological replicates for differential binding



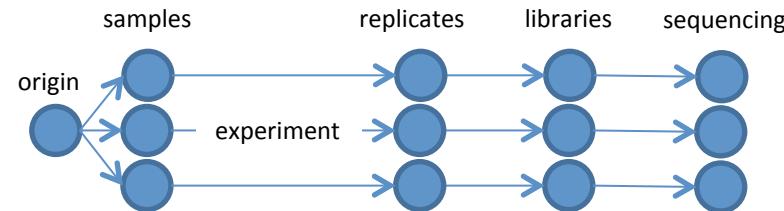
Biological replicates and randomisation

X



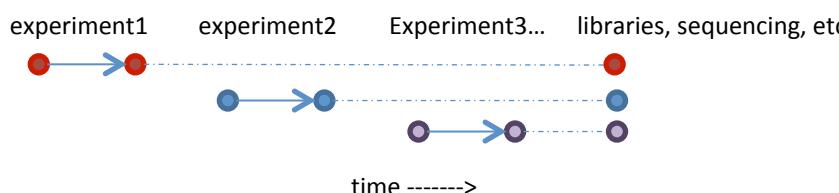
technical replicates are generally a waste of time and money

X



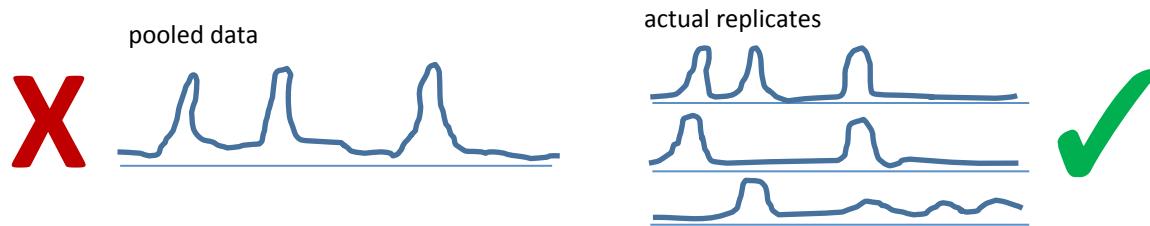
many studies do not account for batch effects
i. time
ii. origin

✓

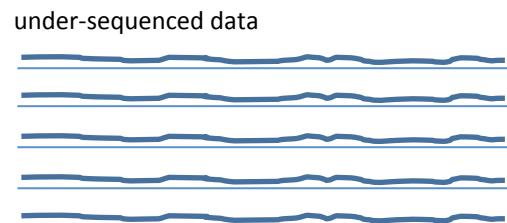


so if you care about reproducibility

Importance of sequencing depth

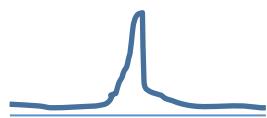


if you need to pool your data, then it is under-sequenced



Sequencing depth depends on data type

Transcription
Factors



point-source

Chromatin
Remodellers
Histone marks



mixed signal

Chromatin
Remodellers
Histone marks
RNA polymerase II



broad signal

Human: TF: 20 M

?

?

H3K4me3: 25 M

H3K36me3: 35 M

H3K27me3: 40 M

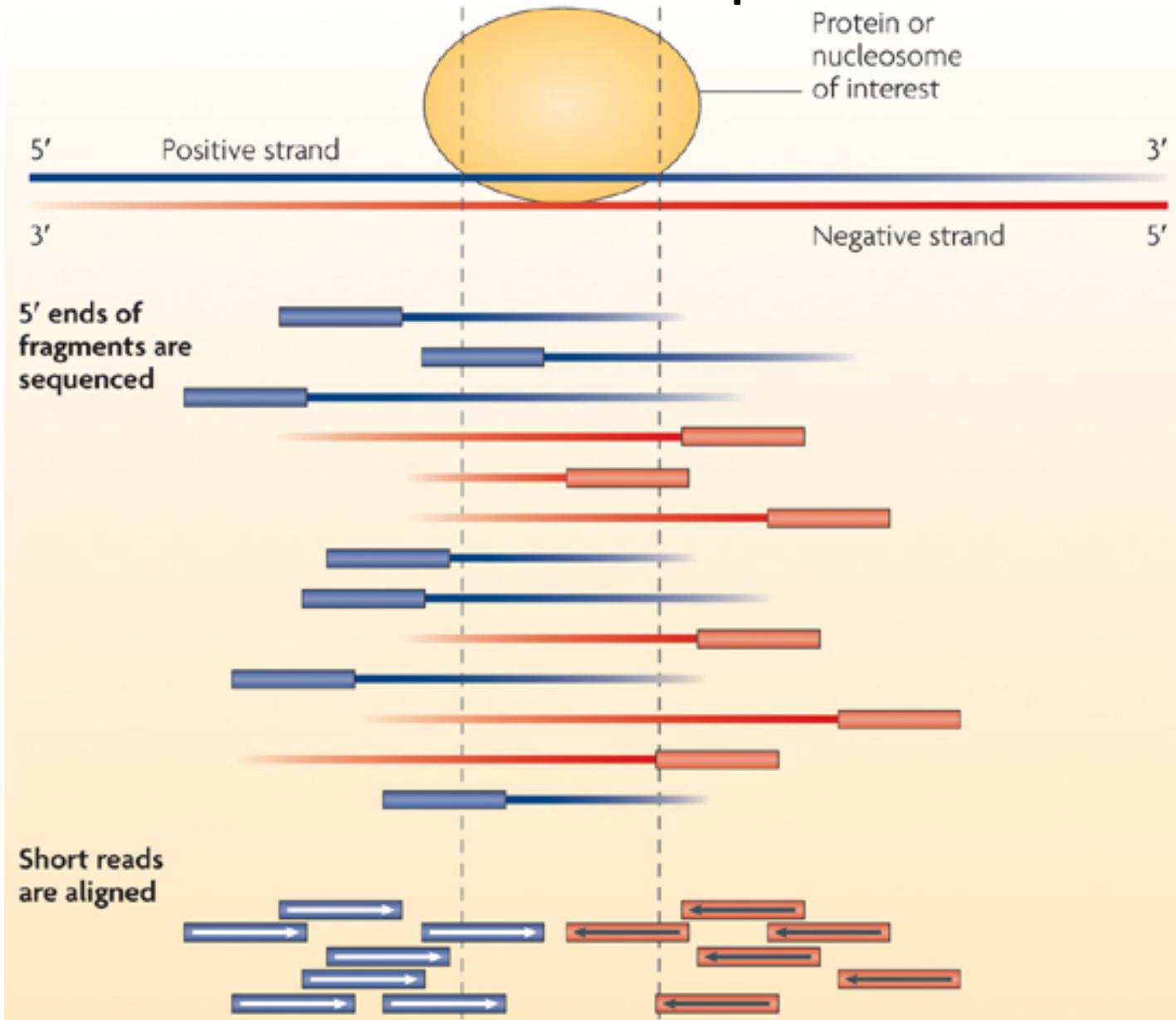
H3K9me3: >55 M

No clear guidelines for mixed and broad type of peaks

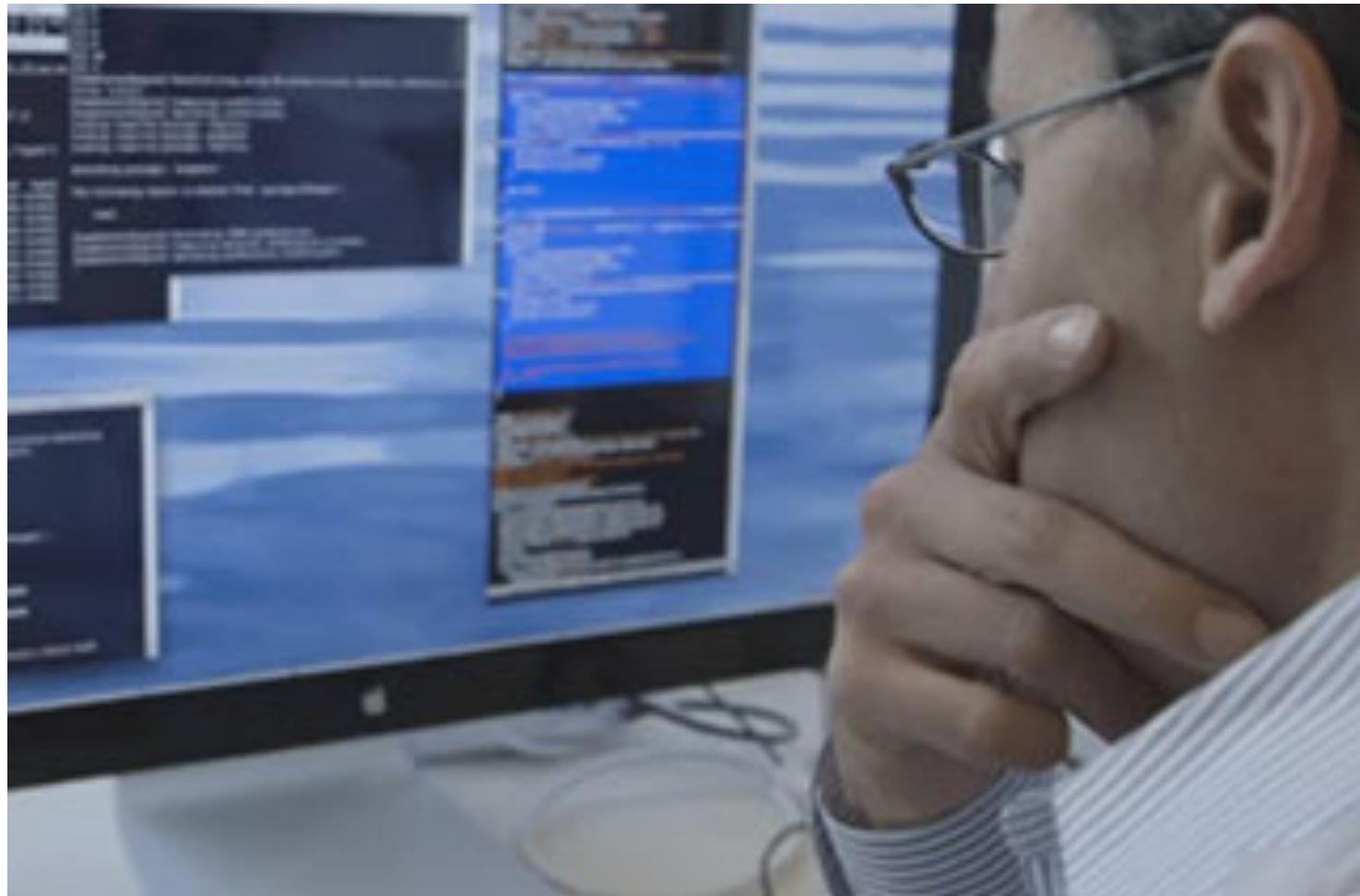
- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

Chromatin = DNA + proteins



Data analysis



Rank the steps of data analysis of a ChIP-seq experiment by importance to the final result

