

Analysis of data from high-throughput molecular biology experiments

Lecture 8, 2012-02-03

Gene expression regulation, brief introduction

Chromatin immuno-precipitation (ChIP)

ChIP-seq – wet lab

ChIP-seq – bioinformatics

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
- ...



Today

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

...

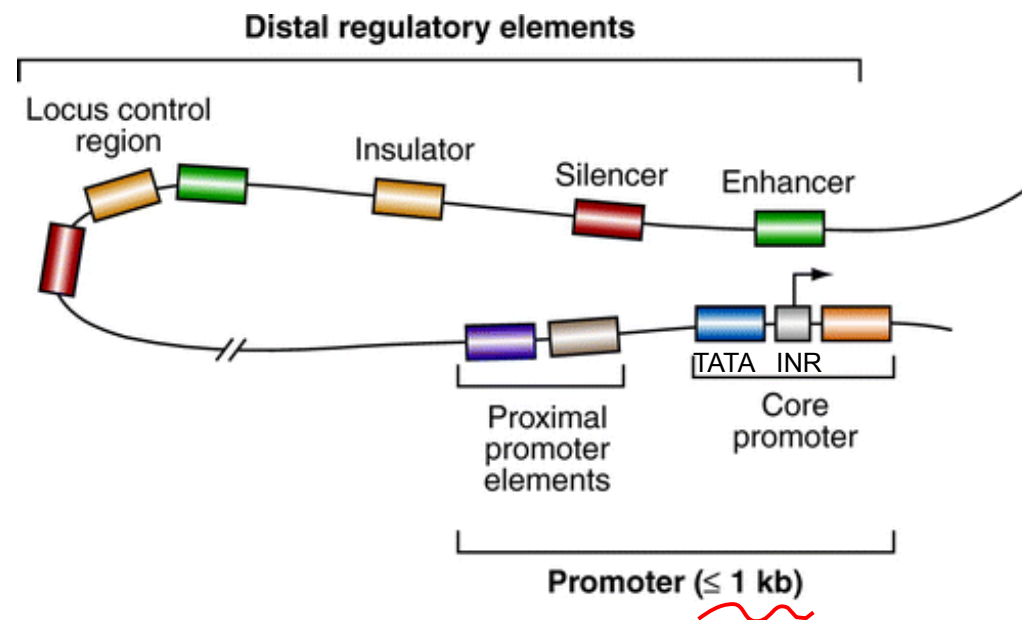



Protein factors binding to genomic DNA regions

Epigenetic modifications

Regulation of gene expression: protein factors

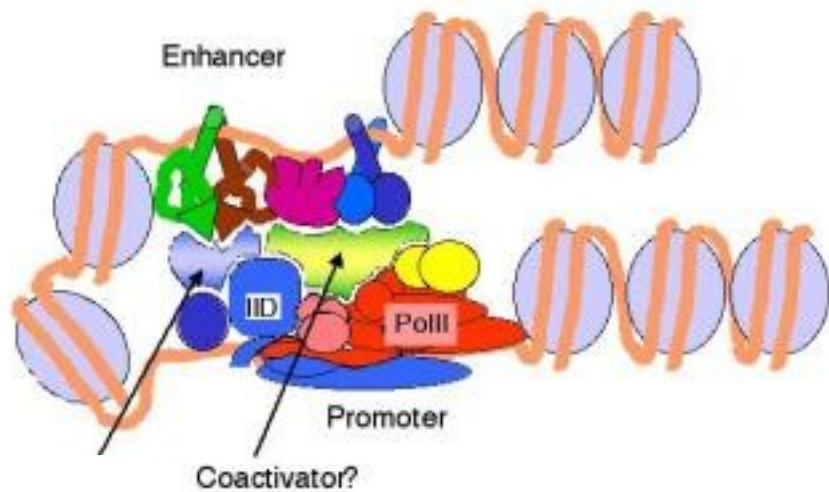
Binding of transcription factors (TFs) to promoters and enhancers/silencers



