

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

Stockholm, November 23 2017

Jakub Orzechowski Westholm

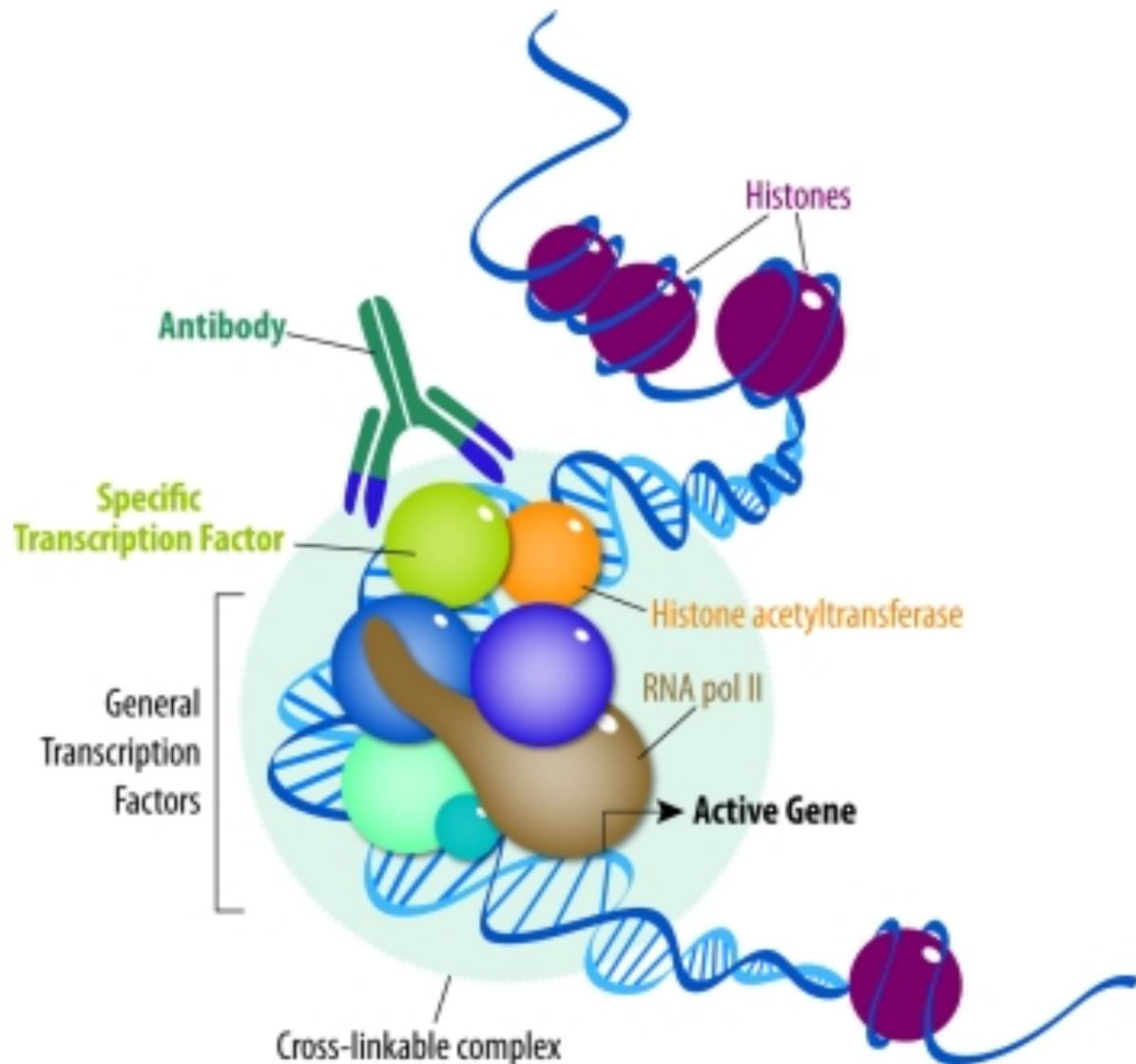
Long-term bioinformatics support

NBIS, SciLifeLab, Stockholm University

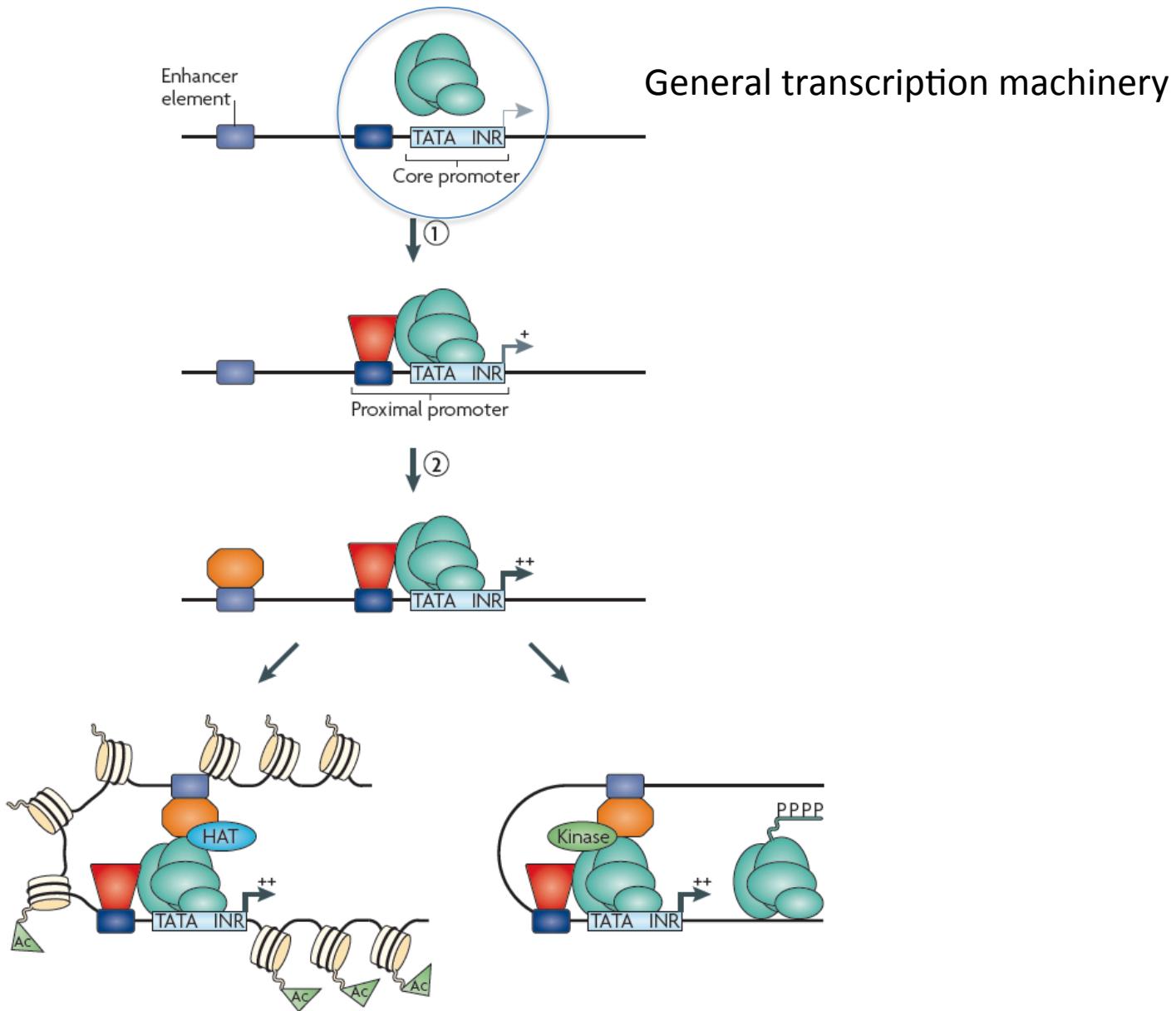
SciLifeLab



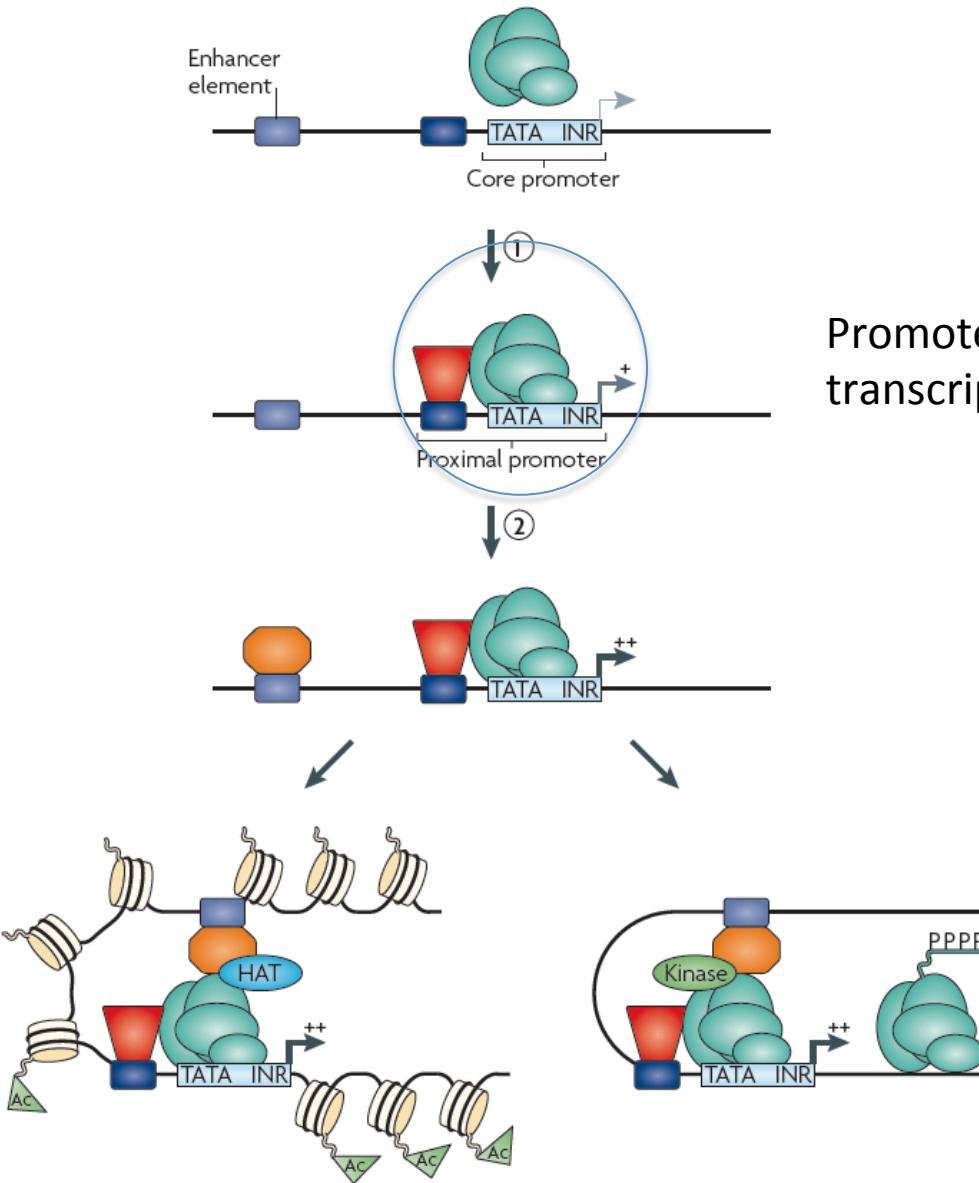
Chromatin immunoprecipitation



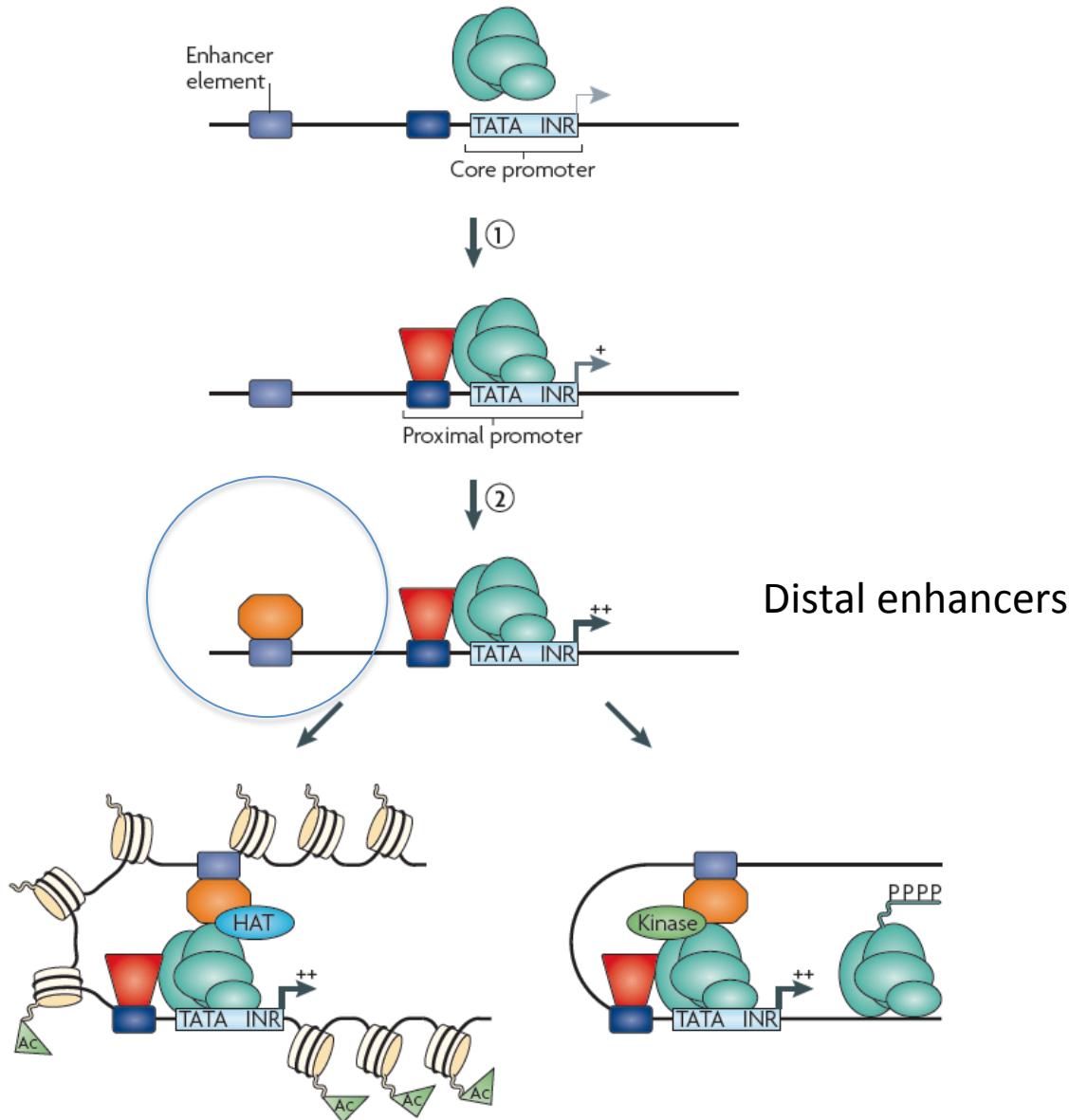
Applications



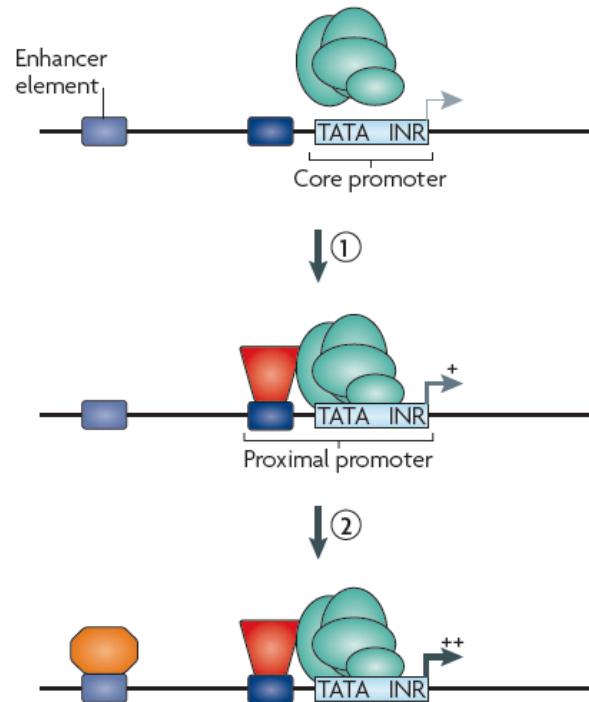
Applications



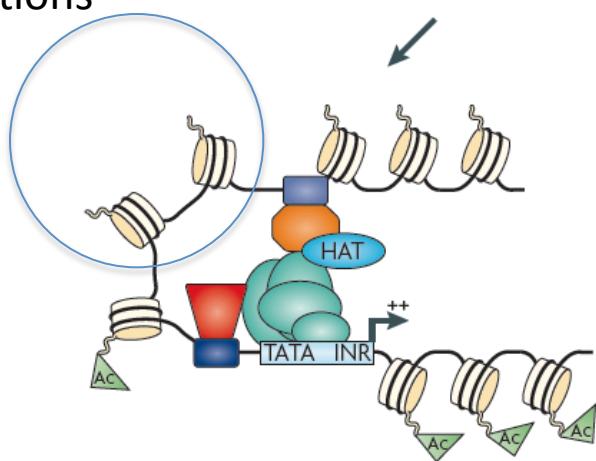
Applications



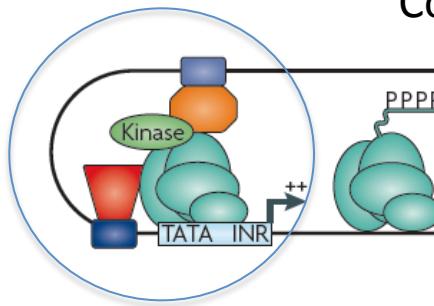
Applications



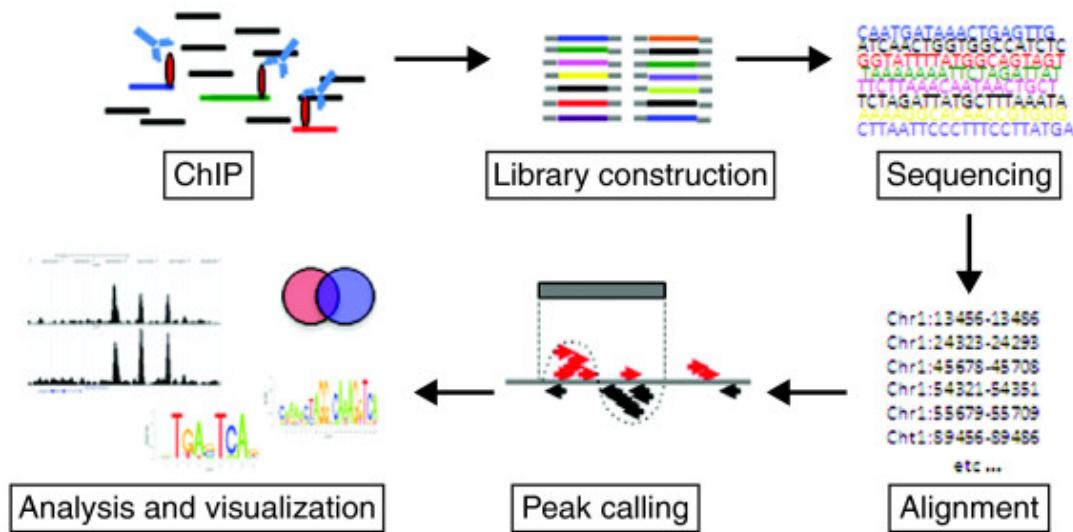
Histone modifications
and variants



Activation states
Co-factors



ChIP-seq workflow



Study design

Critical factors

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

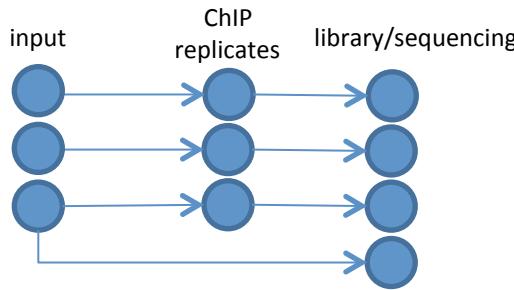
Experiment design

- 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. But some studies show that PE has little effect on end result (Chen et al. Nature Methods 2012).

Experiment design

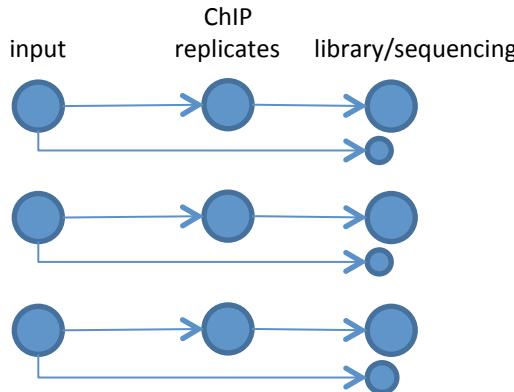
Ideal design:

X

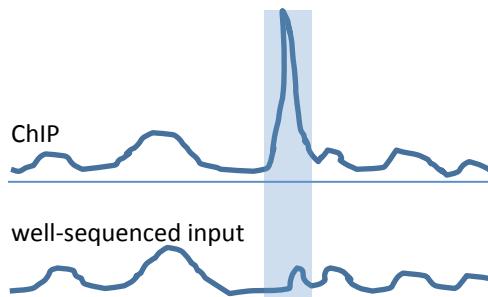
