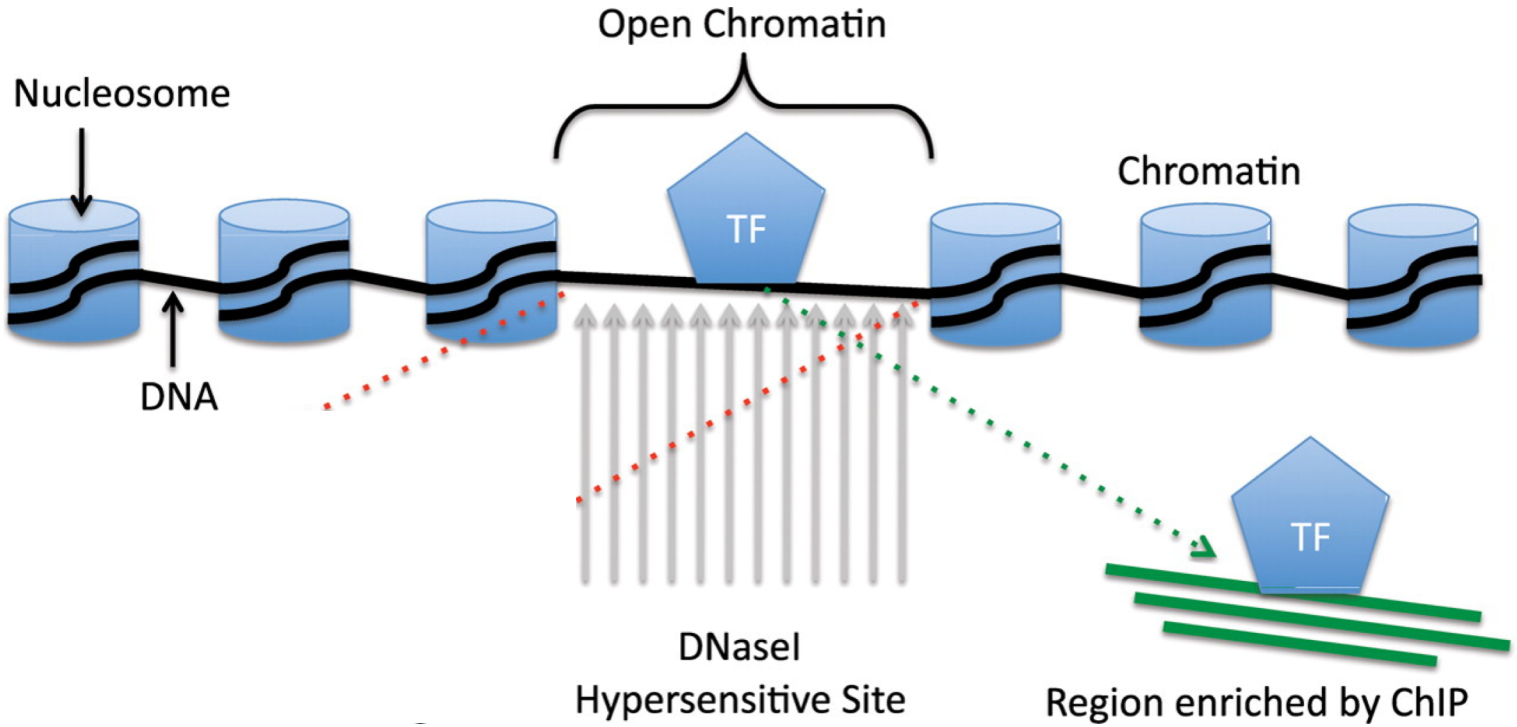
Intro to ChIP-seq

Vivek Iyer

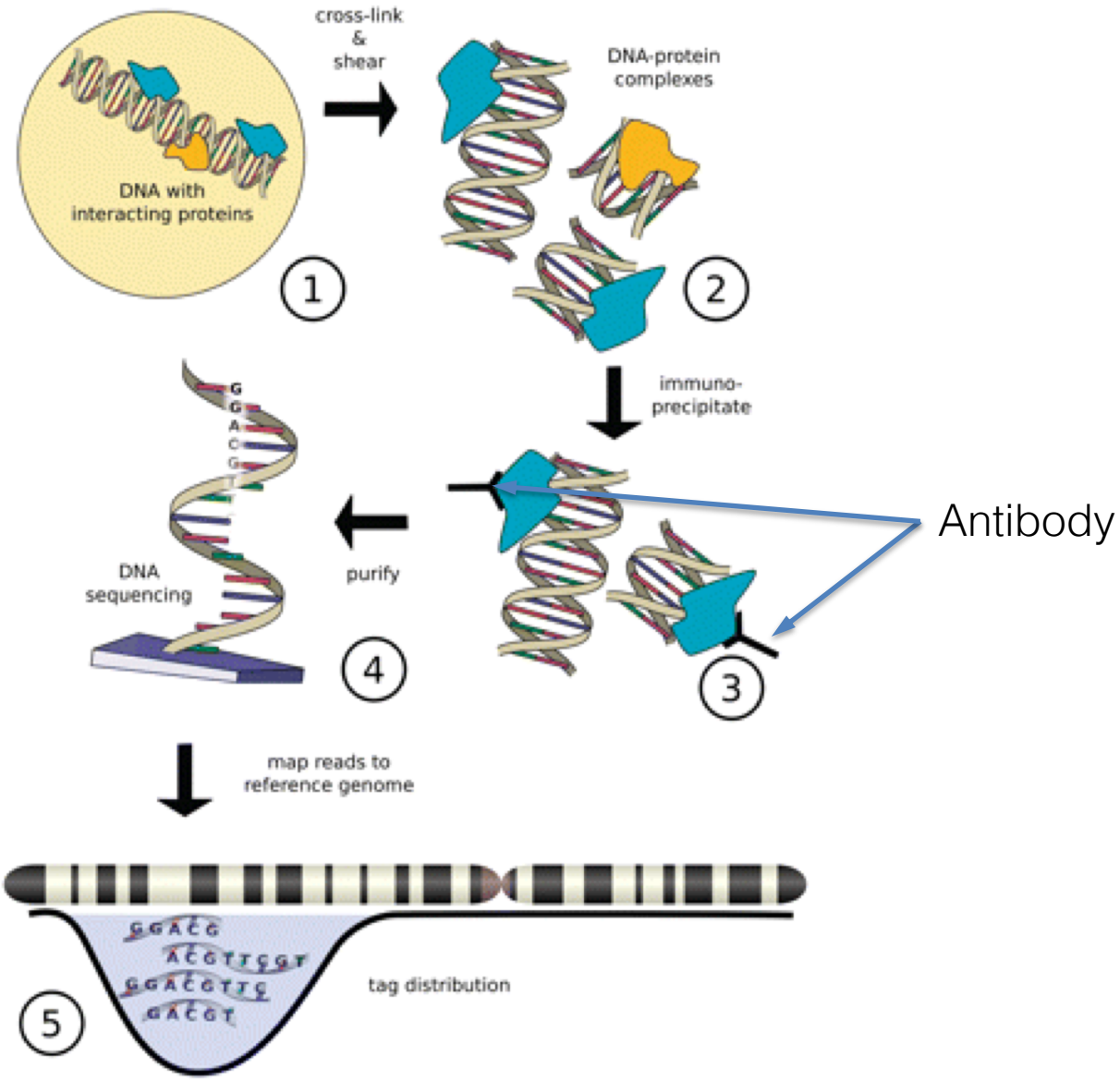
Slides from Daniel Gaffney

Epigenetics/ChiP in one slide

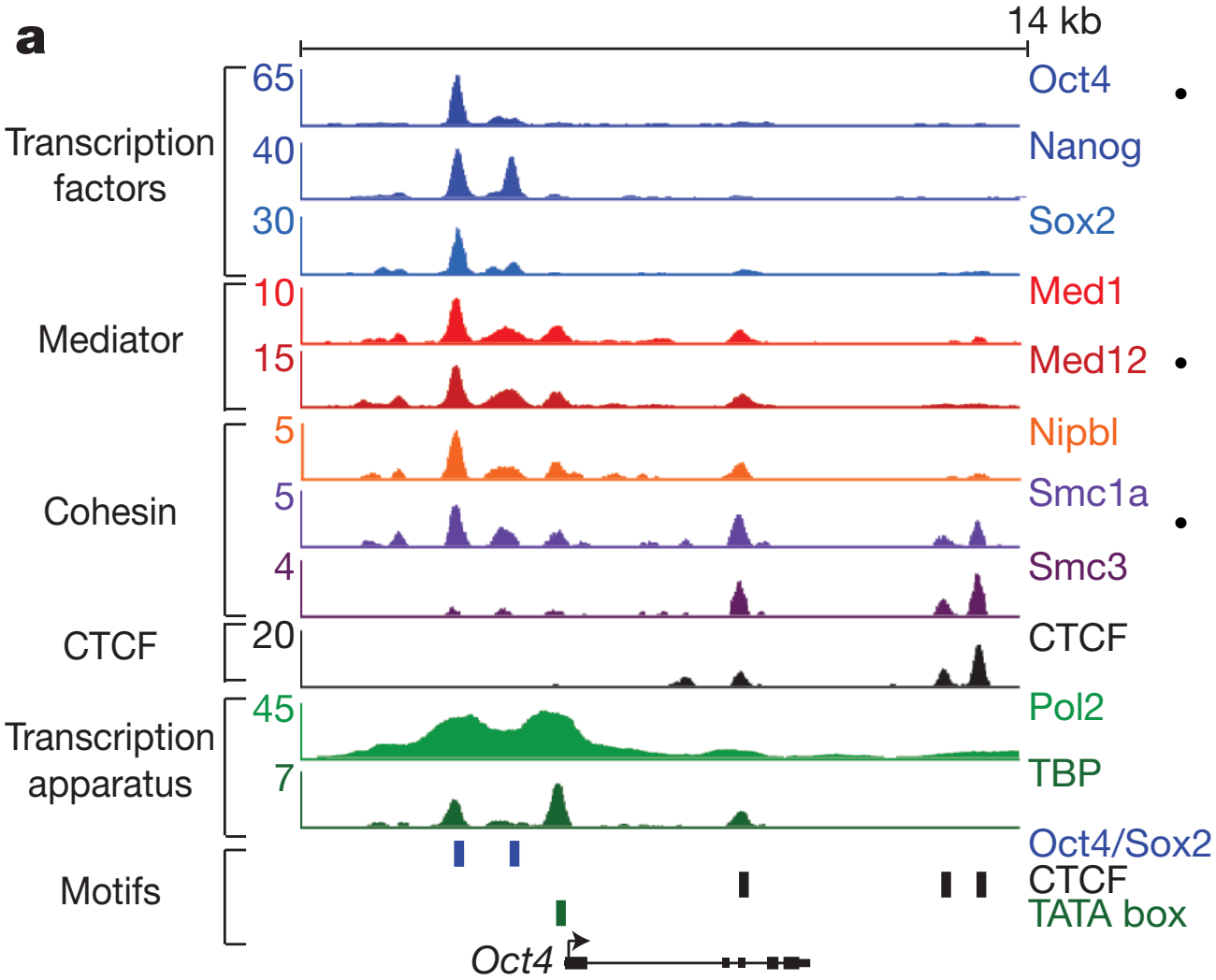


- Regulation of transcription involves interaction of protein and DNA