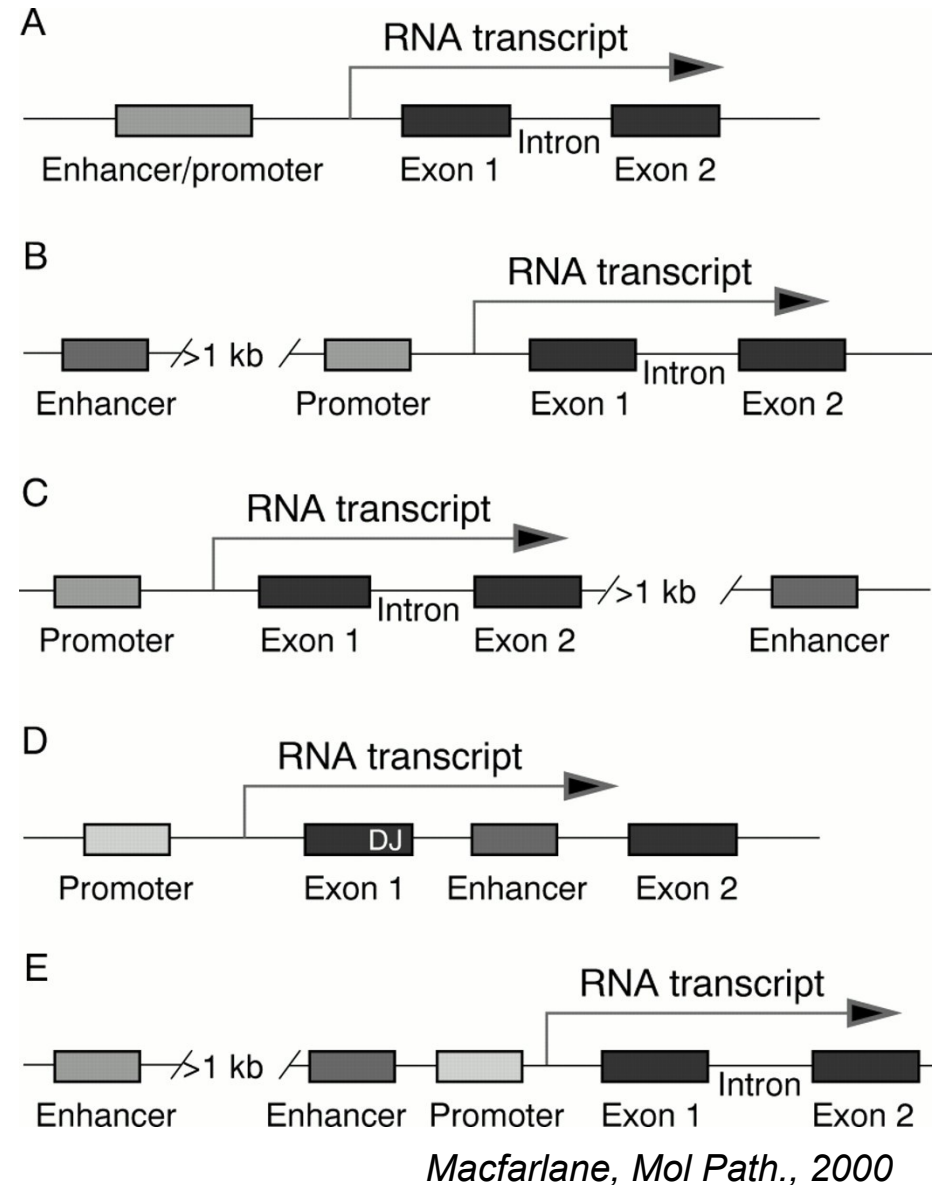
 Maston GA, et al. 2006.
Annu. Rev. Genomics Hum. Genet. 7:29–59

