Lecture 13. ChipSeq, HiC & ChromHMM

Michael Schatz

March 13, 2017

JHU 601.749: Applied Comparative Genomics

Project Proposal!
Due March 15

Project Proposal

Assignment Date: March 7, 2018
Due Date: Thursday, March 15, 2017 @ 11:59pm

Review the Project Ideas page

Work solo or form a team for your class project (no more than 3 people to a team).

The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)

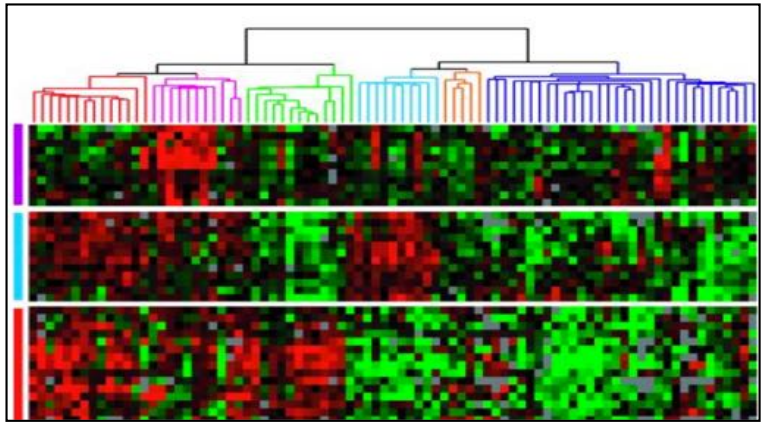
Submit the proposal as a single page PDF on blackboard. After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for the project.

Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

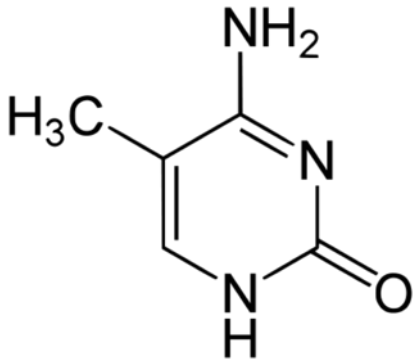
Please use Piazza to coordinate proposal plans!

*-seq in 4 short vignettes

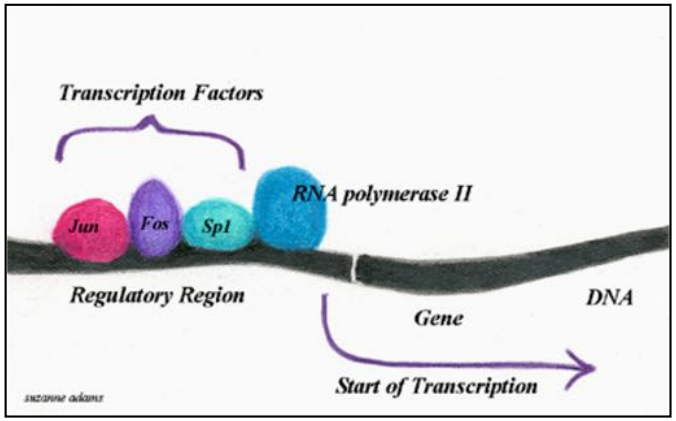
RNA-seq



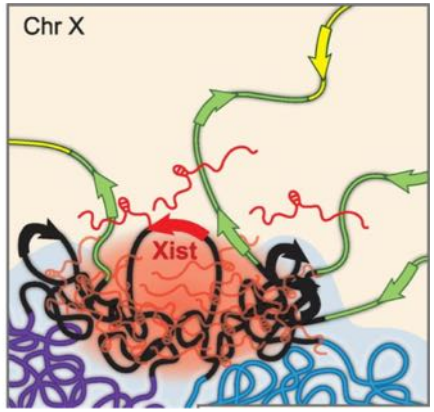
Methyl-seq



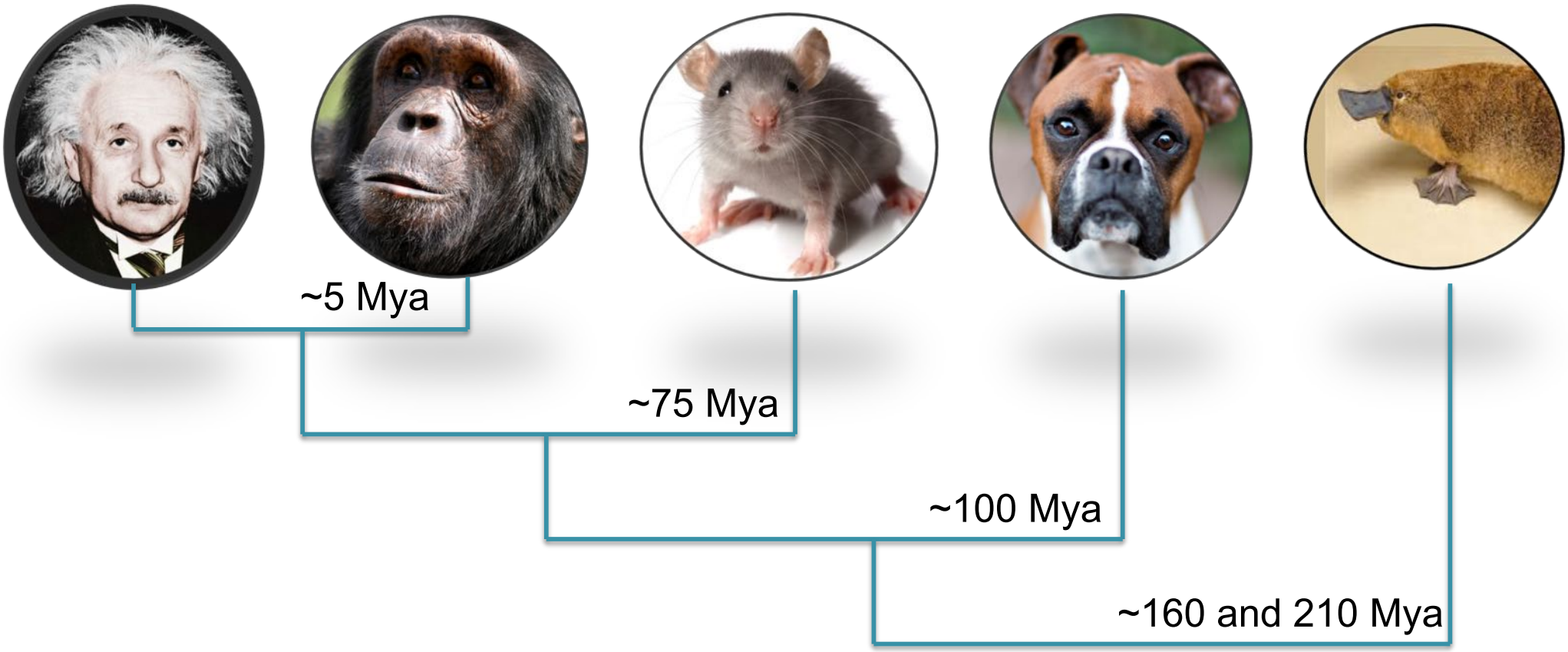
ChIP-seq



Hi-C



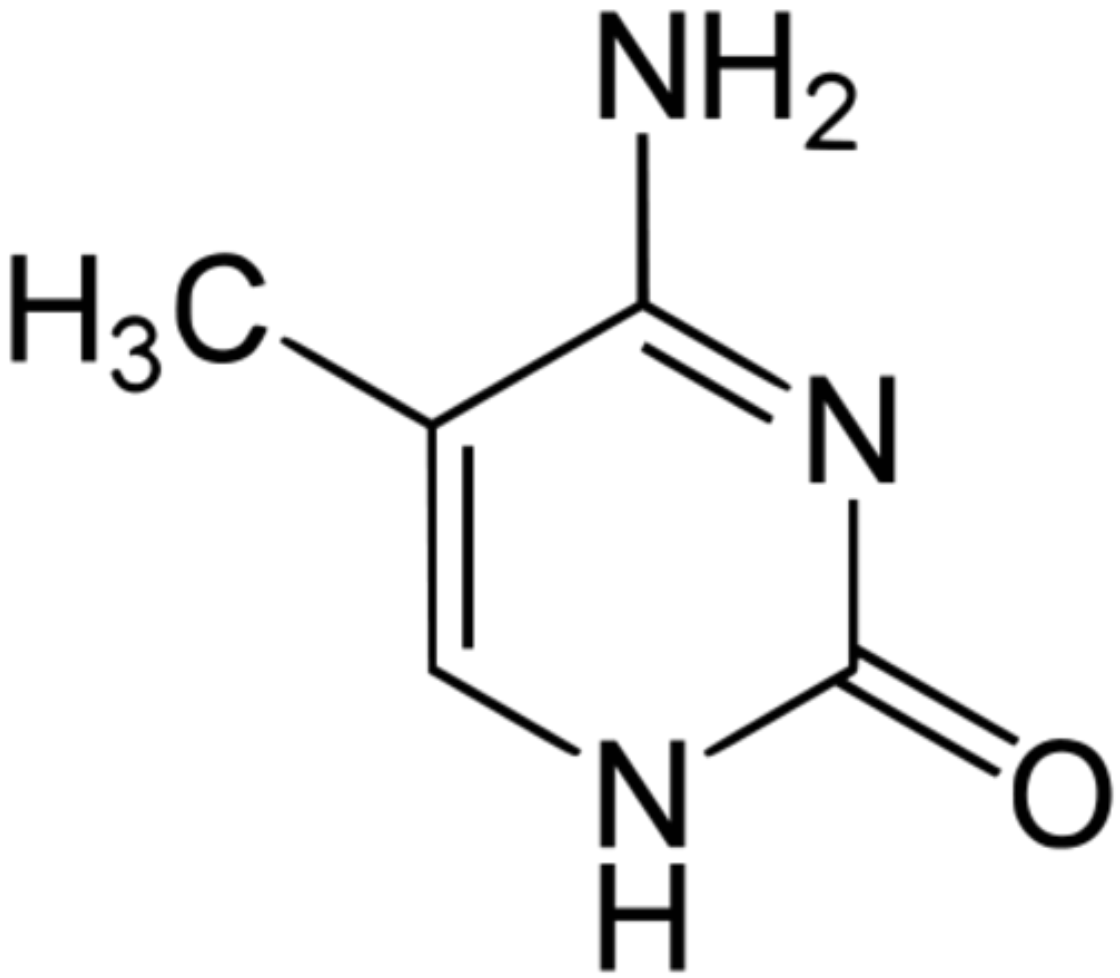
Human Evolution



As expected, the majority of platypus genes (82%; 15,312 out of 18,596) have orthologues in these five other amniotes (Supplementary Table 5). The remaining 'orphan' genes are expected to primarily reflect rapidly evolving genes, for which no other homologues are discernible, erroneous predictions, and true lineage-specific genes that have been lost in each of the other five species under consideration.

Genome analysis of the platypus reveals unique signatures of evolution
(2008) *Nature.* 453, 175-183 doi:10.1038/nature06936

Methyl-seq



Finding the fifth base: Genome-wide sequencing of cytosine methylation
Lister and Ecker (2009) *Genome Research.* 19: 959-966

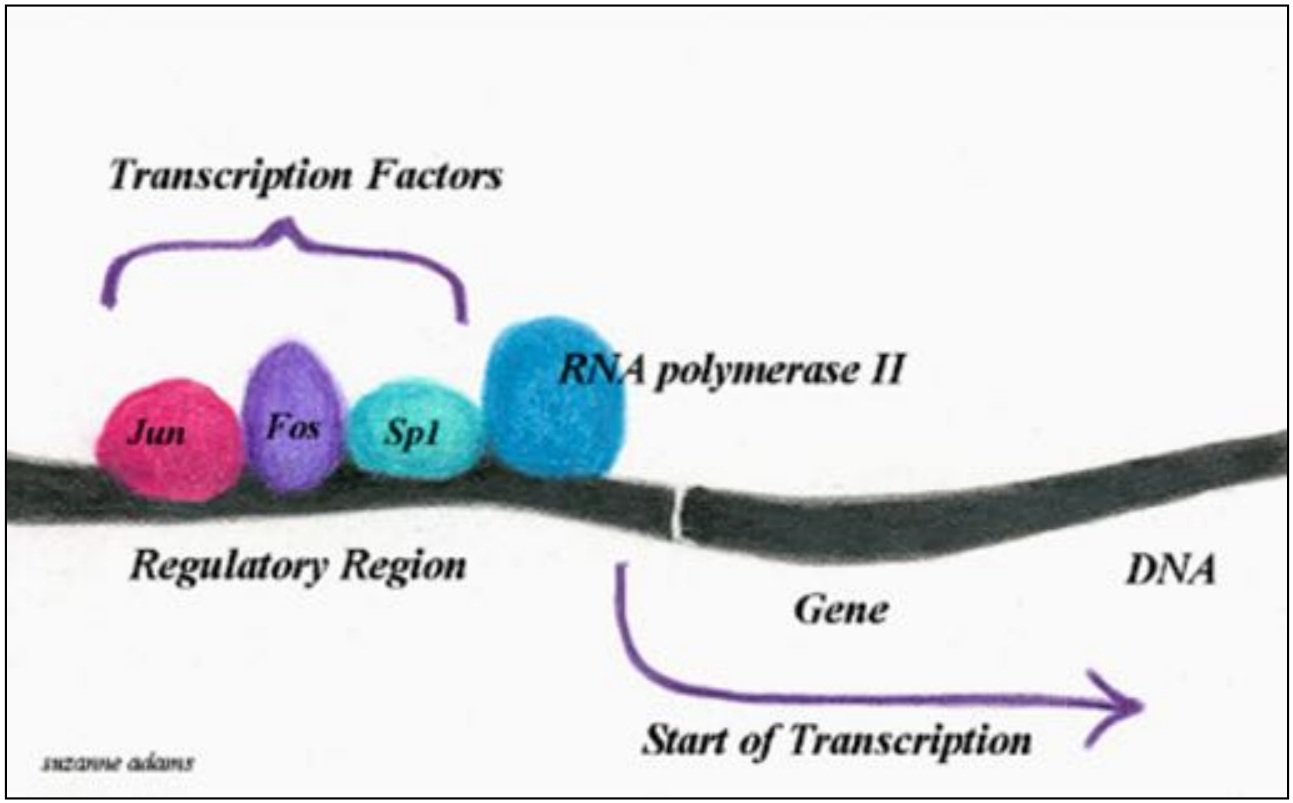
Bisulfite Conversion

Tr
w

-
-
-

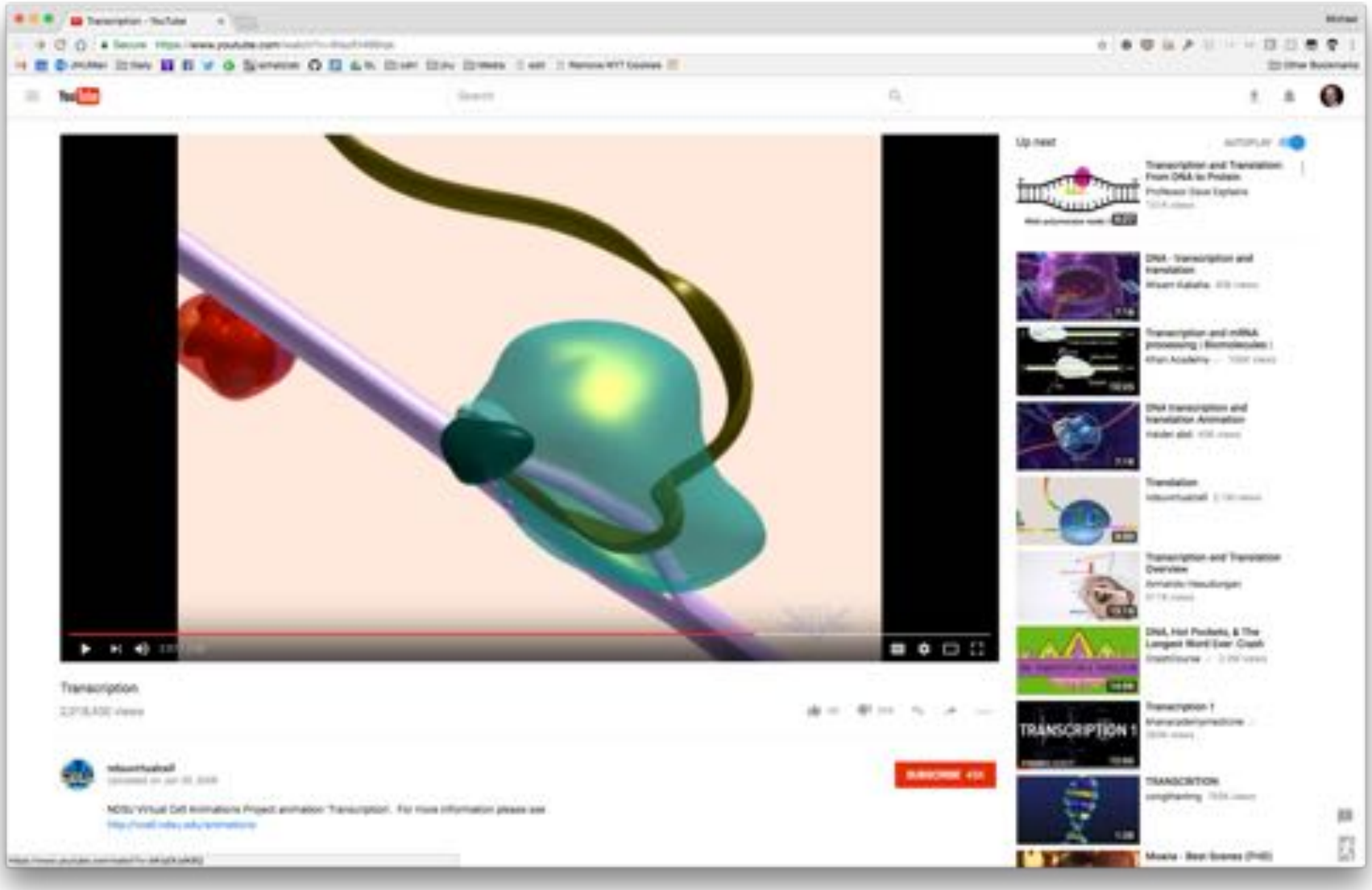
Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