How does ChIP-seq work?



What does ChIP-seq look like?

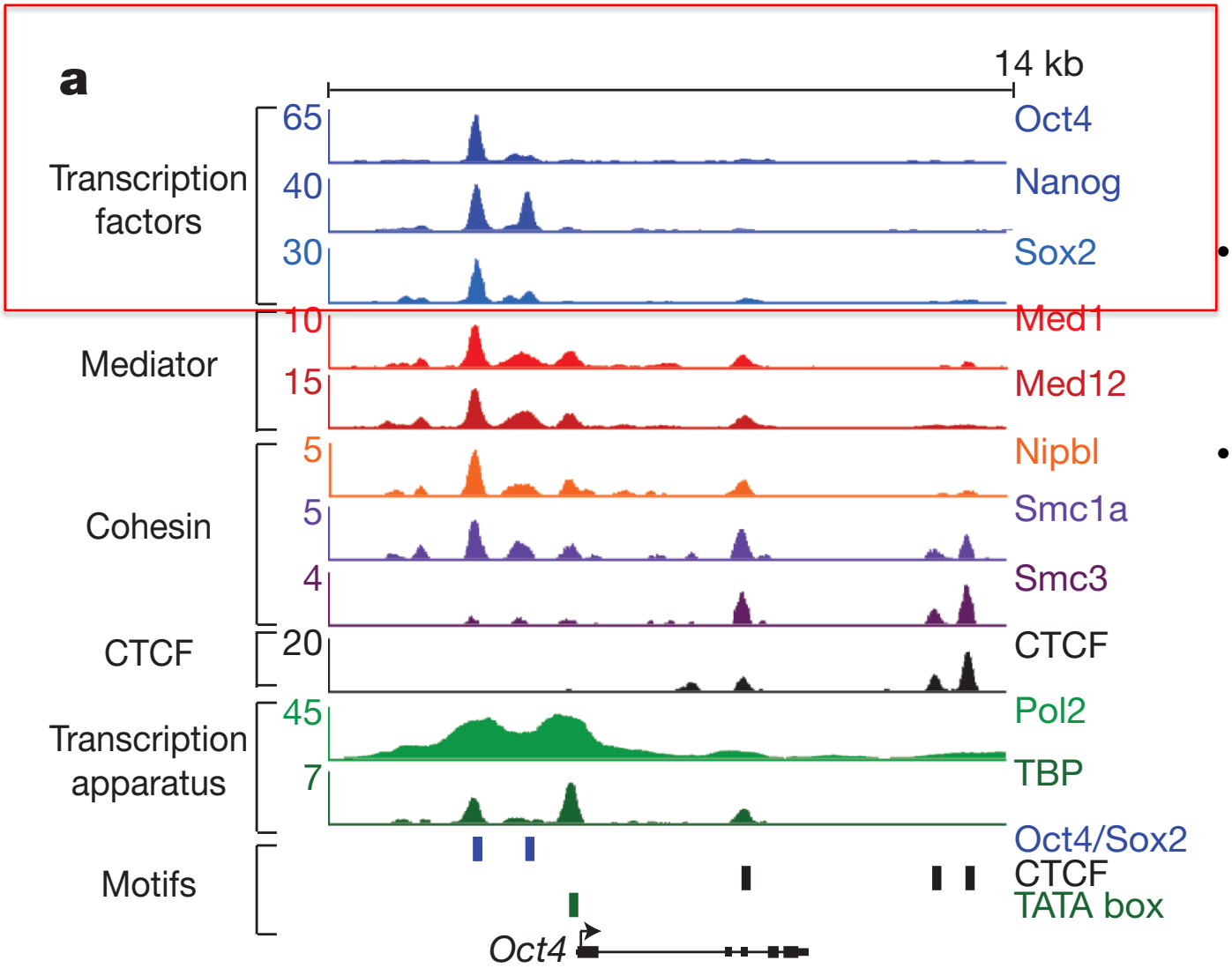


- A statistical procedure (peak calling) is used to call regions of enrichment (peaks)
- Can use a control "Input" sample as a background
- Peak calling quality varies dramatically by quality of the ChIP-seq