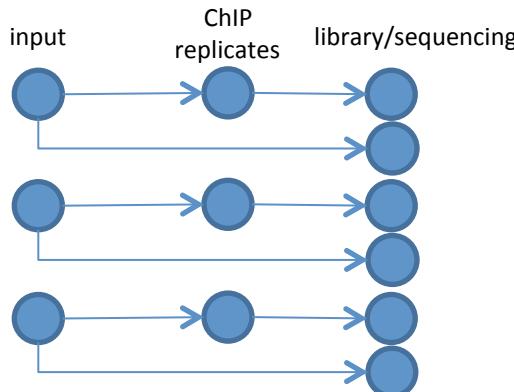
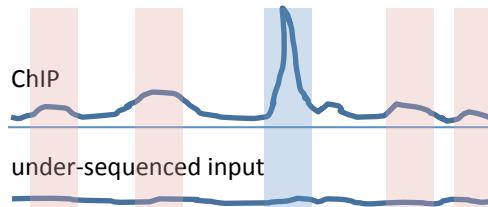


Each sample has a matched input
Input sequenced to a comparable depth
as IP sample

X

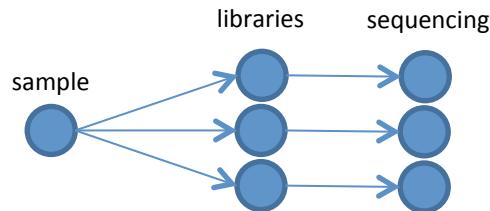


≥2 biological replicates for site identification
≥3 biological replicates for differential binding



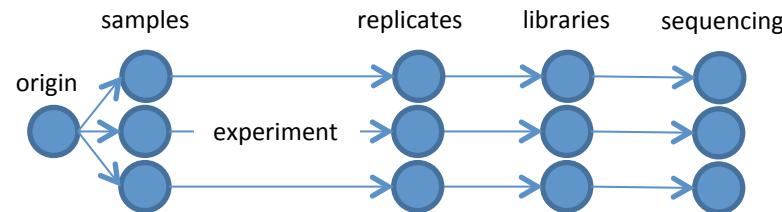
Importance of biological replicates

X



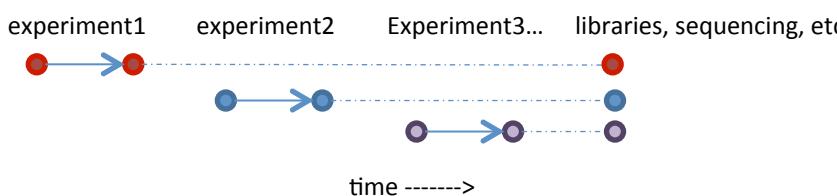
technical replicates are generally a waste of time and money

X



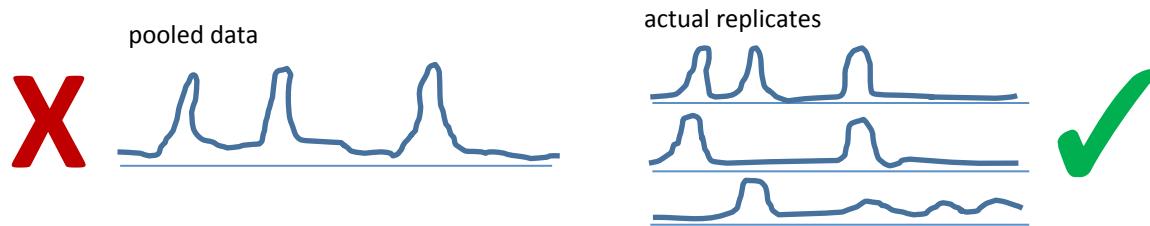
many studies do not account for batch effects
i. time
ii. origin

✓

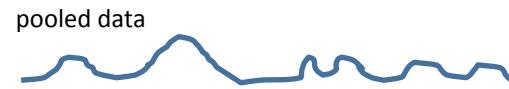
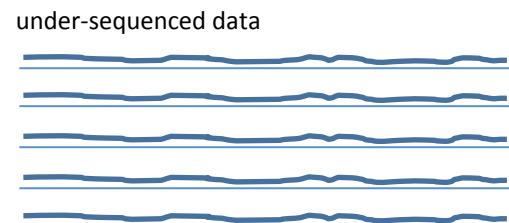


so if you care about reproducibility

Importance of sequencing depth

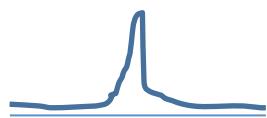


if you need to pool your data, then it is under-sequenced



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

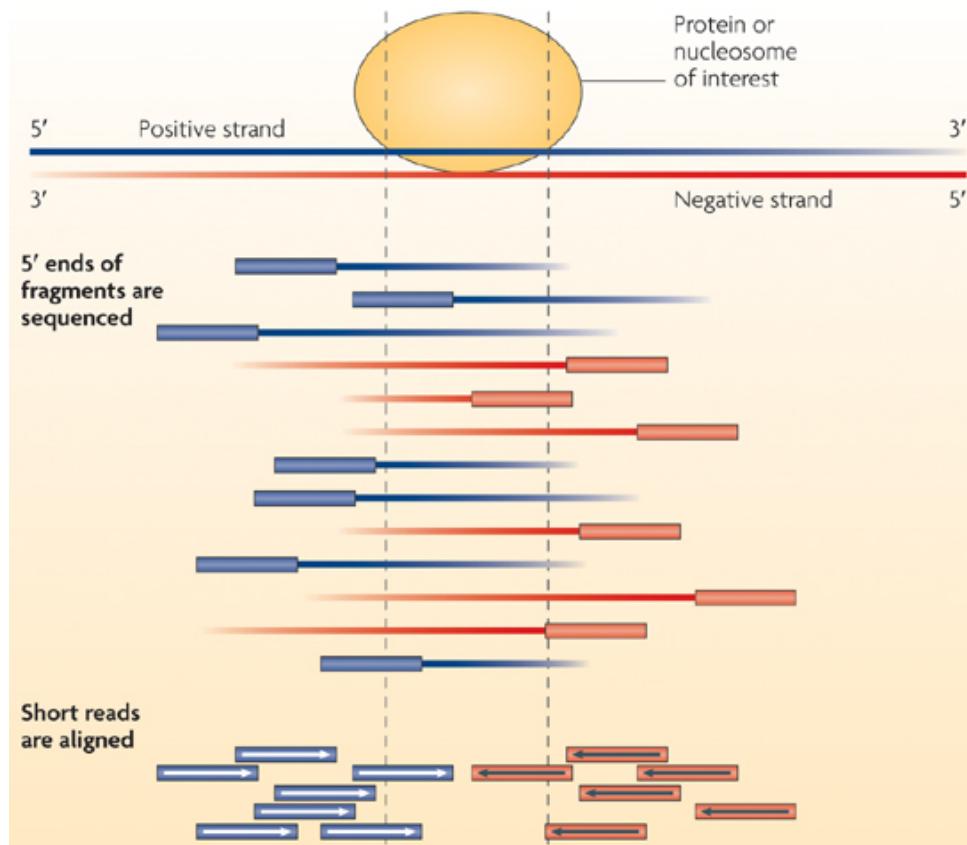
No clear guidelines for mixed and broad type of peaks

ChIP-seq data analysis

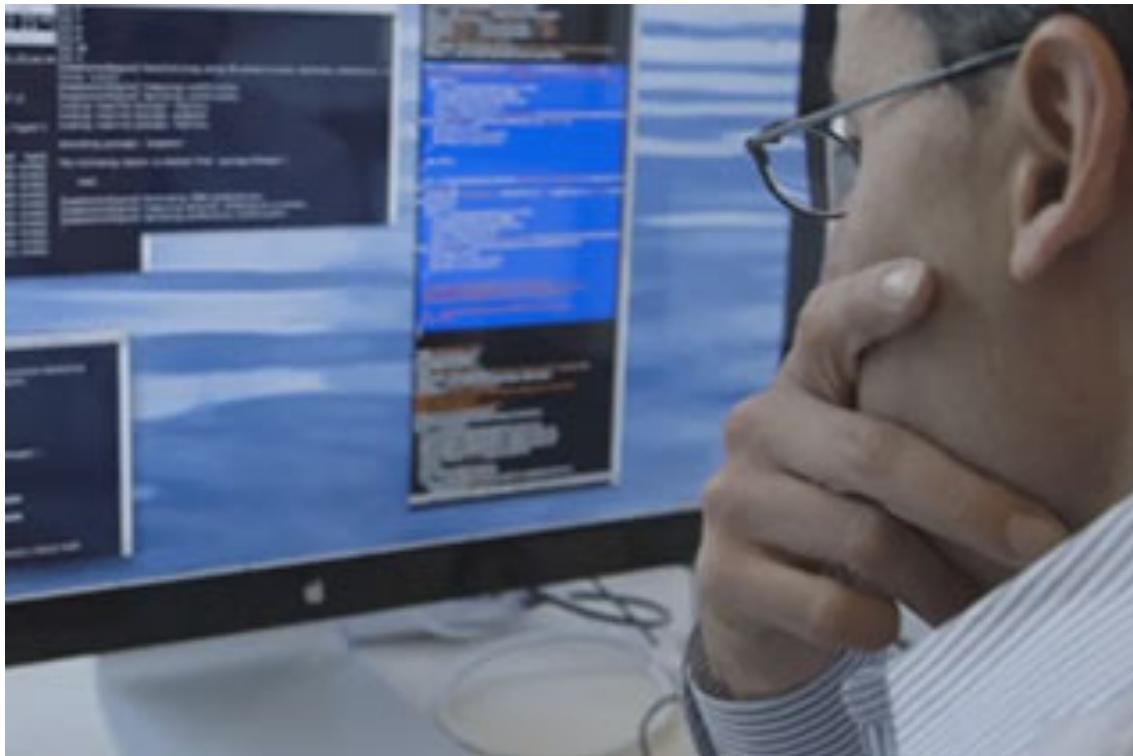
- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources
- Exercise overview

