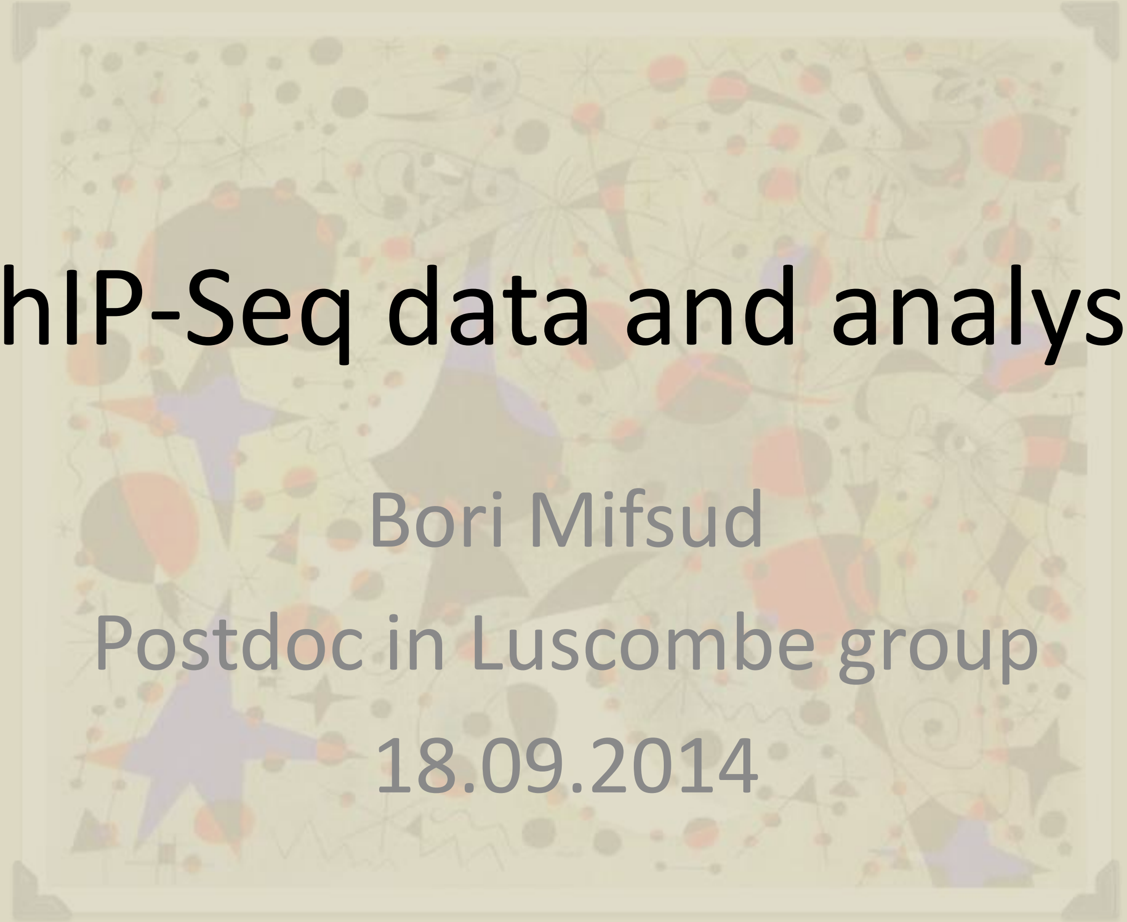


ChIP-Seq data and analysis

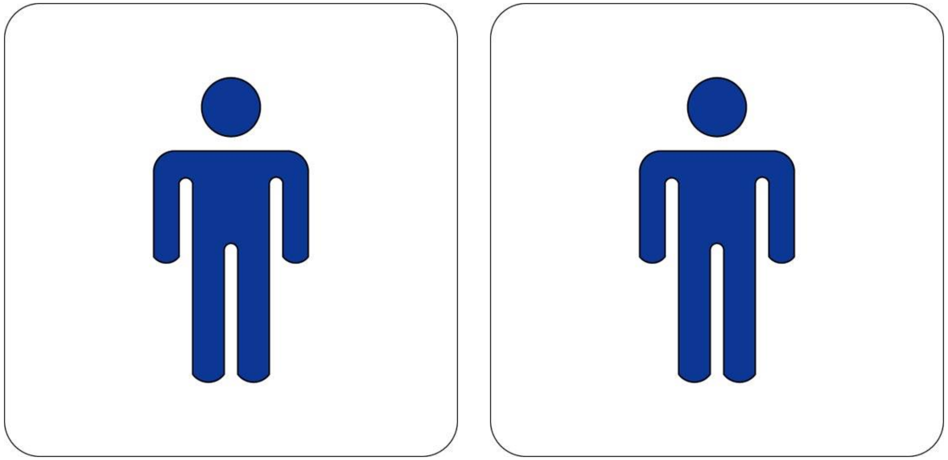
Bori Mifsud

Postdoc in Luscombe group

18.09.2014

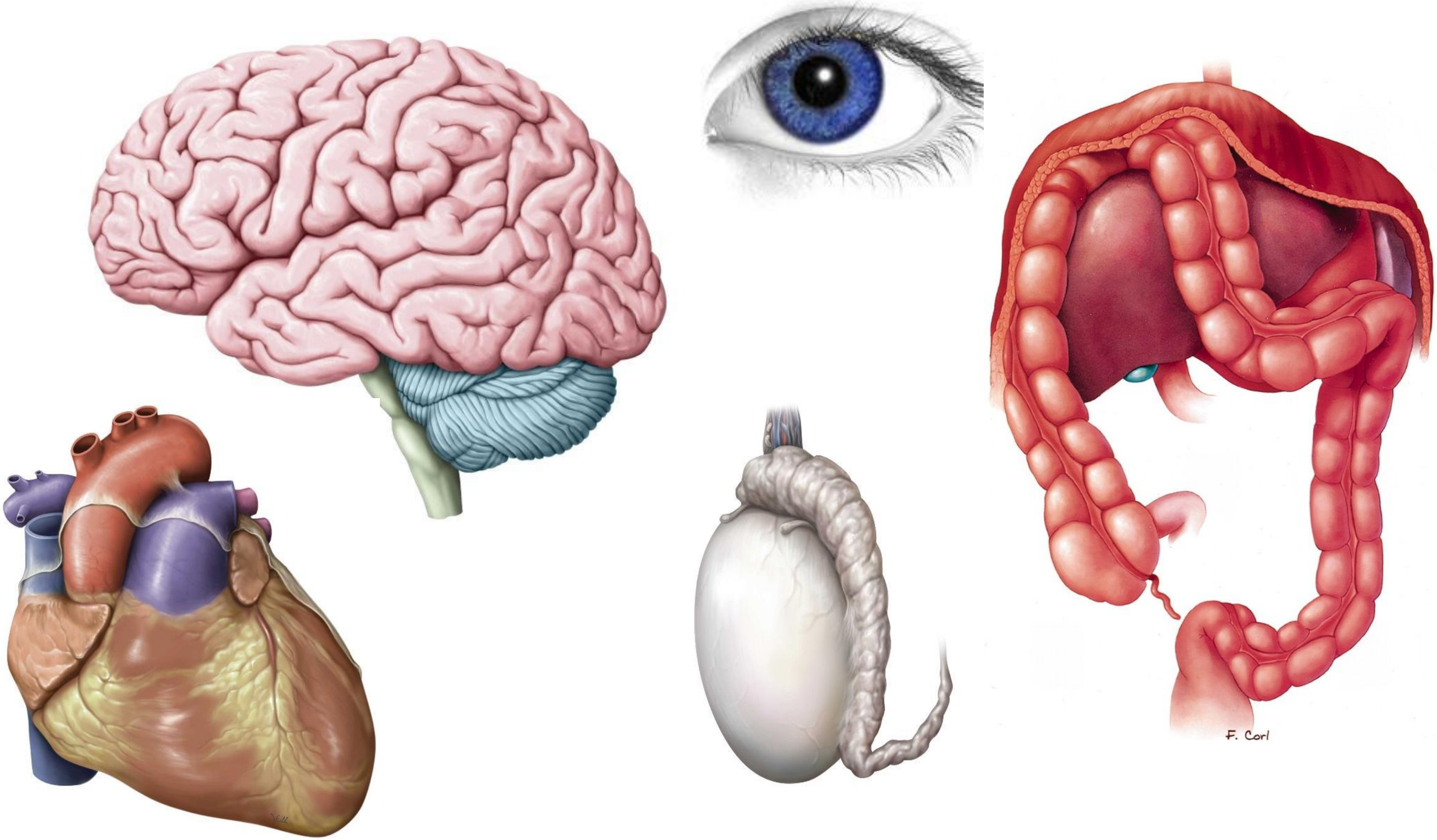
Computational biology UCL-LRI

Why do
Chromatin Immunoprecipitation (ChIP)?

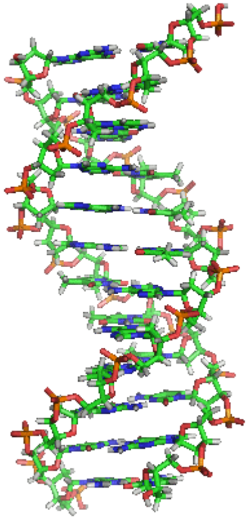


~99.9% identical genetic material

100% identical genetic material

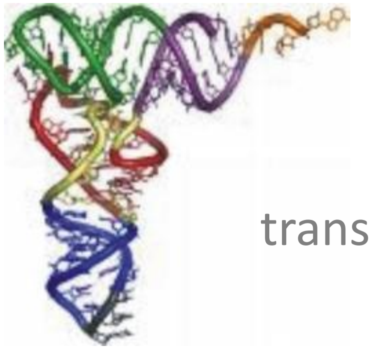


DNA



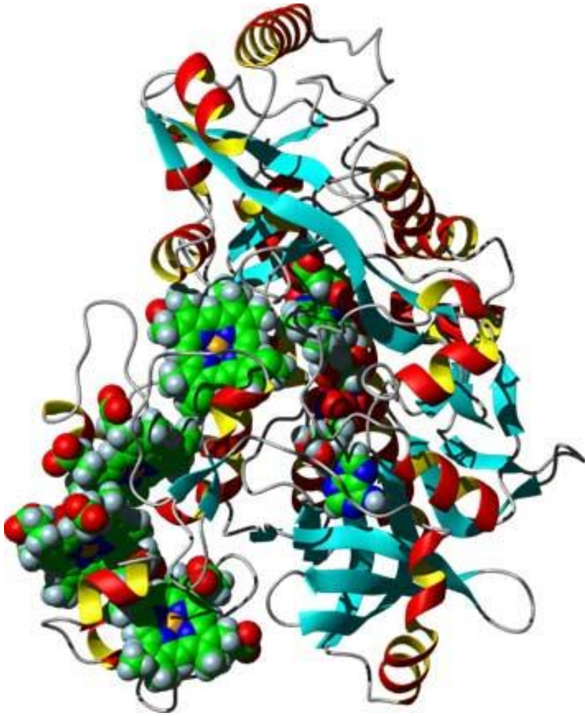
transcription

RNA



translation

Proteins



ChIP to understand transcriptional regulation!