Applications of ChIP-seq

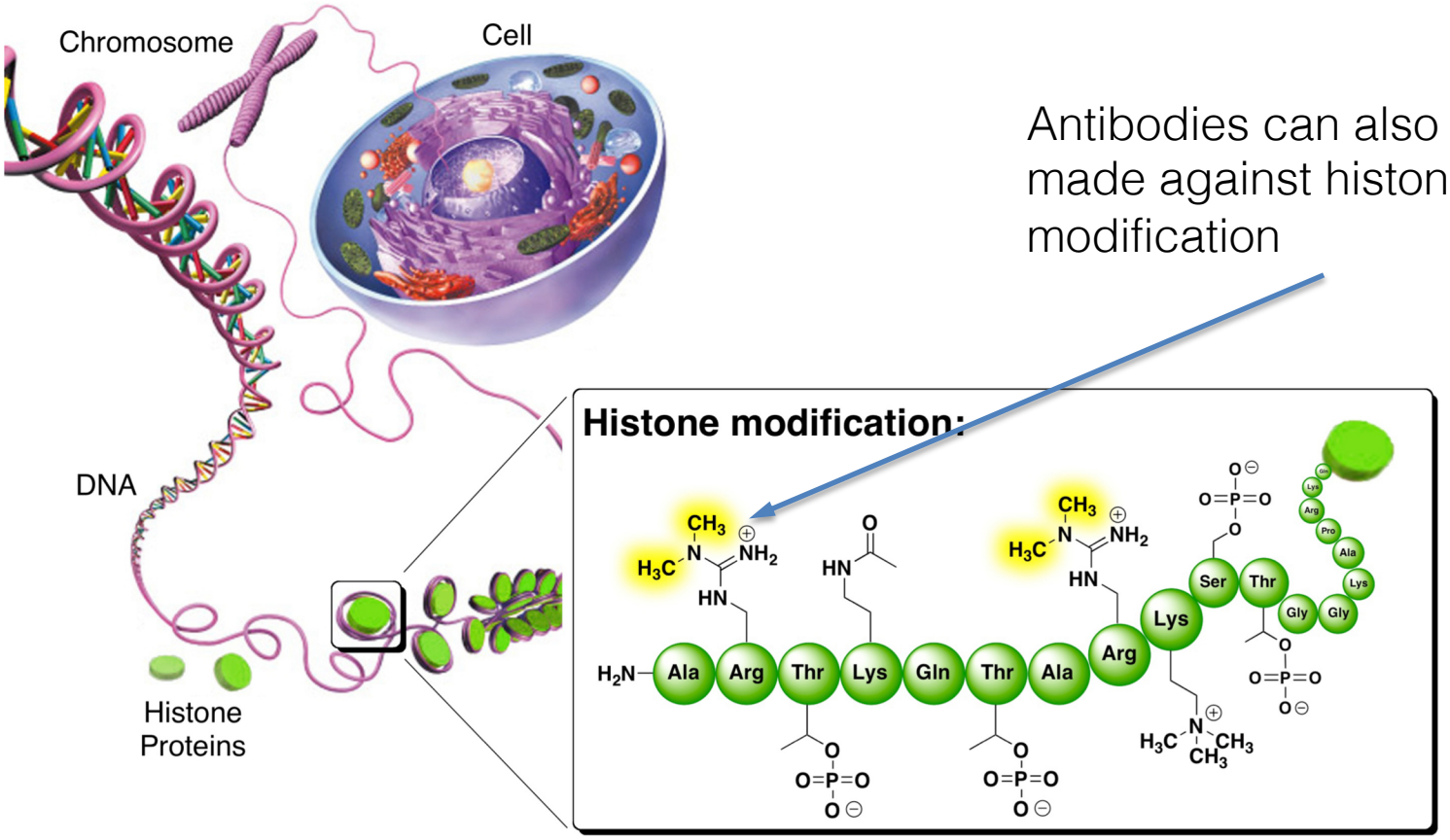
- ChIP-seq is one of the most commonly used approaches for identifying gene regulatory regions
- Two common types:
  1. Transcription factors
  2. Histone modifications

ChIP-seq for transcription factors



- Each of these TFs requires a high quality, ChIP-grade anti-body
- Most antibodies (~60%) are not good enough for ChIP-seq

Histone modifications



Antibodies can also be made against histone modification

EPIGENETIC JARGON CHEAT - SHEET

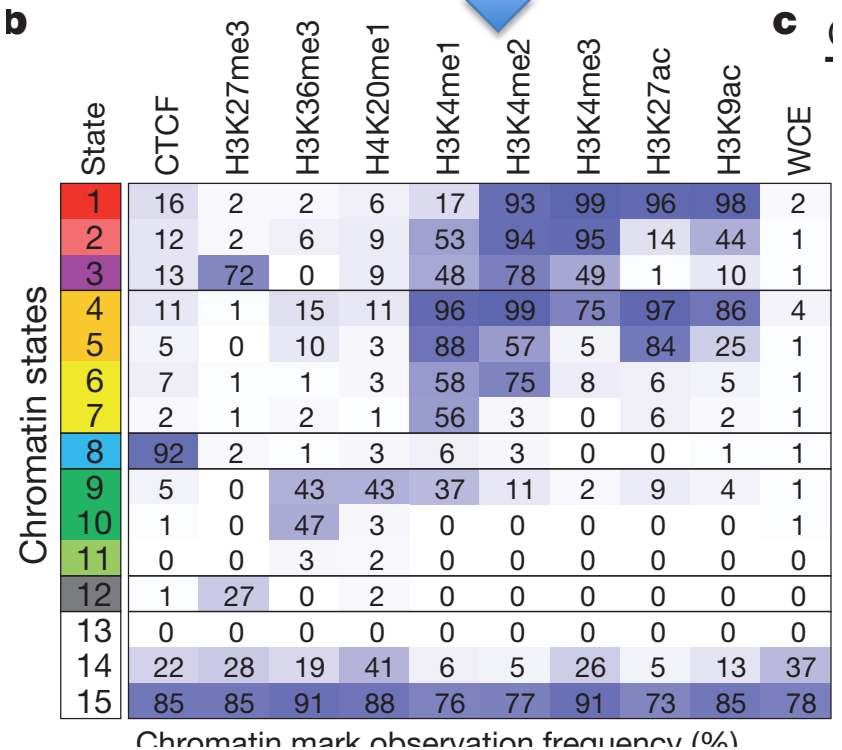
Regulatory Element	Meaning
Promoter	DNA Sequence (100-1kb), initial secure binding site for: RNA Pol complex Transfacs Adjacent regulated gene, defined relative to TSS. Poised: simultaneous activation/repressive histone mods.
Enhancer/Silencer	DNA Seq (50-1.5kb), bound by transfacs *(activator / repressor)* Can act on gene up to 1Mb away: DNA folding brings it close to promoter. Enhancer: Bound by activator, which interacts with complex initiating transcription. Silencer: bound by repressor, which interferes with GTF assembly.
Insulator	DNA, 300-2kb, Block enhancers from acting on promoters: positioned between enhancer and promoter, form chromatin-loop domains.
Polycomb-repressed	Polycomb – group proteins actively remodel chromatin to silence genes.

The histone code

Ernst *et al* 2011

Then: go back and ask what fraction of classified regions contain peaks of a given type.