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

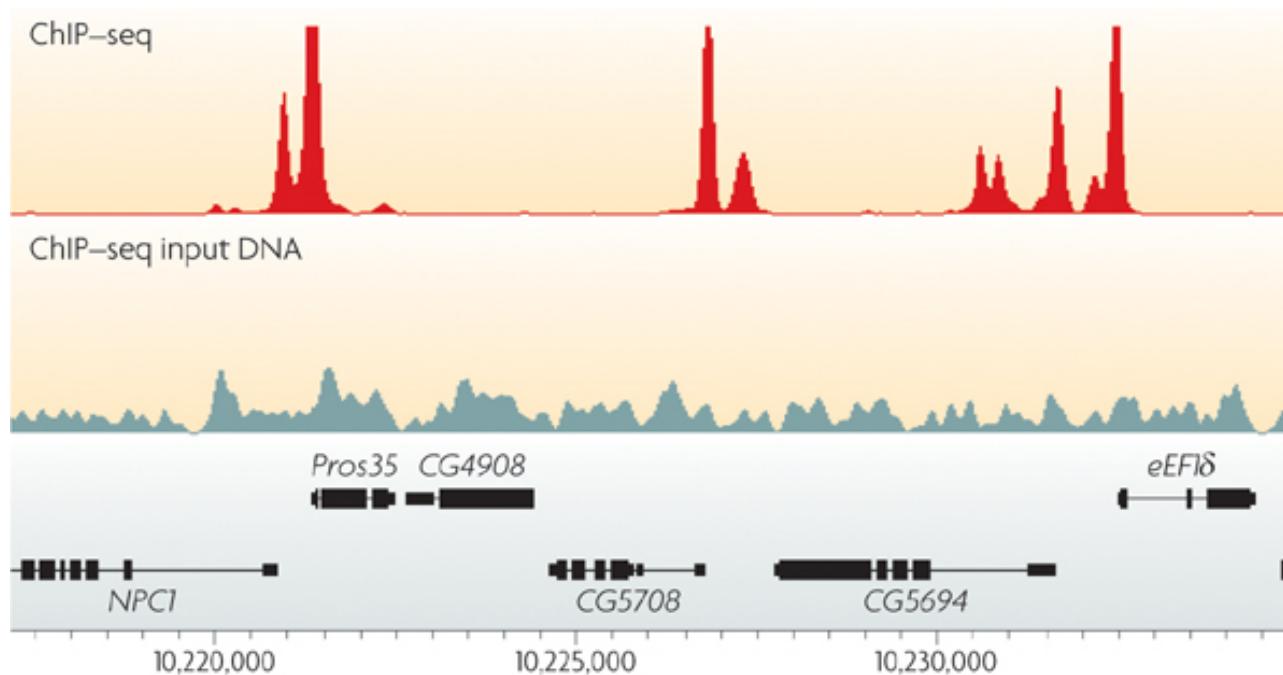
Chromatin = DNA + proteins



Data analysis



Profile of protein binding sites vs. input



Chromator (*Drosophila*) – protein binding methylated histones

design study

obtain input chromatin

perform precipitation

construct library

sequence library

library quality control

filter sequences

align sequences

filter alignments

identify peaks / regions of enrichment

assess data quality

understand the data / results

downstream analyses

Workflow of a ChIP-seq study



Iterative process

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

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

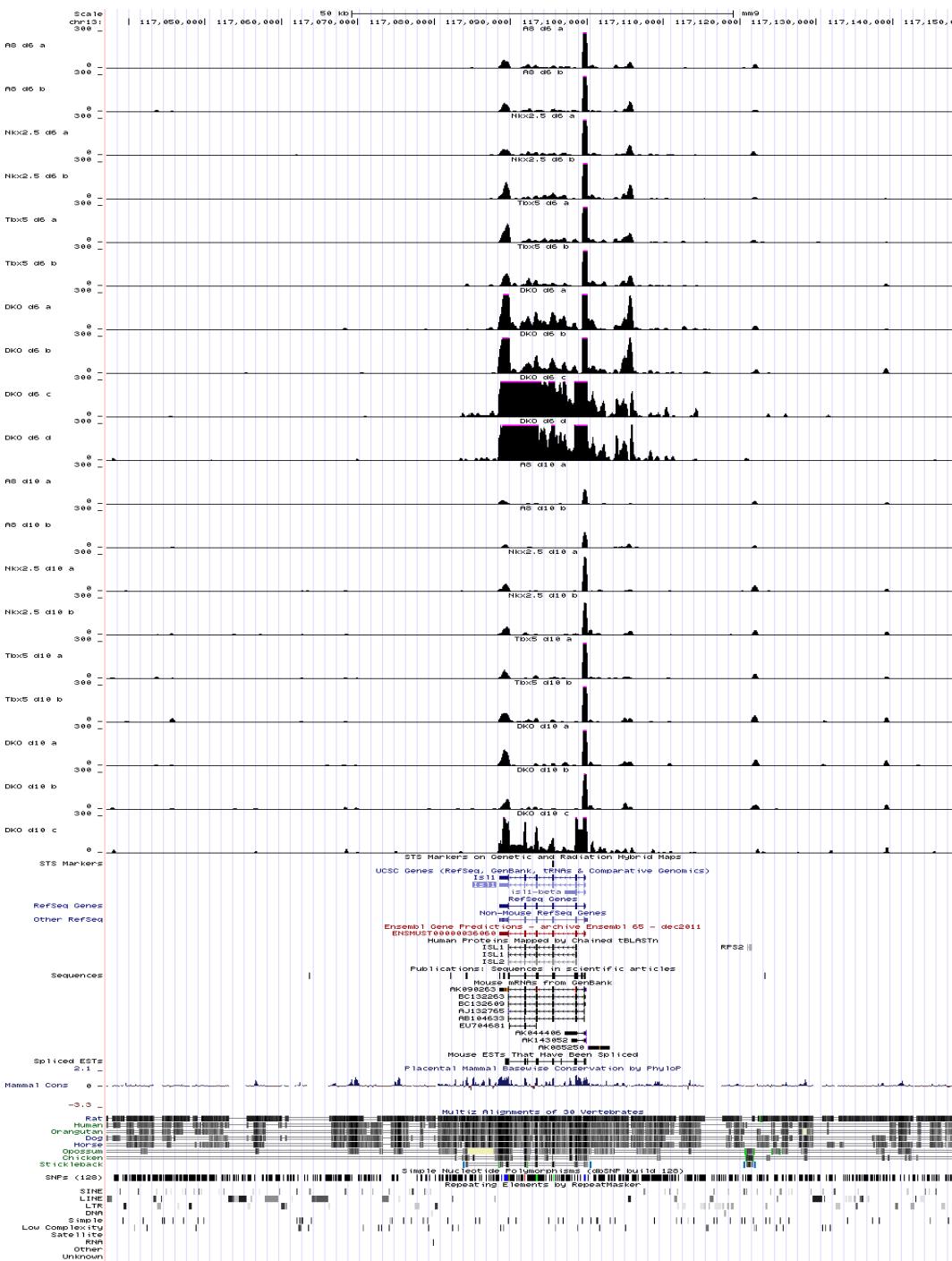
PCR over-amplification

non-specific antibody

other things?

ChIP-seq QC: did the ChIP work?

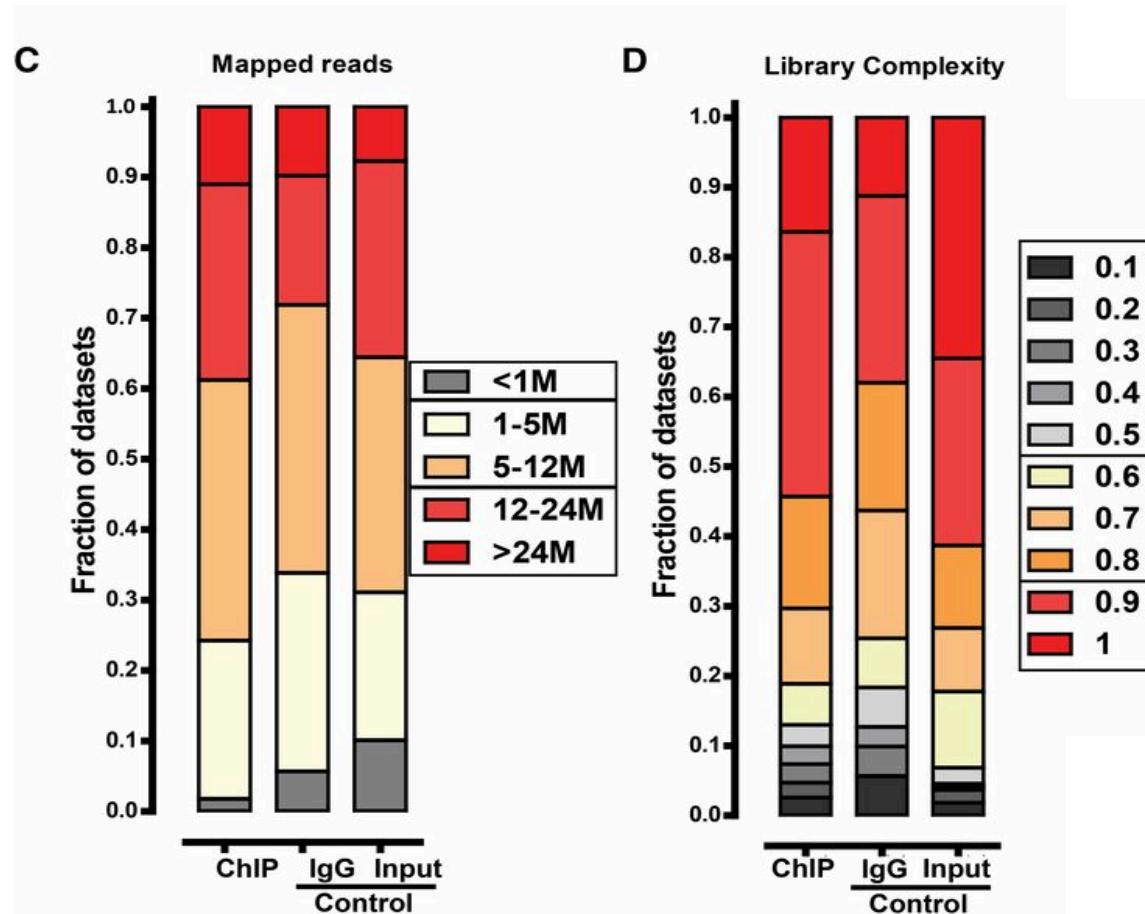
- 1. Inspect the signal (mapped reads, coverage profiles) in genome browser
- 2. Compute peak-independent quality metrics (cross correlation, cumulative enrichment)
- 3. Assess replicate consistency (correlations between replicates of the same condition)



tag density distribution
reproducibility
similarity of coverage
signal at known sites

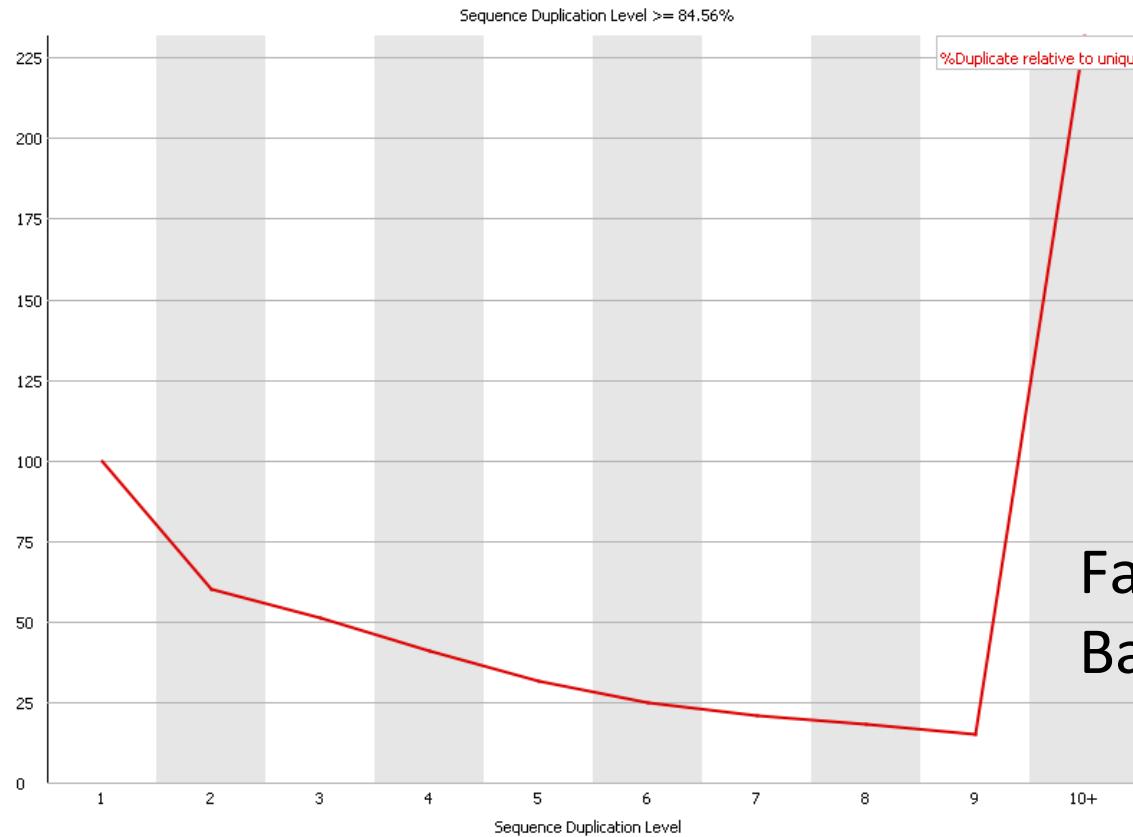
...
Spotting inconsistencies
Confounding factors
Under-sequenced libraries

How do I know my data is of good quality?