ChIP-seq

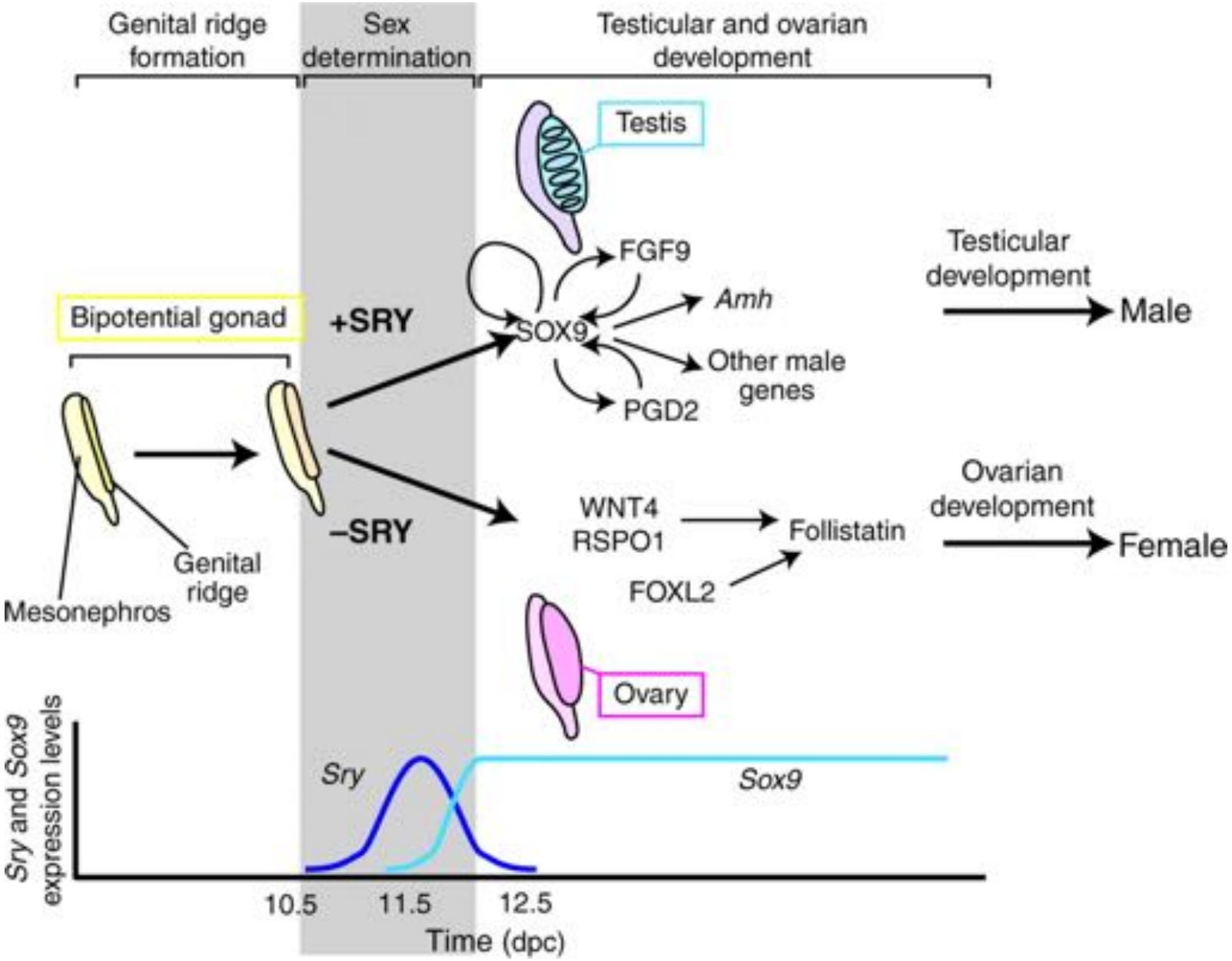


Genome-wide mapping of in vivo protein-DNA interactions.
Johnson *et al* (2007) *Science.* 316(5830):1497-502

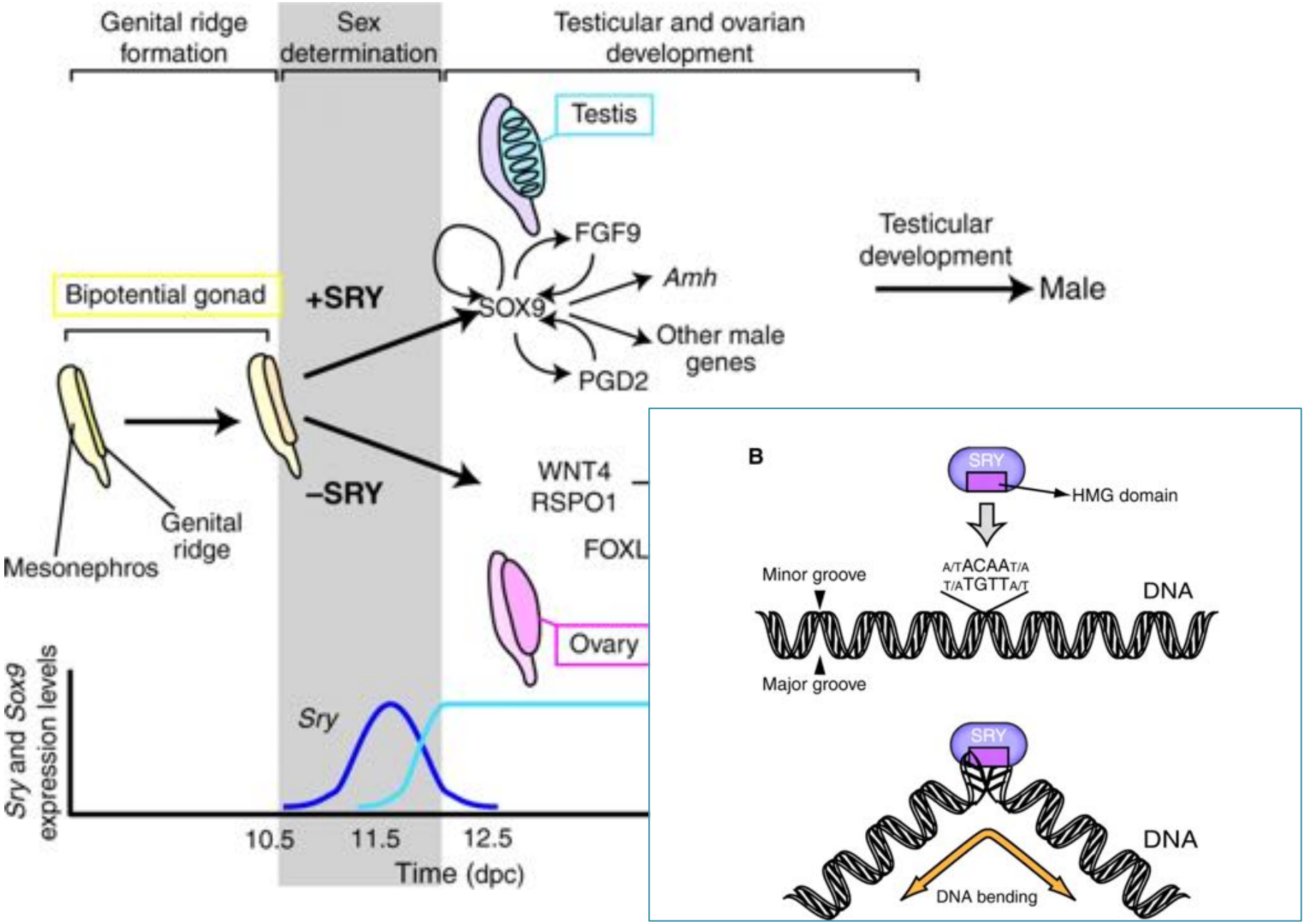
Transcription



https://www.youtube.com/watch?v=WsofH466lqk



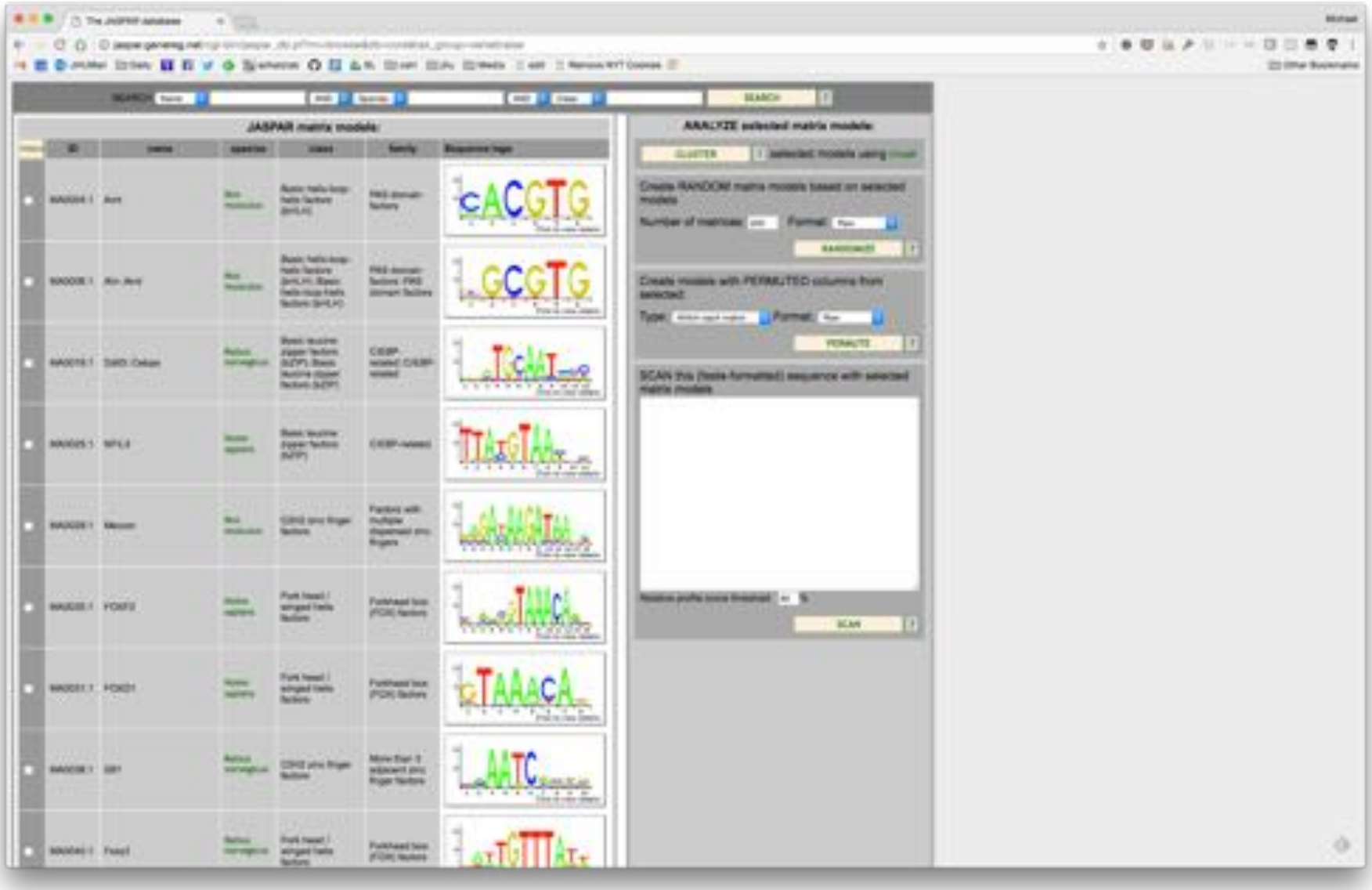
Sry: the master switch in mammalian sex determination
Kashimada and Koopman (2010) Development 137: 3921-3930; doi: 10.1242/dev.048983



Sry: the master switch in mammalian sex determination

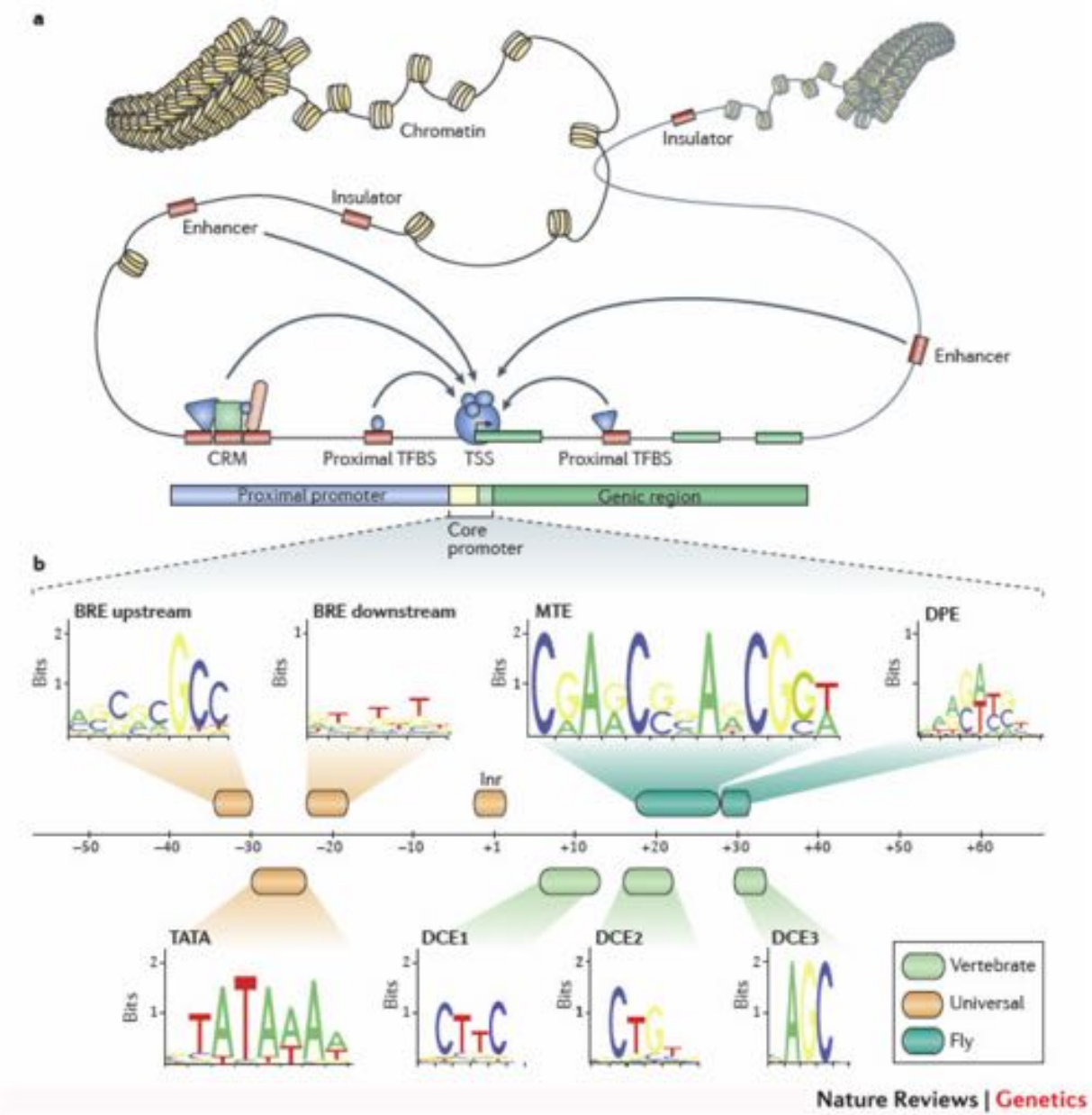
Kashimada and Koopman (2010) Development 137: 3921-3930; doi: 10.1242/dev.048983

Transcription Factors Database



JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles
Anthony Mathelier (2014) Nucleic Acids Res. 42 (D1): D142-D147. DOI: https://doi.org/10.1093/nar/gkt997