Regulation of gene expression: protein factors

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



Ross Hardison, PSU



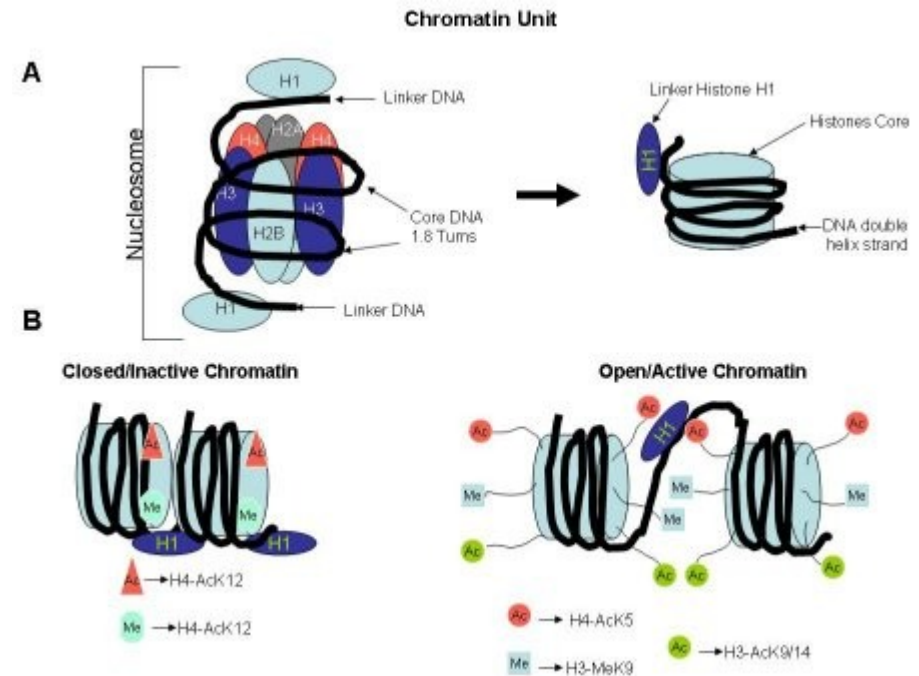
Chromatin

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



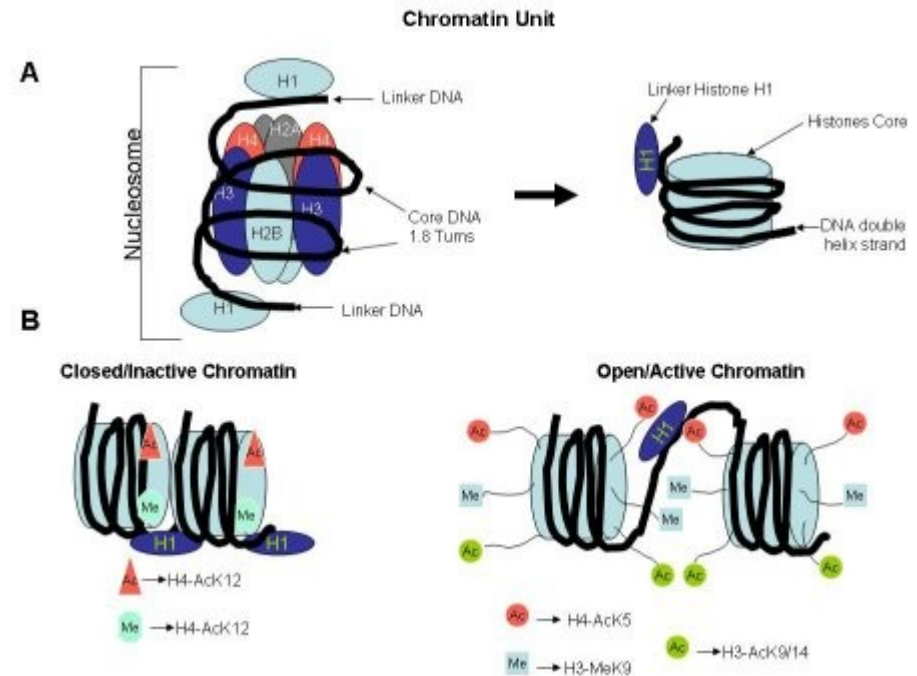
Chromatin

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 chromatin state: *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

(The epigenome

“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.”

Epigenetic phenomena: gene imprinting, X chromosome inactivation, maintenance of cell identity

Epigenetic mechanisms: modifications of histone proteins, methylation of cytosines in DNA, some RNA-mediated mechanisms.

Gradual shift in the meaning of “epigenetics”.

All this taken from Huss, Brief. Bioinformatics vol 11 p 512-523 (2010))

Task: find the regulatory regions in genomic DNA

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

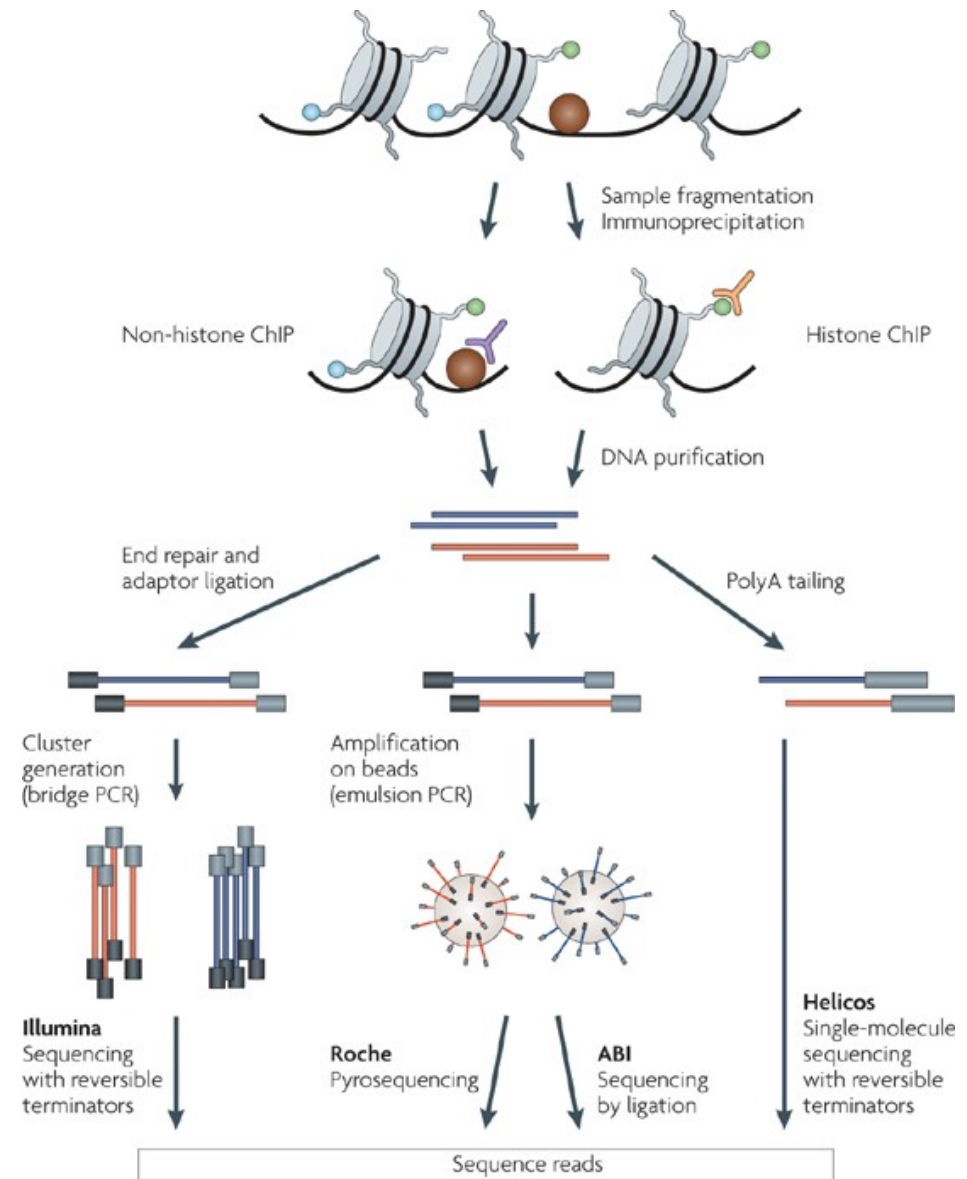
1. core promoter occupancy: what genes have an **RNA Pol II** attached
2. proximal promoter occupancy: what **transcription factors** bind to the promoter regions of the genes
3. enhancer/silencer: what **protein factors** bind to these regions
4. DNA methylation: what bases are **methylated**
=> gene 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**.