First: create these categories by applying HMM classifying stretches of genome to combined peak data:
9 cell lines x 9 chromatin marks.
Apply functional interpretation after categories are created.



(NHLF)

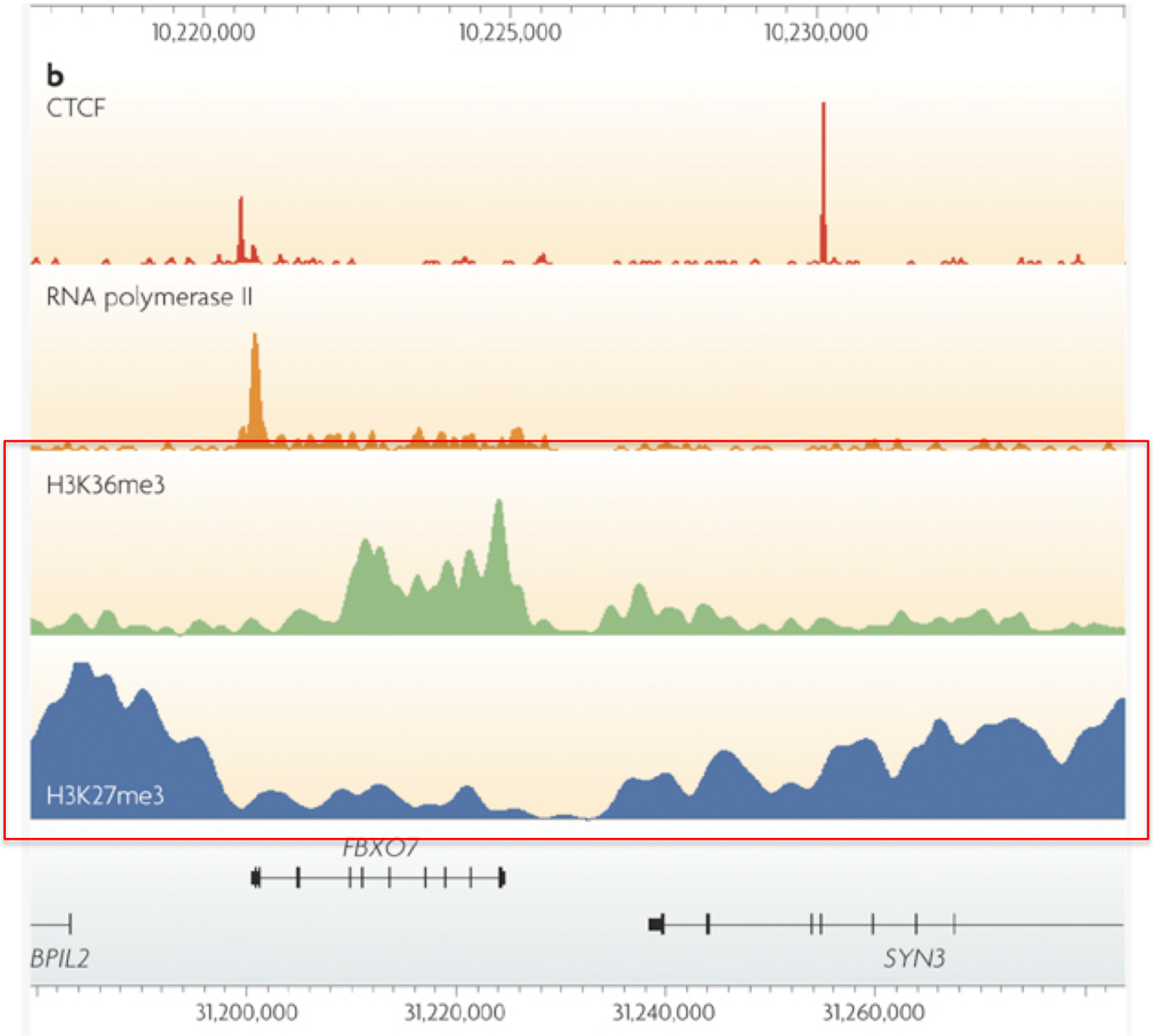
Candidate state annotation

Active promoter
Weak promoter
Inactive/poised promoter
Strong enhancer
Strong enhancer
Weak/poised enhancer
Weak/poised enhancer
Insulator
Transcriptional transition
Transcriptional elongation
Weak transcribed
Polycomb repressed
Heterochrom; low signal
Repetitive/CNV
Repetitive/CNV

Histone mark cheat sheet

Histone mark	Candidate State	Interpretation
H3K9me2,3	-	Silenced genes
H3K27me3	Inactive/poised promoter, polycomb repressed	Downregulation of nearby genes
H3K36me3	Transcriptional transition	Actively transcribed gene bodies.
H4K20me1	Transcriptional transition	Transcriptional activation
H3K4me1,2,3	Strong enhancer	Promoter of active genes
H3K27ac	Active promoter/strong enhancer	Active transcription
H3K9ac	Active promoter	Switch from transcription initiation to elongation.

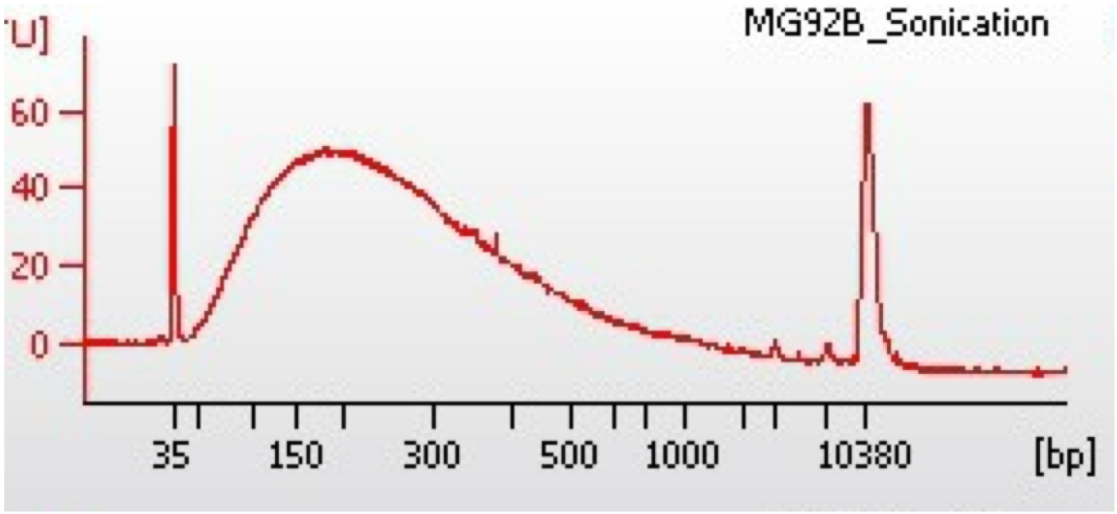
Histone marks



ChIP-seq experimental considerations

- Antibody quality must be high: 60% of antibodies not high enough quality
- Numbers of cells: 2-3M recommended, more for TFs (5-10M)
- Crosslinking time: ~10 mins
- Shearing fragment checked – next slide

Shearing



FRAGMENTS: Aim for fragments in 150-400bp range

- Efficiency varies by cell type
- Optimise by varying number of shearing cycles
- Run input samples on Bioanalyser to check efficiency

