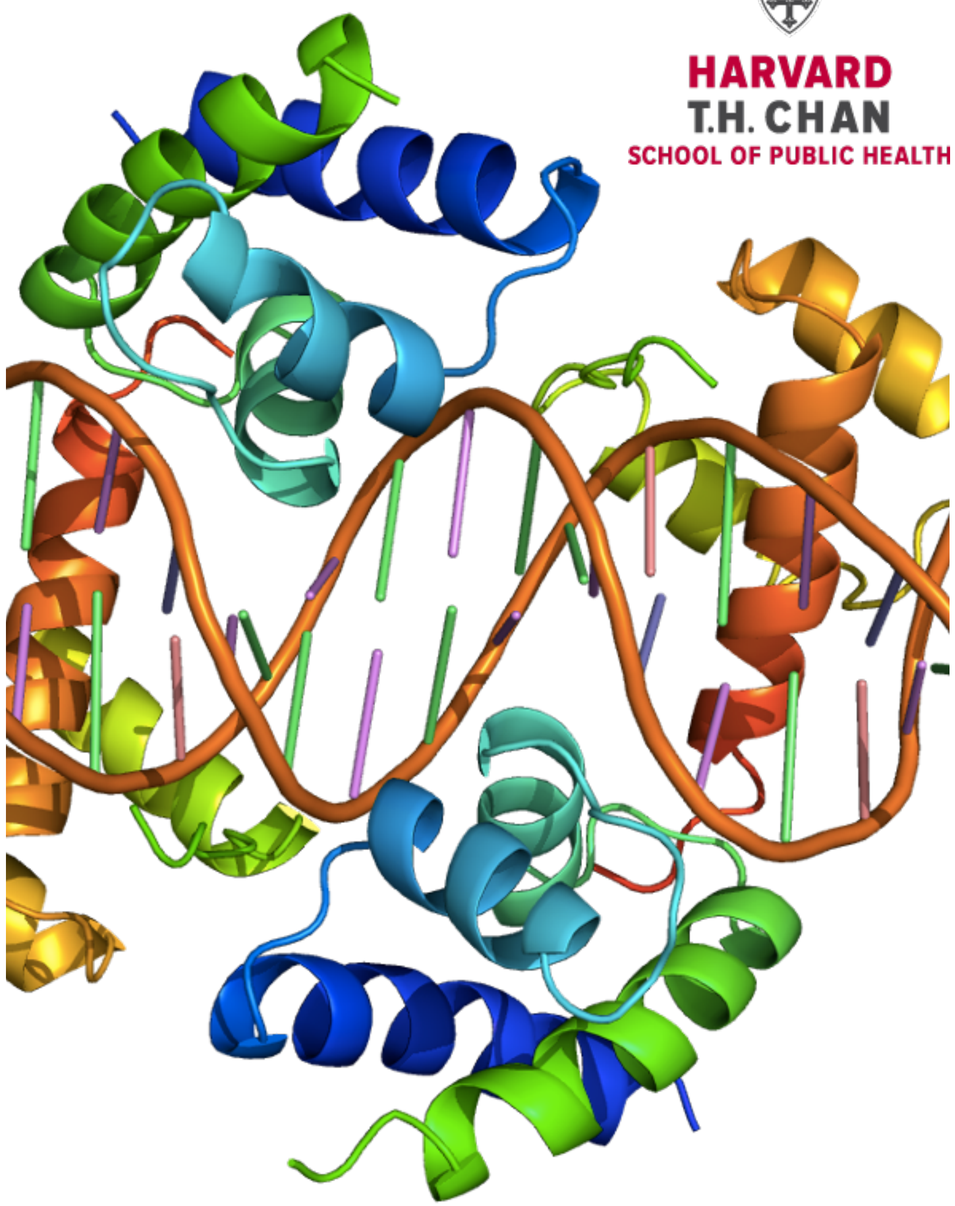
Understanding chromatin biology using high throughput sequencing (HTS)

HSPH Bioinformatics Core

Shannan Ho Sui

August 11, 2023

Slides in collaboration with Dr. Meeta Mistry

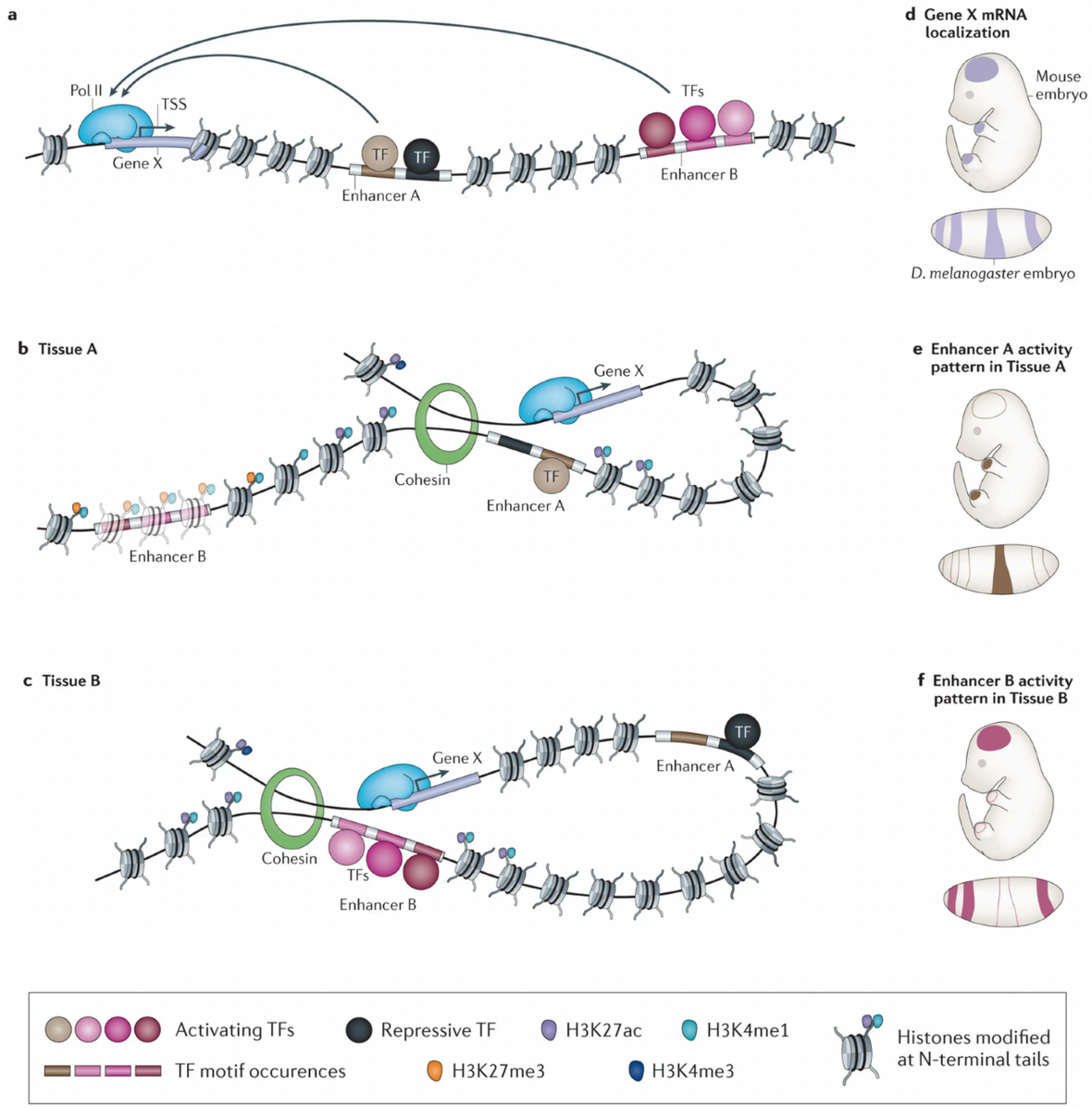


HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Complexity in gene regulation

Diverse mechanisms to ensure that genes are expressed at the right time, in appropriate tissues and under specific conditions

Numerous diseases associated with mutations in the non-coding genome

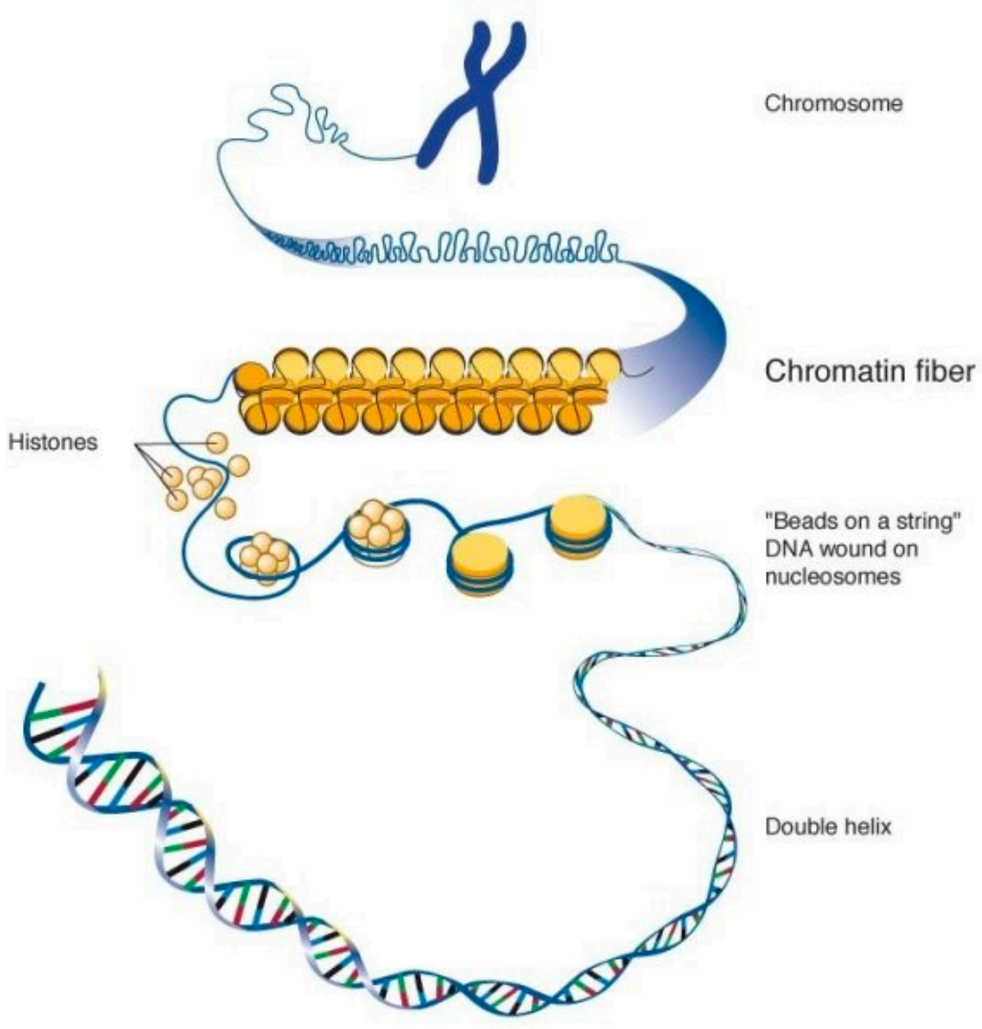


Nature Reviews | Genetics

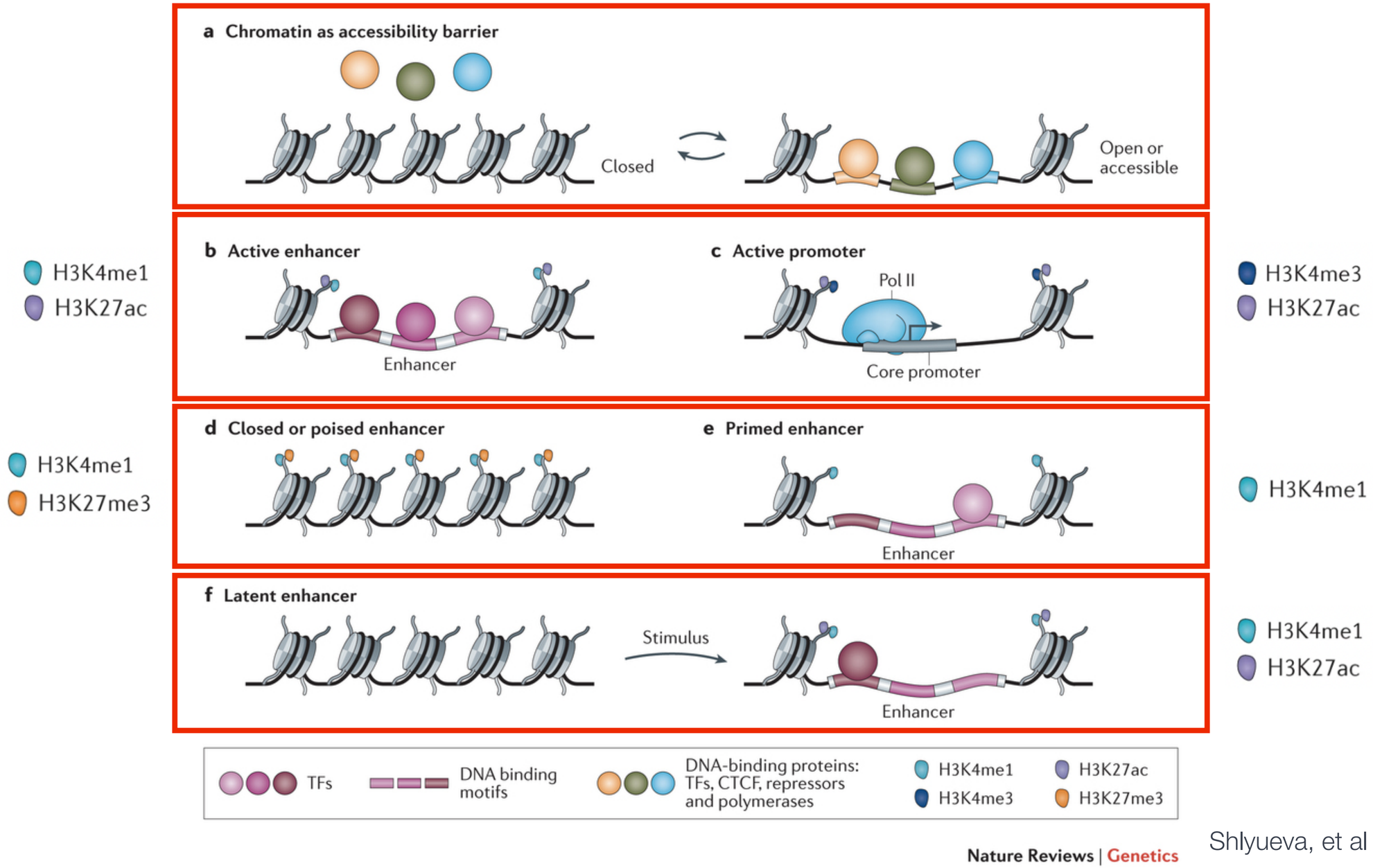
Shlyueva, et al (2014). Transcriptional enhancers: from properties to genome-wide predictions.

What is chromatin?

- **Chromatin:** a mixture of DNA and proteins that form the chromosomes found in the cells of humans and other higher organisms
- **Nucleosome:** 147 bp of DNA wound around 8 histone proteins (octamer) consisting of 2 copies each of the core histones (H2A, H2B, H3, H4)
- **Heterochromatin:** condensed chromatin
- **Euchromatin:** extended chromatin



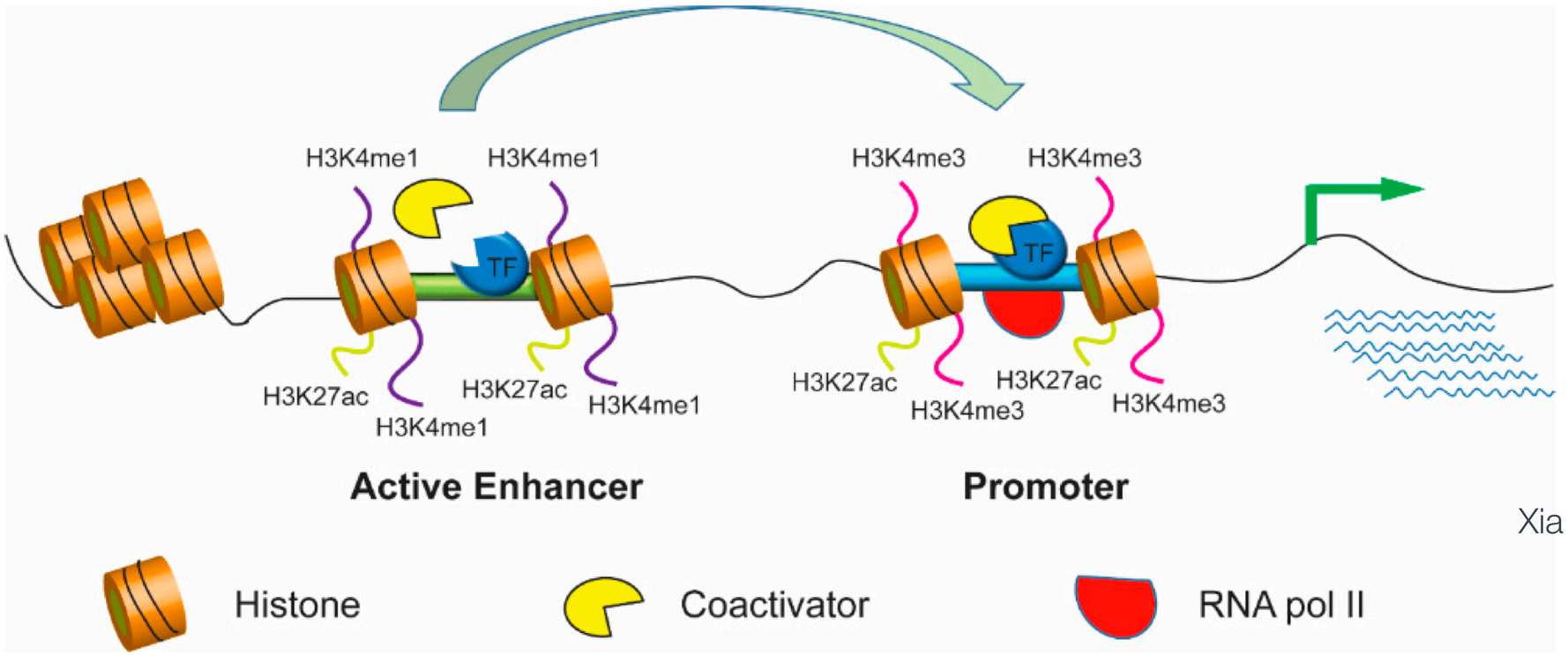
https://www.creative-diagnostics.com/blog/index.php/the-structure-and-function-of-chromatin/



Shlyueva, et al (2014)

Chromatin structure determines if a gene is expressed or not

How do enhancers, repressors and cofactors regulate transcription?