Chromatin ImmunoPrecipitation (ChIP) – sequencing

1. Crosslink any proteins bound to DNA
2. Extract the DNA (now with TFs etc tightly bound to it)
3. Fragment the DNA
4. Immunoprecipitation
Use antibody against the TF whose binding sites you wish to find
=> pull out only the DNA fragments to which the TF of interest is bound
5. Reverse the crosslinks
6. Extract the DNA
7. Prepare a sequencing library
8. Sequence
9. The reads come from the DNA that was pulled out with the TF



ChIP-seq can be used to assess:

A. Transcription factor binding

B. Methylation of cytosines

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.

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.

Sites of covalent modifications in histone N-termini

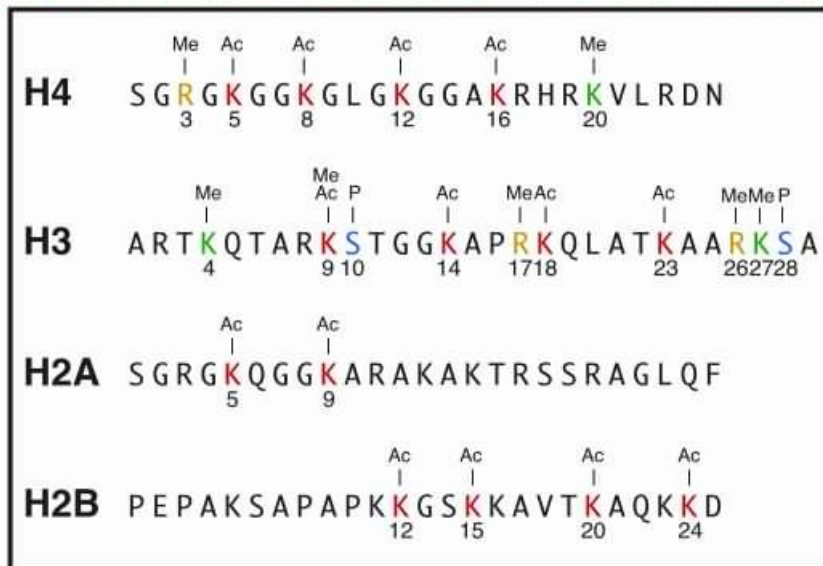


Figure 3

The Histone Code

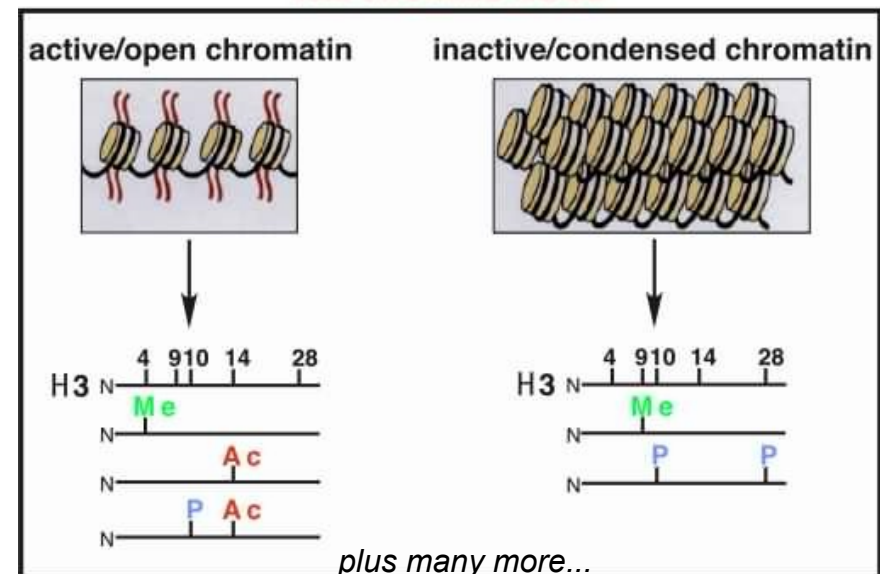


Figure 4

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>

ChIP-seq bioinformatics pipeline

Starting material: a set of sequence reads, representing the regions you extracted with the antibody.

1. Map the reads to the reference, use your favorite aligner – bwa, bowtie, maq...
2. Get your mapped reads into .bed format (or similar, depending on what program is used in the next step)
3. Apply 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 / FDR
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.

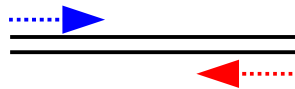
Detecting peaks – clusters of reads

Find regions where many reads map.