Quality control: tag uniqueness – library complexity metric

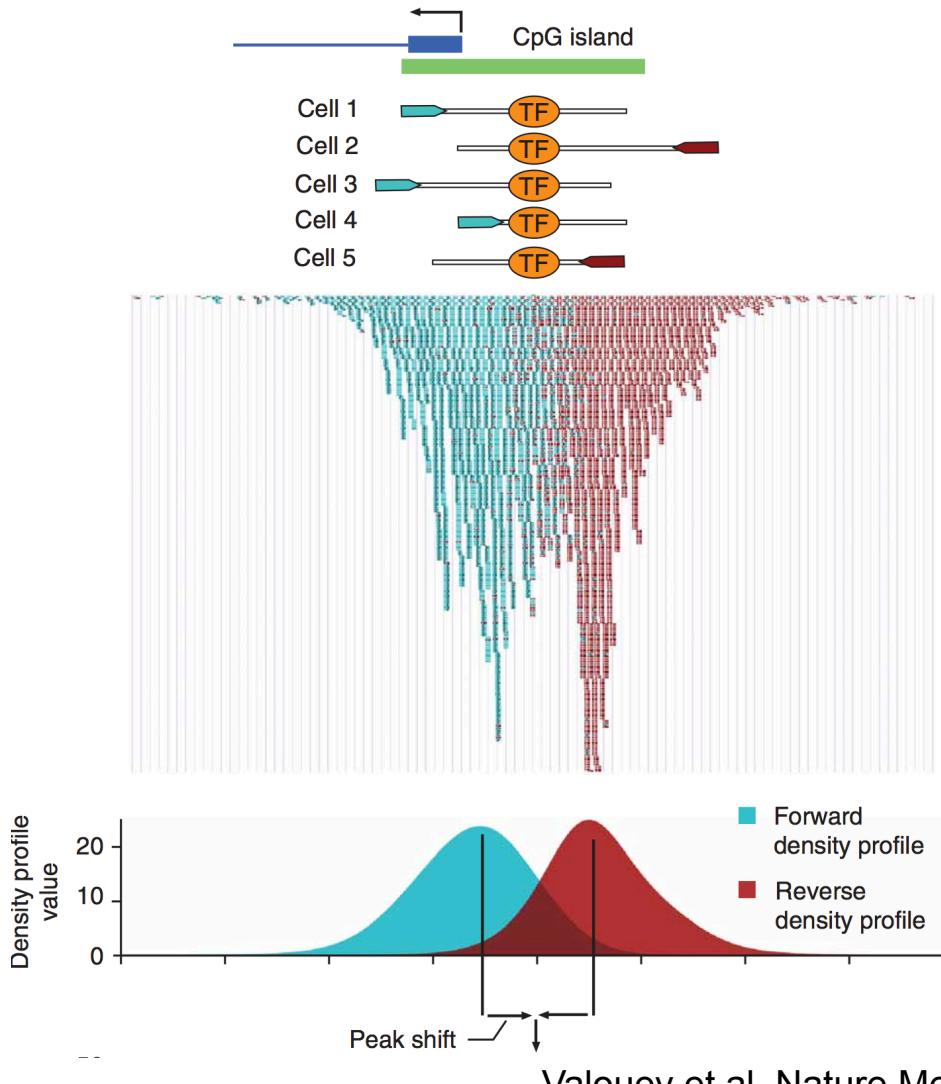
Sequence duplication level > 80% (low complexity library)



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

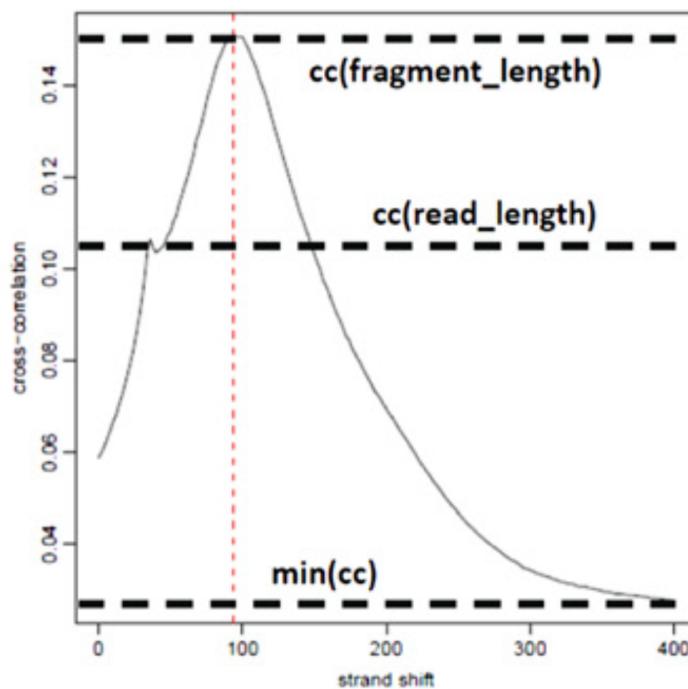
less than 20% of reads should be duplicates for 10 million reads sequenced (ENCODE)

Strand cross-correlation



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.

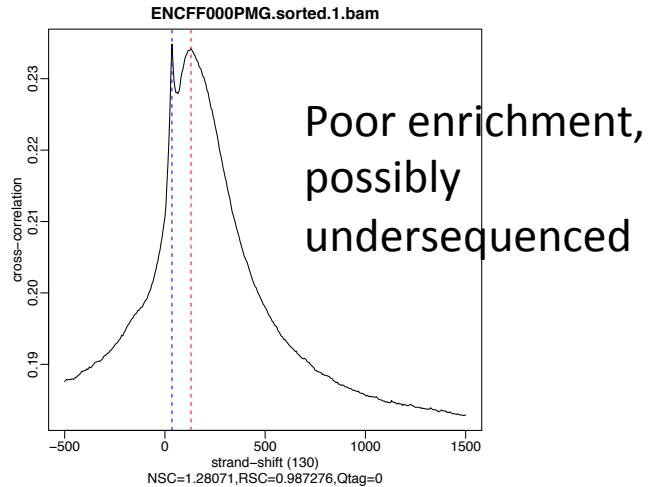
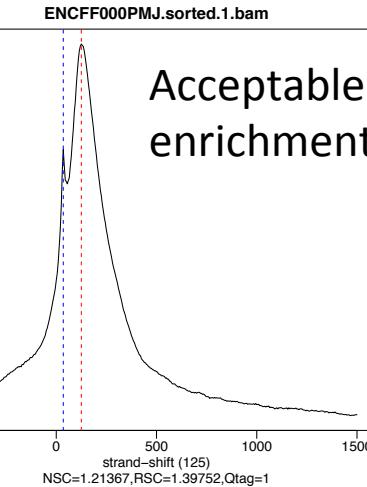
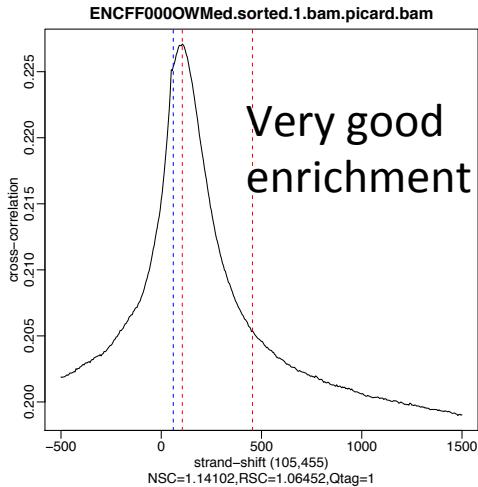


$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

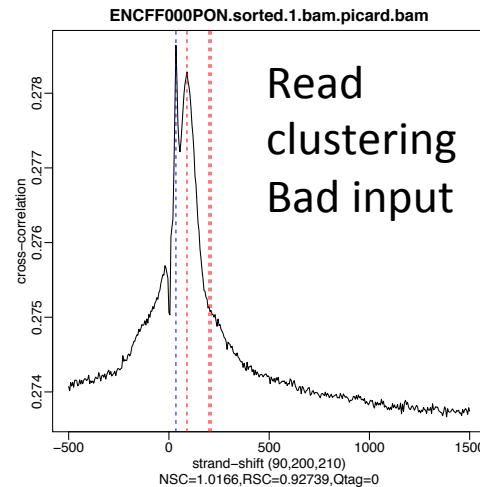
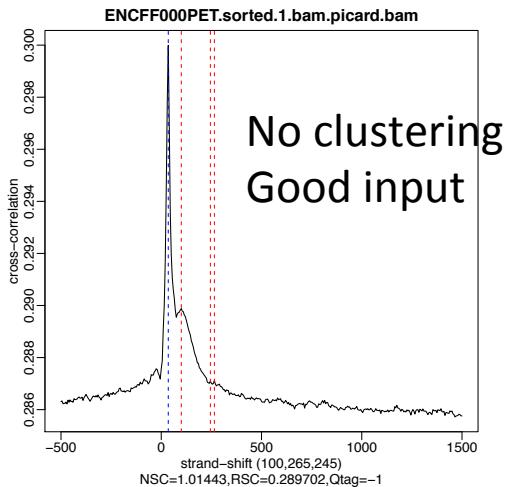
$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Cross-correlation plots

ChIP



Input



How do I know my data is of good quality?

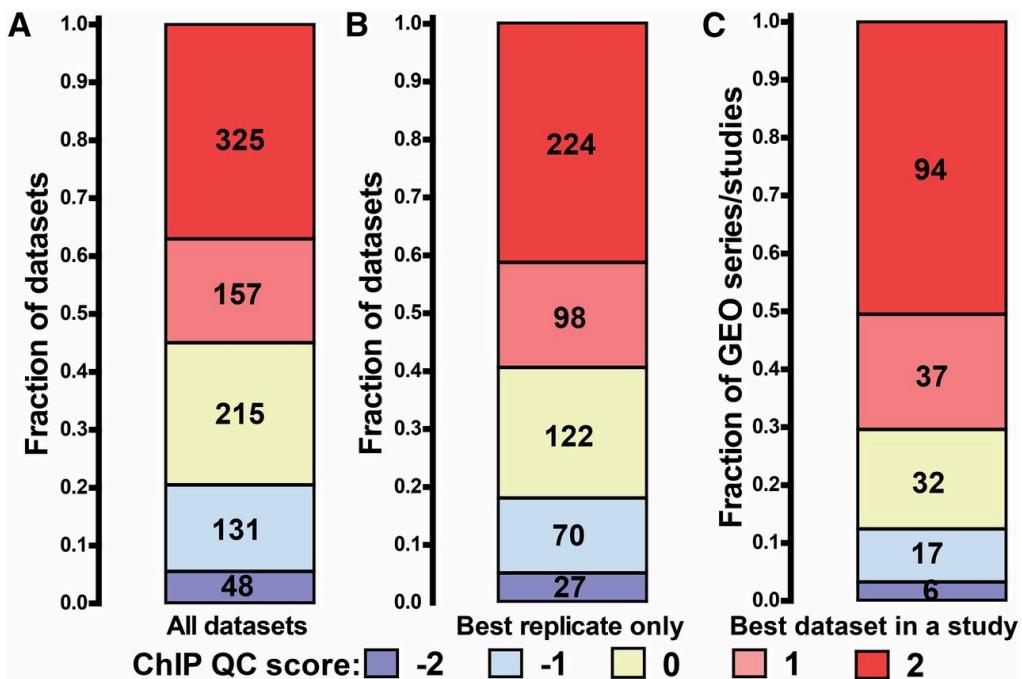
Objective (i.e. peak independent) metrics to quantify enrichment in ChIP-seq;

for TF in mammalian systems:

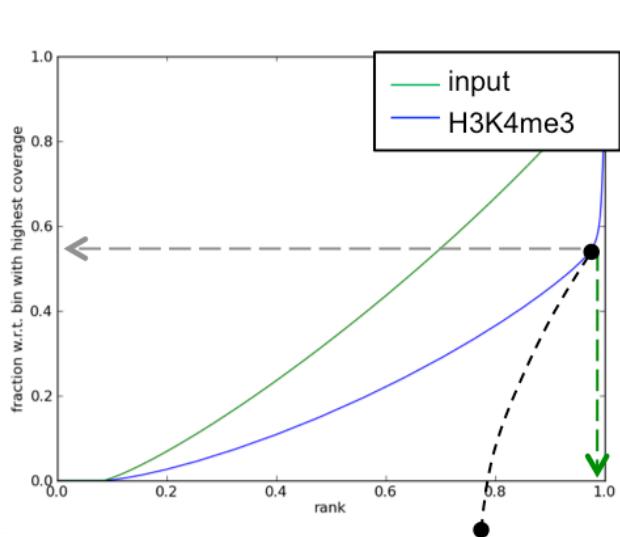
Normalised Strand Correlation NSC
Relative Strand Correlation RSC

Large-scale quality analysis of published ChIP-seq data sets:
20% low quality
25% intermediate quality
55% good quality

30% inputs have metrics similar to IPs

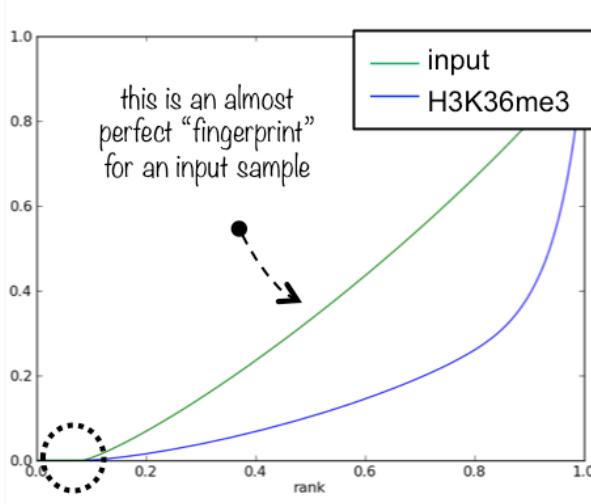


Cumulative enrichment aka “Fingerprint” is another metric for successful enrichment

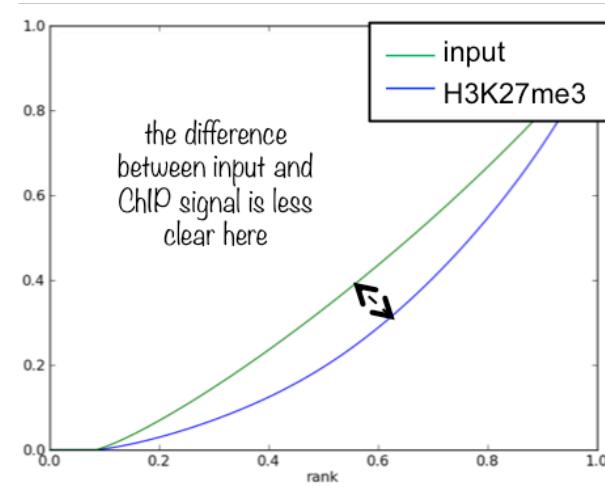


when counting the reads contained in 97% of all genomic bins, only ca. 55% of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

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

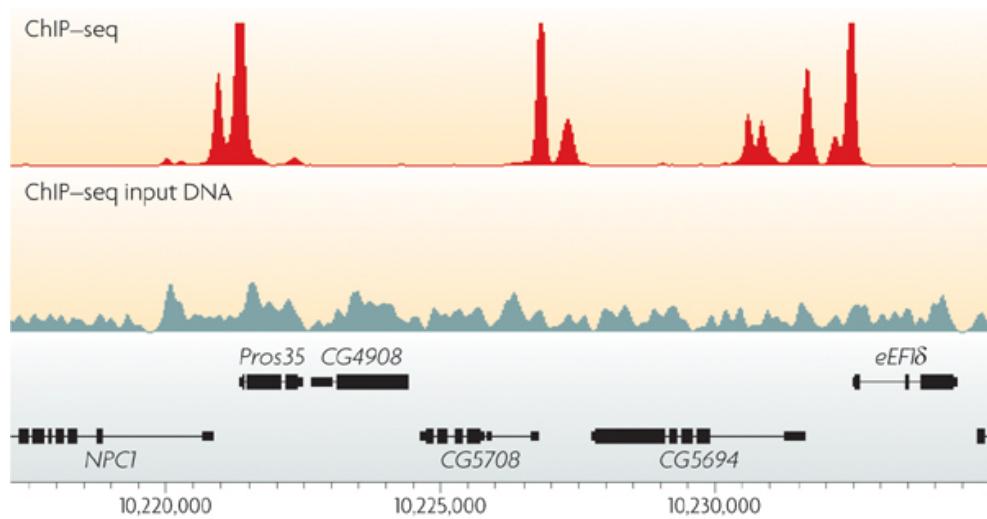
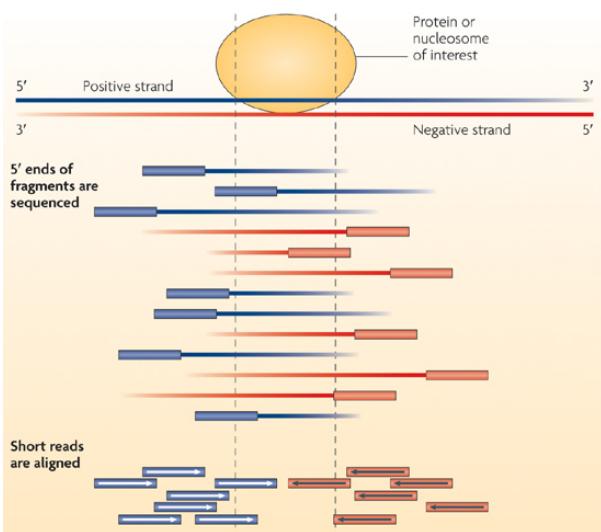


this is an almost perfect “fingerprint” for an input sample
pay attention to where the curves start to rise – this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads – for this example, ca. 10% of the entire genome do not have any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