Promoters

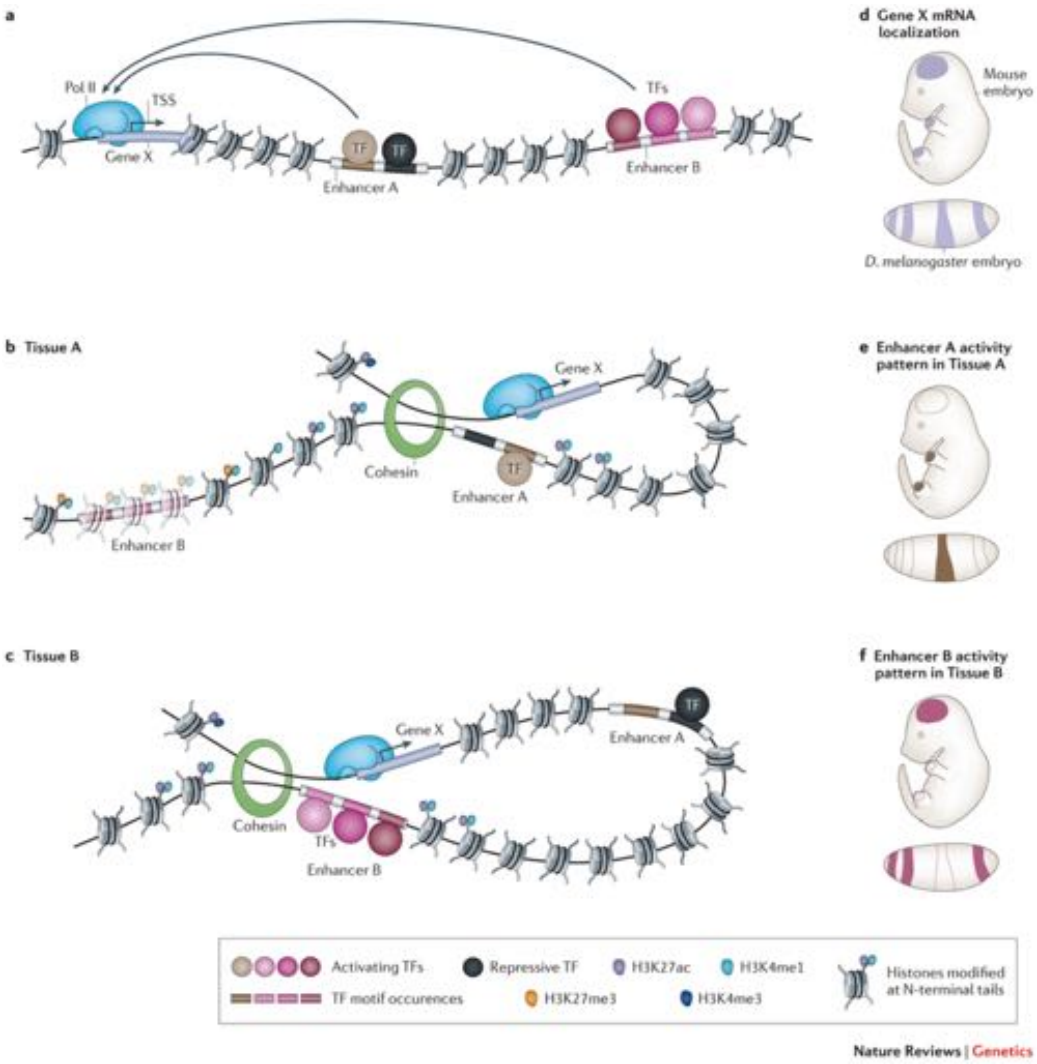


Metazoan promoters: emerging characteristics and insights into transcriptional regulation
Lenhard et al (2014) *Nature Reviews Genetics* 15, 272–286

Enhancers

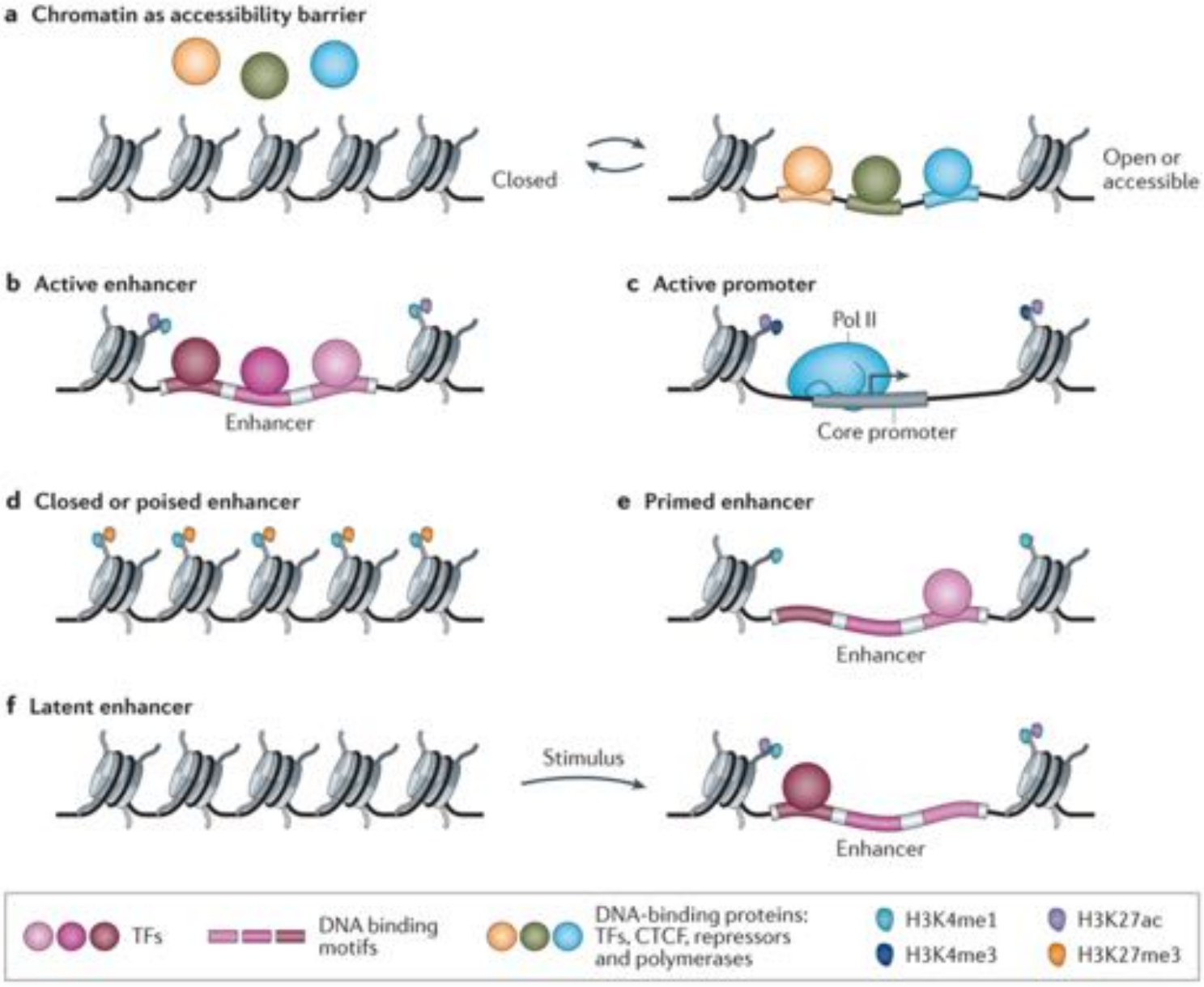
Enhancers are genomic regions that contain binding sites for transcription factors (TFs) and that can upregulate (enhance) the transcription of a target gene.

- Enhancers can be located at any distance from their target genes (up to ~1Mbp)
- In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping
- Active and inactive gene regulatory elements are marked by various biochemical features
- Complex patterns of gene expression result from the additive action of different enhancers with cell-type- or tissue-specific activities



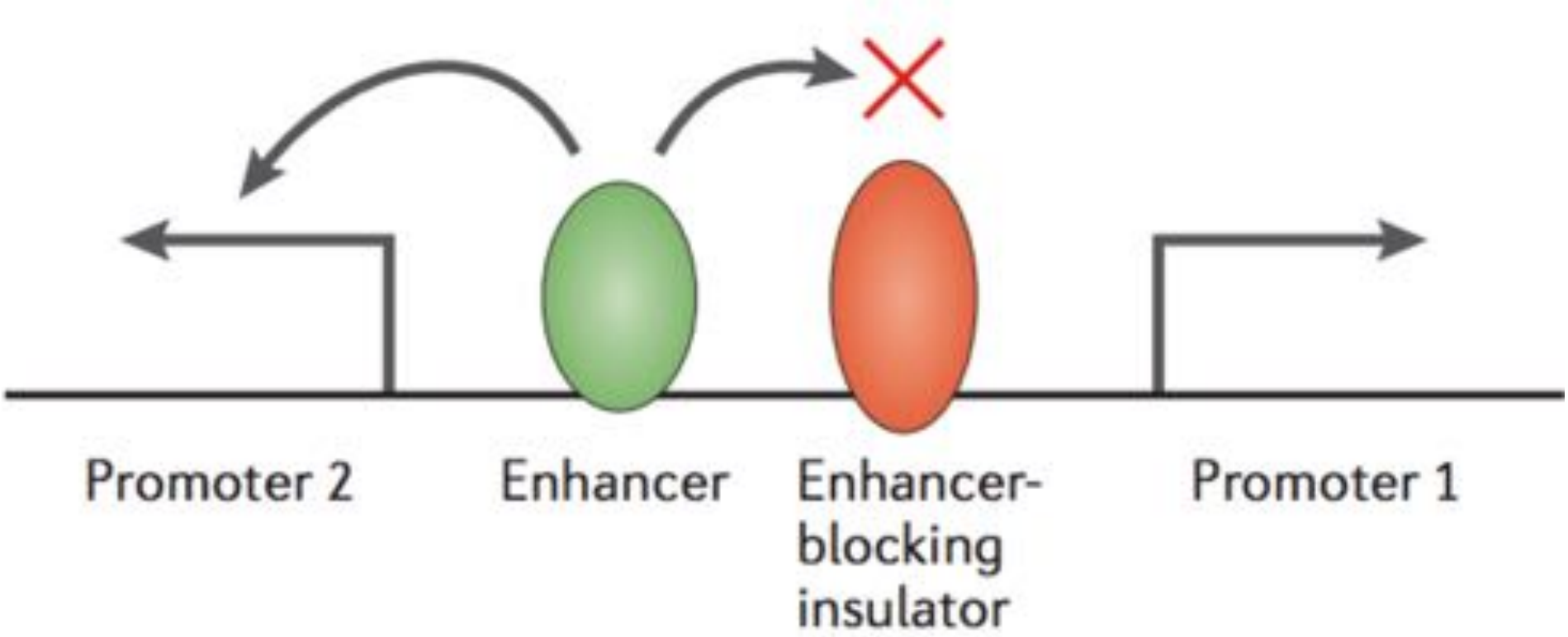
Transcriptional enhancers: from properties to genome-wide predictions
Shlyueva et al (2014) *Nature Reviews Genetics* 15, 272–286

Enhancer States



Nature Reviews | Genetics

Insulators



Insulators are DNA sequence elements that prevent "inappropriate interactions" between adjacent chromatin domains.

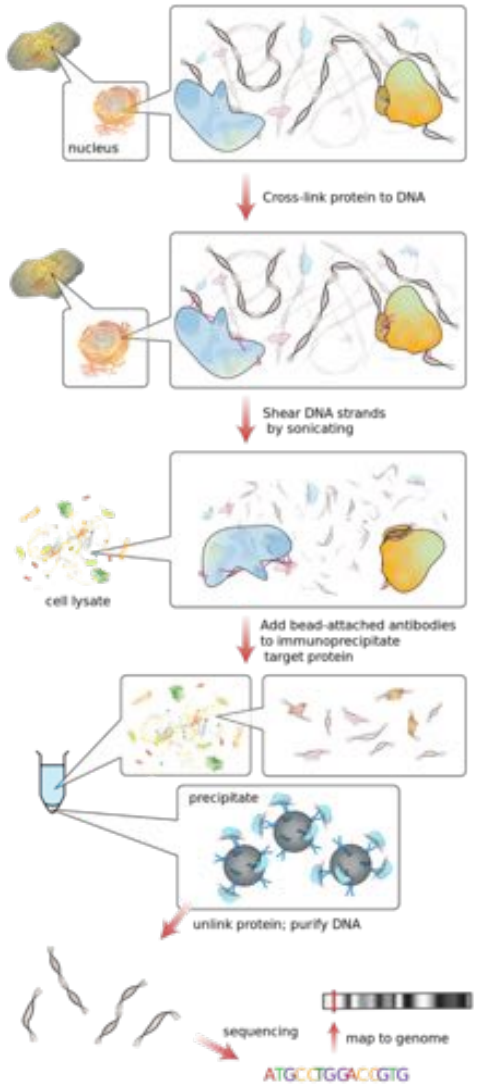
- One type of insulator establishes domains that separate enhancers and promoters to block their interaction,
- Second type creates a barrier against the spread of heterochromatin.

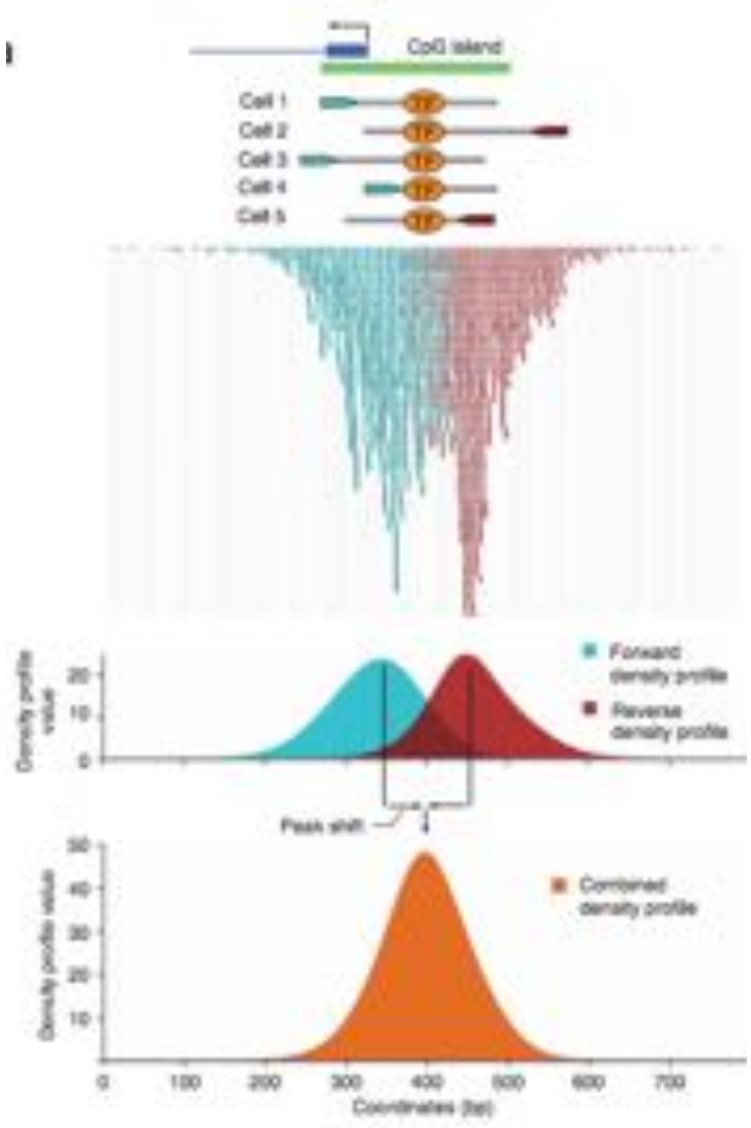
Insulators: exploiting transcriptional and epigenetic mechanisms
Gaszner & Felsenfeld (2006) *Nature Reviews Genetics* 7, 703-713. doi:10.1038/nrg1925

ChIP-seq: TF Binding

Goals:

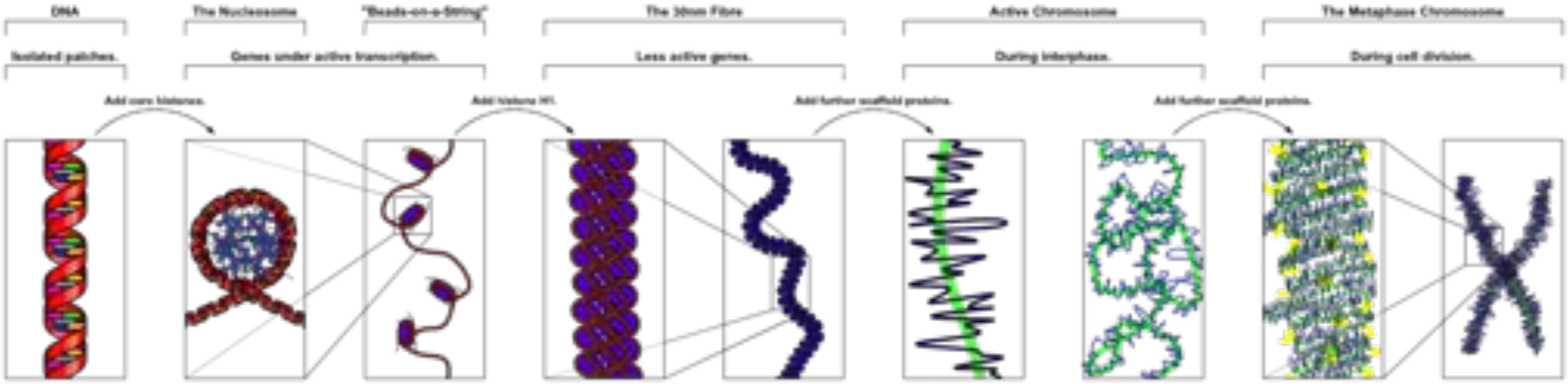
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?





Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data
Valouev et al (2008) *Nature Methods.* 5, 829 - 834

Chromatin compaction model



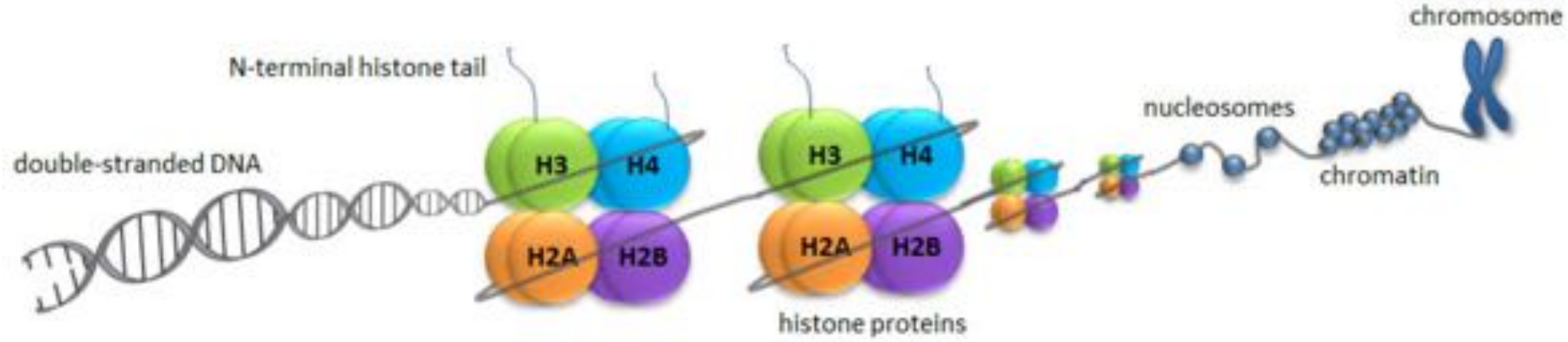
Nucleosome is a basic unit of DNA packaging in eukaryotes

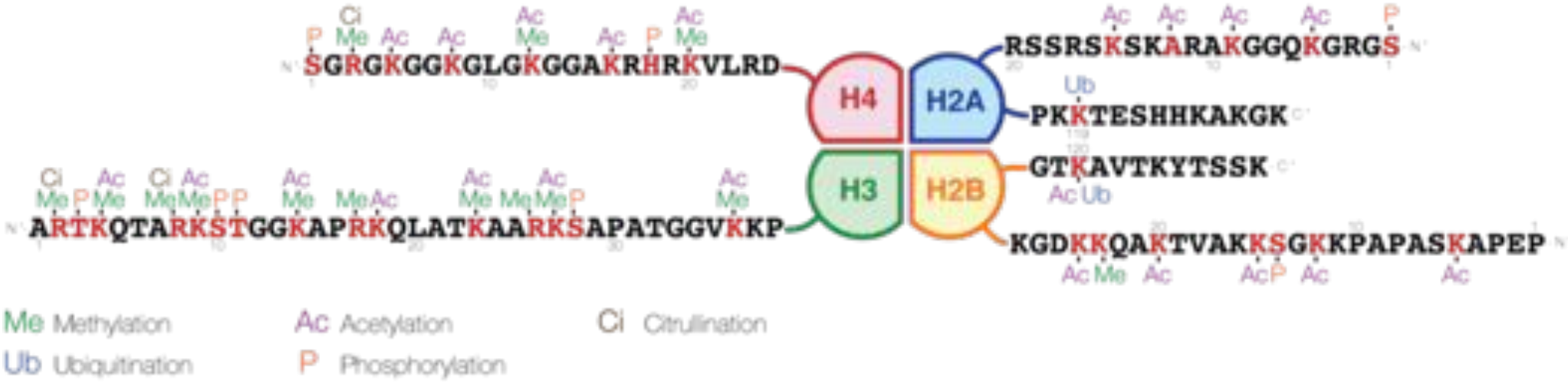
- Consists of a segment of 146bp DNA wound in sequence around eight histone protein cores (thread wrapped around a spool) followed by a ~38bp linker
- Under active transcription, nucleosomes appear as “beads-on-a-string”, but are more densely packed for less active genes

Nucleosomes form the fundamental repeating units of eukaryotic chromatin

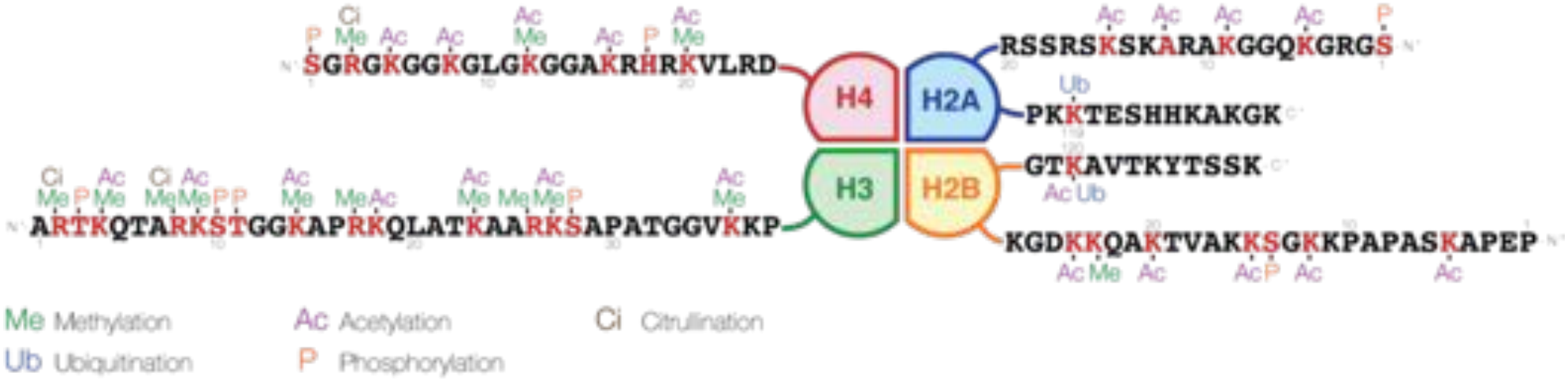
- Used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 μm diameter).

ChIP-seq: Histone Modifications





ChIP-seq: Histone Modifications

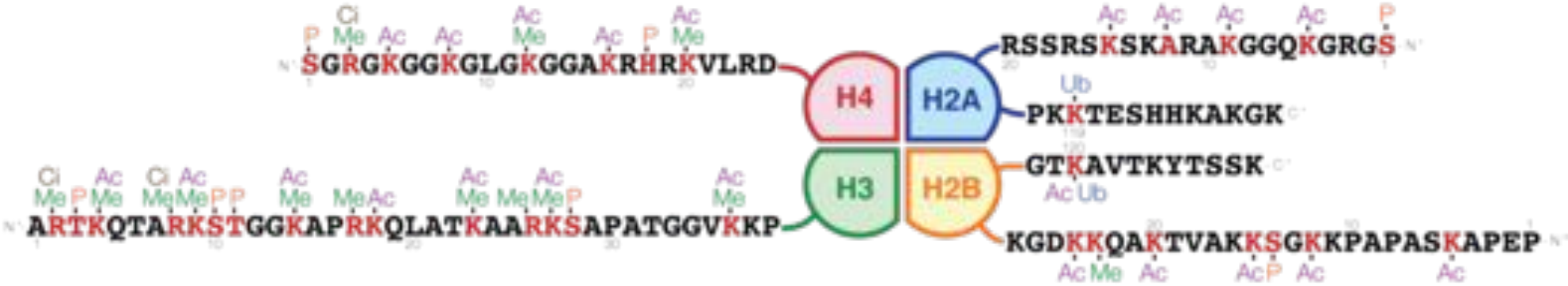


The common nomenclature of histone modifications is:

- The name of the histone (e.g., H3)
- The single-letter amino acid abbreviation (e.g., K for Lysine) and the amino acid position in the protein
- The type of modification (Me: methyl, P: phosphate, Ac: acetyl, Ub: ubiquitin)
- The number of modifications (only Me is known to occur in more than one copy per residue. 1, 2 or 3 is mono-, di- or tri-methylation)

So H3K4me1 denotes the monomethylation of the 4th residue (a lysine) from the start (i.e., the N-terminal) of the H3 protein.

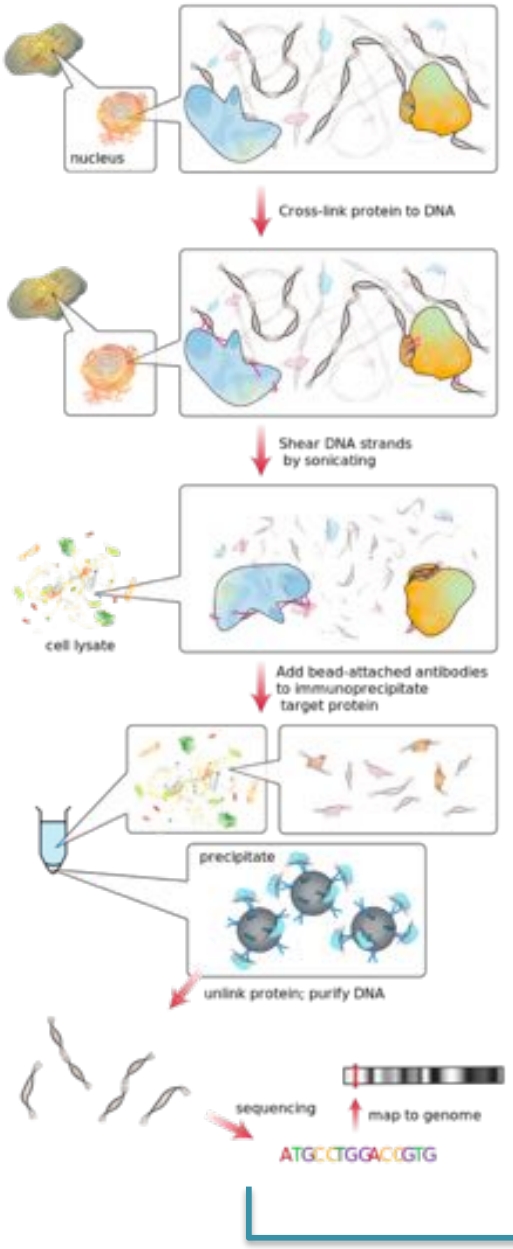
ChIP-seq: Histone Modifications



Type of modification	Histone							
	H3K4	H3K9	H3K14	H3K27	H3K79	H3K122	H4K20	H2BK5
mono-methylation	activation[6]	activation[7]		activation[7]	activation[7][8]		activation[7]	activation[7]
di-methylation	activation	repression[3]		repression[3]	activation[8]			
tri-methylation	activation[9]	repression[7]		repression[7]	activation,[8] repression[7]			repression[3]
acetylation		activation[9]	activation[9]	activation[10]		activation[11]		

- H3K4me3 is enriched in transcriptionally active promoters.[12]
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found in actively transcribed promoters.
- H3K27ac distinguishes active enhancers from poised enhancers.
- H3K122ac is enriched in poised promoters and also found in a different type of putative enhancer that lacks H3K27ac.

General Flow of ChIP-seq Analysis



Peak Calling

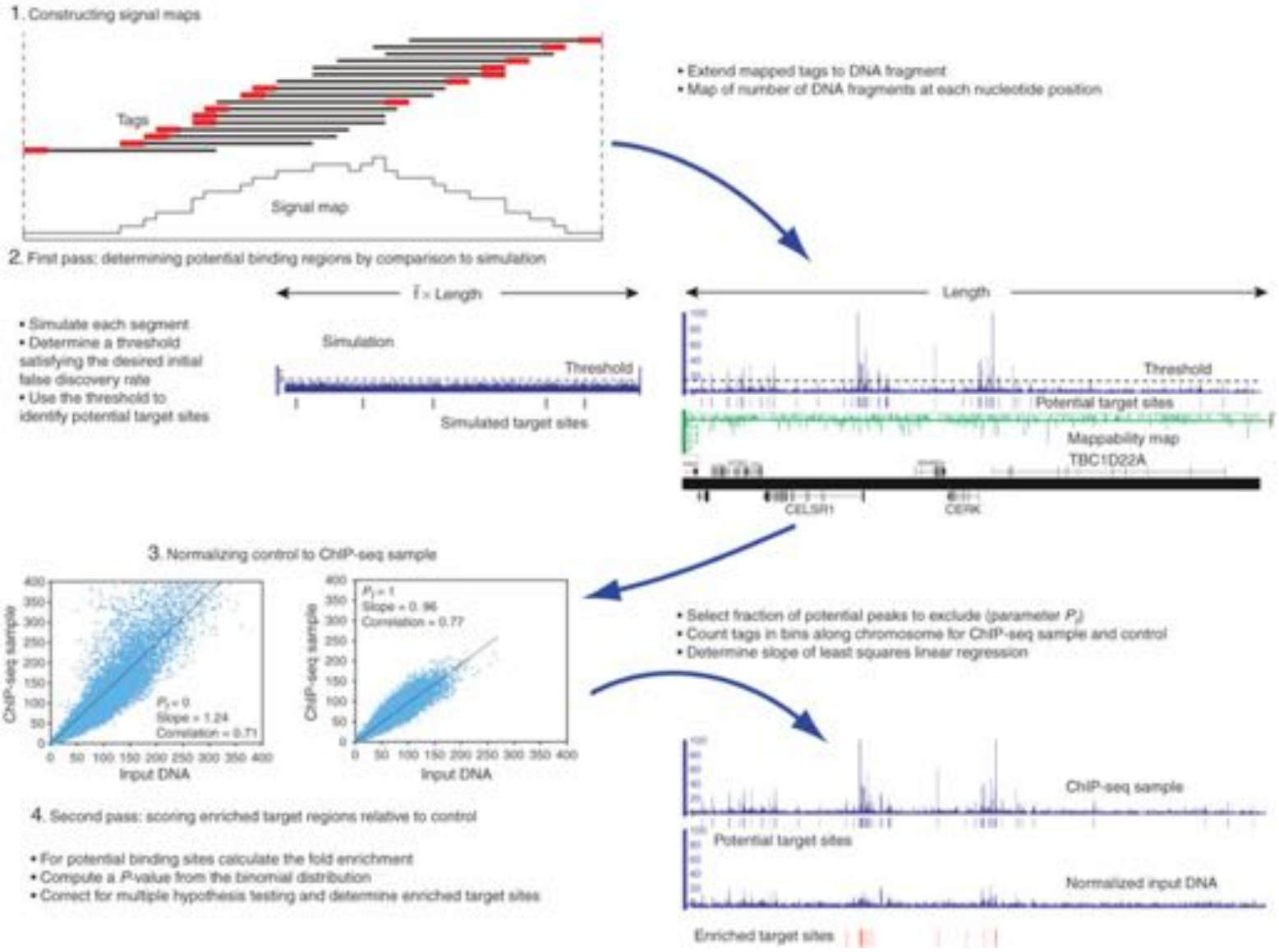
Analyze Enriched Regions

Visualization

Peak-peak comparisons

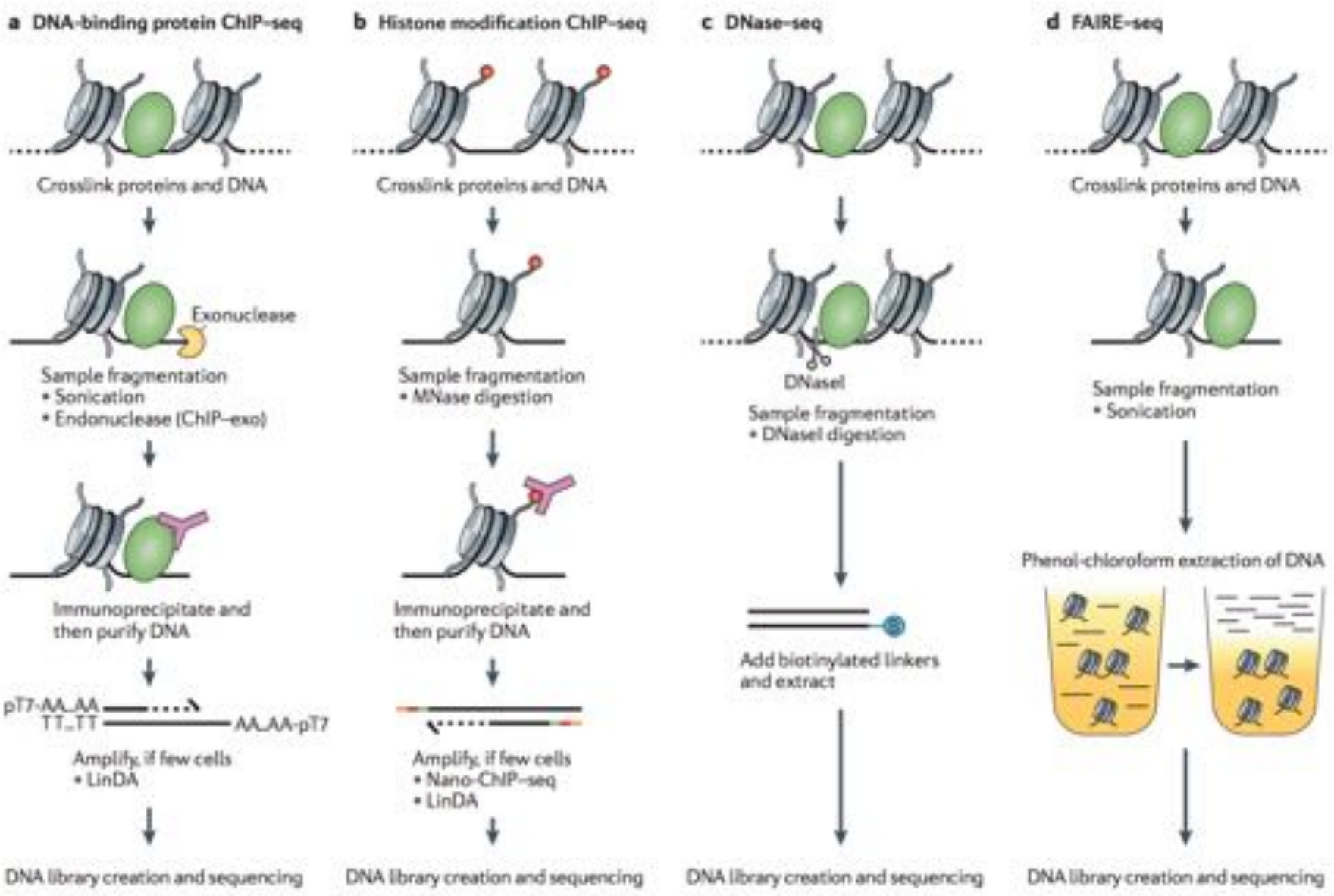
Annotation / Interpretation

PeakSeq



PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls
Rozowsky et al (2009) Nature Biotechnology 27, 66 - 75