These enriched regions are called **peaks**.

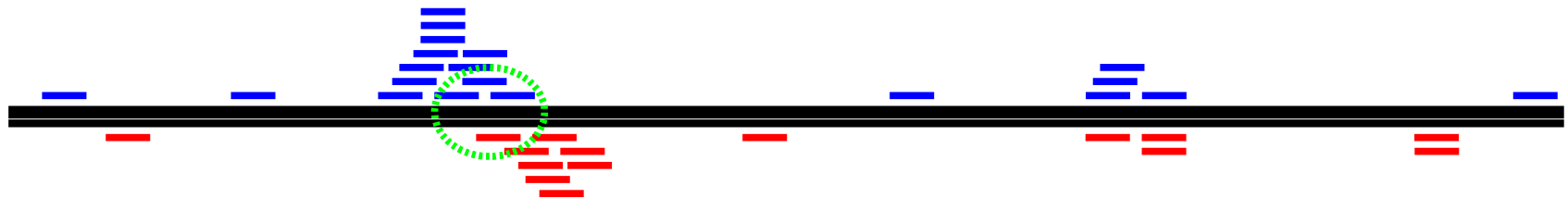
Note: The DNA fragments (200-300bp) are sequenced from both ends



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

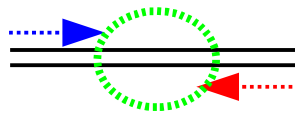
Detecting peaks – clusters of reads

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

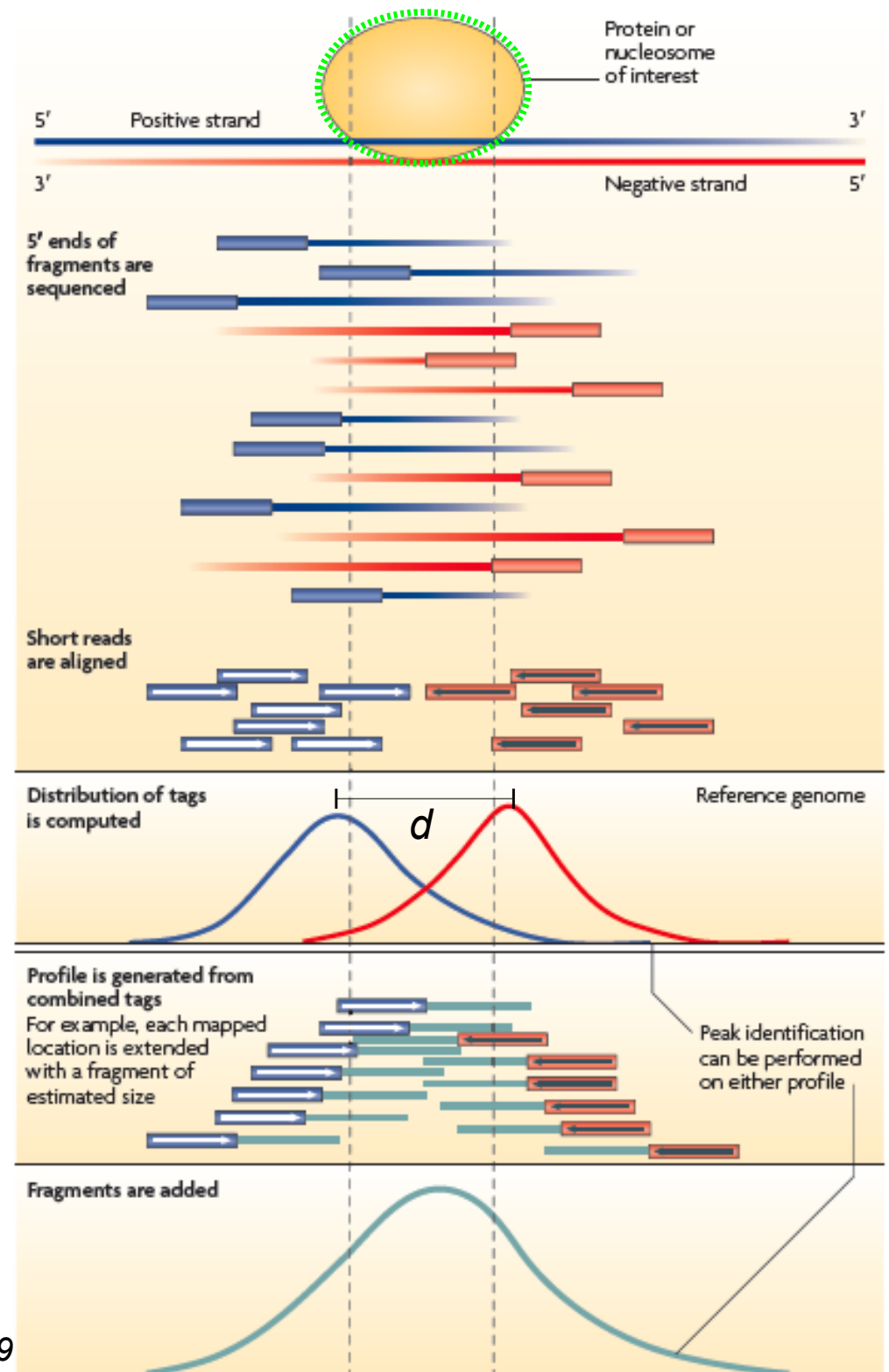
Detecting peaks

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; $d/2$



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)
- 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