Map regulatory elements:

Transcription Factors

–ChIP

Histone marks

–ChIP

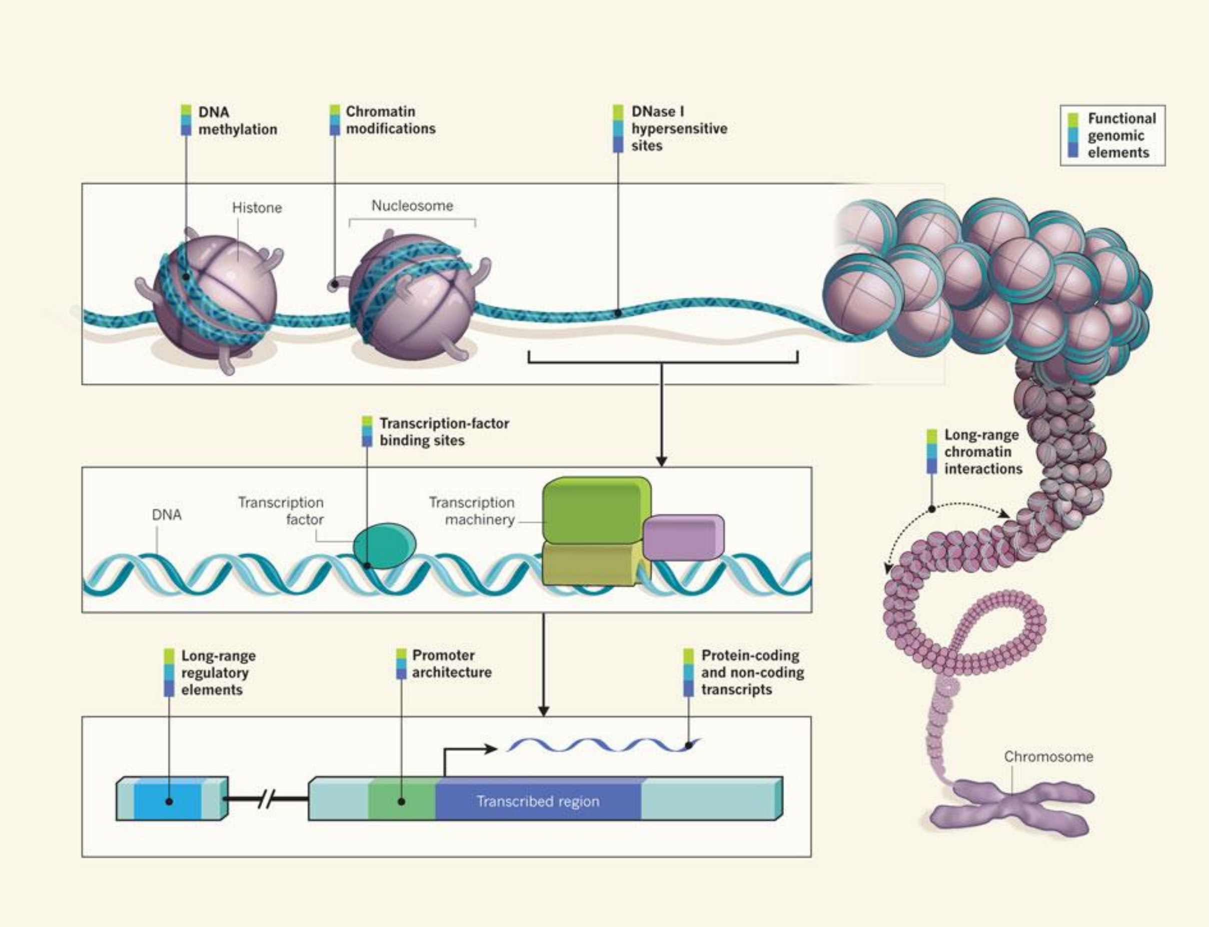
DNA Methylation

–MeDIP etc.

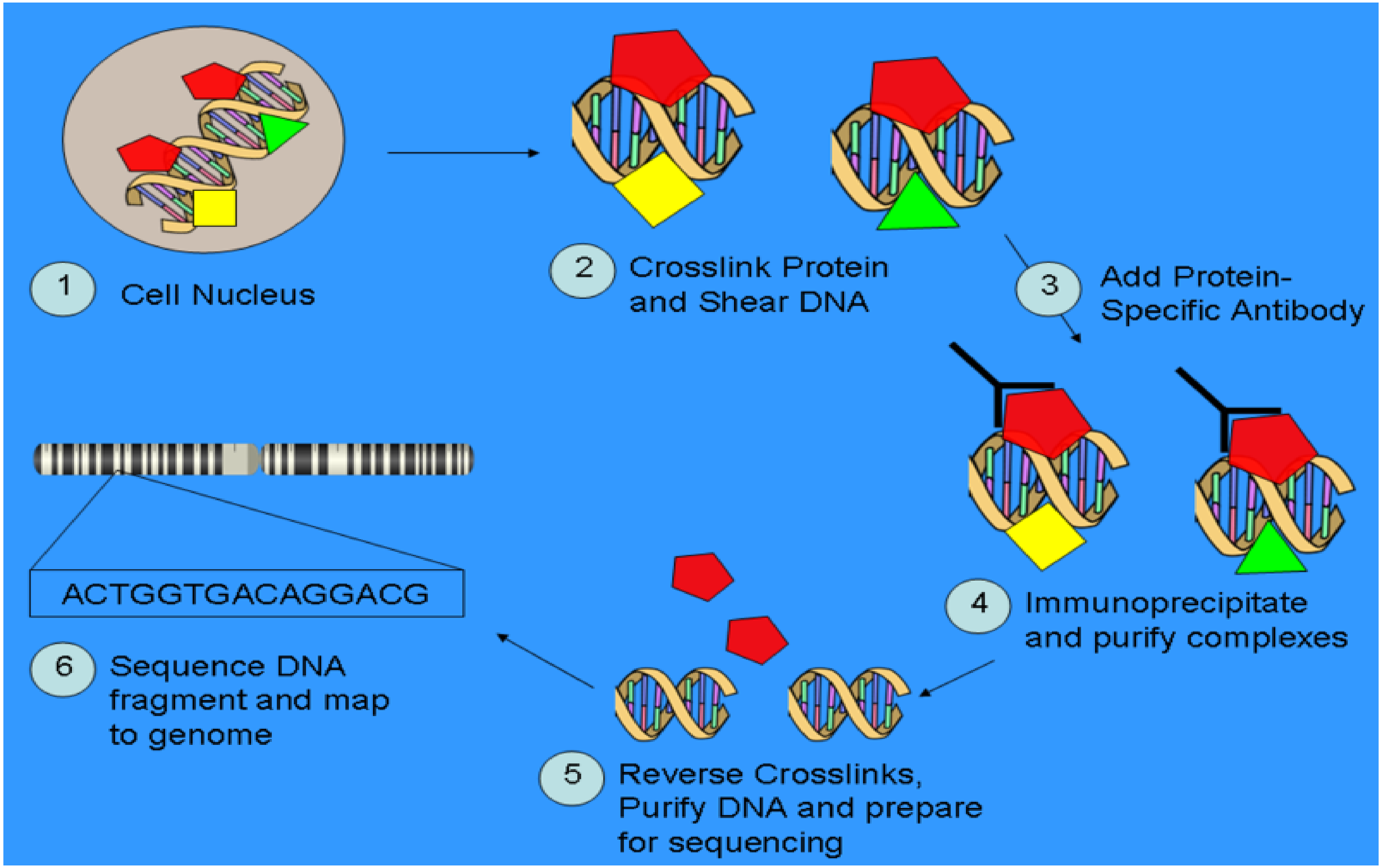
Nucleosomes

RNA Polymerase

–Pol II ChIP


DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
Long-range chromatin interactions
DNA
Transcription factor
Transcription machinery
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

ChIP-seq protocol


1 Cell Nucleus
2 Crosslink Protein and Shear DNA
3 Add Protein-Specific Antibody
4 Immunoprecipitate and purify complexes
5 Reverse Crosslinks, Purify DNA and prepare for sequencing
6 Sequence DNA fragment and map to genome
ACTGGTGACAGGACG

Analysis of ChIP-seq data

Experimental design
- –Controls and replicates

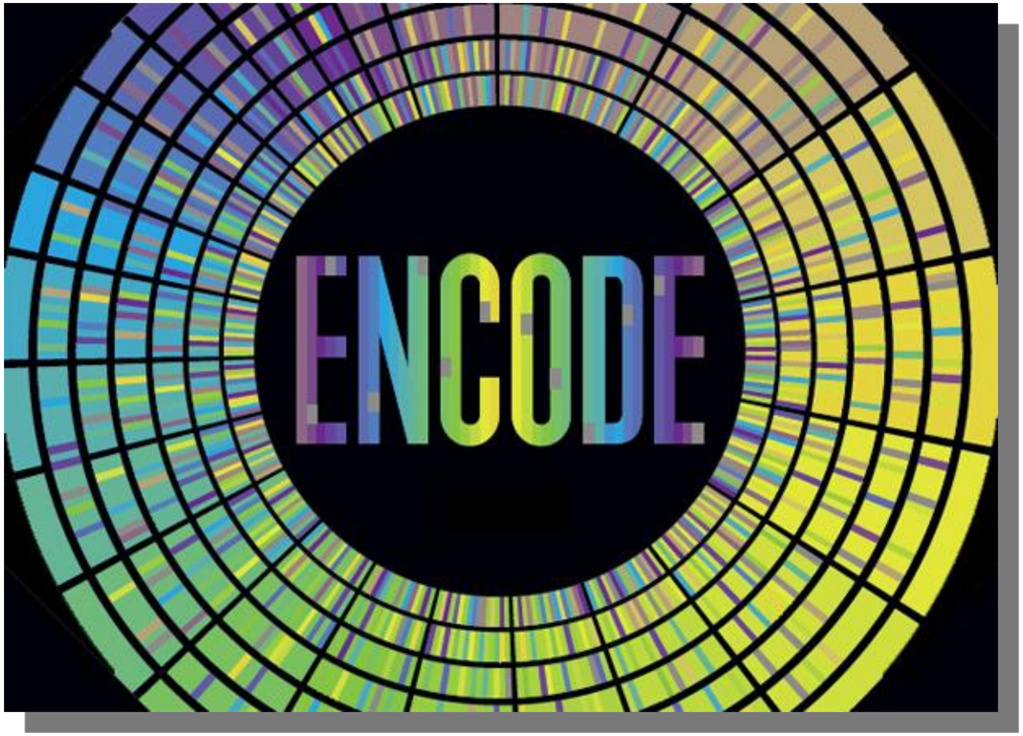
QC/Read processing
- –Library QC
- –Alignment and filtering
- –QC measures and assessment

Peak calling
- –Peak callers

Differential binding analysis
- –Occupancy-based analysis
- –Affinity-based analysis

Validation and downstream analysis
- –Motif analysis
- –Annotation
- –Integrating binding and expression data

ENCODE project



Landt et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE Consortia. Genome Research 22: 1813-1831