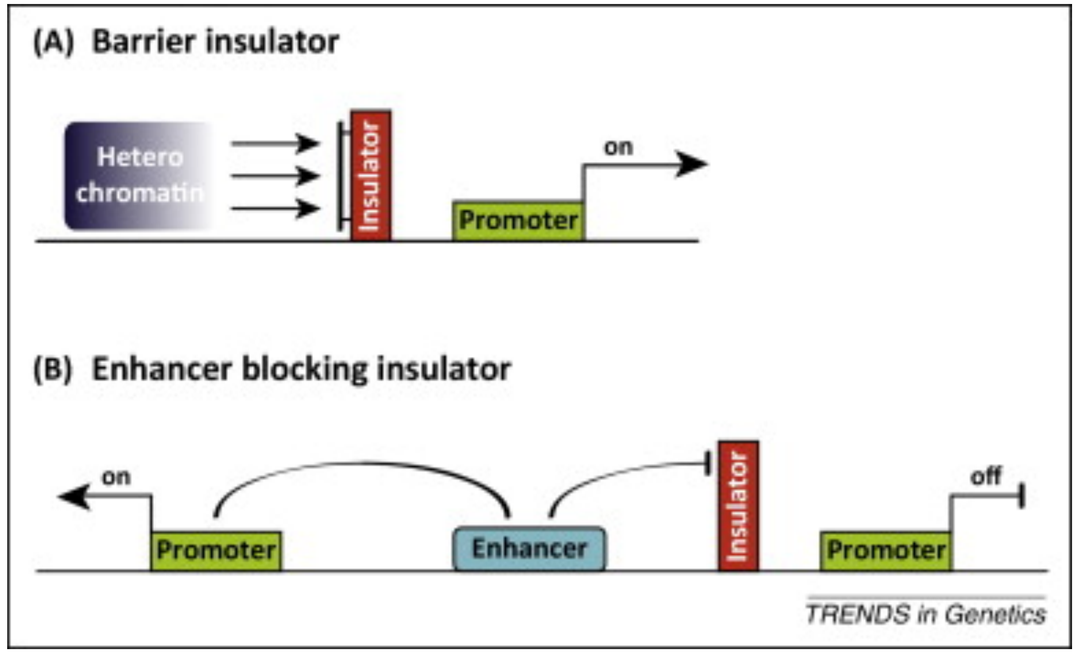


Xia & Wei, Cells, 2019

- Enhancers are DNA regulatory elements that activate transcription to a higher level
- Operate from a distance by forming chromatin loops that bring the enhancer and target gene into proximity
- Silencers reduce transcription from their target promoters
- Cofactors do not bind DNA directly but mediate protein-protein interactions between TFs and the basal transcriptional machinery

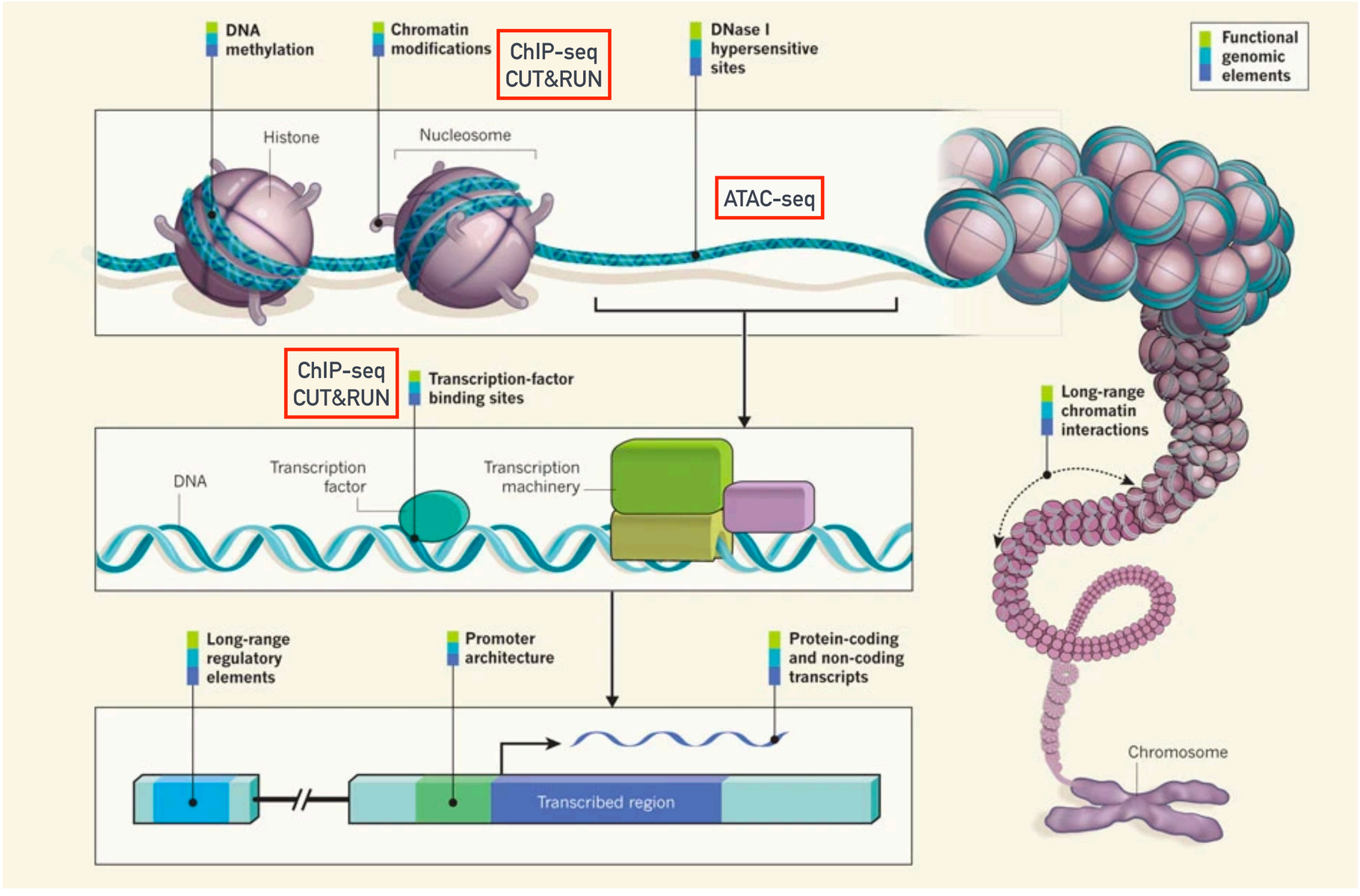
What are insulators?

- Long range regulatory elements
- Block enhancers and silencers from improperly activating or repressing non-cognate promoters
- Barrier insulators prevent silencing of euchromatin by the spread of neighboring heterochromatin
- Enhancer-blocking insulators prevent distal enhancers from acting on promoters of neighboring genes
- Challenging to find based on chromatin features



Trends in Genetics 2014 30161-171

Identifying functional regulatory elements



Ecker, J., Bickmore, W., Barroso, I. et al. ENCODE explained. Nature 489, 52–54 (2012). https://doi.org/10.1038/489052a

The ChIP-seq assay

- Assay genome wide binding of protein to DNA
- Uses a combination of chromatin immunoprecipitation and sequencing
- Identifies how transcription factors and histone modifiers interact with DNA *in vivo*
- Complements DNA accessibility studies and gene expression profiling
- Gain an understanding of gene regulation

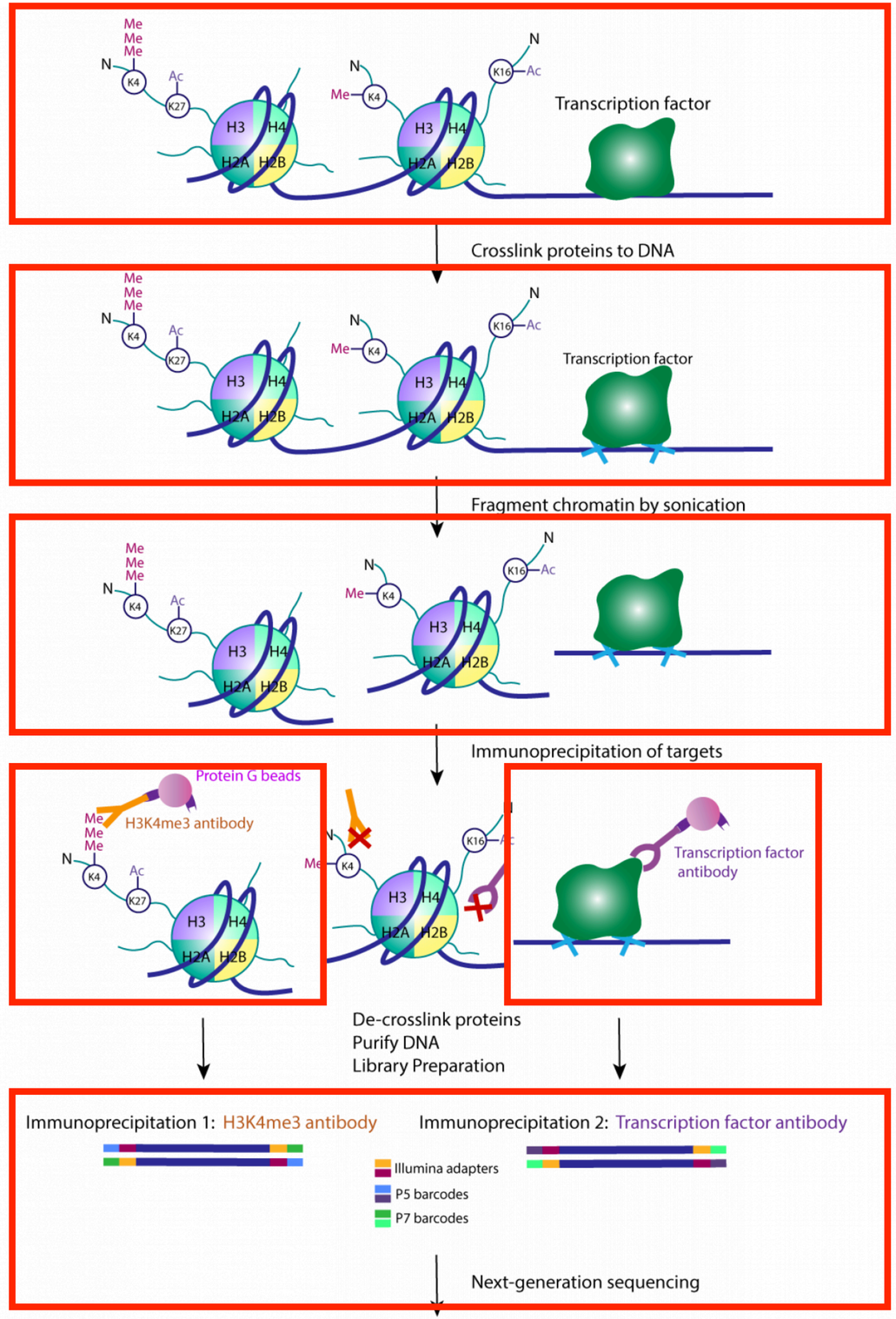
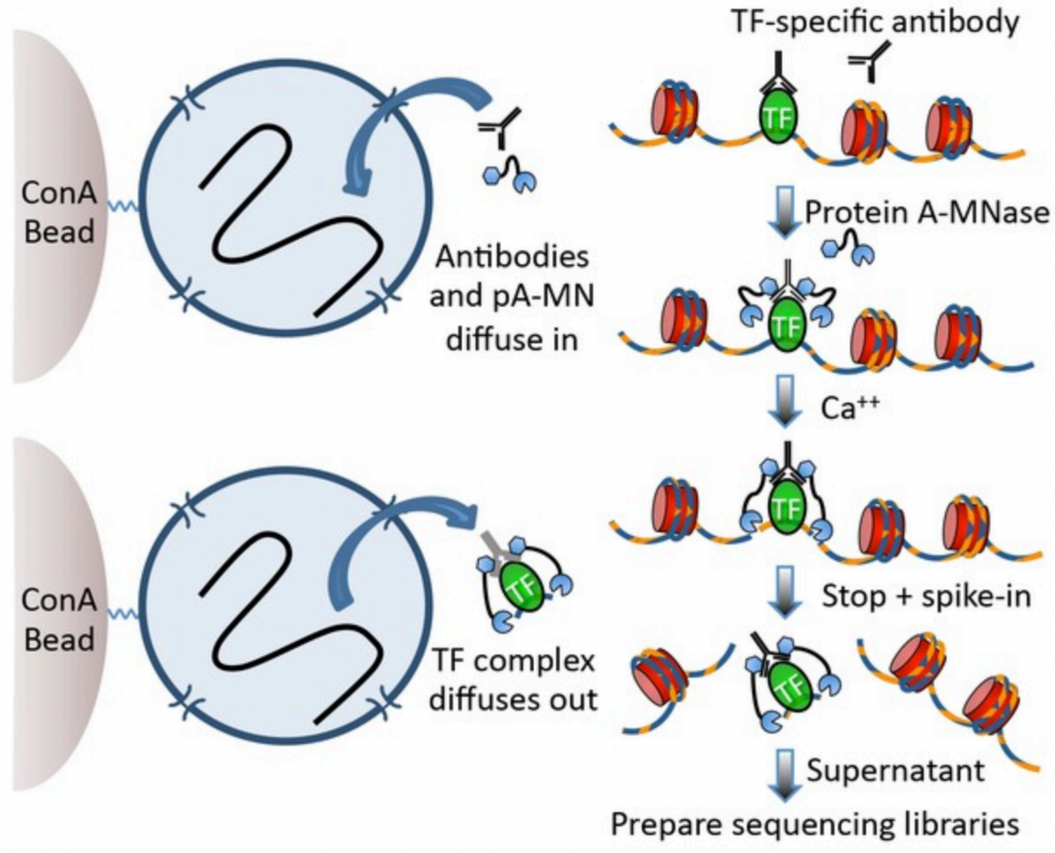


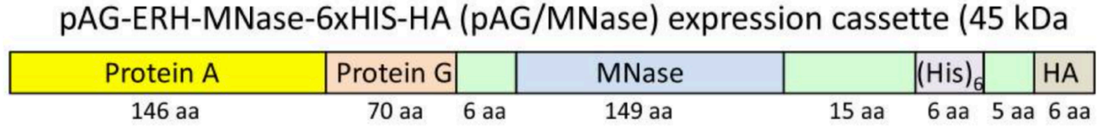
Image: https://brcf.medicine.umich.edu/cores/epigenomics/products-and-services/chip-seq/

Cleavage Under Targets and Release Using Nuclease (CUT&RUN) assay

- Also to assay genome wide binding of protein to DNA
- Combines antibody-targeted controlled cleavage by micrococcal nuclease with sequencing
- Cells are bound to beads and then permeabilized to allow antibodies and pAG-MN to diffuse in
- Antibodies bind to DNA, followed by binding of pAG and activation of MNase with Ca++
- Spike-in added with stop buffer
- Requires fewer cells, lower read depth and is an easier assay to perform

Skene & Henikoff (2017), *eLife* **6**:e21856.

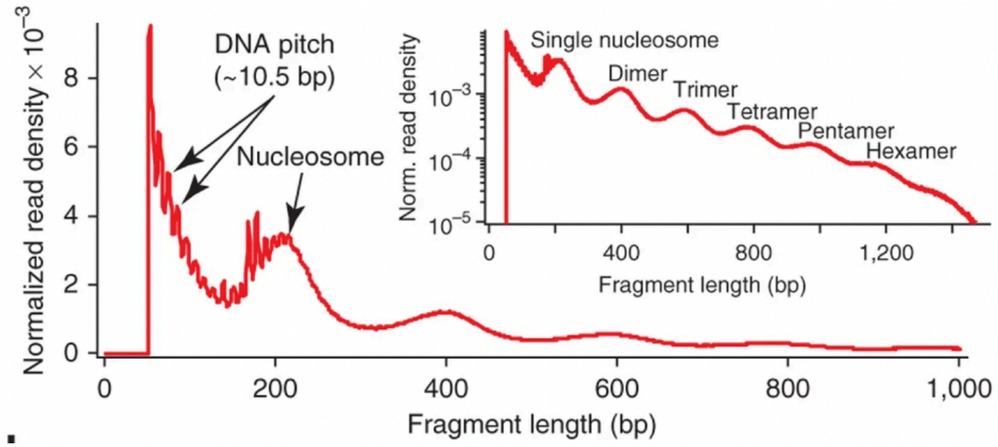


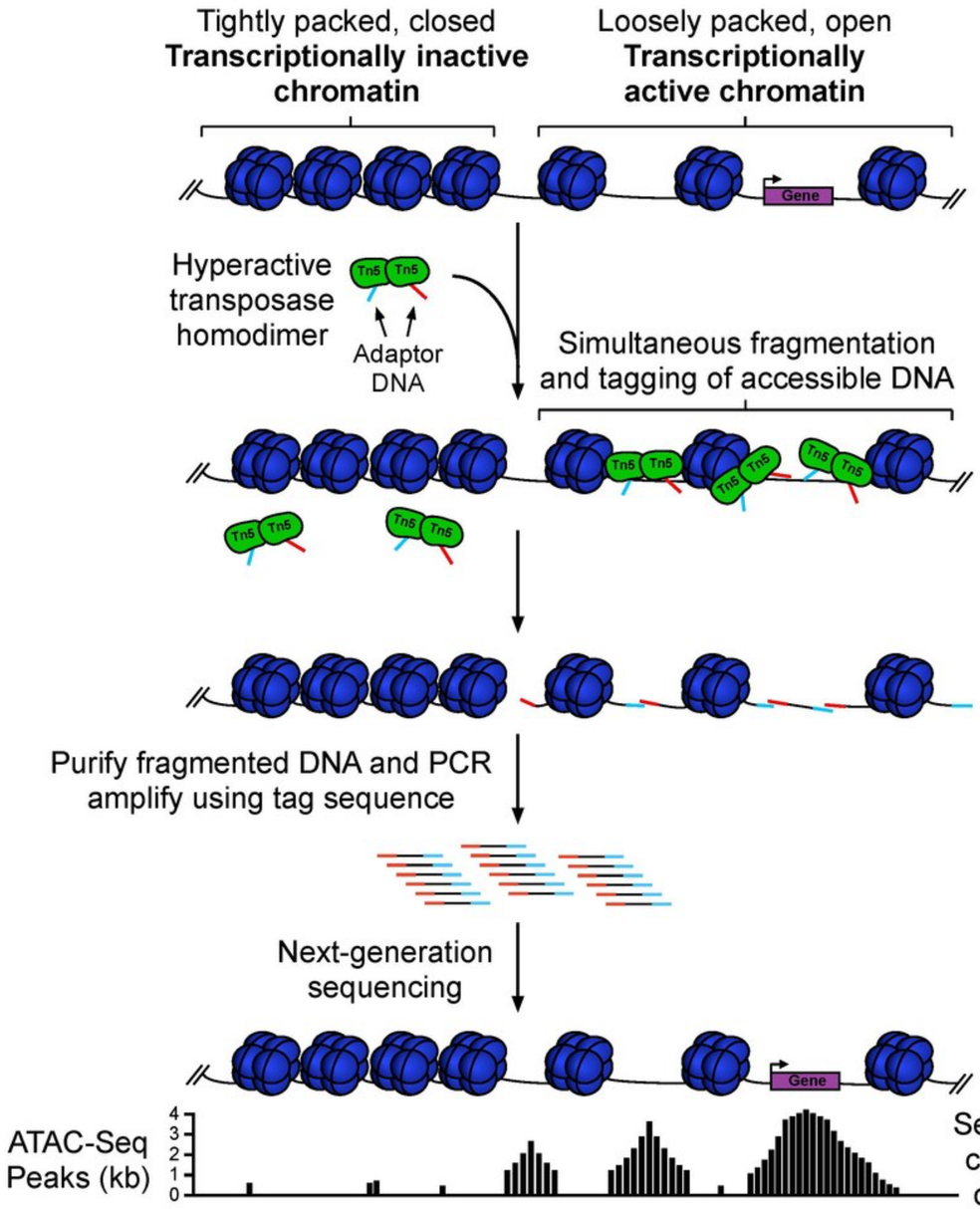


Meers et al. (2019), *eLife* **8**:e46314.