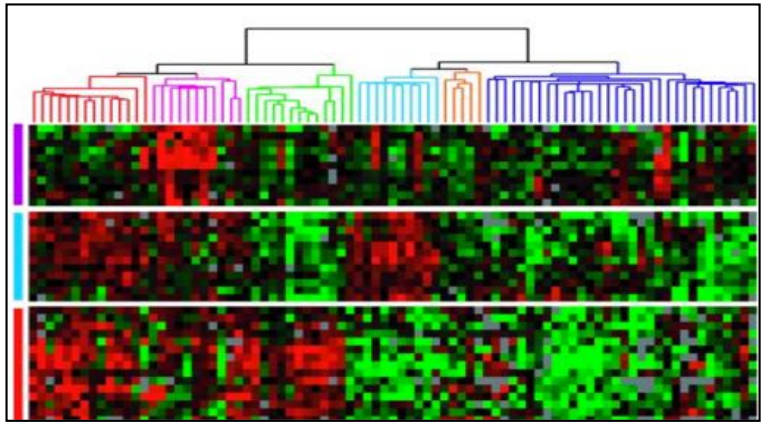
Related Assays



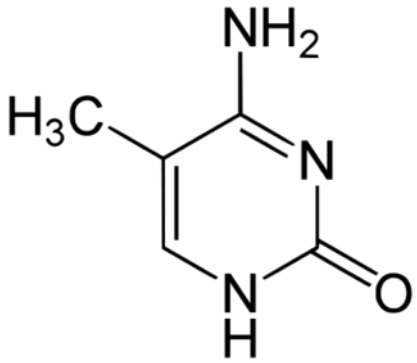
ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions
Furey (2012) *Nature Reviews Genetics*. 13, 840-852

*-seq in 4 short vignettes

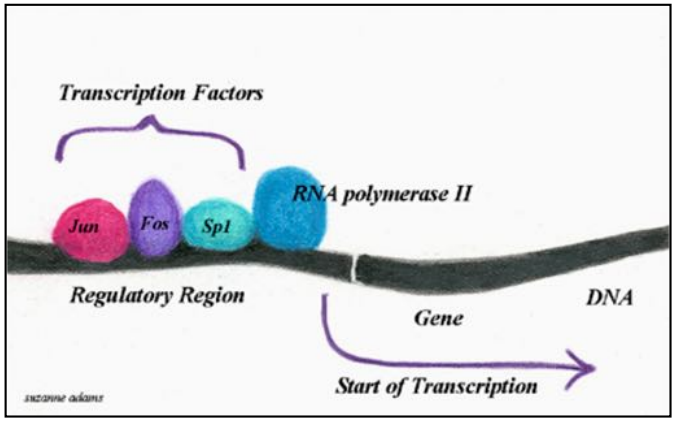
RNA-seq



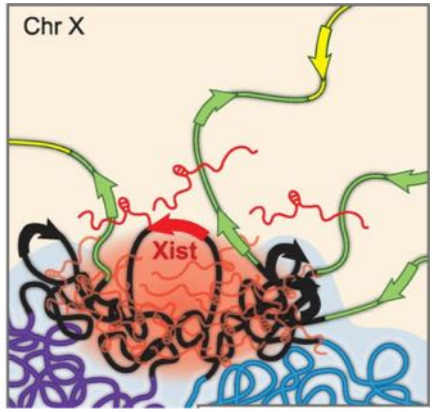
Methyl-seq



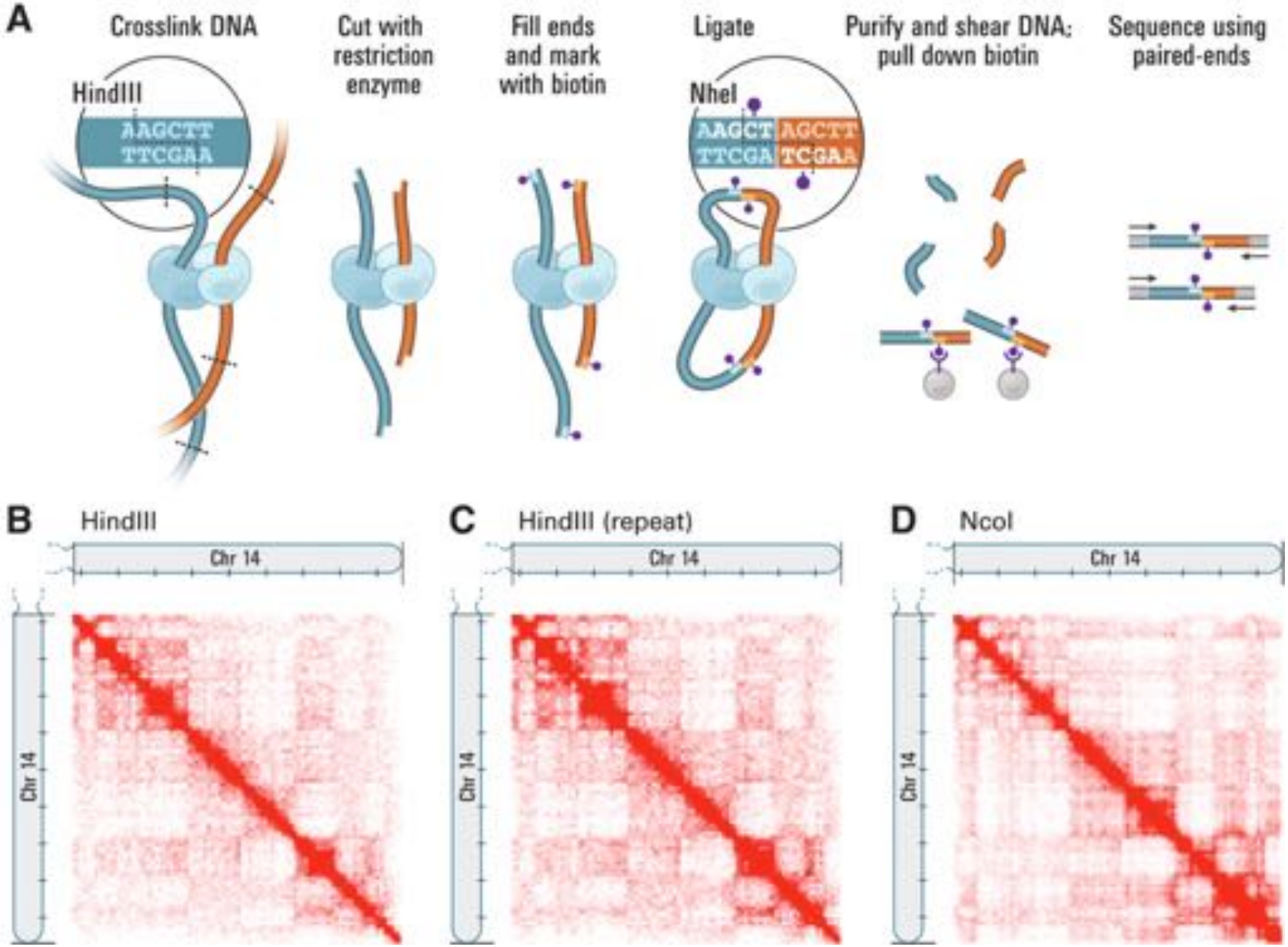
ChIP-seq



Hi-C

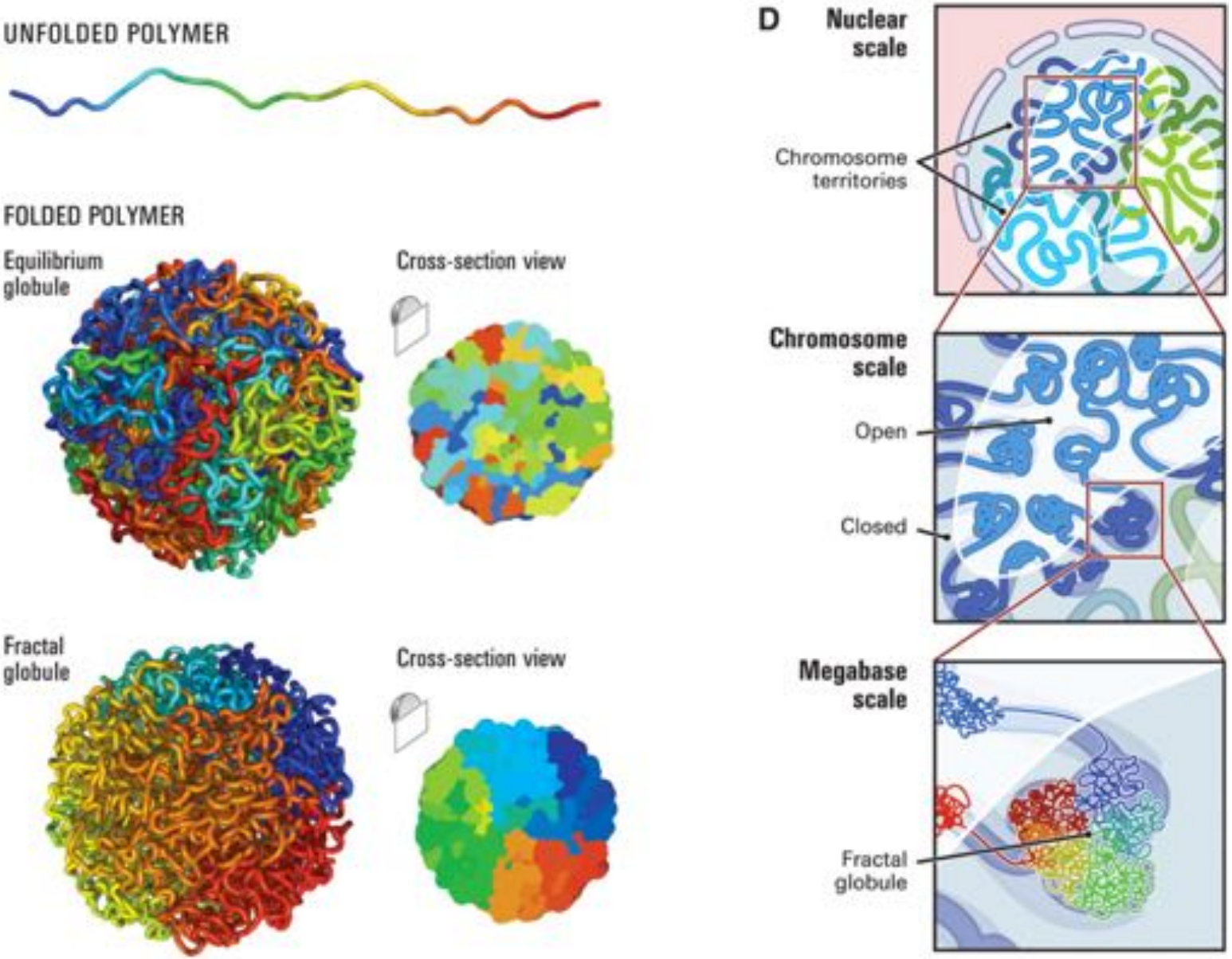


HI-C: Mapping the folding of DNA



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome
Liberman-Aiden et al. (2009) *Science*. 326 (5950): 289-293

HI-C: Mapping the folding of DNA



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome
Liberman-Aiden et al. (2009) *Science*. 326 (5950): 289-293

Gene Regulation in 3-dimensions

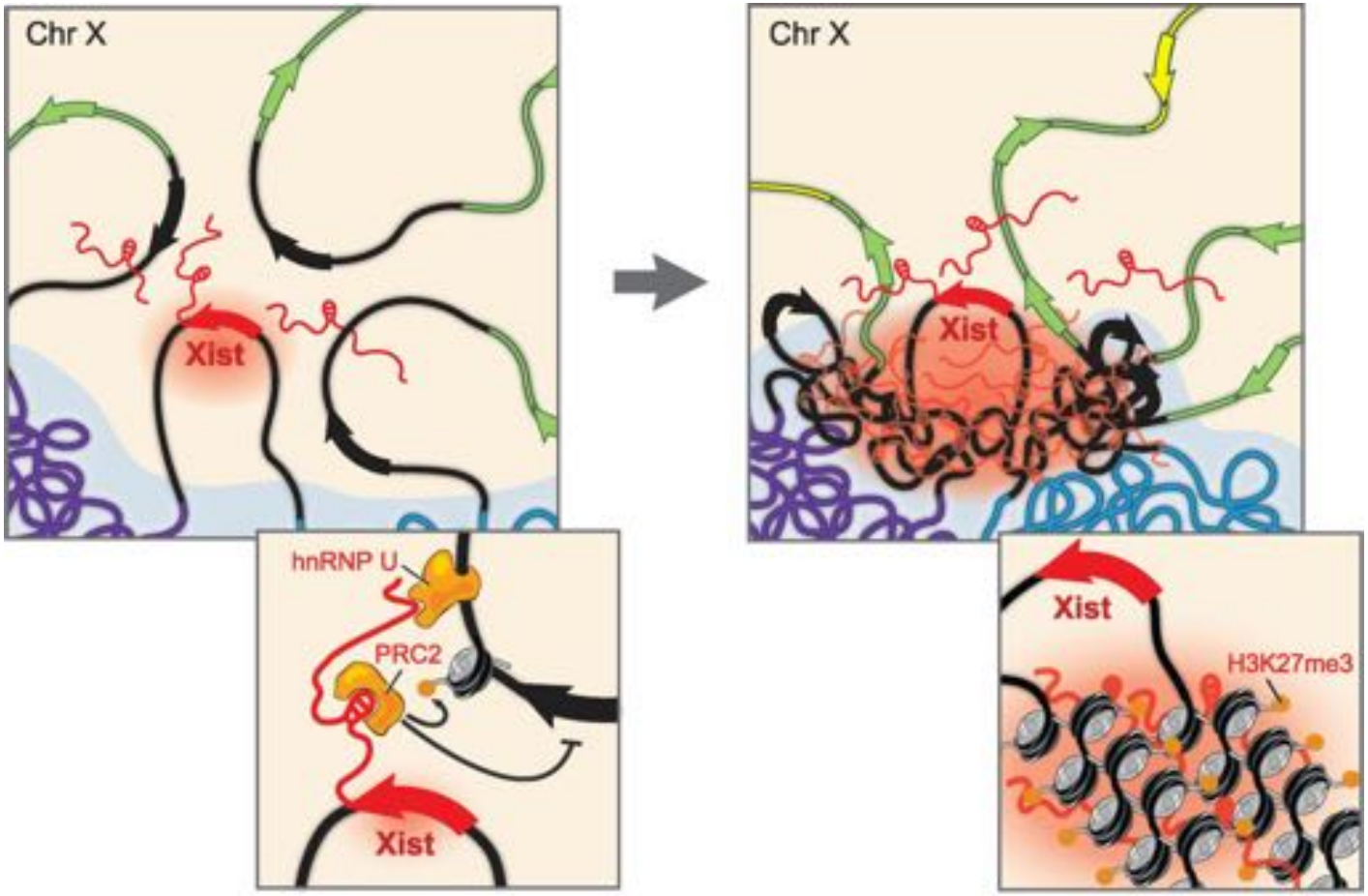
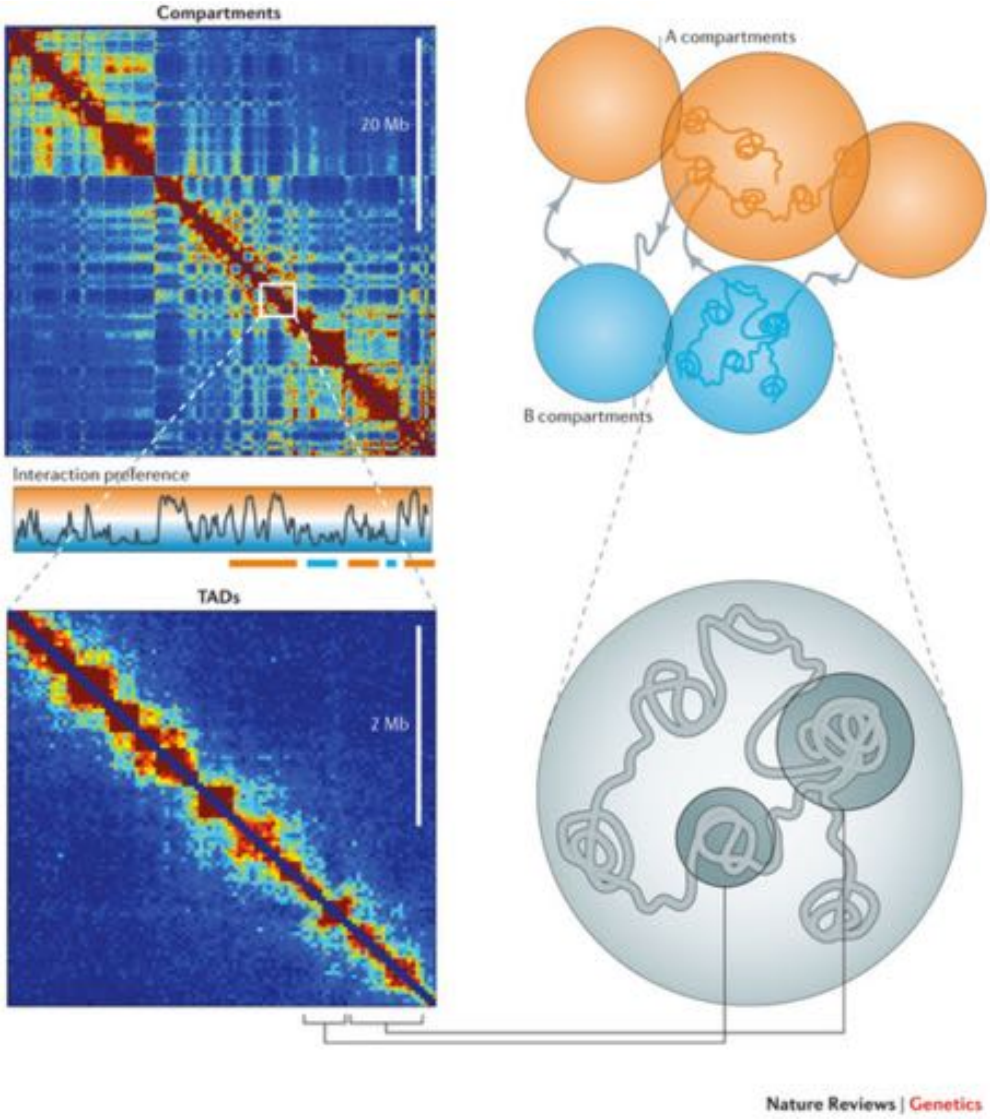


Fig 6. A model for how Xist exploits and alters three-dimensional genome architecture to spread across the X chromosome.