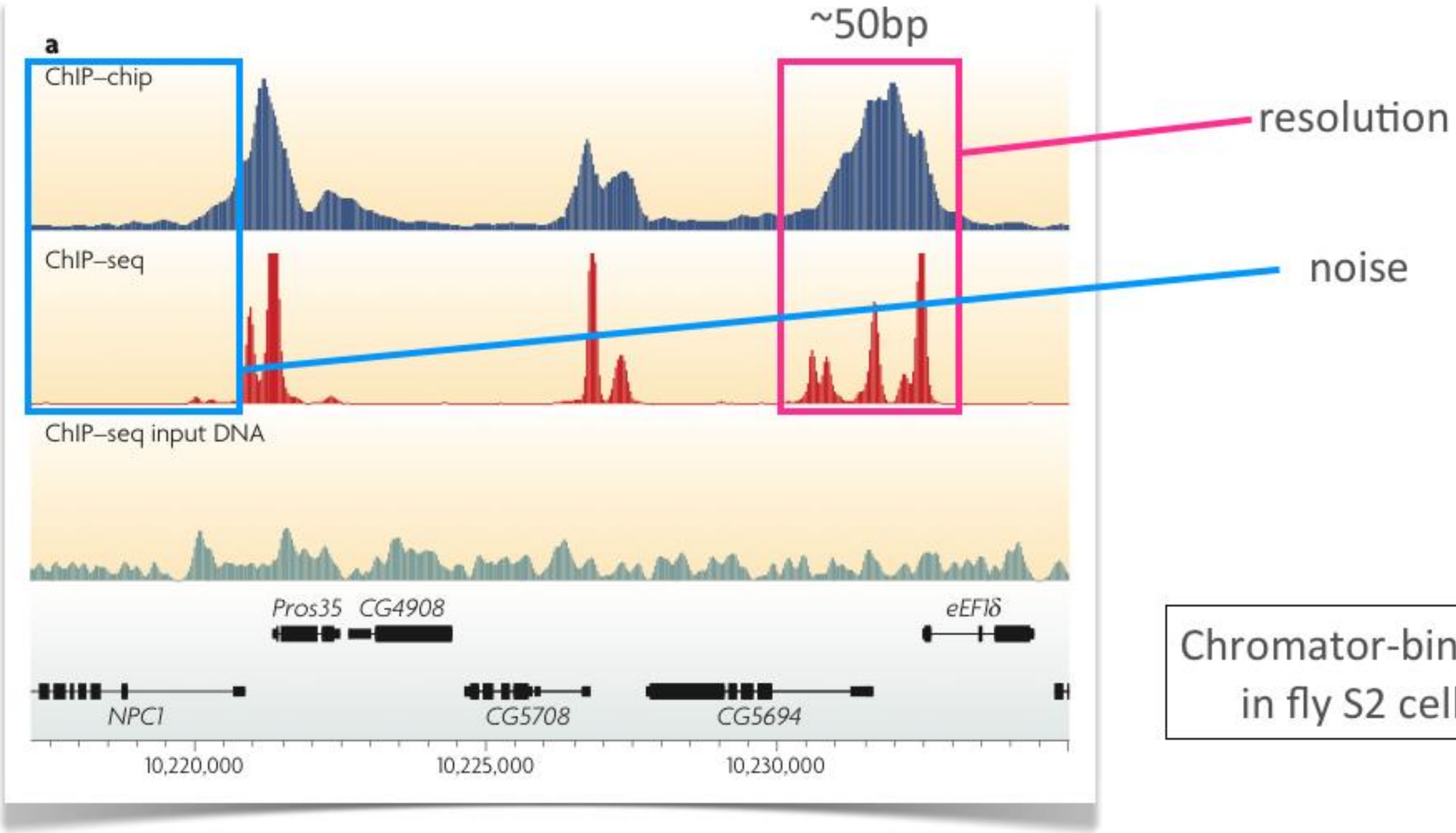
Chen et al. (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods 9: 609

Comparison of ChIP-chip and ChIP-Seq

Comparison of ChIP-chip & chip-Seq

	ChIP-chip	Chip-Seq
coverage	limited by array (genome size, repeats)	whole genome
resolution	30~150bp (array specific)	1bp
noise	cross-hybridisation	sequencing errors and bias
dynamic range	~100x	~10,000x
sample amount	>2micrograms	10-50ng
cost	$400-800 per array	$1000 per lane

Comparison of ChIP-chip & ChIP-Seq



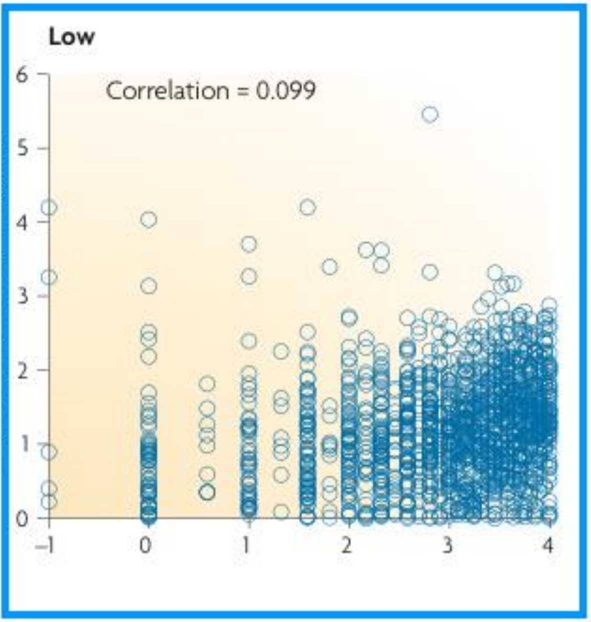
Chromator-binding in fly S2 cells

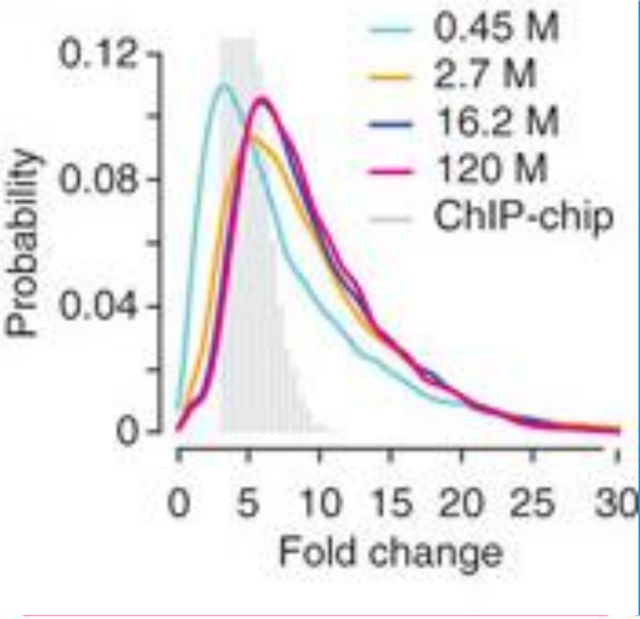
[Park, 2009]

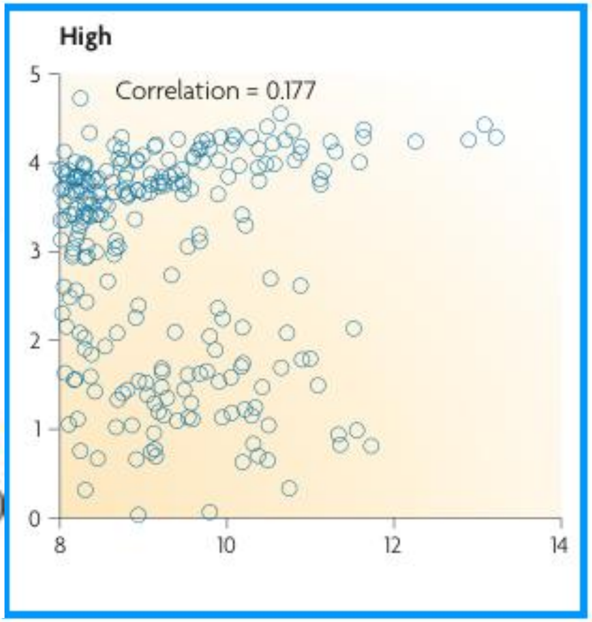
Comparison of ChIP-chip & ChIP-Seq

good correlation for medium expression

poor correlation for high and low expression







Yeast array and RNA-Seq data

[Wang et al, 2008]

Experimental considerations for down-stream analysis

Consideration 1: How good is your antibody?

- ChIP-Seq data depend on antibody quality
- modENCODE project:
  - large-scale screening for histone modifications in flies
  - 20-35% of commercial 'ChIP-grade' antibodies were unusable
- variations between antibodies
  - differences in antibody specificity can make it hard to compare data across multiple transcription factors
  - efforts are made to have a list of 'approved' antibodies for histone modifications

[Celniker et al 2009; Vaquerizas et al, 2008; Egelhofer et al 2011]

Consideration 2: Why do you need controls?

- Controls can be generated by:
  - (cross-linking), lysing and fragmenting the cells but not continuing with IP (that's the most popular way of generating a control sample)
  - (cross-linking), lysing and fragmenting the cells and performing a mock IP (IP without antibody)
  - performing an IP with an antibody that is not known to be involved in DNA or chromatin binding (e.g. IgG)

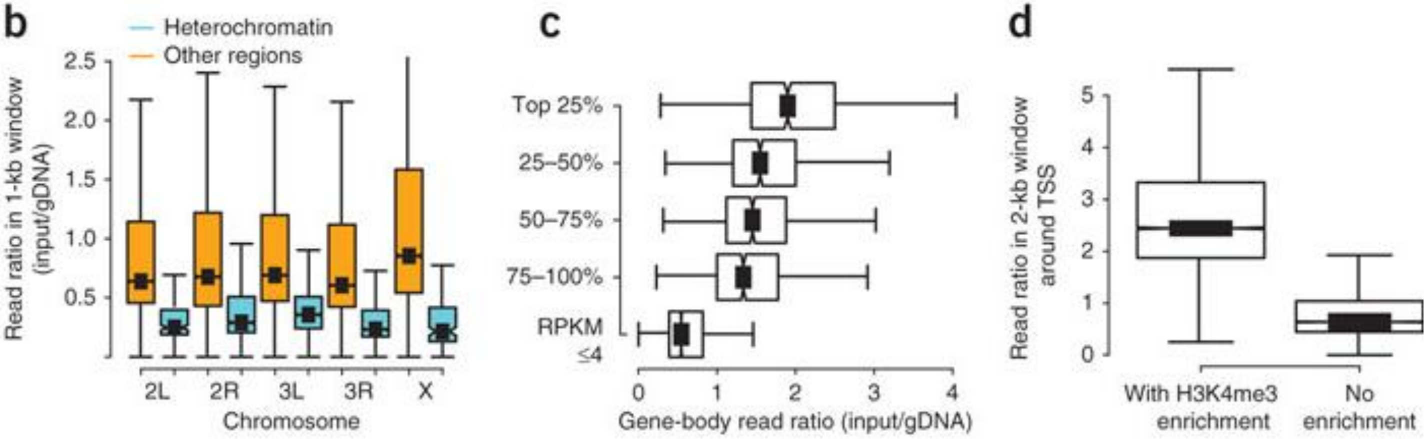
  - (if the genome of the sample being studied has been sequenced using similar technology, one can possibly use this as a control)

Consideration 2: Why do you need controls?

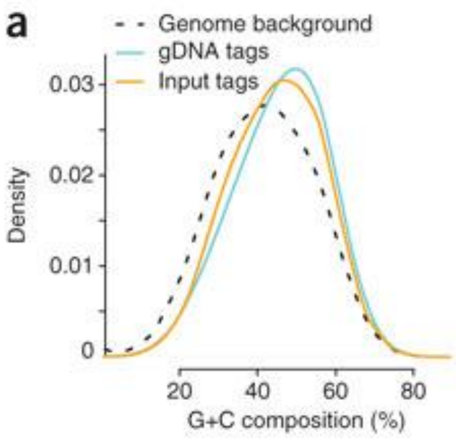
- skipped in early experiments:
  - cost
  - over-confidence in ChIP-Seq data quality
- But there are artefacts from sample preparation & sequencing
  - copy number variation
  - non-uniform fragmentation
  - non-specific pull-down
  - incorrect mapping of repetitive genomic regions
  - GC sequencing bias (http://beads.sourceforge.net [Cheung et al 2011])
- problems become more acute in larger genomes

Consideration 2: Why do you need controls?