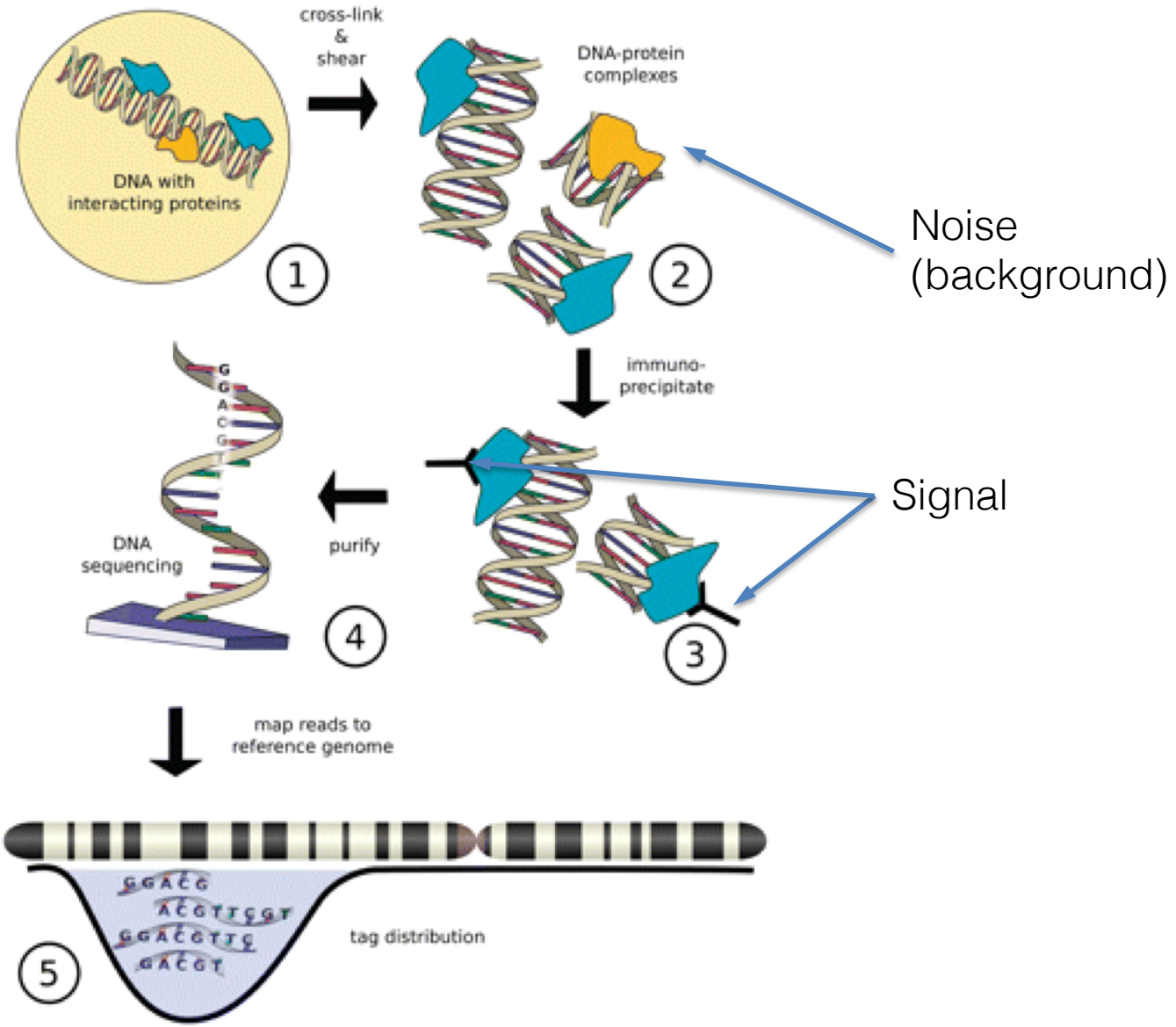
READS: Longer reads better than shorter (uniqueness)

Paired End reads better than single end reads (uniqueness on mapping)