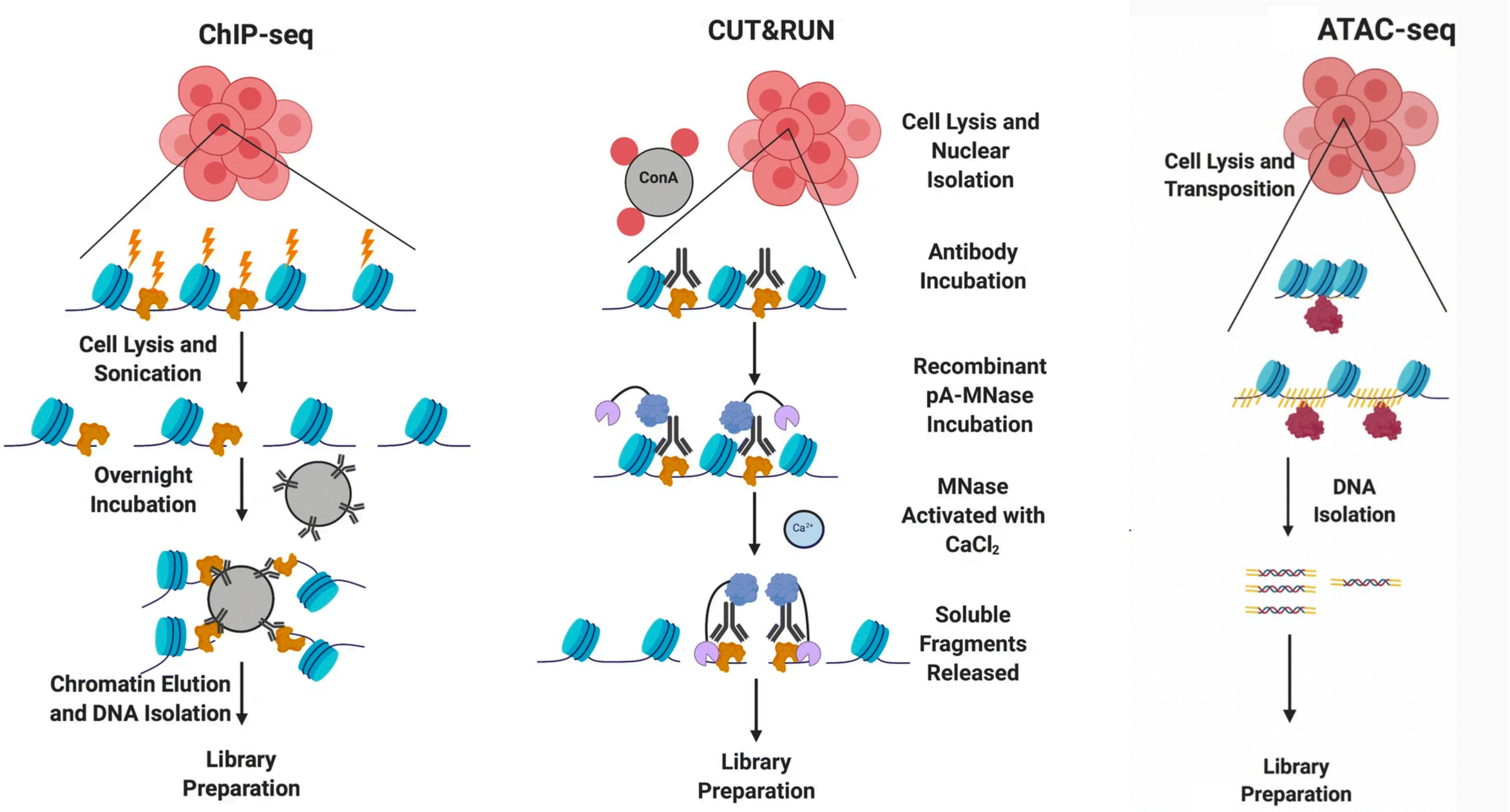
Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq)

- Measure the extent to which DNA is open and accessible genome-wide
- Uses a hyperactive Tn5 transposase that cuts and inserts sequencing adapters into regions of chromatin that are accessible
- Fragment length correlates with nucleosome-free regions (less than 147bp) and mono-, di- and trinucleosome regions





https://seandavi.github.io/AtacSeqWorkshop/articles/Workflow.html



Adapted from Klein & Hainer (2020). *Chromosome Res* **28**, 69–85. https://doi.org/10.1007/s10577-019-09619-9

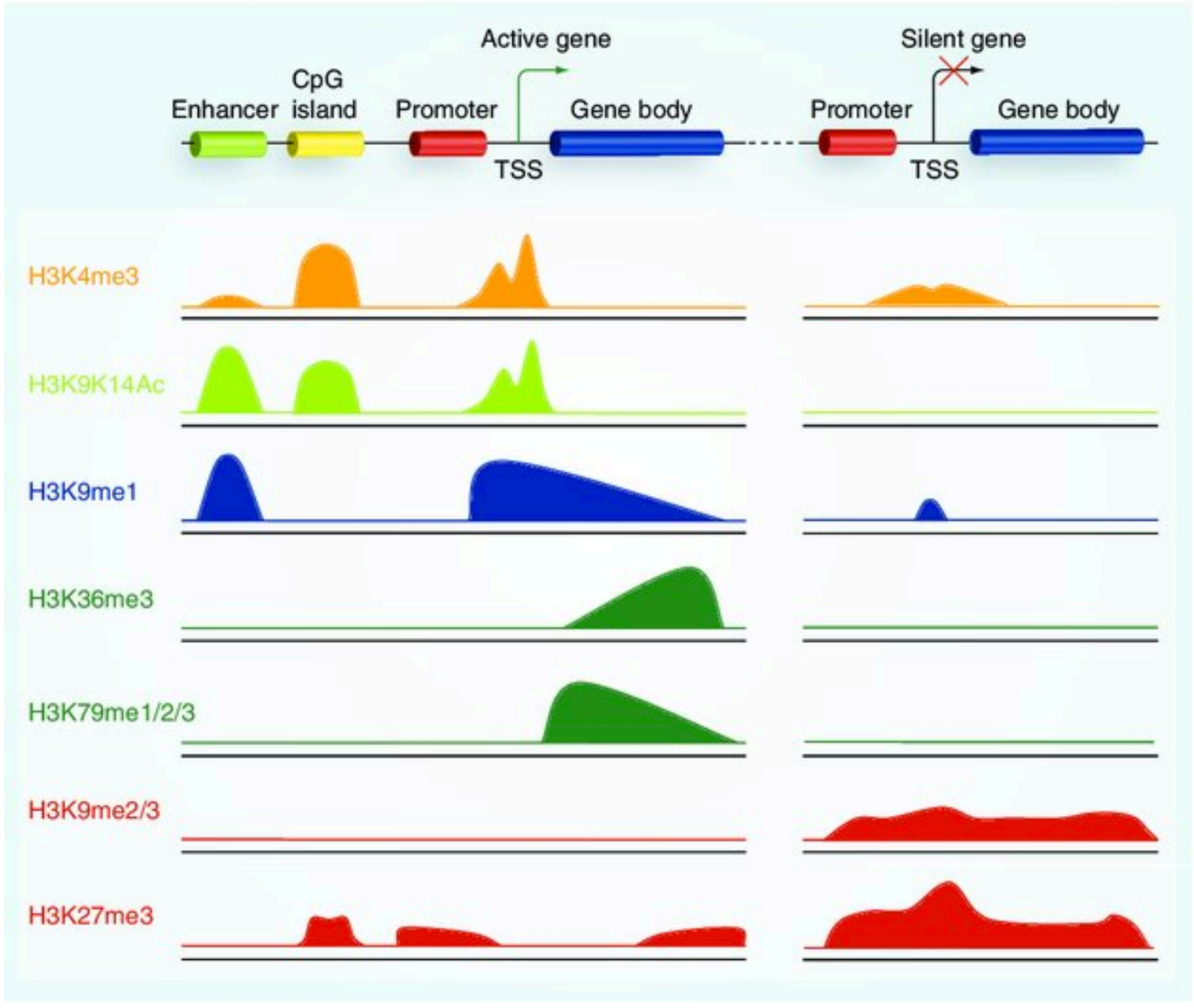
Profiling histone modifications

<u>Active promoters</u>
H3K4me3, H3K9Ac

<u>Active enhancers</u>
H3K27Ac, H3K4me1

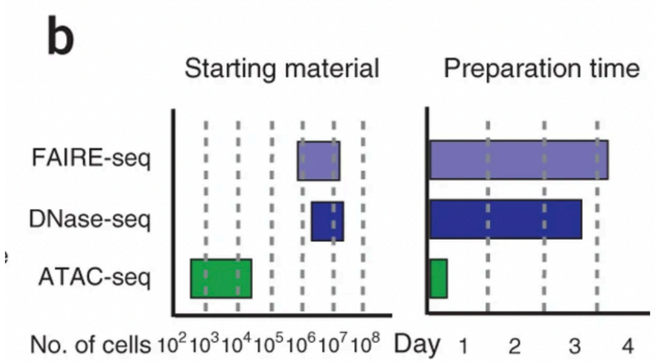
<u>Repressors</u>
H3K9me3, H3K27me3

<u>Transcribed gene bodies</u>
H3K36me3



Human T-cell ChIP-seq data
(Lim et al, 2010, Epigenomics)

Profiling open chromatin



Adapted from Buenrostro, *et al*. *Nat Methods* **10**, 1213–1218 (2013). https://doi.org/10.1038/nmeth.2688

What does it give us?

- Multiple aspects of chromatin architecture simultaneously at high resolution.
  - Maps open chromatin
  - TF occupancy
  - nucleosome occupancy

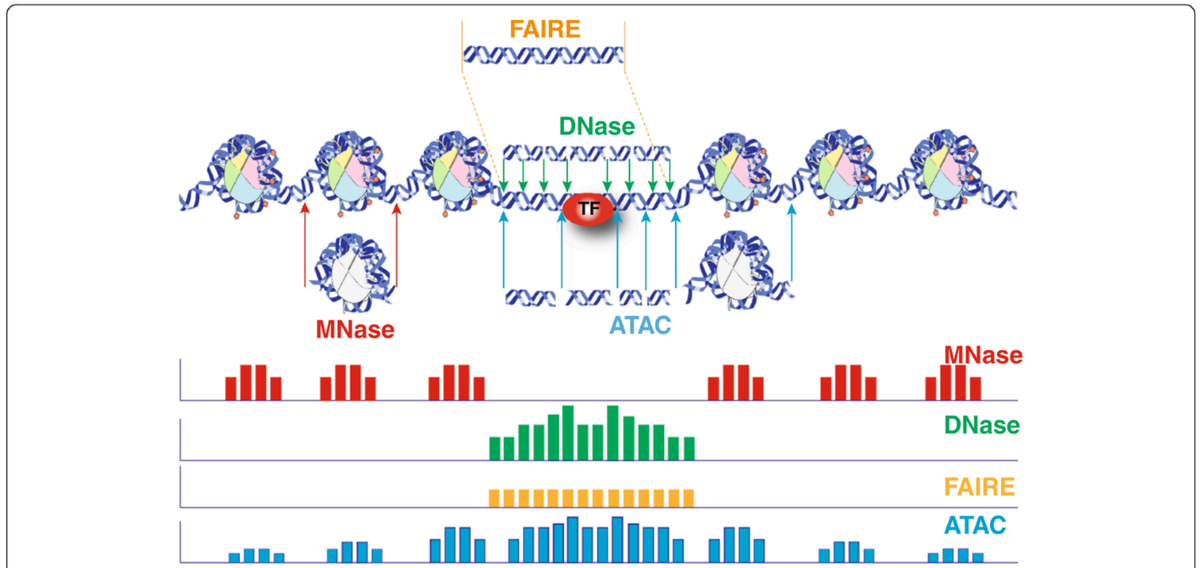
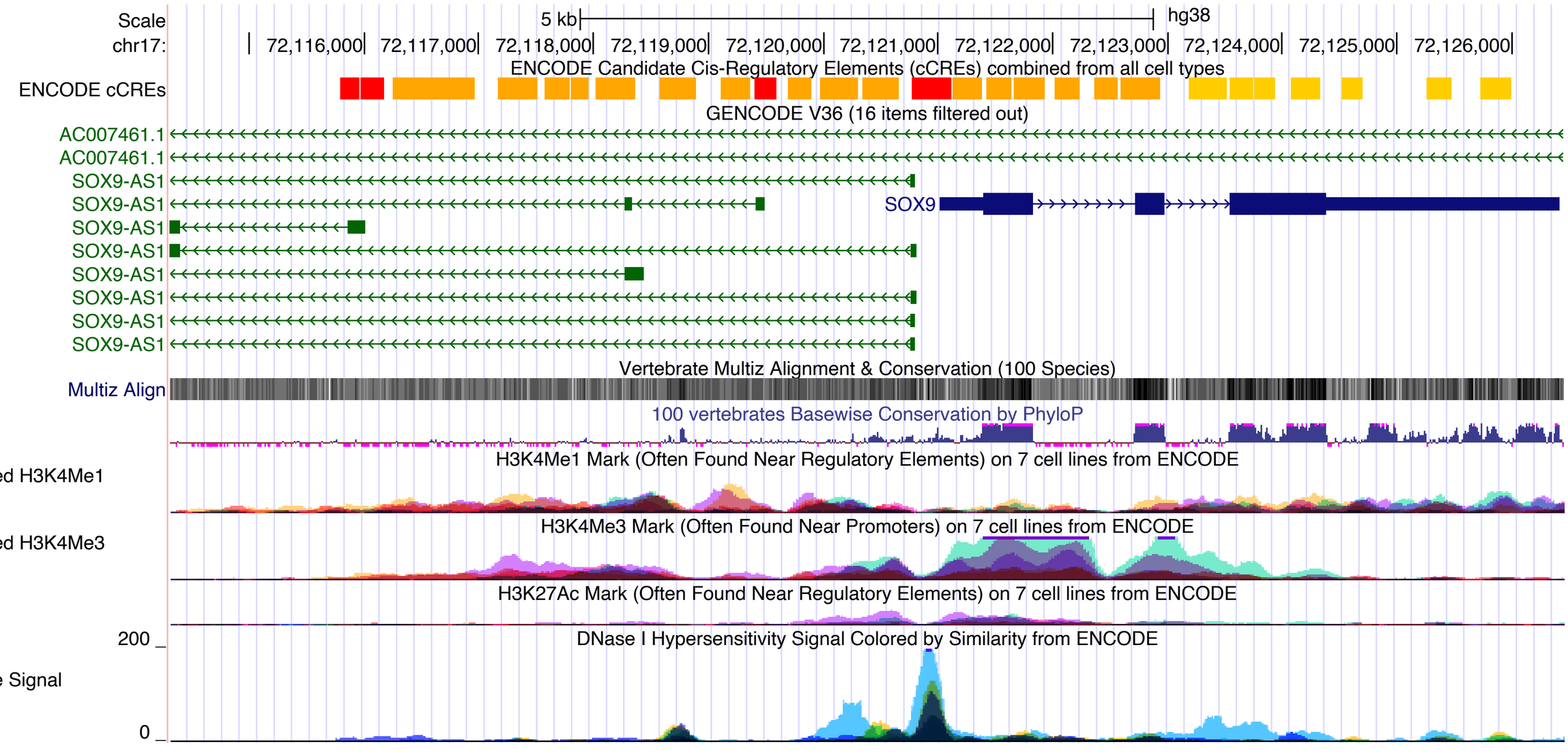


Figure 1 Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each assay across the entire region. The footprint created by a transcription factor (TF) is shown for ATAC-seq and DNase-seq experiments.

Tsompana and Buck, 2014

Slide by Meeta Mistry

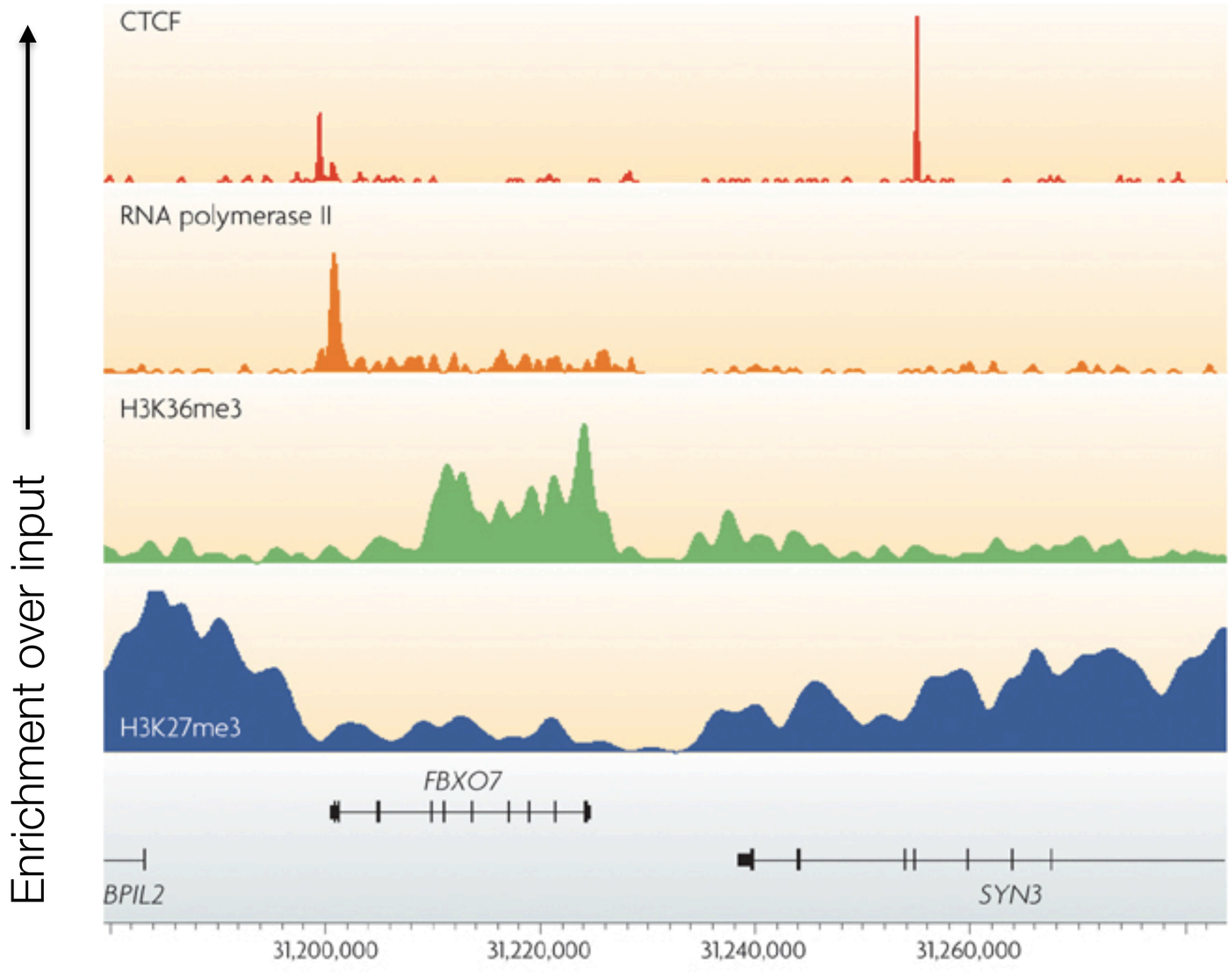
Visualizing peaks in the UCSC genome browser



Types of signals

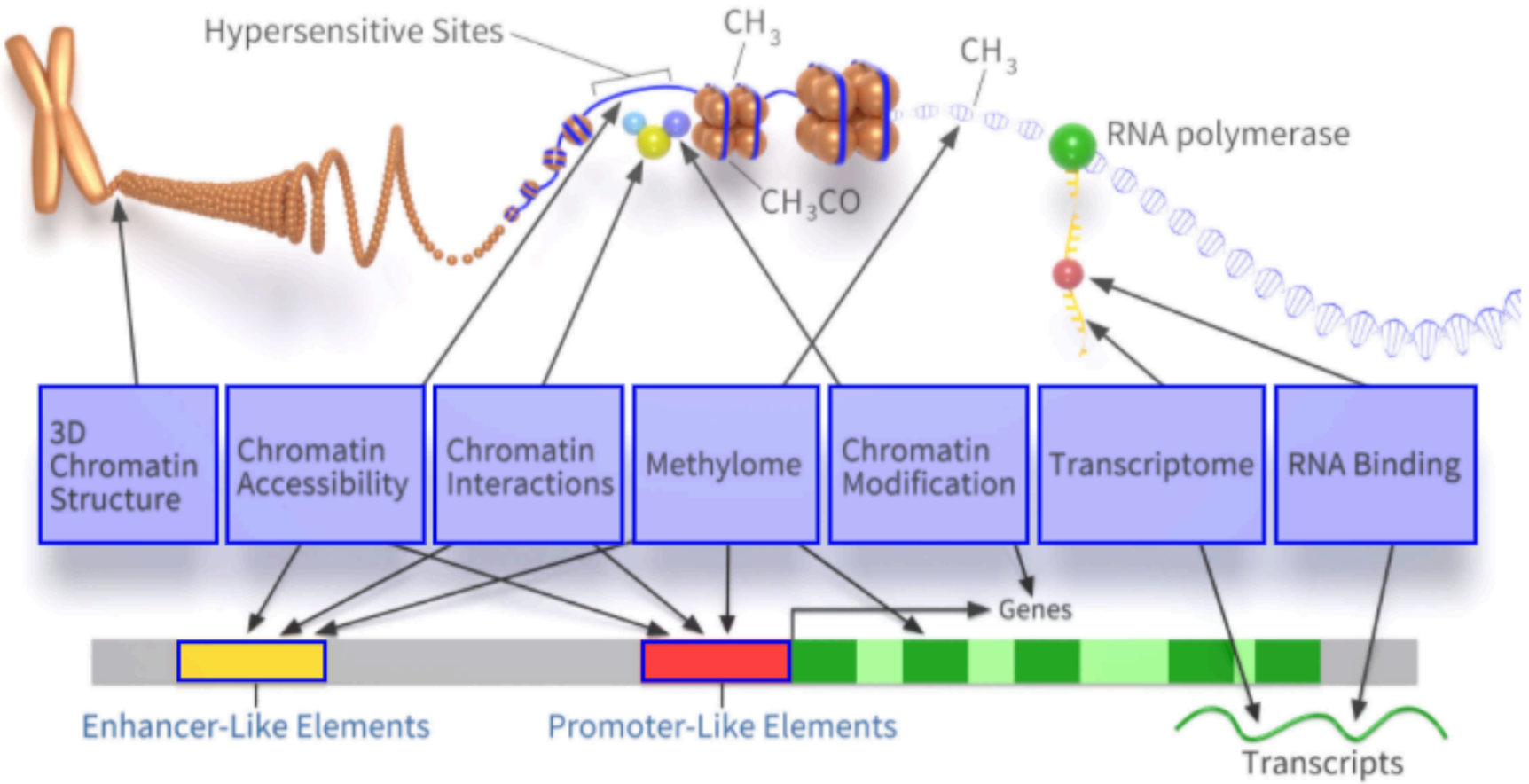
Transcription factors

Histone modifications



Adapted from Park (2009). Nature Reviews Genetics.