- Non-uniform fragmentation (euchromatin-heterochromatin)

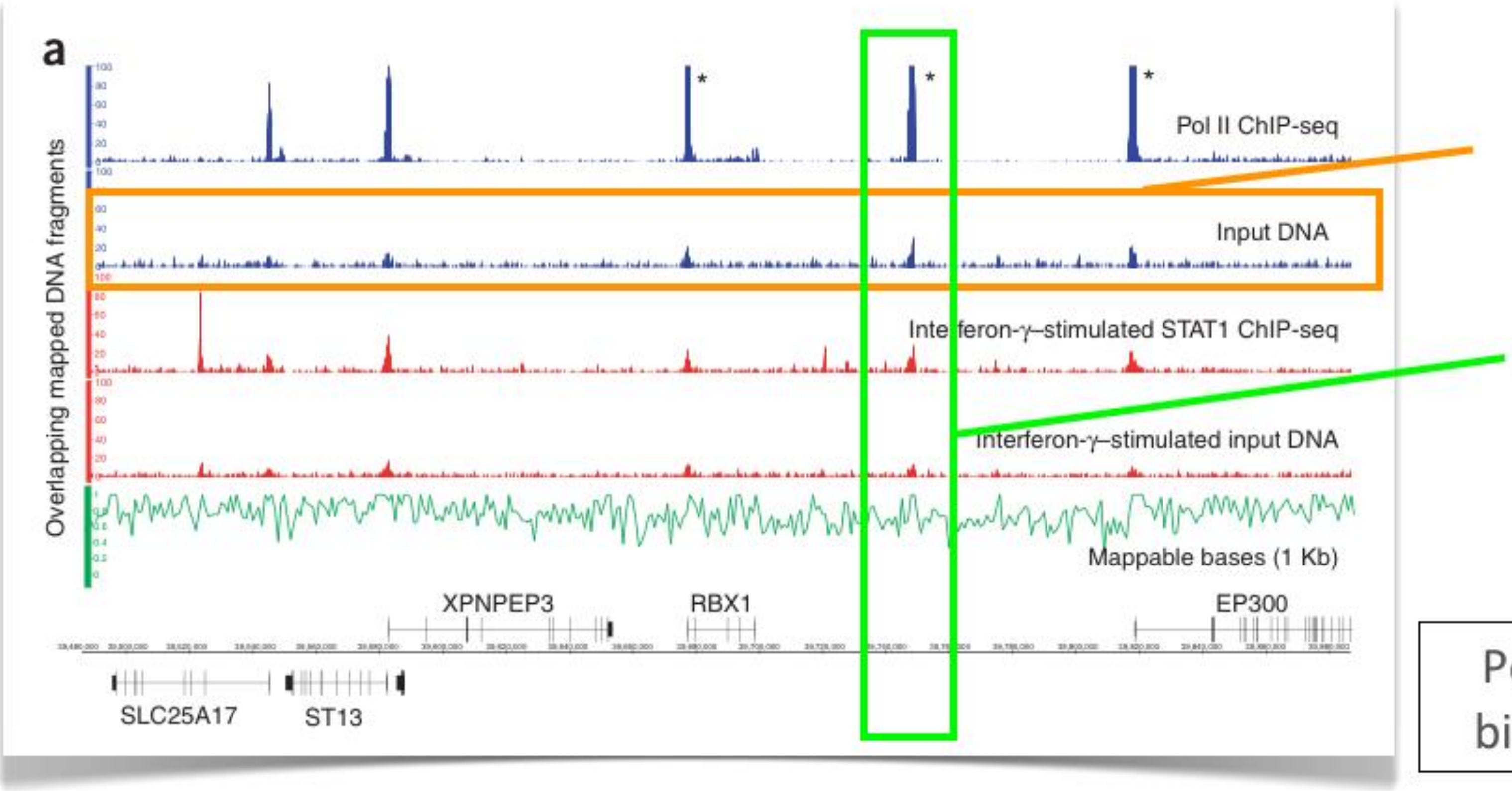


- GC sequncing bias



[Chen et al, 2012]

Consideration 2: Why do you need controls?



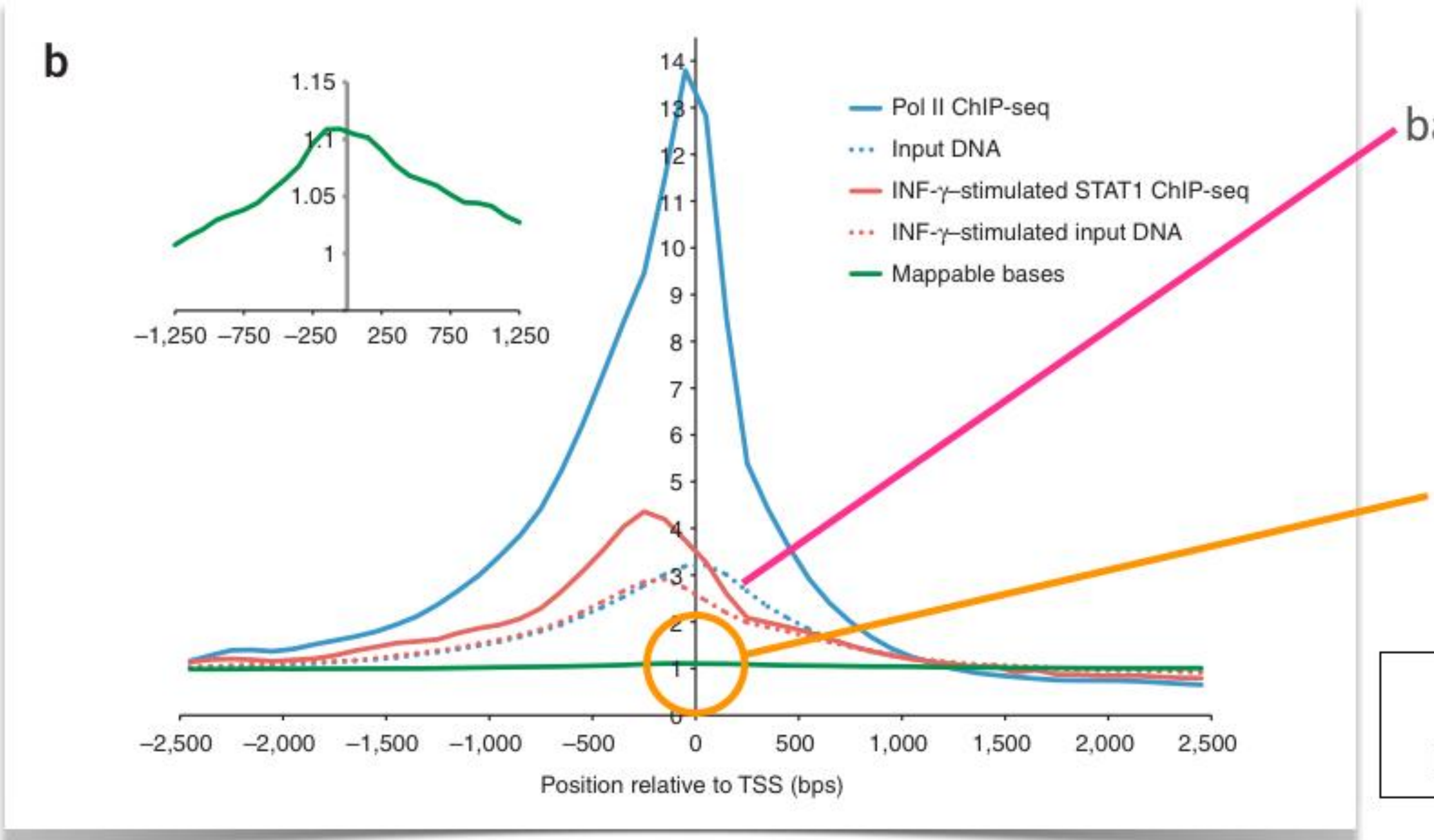
input DNA

peaks align

Pol II & STAT1-binding in HeLa

[Rozowsky et al, 2009]

Consideration 2: Do you need controls?



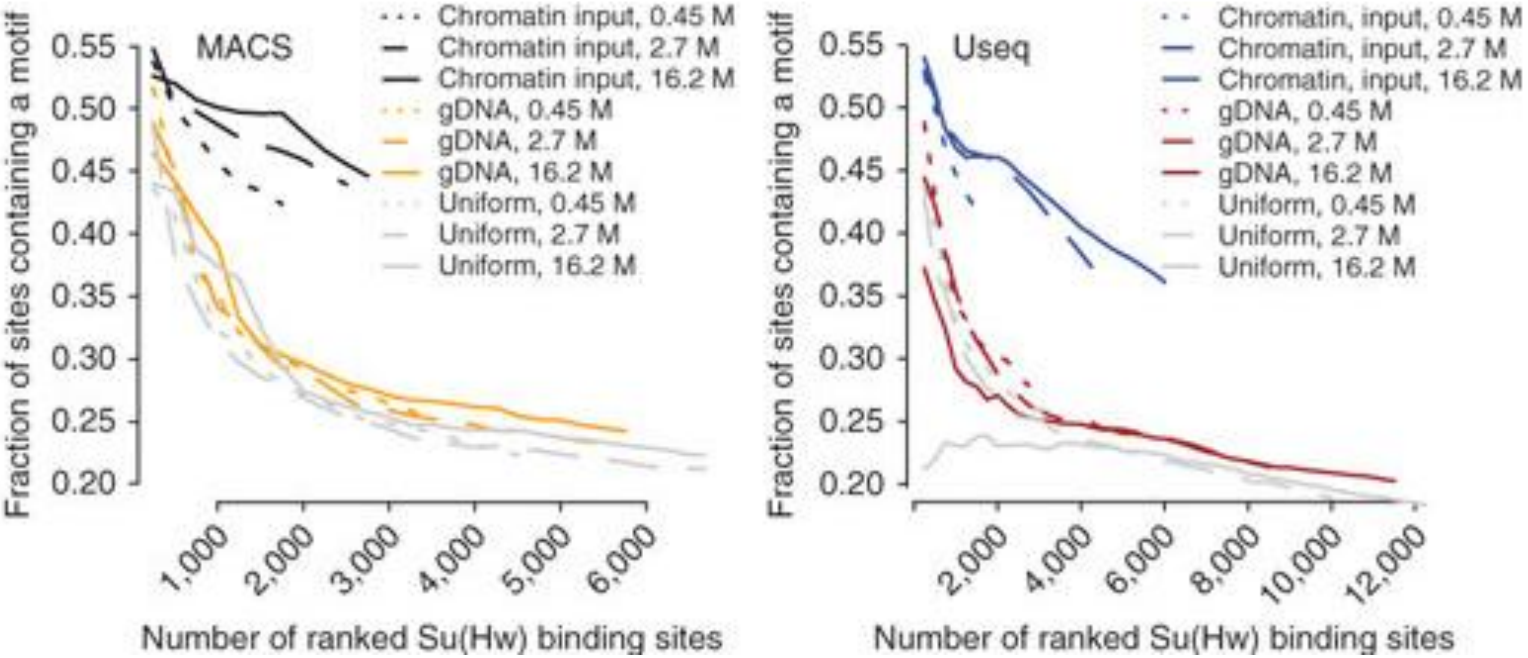
background signal

align signal relative to TSS

Pol II & STAT1-binding in HeLa

[Rozowsky et al, 2009]

Consideration 2: Why do you need controls?



The more sequencing depth you have for the input the better you can identify peaks!

