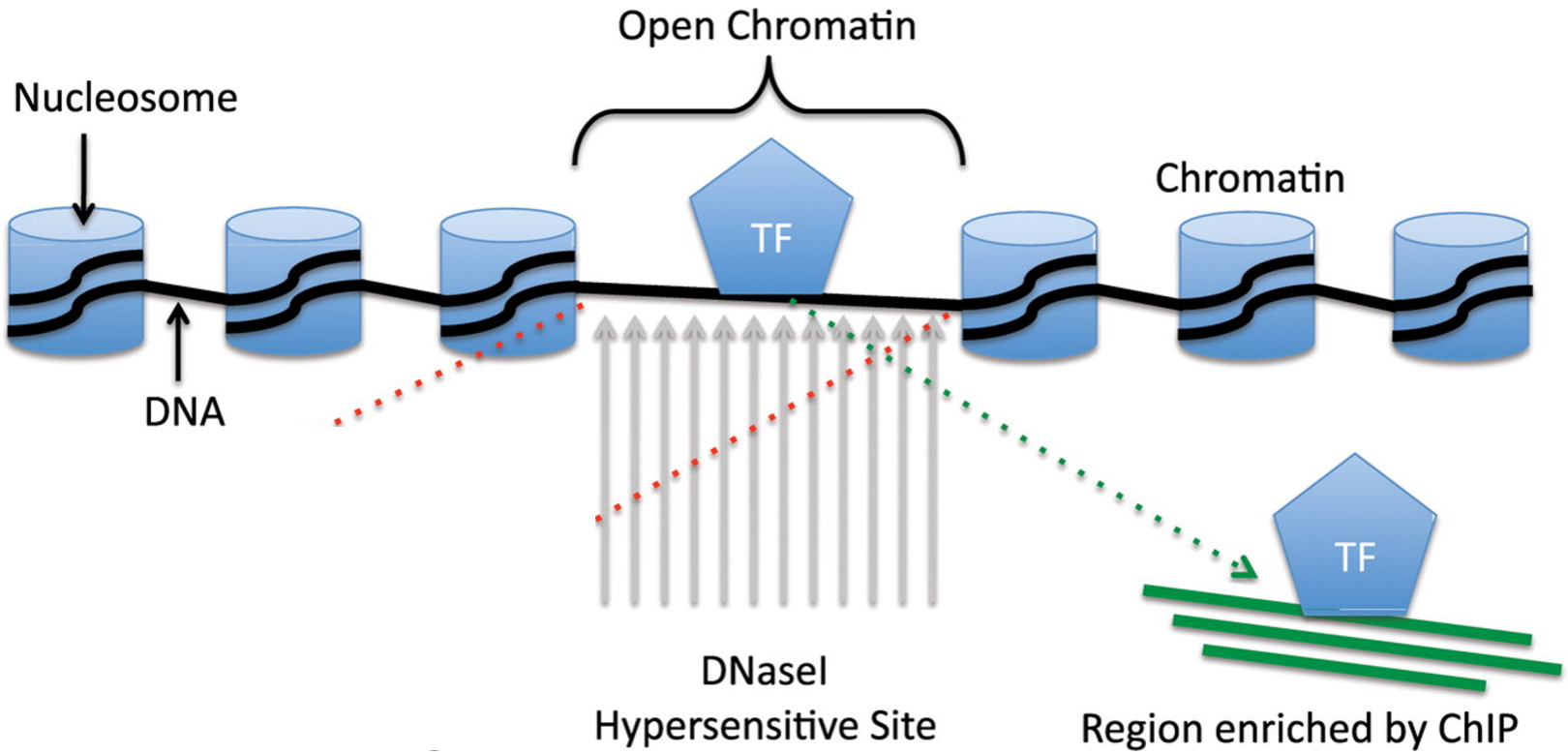
Intro to ChIP-seq

Alejandra Medina Rivera

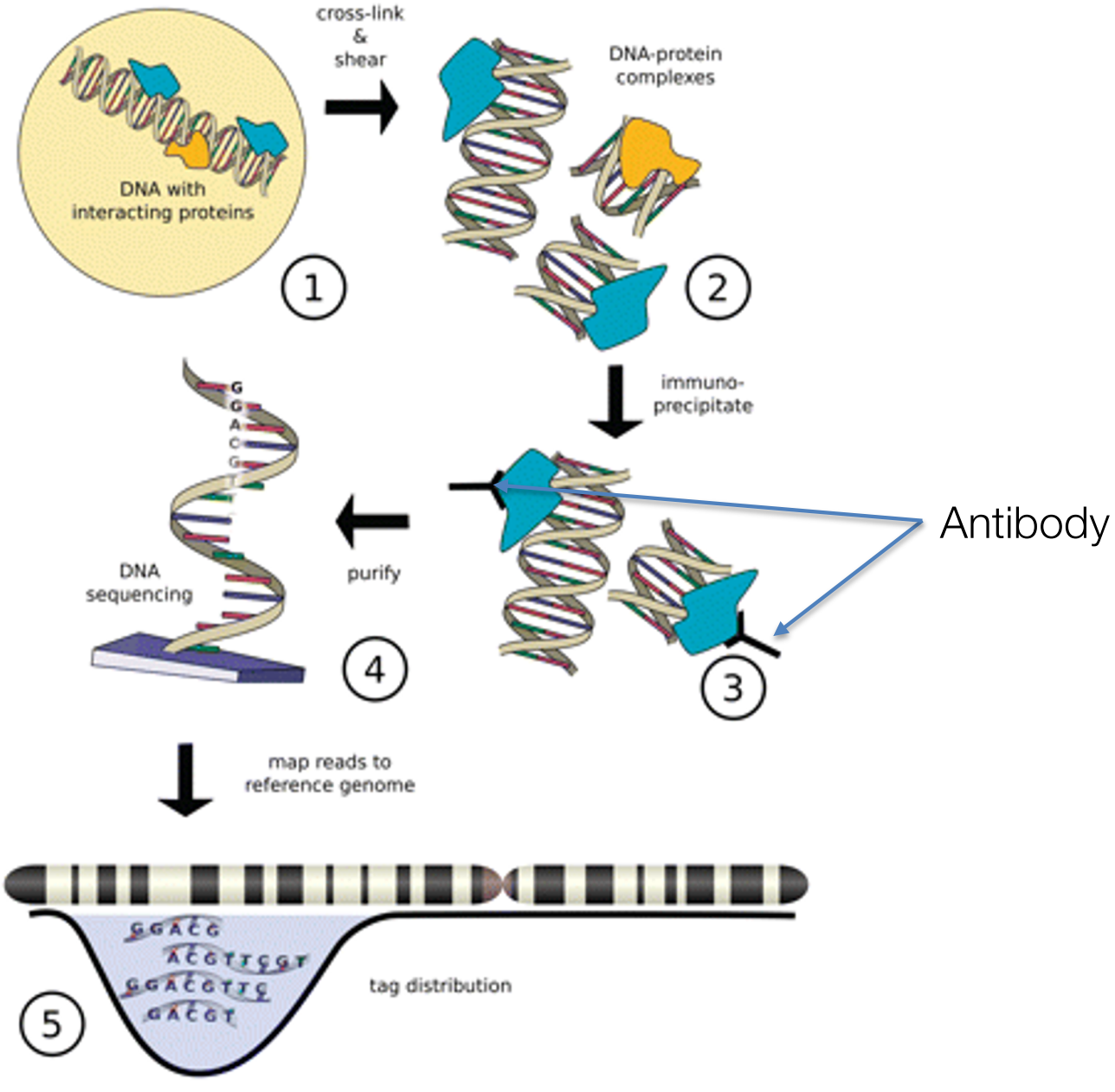
Material by: Daniel Gaffney

Epigenetics/ChiP in one slide

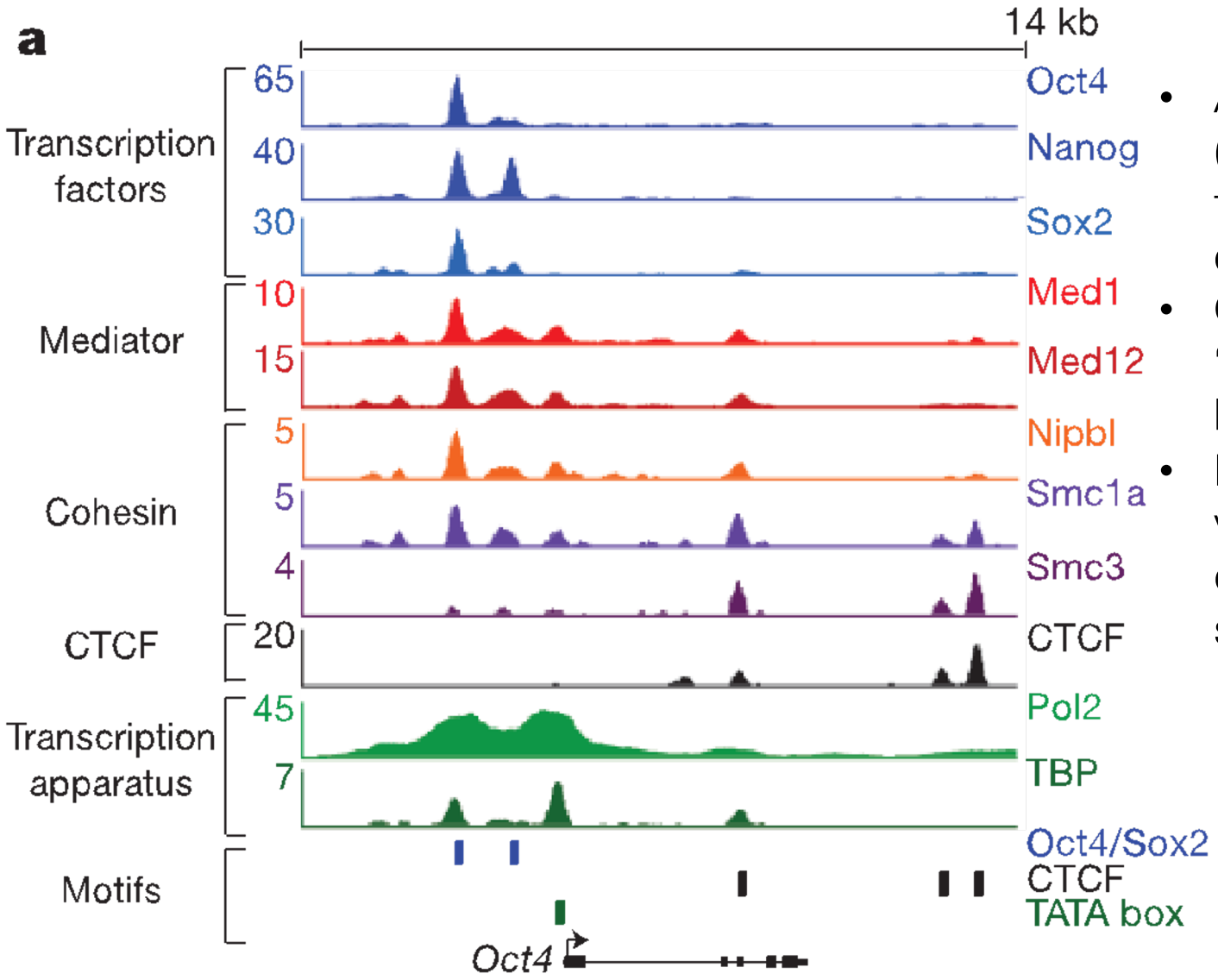


- Regulation of transcription involves interaction of protein and DNA