The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome
Engreitz et al. (2013) *Science*. 341 (6147)

Genome compartments & TADs



Mammalian genomes have a pattern of interactions that can be approximated by two compartments called A and B

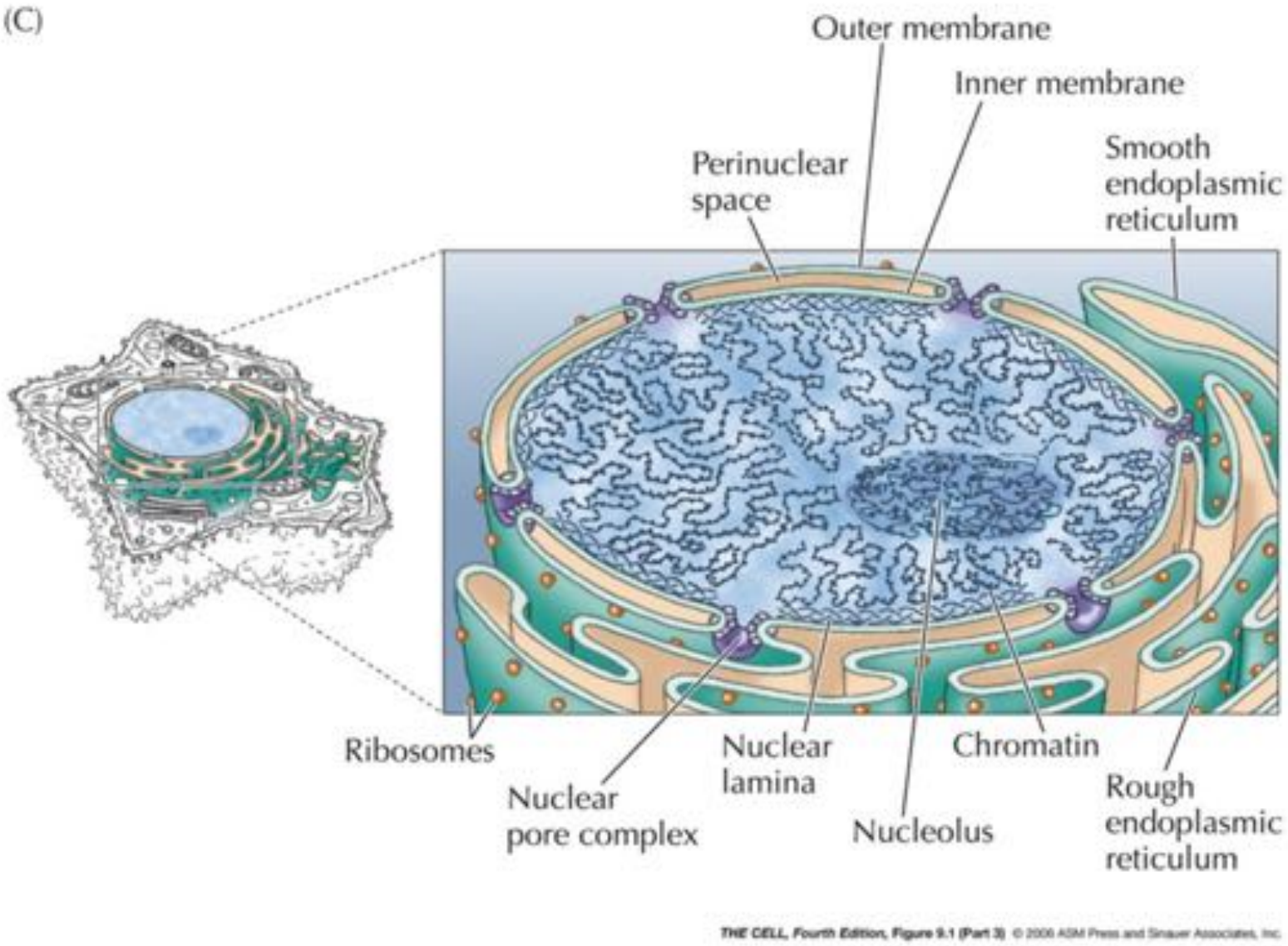
- alternate along chromosomes and have a characteristic size of ~5 Mb each.
- A compartments (orange) preferentially interact with other A compartments; B compartments (blue) associate with other B compartments.
- A compartments are largely euchromatic, transcriptionally active regions.

Topologically associating domains (TADs)

- TADs are smaller (~400–500 kb)
- Can be active or inactive, and adjacent TADs are not necessarily of opposite chromatin status.
- TADs are hard-wired features of chromosomes, and groups of adjacent TADs can organize in A and B compartments

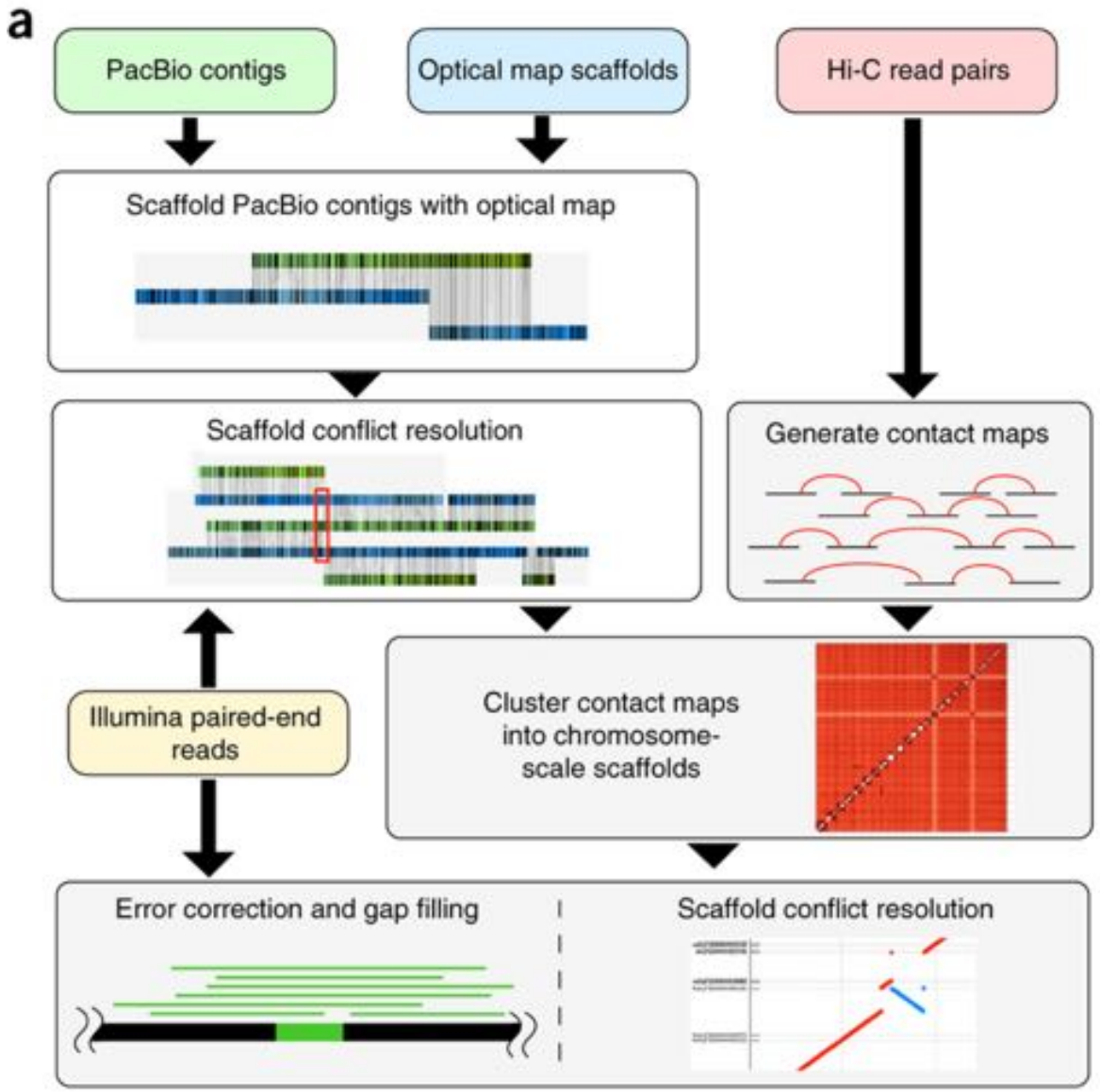
Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data
Dekker et al. (2013) *Nature Reviews Genetics 14, 390–403*

“Lamina-Associated Domains are the B compartment”



Chromosome Conformation Paints Reveal the Role of Lamina Association in Genome Organization and Regulation
Luperchio et al. (2017) bioRxiv. *doi: https://doi.org/10.1101/122226*

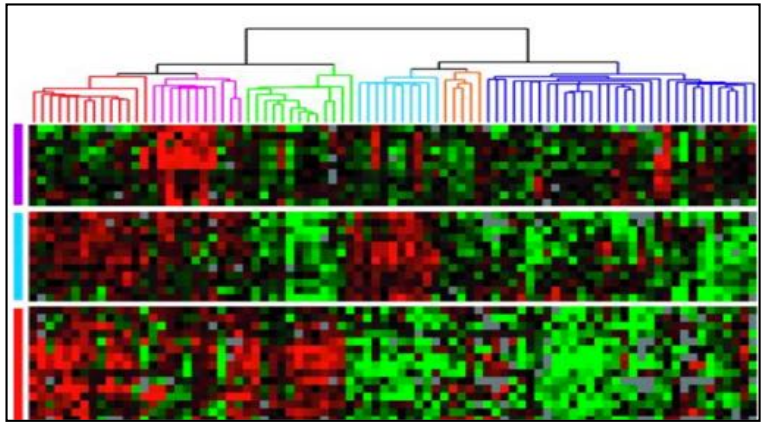
Scaffolding with Hi-C



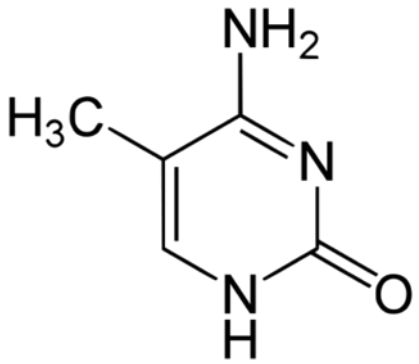
Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome
Bickhart et al (2017) Nature Genetics (2017) doi:10.1038/ng.3802

Putting it all together!

