KTH Royal Institute of Technology
School of Biotechnology

Analysis of data from high-throughput molecular biology experiments BB2490

Transcriptomics: ChIP-seq

Olof Emanuelsson
olofem@kth.se

Lecture 7 BB2490 2014-02-03 10:15-12:00 FD41

Enabler for Life Sciences

Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

Today's lecture:

1. **The goal of life science research**
2. **Regulation of gene expression**
3. **ChIP-seq, experimental procedure**
4. **ChIP-seq, bioinformatics**
5. **Summary**

[1] The goal of life science research

Life science research goals:

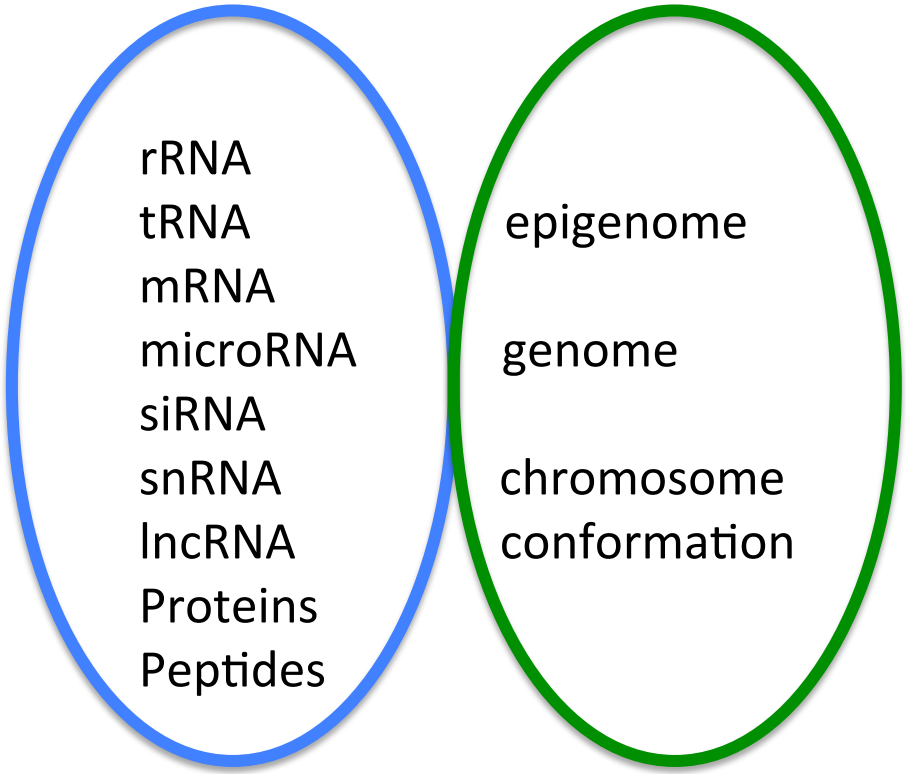
To understand cells, tissues, organisms, populations: the way they function, how they develop, how they interact with each other, how they respond to various stimuli, what physiological and molecular mechanisms are present.

How they [cells, tissues, organisms] function and why they are different.

Molecular biology approach

In molecular biology, we are interested in recording the molecular state of a cell or a collection of cells, i.e., tissue or sample.

This tells us a lot about the functions of the cell.



Thus, we investigate the genome, the epigenome, the transcriptome, and the proteome.

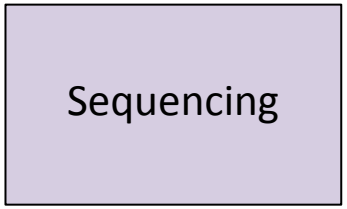
Investigate via DNA sequencing

Several molecular states and processes in the cell can be interrogated via sequencing of DNA.

Biological question → DNA sequencing → Answer or readout

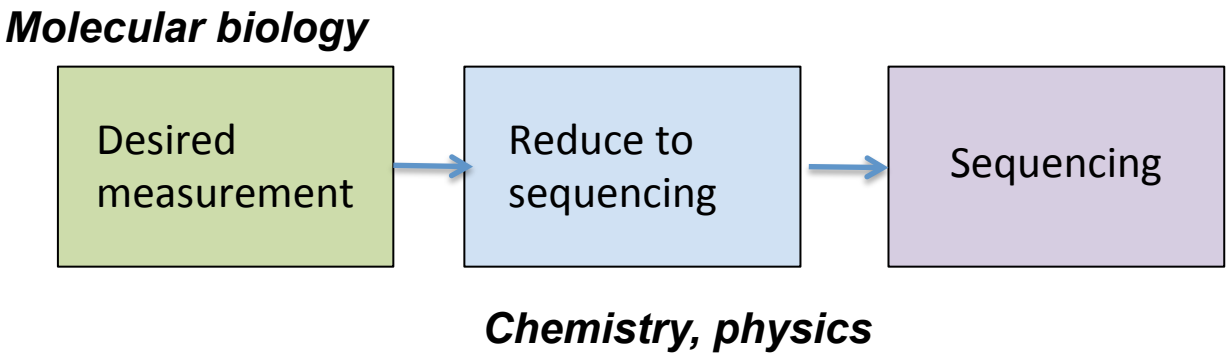
A general scheme for DNA-sequence based interrogation:

`*-seq`



A general scheme for DNA-sequence based interrogation:

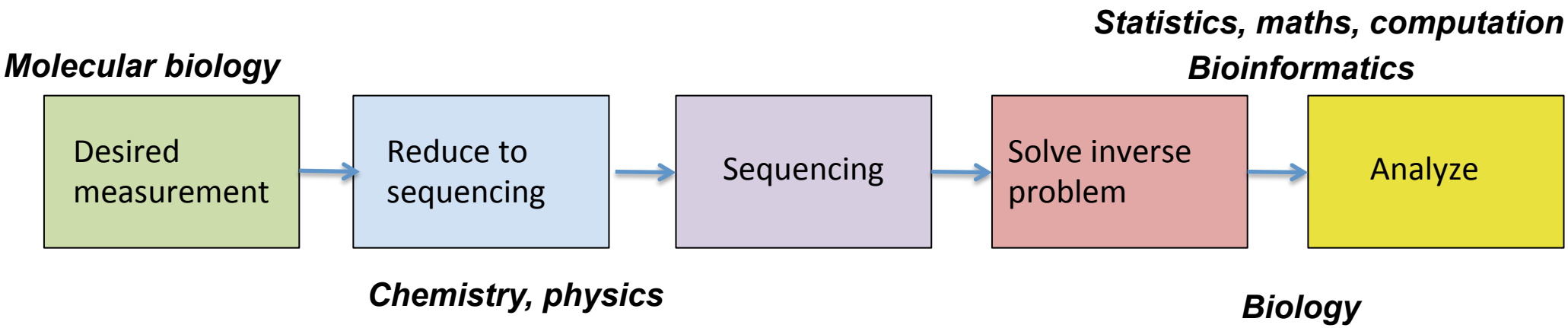
*-seq



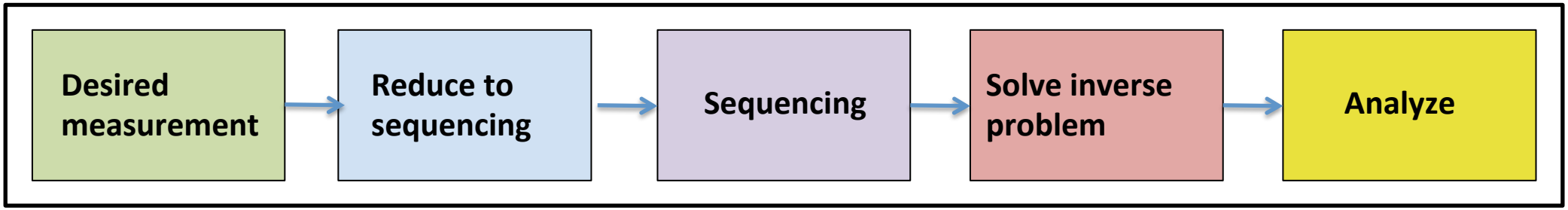
[adopted from Lior Pachter, UC Berkeley]

A general scheme for DNA-sequence based interrogation:

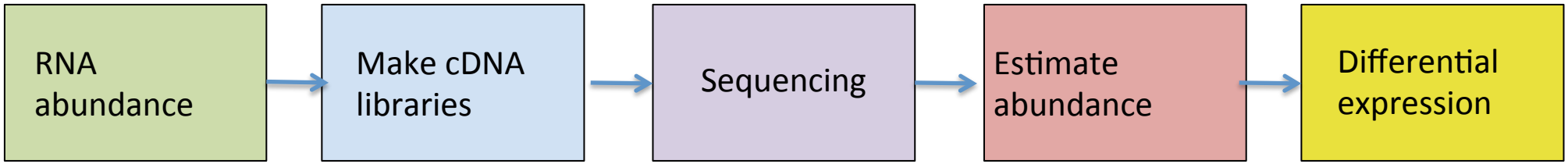
*-seq



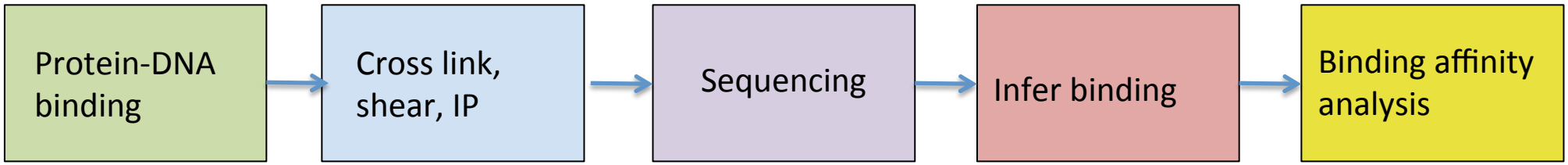
[adopted from Lior Pachter, UC Berkeley]



RNA-seq [*RNA sequencing*: Abundance of RNA molecules]

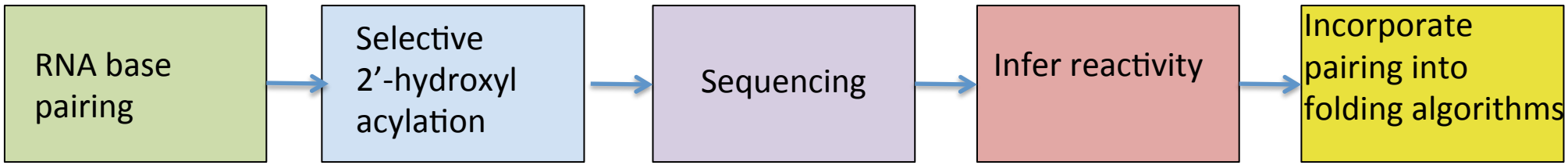


ChIP-seq [*Chromatin Immunoprecipitation sequencing*: Binding of transcription factors; map chromatin modifications]

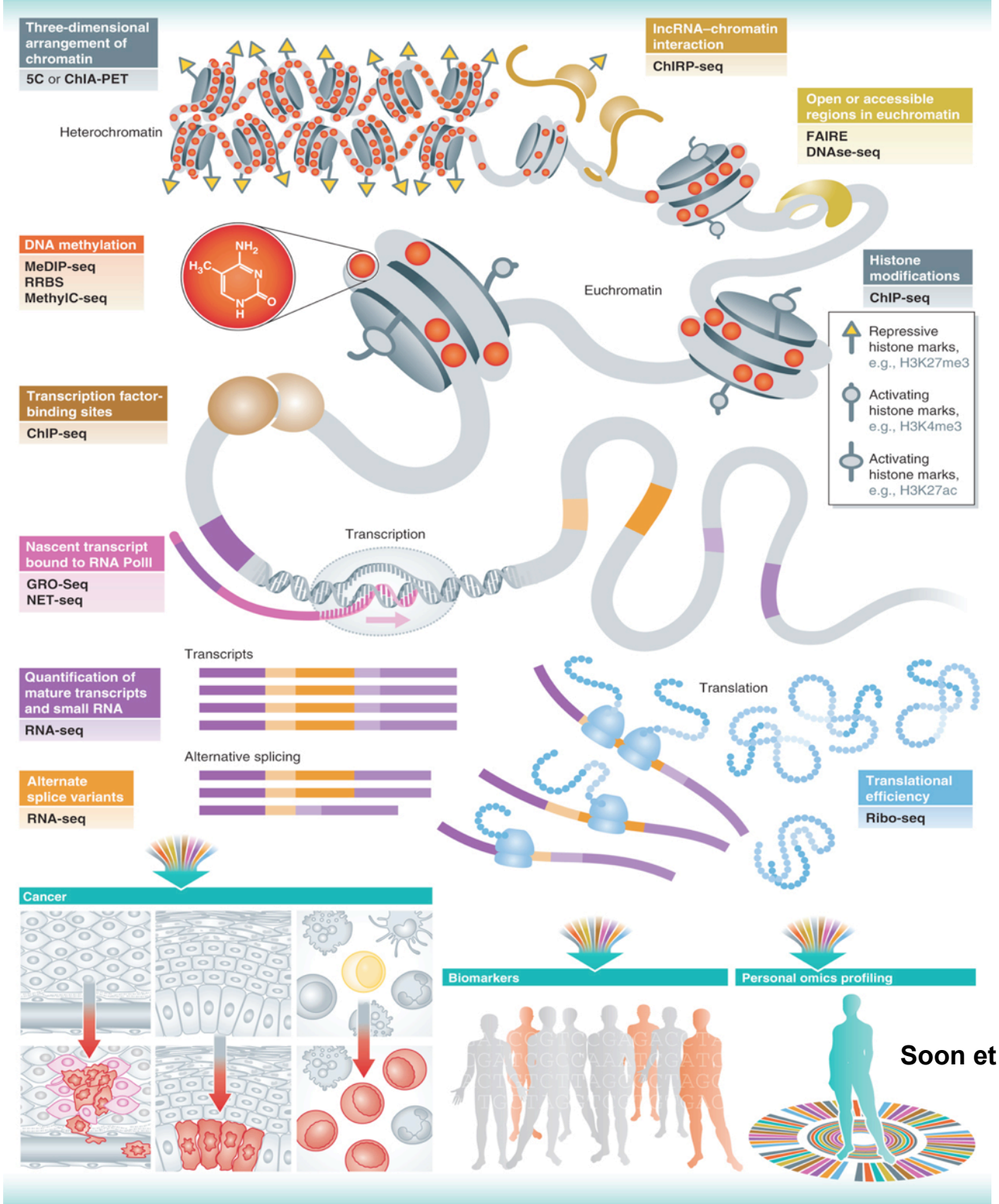


...

SHAPE-seq [RNA structural biology]



~40 different **`*-seq`** assays available



Soon et al., Mol. Syst. Biol. 2013

[2] Regulation of gene expression

Regulation of gene expression

How is gene expression regulated?

1. Promoters
2. Enhancers/silencers
3. Methylation of DNA
4. Histone modifications
5. mRNA degradation
6. RNAi
7. codon bias

…and more

} Today

Regulation of gene expression

How is gene expression regulated?

1. Promoters
2. Enhancers/silencers

} Protein factors binding to genomic DNA regions

3. Methylation of DNA
4. Histone modifications

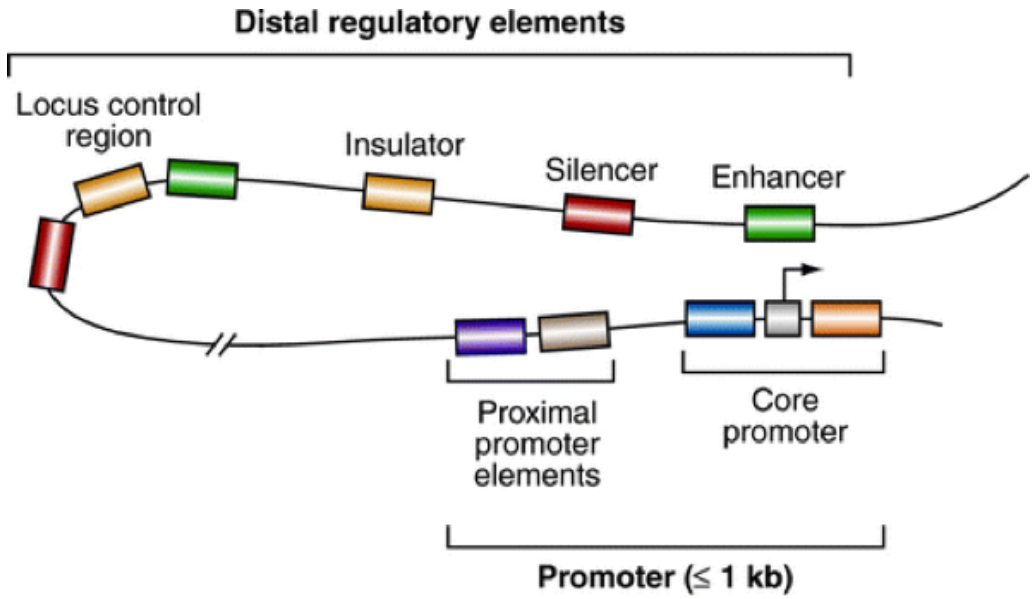
} Epigenetic modifications

5. mRNA degradation
6. RNAi
7. codon bias

…and more

Regulation of gene expression: protein factors

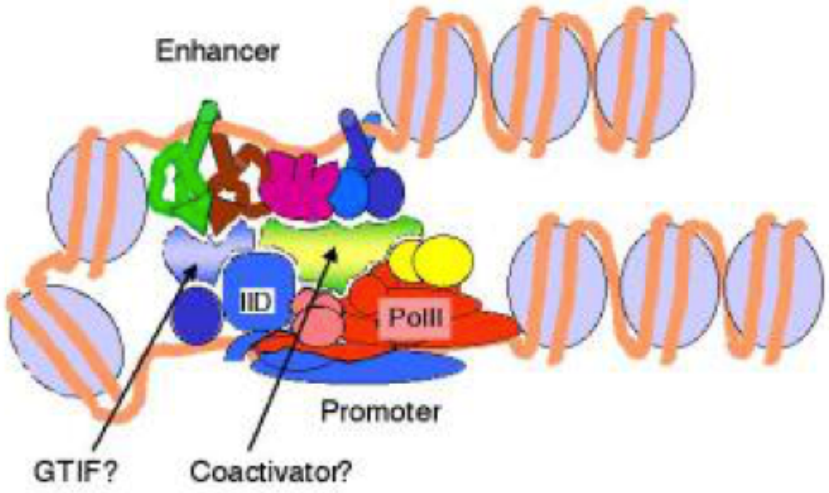
Binding of transcription factors (TFs) to promoters and enhancers/silencers