How does ChIP-seq work?



What does ChIP-seq look like?

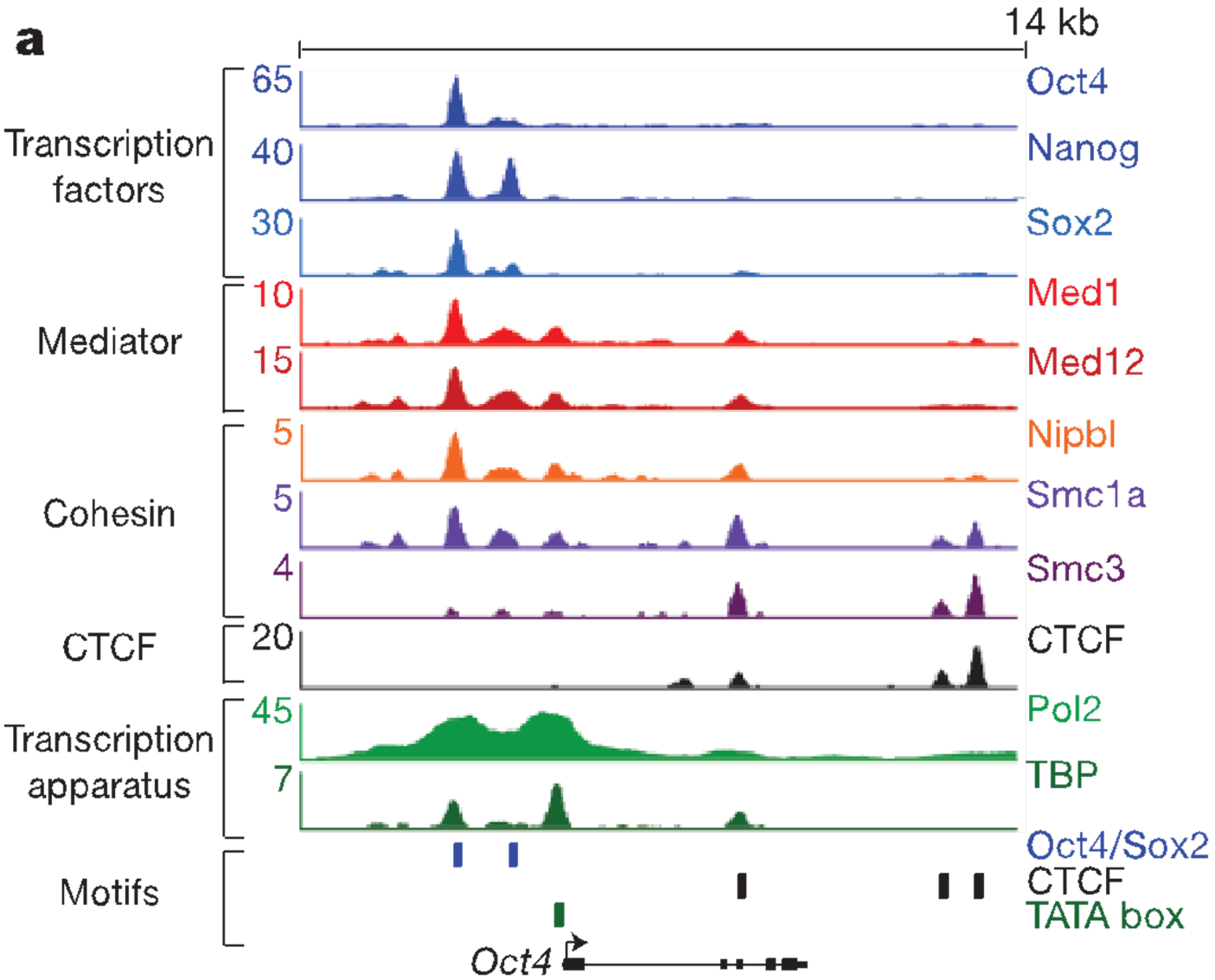


- A statistical procedure (peak calling) is used to call regions of enrichment (peaks)
- Can use a control "Input" sample as a background
- Peak calling quality varies dramatically by quality of the ChIP-seq