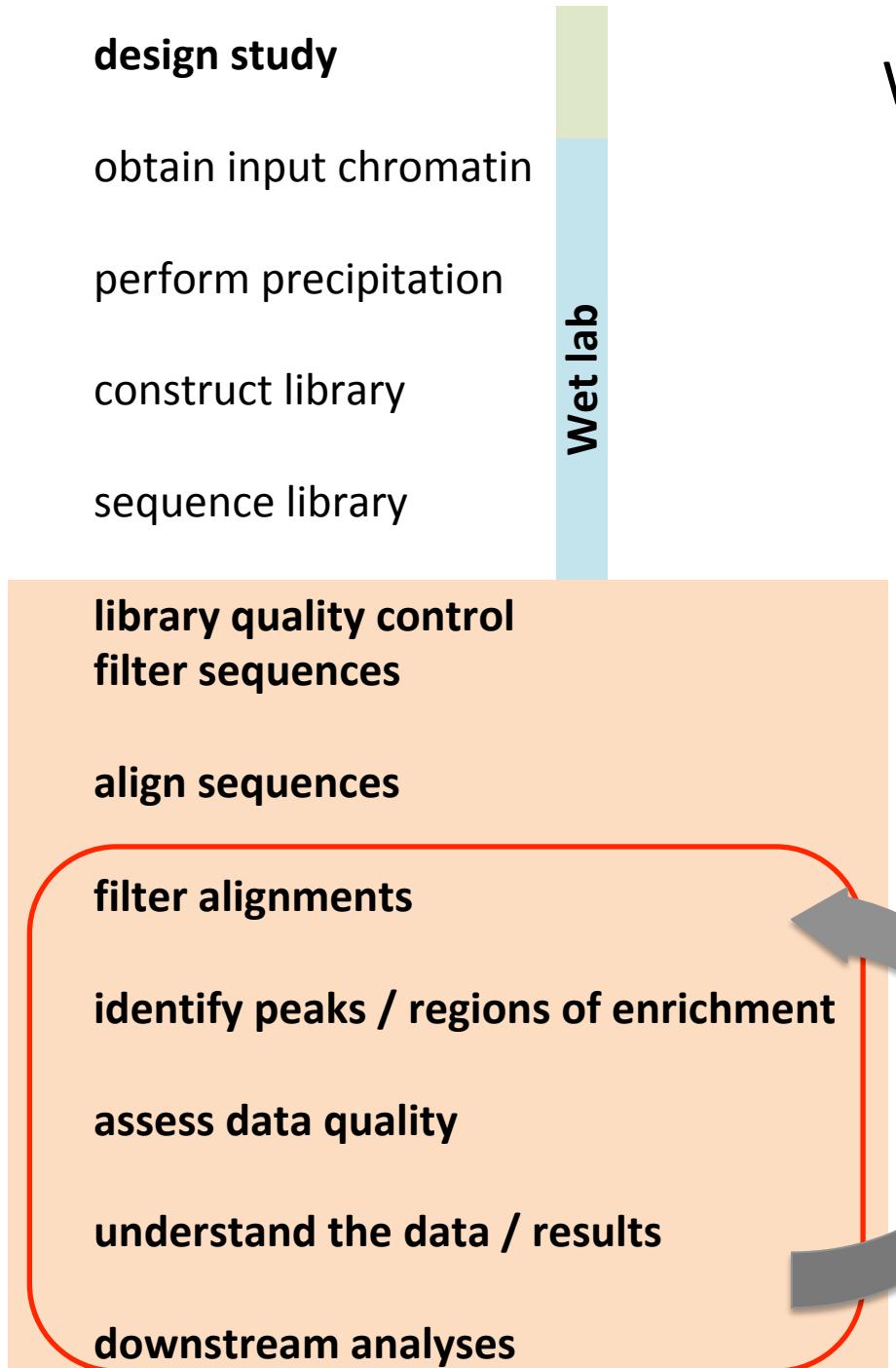
peak calling

read filtering
and mapping

alignment
processing

post-alignment
quality control

Workflow of a ChIP-seq study



- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

Two questions to address

- 1. Did the ChIP part of the ChIP-seq experiment work? Was the enrichment successful?
- 2. Where are the binding sites (of the protein of interest)?

Word of caution!

ChIP-seq experiments are more unpredictable than RNA-seq!

Error sources:

chromatin structure

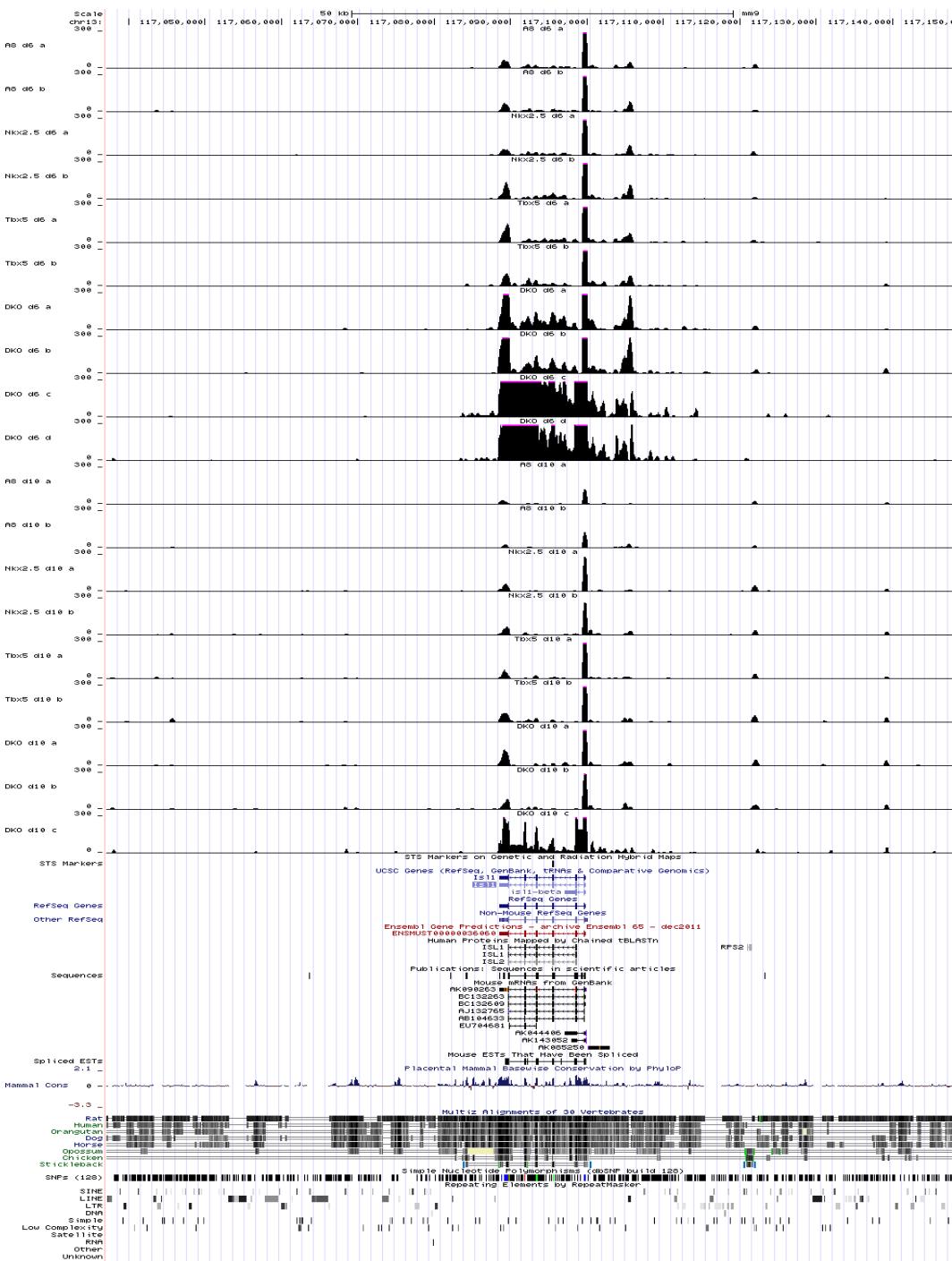
PCR over-amplification

non-specific antibody

other things?

ChIP-seq QC: did the ChIP work?

- 1. Inspect the signal (mapped reads, coverage profiles) in genome browser
- 2. Compute peak-independent quality metrics (cross correlation, cumulative enrichment)
- 3. Assess replicate consistency (correlations between replicates of the same condition)

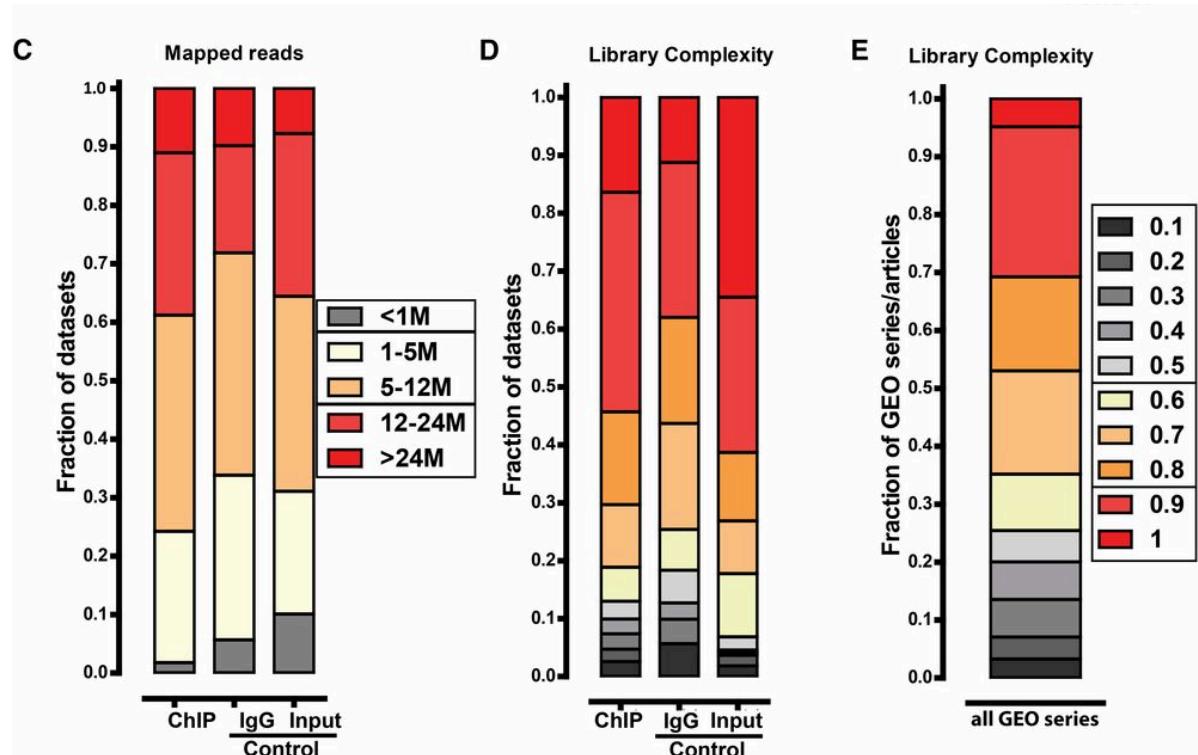
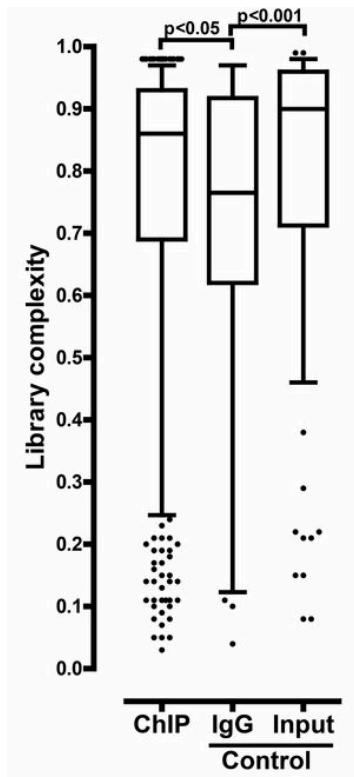


tag density distribution
reproducibility
similarity of coverage
signal at known sites