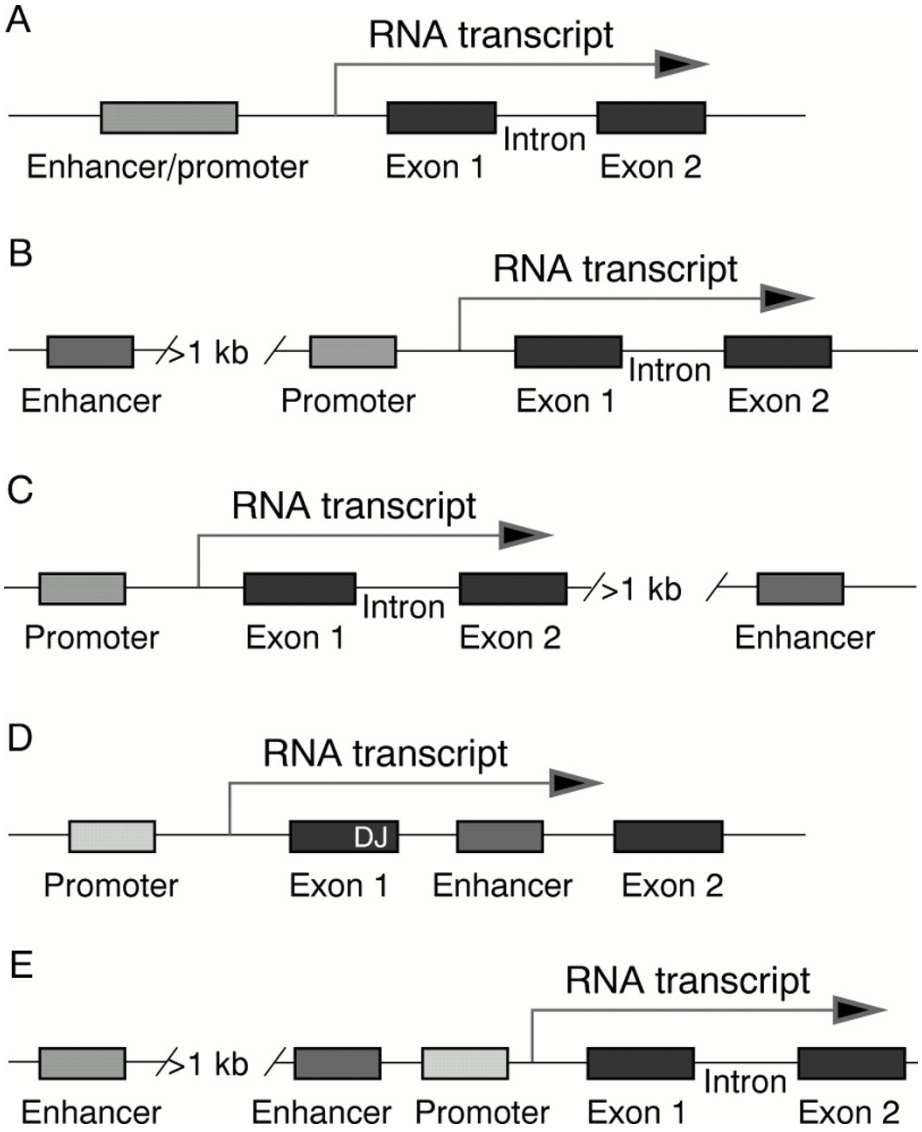
Maston GA, et al. 2006.
Annu. Rev. Genomics Hum. Genet. 7:29–59