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

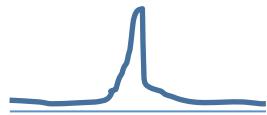
it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed



Peak calling

appropriate methodologies depend on data type

Transcription
Factors



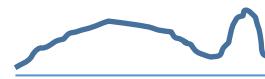
punctate
SPP
MACS2

Chromatin
Remodellers
Histone marks



mixed signal

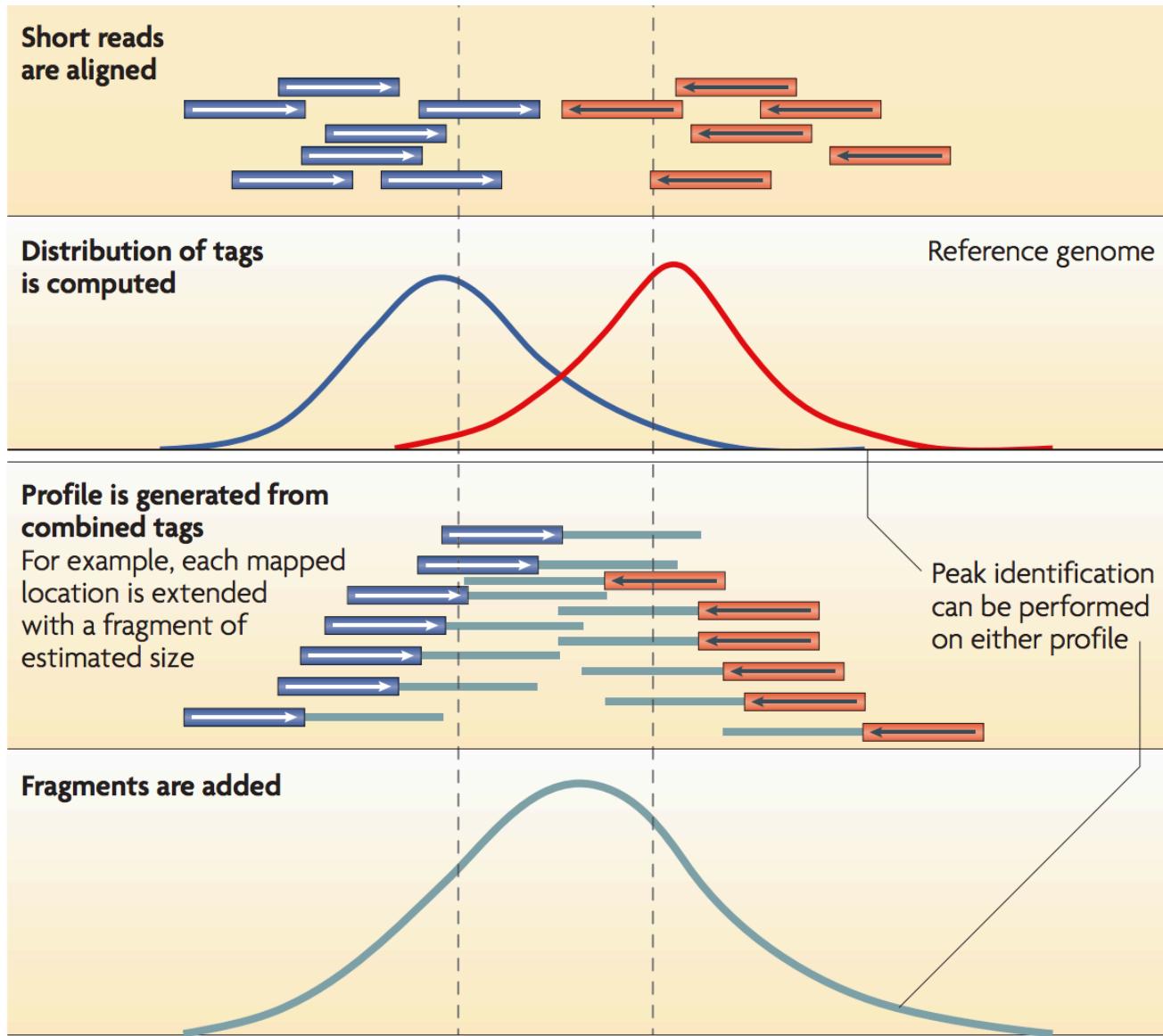
Chromatin
Remodellers
Histone marks
RNA polymerase II



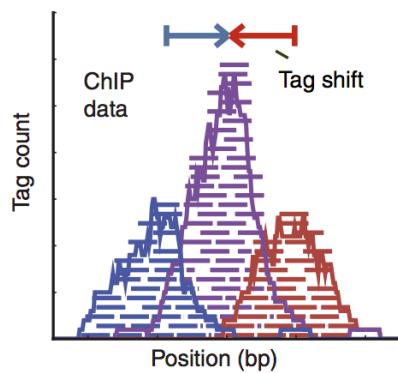
broad signal

This is an active area of algorithm development

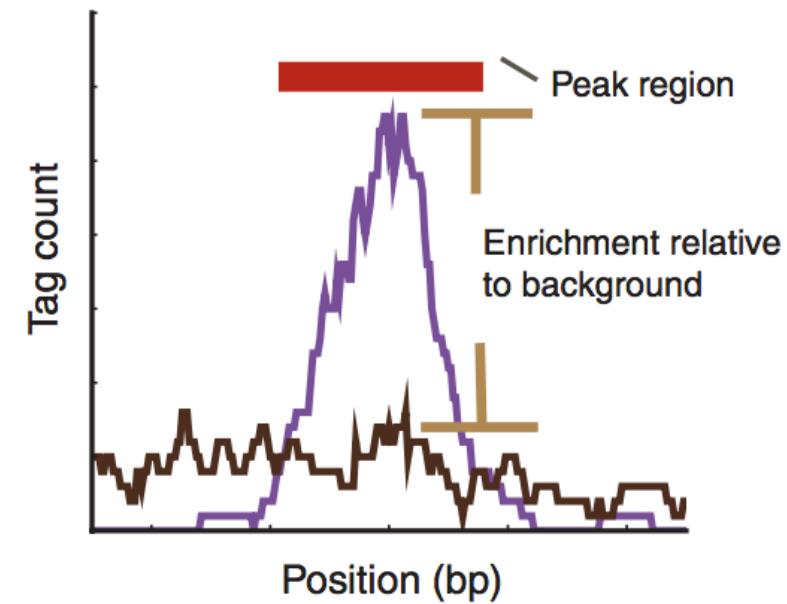
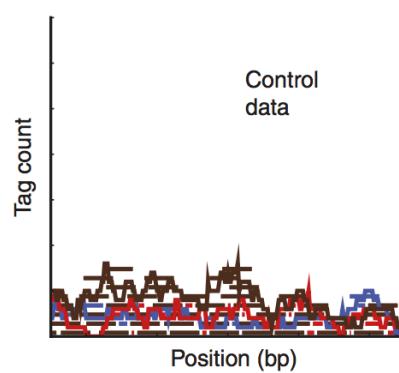
Principle of peak detection



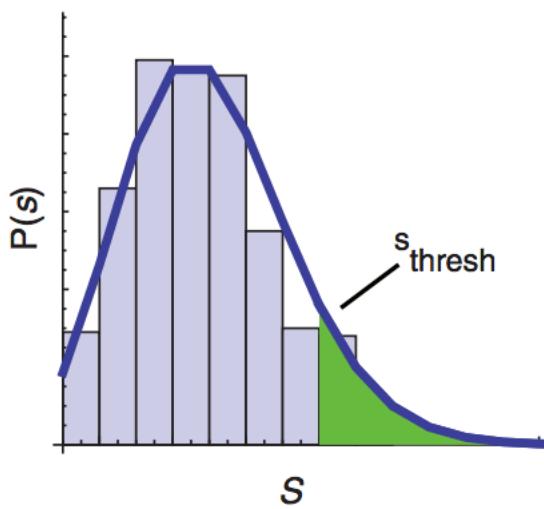
Generate signal profile
along each chromosome



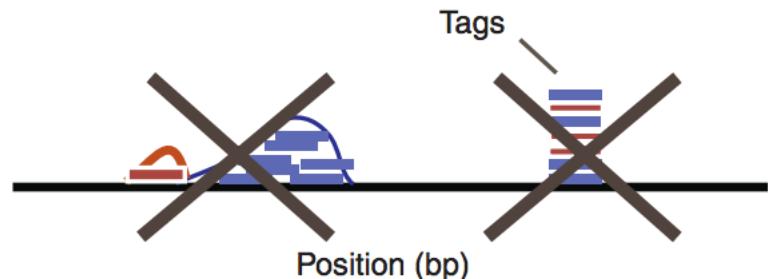
Define background
(model or data)



Assess significance



Filter artifacts



Comparison of peak calling algorithms

Program														
	Reference Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test	
CisGenome	28	1.1	X*	X			X	X		X		X		conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1		X			X			X				
E-RANGE	27	3.1		X			X			X	X			chromosome scale Poisson dist.
MACS	13	1.3.5		X			X			X		X		local Poisson dist.
QuEST	14	2.3			X		X			X**		X		chromosome scale Poisson dist.
HPeak	29	1.1		X			X					X		Hidden Markov Model
Sole-Search	23	1	X	X			X		X			X		One sample t-test
PeakSeq	21	1.01		X			X					X		conditional binomial model
SISSRS	32	1.4		X			X			X				
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
Generating density profiles				Peak assignment		Adjustments w. control data		Significance relative to control data						

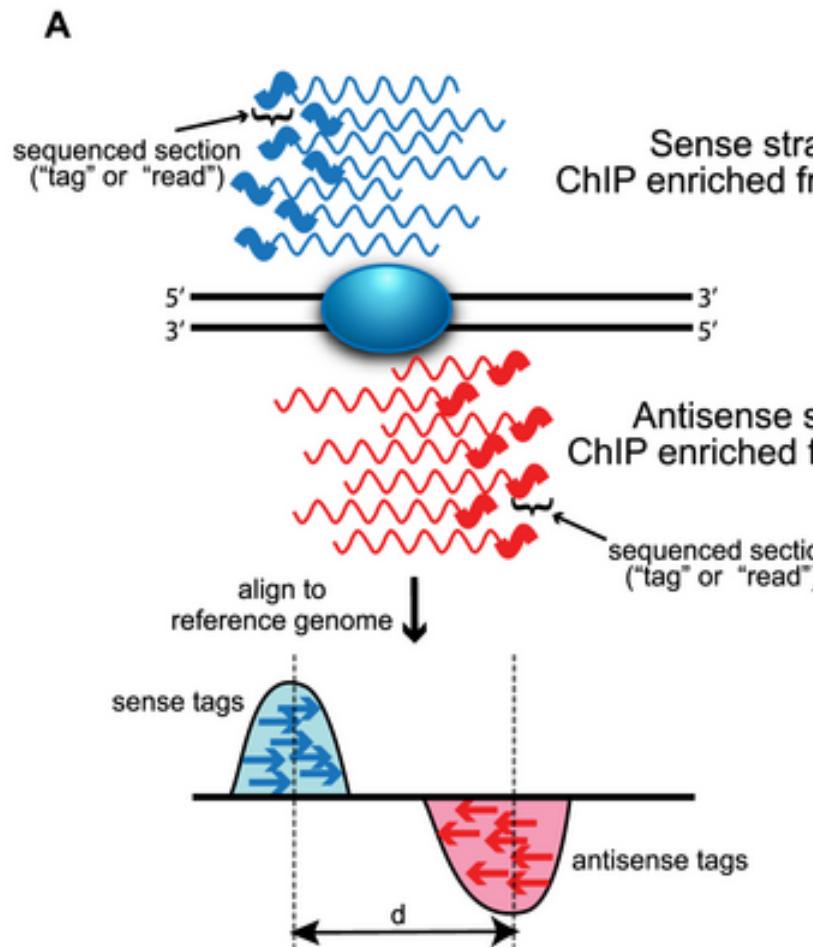
X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

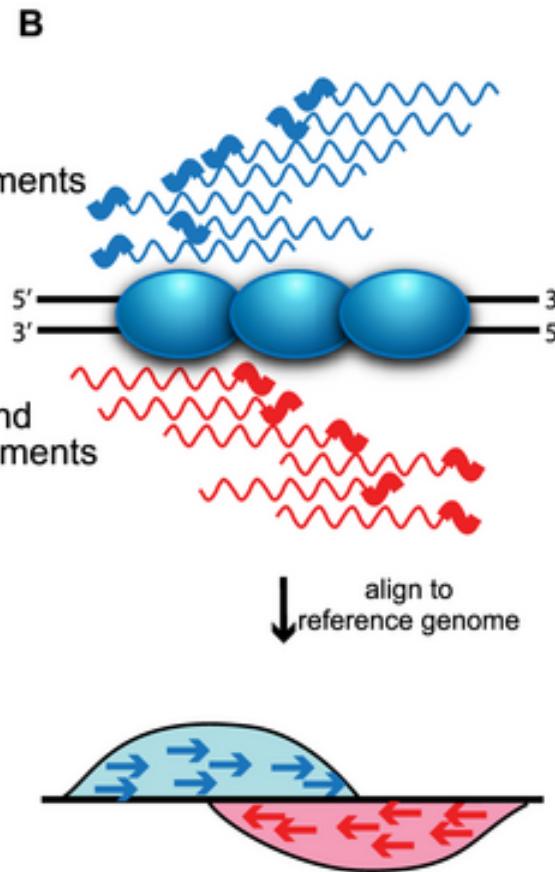
X' = method excludes putative duplicated regions, no treatment of deletions

