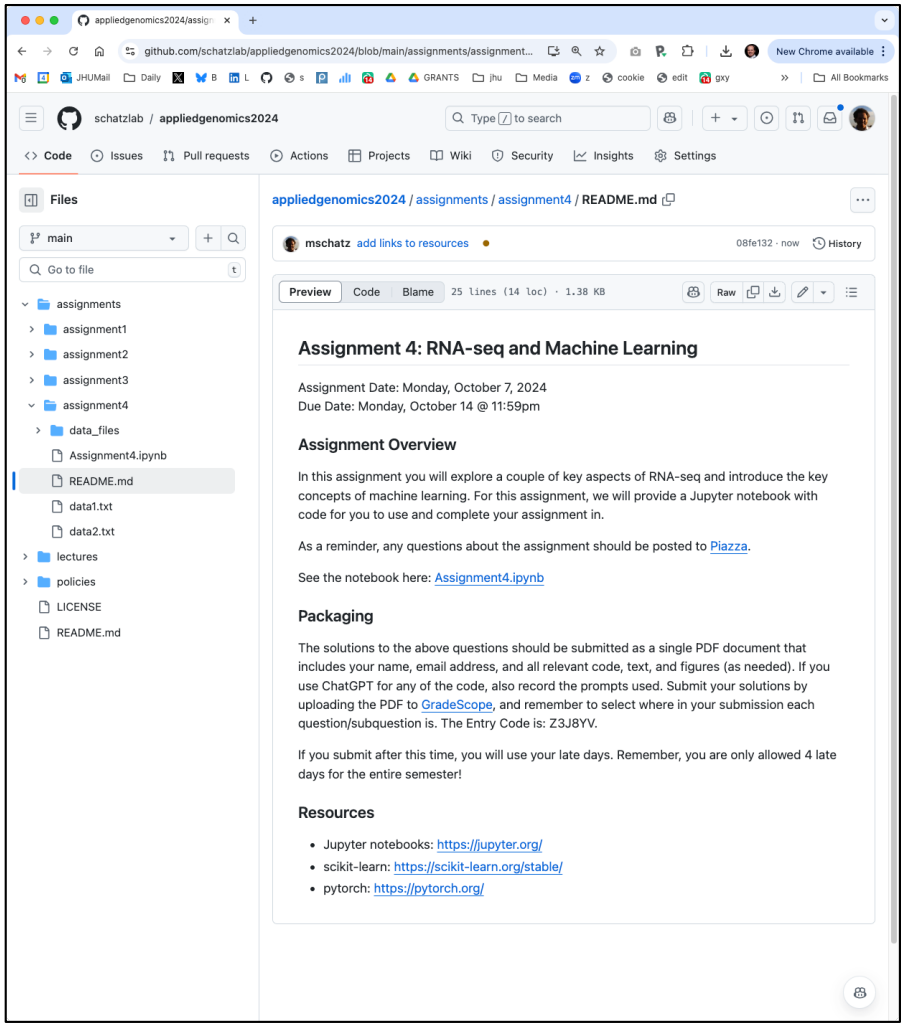
Epigenomics

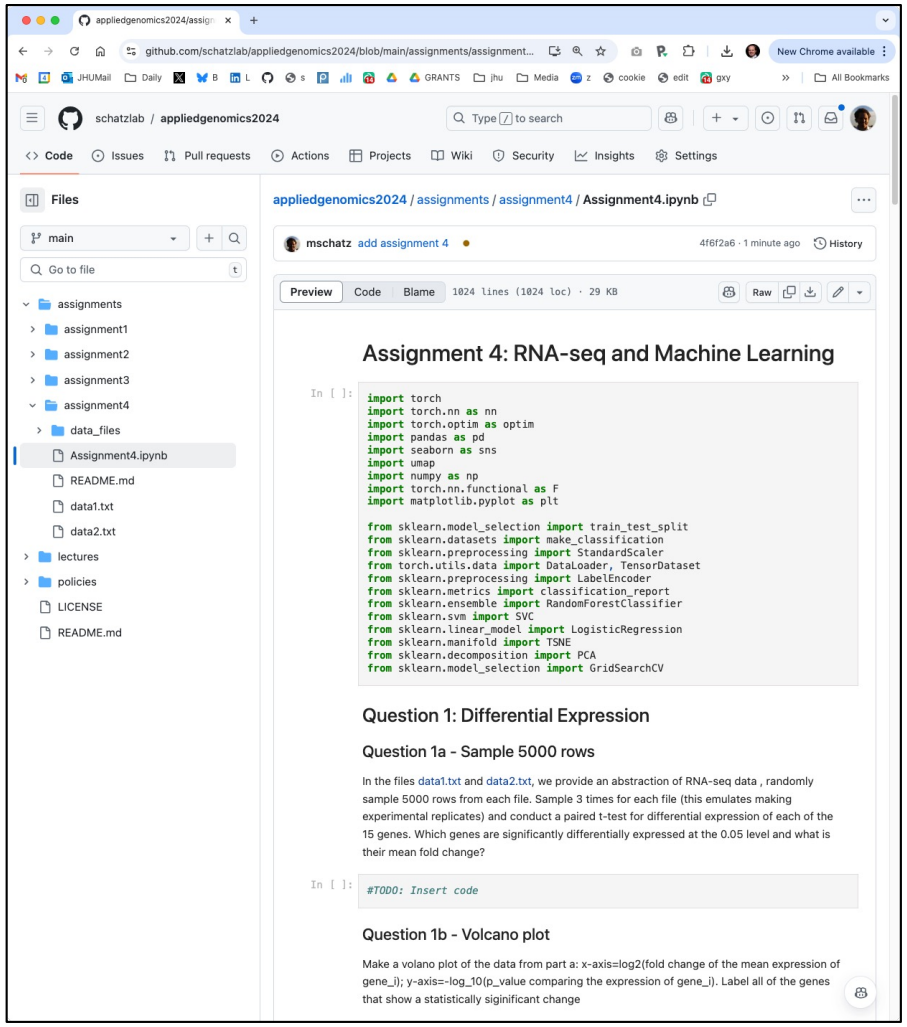
Michael Schatz

October 14, 2024

Lecture 14. Applied Comparative Genomics

Assignment 4
Due: Monday Oct 14, 2024 by 11:59pm





https://schatz-lab.org/appliedgenomics2024/assignments/assignment4/

Check Piazza for questions!

Project Proposal
Due: Monday Oct 21, 2024 by 11:59pm

Project Proposal

Assignment Date: Monday October 14, 2024
Due Date: Monday, October 21 2024 @ 11:59pm

Review the Project Ideas page

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

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 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)
- Please add a note if you need me to sponsor you for a MARCC account (high RAM, GPUs, many cores, etc)
- Please add a note if you need me to sponsor you for an AnVIL cloud billing account (dynamic resources)