Regulation of gene expression: protein factors

Binding of transcription factors (TFs) to promoters and enhancers/silencers



Ross Hardison, PSU



Macfarlane, Mol Path., 2000

Epigenome

DNA methylation

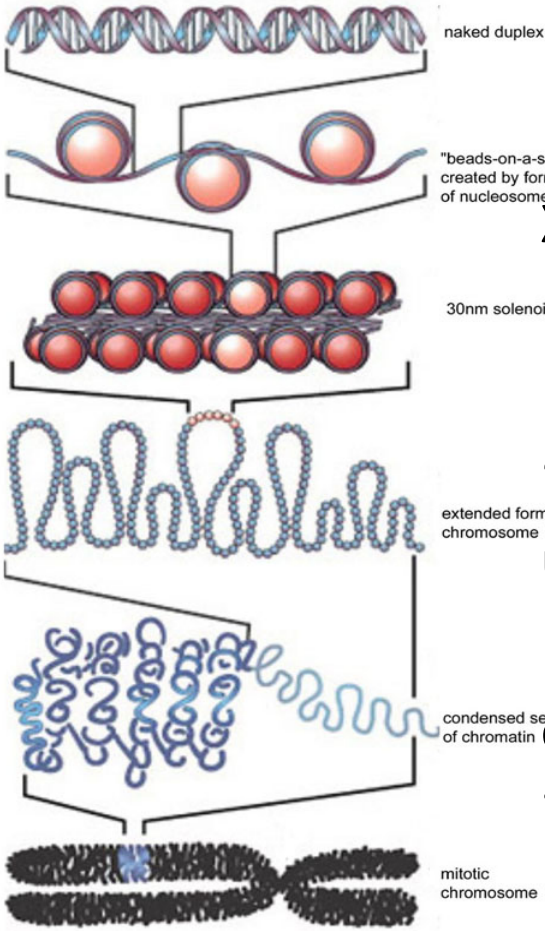
represses gene expression

Histone modifications

repress or enable gene expression

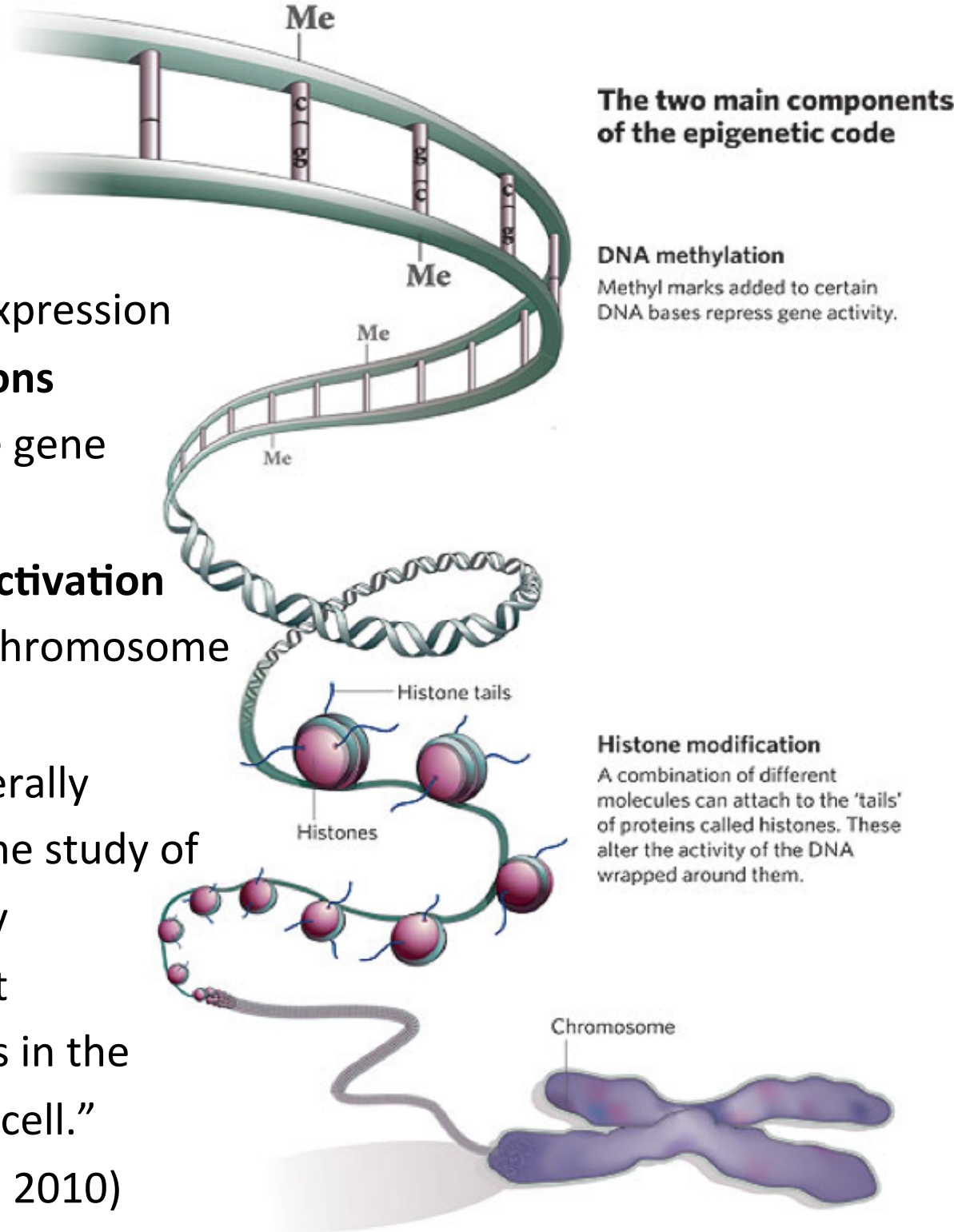
X-chromosome inactivation

inactivates an X chromosome



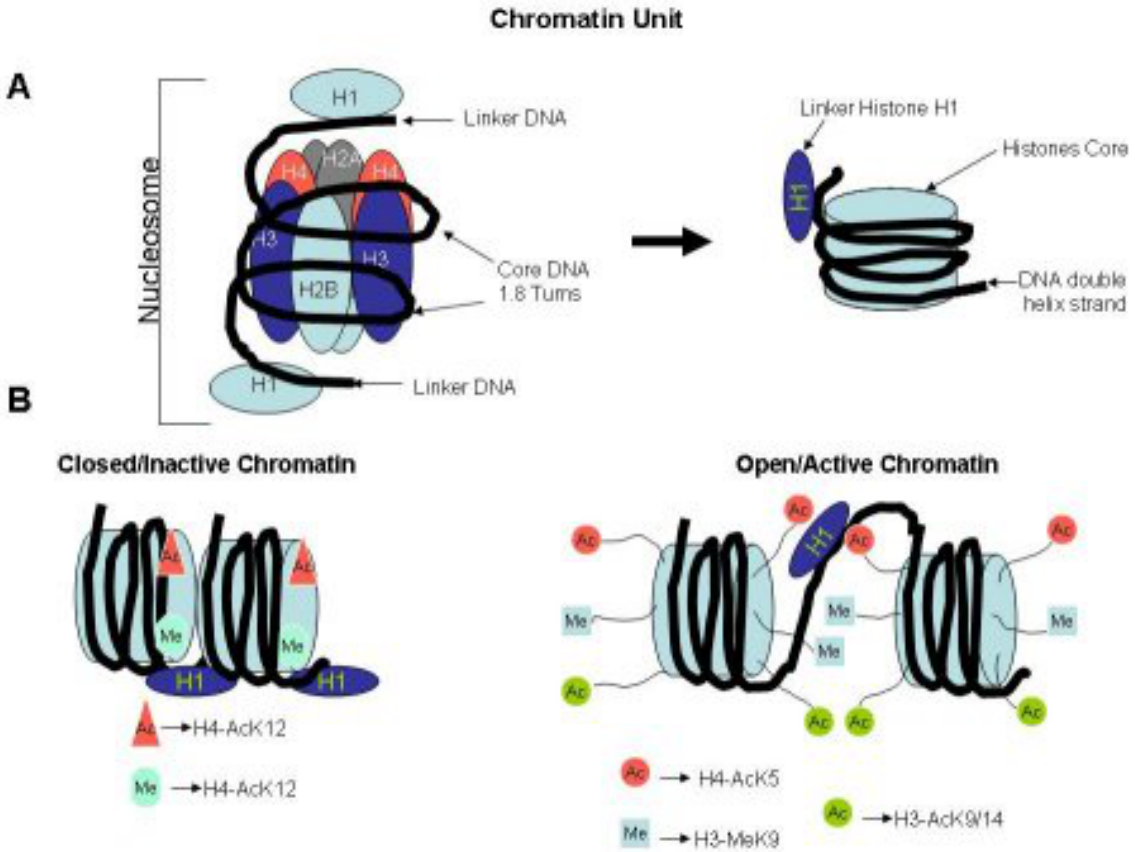
"Epigenetics is generally understood to be the study of heritable regulatory changes that do not involve any changes in the DNA sequence of a cell."

(Huss, Brief. Bioinf., 2010)



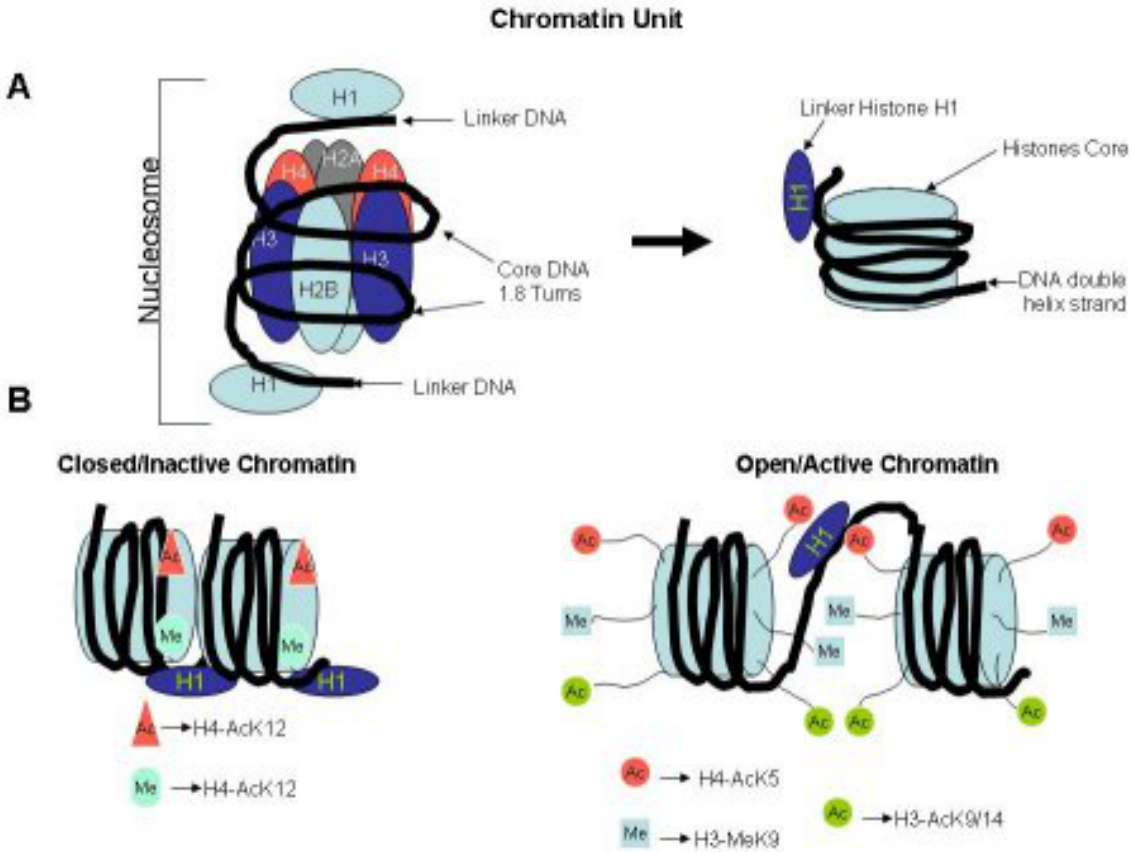
Chromatin

Chromatin: the complex of genomic DNA and its associated protein factors
Nucleosome: the basic unit of chromatin, DNA wrapped around core histone proteins
Core histones: protein complexes of 2x4 subunits (H2A, H2B, H3, H4) around which DNA (146 bases) is wrapped.
Linker histone: H1



Chromatin

Chromatin: the complex of genomic DNA and its associated protein factors
Nucleosome: the basic unit of chromatin, DNA wrapped around core histone proteins
Core histones: protein complexes of 2x4 subunits (H2A, H2B, H3, H4) around which DNA (146 bases) is wrapped.
Linker histone: H1



⇒ Acetylation and methylation of the tails of histone proteins are markers of <u>chromatin state:</u> *open* or *closed*
⇒ "Open" conformation exposes the DNA to the transcription machinery of the cell; thus, this *enables transcription.*
⇒ Chromatin structure conformation is primarily regulated by proteins through acetylation and methylation of the histones

Task: find the regulatory regions in genomic DNA

For a given organism-tissue-developmental stage-condition:

1. core promoter occupancy: what genes have an **RNA Pol II** attached
2. proximal promoter occupancy: what **transcription factors** bind to the promoter regions of the genes
3. enhancer/silencer: what **protein factors** bind to these regions
4. DNA methylation: what bases are **methylated**
   => gene repression
5. histone modifications: how are the histone tails modified (**acetylated**/**methylated**)
   => open/closed chromatin
6. DNAse hypersensitive sites: regions exposed to DNAse degradation
7. FAIRE - formaldehyde-assisted isolation of regulatory elements

We want to know what genomic DNA regions are associated with these factors/ modifications

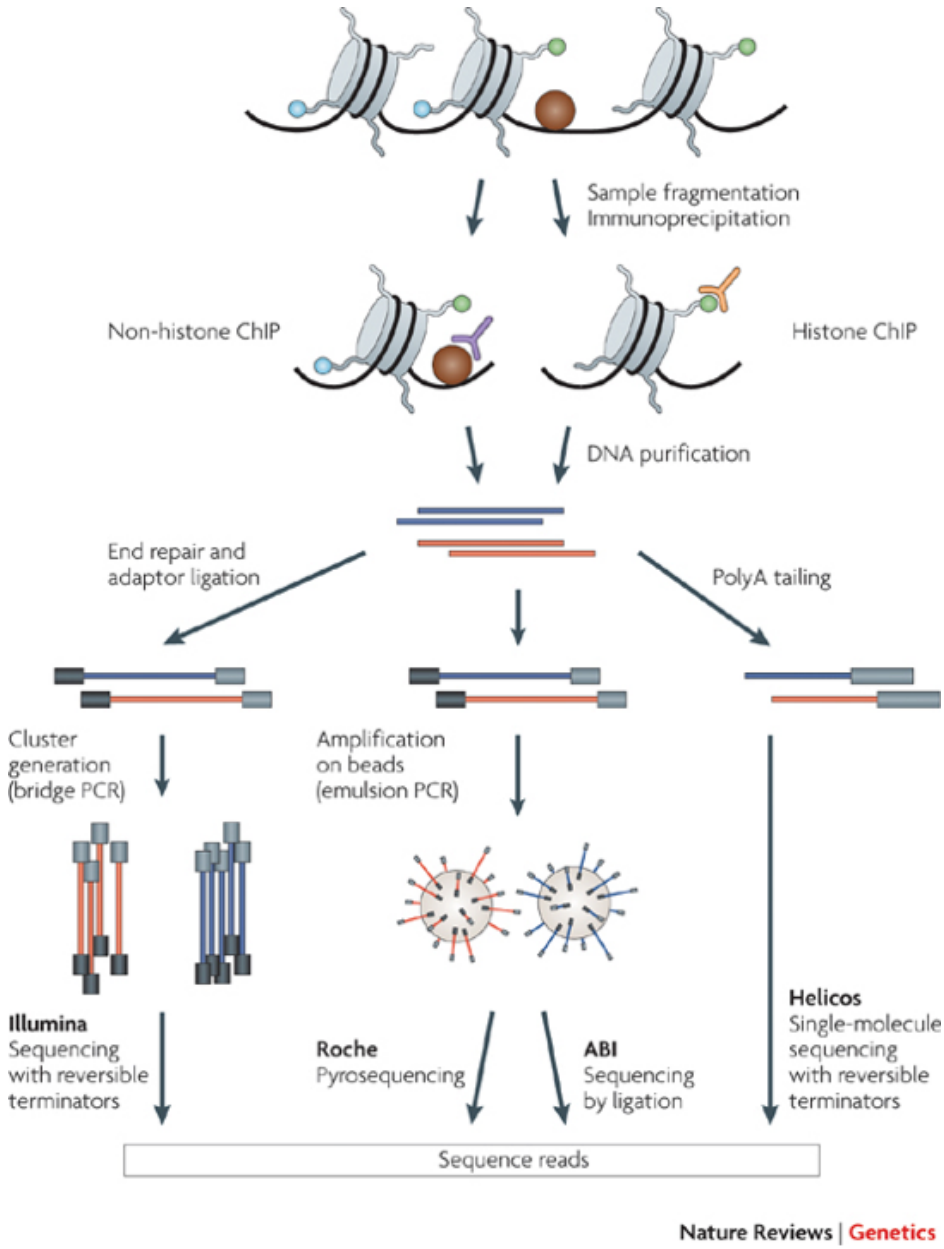
To do this (points 1-5): Chromatin immunoprecipitation, **ChIP**.

[3] ChIP-seq, experimental procedure

Chromatin ImmunoPrecipitation (ChIP) – sequencing

1. Crosslink proteins bound to DNA
2. Extract the DNA (including the proteins now tightly bound to it)
3. Fragment the DNA
4. Immunoprecipitation:
   Use antibody against the protein of interest
   => pull out only the DNA fragments to which the protein of interest is bound
5. Reverse the crosslinks
6. Extract the DNA (now without any proteins bound to it)
7. Prepare a sequencing library
8. Sequence
9. The reads come from the DNA that was pulled out with the protein.

[The protein is typically a transcription factor]



Park, Nat. Rev. Genet., 2009

ChIP-seq can be used to assess:

A. Transcription factor binding

B. Methylation of cytosines in DNA

C. Histone modifications

A. Transcription factor binding

Occurs at any promoter/enhancer/silencer region.
Use antibody against the transcription factor you'd like to assay.
There are antibodies for many, but not all, transcription factors

B. Methylation of cytosines

Occurs at CpG dinucleotides.
To capture methylation status, use antibody against 5-methyl-cytosine.
Other possibility: bisulphite treatment of DNA – unmethylated C is changed into U, while methylated C is unchanged. This can then be assessed using regular DNA sequencing (no antibodies involved).

C. Histone modifications

Occurs at the tails of the histone core proteins
Some histone modifications are markers of open chromatin (activation), some of closed chromatin (repression)
Use antibody against the modification you want to investigate.

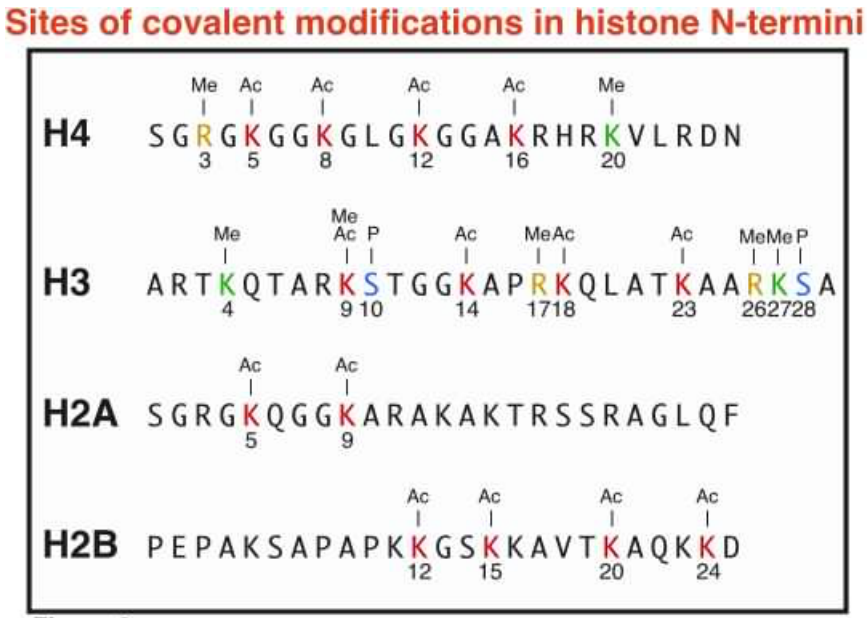


Figure 3

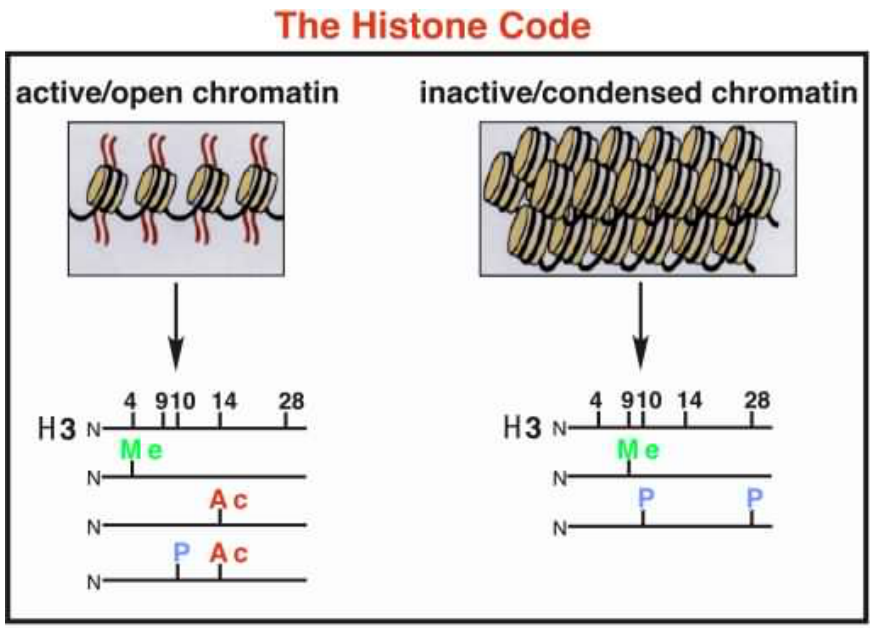


Figure 4

Figures from: Uta-Maria Bauer, U. Marburg

C. Histone modifications

Type of modification	Histone						
	H3K4	H3K9	H3K14	H3K27	H3K79	H4K20	H2BK5
mono-methylation	activation[11]	activation[12]		activation[12]	activation[12][13]	activation[12]	activation[12]
di-methylation		repression[14]		repression[14]	activation[13]		
tri-methylation	activation[15]	repression[12]		repression[12]	activation,[13] repression[12]		repression[14]
acetylation		activation[15]	activation[15]				

From http://en.wikipedia.org/wiki/Histone

[4] ChIP-seq, bioinformatics

ChIP-seq bioinformatics pipeline

Starting material: a set of sequence reads, originating from the regions you extracted with the antibody. (Typically, single-end reads are used).