...
Spotting inconsistencies
Confounding factors
Under-sequenced libraries

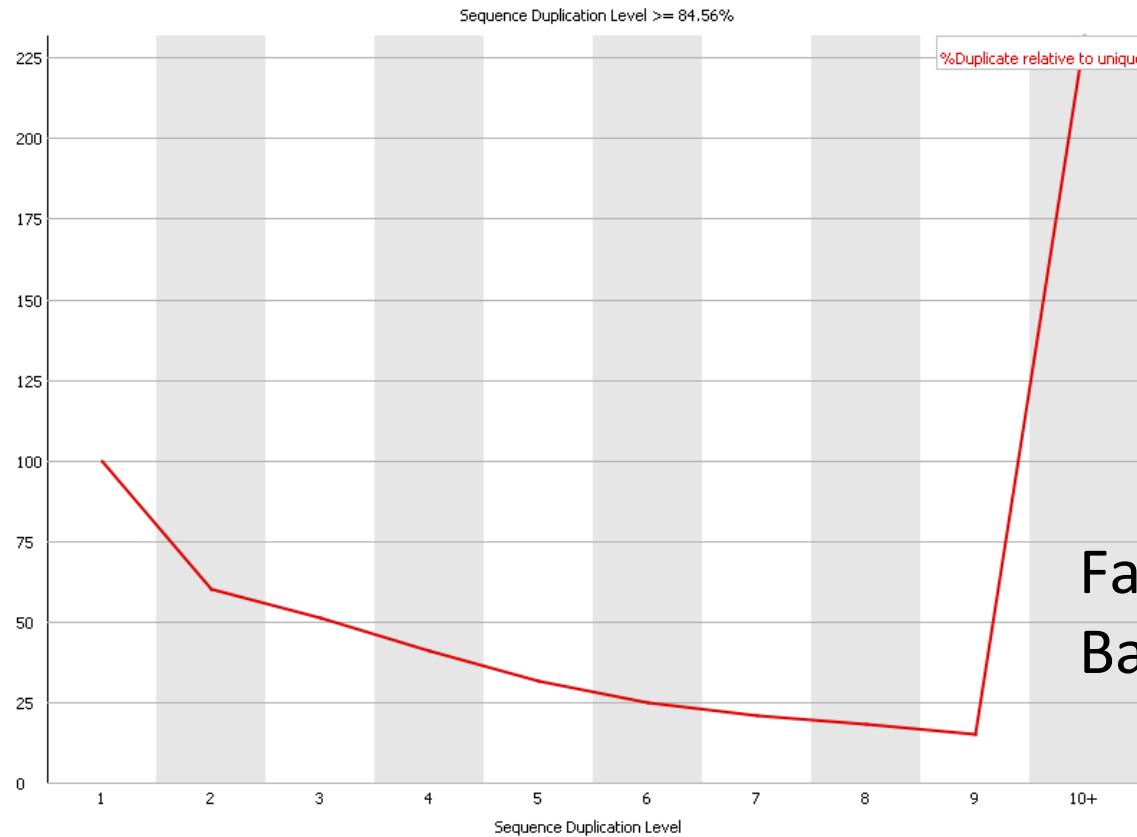
How do I know my data is of good quality?

Library complexity



Quality control: tag uniqueness – library complexity metric

Sequence duplication level > 80% (low complexity library)



NRF: Non-redundant fraction (of reads): proportion of unique tags / total

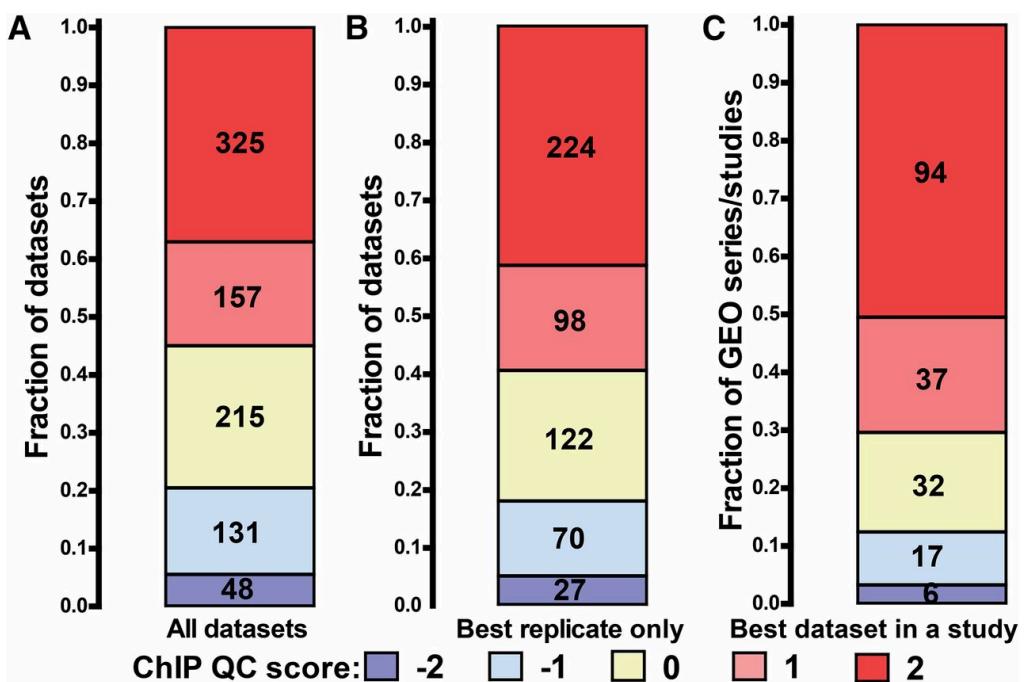
How do I know my data is of good quality?

Objective (i.e. peak independent) metrics to quantify enrichment in ChIP-seq;

for TF in mammalian systems:

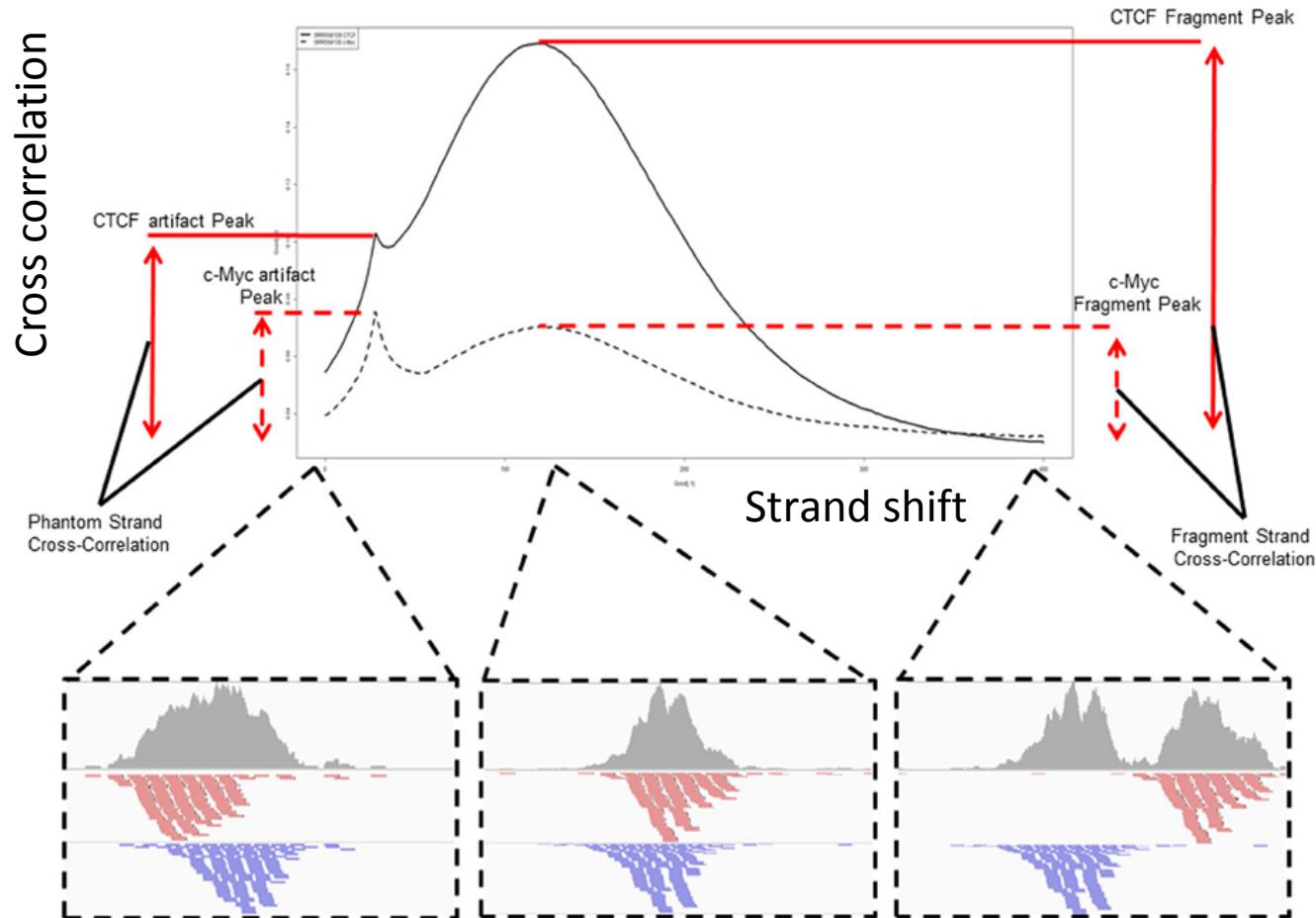
Normalised Strand Correlation NSC
Relative Strand Correlation RSC

Large-scale quality analysis of published ChIP-seq data sets:
20% low quality
25% intermediate quality
30% inputs have metrics similar to IPs