Applications of ChIP-seq

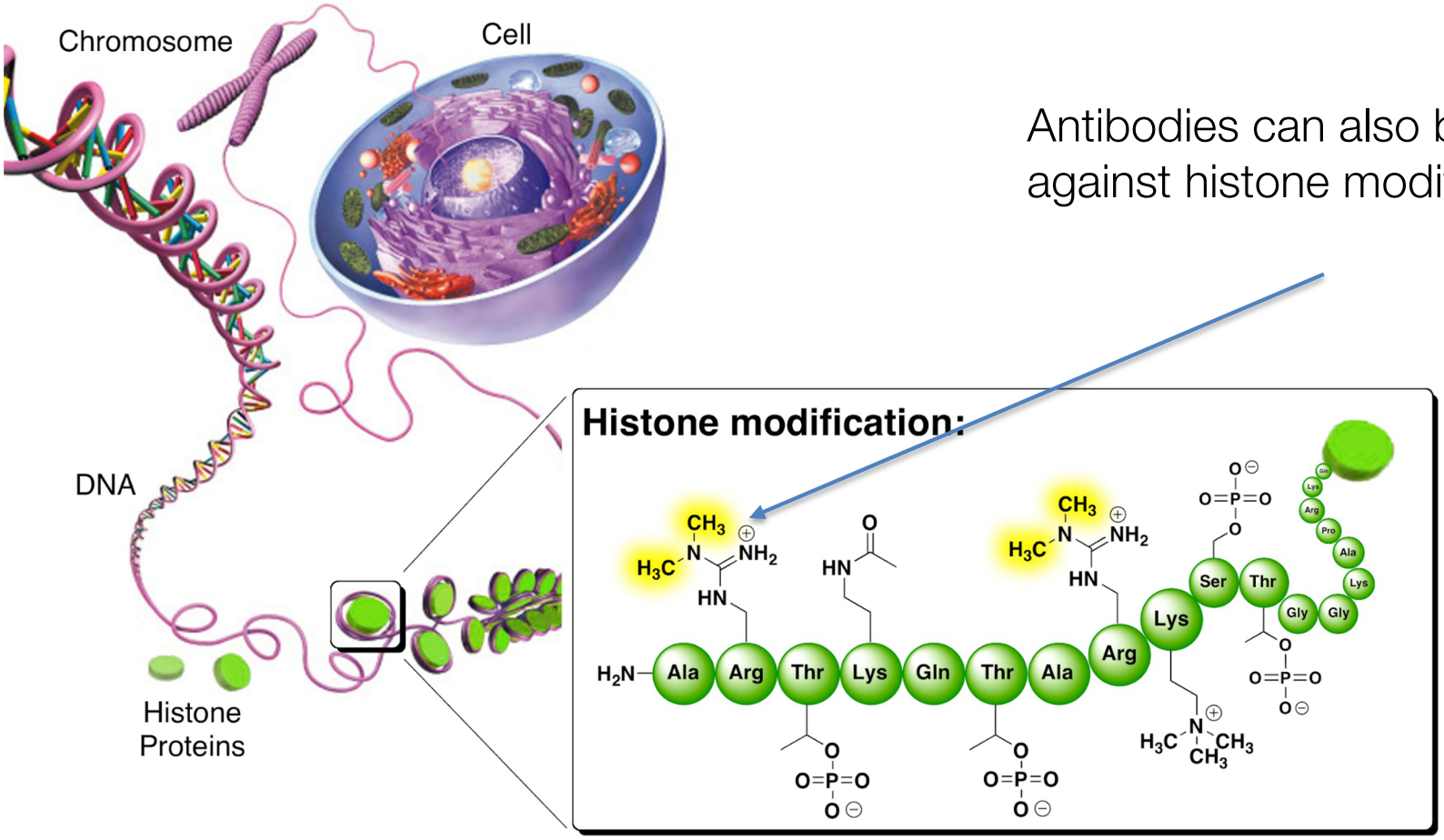
- ChIP-seq is one of the most commonly used approaches for identifying gene regulatory regions
- Two common types:
  1. Transcription factors
  2. Histone modifications

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

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.

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

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.

Ernst *et al* 2011

b

State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
1	16	2	2	6	17	93	99	96	98	2
2	12	2	6	9	53	94	95	14	44	1
3	13	72	0	9	48	78	49	1	10	1
4	11	1	15	11	96	99	75	97	86	4
5	5	0	10	3	88	57	5	84	25	1
6	7	1	1	3	58	75	8	6	5	1
7	2	1	2	1	56	3	0	6	2	1
8	92	2	1	3	6	3	0	0	1	1
9	5	0	43	43	37	11	2	9	4	1
10	1	0	47	3	0	0	0	0	0	1
11	0	0	3	2	0	0	0	0	0	0
12	1	27	0	2	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	22	28	19	41	6	5	26	5	13	37
15	85	85	91	88	76	77	91	73	85	78

Chromatin states

c

(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

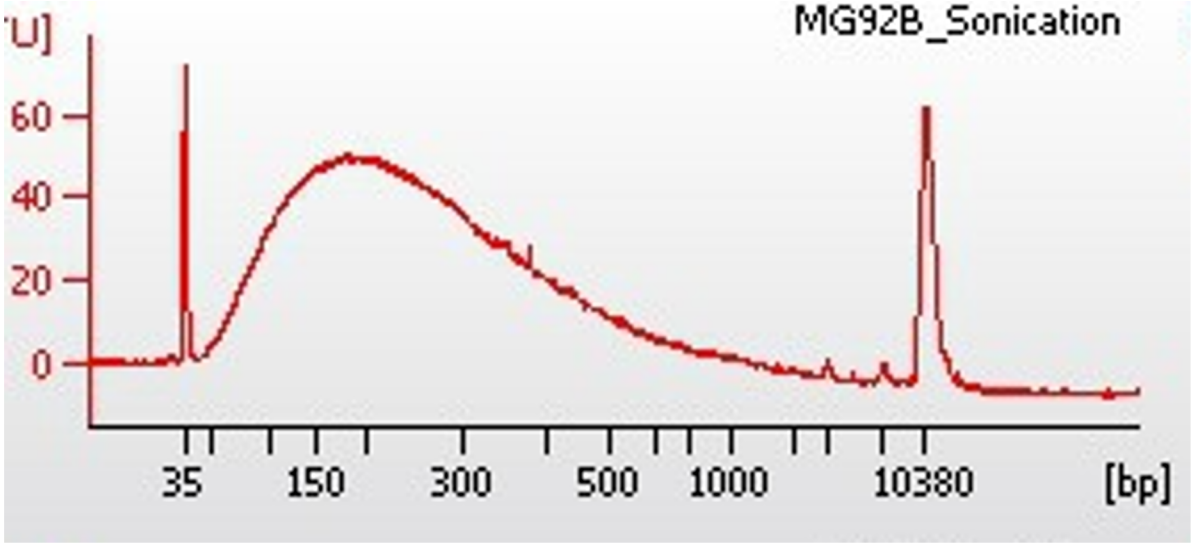
Histone mark cheat sheet

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.