[Chen et al, 2012]

Consideration 3: Sequencing depth

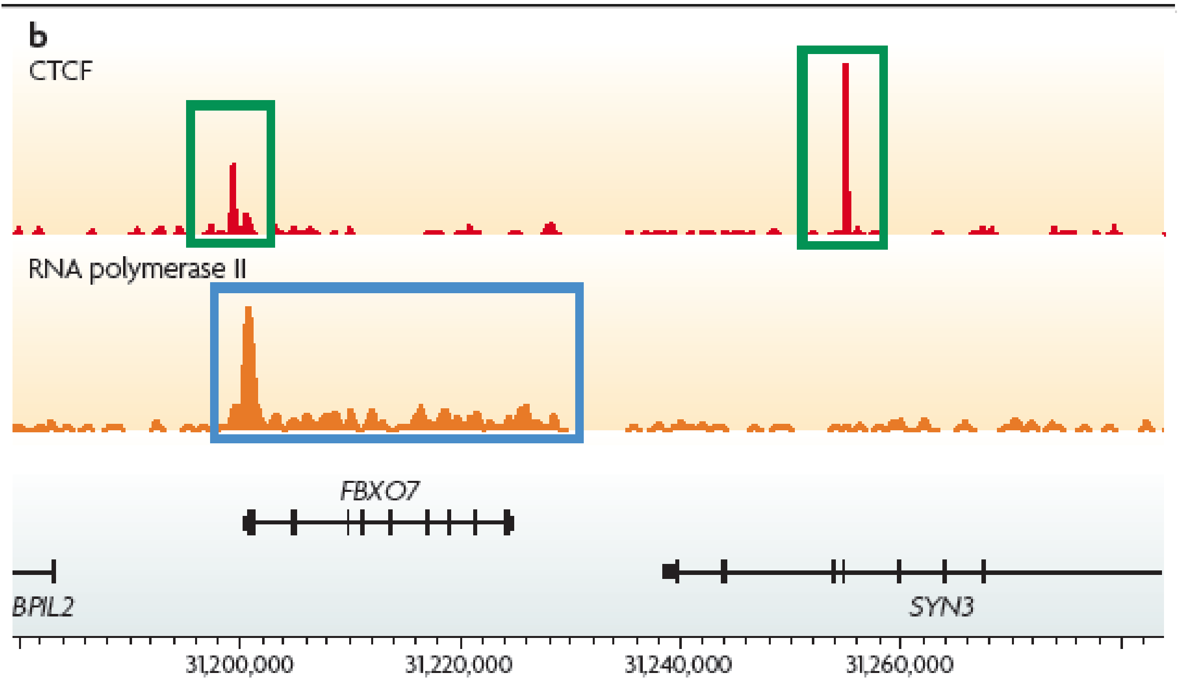
- sequencing depth depends on genome size, protein & biological question
- one lane gives ~35 million reads (over 100 million reads – HiSeq)
  - ~270x genomic coverage for bacteria
  - ~10x coverage for fly
  - ~0.4x coverage for human
- proteins bind genome in different ways
  - chromatin & Pol II cover the genome
  - sequence-specific TFs are more confined

Consideration 3: Sequencing depth

Proteins bind in different ways

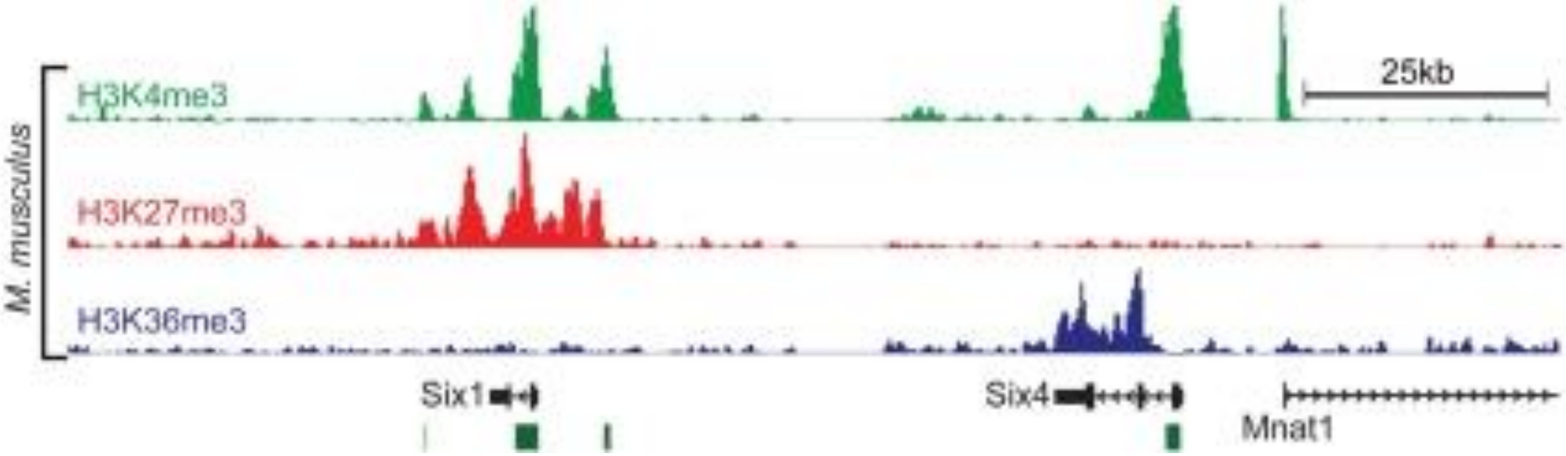
Transcription factor – tight, highly-peaked binding region

RNA PolII – enriched at TSS but bound throughout gene body



ChIP-Seq data from fly S2 cells

Consideration 3: Sequencing depth



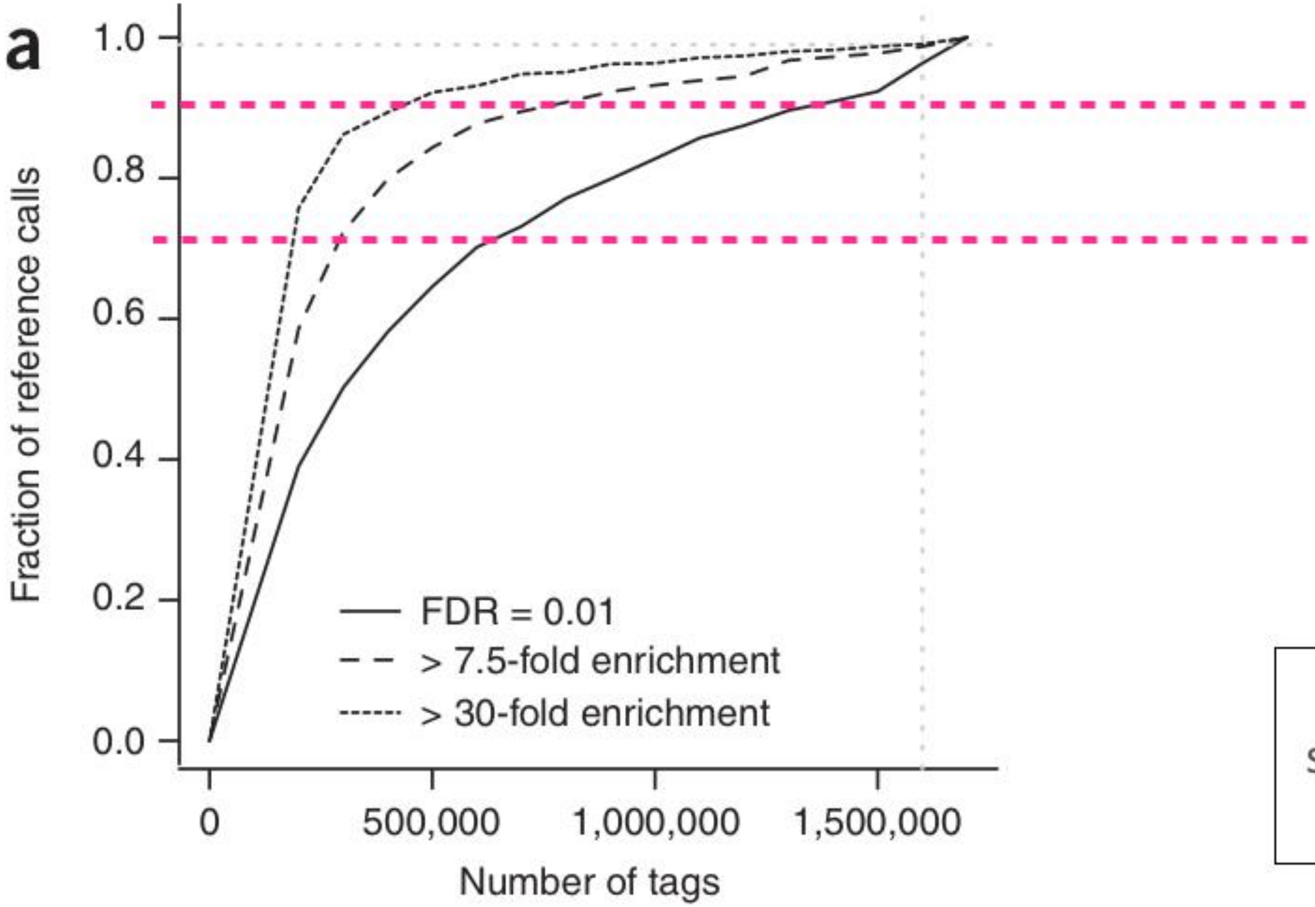
Activating mark (near TSS)

Peaks within body of inactive genes

Peaks within body of active genes

[Ku et al, 2009]

Consideration 3: Sequencing depth

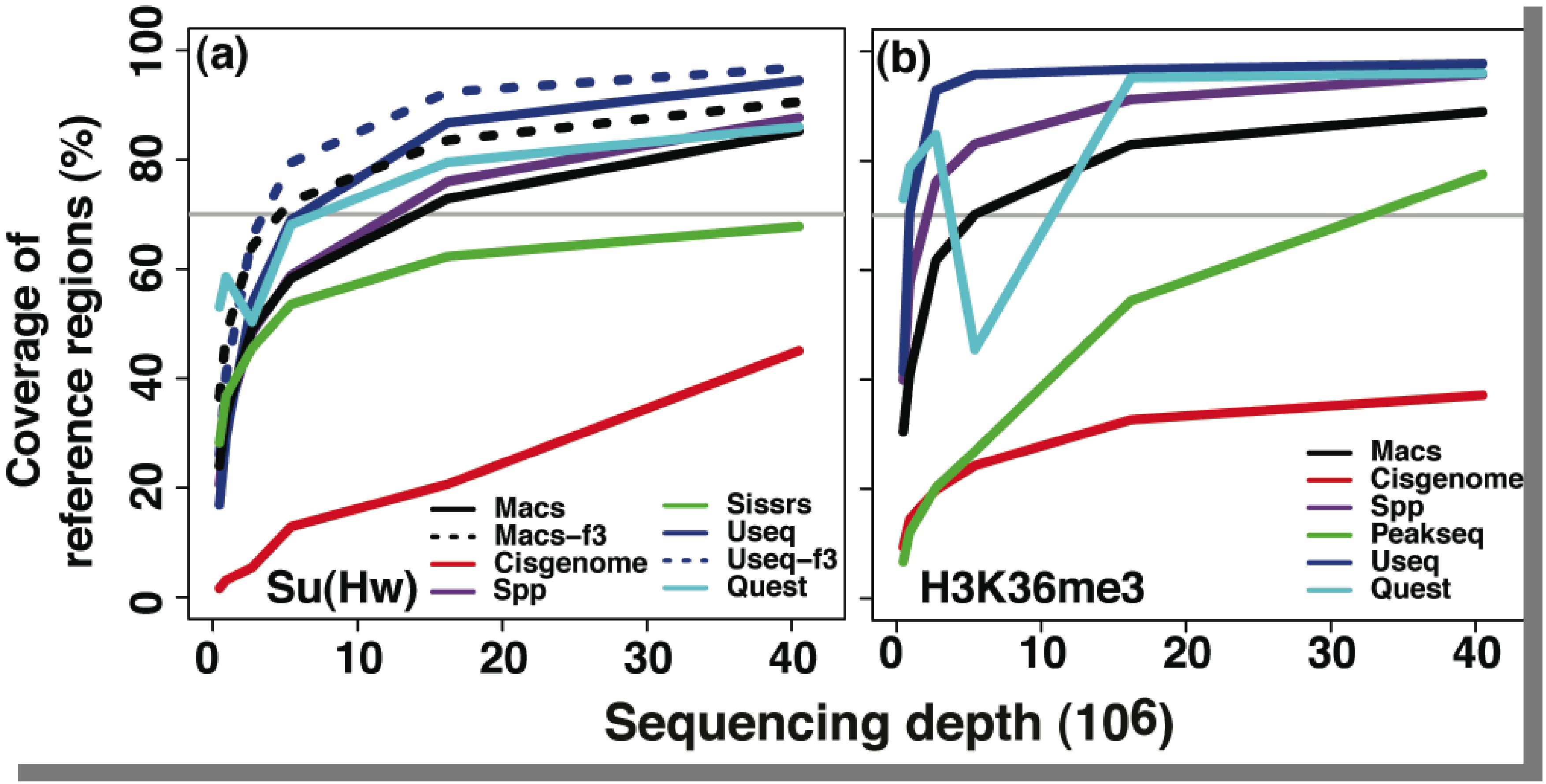


simulation of sequence depth v binding sites

[Kharchenko et al, 2008]