Submit the proposal as a 1 to 2 page PDF on GradeScope (each team should submit one proposal and tag all people in the team). After submitting your proposal, I will provide feedback. If necessary, we can 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.

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!

https://github.com/schatzlab/appliedgenomics2024/blob/main/project/proposal.md

Check Piazza for questions!

Class Schedule

M	Oct 14	Epigenome	Project Proposal Assigned
W	Oct 16	Single cell	
M	Oct 21	Transformers	Assignment 5 Assigned
W	Oct 23	Enformer	
M	Oct 28	DL in Genomics	Preliminary Report Assigned
W	Oct 30	Midterm Review	
M	Nov 4	Midterm!	
W	Nov 6	Disease Genomics	
M	Nov 11	Metagenomics	Final Report Assigned
W	Nov 13	No Class (BIODATA24)	
M	Nov 18	Cancer Genomics	
W	Nov 20	Project Presentation 1	
M	Nov 25	Thanksgiving Break	
W	Nov 27	Thanksgiving Break	
M	Dec 2	Project Presentation 2	
W	Dec 4	Project Presentation 3	
M	Dec 16	Project Report Due	

Midterm

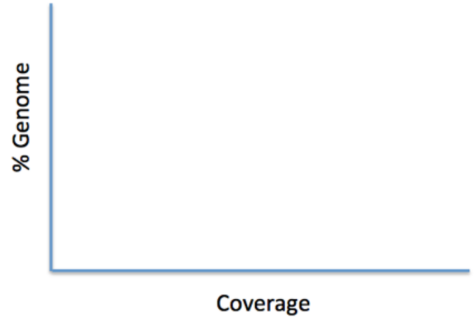
In class exam on November 4

Short answer & analysis
- No coding or complex calculations
- Bring a pen!