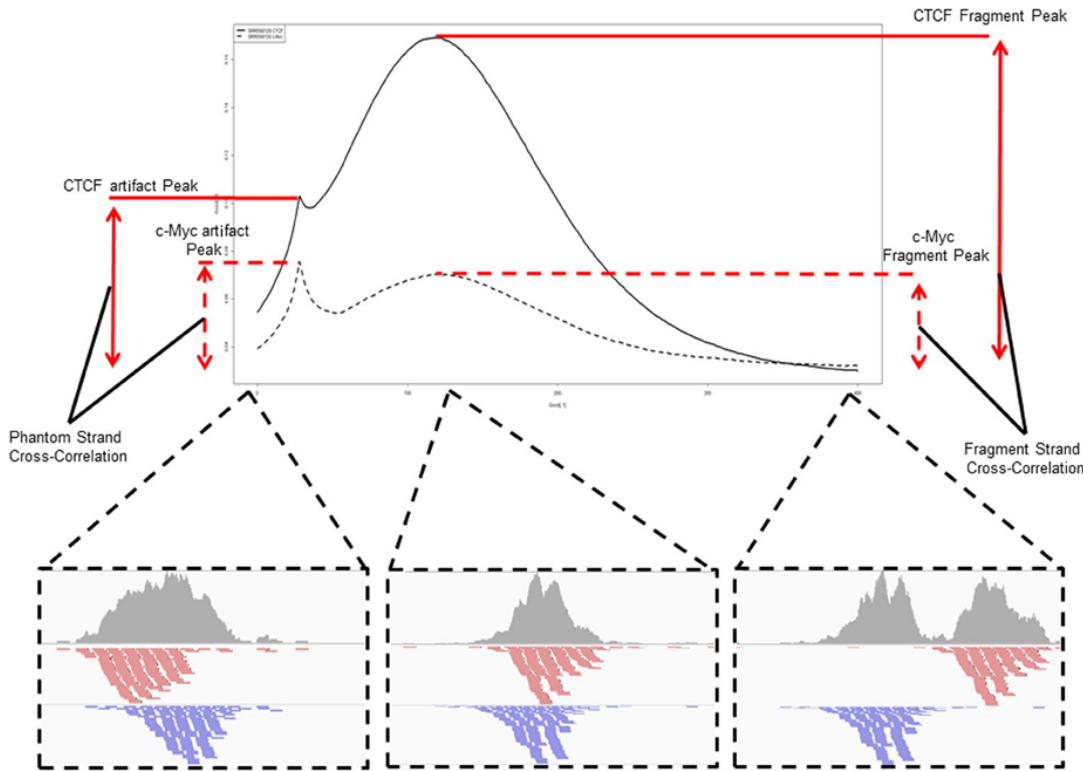


Strand cross-correlation

The correlation between signal of the 5' end of reads on the (+) and (-) strands is assessed after successive shifts of the reads on the (+) strand and the point of maximum correlation between the two strands is used as an estimation of fragment length.



Strand cross-correlation

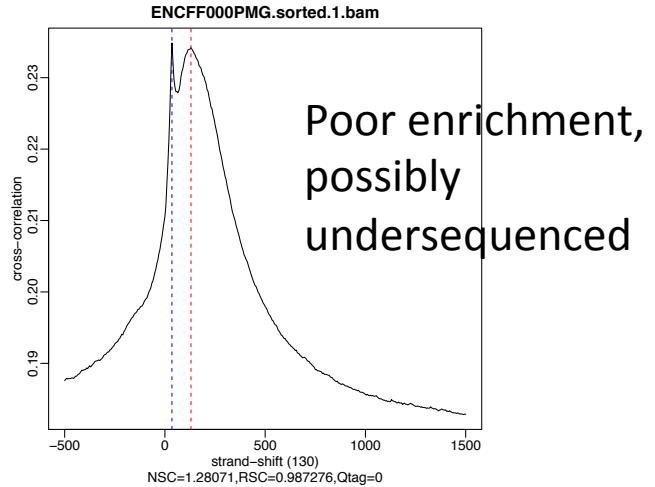
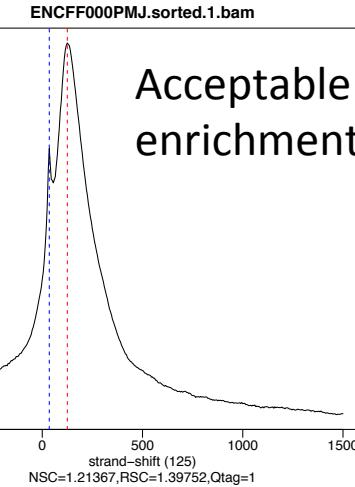
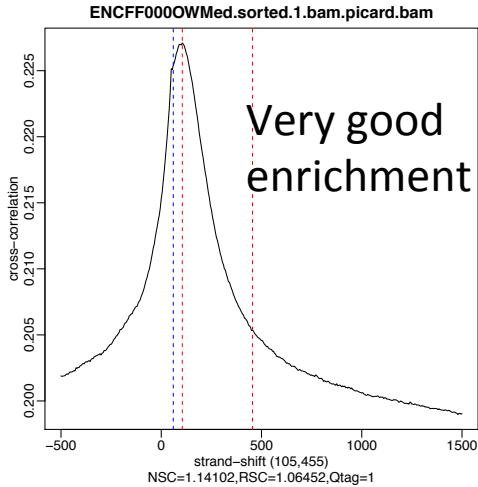


$$NSC = \frac{\text{Max CC value (fLen)}}{\text{Min CC}}$$

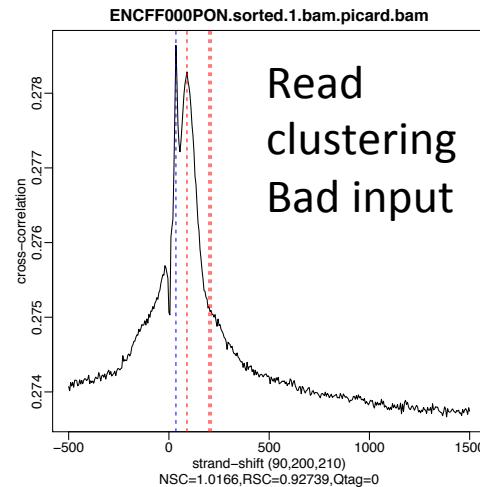
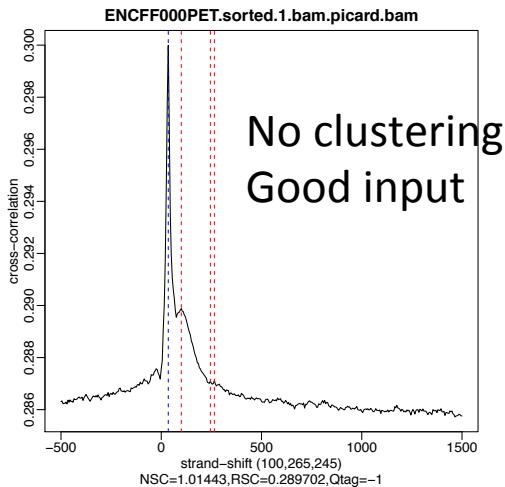
$$RSC = \frac{\text{Max CC} - \text{Min CC}}{\text{Phantom CC} - \text{Min CC}}$$

Cross-correlation plots

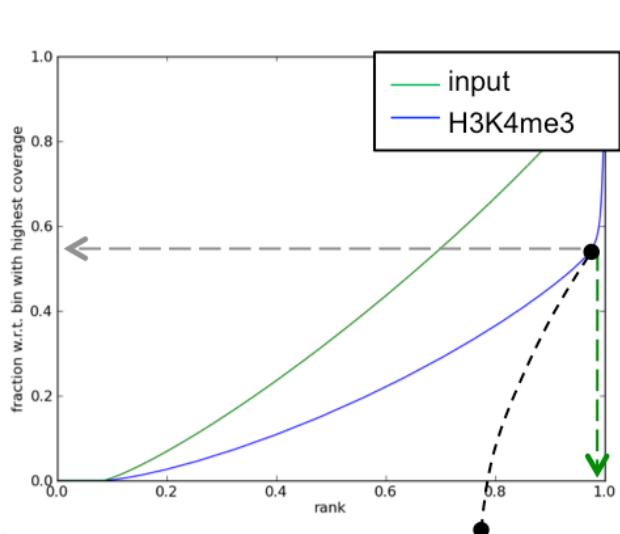
ChIP



Input

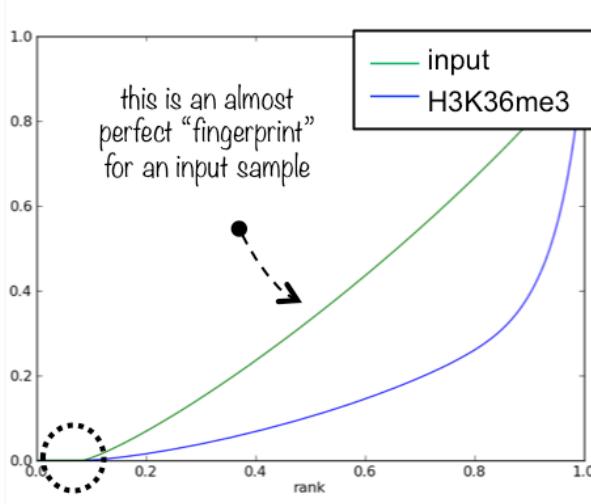


Cumulative enrichment aka “Fingerprint” is another metric for successful enrichment

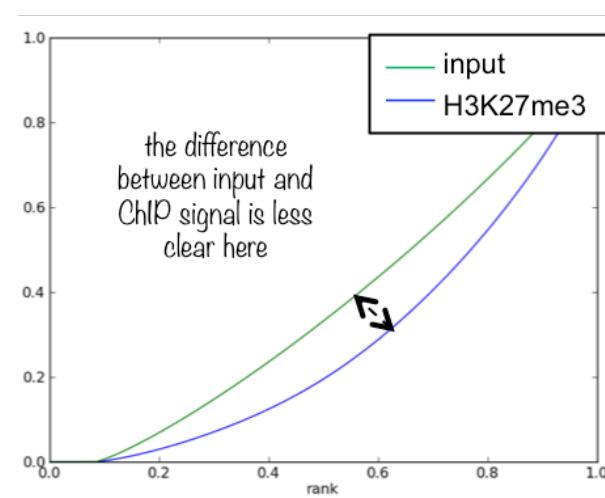


when counting the reads contained in 97% of all genomic bins, only ca. 55% of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

→ this indicates very localized, very strong enrichments!
(as every biologist hopes for in a ChIP for H3K4me3)