Consideration 3: Sequencing depth

(optimum is different for different peak finder software)



Plateau for most peak finders ~16.2 M reads in Drosophila (corresponding to ~327 M reads in human)

[Chen et al, 2012]

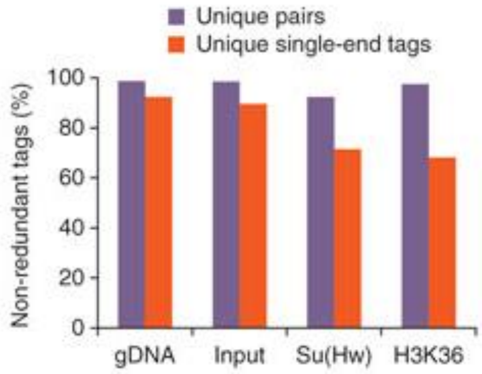
- How many replicates?

Reproducibility information gives confidence in peaks, helps choosing thresholds (IDR)

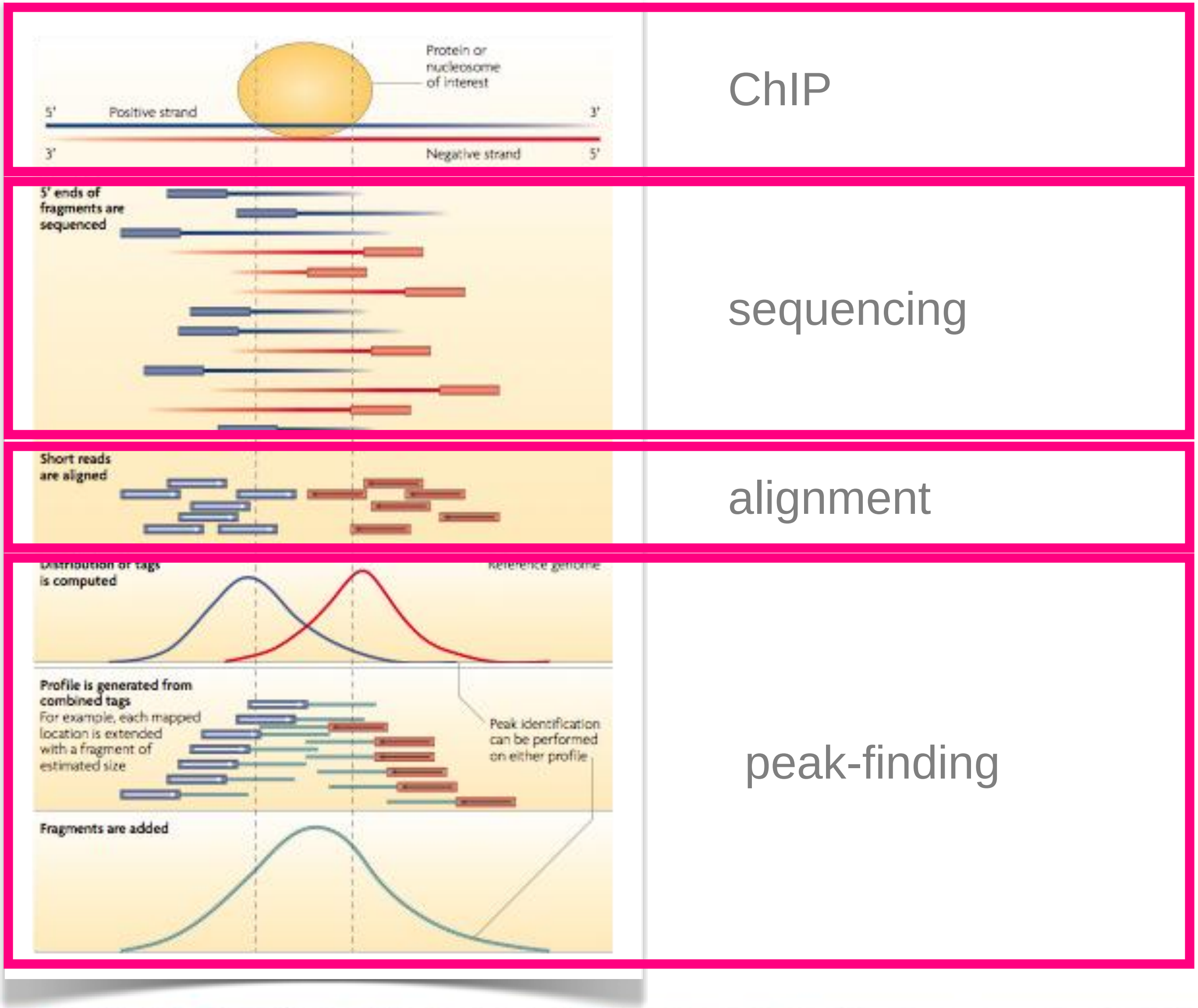
- How many reads do you need?
  - The more the better!

- How long should reads be?

- Do you need paired end reads?
  - Can help with mapping but not nearly as important as for identifying indels in DNA sequencing or multiple isoforms in RNA-seq (can be important for proteins/modifications that are in repetitive elements)
  - There is a difference when you assess the complexity of the sample



Data processing steps


Protein or nucleosome of interest
5'
Positive strand
3'
3'
Negative strand
5'
5' ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

ChIP

sequencing

alignment

peak-finding

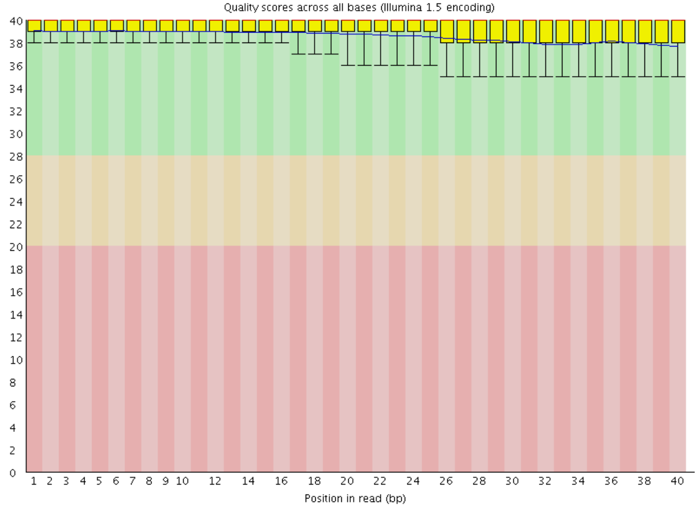
schematic of ChIP-seq experiments

[Park et al, 2009]

Quality control

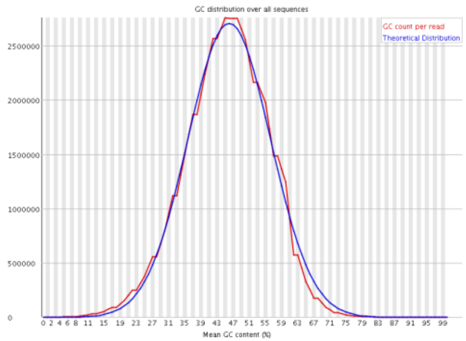
Many tools (SAMstat, htSeqTools, fastQC etc.)

- Read quality

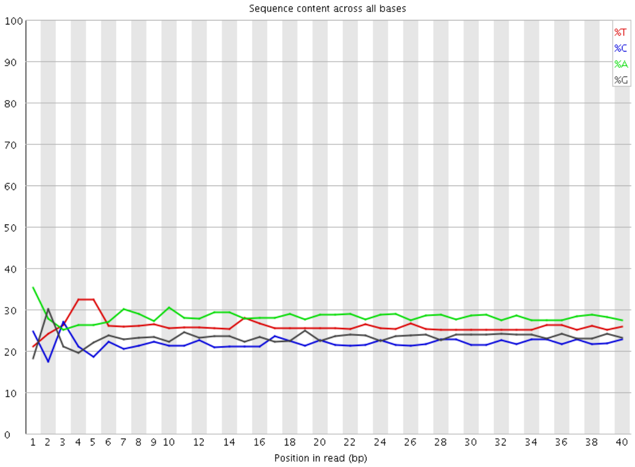


- Sequence content
- Duplication (PCR artefacts)
- Library complexity (overrepresented sequences)
- Contamination

Quality control



Many tools (SAMstat, htSeqTools, fastQC etc.)

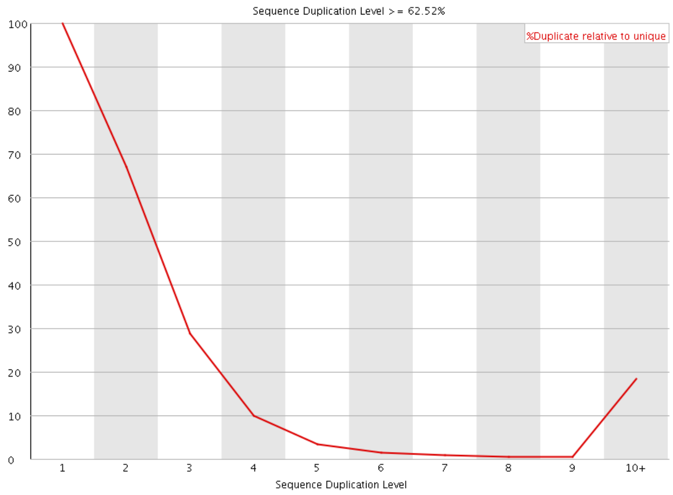


- Read quality
- Sequence content
- Duplication (PCR artefacts)
- Library complexity (overrepresented sequences)
- Contamination

Quality control

Many tools (SAMstat, htSeqTools, fastQC etc.)

- Read quality
- Sequence content
- Duplication (PCR artefacts)

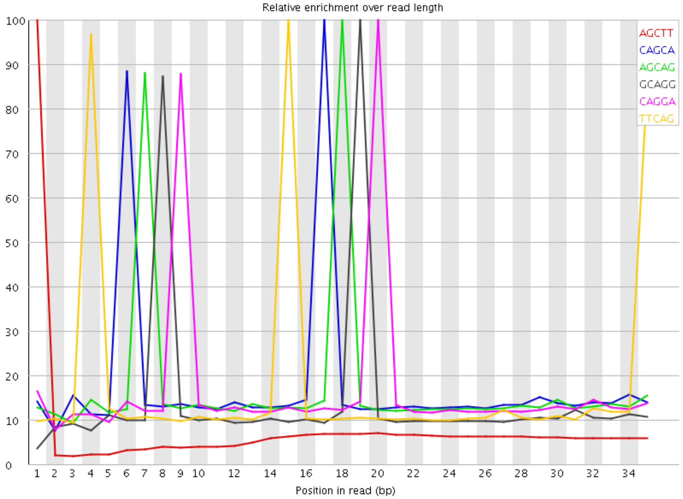


- Library complexity (overrepresented sequences)
- Contamination

Quality control

Many tools (SAMstat, htSeqTools, fastQC etc.)