One page of notes
- Equations, diagrams, key facts, algorithms, data structures
- Recommend powerpoint ☺

Review the assignments & lectures to identify key topics!
- Review primary papers for more details

Q3. The Maryland blue crab genome is 1 Gbp in size. Approximately how many 100bp reads should we sequence so that we expect at least 99.85% of the genome will be sequenced at least 40 times? Sketch the expected coverage distribution for this number of reads; be sure to clearly label the mean coverage, and how 40 fold coverage relates to the mean. (Hint: In a normal distribution, 68.2% of the data is within 1 standard deviation of the mean, 95.4% within 2, 99.7% within 3, and 99.9% within 4)

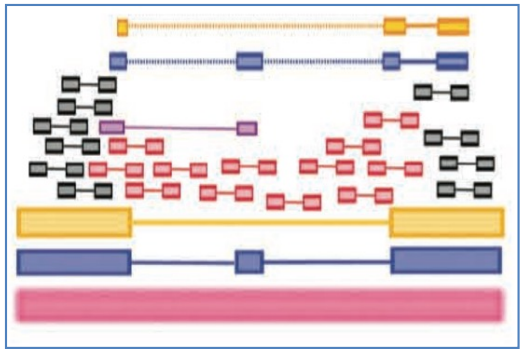


Annotation Summary

- Three major approaches to annotate a genome
  1. Alignment:
     - Does this sequence align to any other sequences of known function?
     - Great for projecting knowledge from one species to another
  2. Prediction:
     - Does this sequence statistically resemble other known sequences?
     - Potentially most flexible but dependent on good training data
  3. Experimental:
     - Lets test to see if it is transcribed/methylated/bound/etc
     - Strongest but expensive and context dependent

- Many great resources available
  - Learn to love the literature and the databases
  - Standard formats let you rapidly query and cross reference
  - Google is your number one resource ☺

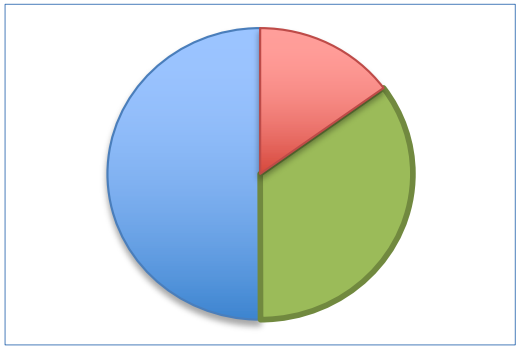
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

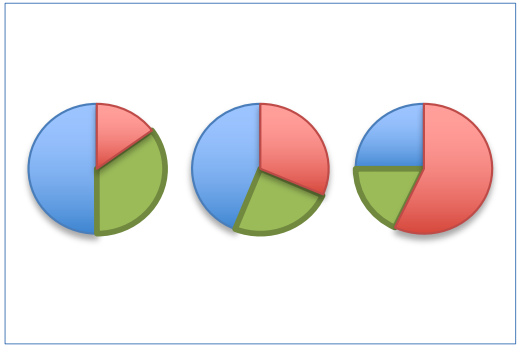
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. TPM)

Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



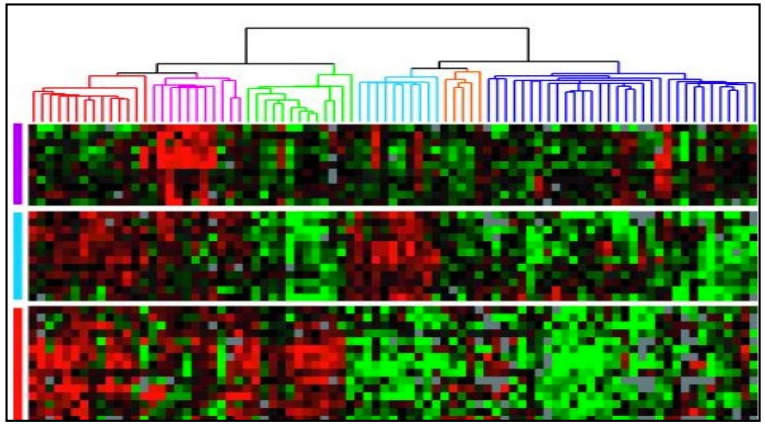
Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