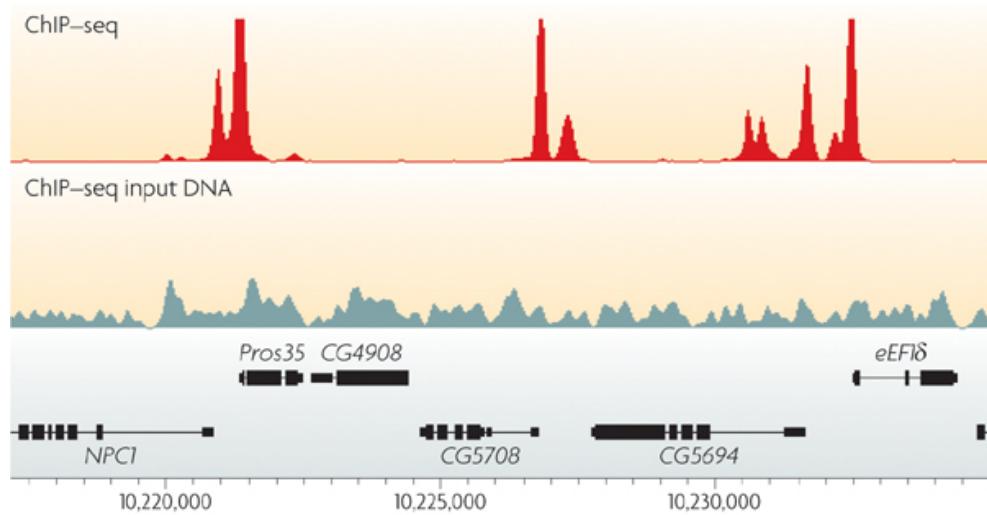
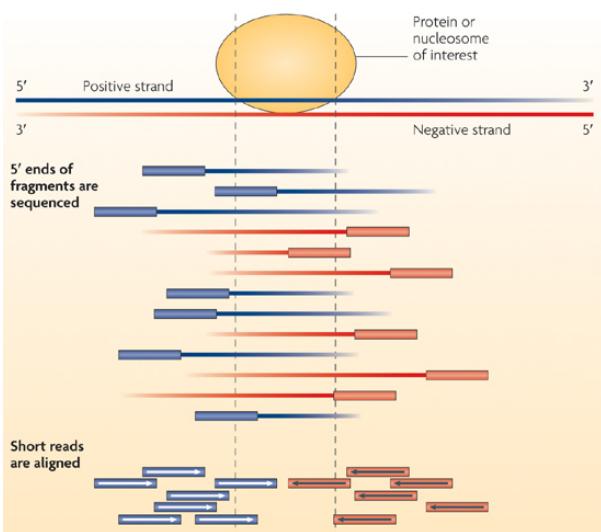


pay attention to where the curves start to rise – this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads – for this example, ca. 10% of the entire genome do not have any read)



H3K27me3 is a mark that yields broad domains instead of narrow peaks

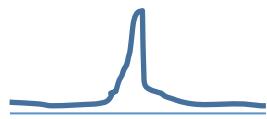
it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed



Peak calling

appropriate methodologies depend on data type

Transcription
Factors



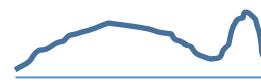
punctate
SPP
MACS2

Chromatin
Remodellers
Histone marks



mixed signal

Chromatin
Remodellers
Histone marks
RNA polymerase II

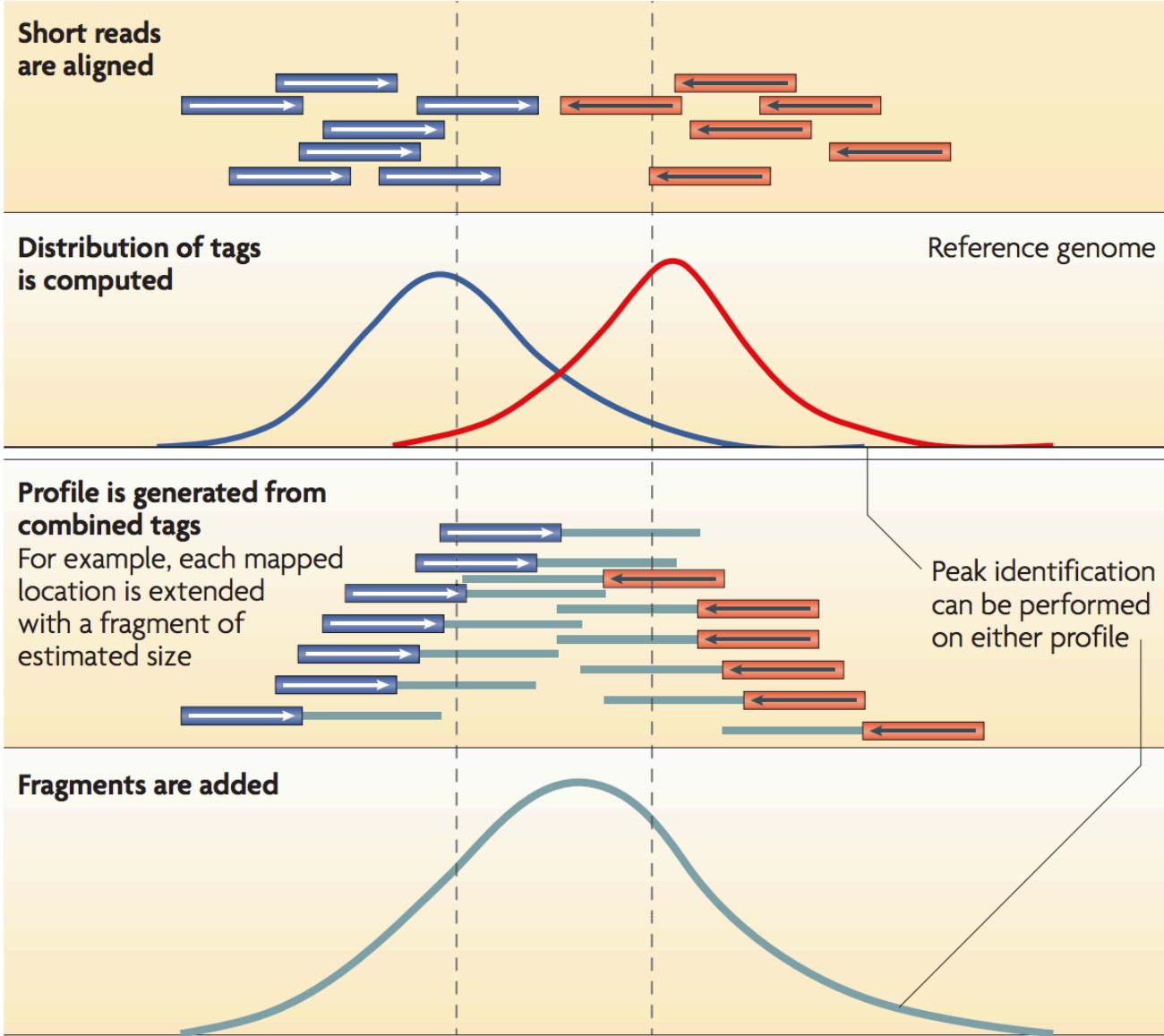


broad signal

MACS2 in broad mode, windows approaches

This is an active area of algorithm development

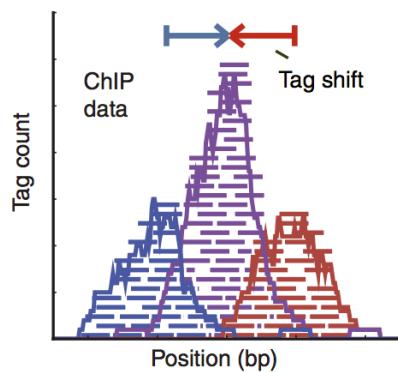
Principle of peak detection



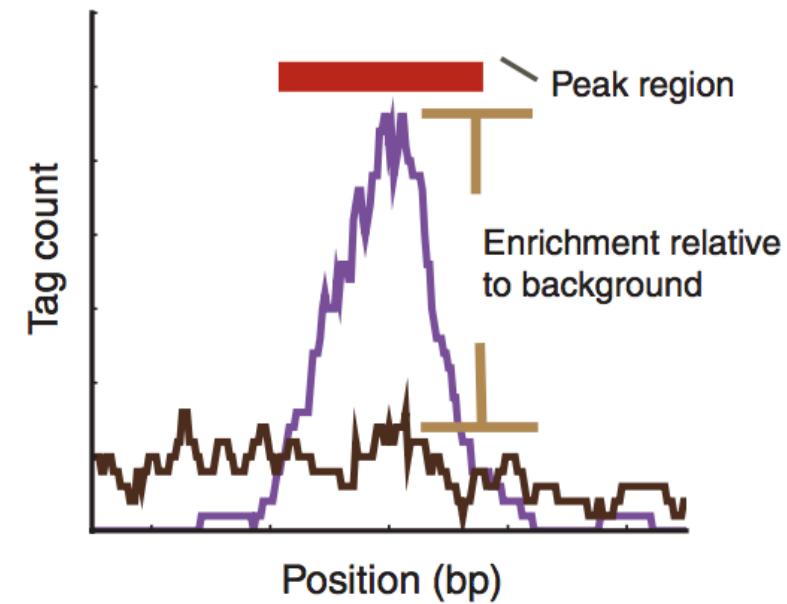
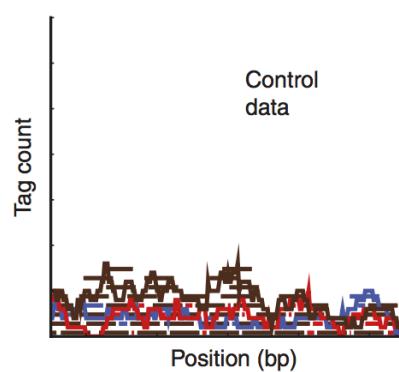
Symmetry in reads mapped to opposite DNA strands

Computation of enrichment model

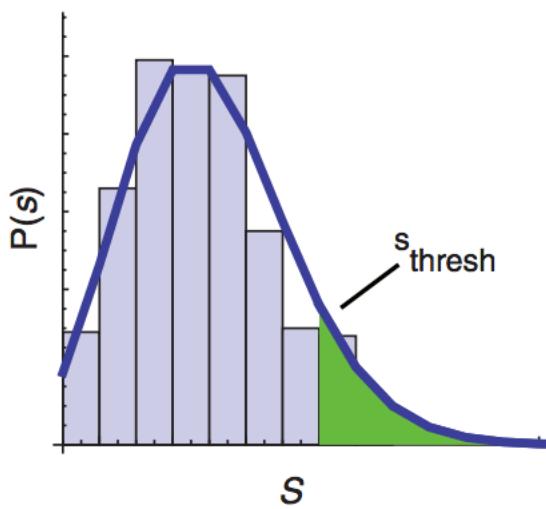
Generate signal profile
along each chromosome



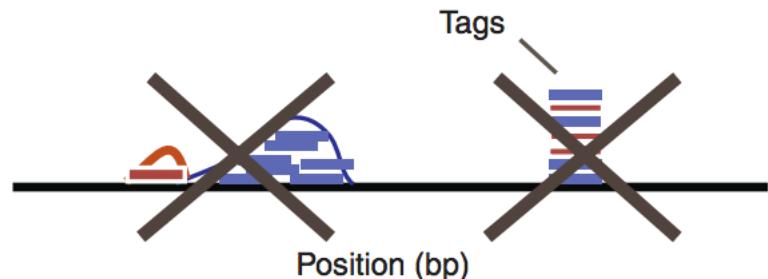
Define background
(model or data)