ENCODE: Encyclopedia of DNA Elements



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI) [encodeproject.org]

Features	ChIP-Seq	CUTANA™ CUT&RUN	CUTANA™ CUT&Tag
Sample Input	Sheared Chromatin	Cells OR nuclei	Nuclei (recommended)
Typical Required Cell #	> 1 Million	500K	100K
Ideal Targets	Histone PTMs & chromatin-interacting proteins	Histone PTMs & chromatin-interacting proteins, including remodelers	Histone PTMs & select validated targets
Secondary Antibody	No	No	Yes
Library Preparation	Yes	Yes	No (Direct to PCR)
Protocol Time (Cells → NGS libraries)	~ 1 week	2 days (can be automated)	2 days (can be automated)
Sequencing Depth	> 30 million	3-5 million	3-5 million
Signal : Noise	Low	High	High
Experimental Throughput	Low	High	High

Comparison of ChIP-seq, CUT&RUN and CUT&TAG

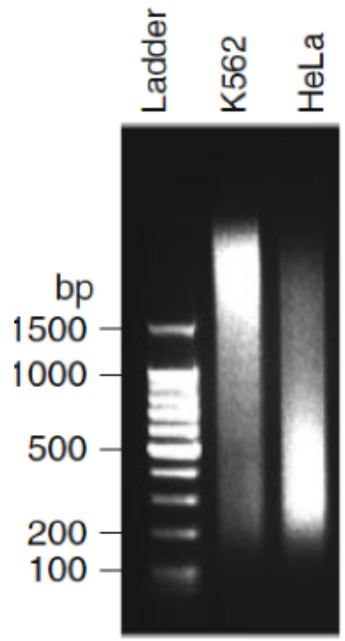
https://www.epicypher.com/resources/blog/cut-and-run-vs-cut-and-tag-which-one-is-right-for-you/

Parameters for a successful ChIP-seq

- Efficient and specific antibody
  - Antibody may work for ChIP-seq yet fail in CUT&RUN because it is in its native form, not fixed
- Amount of starting material
  - ChIP DNA yield depends on various factors
    - Cell type in question
    - Abundance of the mark or protein (histones have high binding coverage than TFs)
    - Antibody quality

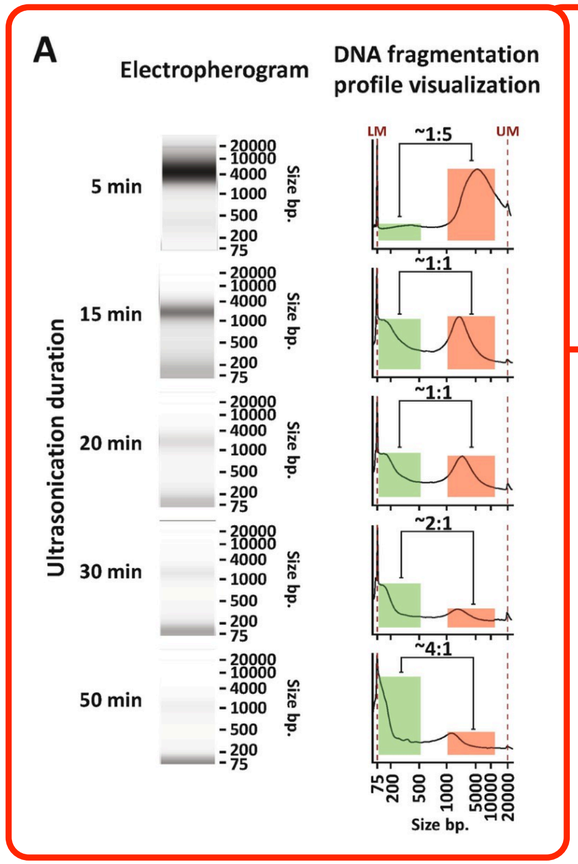
Parameters for a successful ChIP

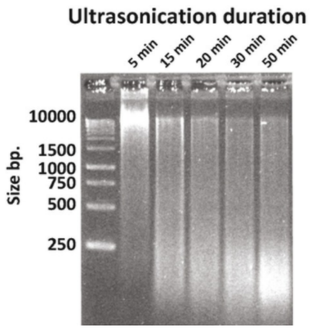
- Chromatin fragmentation
  - Size matters (not too big and not too small)
  - Can vary between cell types
- Stringency of washes

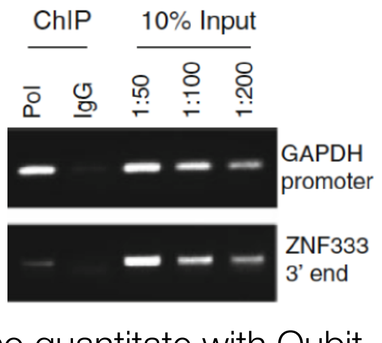


Fragments too big:
Reduced signal to noise ratio in ChIP-seq

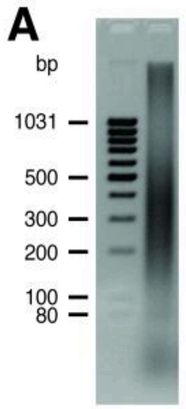
Oversonication:
Fragmentation biased towards promoter regions causes ChIP-seq enrichments at promoters in both, ChIP AND control (input) sample

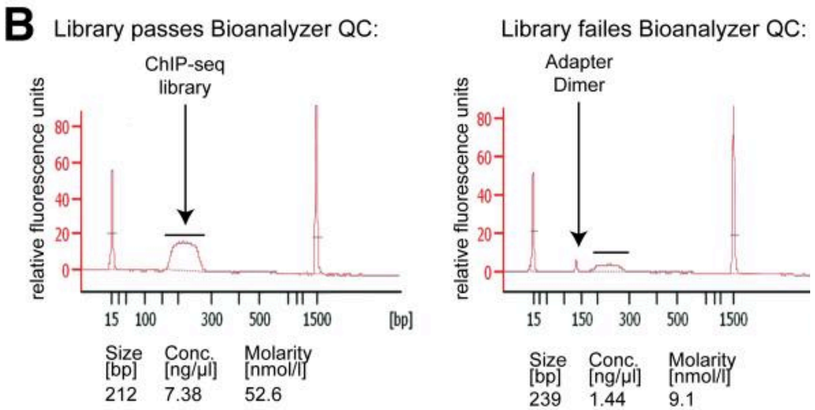


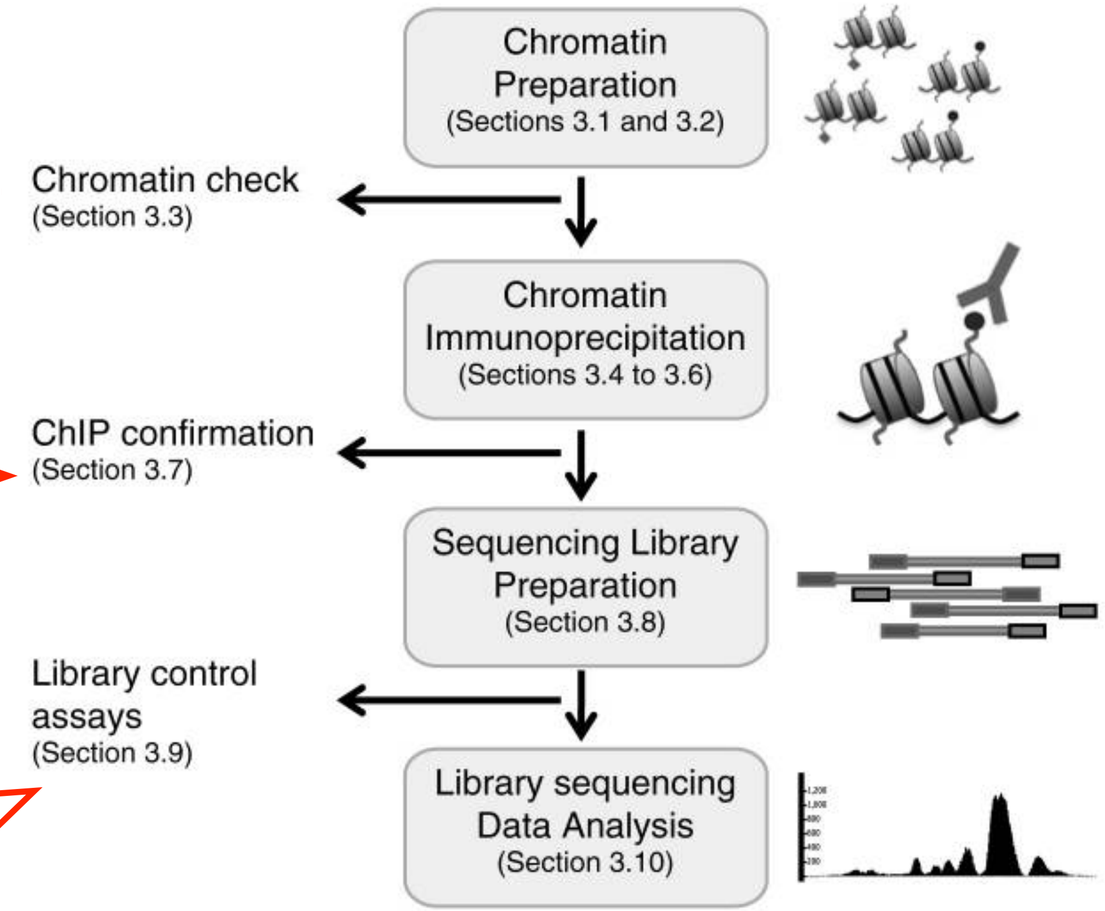




Also quantitate with Qubit (ChIP) / NanoDrop (WCE)







Quality check points

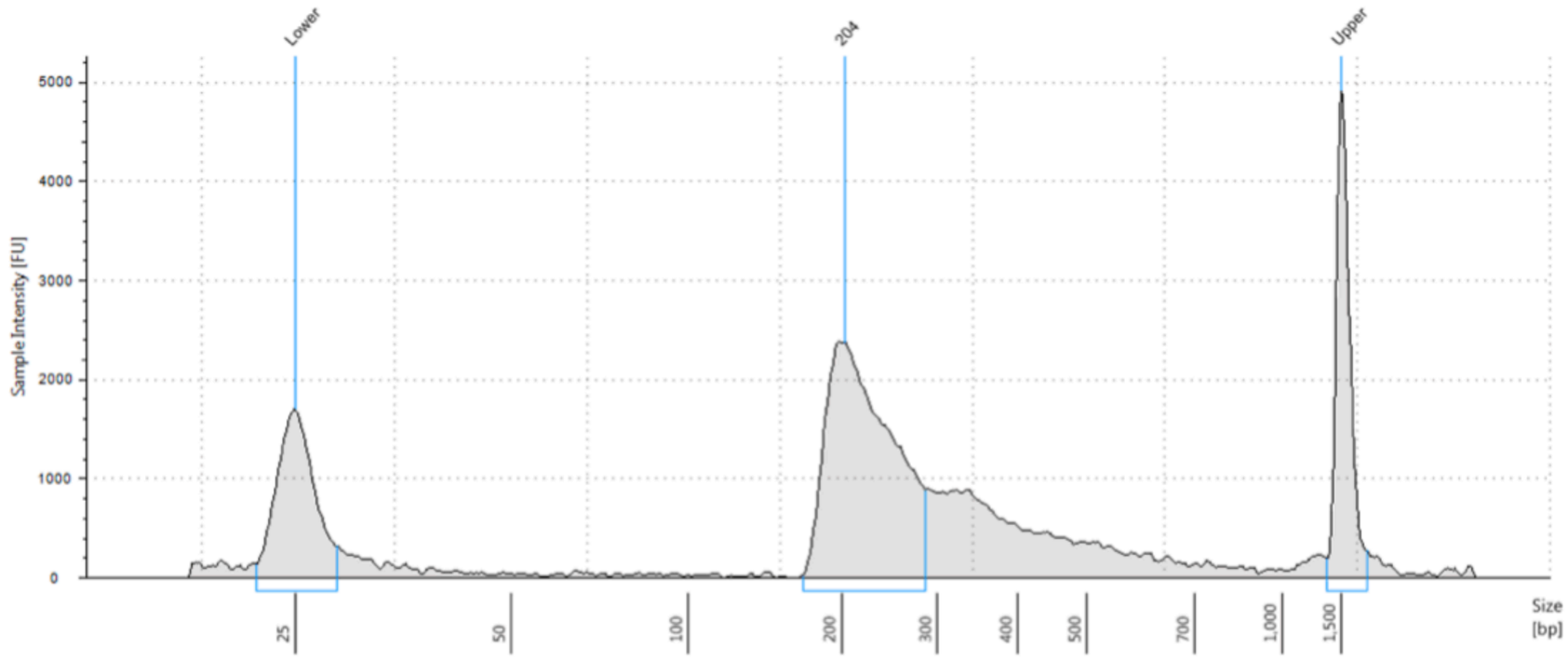
O'Geen et al (2011), Methods Mol Biol
Schmidt et al (2009), Methods
Larsen et al (2018), Two-Hybrid Systems. Methods Mol Bology, v 1794

Maximizing success

	ChIP-seq	CUT&RUN	ATAC-seq
Number of cells	1-10 million	200-500K for TFs * Can use fewer for histone marks (>5000)	50-500K
Antibody QC	Western blot	Western blot	N/A
IP DNA	> 10 ng	> 1 ng	> 3 ng
IP QC	qPCR	qPCR with custom primers	N/A
Library QC	Tape station / BioAnalyzer	Tape station / BioAnalyzer	Tape station / BioAnalyzer
Negative control	Input DNA or IgG	Non-specific IgG	N/A
Positive control	H3K4me3 or known protein	H3K4me3	N/A
Replicates	3	3	3

* Check cell count before and after bead purification
  Check >90% of cells permeabilized (cell counter or hemocytometer)

Library QC for ATAC-seq



Library fragments contain the original DNA insert (< 90bp) + 135 bp from the adapters on each end. This creates library fragments starting at around 200 bp which then increase to around 1000 bp. Because of the periodicity of neighboring nucleosomes, fragments pile up with peaks between 160-200 bp apart. Important to see a good spread of fragments over the range between 200-1000 bp, with the majority under 600 bp.

https://www.activemotif.com/blog-library-qc

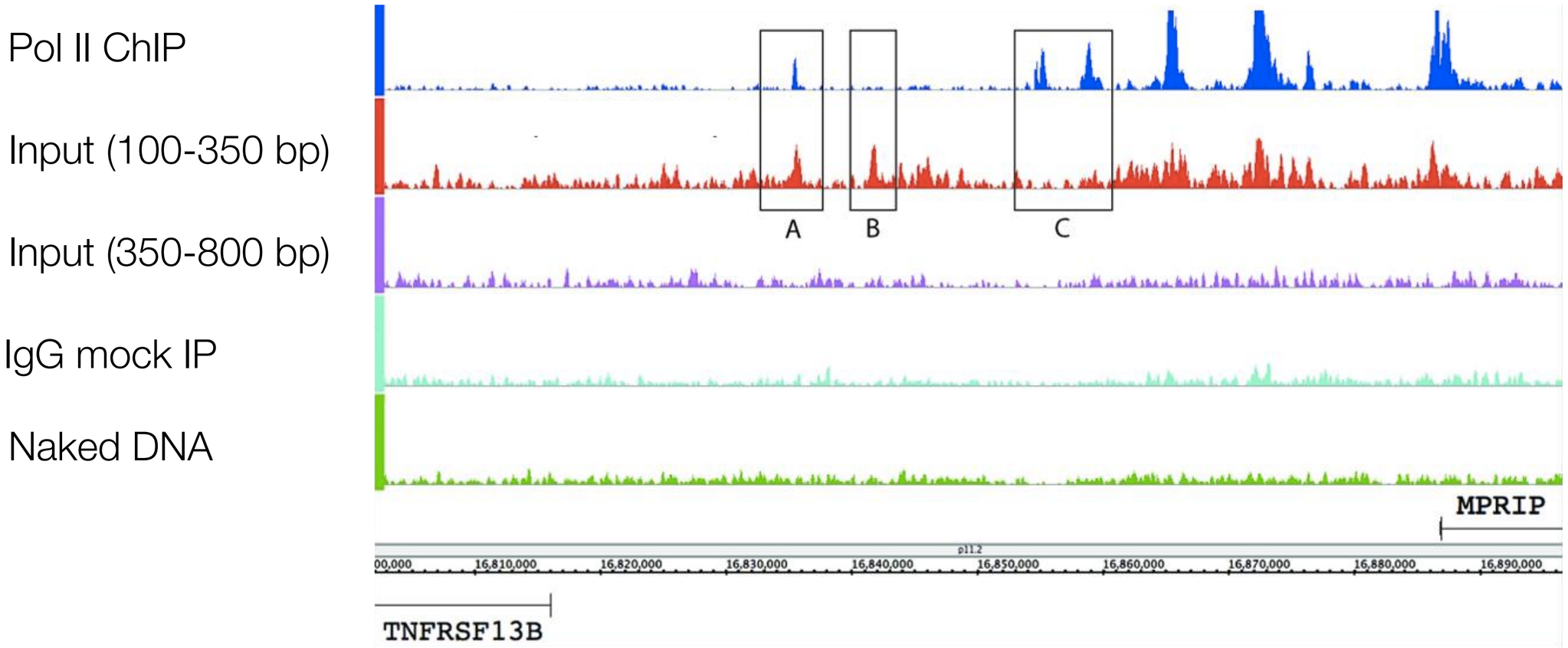
Controls

- ChIP-seq assays require input controls
- CUT&RUN uses a non-specific IgG control
  - Also recommend using a H3K4me3 positive control
- Controls are not typically used for ATAC-seq
  - Expensive and of limited value. A control for a given sample would be genomic DNA from the sample that, instead of transposase treatment, is fragmented (e.g. by sonication), has adapters ligated, and is sequenced along with the ATAC sample.

Why are ChIP-seq controls necessary?