ChIP-seq experimental considerations

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

Shearing

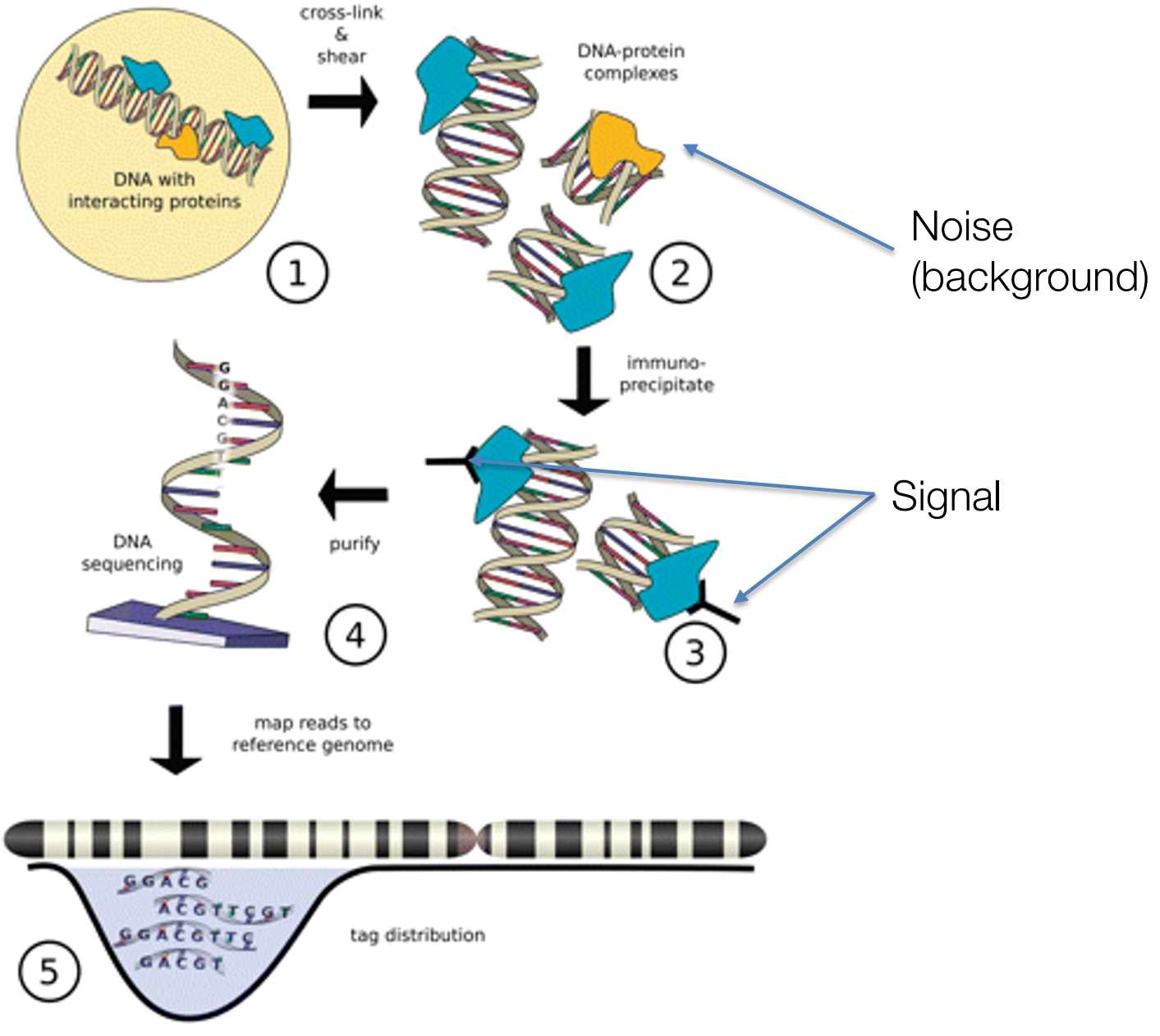


- Aim for fragments in 150-400bp range
- Efficiency varies by cell type
- Optimise by varying number of shearing cycles
- Run input samples on Bioanalyser to check efficiency

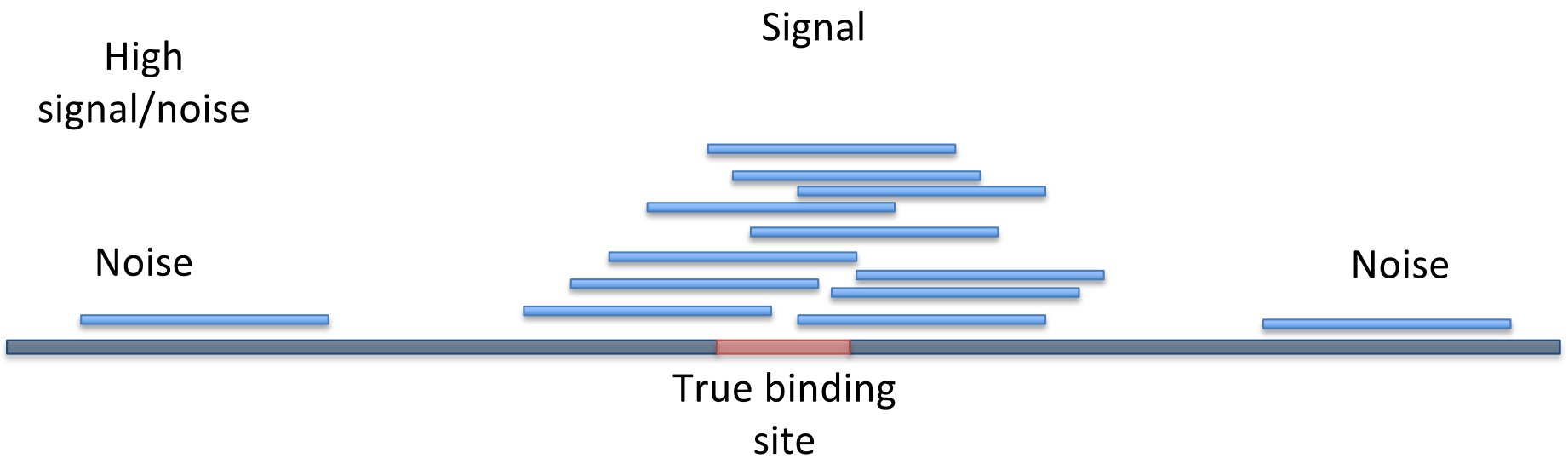
ChIP-seq technical issues

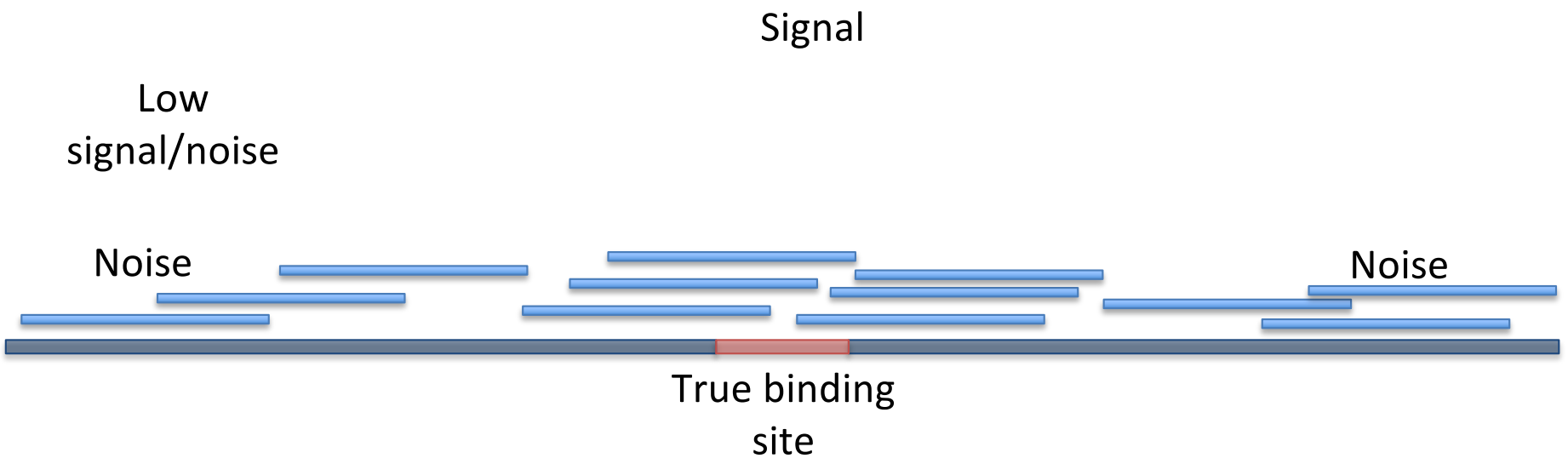
1. Signal / noise: Does my antibody work?
2. Library complexity: Did I have enough starting material?