Point-source vs. broad peak detection

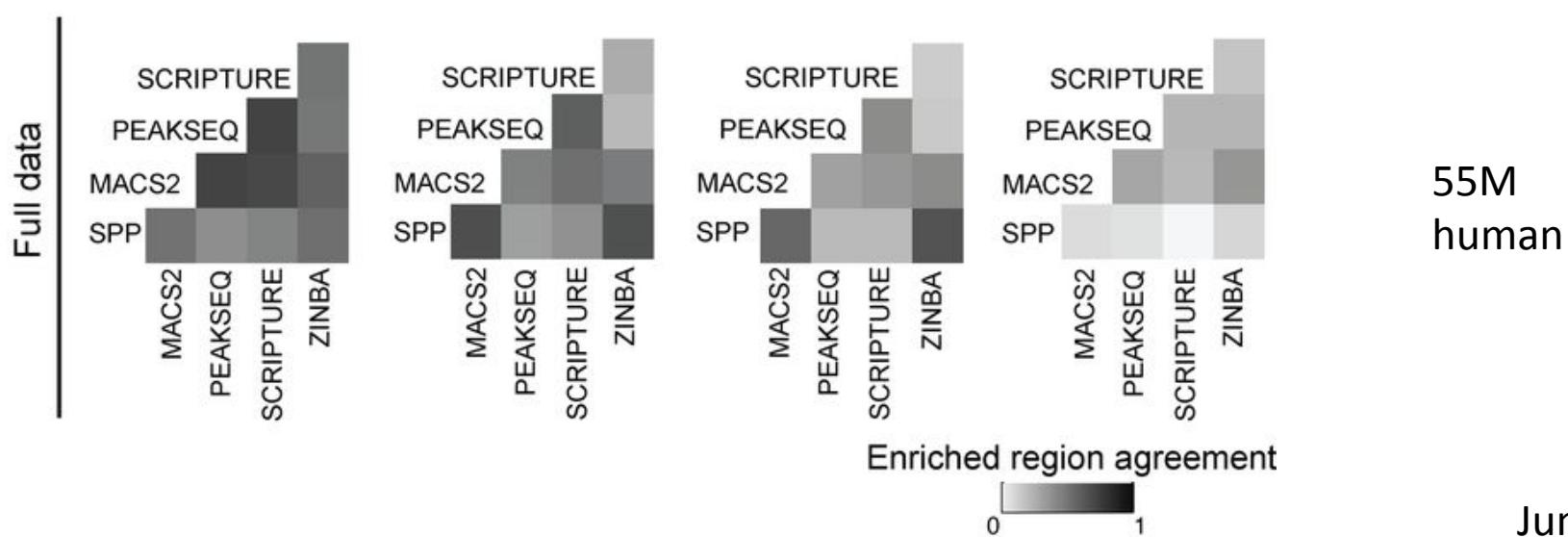
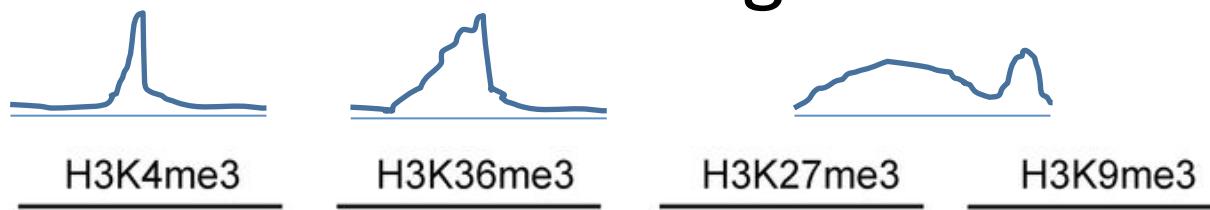
Sequence-specific binding (TFs)



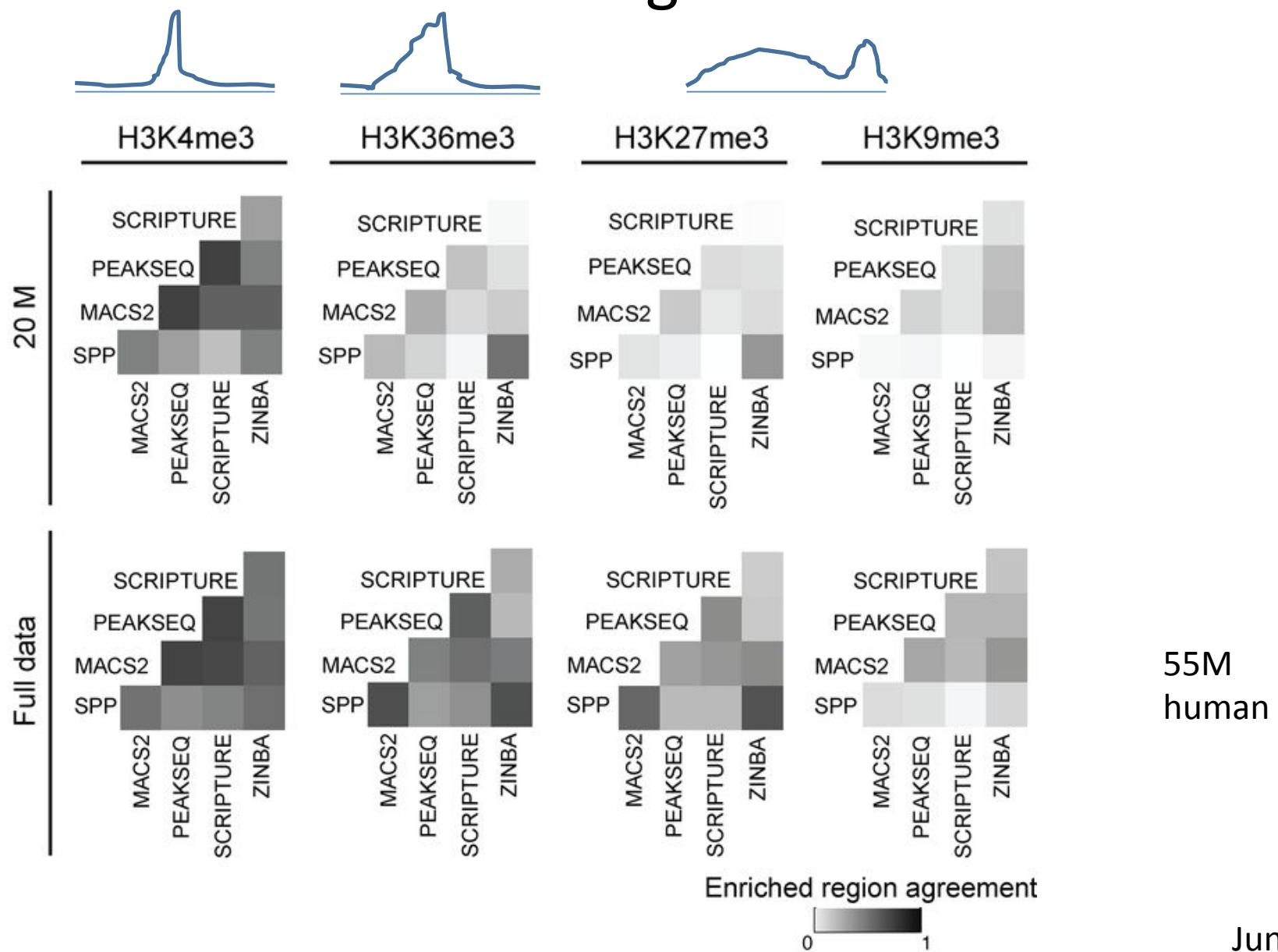
Distributed binding (histones, RNAPol2)



Comparison of enriched regions detected by various algorithms



Comparison of enriched regions detected by various algorithms

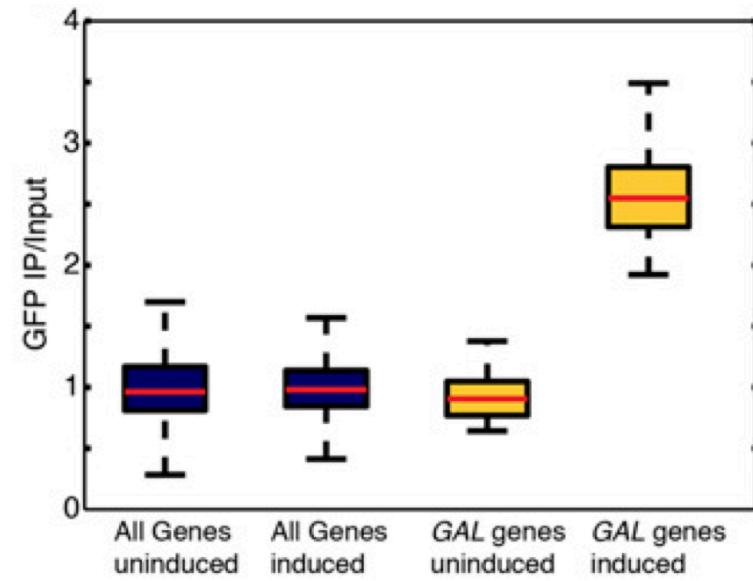
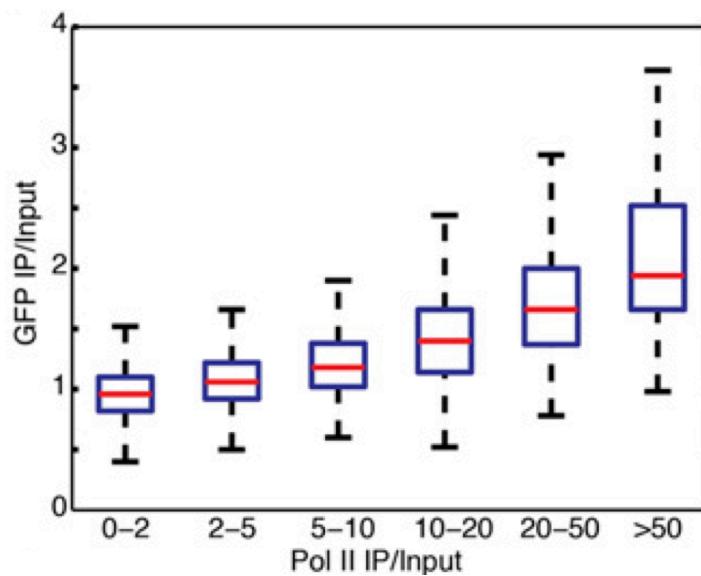


Blacklist regions

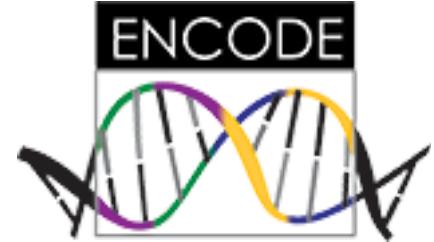
- Some regions consistently give a strong ChIP signal, regardless of antibody.
- Many are repeats.
- These are called “blacklist” regions, and are usually removed from the analysis.
- Lists are available for common genomes: human, mouse, fly and worm

Other – “hyper-chippable regions”

- Regions with highly expressed genes gave very strong signal
 - Even for GFP or other non-biologically relevant proteins.
 - This artifact depended on gene expression, i.e. it's dynamic and depends on the state of the cells.



Quality considerations



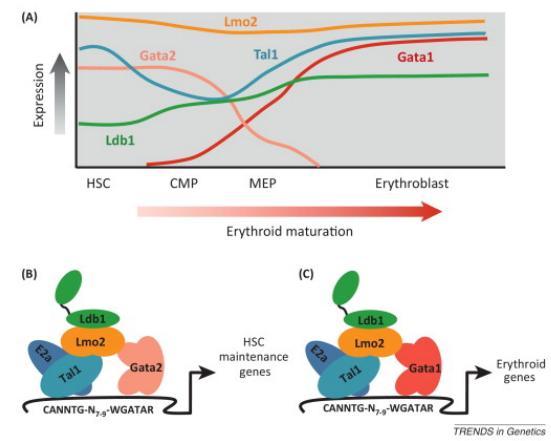
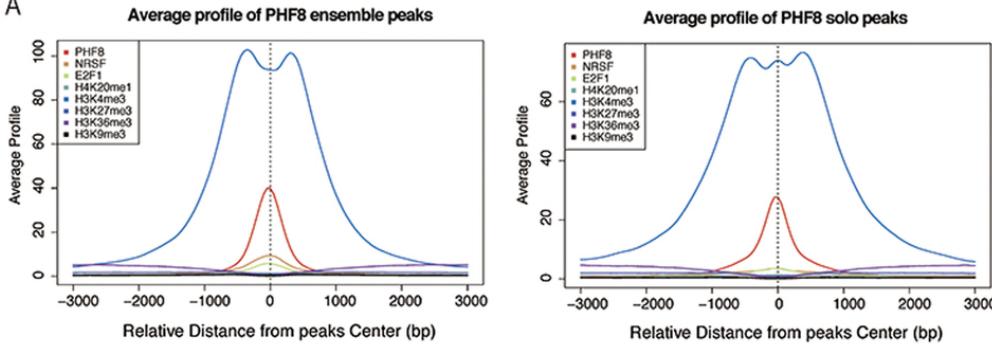
- ChIP-seq quality guidelines from the ENCODE project (Relative strand cross-correlation, Irreproducible discovery rate)
- Antibody validation
- Appropriate sequencing depth (depending on genome size and peak type). For human genome and broad-source peaks, min. 40-50M reads is required.
- Experimental 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 – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources
- Exercise overview

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



TRENDS in Genetics

Peak annotation

Identification of nearest genomic features

- BEDtools,
- BEDops,
- PeakAnnotator,
- CisGenome,
- In R / Bioconductor: ChIPpeakAnno

Motif detection

- Enrichment of known sequence motifs (CEAS, Transfac Match, HOMER, RSAT)
- *De novo* motif detection (MEME, CisFinder, HMS, DREME, ChIPMunk, HOMER, RSAT)

Enrichment of known motifs (Homer):

Homer Known Motif Enrichment Results

Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)

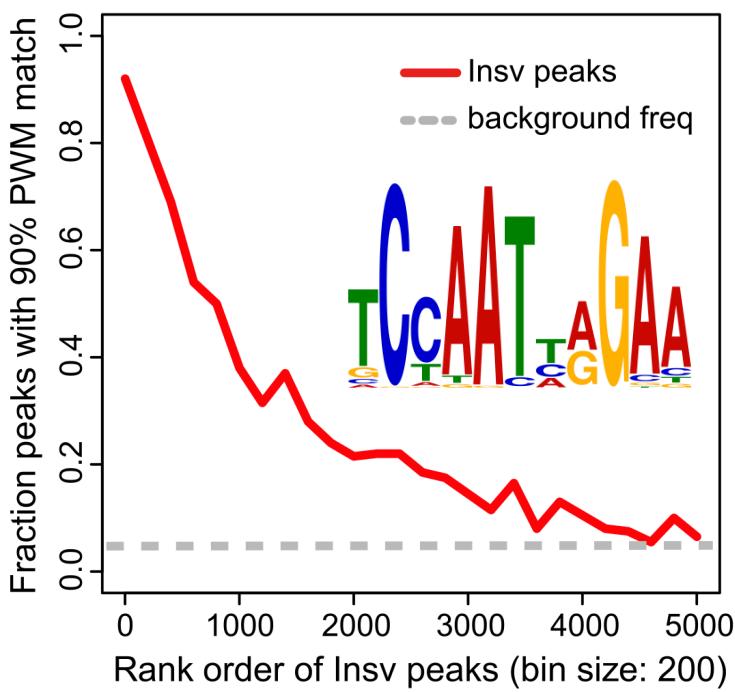
Total Target Sequences = 900, Total Background Sequences = 45419



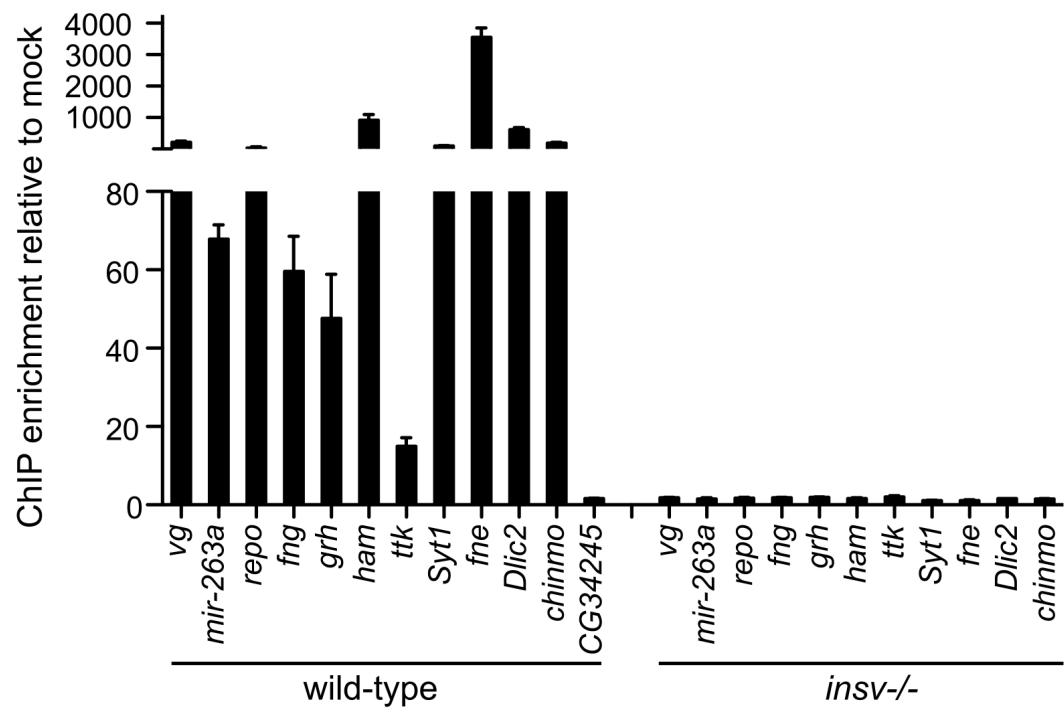
Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Motif File	PDF
1		Lhx3(Homeobox)/Forebrain-p300-ChIP-Seq/Homer	1e-178	-4.114e+02	0.0000	512.0	56.89%	6985.5	15.38%	motif file (matrix)	pdf
2		Sox3(HMG)/NPC-Sox3-ChIP-Seq(GSE33059)/Homer	1e-128	-2.955e+02	0.0000	515.0	57.22%	9264.1	20.40%	motif file (matrix)	pdf