It allows us to compare with the same region in a matched control and identify a the presence of artefacts that tend to generate false positive peaks. These include:

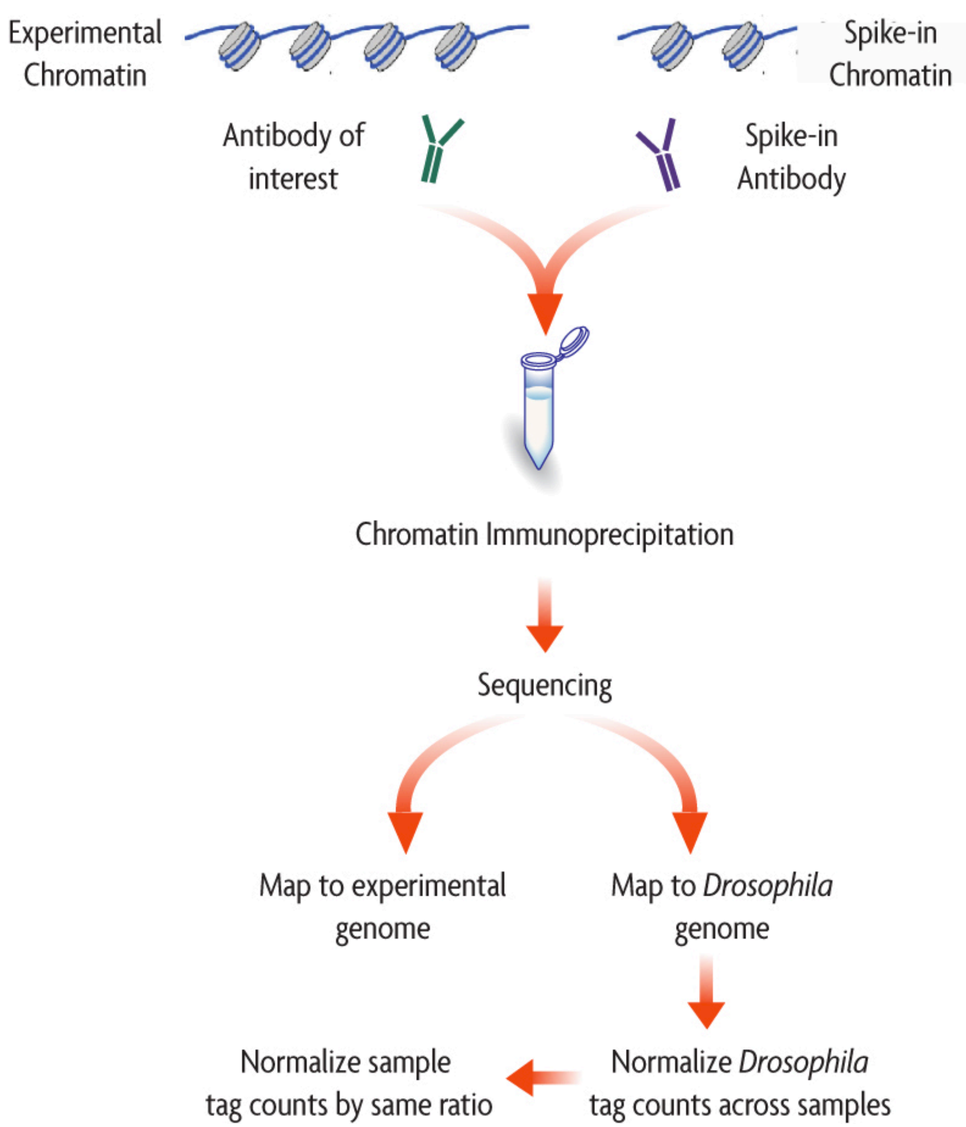
- Open chromatin regions fragment more easily than closed regions
- Repetitive sequences might seem to be enriched (ENCODE also provides a “Black List”)
- Uneven distribution of sequence tags across the genome
- Hyper-ChIPable regions



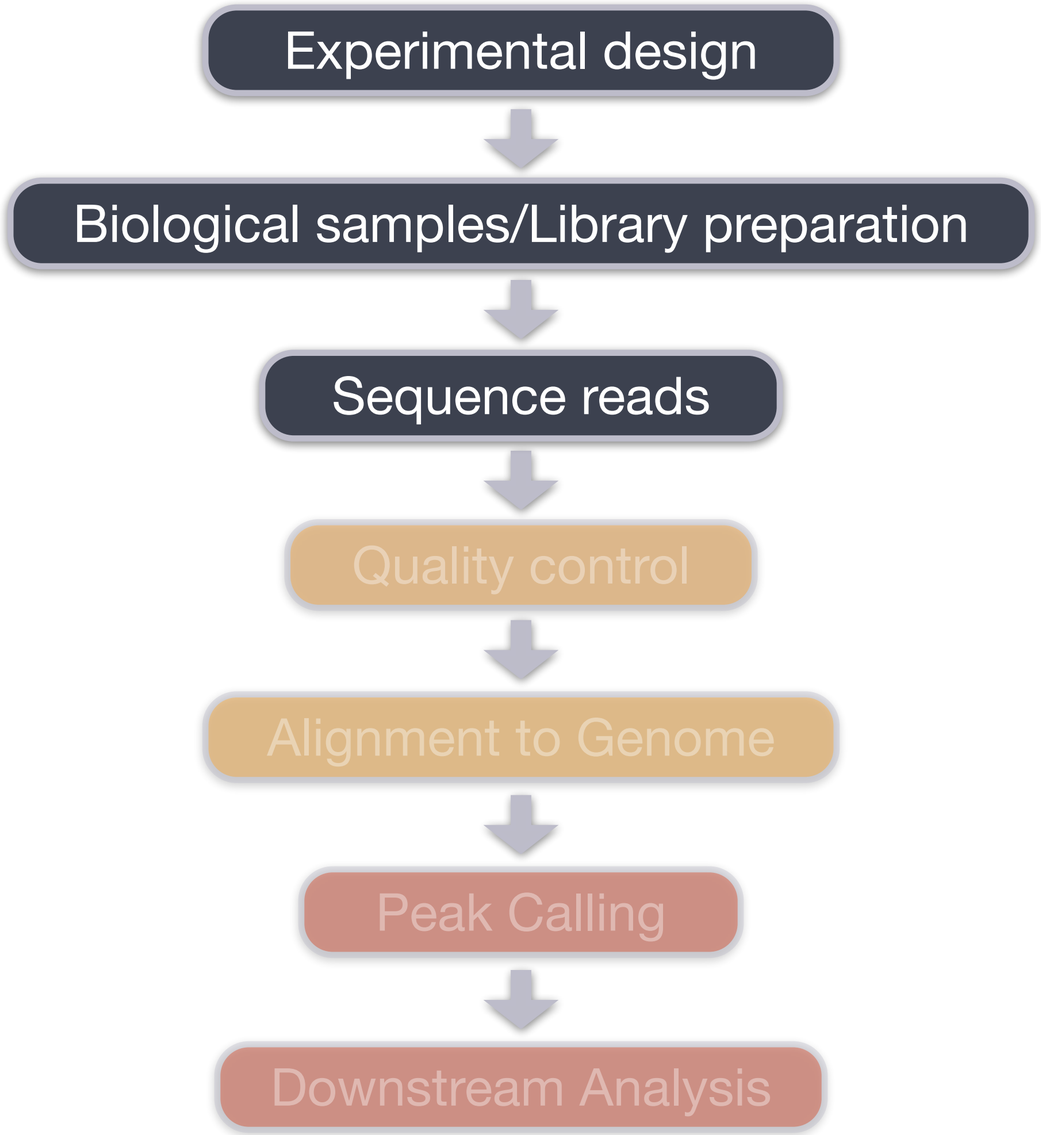
Map of ChIP-seq versus control signals

A note on spike-in controls

- Reduce the effects of technical variation
- Detect subtle biological differences that are not observed with standard ChIP analysis
- Theoretically, can be applied across different antibodies and samples without bias
  - However, does not always work well with different antibodies or with variable cell numbers
  - Works best within a single experiment with the same antibody (e.g. KO vs WT with one antibody)



https://www.activemotif.com/catalog/1091/chip-normalization


Experimental design
Biological samples/Library preparation
Sequence reads
Quality control
Alignment to Genome
Peak Calling
Downstream Analysis

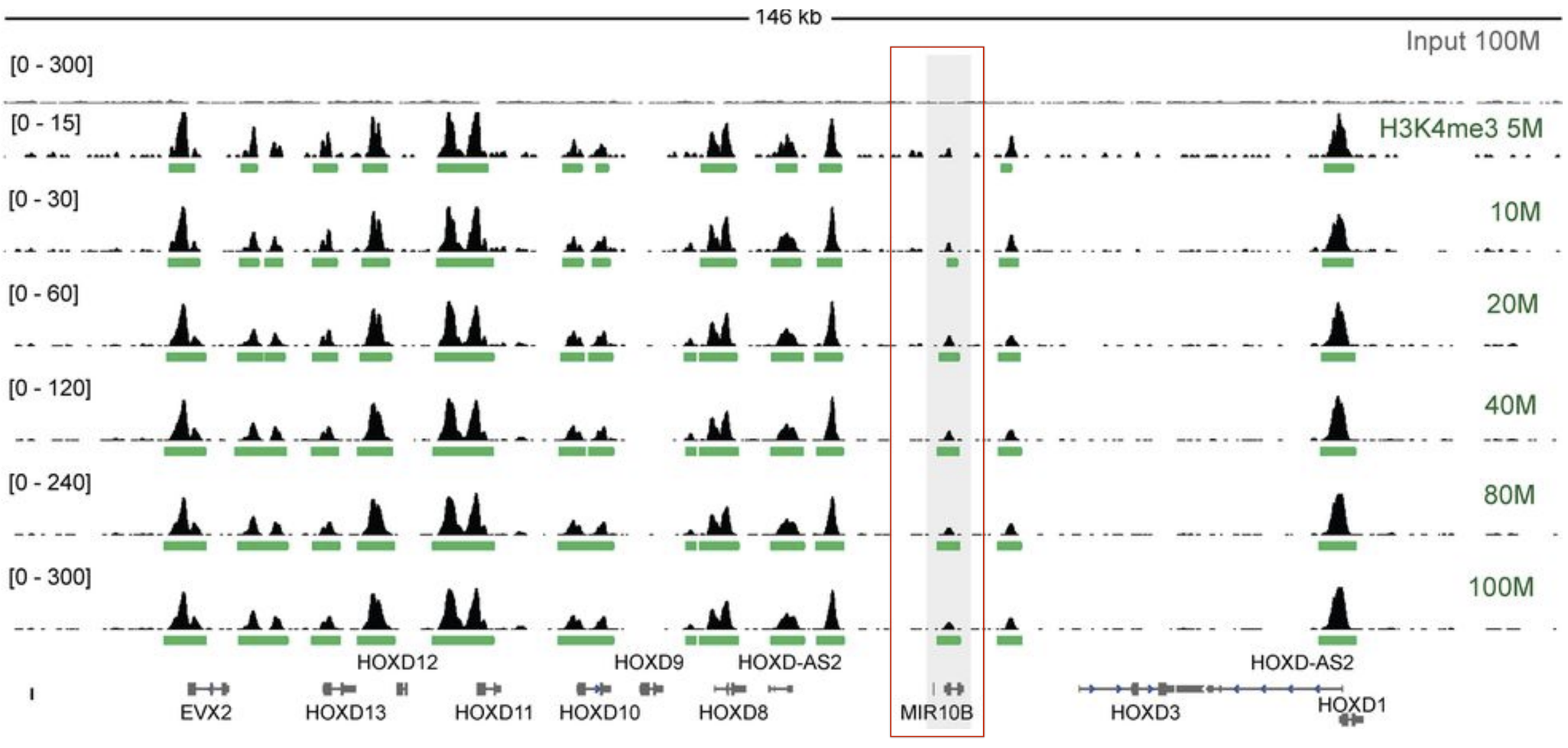
Workflow

Sequencing recommendations

	ChIP-seq	CUT&RUN	ATAC-seq
Read length	50-150 bp	50-75 bp	50-75 bp
Sequencing mode	Single-end in most cases. Paired-end for allele-specific chromatin events or transposable elements	Paired-end recommended for accurate fragment size information.	Paired-end recommended for accurate fragment size information.
Sequencing depth	20-40 million for TFs; 45 million for broad histone profiles Control sequenced to equal or higher depth	2-8 million paired-end reads Control sequenced to equal or higher depth	50 million paired-end reads for changes in accessibility; 200 million for TF footprinting

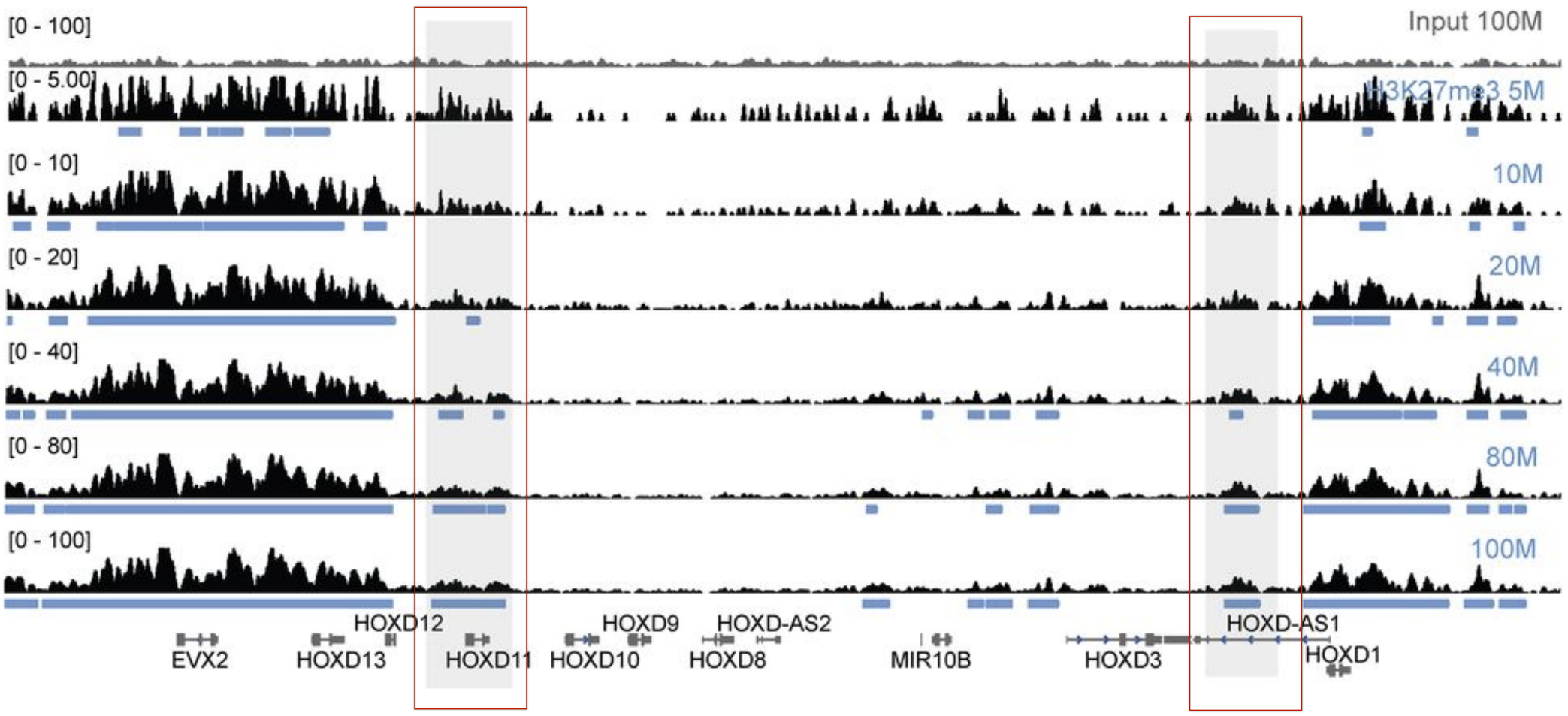
- Balance cost with value of more informative reads.
- Sequencing depth guidelines are for mammalian cells. Smaller genomes require lower depth.

Impact of sequencing depth (ChIP-seq)



Adapted from Jung et al (2014). NAR.

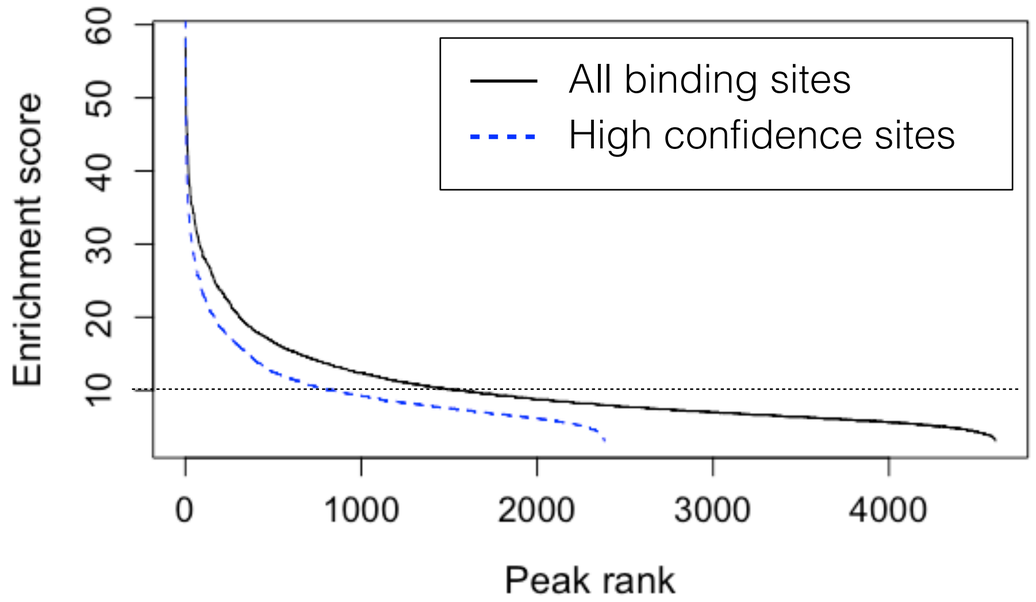
Impact of sequencing depth

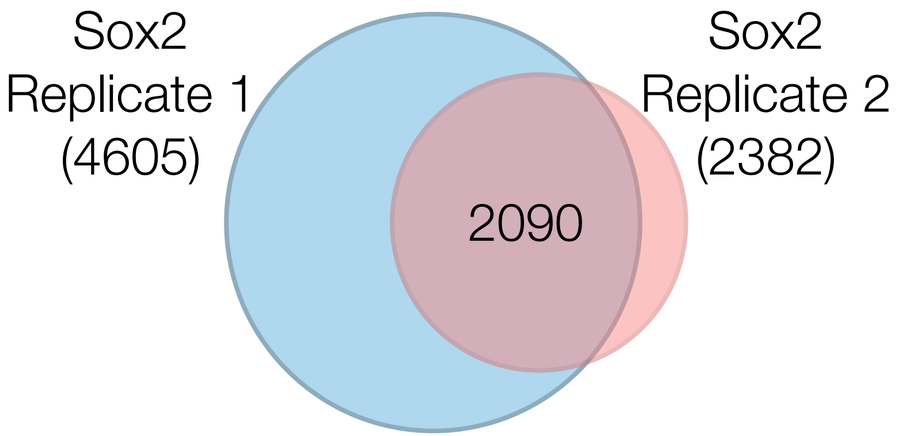


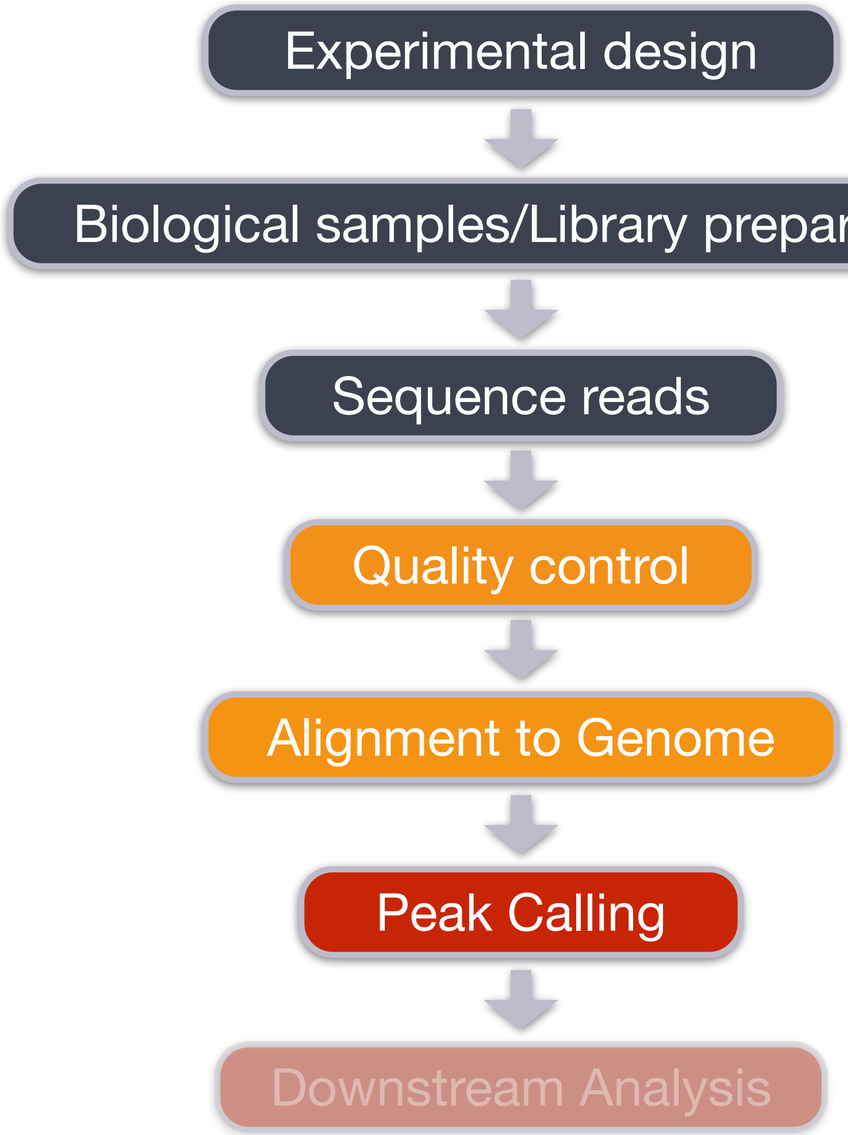
Adapted from Jung et al (2014). NAR.