Signal / noise

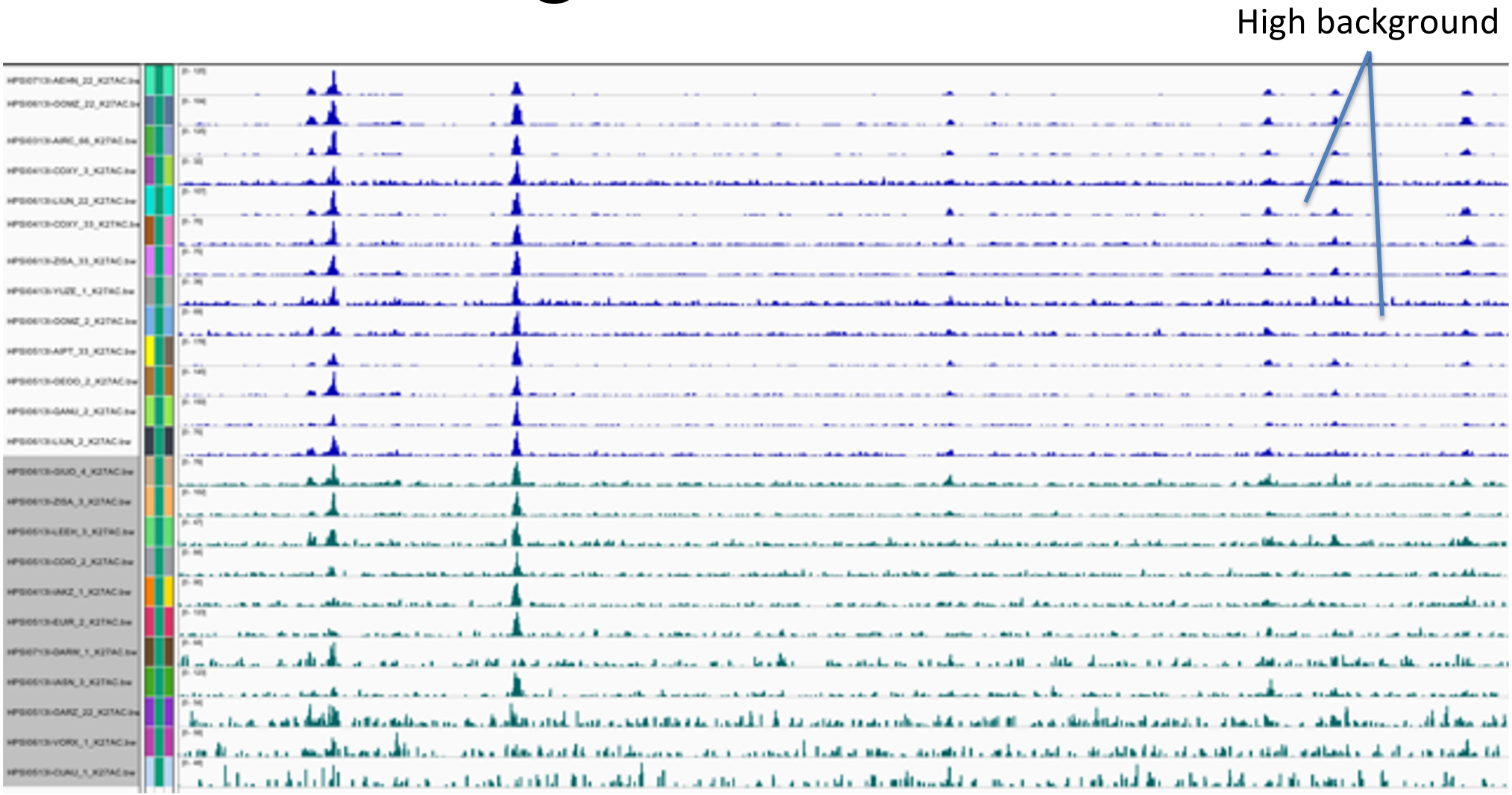


Signal / noise





Signal-to-noise



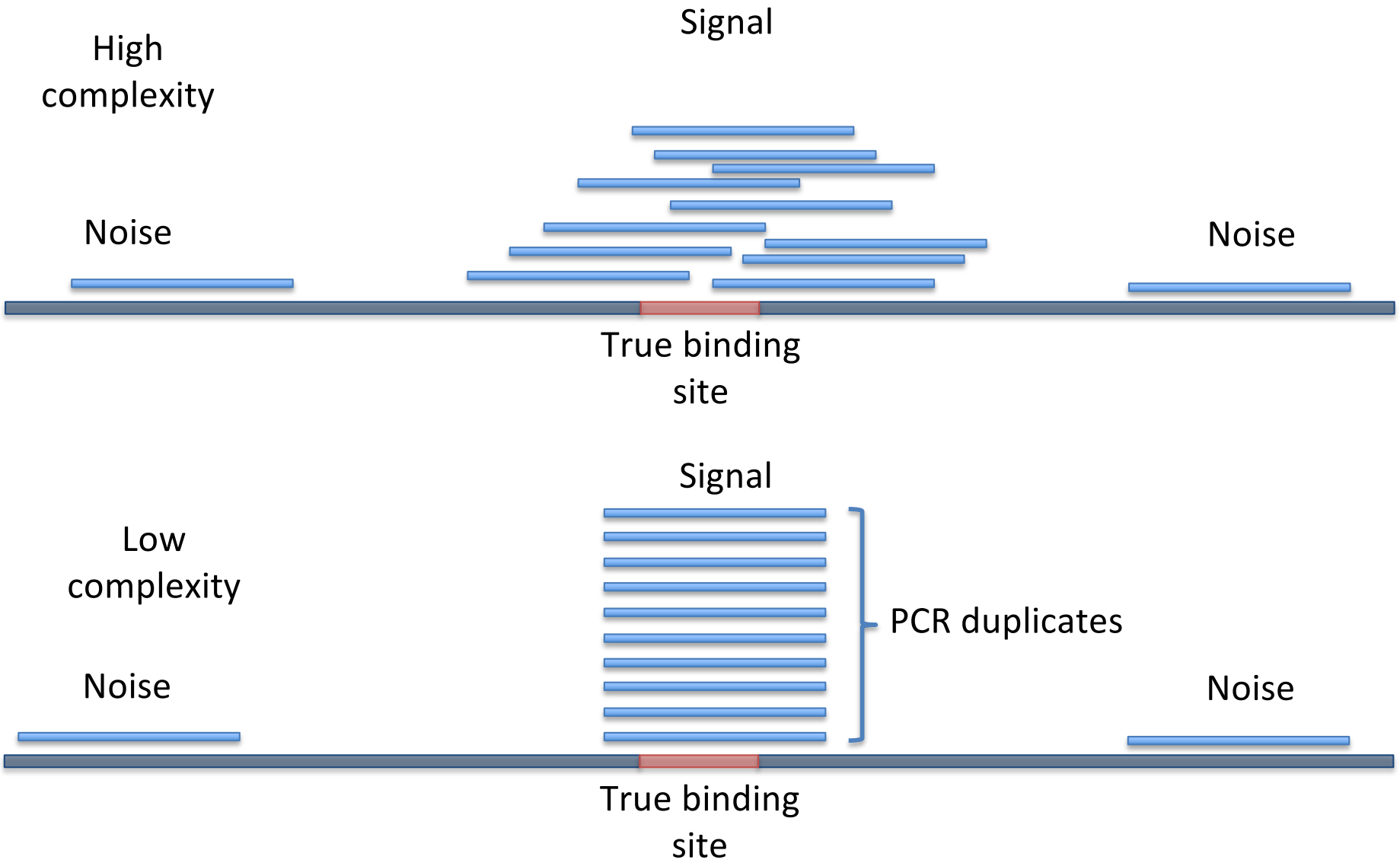
FRIP

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

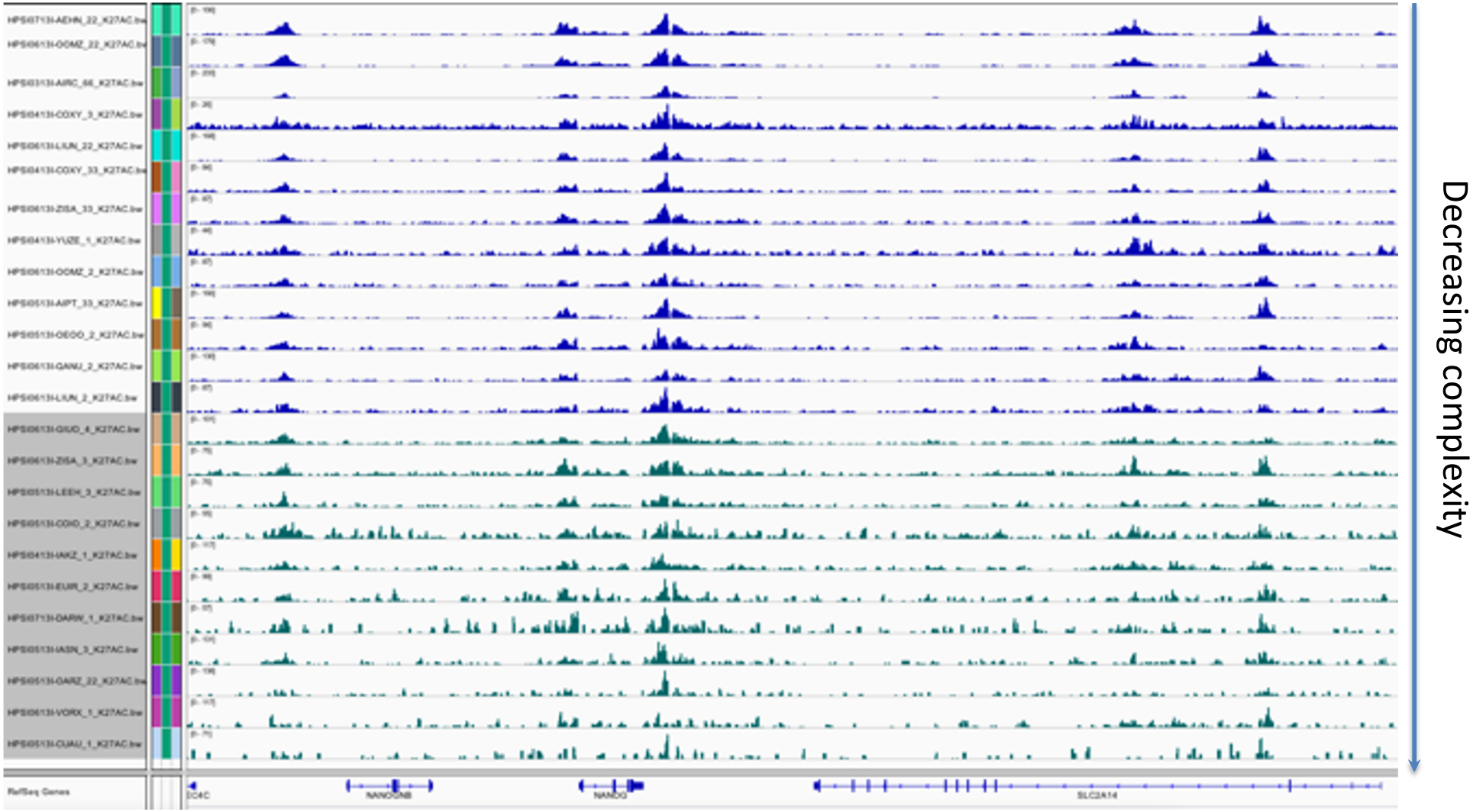
Library complexity

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

Library complexity



Library complexity