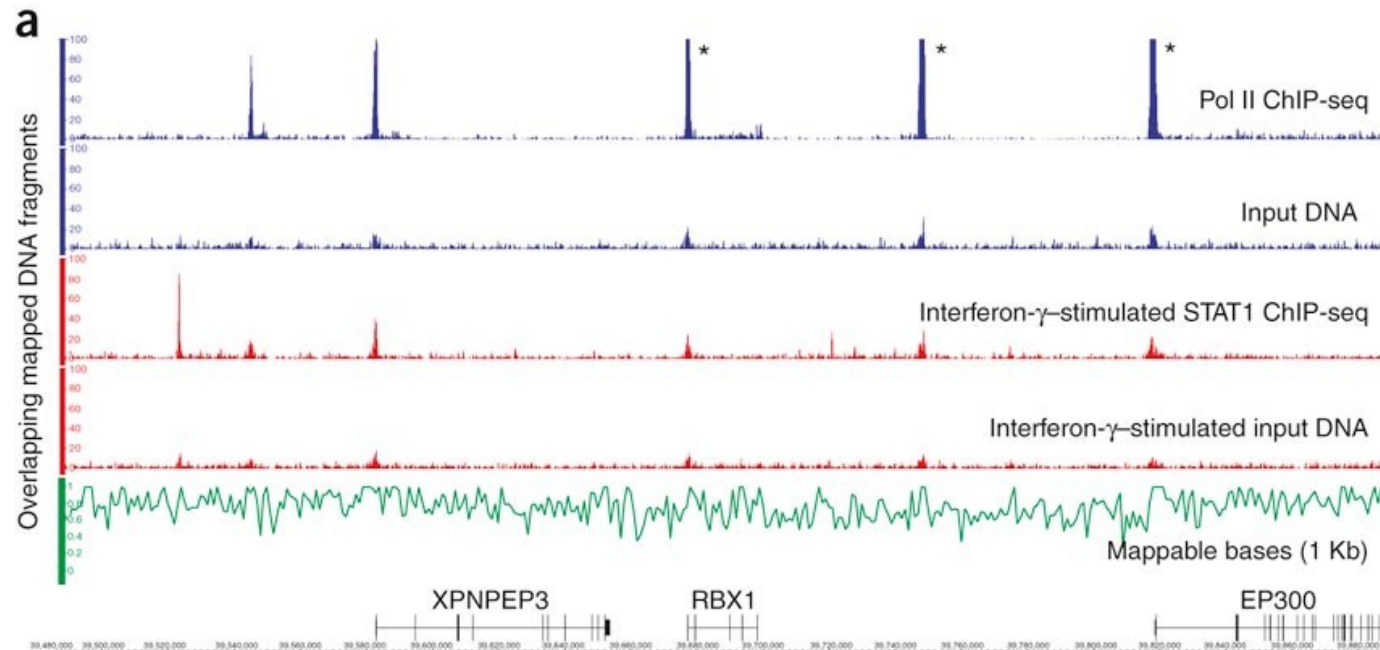
genomic DNA removed before IP

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

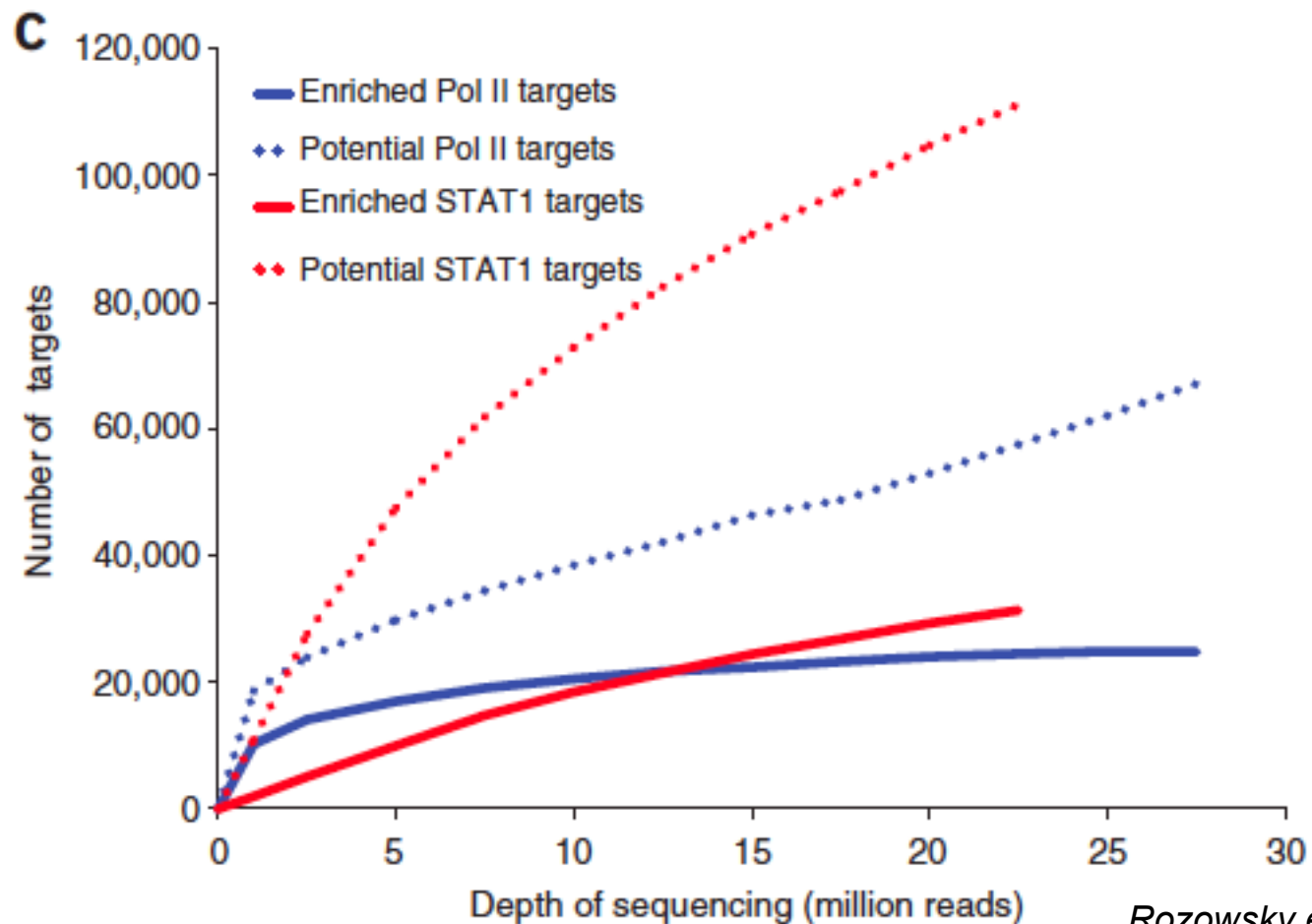
Having a control sample also enables the estimation of FDR by “sample swap”, trying to find peaks in input DNA with the actual ChIP sample as control sample.