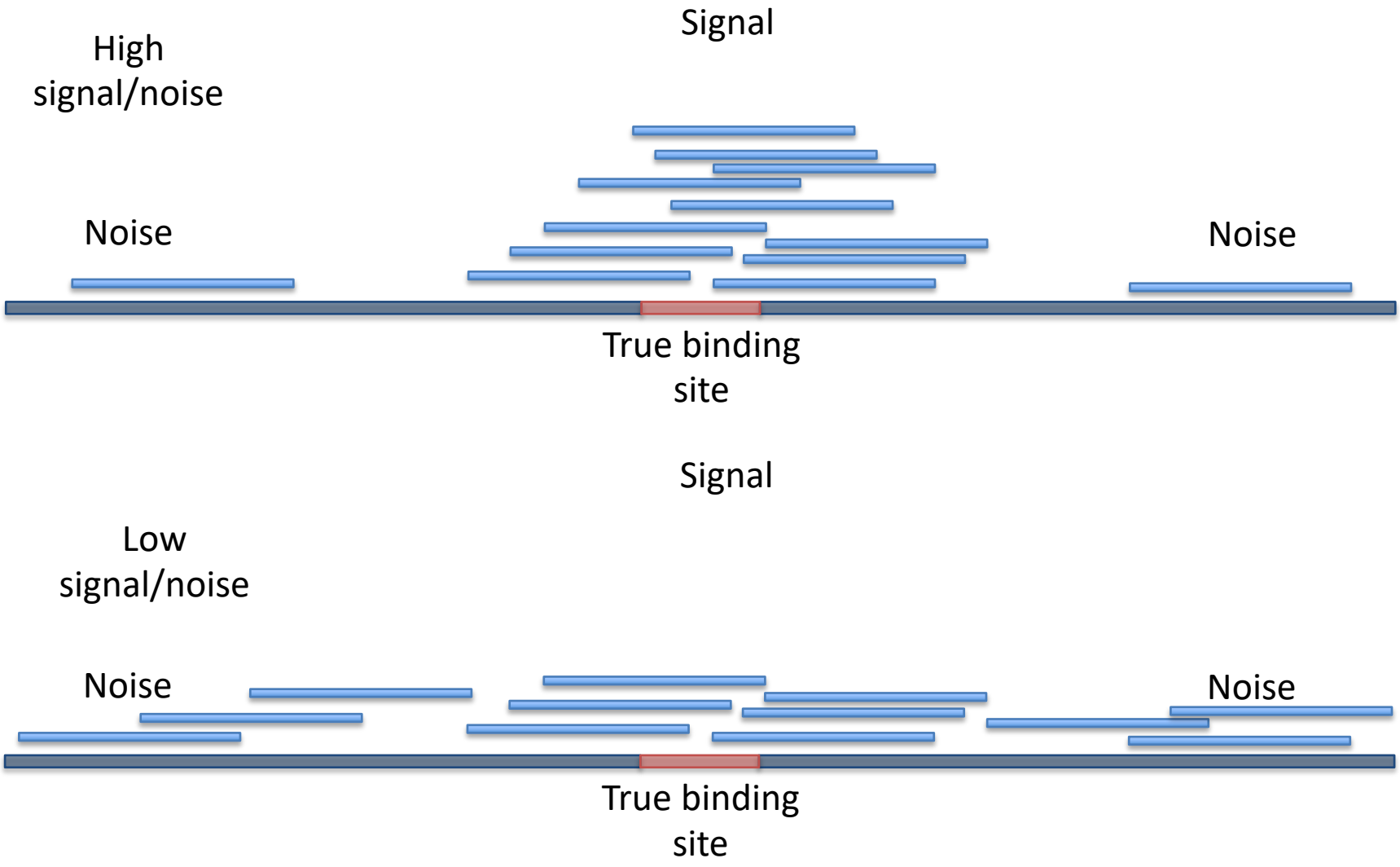
ChIP-seq technical issues

1. Signal / noise: Does my antibody work?
2. Library complexity: Did I have enough starting material?

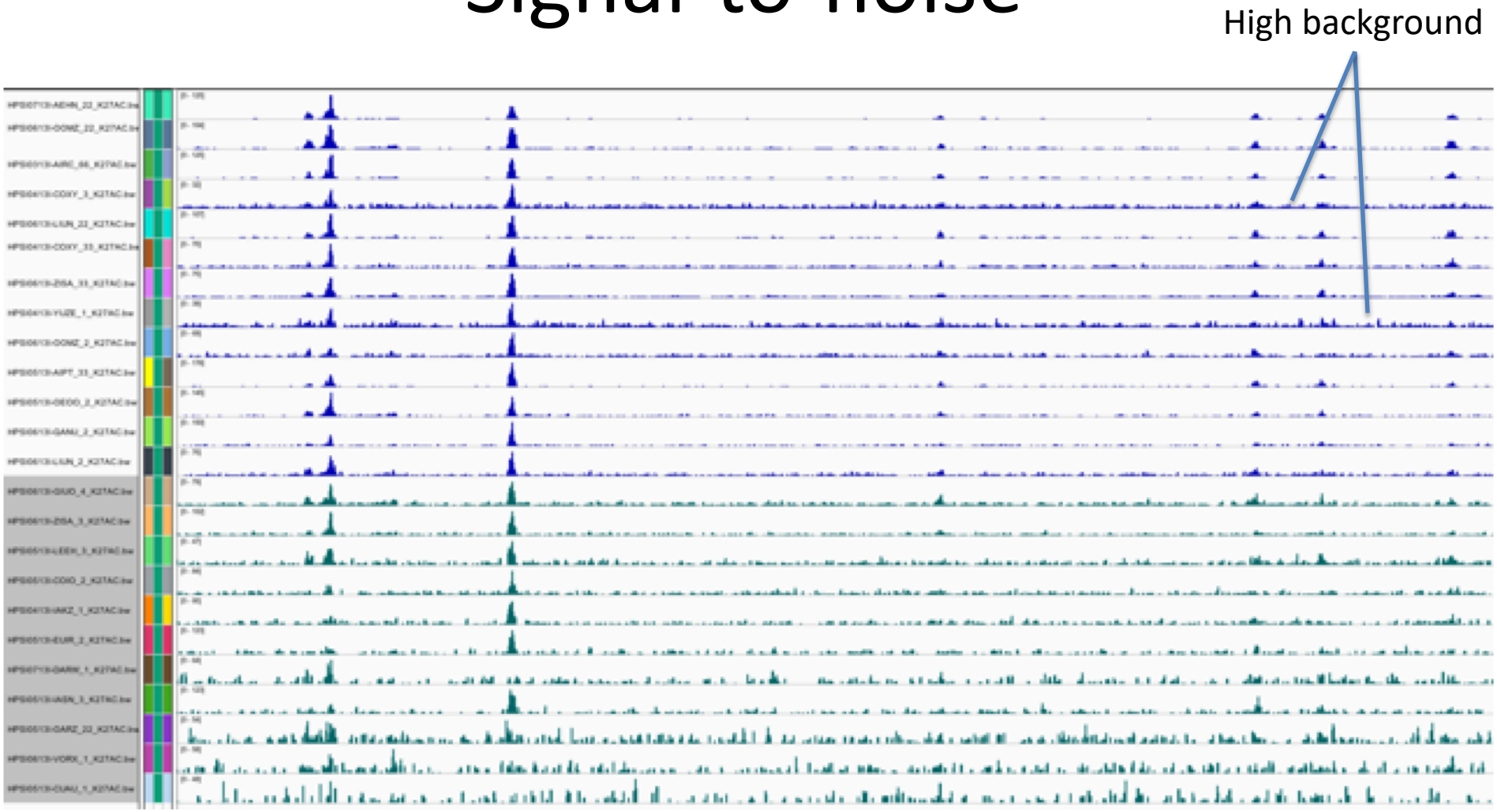
Signal / noise



Signal / noise



Signal-to-noise



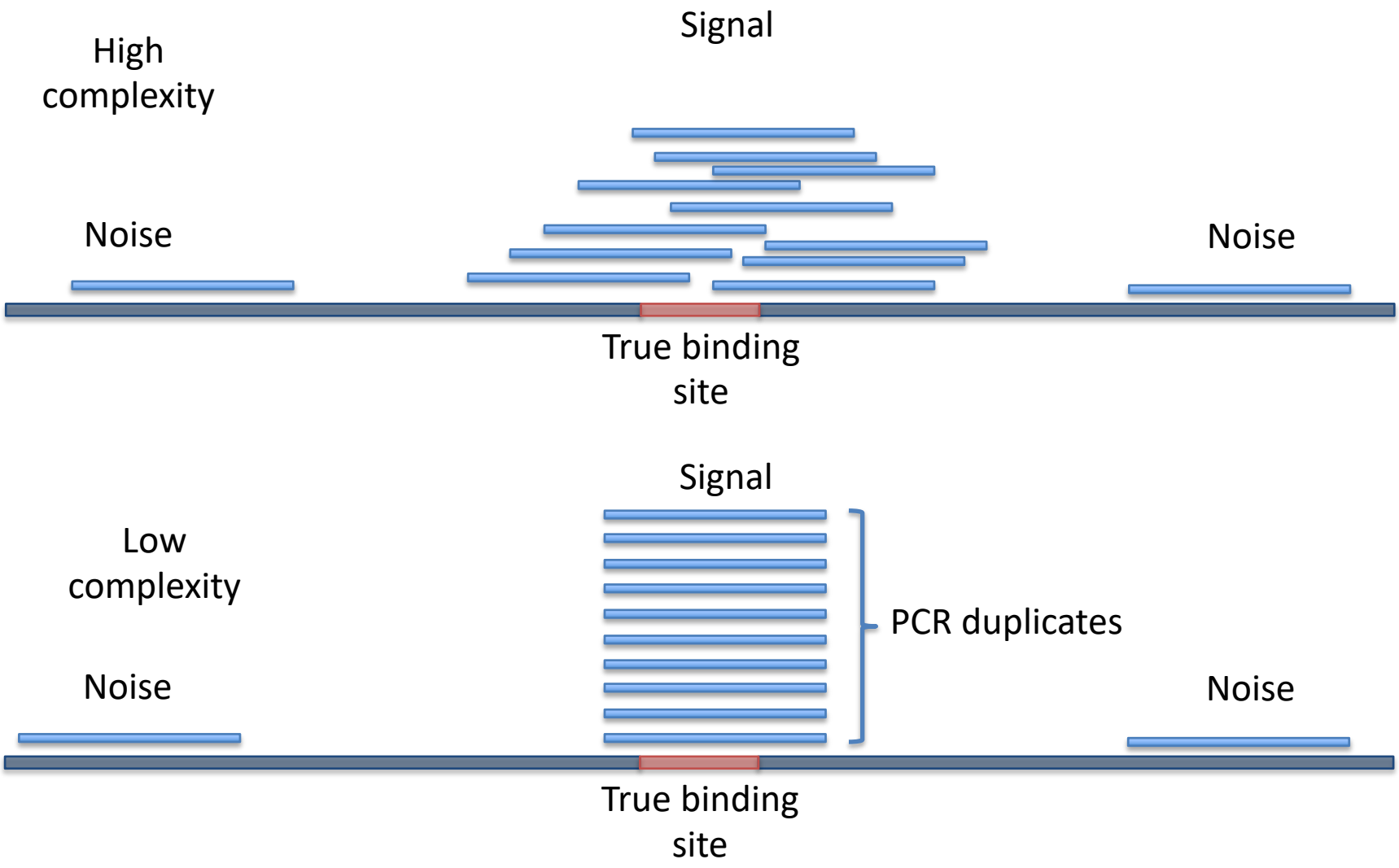
Measuring S/N via FRIP

- Fragments In Peaks