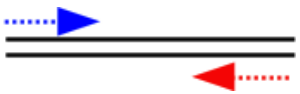
1. Map the reads to the reference, use your favorite aligner – bwa, bowtie, ...
2. Get your mapped reads into .bed format (or similar, depending on what program is used in the next step)
3. Apply an algorithm that finds clusters of reads – these are called **peaks**, and the software is often called a **peak finder**
4. Assign p-values to each cluster (peak)
5. If possible from experimental setup and the software: estimate FDR => a list of regions with *P*-values and possibly also FDRs
6. Further analyses, e.g.
   - look for presence of the TF binding site motif within or near the peaks
   - look for *any* overrepresented motif within or near the peaks
   - correlate the peaks with other genomic features; TSSs, exon/intron boundaries, other TF binding profiles, methylation status, etc.

Detecting peaks – clusters of reads

Find regions where many reads map.

These enriched regions are called **peaks**.

Note: The DNA fragments (200-300bp) are sequenced from both ends



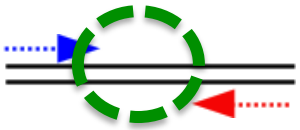
=> strand-specific pattern of mapped reads on the genomic DNA

Detecting peaks – clusters of reads

Find regions where many reads map.

These enriched regions are called **peaks**.

Note: The DNA fragments (200-300bp) are sequenced from both ends



=> strand-specific pattern of mapped reads on the genomic DNA

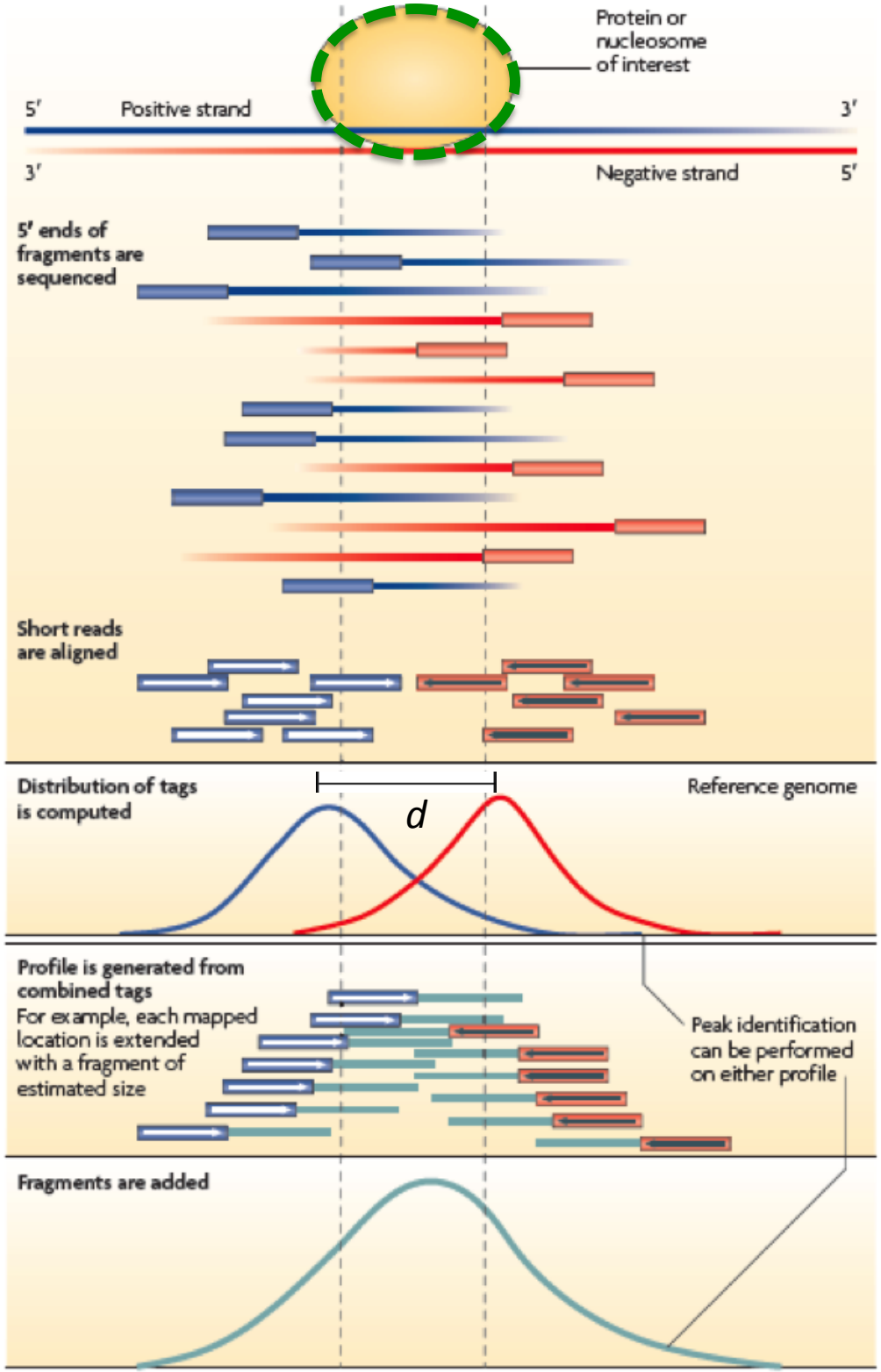
Detecting peaks

Scan the genomic DNA and look for enriched regions using a *window approach*.

Strand-specific patterns emerge and are used to locate the peaks

(1) **extend** the reads to the estimated fragment length, see two bottom panels at right.

(2) **shift** reads towards the middle of the two peaks; d/2



Park, Nat. Rev. Genet., 2009

Detecting peaks – what peaks are significant

When is a read enrichment also statistically significant?
Compare the read count with a background distribution

- Poisson distribution (e.g. MACS, HOMER)
- Binomial distribution (e.g. PeakSeq, CisGenome)

=> output is, for each peak, a *P*-value describing the probability that the read enrichment at this peak is due to chance.

Exactly what background distribution to compare with?

1. read distribution in **ChIP-sample** DNA

Detecting peaks – what peaks are significant

When is a read enrichment also statistically significant?
Compare the read count with a background distribution

- Poisson distribution (e.g. MACS, HOMER)
- Binomial distribution (e.g. PeakSeq, CisGenome)

=> output is, for each peak, a *P*-value describing the probability that the read enrichment at this peak is due to chance.