Input DNA most commonly used.

Control sample

Dashed lines: potential peaks without using control sample (input DNA)

Solid lines: peaks actually called (enriched) using control sample (input DNA)



Example output from a peak caller (MACS)

```
# This file is generated by MACS
# ARGUMENTS LIST:
# name = /Users/chip_seq/data/2009-09-17/MACS_analysis/s_8.vs.s_7.res
# format = BED
# ChIP-seq file = /Users/chip_seq/data/2009-09-17/bwa_alignment/s_8.aln.bed
# control file = /Users/chip_seq/data/2009-09-17/bwa_alignment/s_7.aln.bed
# effective genome size = 2.70e+09
# tag size = 75
# band width = 300
# model fold = 32
# pvalue cutoff = 1.00e-05
# Ranges for calculating regional lambda are : peak_region,1000,5000,10000
# unique tags in treatment: 2181922
# total tags in treatment: 3129915
# unique tags in control: 7203468
# total tags in control: 7401694
# d = 140
chr    start    end      length  summit  tags    -10*log10(pvalue)    fold_enrichment  FDR(%)
chr1   714086   714863   778     434     28      102.75  12.20  4.61
chr1   762023   762919   897     573     25      98.43   13.10  5.13
chr1   901089   902654   1566    1072    48      83.19   8.33   7.14
chr1   911241   912056   816     239     17      55.87   8.62   25.81
chr1   948906   950059   1154    308     77      190.83  8.02   1.37
...
```


Narrow and wide peaks

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

CTCF

RNA pol II

Histone modification peaks are **wide**:

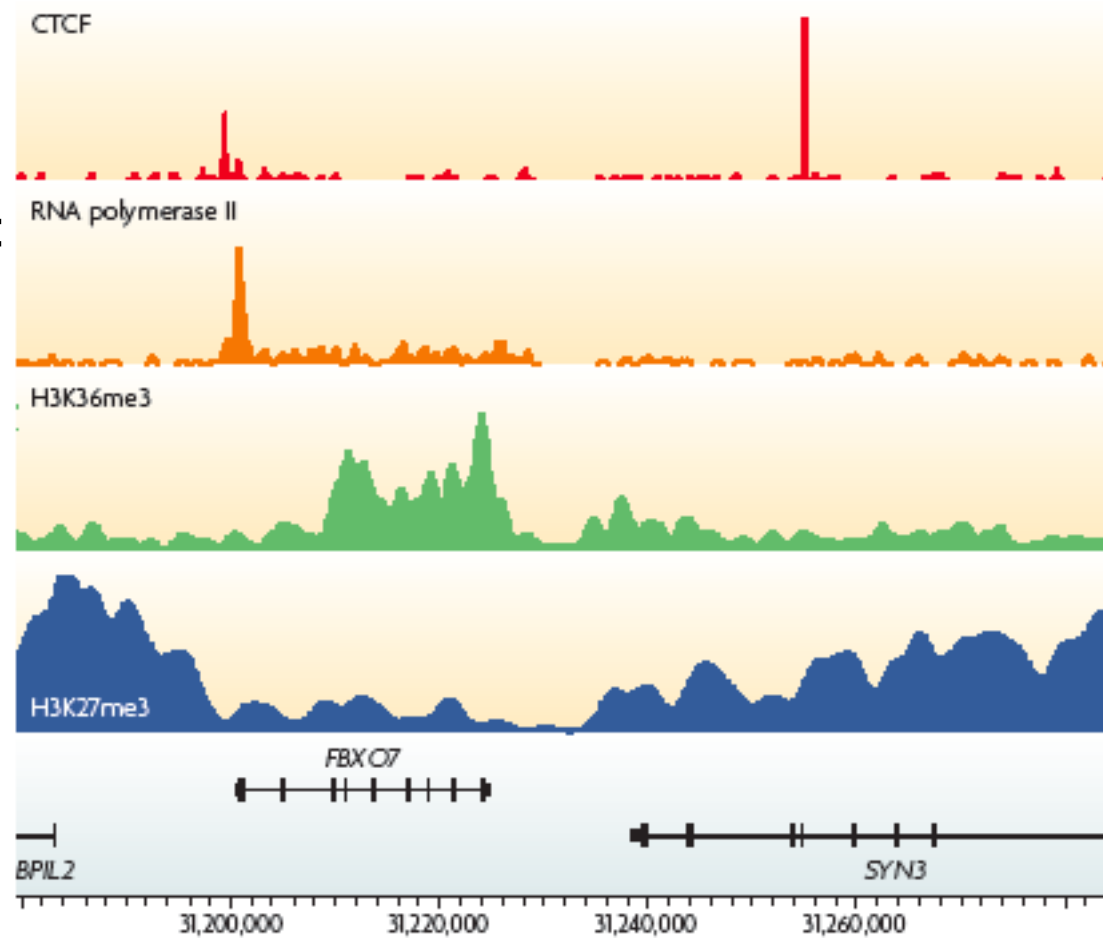
In particular for repressive histone marks, e.g.:

H3K36me3 (associated with transcription elongation)

H3K27me3 (associated with gene silencing)

For such modifications:

- distinct peaks lacking
- sequencing coverage perhaps not enough to provide a high, continuous coverage of the entire region with modified histones



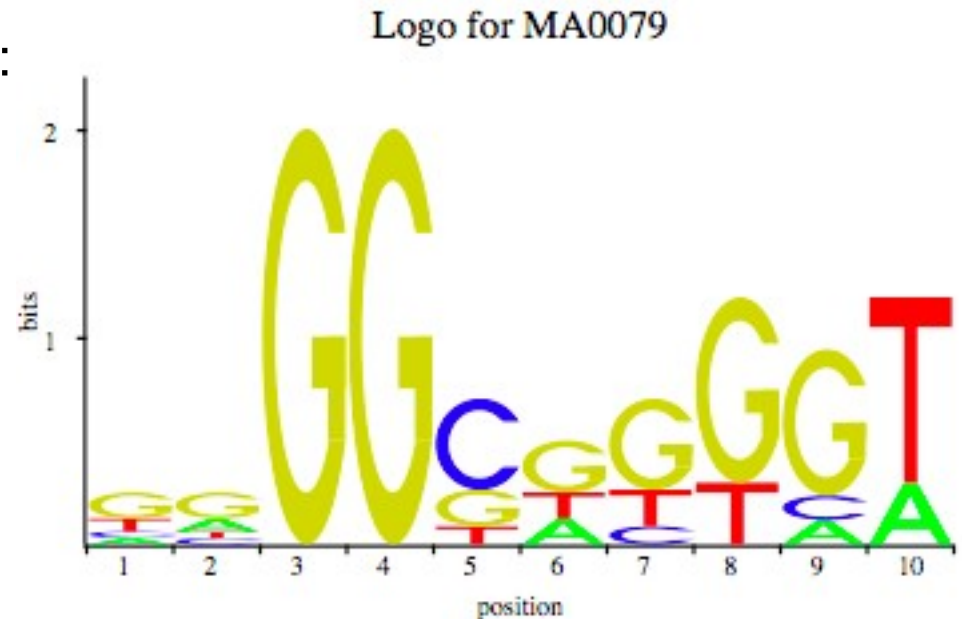
Park 2009 (data from Barski et al 2007)

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 this motif is present in most of the peak regions



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, 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/>

FindPeaks, <http://vancouvershorttr.sourceforge.net/>

PeakSeq, <http://archive.gersteinlab.org/proj/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://biowhat.ucsd.edu/homer/chipseq/index.html>

F10 Thursday 9 Feb., 13:15, 3 papers to be presented (2 ChIP-seq, 1 MaxQuant)

MACS: **Sayed Awn Muhammad, Dimitra Lappa**

Zhang et al. “Model-based analysis of ChIP-Seq (MACS)”. *Genome Biol* (2008) vol. **9** (9) pp. R137

SICER: **Athanasia Palasantza, Md Ali Shabibur Rahman**

Zang et al. “A clustering approach for identification of enriched domains from histone modification ChIP-Seq data” *Bioinformatics* (2009) vol. **25** (15) p. 1952-8

MaxQuant: **Robin Andeer, Robert Lindroos**

Cox and Mann. “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification”. *Nat Biotechnol* (2008) vol. **26** (12) pp. 1367-72

~12 minutes presentation, in pairs!

Email your presentation to me 1 hour before if you would like to use my computer for the presentation! [Acceptable formats: pdf, ppt, pptx, odp. **NOT** keynote].