Example - Validating peaks

Look for sequence motif, especially in top peaks

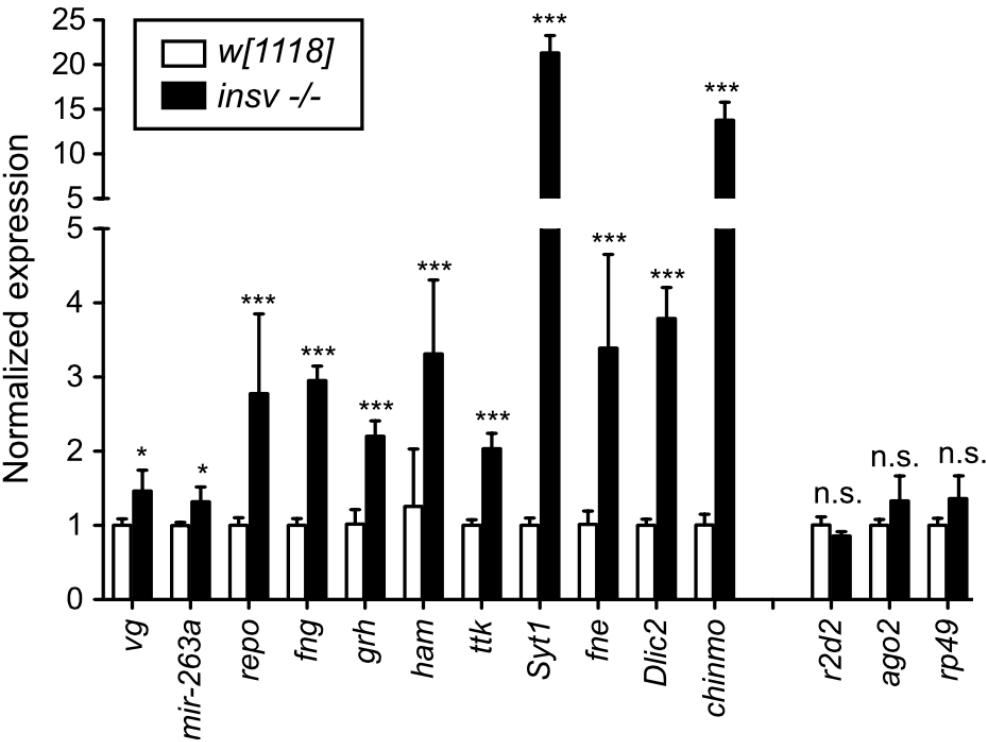


Validation by ChIP-qPCR

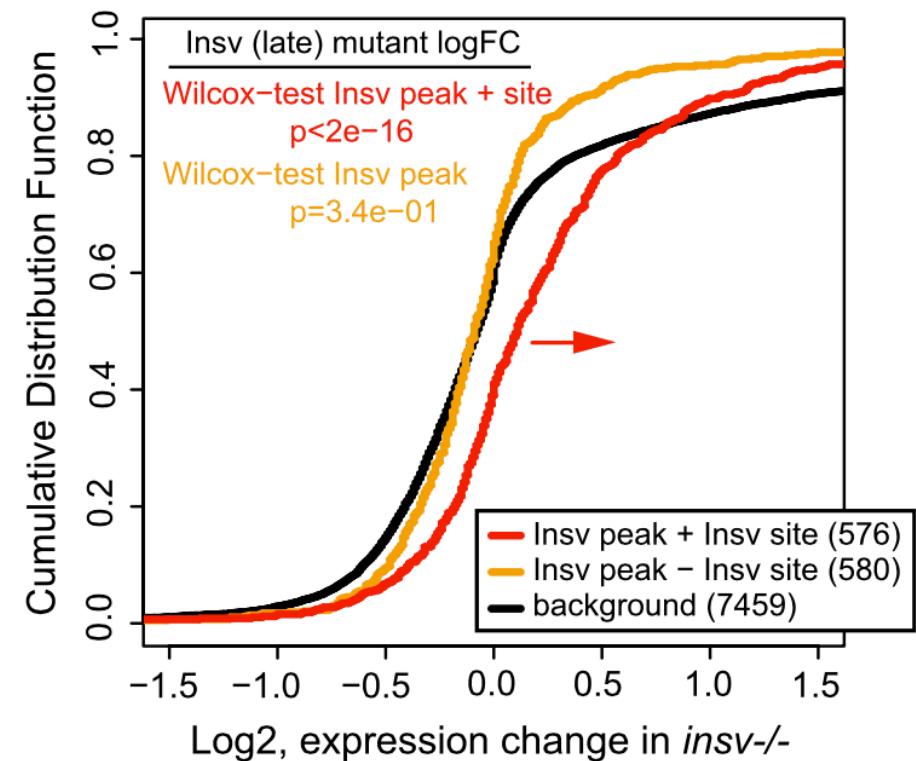


Example – Integration with gene expression data

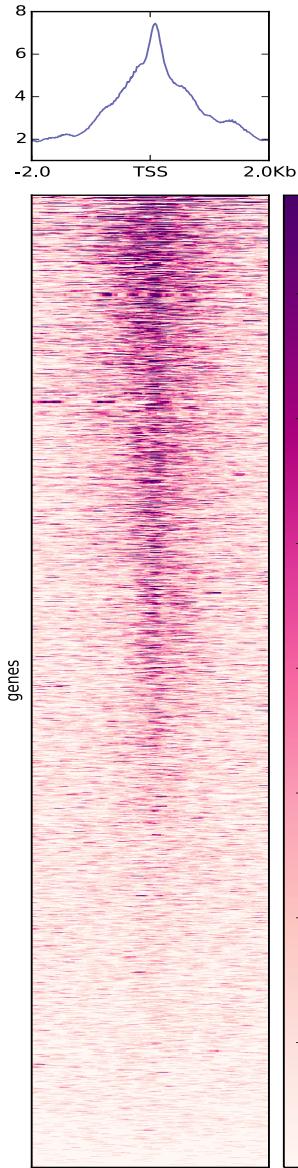
rt-qPCR of nearby genes, in wt mutant



Expression of all genes, in mutant and wt



Signal visualisation and interpretation



deepTools

ngsplots

seqMiner

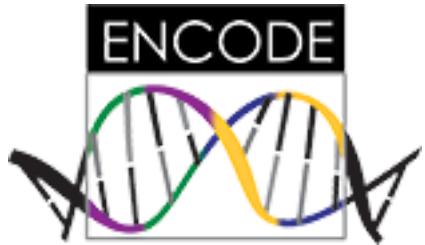
- Clustering
- Heatmaps
- Profiles
- Comparison of different datasets

Differential occupancy

- Use algorithms developed for differential expression and summarise reads mapped in peaks; normalisation; statistical testing; R environment
 - edgeR / csaw
 - DiffBind (implements several normalisation methods)
- Calculate enrichment in sliding windows
 - DROMPA
 - Diffreps

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

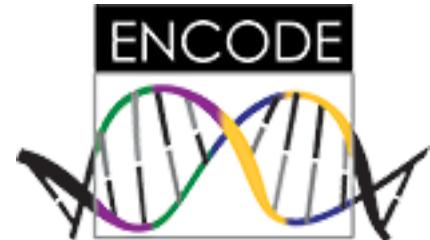
Where to obtain data?



The *ENCODE* (Encyclopedia of DNA Elements) Consortium and the *Roadmap Epigenomics* Consortium are a vast resource of various kinds of functional genomics data (as well as RNA-seq data).

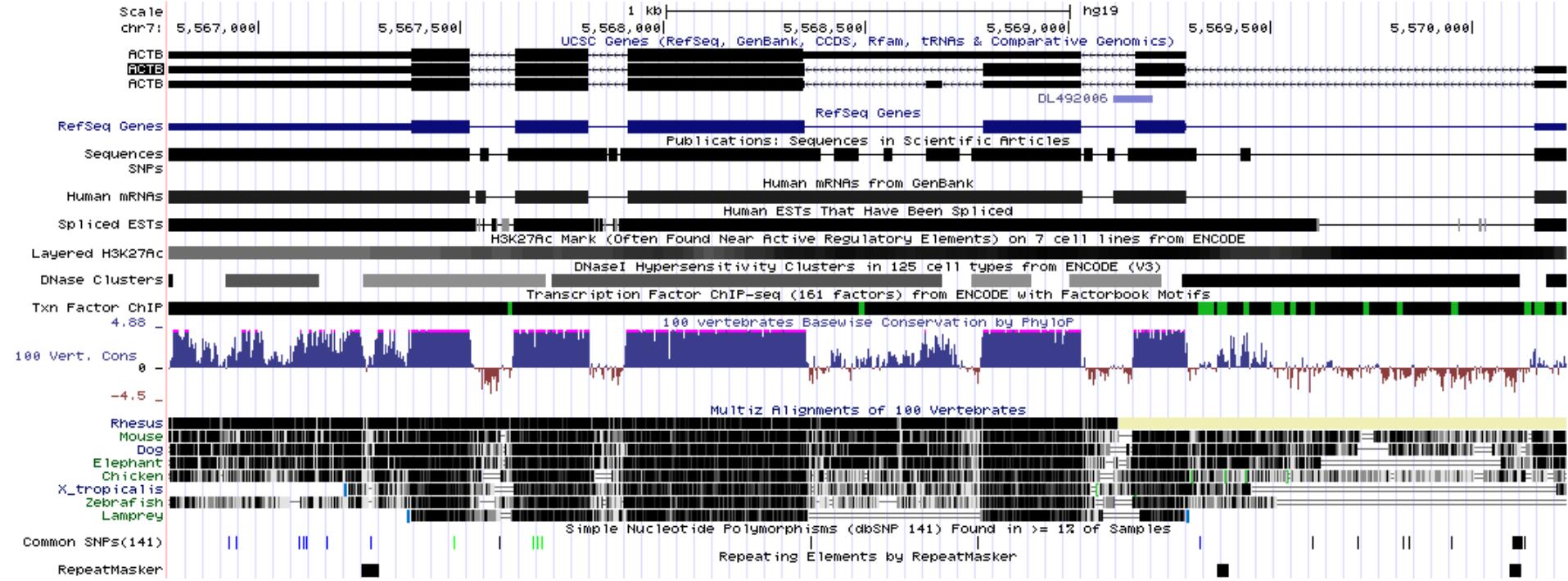
The ENCODE project

www.encodeproject.org



- Encyclopedia of DNA elements
- Identification of regulatory DNA elements in human (and mouse) genome
- 240 human and 55 mouse DNA binding proteins
- 1464 human and 432 mouse samples
- RNA profiling, protein-DNA interaction, chromatin condensation, DNA methylation, ...
- 2009 - ongoing

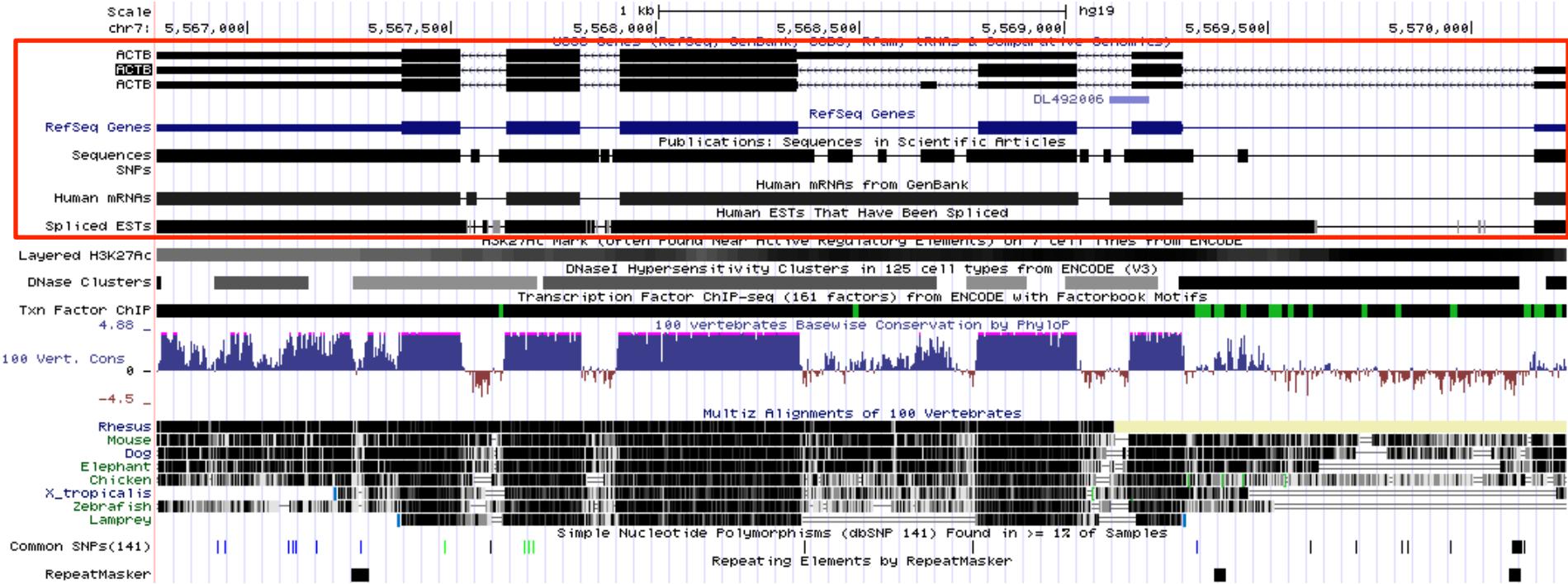
Human ACTB locus as seen in the UCSC Genome Browser



Gene model
Alternative transcripts
Histone modifications
Chromatin structure

Transcription factor binding sites
DNA conservation
Single nucleotide polymorphisms (SNP)
Repeats