Assess significance

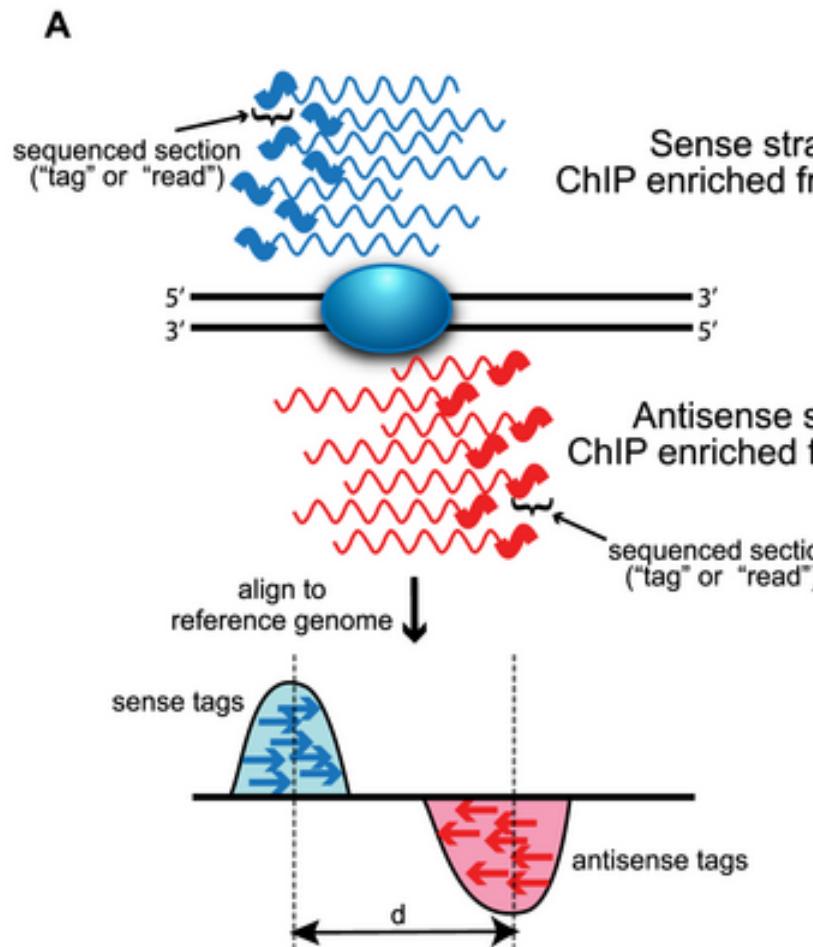


Filter artifacts

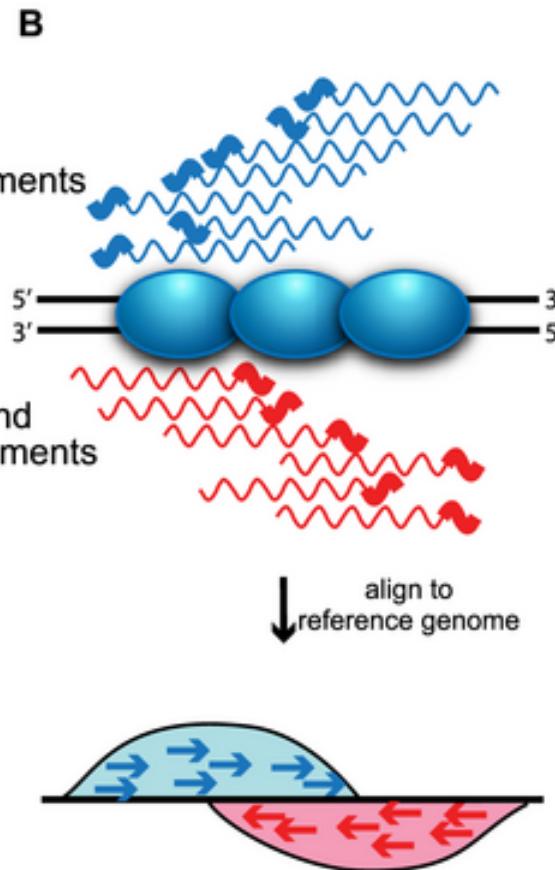


Point-source vs. broad peak detection

Sequence-specific binding (TFs)

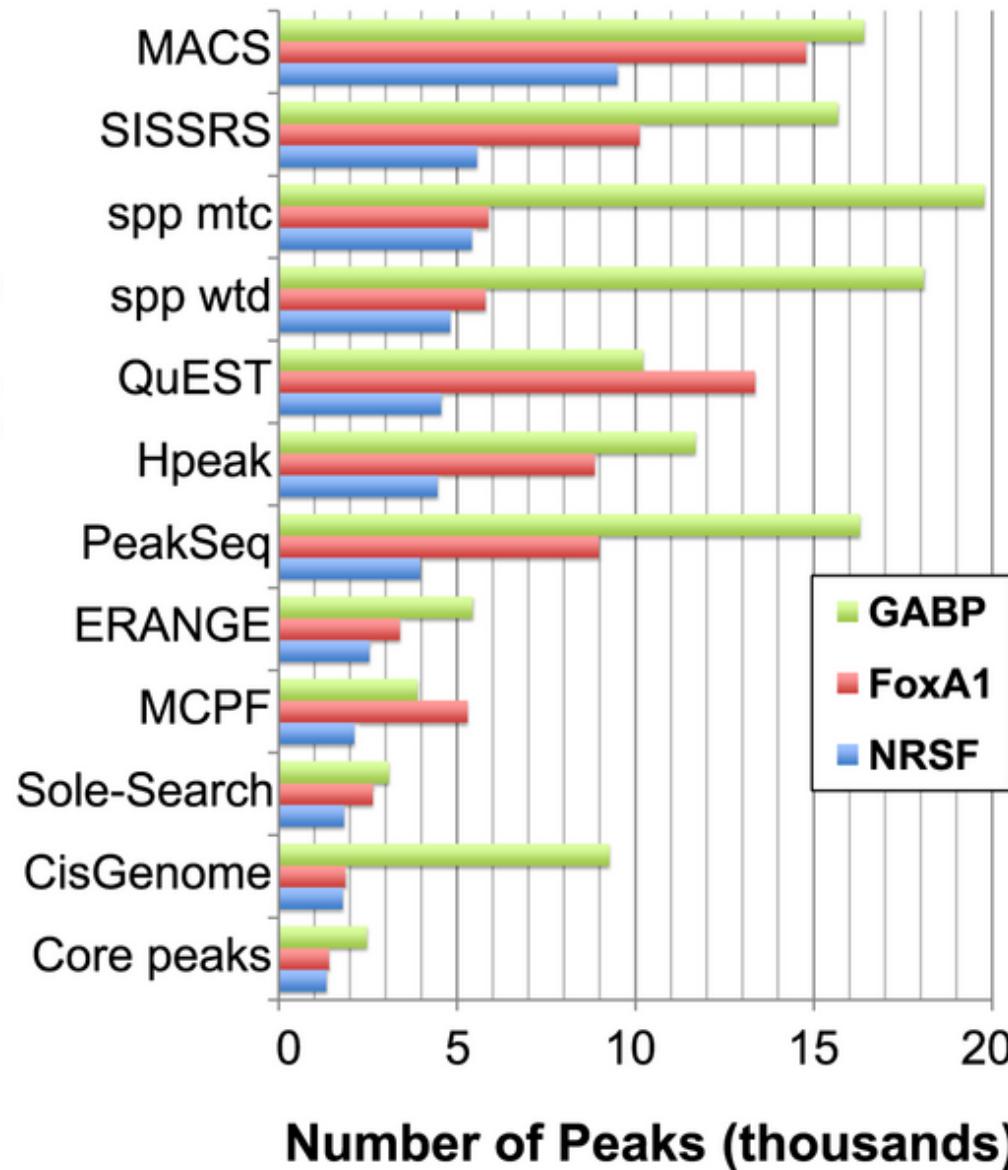


Distributed binding (histones, RNAPol2)

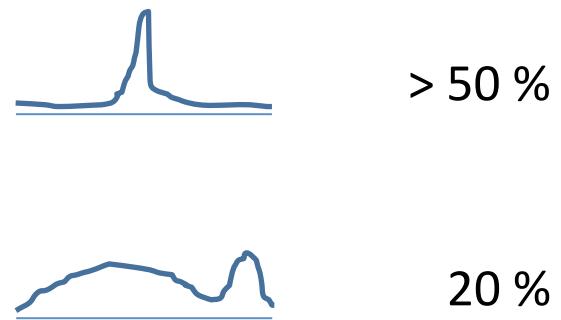


Comparison of peak calling algorithms

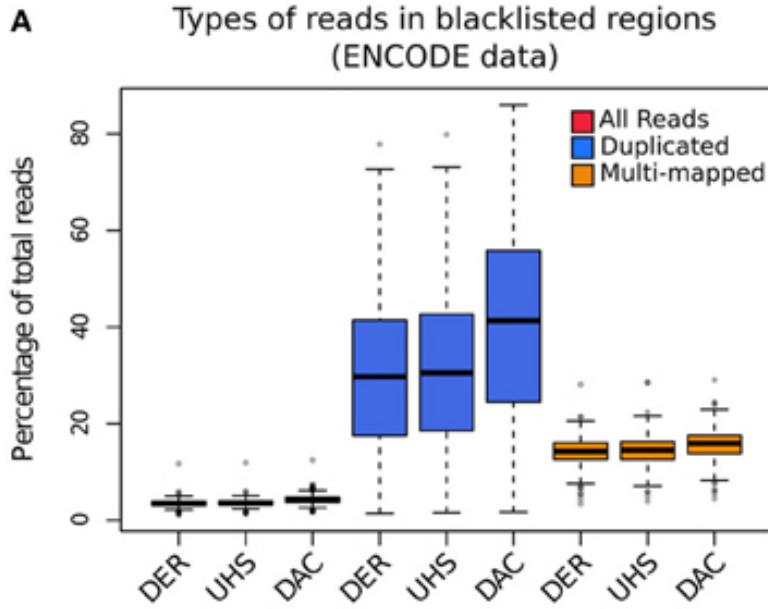
Peak calling program



Peak overlap (Ho et al, 2012)



“Hyper-chippable” regions



Reads mapped to these regions should be filtered out prior to peak calling

Tracks available from UCSC for human, mouse, fly and worm

DER – Duke Excluded Regions

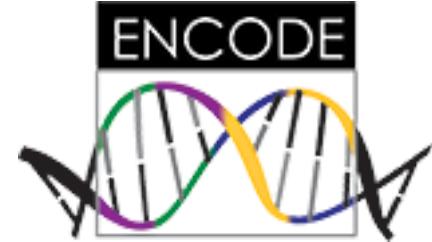
(11 repeat classes)

UHS – Ultra High Signal

(open chromatin)