Replicates and reproducibility

- Biological replicates are essential to understand variation and for differential binding analysis
- More replicates is often preferable to greater depth
- Better to sequence high-quality sample at lower depth than low-quality sample to higher depth







FASTQC

BWA, Bowtie2

Filter duplicates, multi mappers, blacklist

ChIP-seq workflow

Quality check and filtering

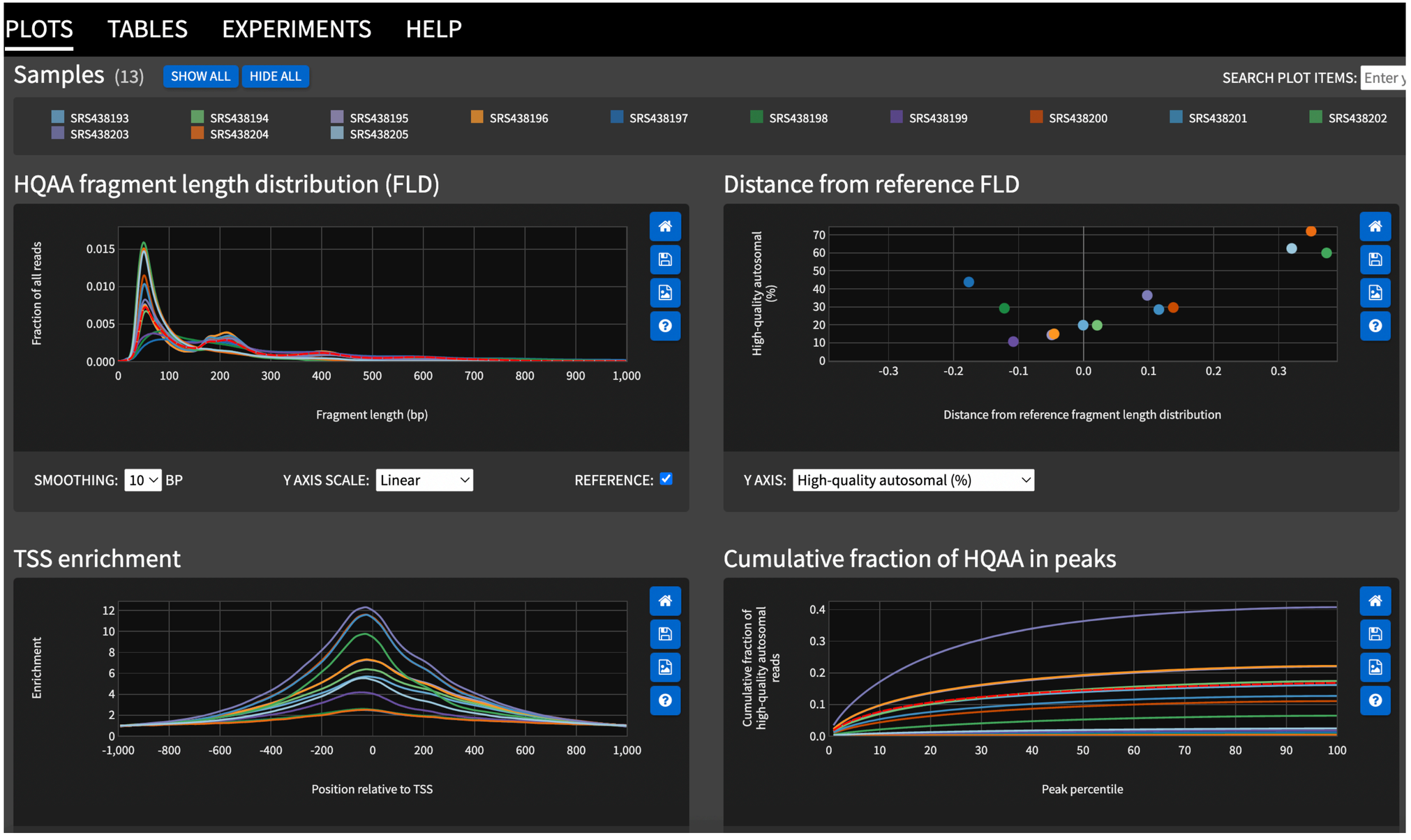
- Raw sequence QC performed with FASTQC
  - Explore duplication rates and possibly remove duplicates
  - Don’t be surprised to see over-represented sequences
  - Remove blacklisted regions
  - Assess cross correlation scores and Fraction of Reads in Peaks (FRiP)

Software: ChIPQC, Homer, ChiLin, DiffBind

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAGCAGTGGTATCAACGCAGAGTACATGGGAAGCAGTGGTATCAACGCAG	42089	0.41441946173421285	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGGATGTGAGGGCGATCTGGC	32502	0.32002331595631606	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGCGCGACCTCAGATCAGACGT	23822	0.2345577328383288	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGTACCTGGTTGATCCTGCCAG	20383	0.20069642634722756	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGAGATTCTGAAACCATTTACT	16026	0.15779624827751895	No Hit
AAGCAGTGGTATCAACGCAGAGTACTGGGTCAATAAGATATGTTGATTTT	15612	0.15371989442834305	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGGGCGGAGGAAGCTCATCAG	15338	0.1510220177262315	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGACTGACACGCTGTCCTTTCC	13227	0.13023655160156888	No Hit
AAGCAGTGGTATCAACGCAGAGTACTGGCCGTGAGTCTGTTCCAAGCTCC	12826	0.12628819920176326	No Hit
AAGCAGTGGTATCAACGCAGAGTACATGGGGGGGGTGTACTGGCTTCGAC	10313	0.10154453441195888	No Hit

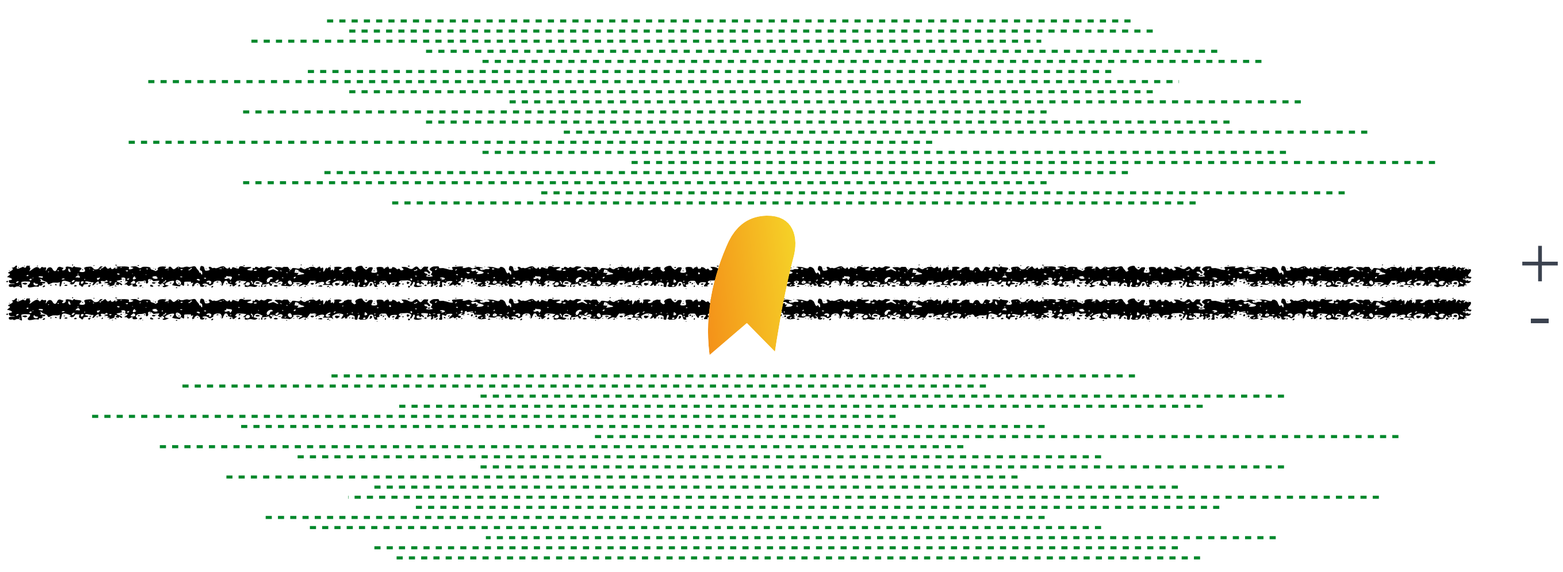
Quality control with atacqv



Understanding strand cross-correlation

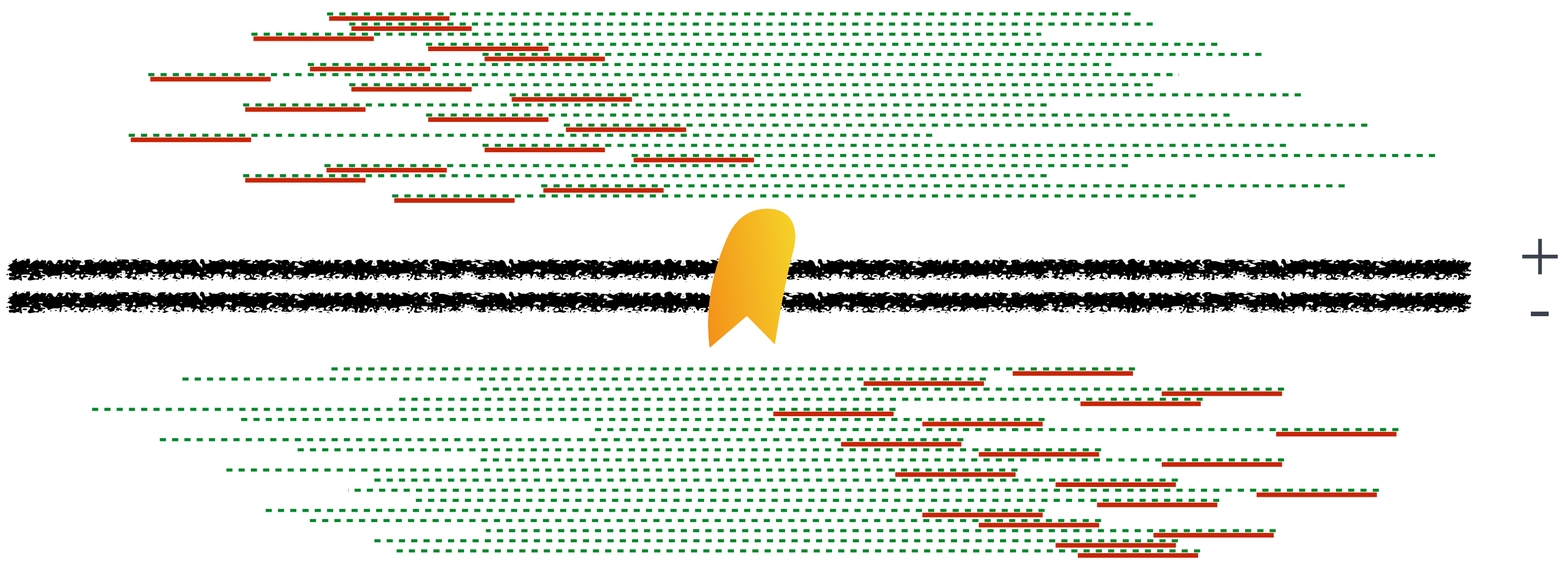
= binding site

---- = size selected DNA fragment



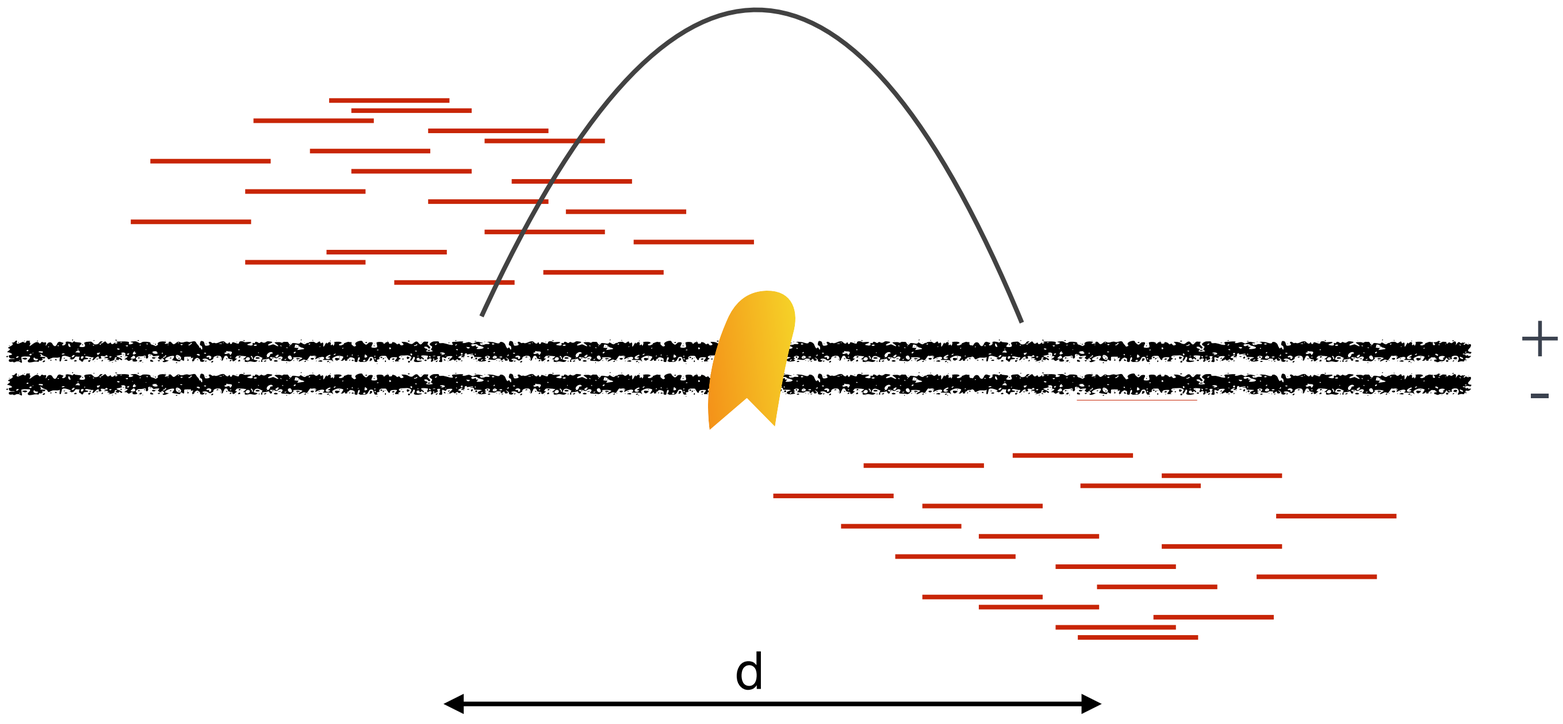
Understanding strand cross-correlation

ChIP-seq fragments are sequenced from the 5’ end



Understanding strand cross-correlation

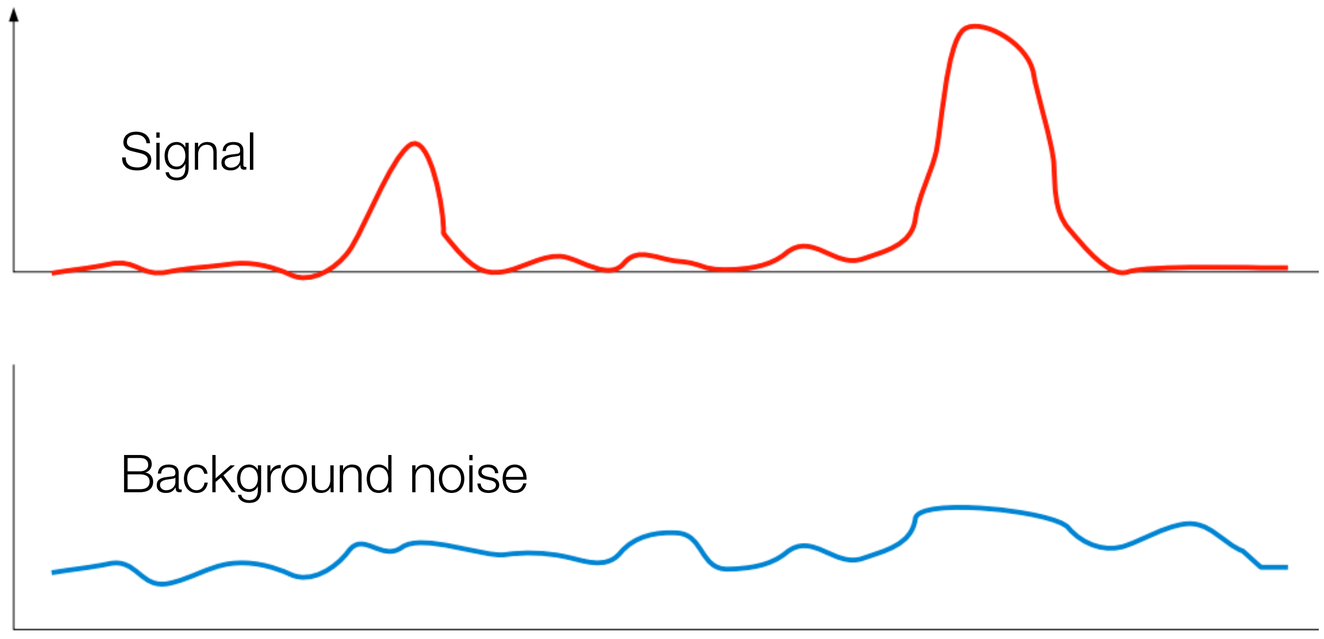
Alignment generates a **bimodal pattern** on the plus and minus strands around binding sites



Peak calling algorithms use this pattern to estimate the relative strand shift

Modeling noise to detect real peaks

- Noise is not uniform (chromatin conformation, local biases, mappability)
- Input data is mandatory for a reliable estimation of noise (even though some tools don't require it)



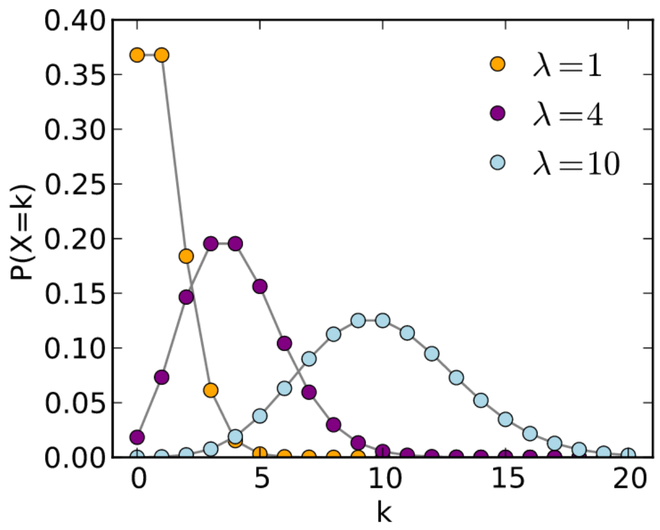
Peak detection

- Most algorithms model the number of reads from a genomic region/ window using a Poisson distribution
- One parameter model for estimating the expected number of reads in the window
- Often more variance in real data than assumed by the Poisson (overdispersion)
- MACS (model-based analysis of ChIP-Seq) uses multiple Poisson distributions to model the local background noise within each region from the input data