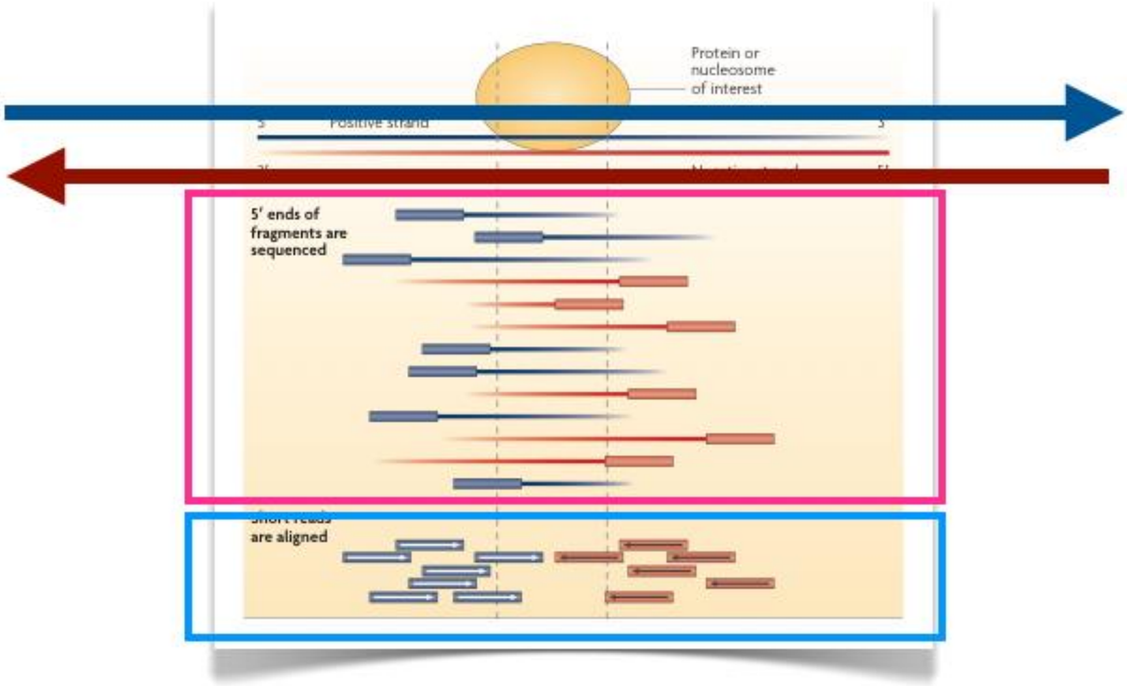
- Read quality
- Sequence content
- Duplication (PCR artefacts)
- Library complexity (overrepresented sequences)
- Contamination



Genome alignment

- Choice of software depends on:
  - accuracy, speed, memory, flexibility e.g. BWA, Bowtie
- Questions:
  - allow for mis-matches between reads and reference genome?
    - (if you are interested in allele-specific binding care must be taken, since in some regions reads containing the non-reference allele might not be aligned well)
  - multiple matches to reference?

Genome alignment

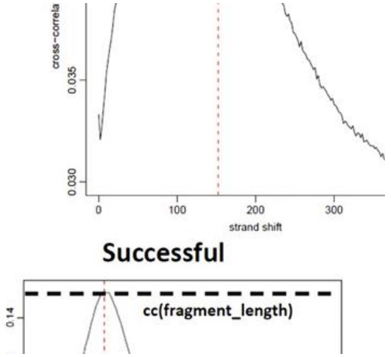

Protein or nucleosome of interest
5'
Positive strand
3'
3'
Negative strand
5'
5′ ends of fragments are sequenced
Short reads are aligned

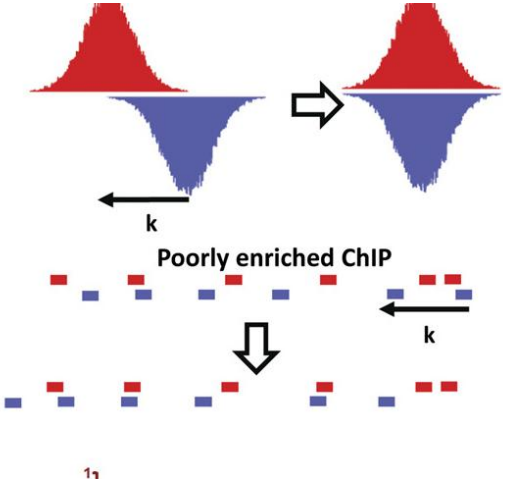
sequencing occurs 5′ to 3′ on both strands

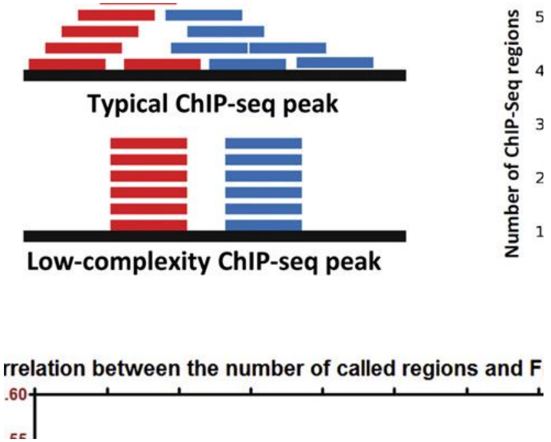
positive and reverse strand sequences

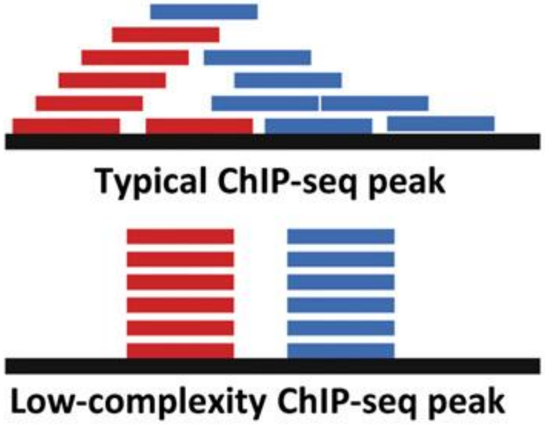
sequences map to either strand

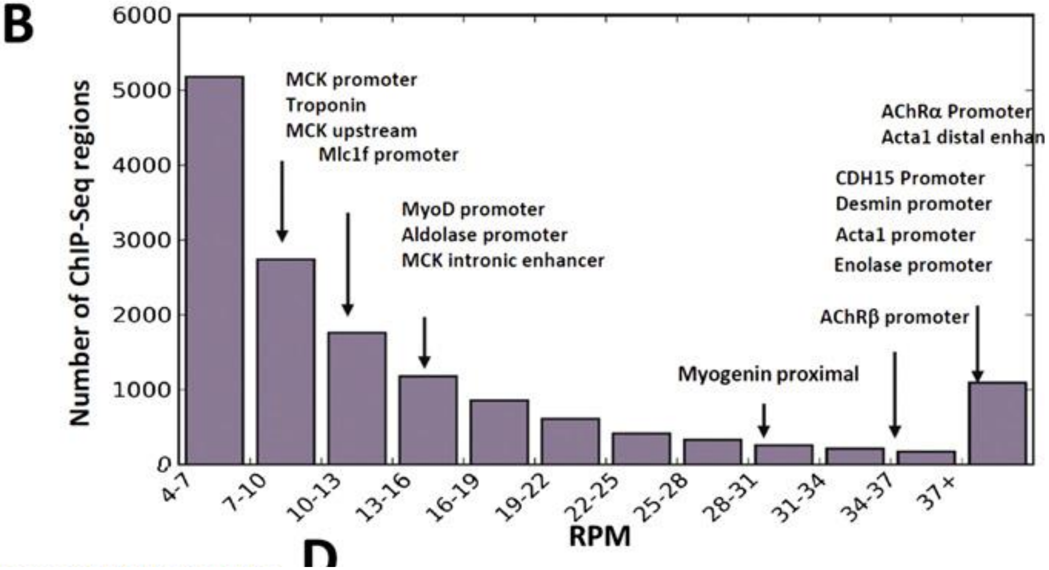
Strand information for quality control





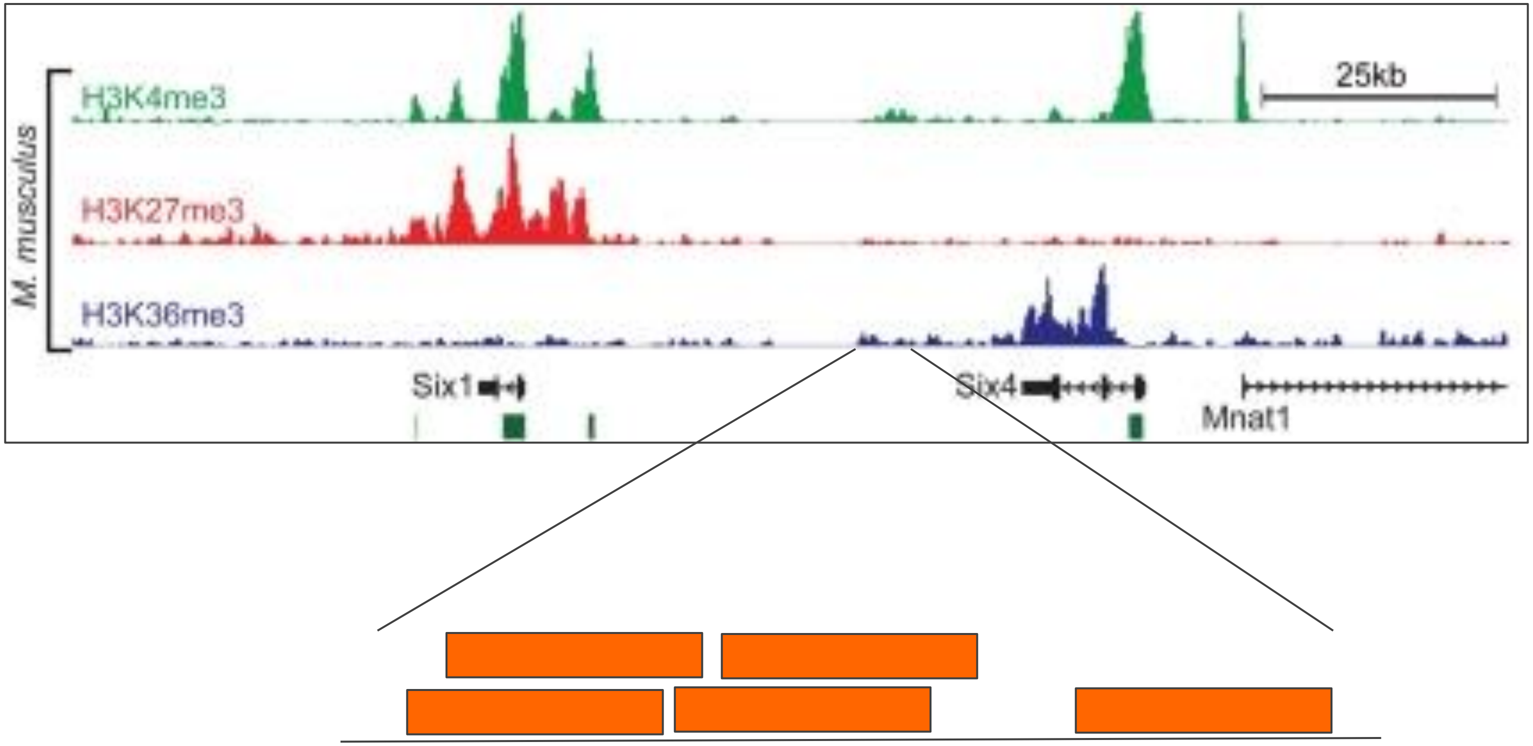

Typical ChIP-seq peak
Low-complexity ChIP-seq peak
Number of ChIP-Seq regions
5
4
3
2
1
rrelation between the number of called regions and F
.60