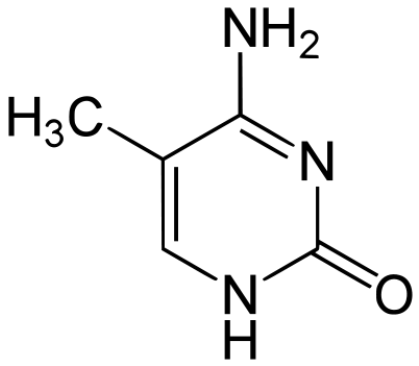
RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt688

*-seq in 4 short vignettes

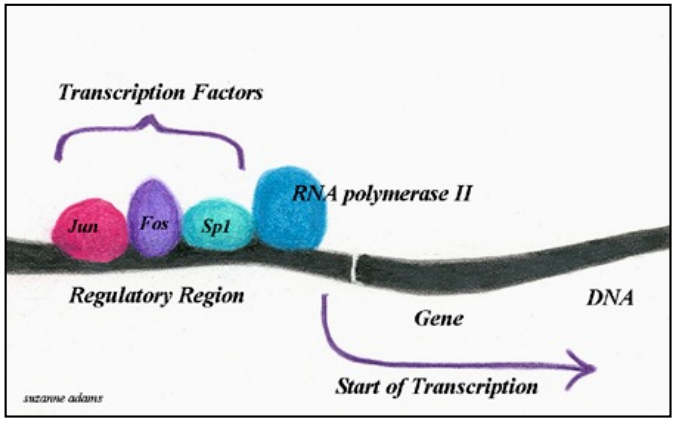
RNA-seq



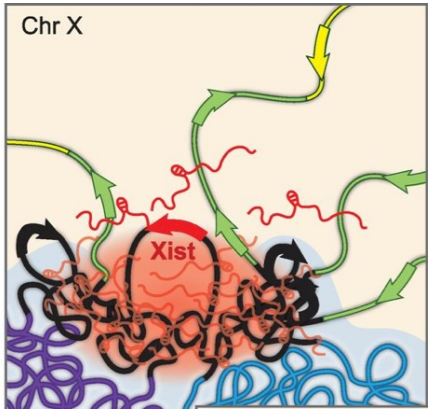
Methyl-seq



ChIP-seq



Hi-C



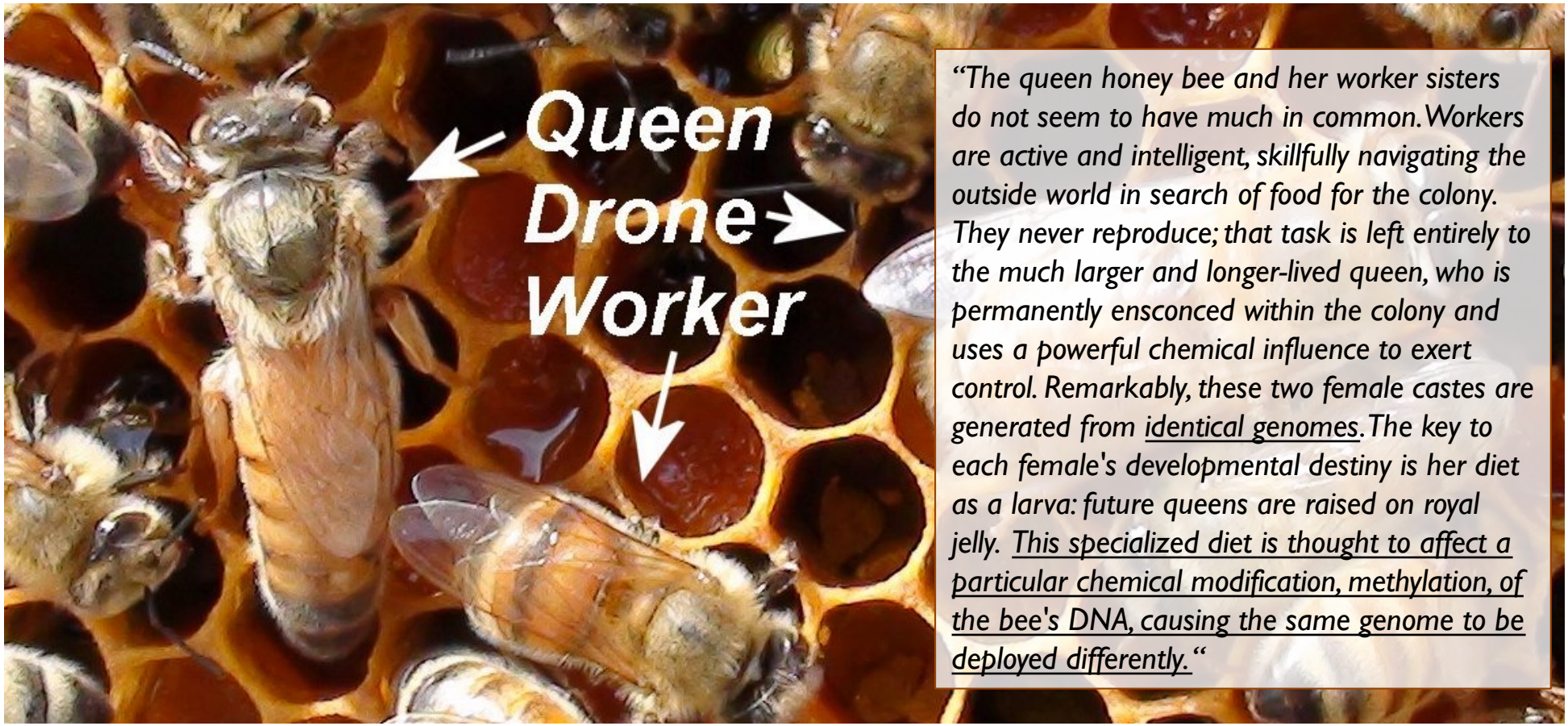
OPEN ACCESS Freely available online

PLoS BIOLOGY

The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers

Frank Lyko[1☯], Sylvain Foret[2☯], Robert Kucharski[3], Stephan Wolf[4], Cassandra Falckenhayn[1], Ryszard Maleszka[3*]

1 Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany, **2** ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia, **3** Research School of Biology, the Australian National University, Canberra, Australia, **4** Genomics and Proteomics Core Facility, German Cancer Research Center, Heidelberg, Germany



"The queen honey bee and her worker sisters do not seem to have much in common. Workers are active and intelligent, skillfully navigating the outside world in search of food for the colony. They never reproduce; that task is left entirely to the much larger and longer-lived queen, who is permanently ensconced within the colony and uses a powerful chemical influence to exert control. Remarkably, these two female castes are generated from identical genomes. The key to each female's developmental destiny is her diet as a larva: future queens are raised on royal jelly. This specialized diet is thought to affect a particular chemical modification, methylation, of the bee's DNA, causing the same genome to be deployed differently."