- # Fragments found in peaks / Total # fragments
- >1%

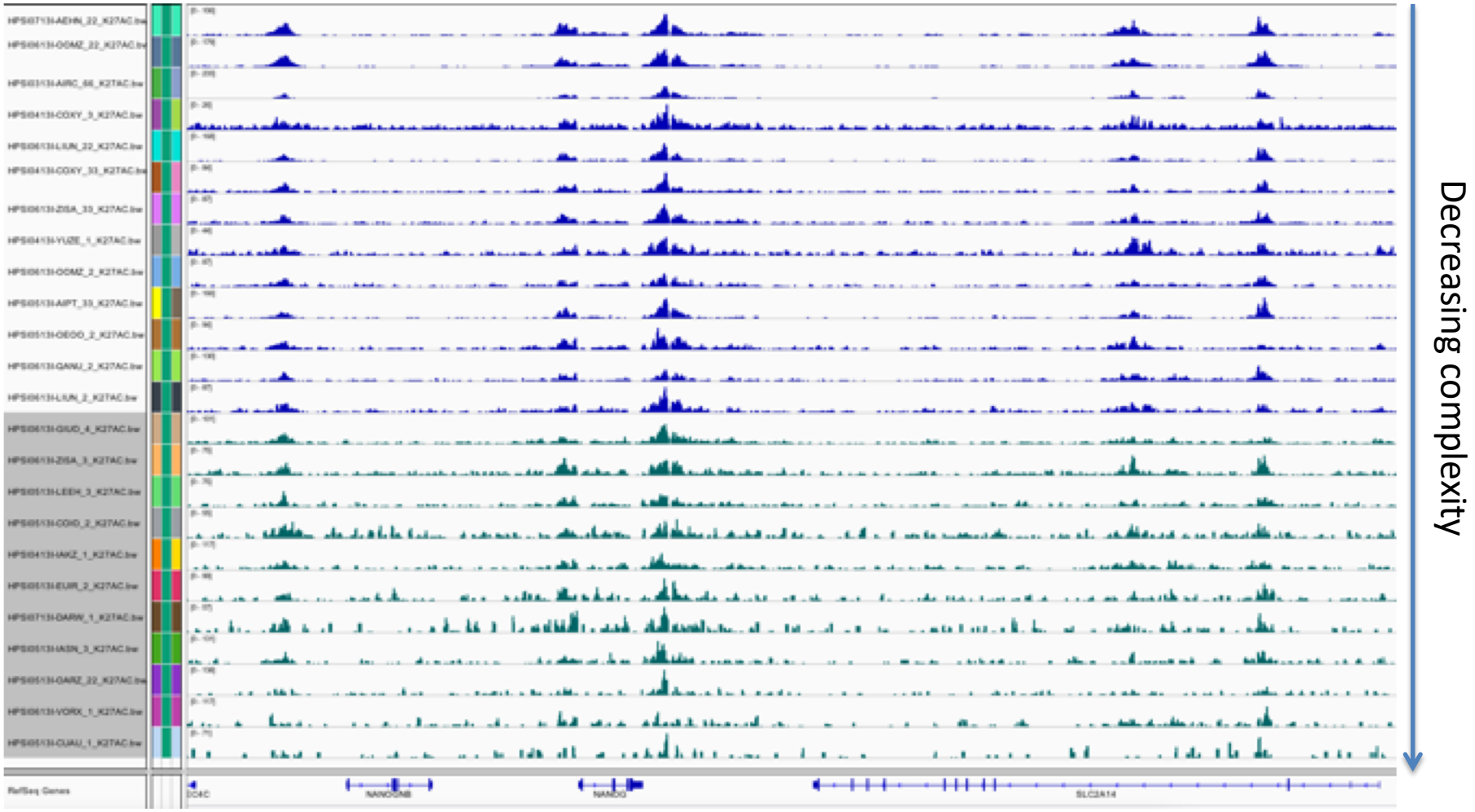
Library complexity

- Problem: Not enough starting material
  - Not enough cells
  - Antibody efficiency
- More PCR required

Library complexity



Library complexity



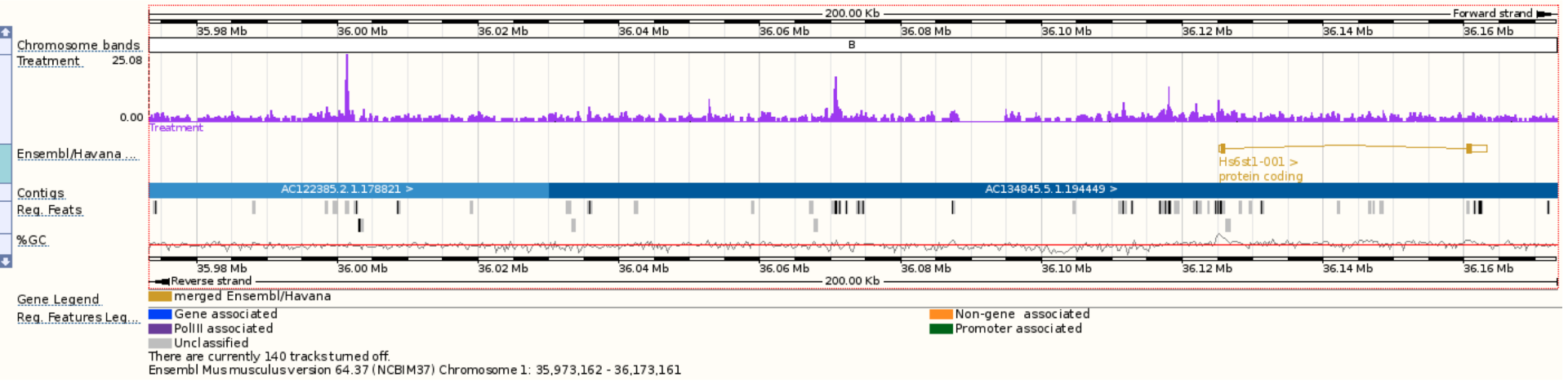
Measuring complexity – Nonredundant fraction