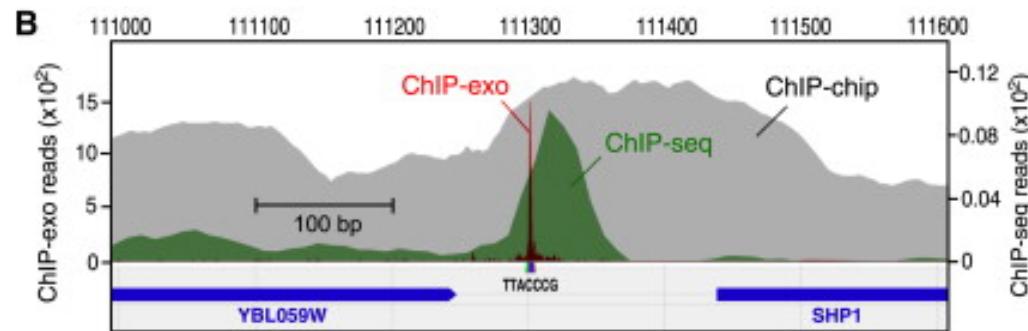
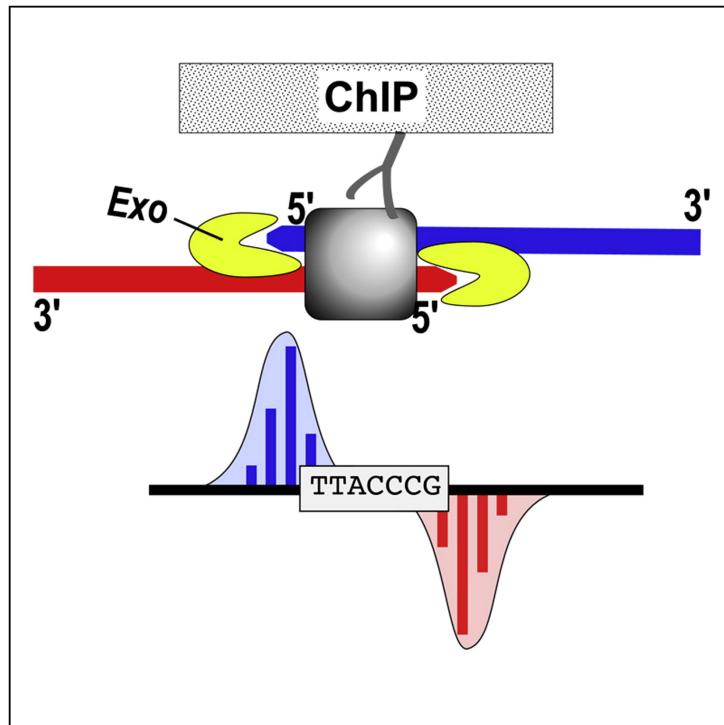
DAC – consensus excluded regions

Quality considerations

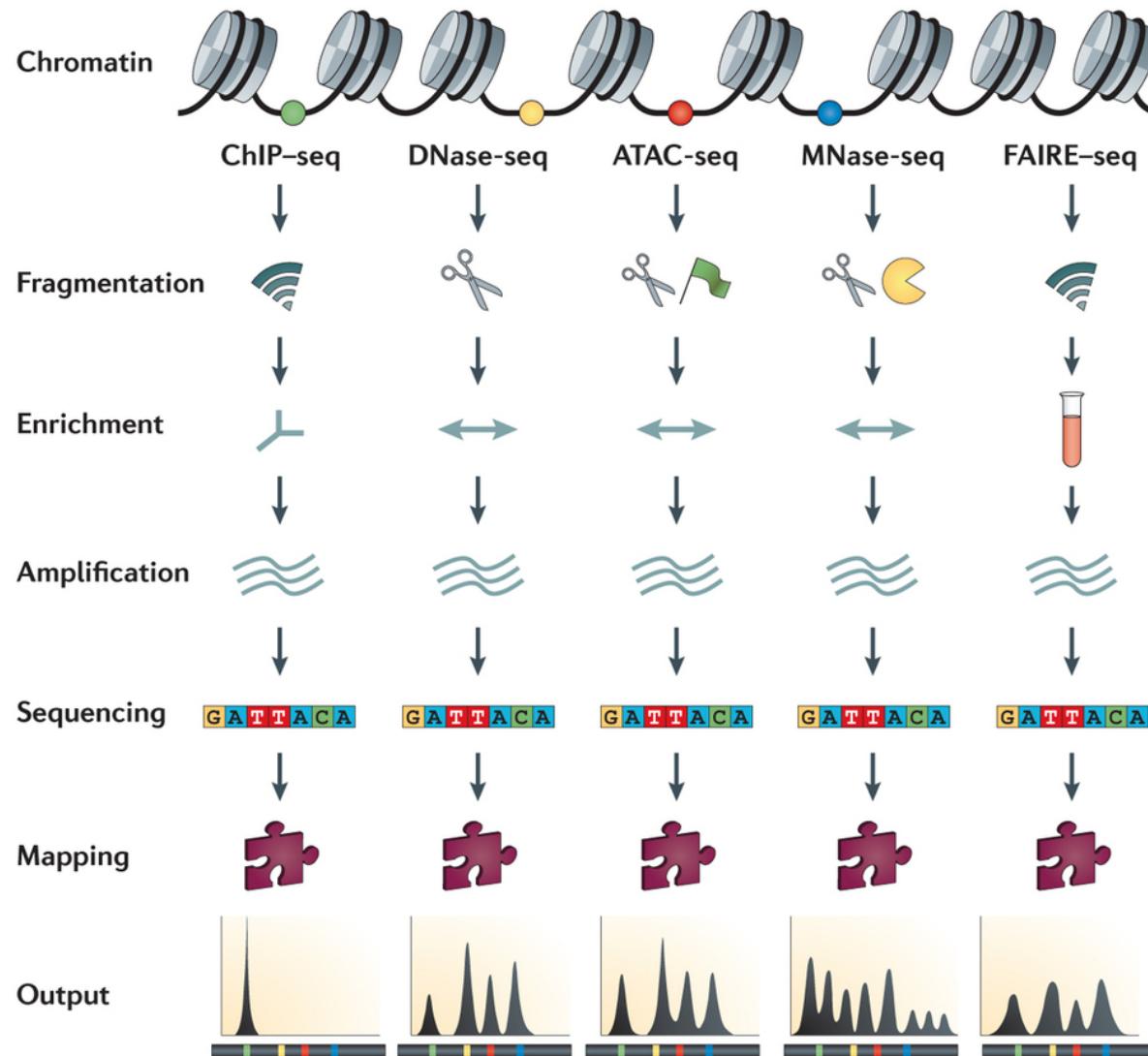


- ChIP-seq quality guidelines from the ENCODE project (Relative strand cross-correlation, Irreproducible discovery rate)
- Antibody validation
- Appropriate sequencing depth (depending on genome size and peak type). For human genome and broad-source peaks, min. 40-50M reads is required.
- Experimental replication
- Fraction of reads in peaks (FRiP) > 1%
- Cross correlation (correlation of the density of sequences aligned to opposite DNA strands after shifting by the fragment size)
- Experimental verification of known binding sites (and sites not bound as negative controls)

ChIP-exo: improvement in binding site identification



Other functional genomics techniques

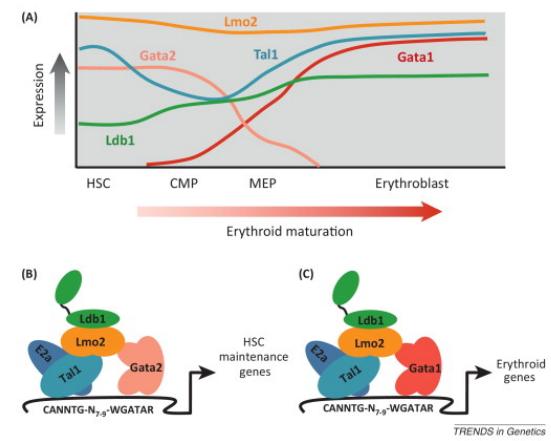
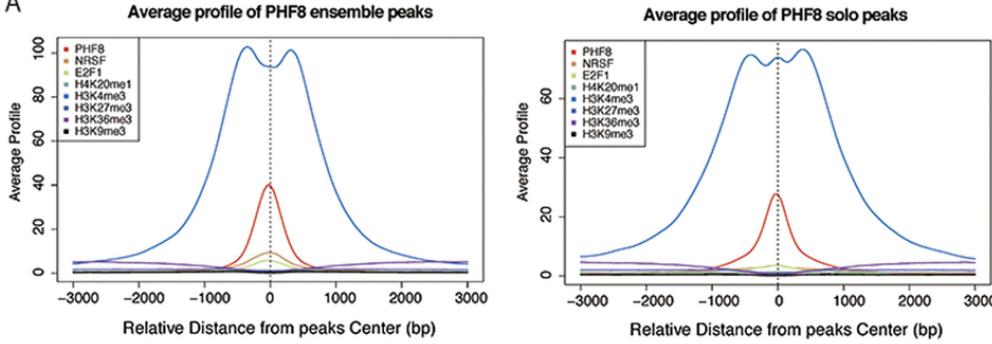


- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

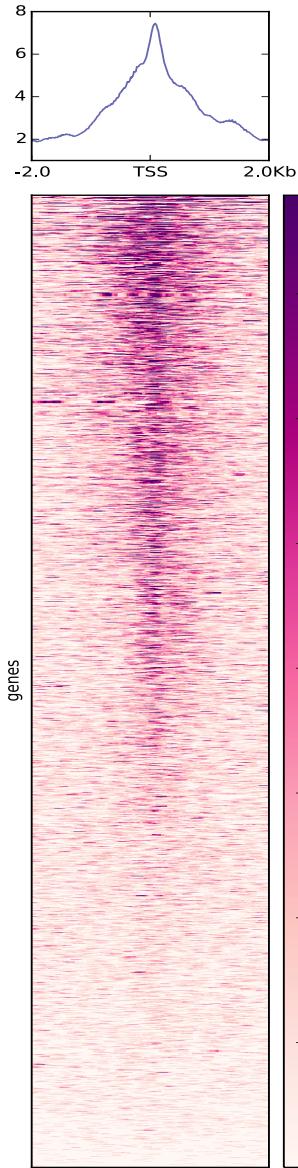
ChIPseq downstream analyses

- Validation (wet lab)
- Downstream analysis
 - Motif discovery
 - Annotation
 - Integration of binding and expression data
 - Integration of various binding datasets
 - Differential binding

A



Signal visualisation and interpretation



deepTools

ngsplots

seqMiner

- Clustering
- Heatmaps
- Profiles
- Comparison of different datasets

Binding profile of a TF in relation to the transcription start site

- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

Further reading

- Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Carroll et al, Front. Genet. 2014
- Impact of sequencing depth in ChIP-seq experiments. Jung et al, NAR 2014
- ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Landt et al, Genome Res. 2012
- <http://genome.ucsc.edu/ENCODE/qualityMetrics.html#definitions>
- <https://www.encodeproject.org/data-standards>

Bioconductor ChIP-seq resources

- General purpose tools:
 - Rsubread (read mapping; not ideal for global alignment)
 - Rbowtie (global alignment)
 - GenomicRanges (tools for manipulating range data)
 - Rsamtools (SAM / BAM support)
 - htSeqTools (tools for NGS data; post-alignment QC)
 - chipseq (utilities for ChIP-seq analysis)
- Peak calling
 - SPP
 - BayesPeak (HMM and Bayesian statistics)
 - MOSAiCS (model-based one and two Sample Analysis and Inference for ChIP-Seq)
 - iSeq (Hidden Ising models)
 - ChIPseqR (developed to analyse nucleosome positioning data)
 - CsaW (a pipeline for ChIP-seq analysis, including statistical analysis of differential occupancy)
- Quality control
 - ChIPQC
- Differential occupancy
 - edgeR
 - DESeq2
 - DiffBind (compatible with objects used for ChIPQC, wrapper for DESeq and edgeR DE functions)
- Peak Annotation
 - ChIPpeakAnno (annotating peaks with genome context information)
 - ChIPSeeker (functional annotation of peaks)

The Epigenomics Roadmap Project



<http://www.roadmapepigenomics.org/>

- Reference human epigenomes
- DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts
- Stem cells and primary *ex vivo* tissues
- 111 tissue and cell types
- 2,804 genome-wide datasets

Questions?