Human ACTB locus as seen in the UCSC Genome Browser



Gene model

Alternative transcripts

Histone modifications

Chromatin structure

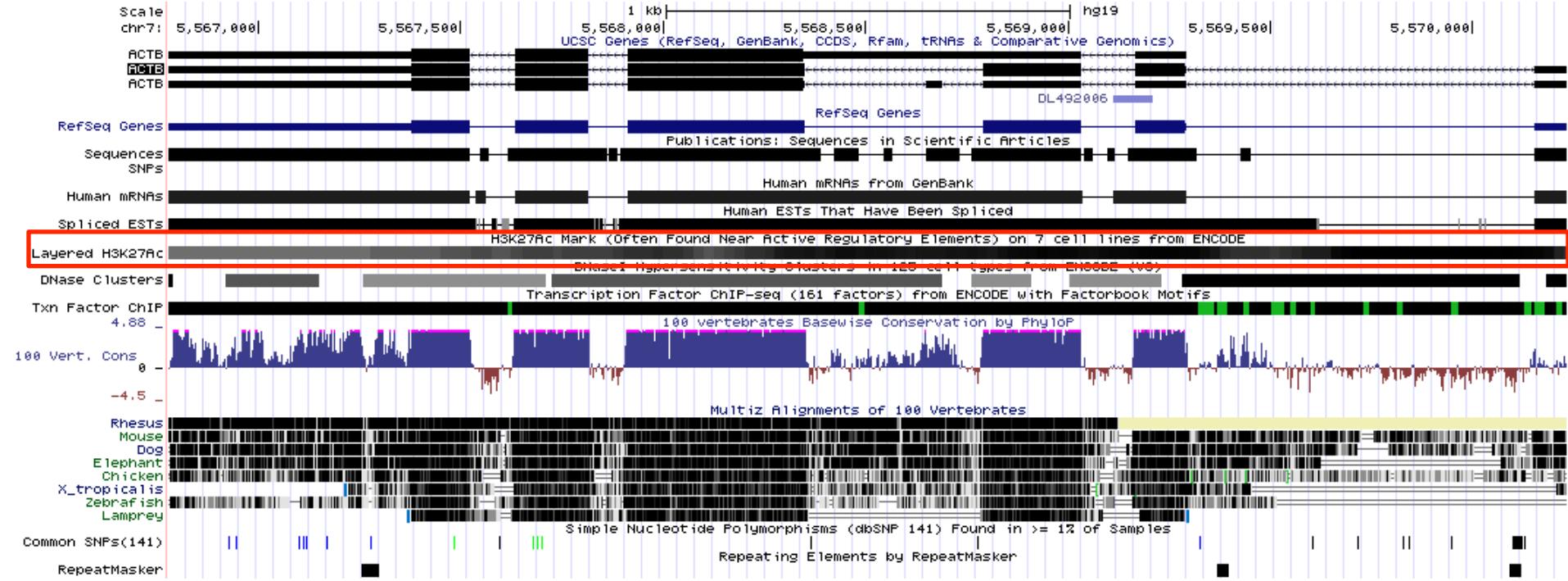
Transcription factor binding sites

DNA conservation

Single nucleotide polymorphisms (SNP)

Repeats

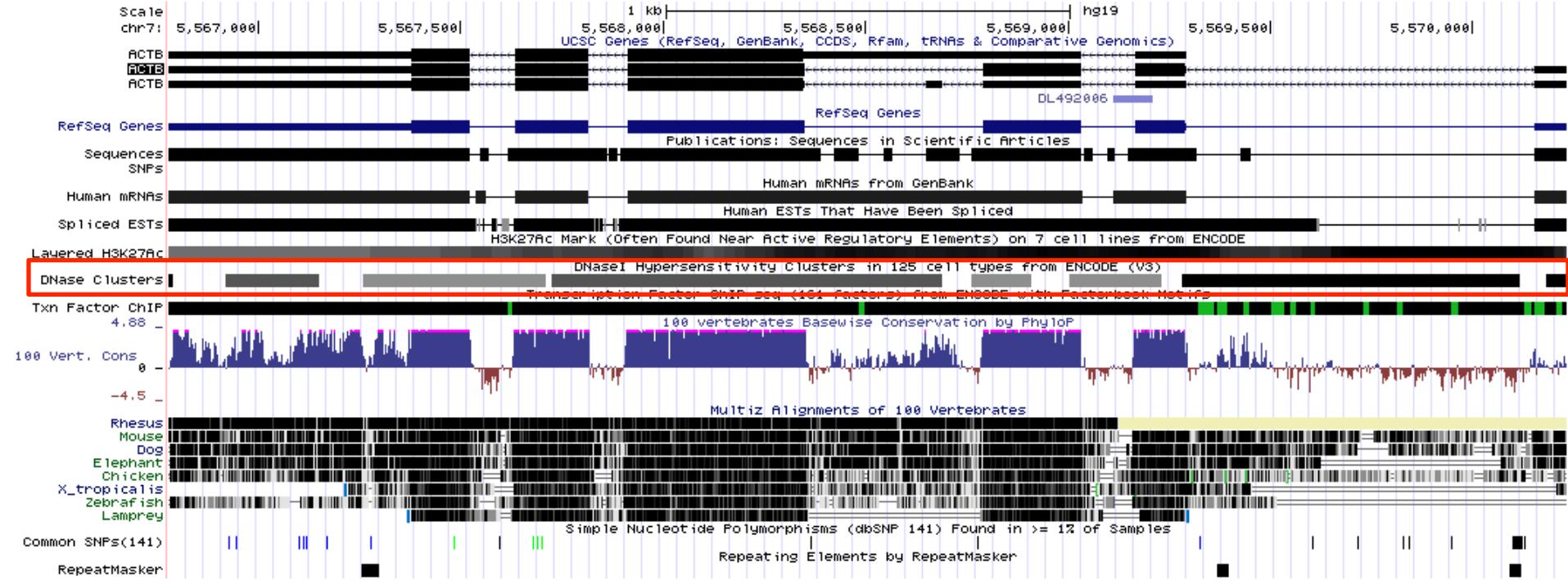
Human ACTB locus as seen in the UCSC Genome Browser



Gene model
Alternative transcripts
Histone modifications
Chromatin structure

Transcription factor binding sites
DNA conservation
Single nucleotide polymorphisms (SNP)
Repeats

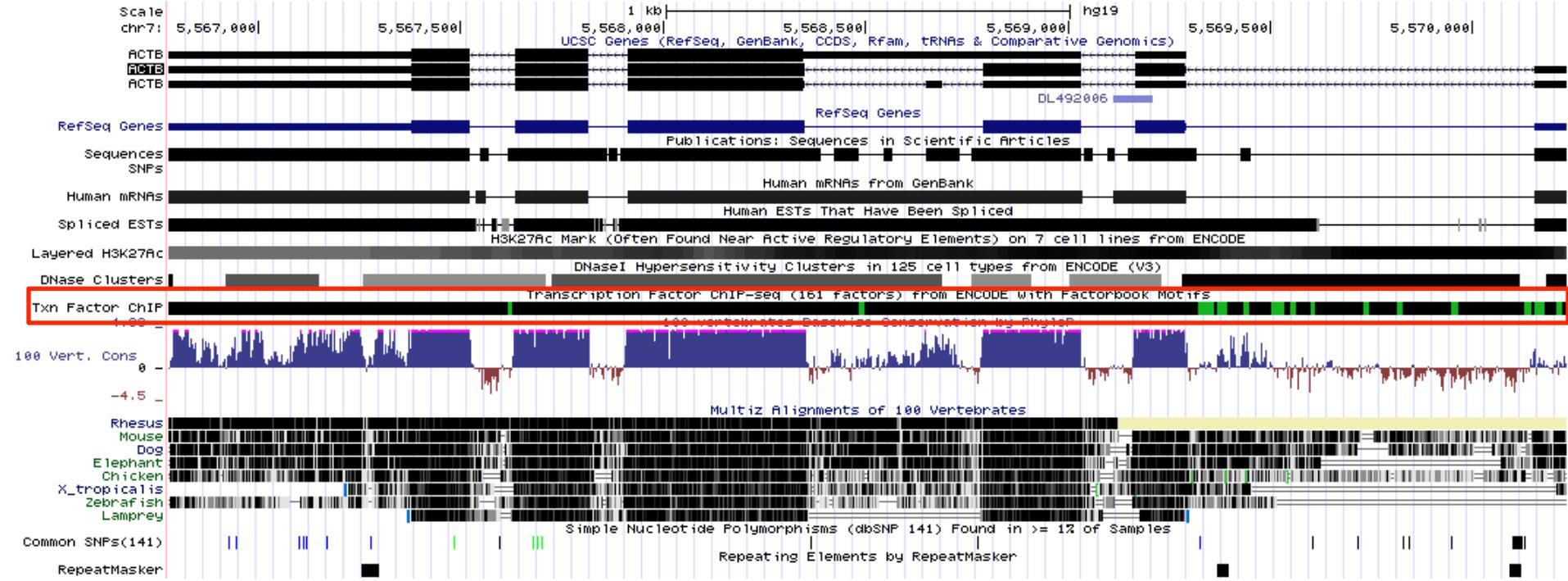
Human ACTB locus as seen in the UCSC Genome Browser



Gene model
Alternative transcripts
Histone modifications
Chromatin structure

Transcription factor binding sites
DNA conservation
Single nucleotide polymorphisms (SNP)
Repeats

Human ACTB locus as seen in the UCSC Genome Browser



Gene model
Alternative transcripts
Histone modifications
Chromatin structure

Transcription factor binding sites
DNA conservation
Single nucleotide polymorphisms (SNP)
Repeats

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

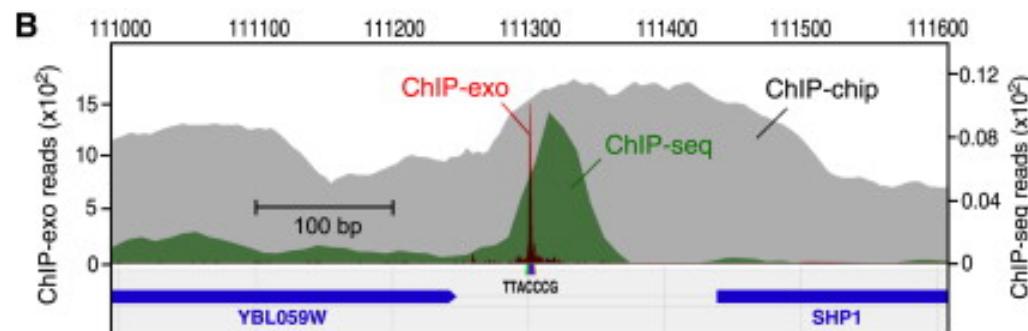
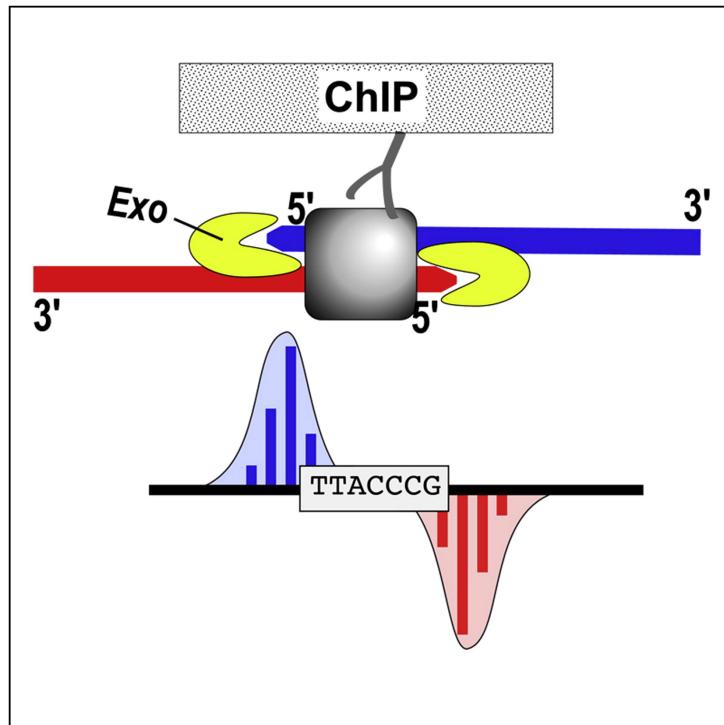
Further reading

- 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>

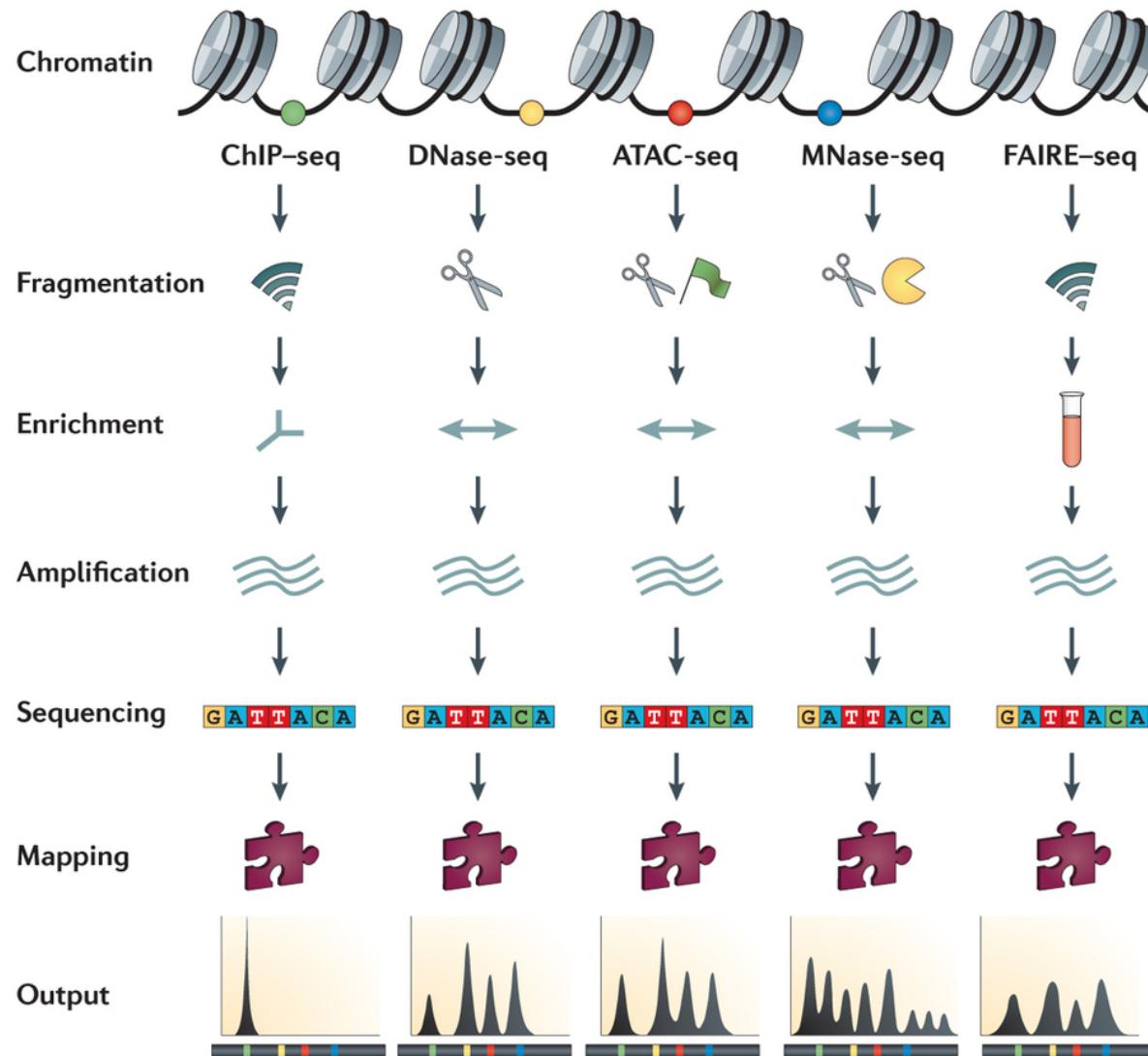
Bioconductor ChIP-seq resources

- 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)
 - Csaaw (a pipeline for ChIP-seq analysis, including statistical analysis of differential occupancy)
- 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)
- Quality control
 - ChIPQC
- Differential occupancy
 - edgeR
 - DESeq, DESeq2
 - DiffBind (compatible with objects used for ChIPQC, wrapper for DESeq and edgeR DE functions)
- Peak Annotation
 - ChIPpeakAnno (annotating peaks with genome context information)

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
- Exercise overview

Exercise

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

Questions?

Jakub.Westholm@scilifelab.se