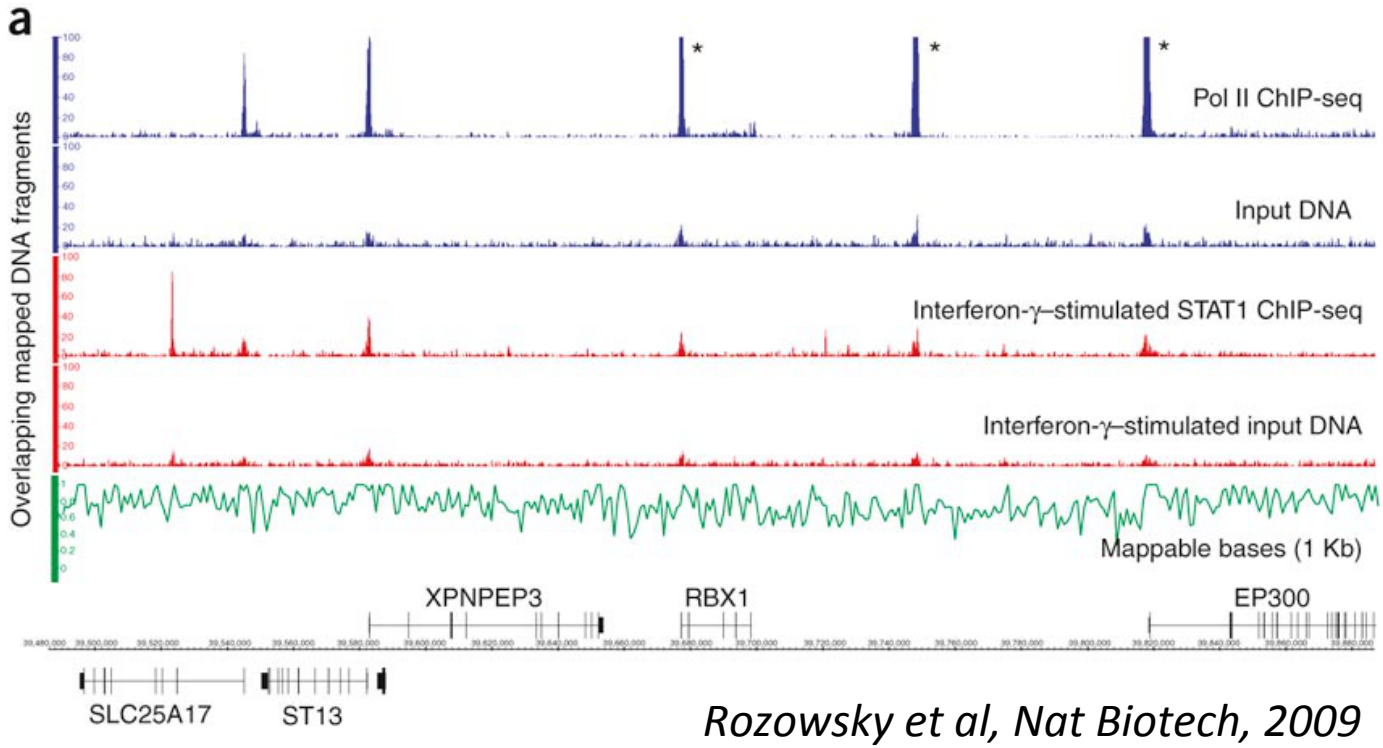
Exactly what background distribution to compare with?

1. read distribution in **ChIP-sample** DNA

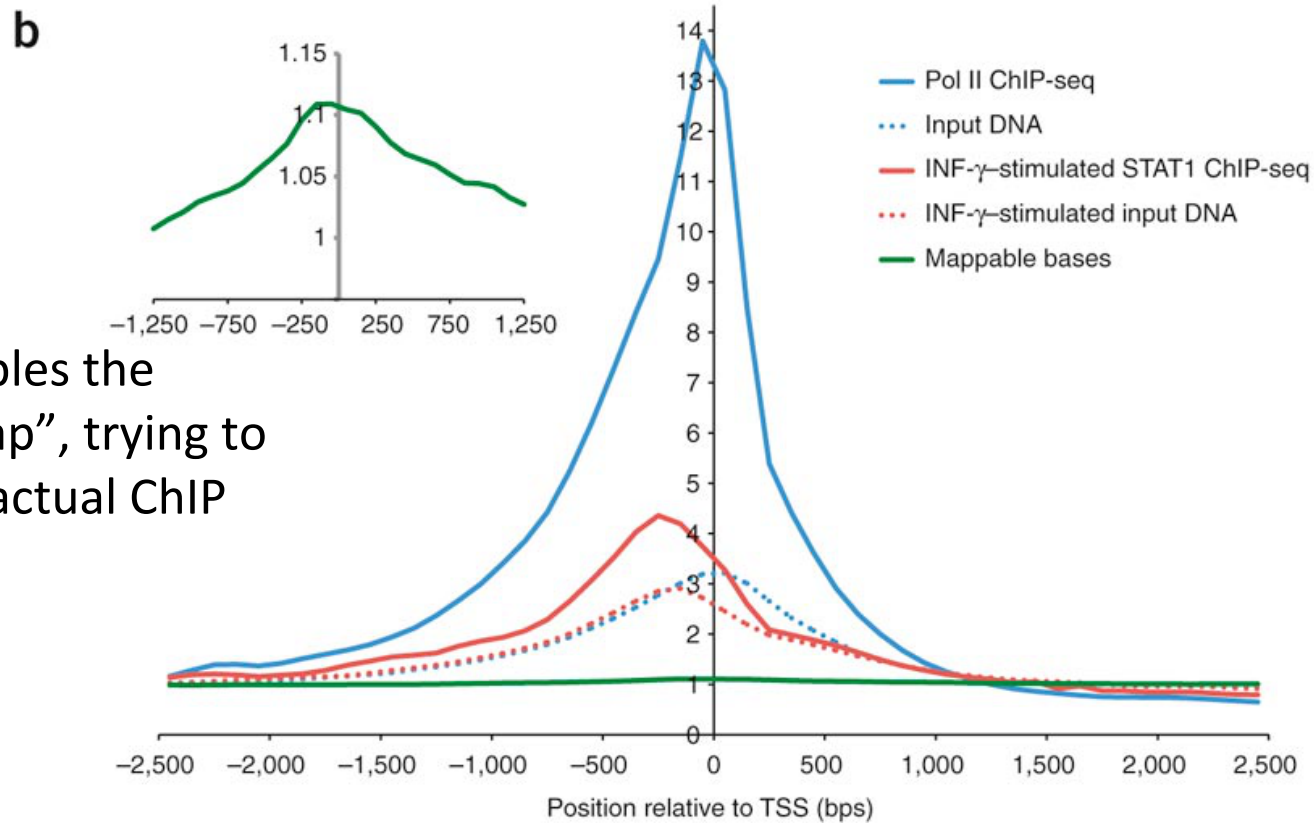
2. read distribution in **control sample** DNA

Control sample

1. *input DNA*: genomic DNA removed before IP
2. *mock IP DNA*: DNA precipitated without antibodies
3. *nonspecific IP DNA*: DNA precipitated with antibody (e.g. IgG) known to not associate with any DNA binding protein



Rozowsky et al, Nat Biotech, 2009

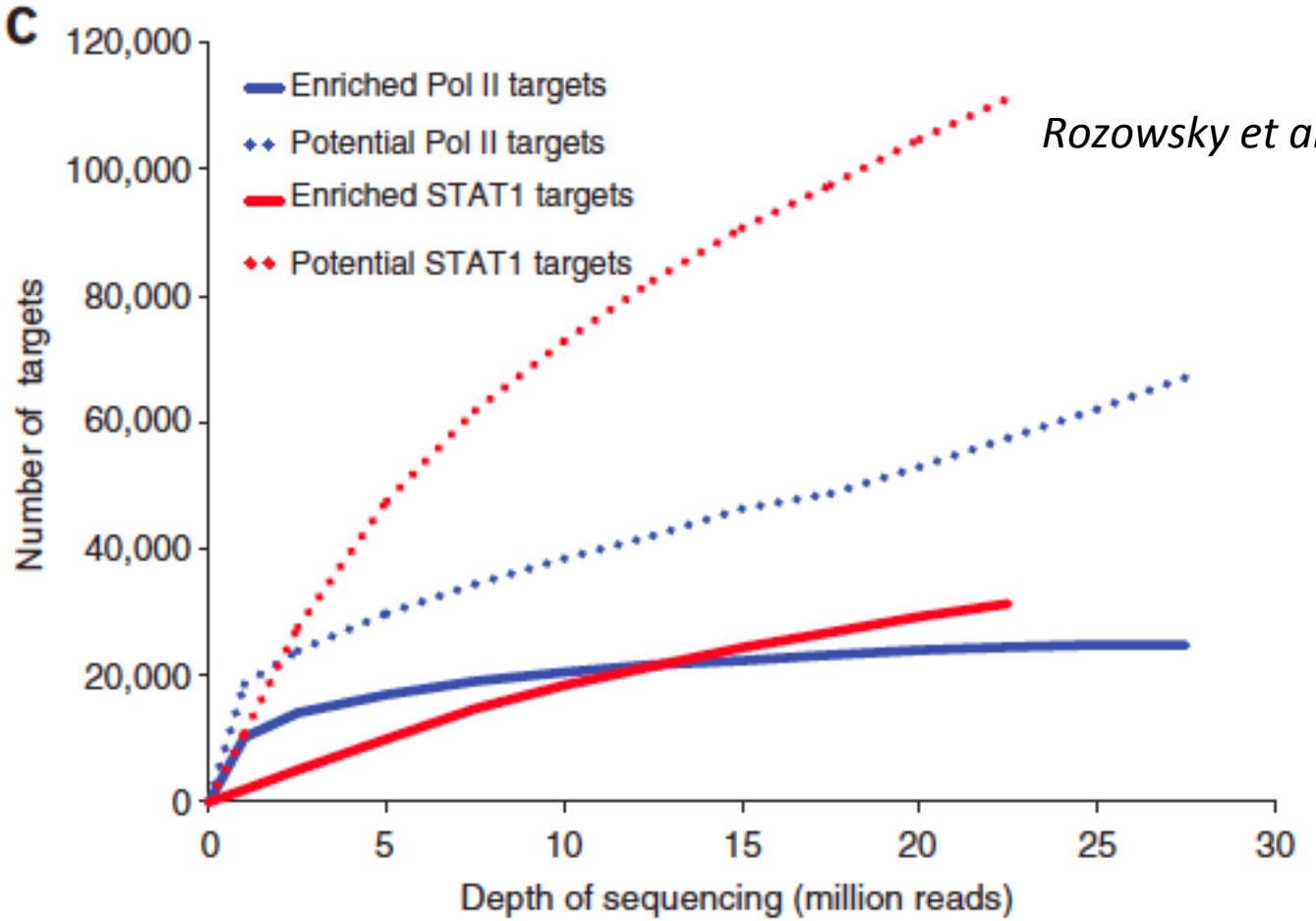


Having a control sample also enables the estimation of FDR by "sample swap", trying to find peaks in input DNA with the actual ChIP sample as control sample.