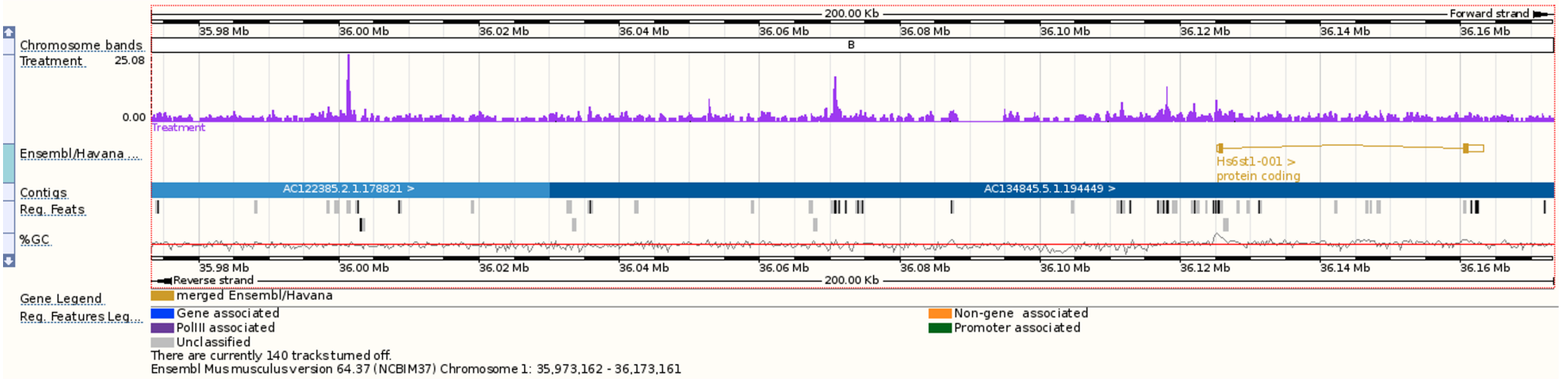
Nonredundant fraction

- # unique fragments positions / total # fragments
- >0.8

Basic analysis of ChIP-seq

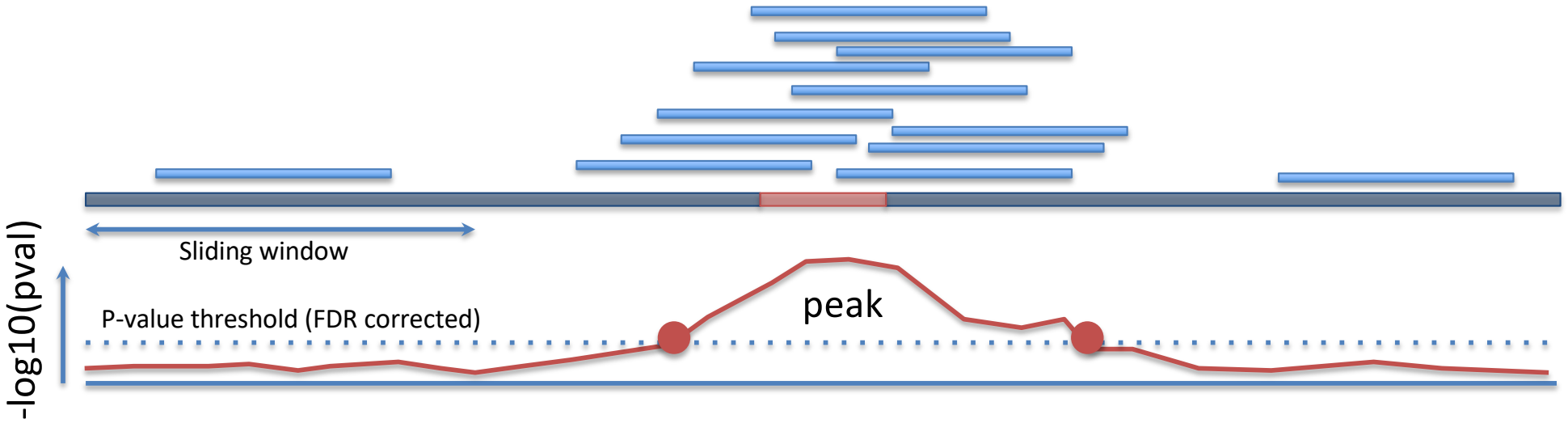
1. Read alignment
2. Visualisation
3. Peak calling
   - Peak annotation (mapping peaks to genes etc)
   - Motif analysis
4. Differential binding
   - Case / control
   - Naïve / stimulated