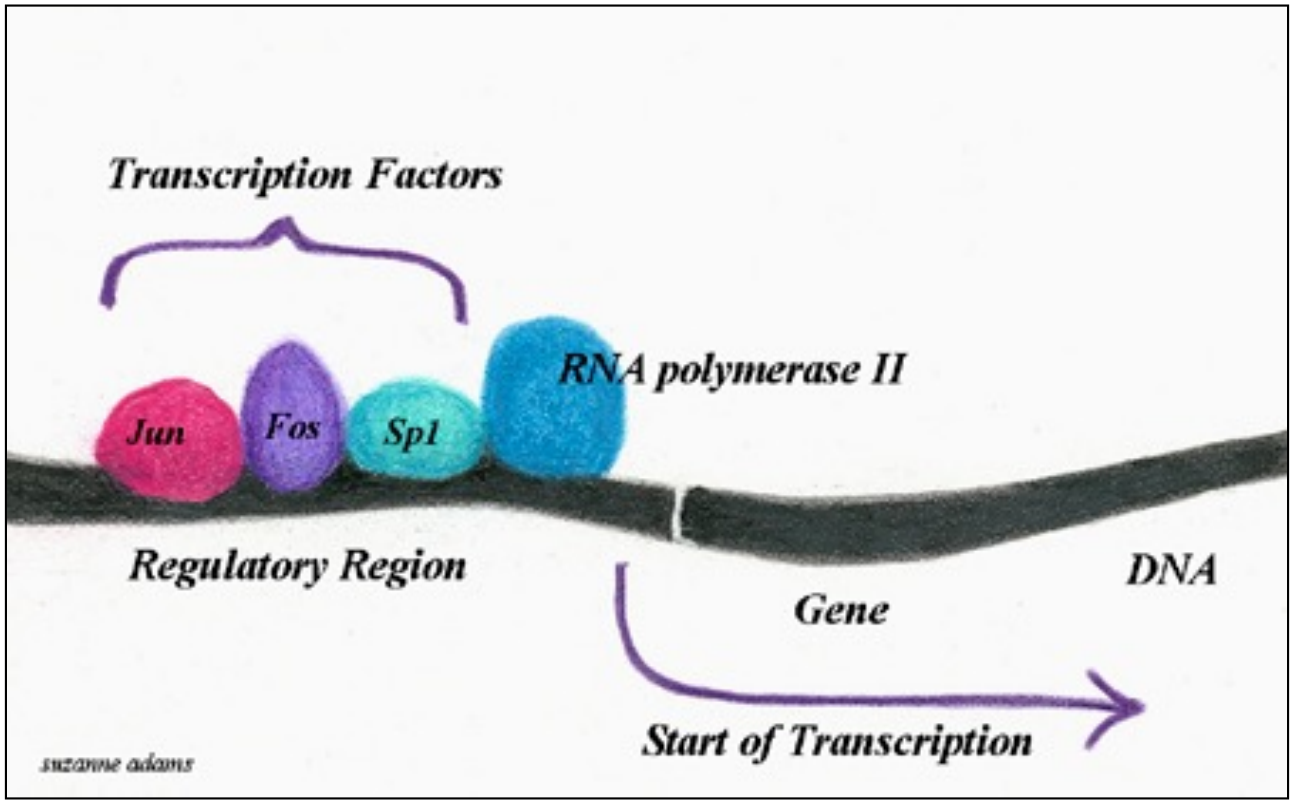
Bisulfite Conversion

Tr
w

-
-
-