$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

where

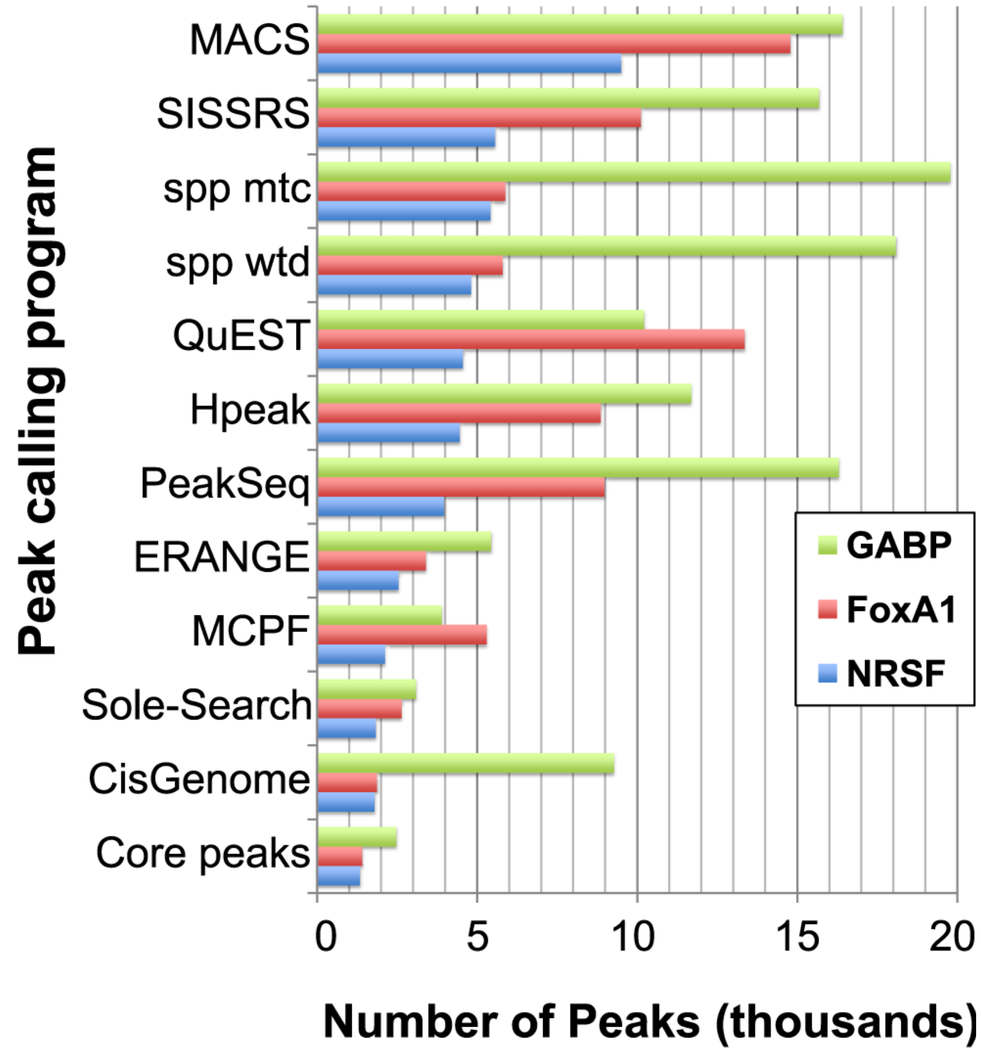
- λ is the average number of events per interval
- e is the number 2.71828... (Euler's number) the base of the natural logarithms
- k takes values 0, 1, 2, …
- $k! = k \times (k-1) \times (k-2) \times \ldots \times 2 \times 1$ is the factorial of k.



http://en.wikipedia.org/wiki/Poisson_distribution

Peak callers

- Variability in number of peaks called
- Tend to agree on the strongest signals



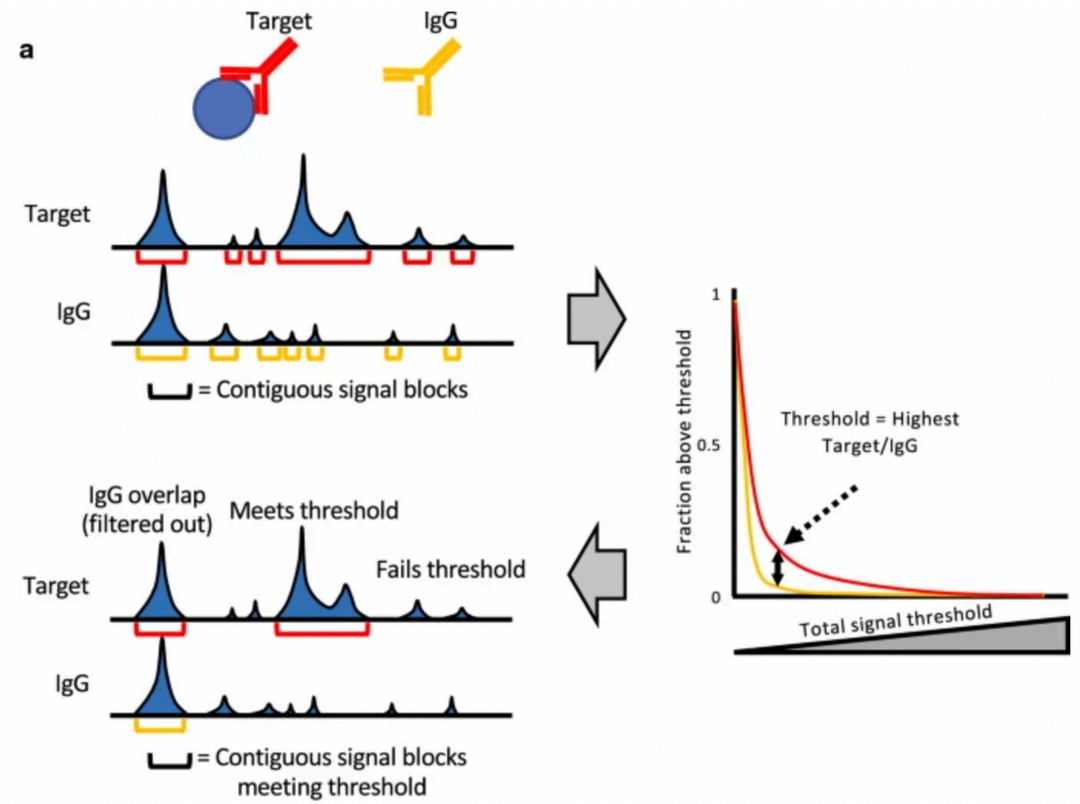
Wllbanks & Facciotti (2010). PLoS ONE.

How to choose one

- Widely used
- Actively maintained and updated
- Default settings are a good start but know your parameters for your peak caller
- Be critical! Visually inspect your data (IGV)

CUT&RUN peak calling

- Consider using SEACR (Henikoff Lab)
- Useful for identifying large domains (H3K27me3)
- Fewer false positive calls



Meers, et al (2019). *Epigenetics & Chromatin* **12**, 42 (2019). https://doi.org/10.1186/s13072-019-0287-4

Downstream analysis

- Detecting differential enrichment across samples
  - Steinhauser et al, Brief Bioinform. (2016)

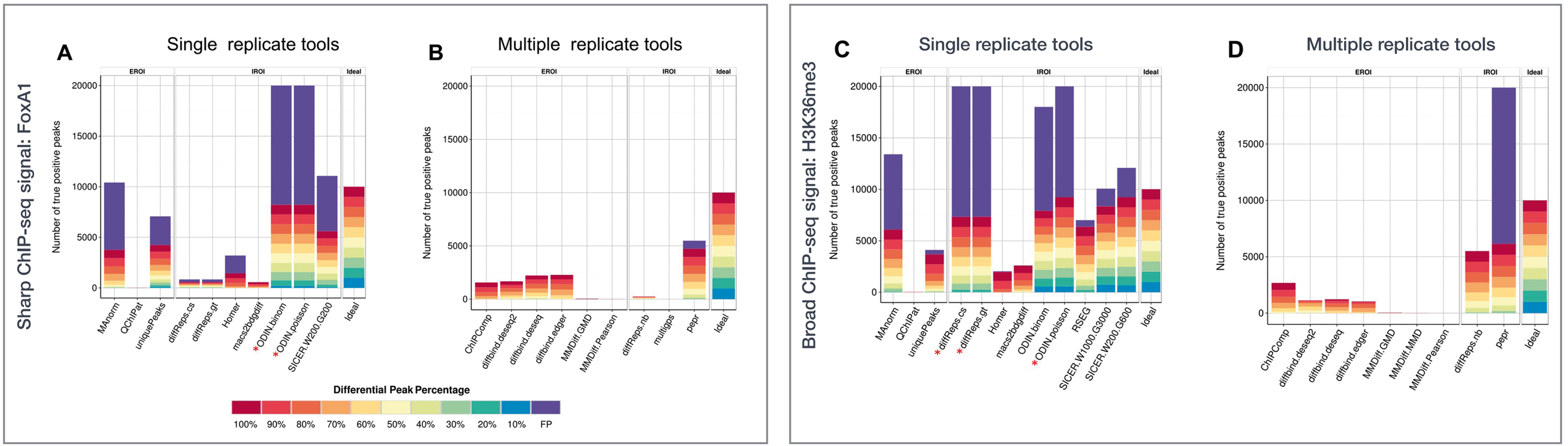


Figure 4. Proportion of true and false positives for each tool on the simulated FoxA1 data set (A, B) and H3K36me3 data (C, D)

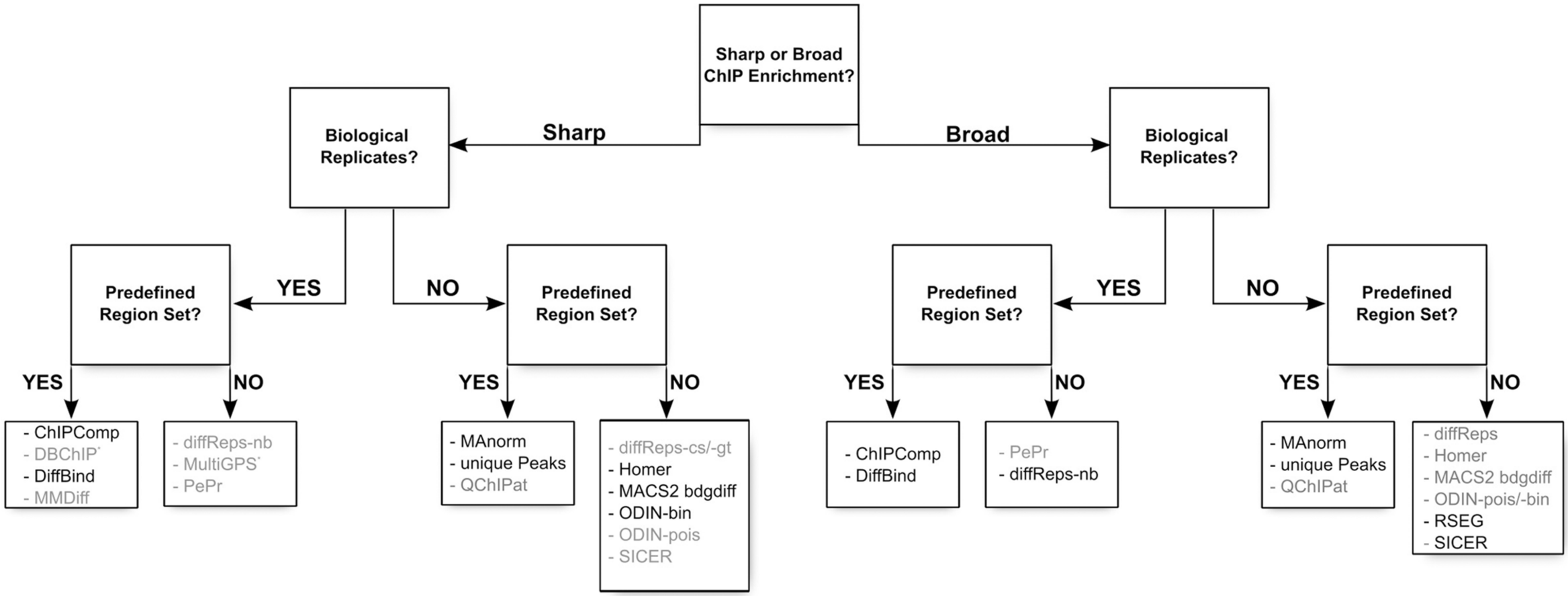
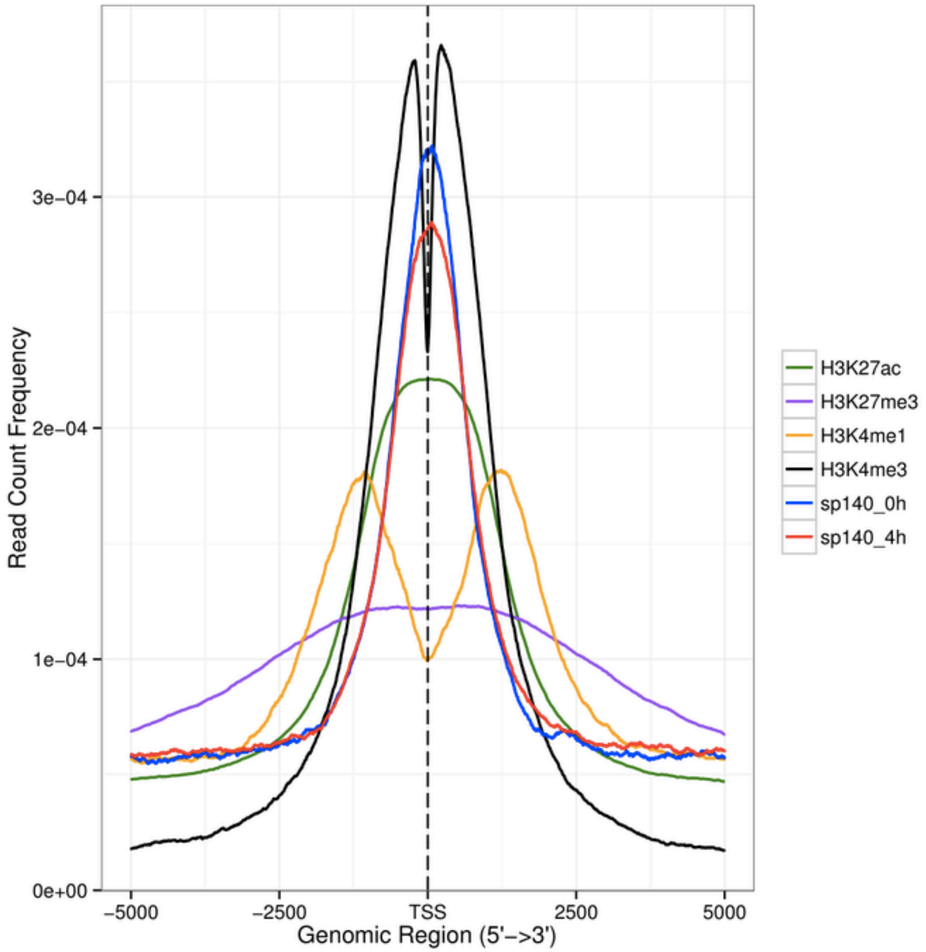
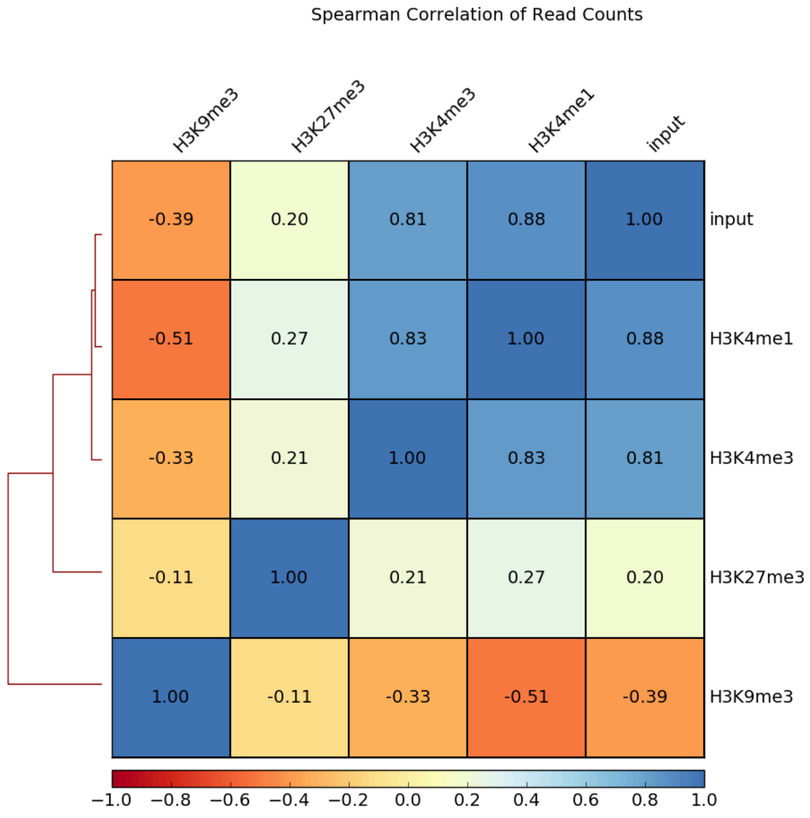


Figure 7. Decision tree indicating the proper choice of tool depending on the data set: shape of the signal (sharp peaks or broad enrichments), presence of replicates and presence of an external set of regions of interest [*Steinhauser, et al, 2016*].

Downstream analysis

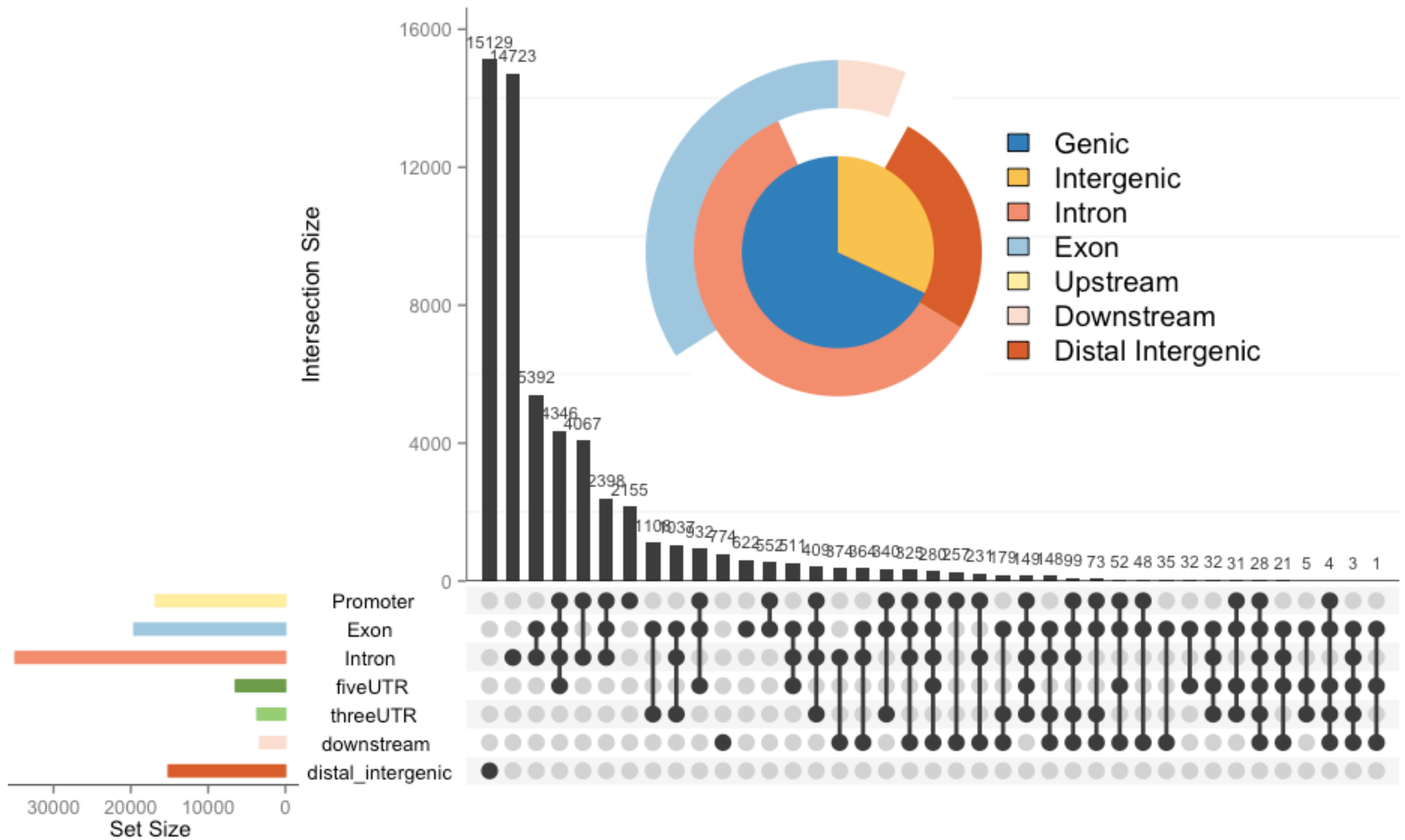
- Annotation of peaks - distance from TSS
  - ChIPseeker, Homer, ChiLin, DeepTools