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

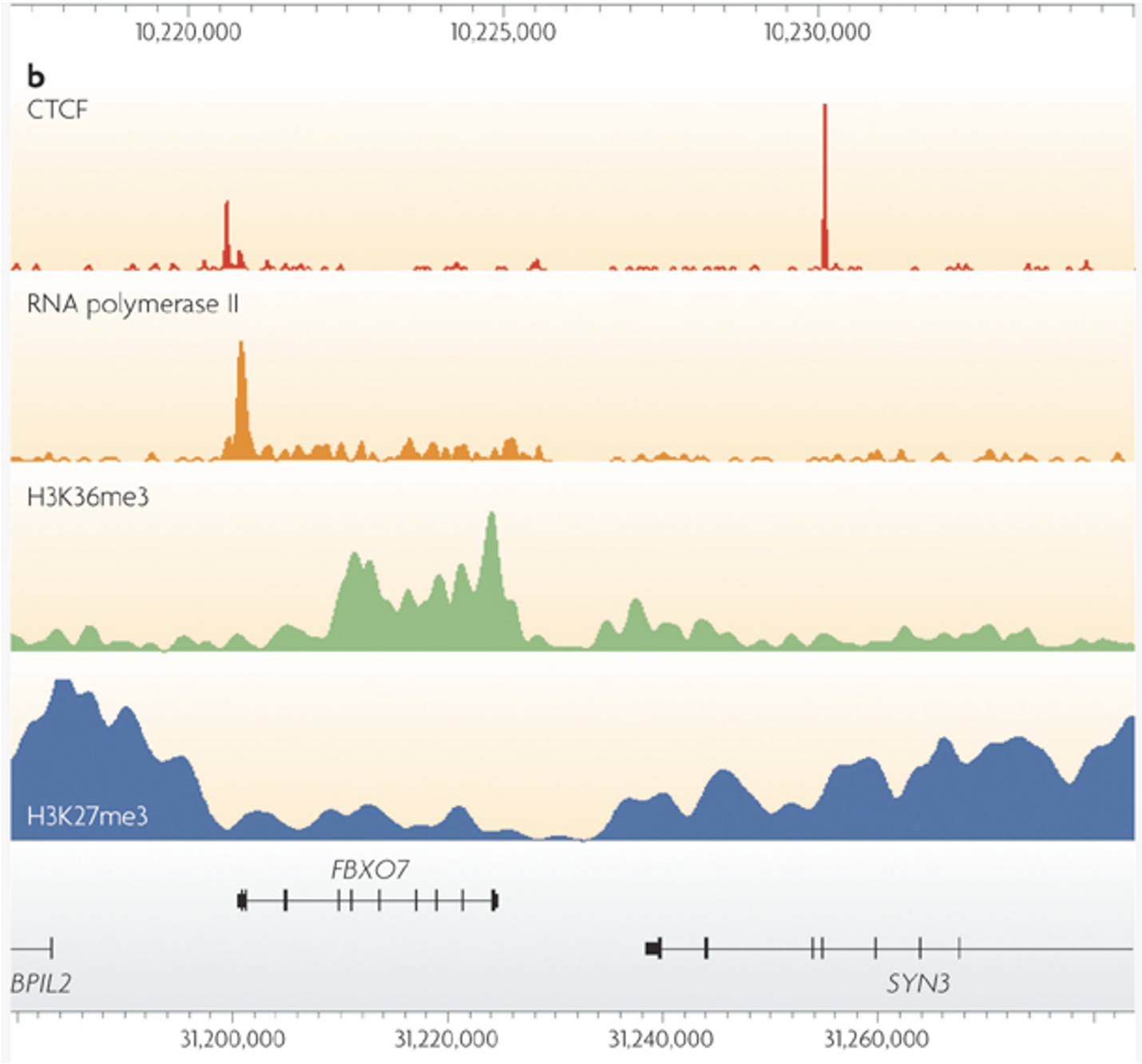
Peak calling



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

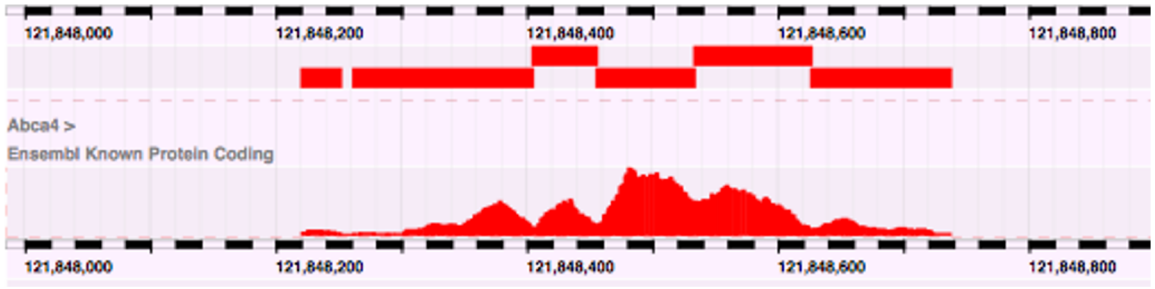
Peak calling challenges

- What's expected?
  - Treatment sample (with antibody)
  - Input 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 analysis