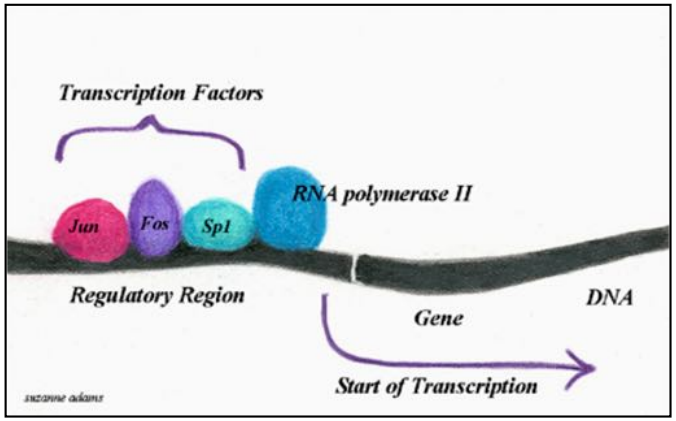
RNA-seq



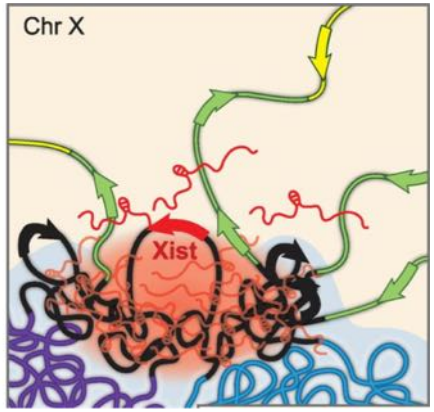
Methyl-seq



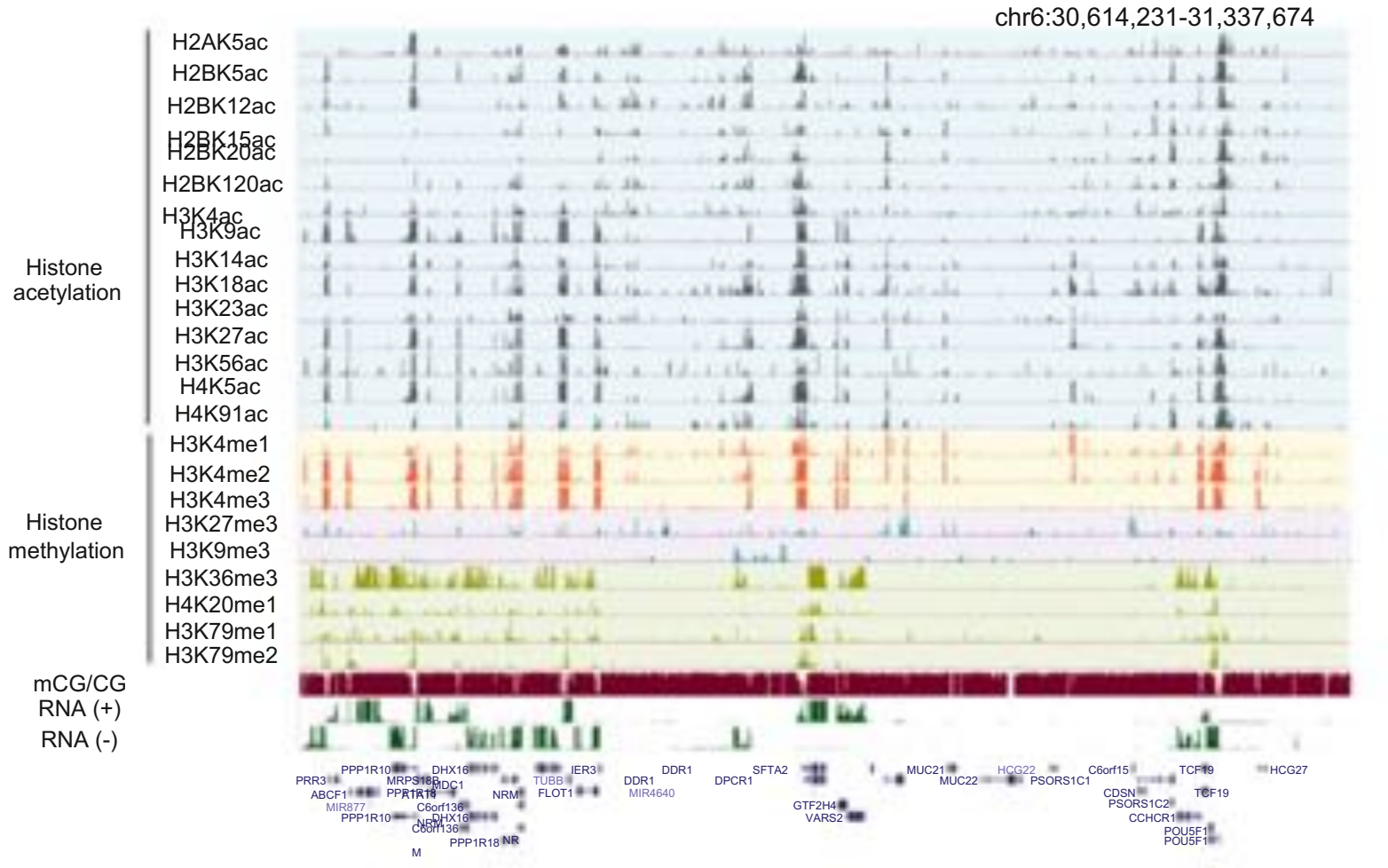
ChIP-seq



Hi-C

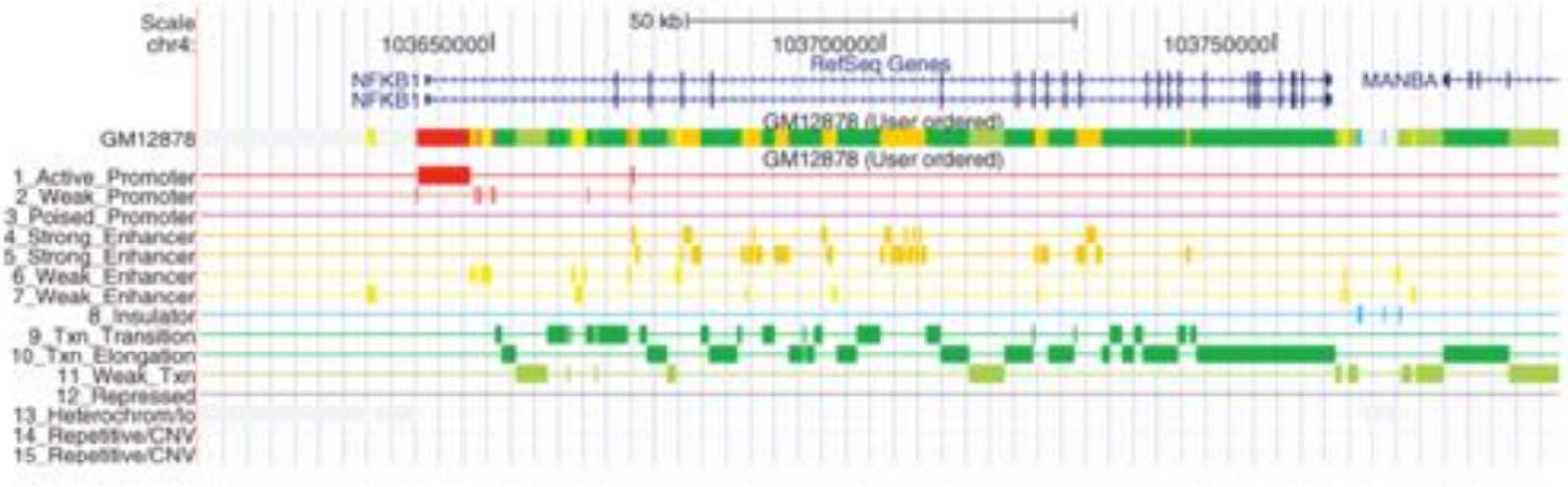


We can call peaks, but…



We need a way to summarize the combinatorial patterns of multiple histone marks into meaningful biological units

ChromHMM



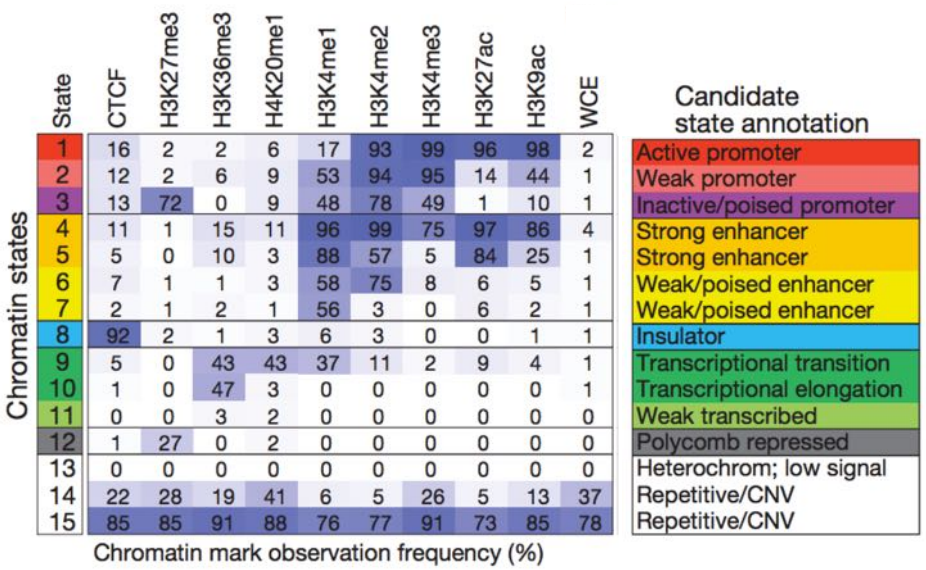
ChromHMM is software for learning and characterizing chromatin states.

- ChromHMM can integrate multiple chromatin datasets such as ChIP-seq data of various histone modifications to discover de novo the major re-occuring combinatorial and spatial patterns of marks.
- ChromHMM is based on a multivariate Hidden Markov Model that explicitly models the presence or absence of each chromatin mark.
- The resulting model can then be used to systematically annotate a genome in one or more cell types.

ChromHMM: automating chromatin-state discovery and characterization
Ernst & Kellis (2012) Nature Methods 9, 215–216. doi:10.1038/nmeth.1906

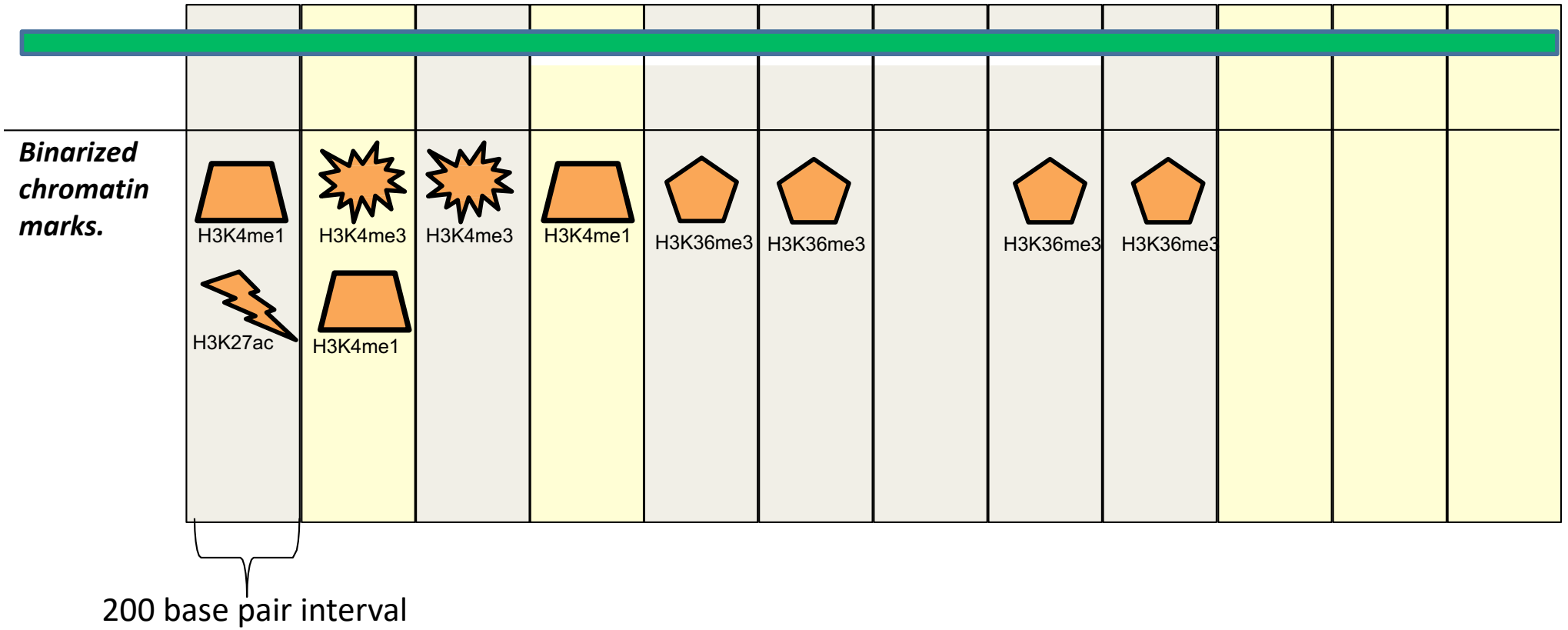
ChromHMM and Chromatin States

Chromatin states are defined based on different combinations of histone modifications and correspond to different functional regions



The goal is to segment every base of the the genome into biologically meaningful units: reveal & annotate *functional* elements

ChromHMM : Multivariate Hidden Markov Model

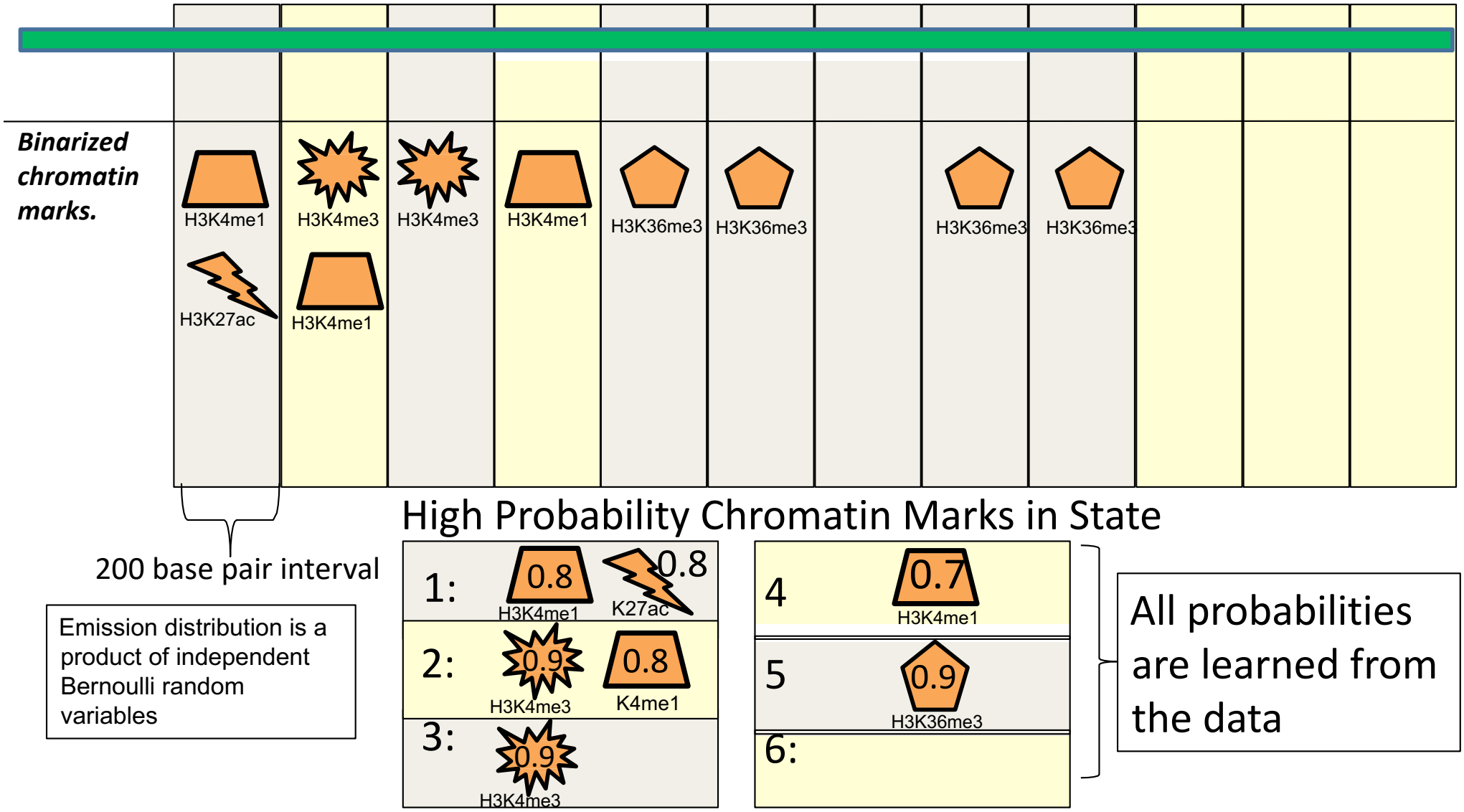


Emission distribution is a product of independent Bernoulli random variables

Binarization leads to explicit modeling of mark combinations and interpretable parameters

Ernst and Kellis, Nat Biotech 2010 ; Ernst and Kellis, Nature Methods 2012

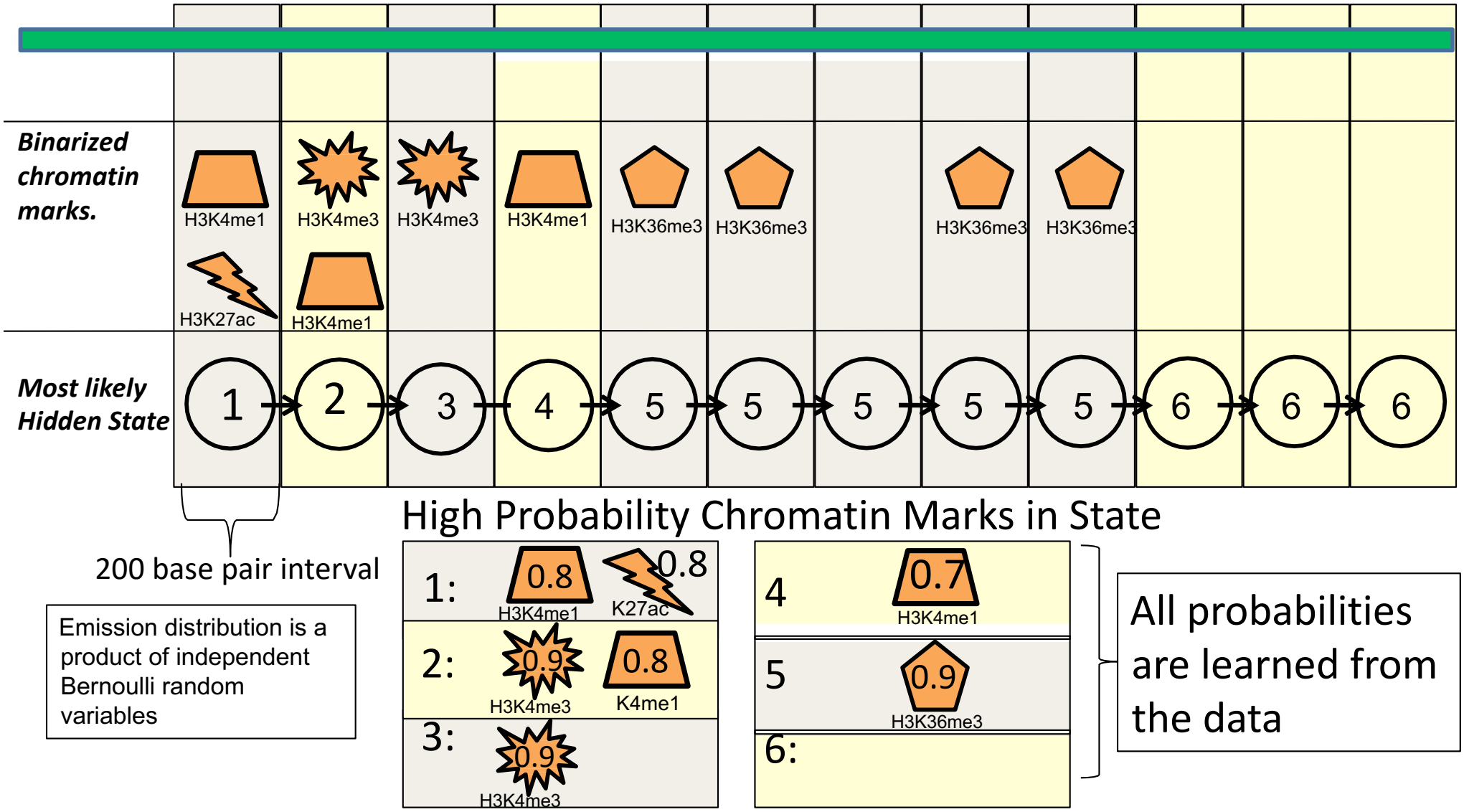
ChromHMM : Multivariate Hidden Markov Model



Binarization leads to explicit modeling of mark combinations and interpretable parameters