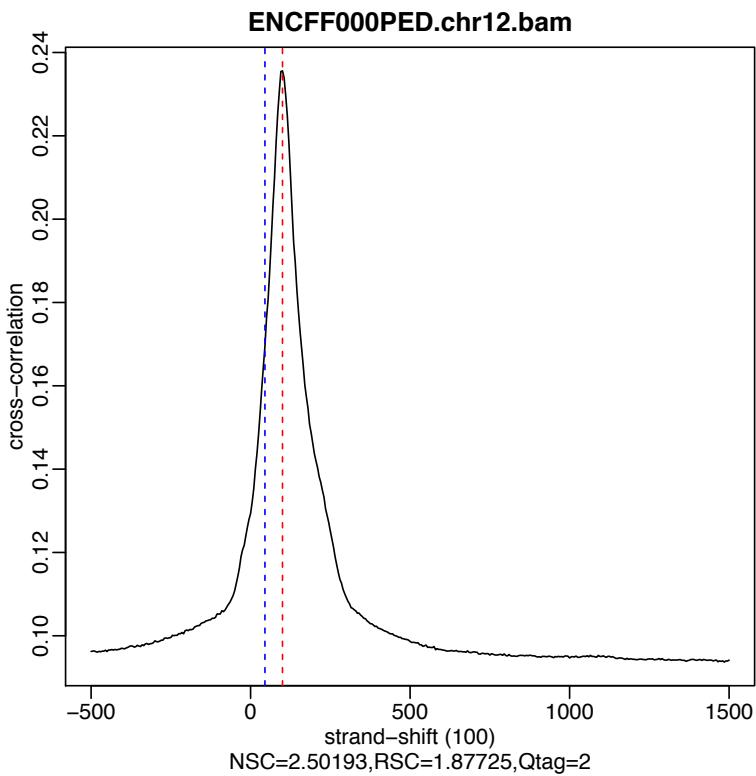
agata.smialowska@nbis.se

- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources
- Exercise overview

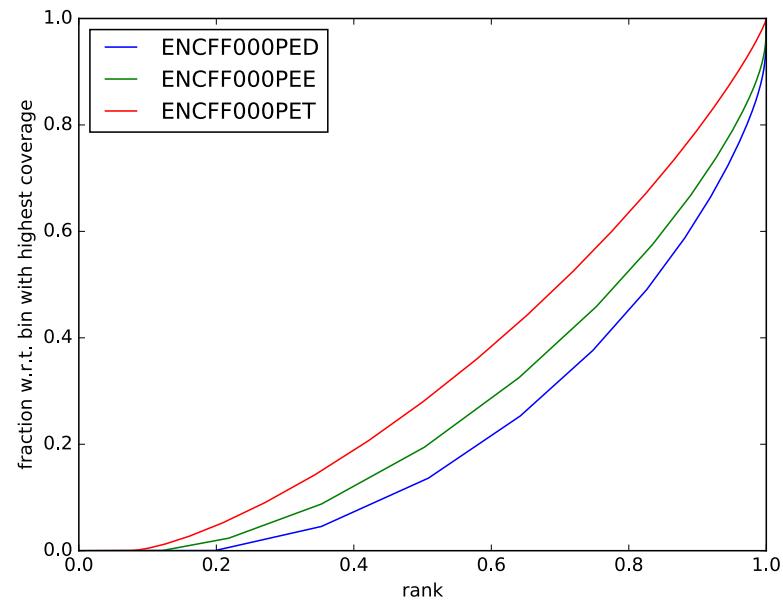
Exercise

- 1. Quality control
- 2. Read preprocessing
- 3. Peak calling
- 4. Exploratory analysis (sample clustering)
- 5. Visualisation

Did my ChIP work?

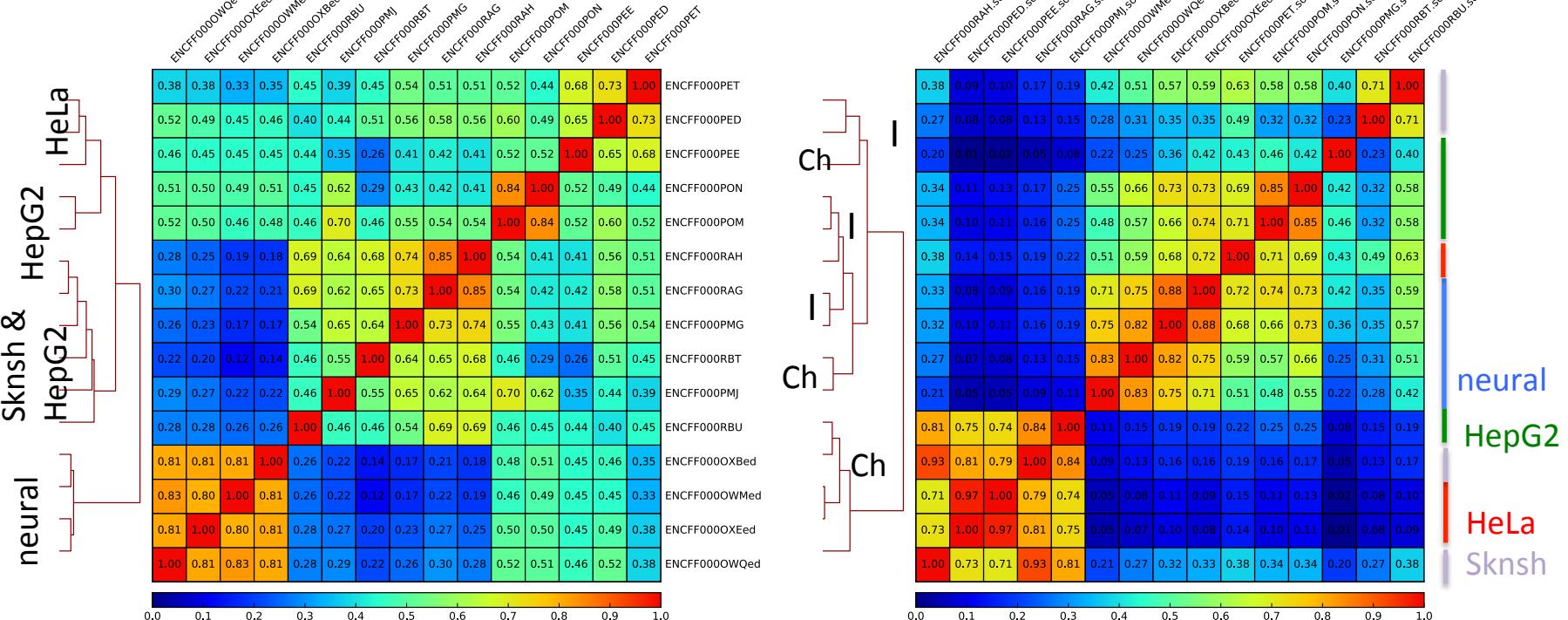


Cross-correlation



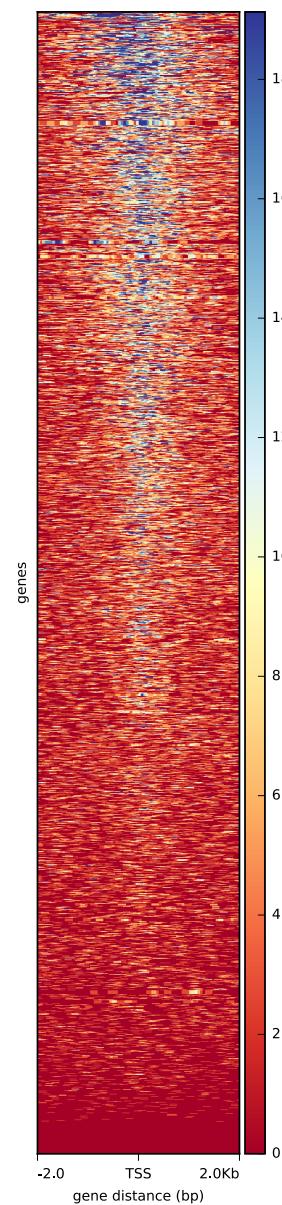
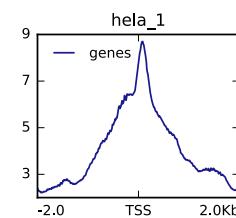
Cumulative enrichment

Exploratory analysis



Clustering of libraries
by reads mapped in bins,
genome – wide (spearman)

Clustering of libraries
by reads mapped in peaks
(pearson)



Binding profile around TSS

That's all for now,
time to do some hands-on work

Library quality control and preprocessing

- FastQC / Prinseq
- Trim adapters if any adapter sequences are present in the reads (as determined by the QC)
- In some cases, you'll observe k-mer enrichment (especially if the data is ChIP-exo, a new variation of ChIP-seq) – it is not necessarily a bad thing, if sequence duplication levels are low; however it may indicate **low complexity of the library** – a warning sign that the enrichment in ChIP was not successful or the libraries are over-amplified (often the latter is the consequence of the former)

Mapping reads to the reference genome

- Choose the right reference: assembly version (not always the newest is best) and type (primary assembly, or assembly from individual chromosome sequences + non-chromosomal contigs; not the top level assembly); choose the matching annotation file (GTF, GFF)
- Read mapping: **global alignment**
- Mappers (= aligners): Bowtie, BWA, BBMap, Novoalign, ... (lots of tools are available)
- Visualise data in genome browser
 - BAM files or tracks (wig, bedgraph, bigWig)
 - Local (IGV) or web-based (UCSC genome browser)
 - Data quality assessment

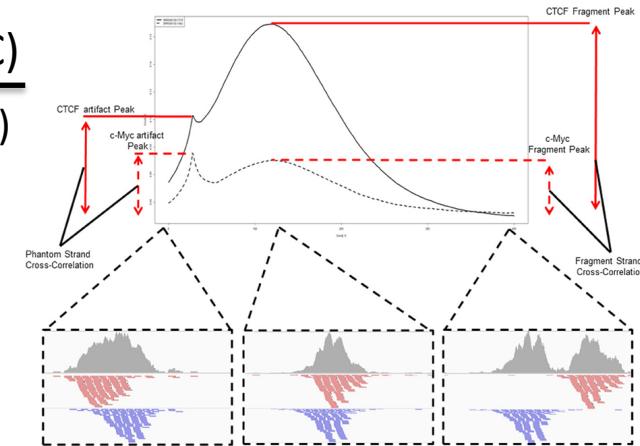
Cross-correlation profiles, RSC and NSC

- Metrics to quantify the fragment length signal and the ratio of fragment length signal to read length signal
- Relative Cross Correlation (RSC) - ChIP to artifact signal

$$\frac{\text{CC(Fragment length)} - \min(\text{CC})}{\text{CC(read length)} - \min(\text{CC})}$$

- Normalised Cross Correlation (NSC)

$$\frac{\text{CC(Fragment length)}}{\min(\text{CC})}$$



- TFs: fragment lengths are often greater than the size of the DNA binding event, the distinct clustering of (+) and (-) reads around this site is very apparent
- NSC>1.1 (higher values indicate more enrichment; 1 = no enrichment)
- RSC>0.8 (0 = no signal; <1 low quality ChIP; >1 high enrichment)
- Broad peaks: this clustering may be more diffuse (fragment length < peak)