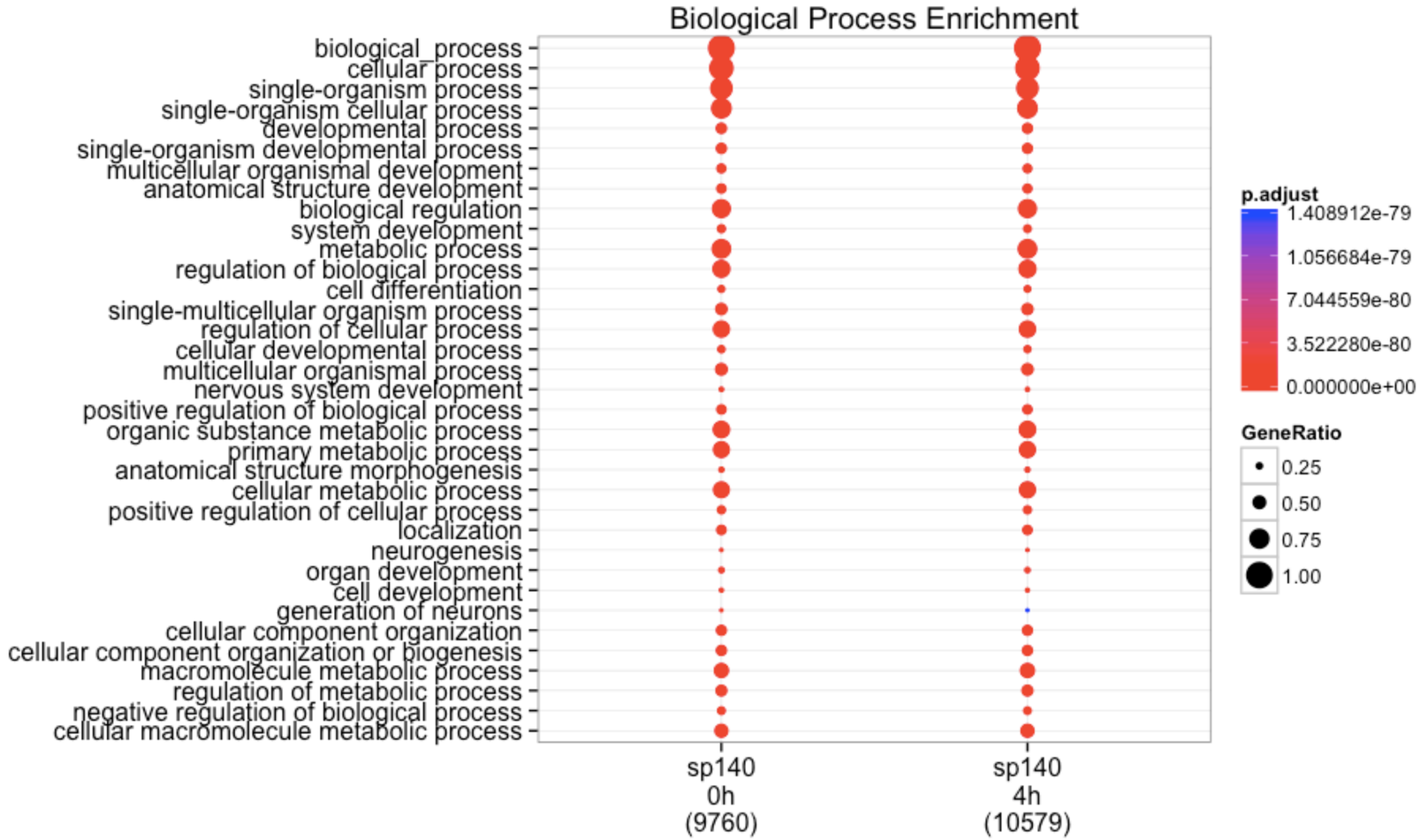
Downstream analysis

- Annotation of peaks - genomic context
  - ChIPseeker, Homer



Downstream analysis

- Functional enrichment analysis
  - ChIPseeker, GREAT, Homer



Downstream analysis

- Motif discovery
  - MEME suite, ChiLin, Homer

For further information on how to interpret these results or to get a copy of the MEME software please access http://meme.nbcr.net.

If you use DREME in your research please cite the following paper:
Timothy L. Bailey, "DREME: Motif discovery in transcription factor ChIP-seq data", *Bioinformatics*, **27**(12):1653-1659, 2011. [full text]

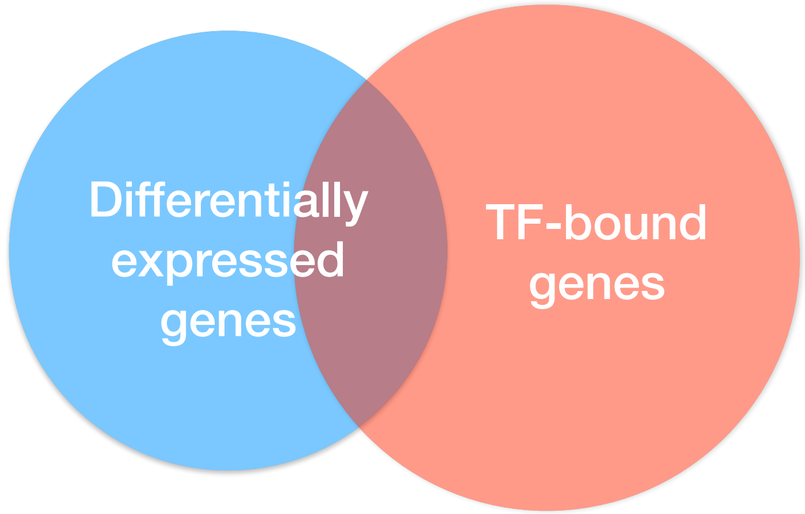
DISCOVERED MOTIFS | INPUTS & SETTINGS | PROGRAM INFORMATION

DISCOVERED MOTIFS

	Motif	Logo	RC Logo	E-value	Unerased E-value	More	Submit/Download
1.	CYWTTGTB			4.2e-299	4.2e-299		
2.	ATGBWAAT			8.4e-179	1.1e-179		
3.	CCMCDCCC			1.3e-130	1.1e-131		

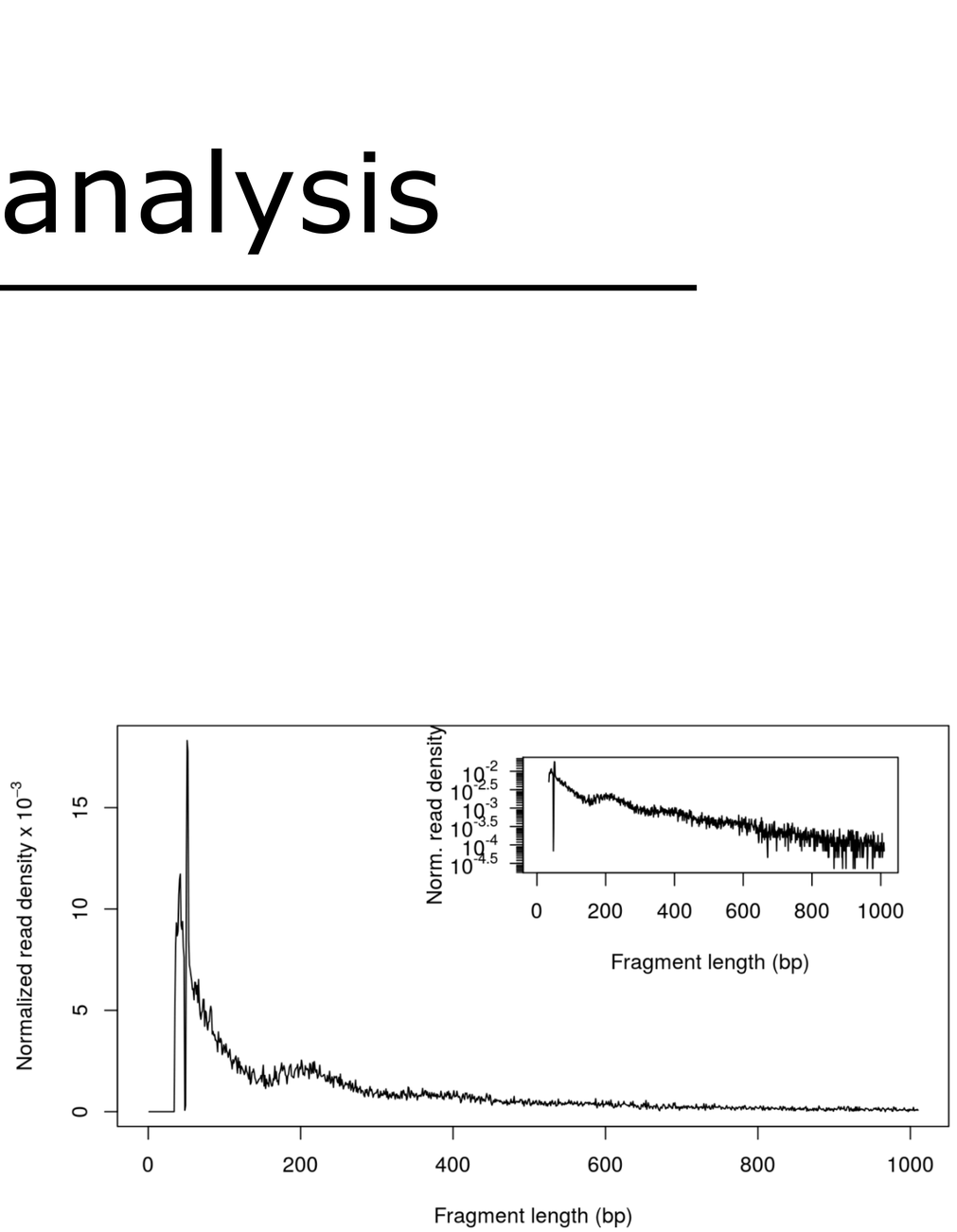
Downstream analysis

- Integrative analysis of RNA-seq and ChIP-seq
  - Which of the regulated genes are direct targets of the TF?
  - Is the TF an activator, repressor, or both?
  - Does the TF have different binding partners depending on the direction of regulation?



ATAC-seq data analysis

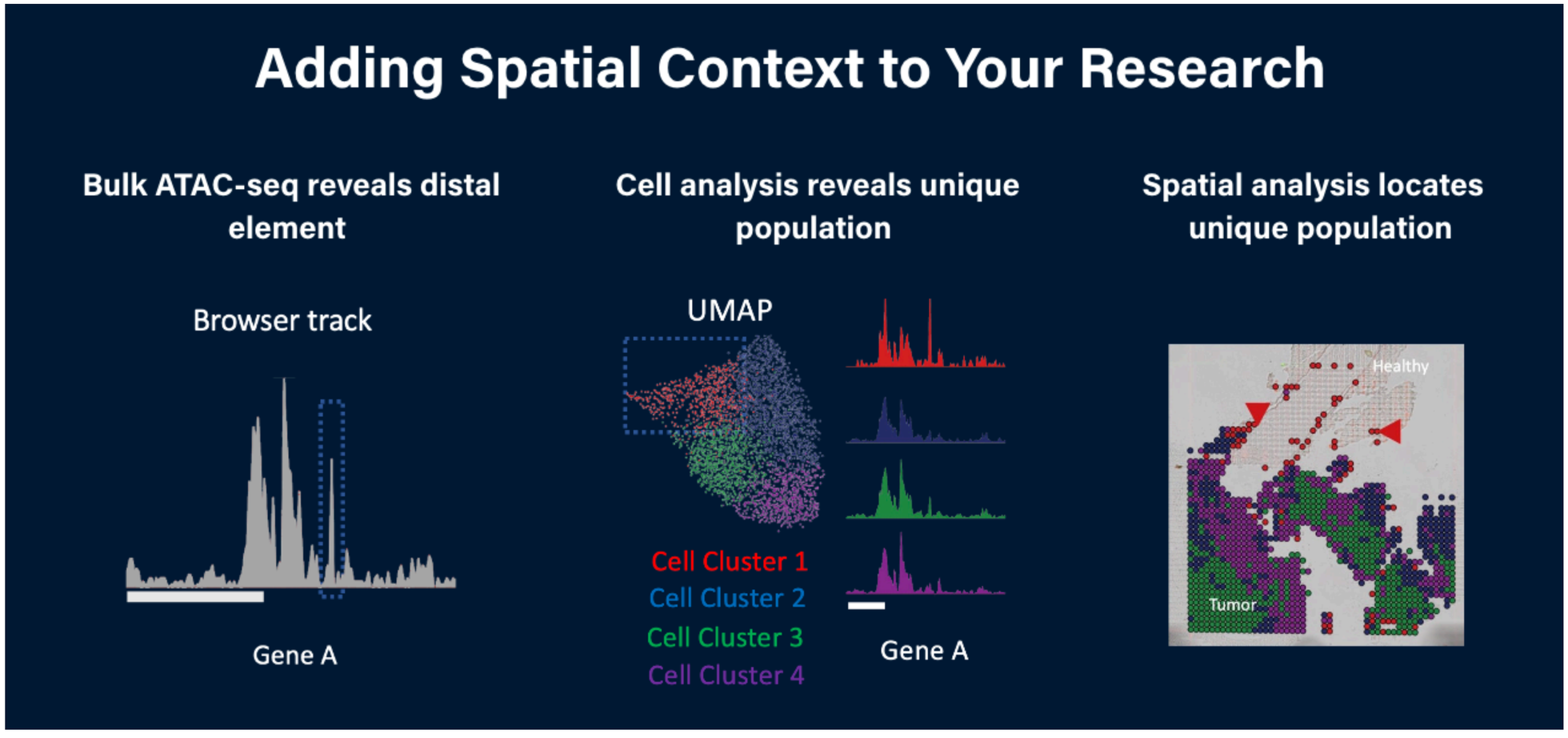
- Peak calling using MACS2 with PE settings and without model building
- Remove mitochondrial reads
- Shift alignments
- Separate nucleosome free regions (NFR) from nucleosome containing regions



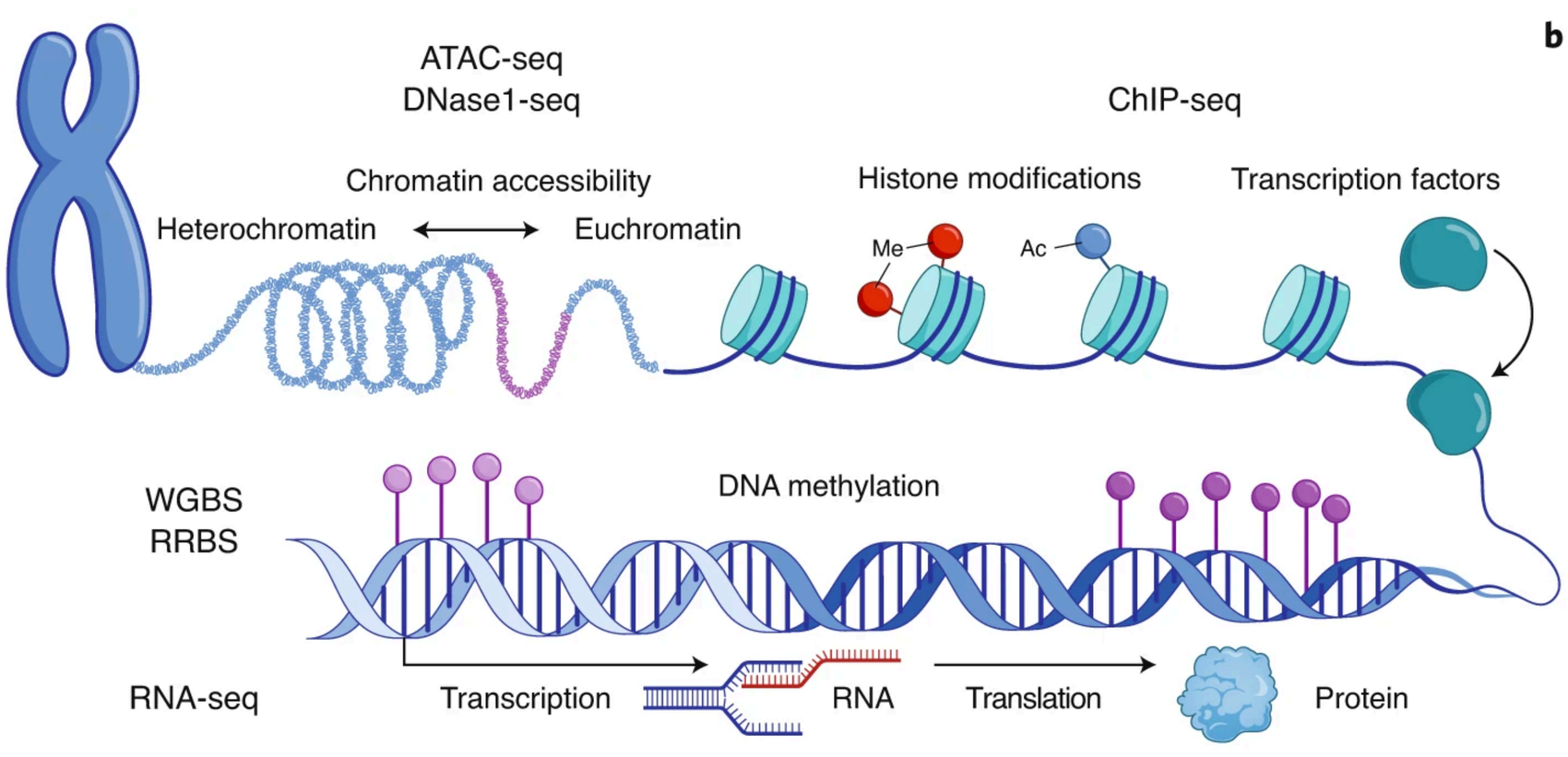
Ou et al, Bioconductor 2018

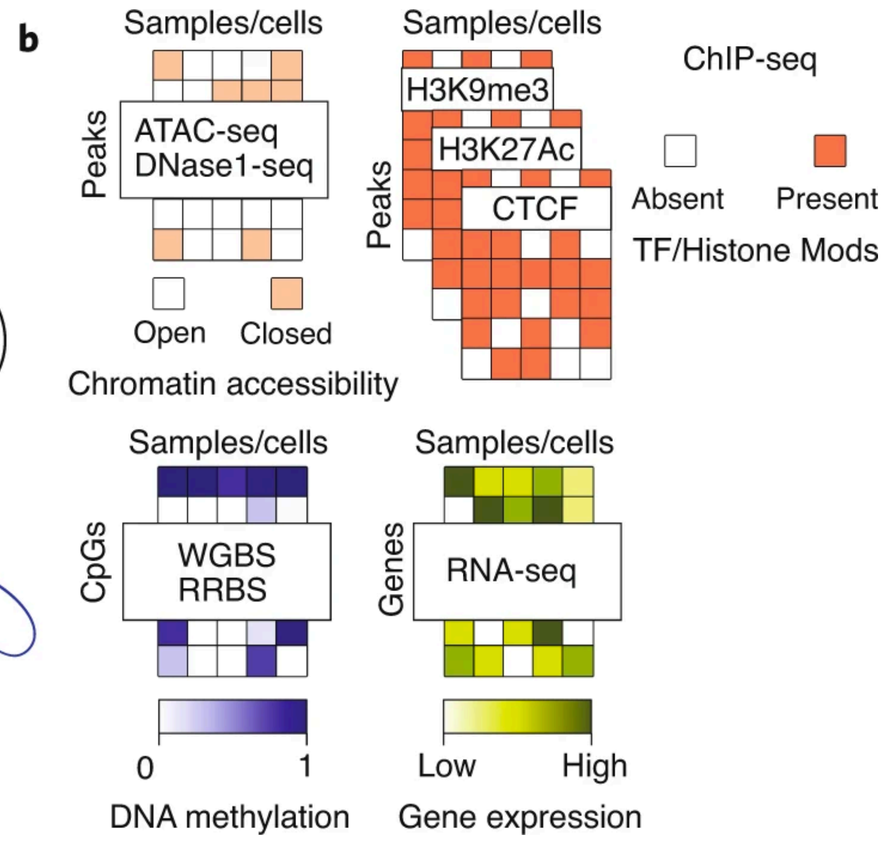
Advances in technology

- 10X Single cell ATAC-seq
- 10X Multiome (sc ATAC-seq + RNA-seq)
- Spatial epigenomics (AtlasXomics)



Inputs for machine learning





https://www.nature.com/articles/s43588-021-00038-7

Summary

- A review of chromatin structure
- Basics of the ChIP-seq, CUT&RUN and ATAC-seq protocols
- Better understanding of how to design epigenomic experiments
- How to analyze the data
- What to look for in a good ChIP data set
- Emerging methods to improve signals and characterize regulatory domains

Ask us questions

shosui@hsph.harvard.edu

bioinformatics.sph.harvard.edu