Input DNA most commonly used.

Control sample

Dashed lines: potential peaks *without* using control sample
Solid lines: peaks actually called (enriched) *using control sample* (in this case: input DNA)



Rozowsky et al, Nat Biotech, 2009

Narrow and broad peaks

<u>Point source:</u>

TF binding site peaks are **narrow**, e.g.

CTCF

<u>Mixed source:</u>

RNA pol II

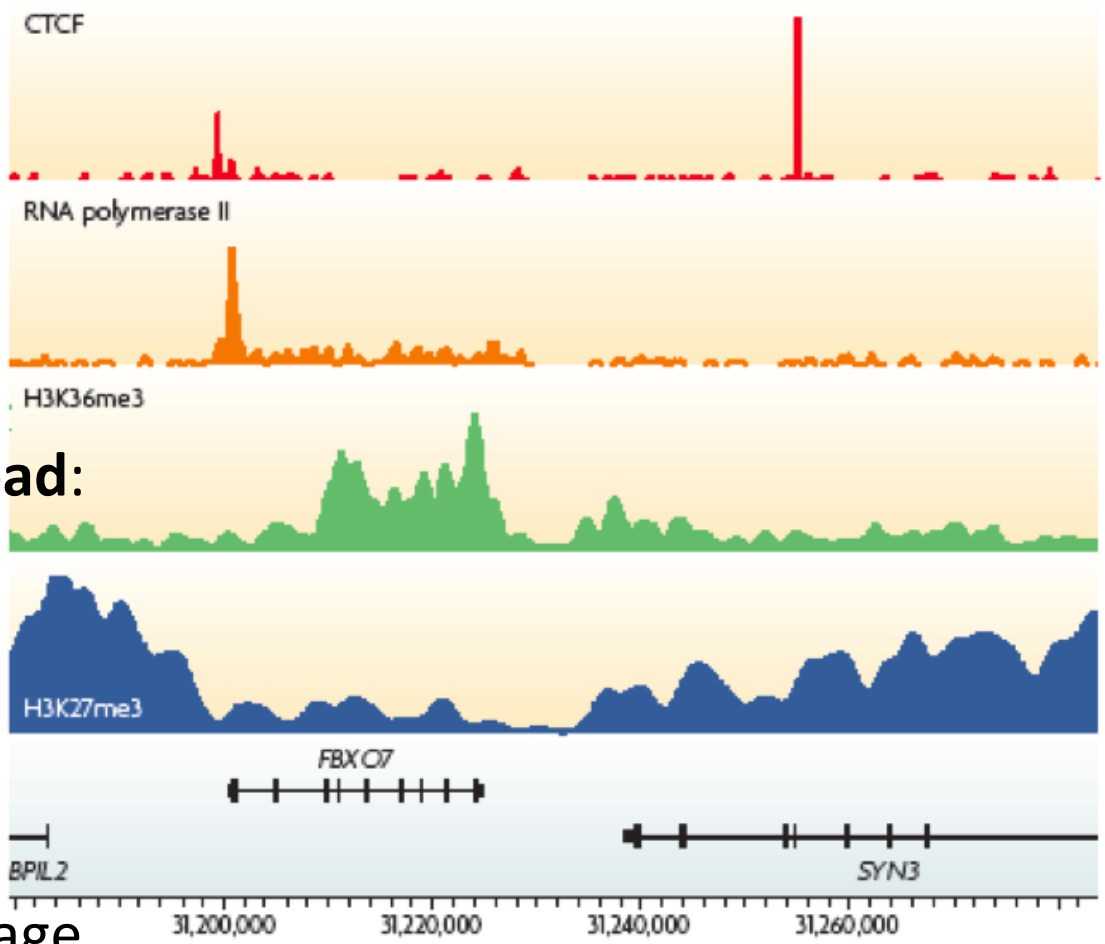
<u>Broad source:</u>

[a] Histone modification peaks are **broad**:

E.g.: H3K36me3 (associated with transcription elongation) and H3K27me3 (associated with gene silencing) For such modifications:

- distinct peaks lacking
- assure that sequencing coverage is enough to provide a high, continuous coverage of the entire region with modified histones

[b] DNAse-seq and [c] FAIRE-seq

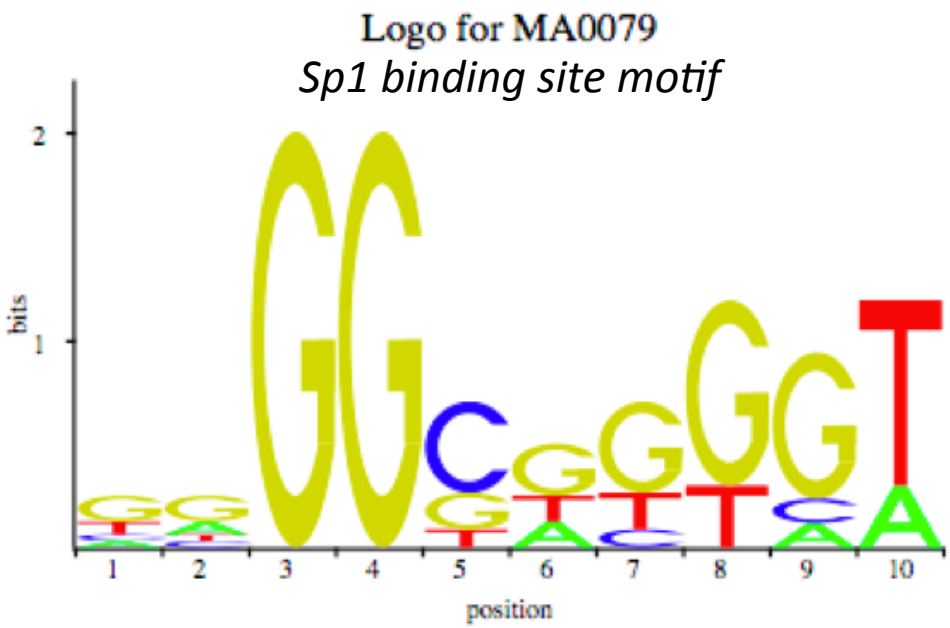


Park 2009 (data from Barski et al 2007)

Validation of results

Informatics: look for enrichment of TF binding motifs in or near the set of regions

E.g., Sp1 binding site motif:
If Sp1 was the TF you wanted to pull out in the ChIP reaction, hopefully its binding motif is present in most of the peak regions (see motif to the right)



Wet lab: quantitative PCR, i.e., go back to the sample and verify that the DNA-sequences your peak finding program picked up actually are there.

ChIP-seq considerations

The antibody is crucial:
- cases of bad sensitivity (low yield) or bad specificity (cross-reaction)
- quality may even differ between different batches of presumably identical antibodies

Sequencing errors and GC bias
- just like in any other MPS setup

Reads mapping to >1 genomic region (multireads)
- handled by the aligner

Many reads mapping to the exact same region
- PCR artefact?
- on the other hand, if sequencing depth is large enough, it might result from >1 identical fragment in the sample
- handled (in some cases) by the peak finder

Software available (a selection thereof)

MACS, http://liulab.dfci.harvard.edu/MACS/
SISSRs, http://sissrs.rajajothi.com/
FindPeaks, http://vancouvershortr.sourceforge.net/
PeakSeq, http://info.gersteinlab.org/PeakSeq
SICER, http://home.gwu.edu/~wpeng/Software.htm
CisGenome, http://www.biostat.jhsph.edu/~hji/cisgenome/
QuEST, http://mendel.stanford.edu/SidowLab/downloads/quest/
HOMER, http://homer.salk.edu/homer/ngs/index.html

[5] Summary

Concluding task

Write down your reflections from the **RNA-seq and/or ChIP-seq** lecture on:

1. Something that you found interesting and/or fun.
2. Something that you found hard to grasp.
3. Something that you think these lectures should cover better (either something that wasn't covered at all, or something that you'd like to be covered in more detail).

Format: one-two sentence(s) per question.

Time: 5 minutes.

Hand in your paper to me when you leave the room today.

Please write your name on it!