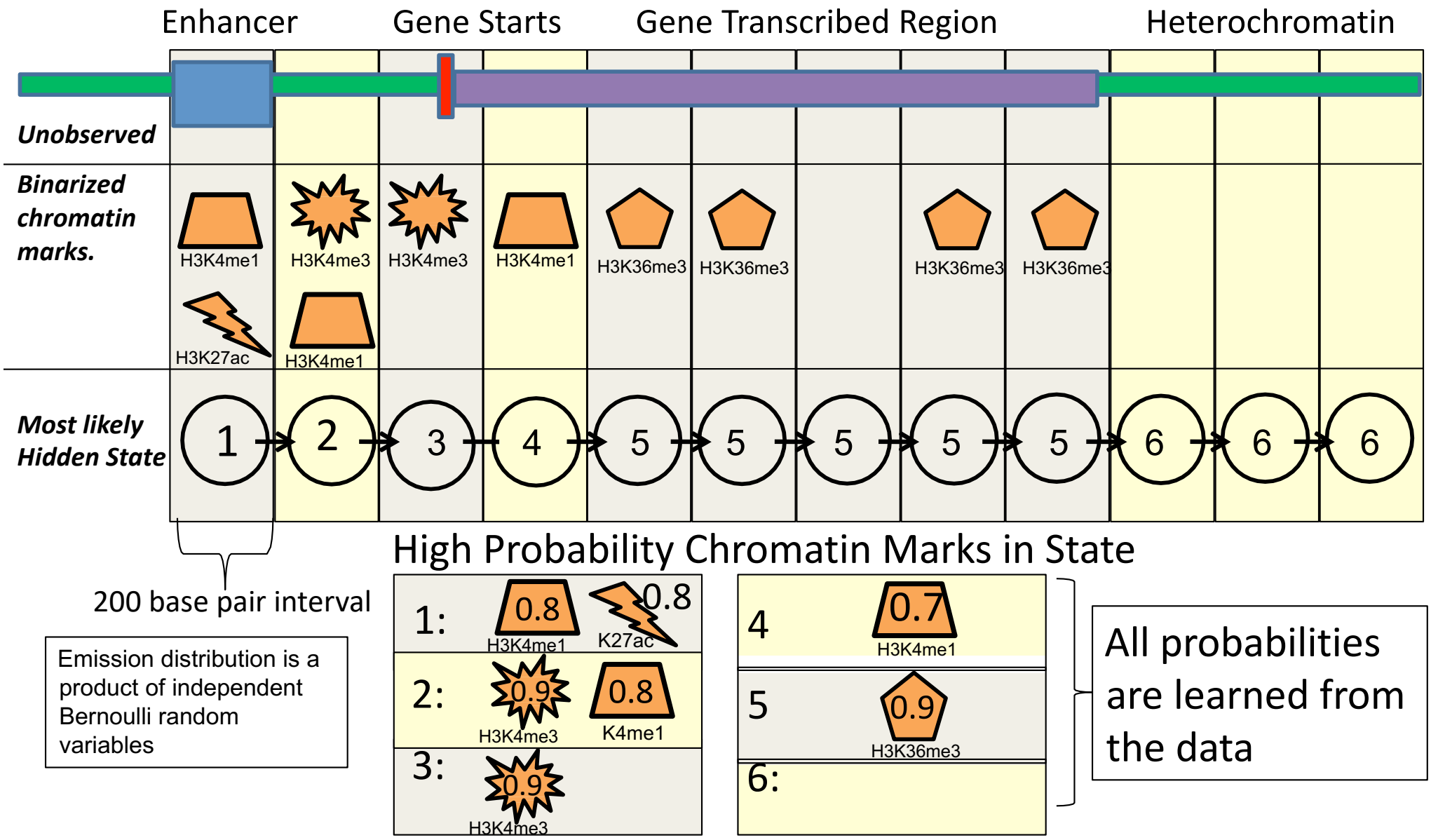
Ernst and Kellis, Nat Biotech 2010 ; Ernst and Kellis, Nature Methods 2012

ChromHMM : Multivariate Hidden Markov Model



Ernst and Kellis, Nat Biotech 2010 ; Ernst and Kellis, Nature Methods 2012

ChromHMM : Multivariate Hidden Markov Model



Binarization leads to explicit modeling of mark combinations and interpretable parameters

Ernst and Kellis, Nat Biotech 2010 ; Ernst and Kellis, Nature Methods 2012

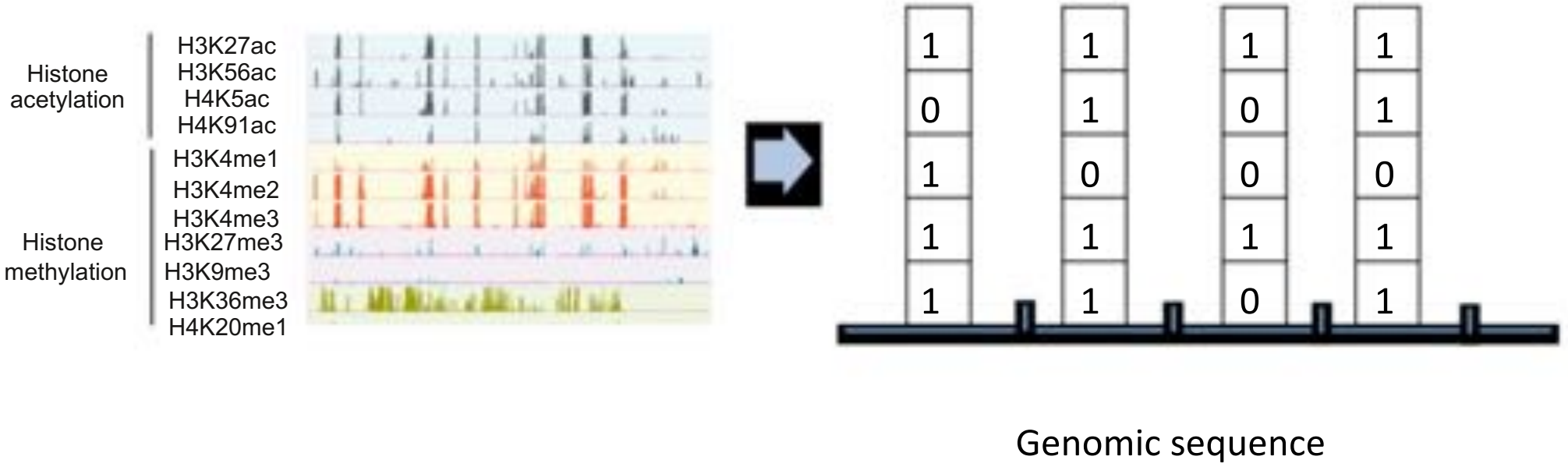
The Workflow

1. Get ChIP-seq raw reads for different histone modifications
2. Align the reads to a reference genome
3. Convert aligned reads in bed format
4. Create Binned and Binarized Tracks
5. Train the model
6. Infer the states
7. Interpretation

Thanks Sri!

Create Binned and Binarized Tracks

- ChromHMM quantify the presence or absence of each mark in bins of fixed size



Genomic sequence

Train the model and segment the genome

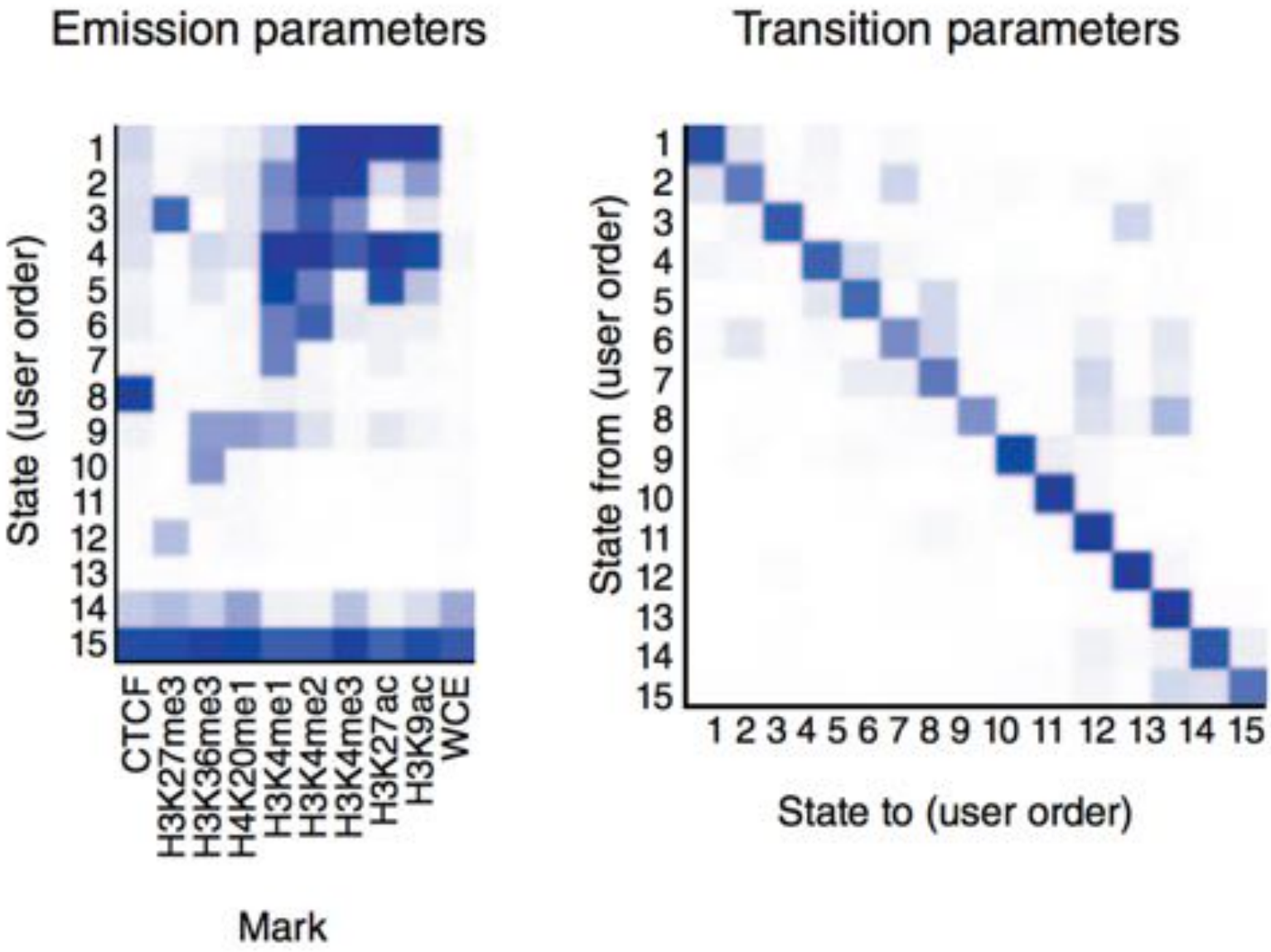

```
java -mx1600M -jar ChromHMM.jar LearnModel
SAMPLEDATA_HG18 OUTPUTSAMPLE 10 hg18
```

Output of ChromHMM

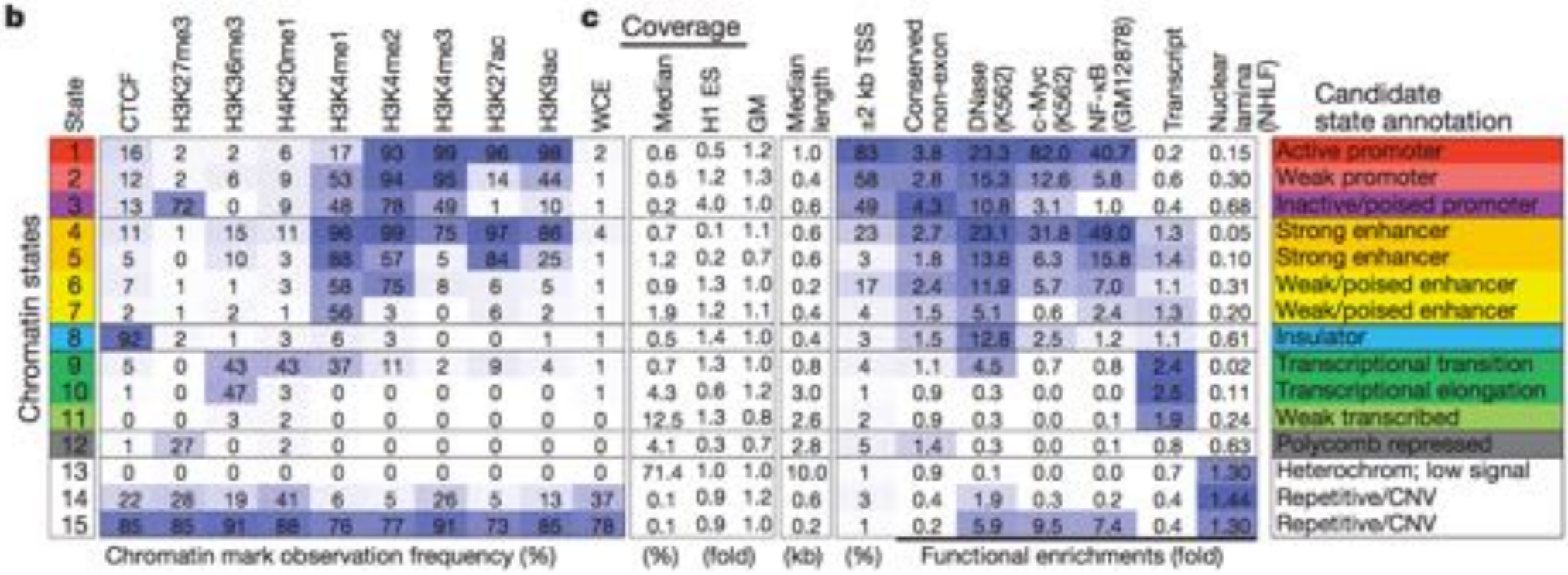
- ChromHMM generates an HTML report called **webpage_N.html** (N is the number of states used) with many useful information :

1. Model learned: transition and emission parameters
2. Enriched functional categories
3. Bed files to visualize the segmentation

Transition and Emission Parameters



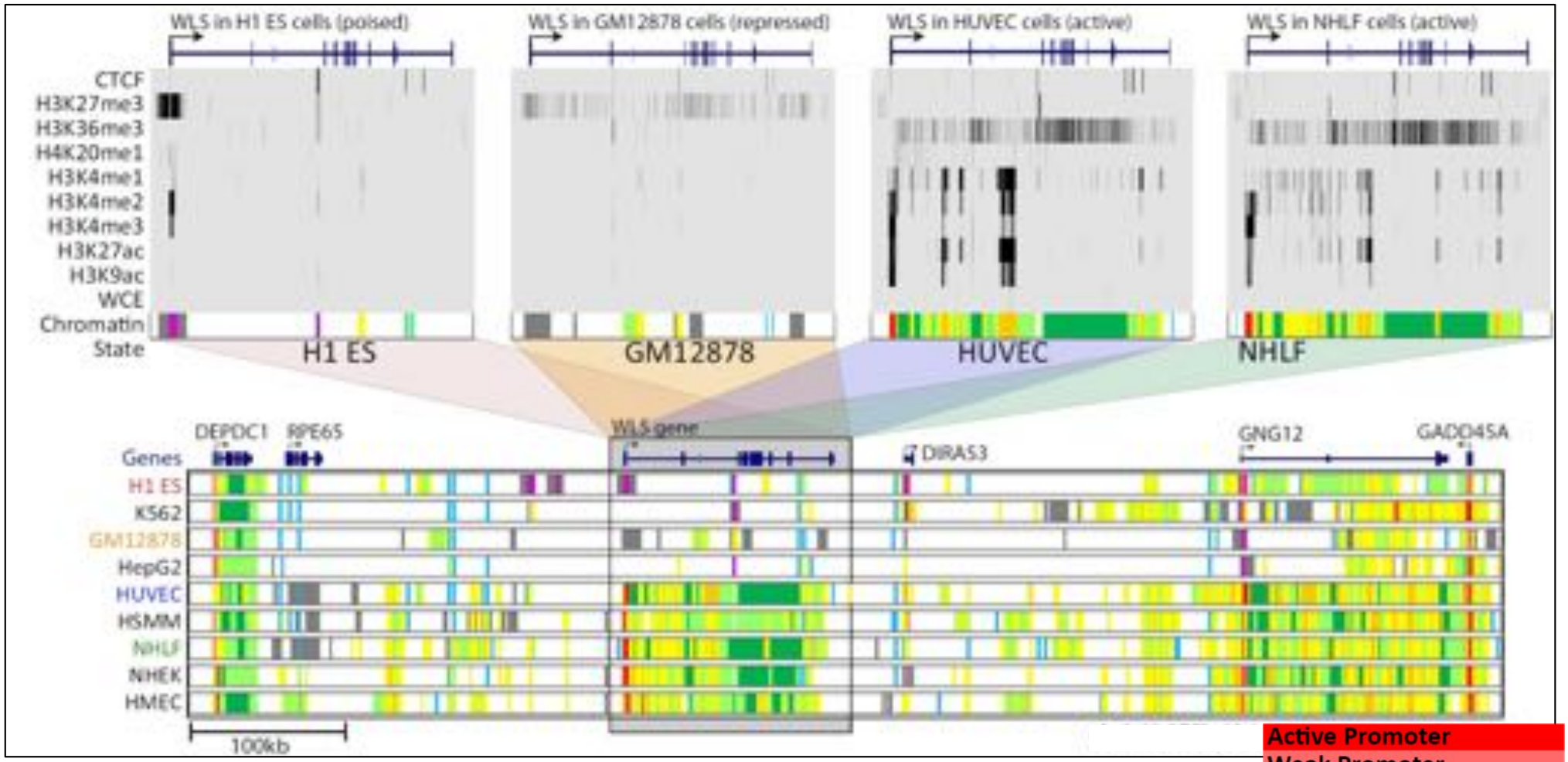
Enriched functional category



The states predicted by the HMM are ***statistical*** entities (#1 – #15)
The states we want are ***biological*** entities (Active/Weak/Poised promoter)

Investigate the properties of the statistical entities to label them with biological functions
=> Supervised learning problem ☺

Chromatin states dynamics across nine ENCODE cell types



- **Single annotation track for each cell type**
- **Summarize cell-type activity at a glance**
- **Can study 9-cell activity pattern across ↓**

Ernst et al, Nature 2011