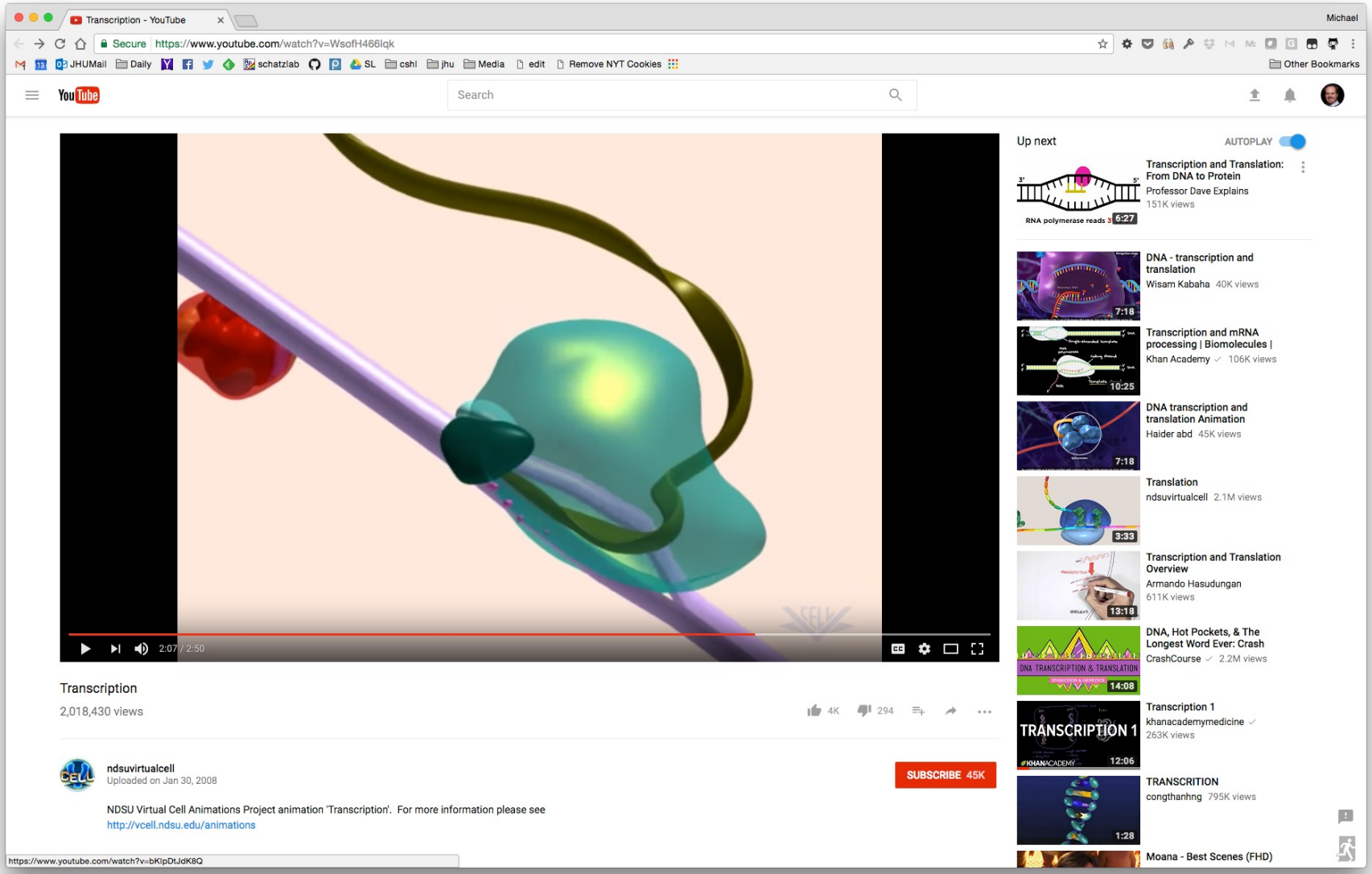
Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