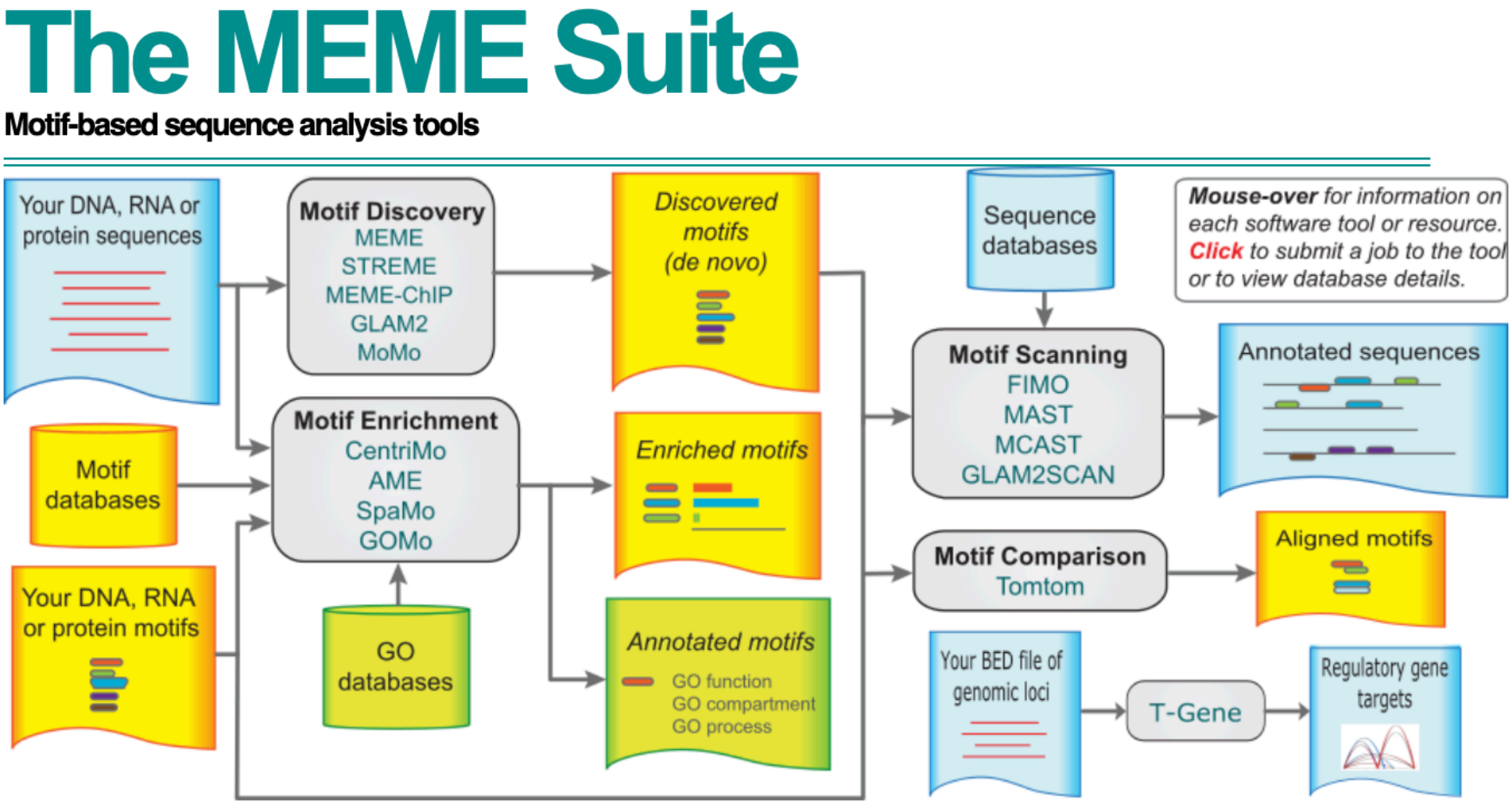


GAATCCCACA TTTGCATAACAAAAG ACTCCTGGTG
CAGCTGCTCT TCTGCATAACAAAGG GTGGCCCTGC
CCGGTTTTTC TTTGCATAACAATAA GATCTGGCTA
TTATTCTCAC TTTGCATAGGAATGG GGCAGTTAGA
CACAGCCACA TTTGCATAACAGAAG CCGAGCCCGC
CTTGGGTGAA TTTGCAAGACAAAGG ACAATGATCA

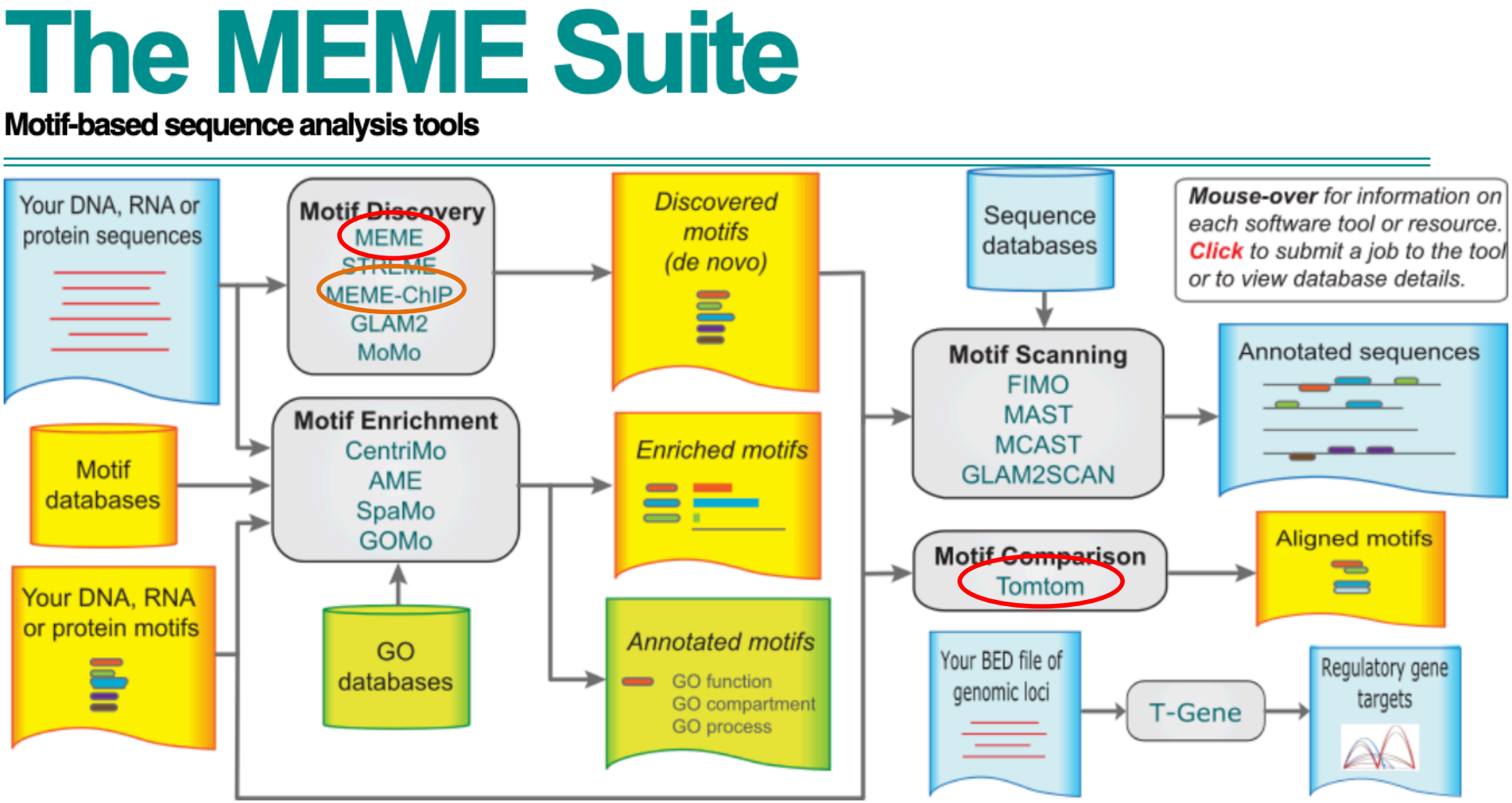
Align sequences from multiple peaks

Discover motifs

Motif analysis tools



Motif analysis tools



Motif analysis tools

HOMER

Software for motif discovery and next-gen sequencing analysis

---

Homer *de novo* Motif Results

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 1351
Total background sequences = 19001
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-116	-2.684e+02	55.29%	24.70%	1843.3bp (1700.6bp)	hsa-miR-206 MIMAT0000462 Homo sapiens miR-206 Targets (miRBase) More Information | Similar Motifs Found
2		1e-41	-9.637e+01	68.39%	49.40%	1877.0bp (1667.5bp)	hsa-miR-485-3p MIMAT0002176 Homo sapiens miR-485-3p Targets (mi More Information | Similar Motifs Found
3		1e-28	-6.530e+01	60.92%	45.25%	1651.5bp (1428.1bp)	hsa-miR-181a* MIMAT0000270 Homo sapiens miR-181a* Targets (miR More Information | Similar Motifs Found

Motif analysis tools