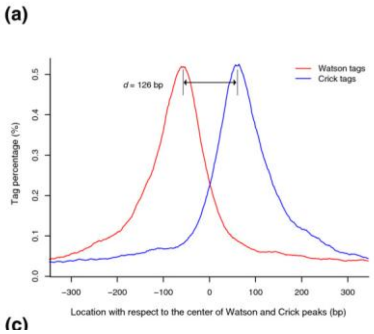
[Landt et al, 2012]

Peak-finding



Basic idea: Count the number of reads in windows and determine whether this number is above background – if so, define that region as bound

MACS 2.0



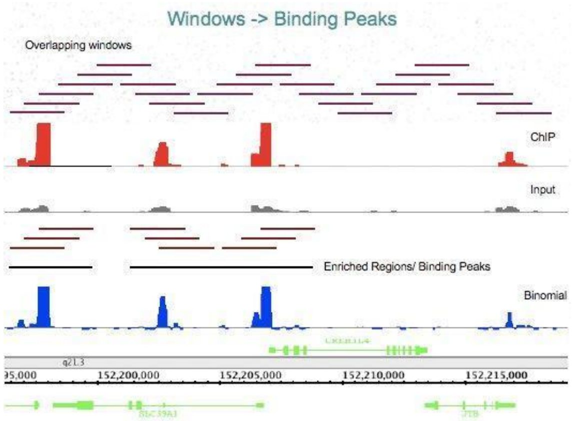
Calculating peakshift for 1000 best peaks

Shift reads 3'

Identify potentially bound regions

Calculate enrichment and significance using poisson distribution with local λ

USeq



Calculating peakshift

Shift reads 3'

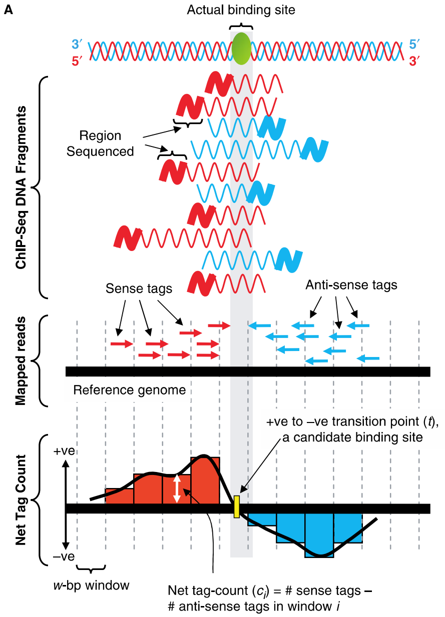
Define windows

Calculate enrichment per window, significance using negative binomial

Join regions that are within max gap

eFDR

SISSRs



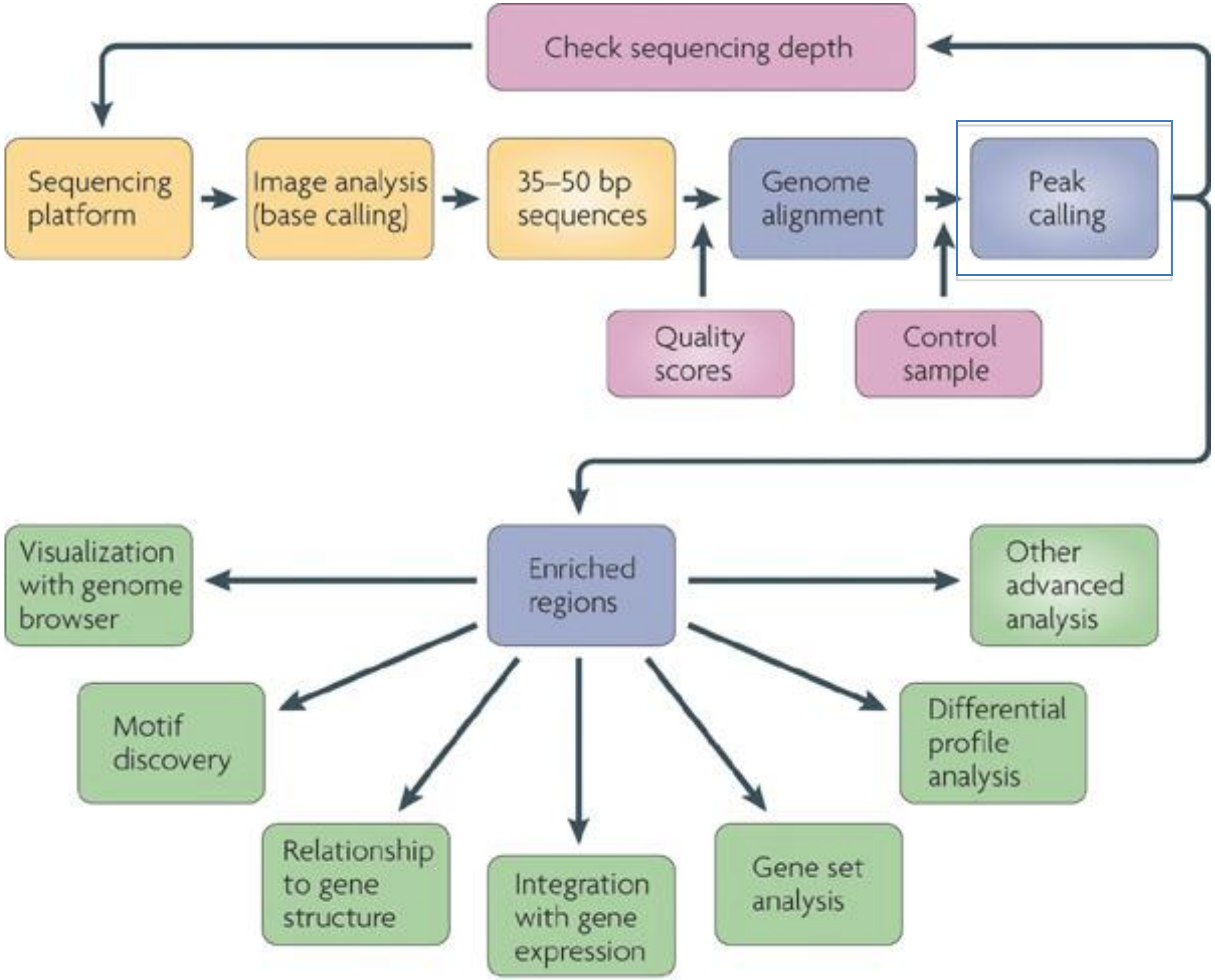
Estimate fragment length (mean sense-antisense dist)

Windows with w/2 shift through genome

Define potential peaks by transition in net tag count $(n_{sense}-n_{antisense})$

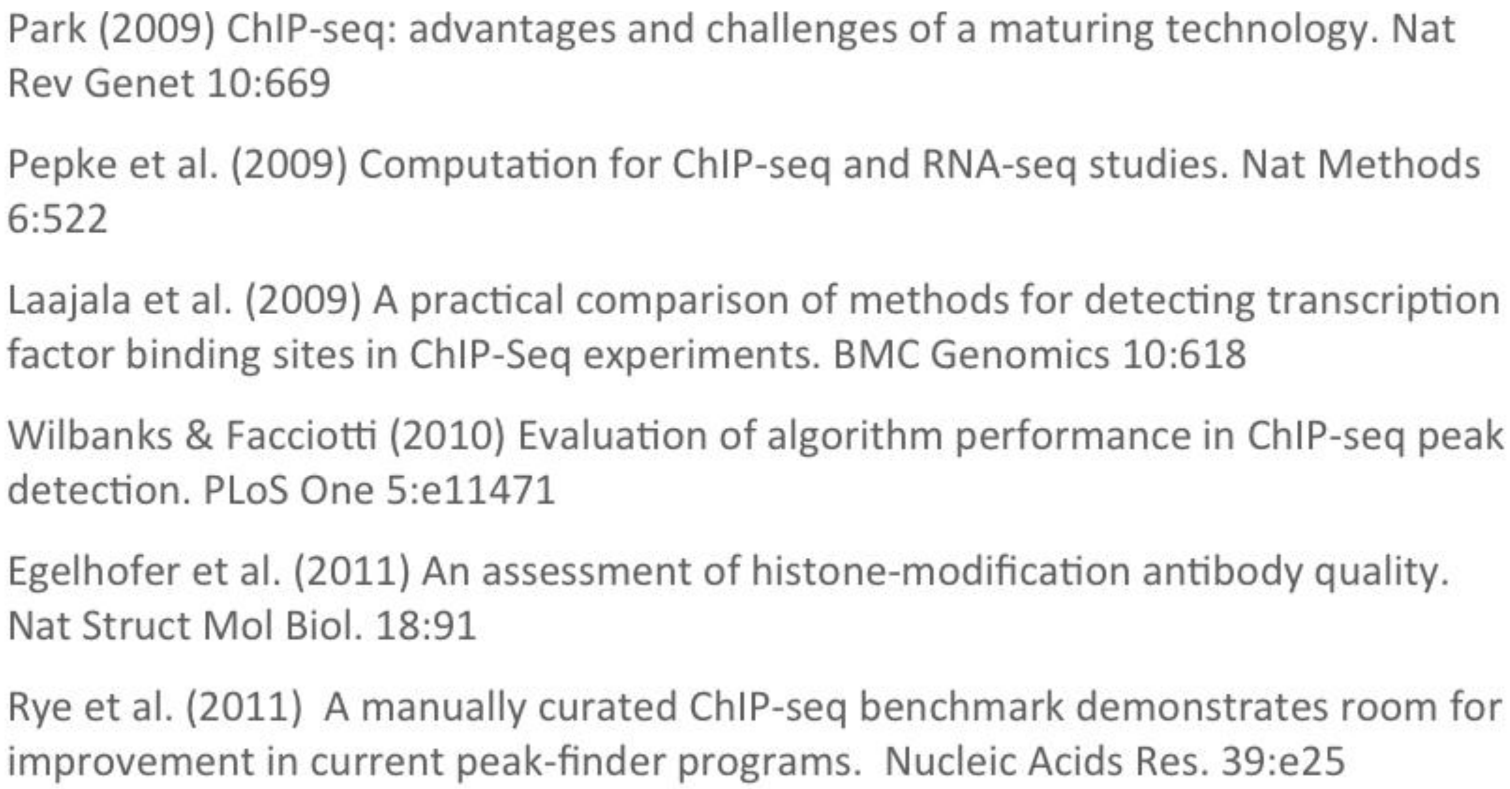
Calculate enrichment and significance using poisson

Downstream of ChIP



[Park 2009]

References:



Park (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669

Pepke et al. (2009) Computation for ChIP-seq and RNA-seq studies. Nat Methods 6:522

Laajala et al. (2009) A practical comparison of methods for detecting transcription factor binding sites in ChIP-Seq experiments. BMC Genomics 10:618

Wilbanks & Facciotti (2010) Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One 5:e11471

Egelhofer et al. (2011) An assessment of histone-modification antibody quality. Nat Struct Mol Biol. 18:91

Rye et al. (2011) A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. Nucleic Acids Res. 39:e25

Landt et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE Consortia. Genome Research 22: 1813-1831

Chen et al. (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods 9: 609

Meyer & Liu (2014) Identifying and mitigating bias in next generation sequencing methods for chromatin biology. Nature Reviews Genetics doi:10.1038/nrg3788