- NRF = # unique fragments positions / total # fragments
- Encode goal: NRF >0.8

Basic analysis of ChIP-seq

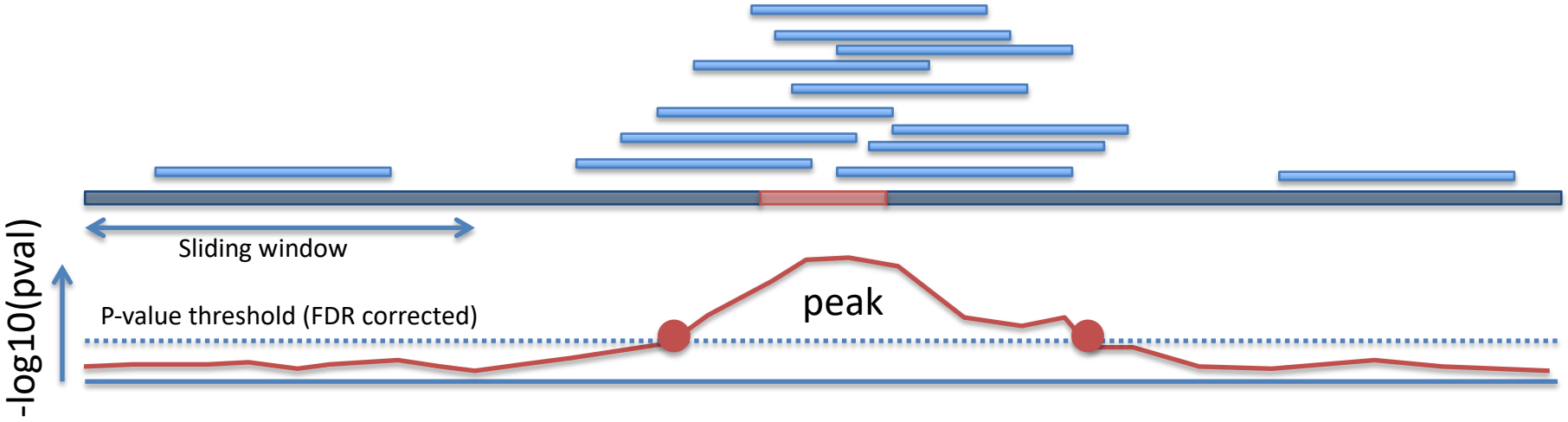
1. Read alignment
2. Visualisation
3. Peak calling
   - Peak annotation (mapping peaks to genes etc)
   - Motif analysis
4. Differential binding
   - Case / control
   - Naïve / stimulated

Visualisation in a genome browser



- Convert mapped reads to “signal” – e.g. read depth at each bp or in windows
- BAM files to e.g. wig, bedgraph
- IGV, ensembl, UCSC

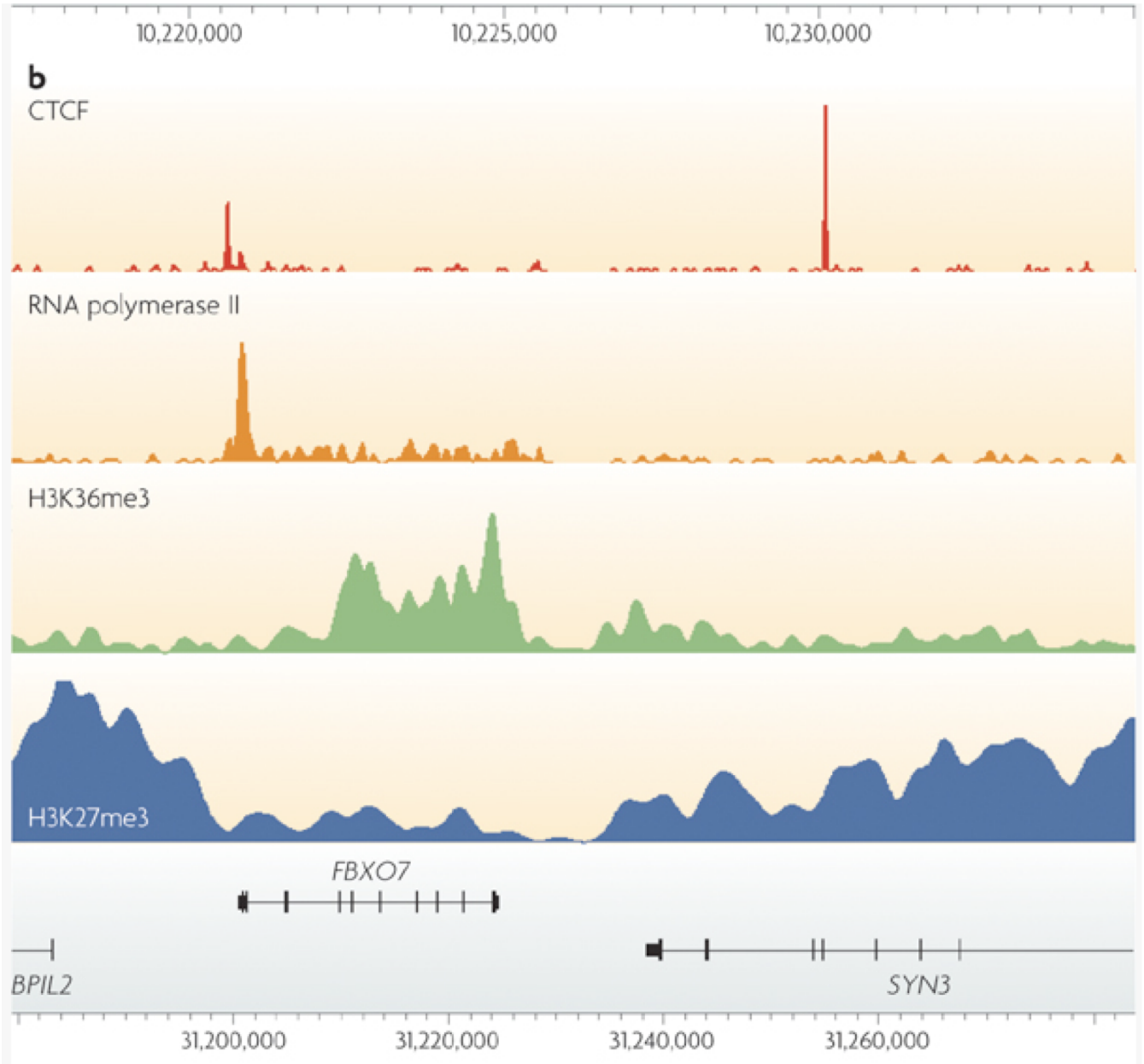
Peak calling



- Observed counts
- Expected counts
- Poisson test: p-value – prob(observing frag count at least as extreme under null)

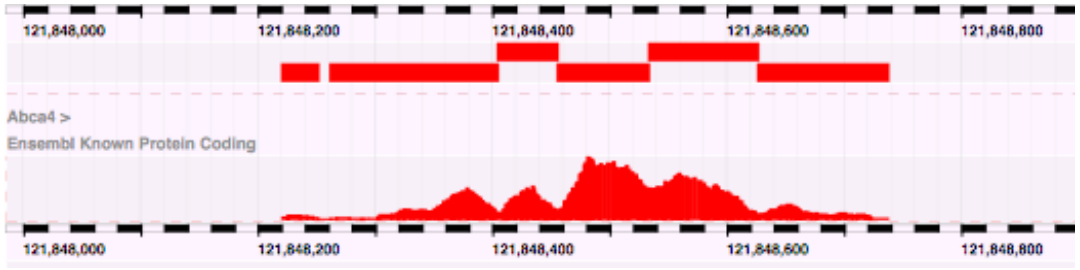
Peak calling challenges

- What's expected?
  - Treatment sample (with antibody)
  - Input / control sample (no antibody)
- Replicates
  - Yes! (min 2, more = better)
- Peak sizes
  - These are variable: small for TFs, large for some Histone mods, and for Pol2 etc.



Park J, Nature Reviews Genetics, 2009

Motif analysisr



EXTRACT SEQ FROM *SUMMITS*

Align sequences from multiple peaks

```
GAATCCCACA TTTGCATAACAAAAG ACTCCTGGTG
CAGCTGCTCT TCTGCATAACAAAGG GTGGCCCTGC
CCGGTTTTTC TTTGCATAACAATAA GATCTGGCTA
TTATTCTCAC TTTGCATAGGAATGG GGCAGTTAGA
CACAGCCACA TTTGCATAACAGAAG CCGAGCCCGC
CTTGGGTGAA TTTGCAAGACAAAGG ACAATGATCA
```

CONSENSUS LOGO
***Letter Height**: Relative Frequency*
***Stack Height** = 2 – Shannon entropy**
**sum_i(p_i log p_i) – how smeared out probs are*
More smeared out => lower stack height

Discover motifs