ChIP-seq



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

Transcription

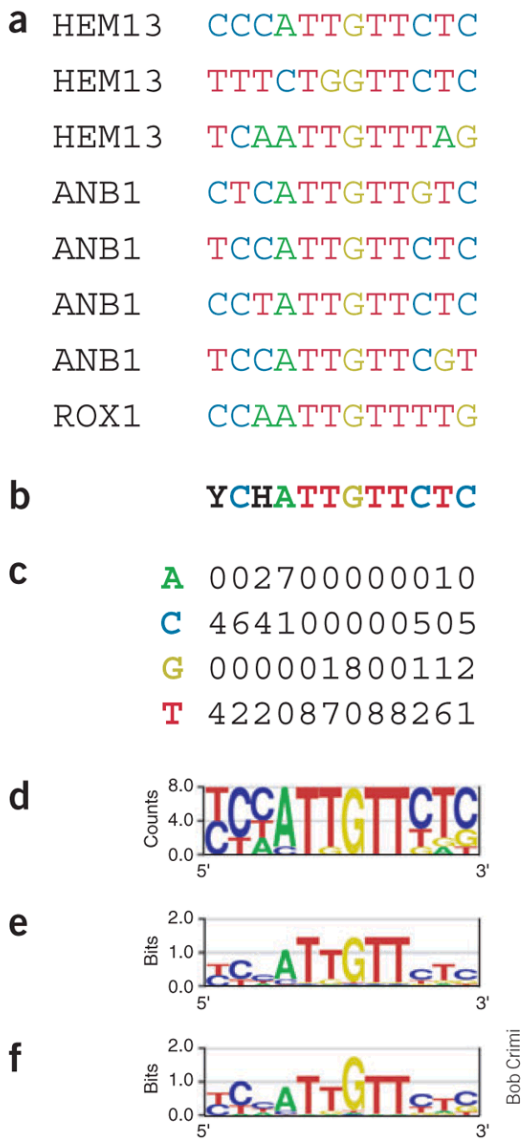


https://www.youtube.com/watch?v=WsofH466lqk

Transcription Factors

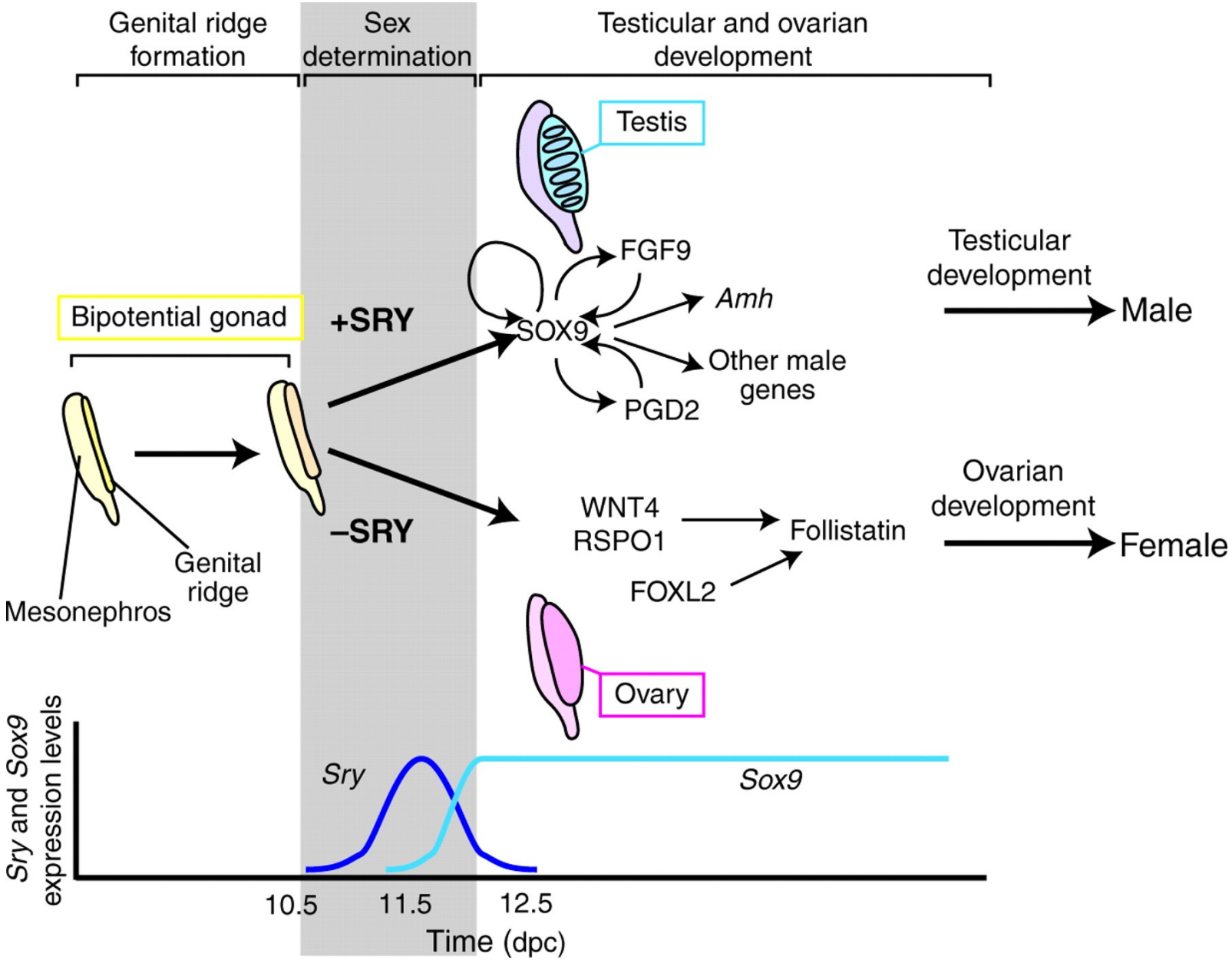
A transcription factor (or sequence-specific DNA-binding factor) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.

- Transcription factors work alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase to specific genes.
- A defining feature of transcription factors is that they contain at least one DNA-binding domain (DBD)
- Figure (a) Eight known genomic binding sites in three S. cerevisiae genes. (b) Degenerate consensus sequence. (c,d) Frequencies of nucleotides at each position. (e) Sequence logo (f) Energy normalized logo using relative entropy to adjust for low GC content in S. cerevisiae.



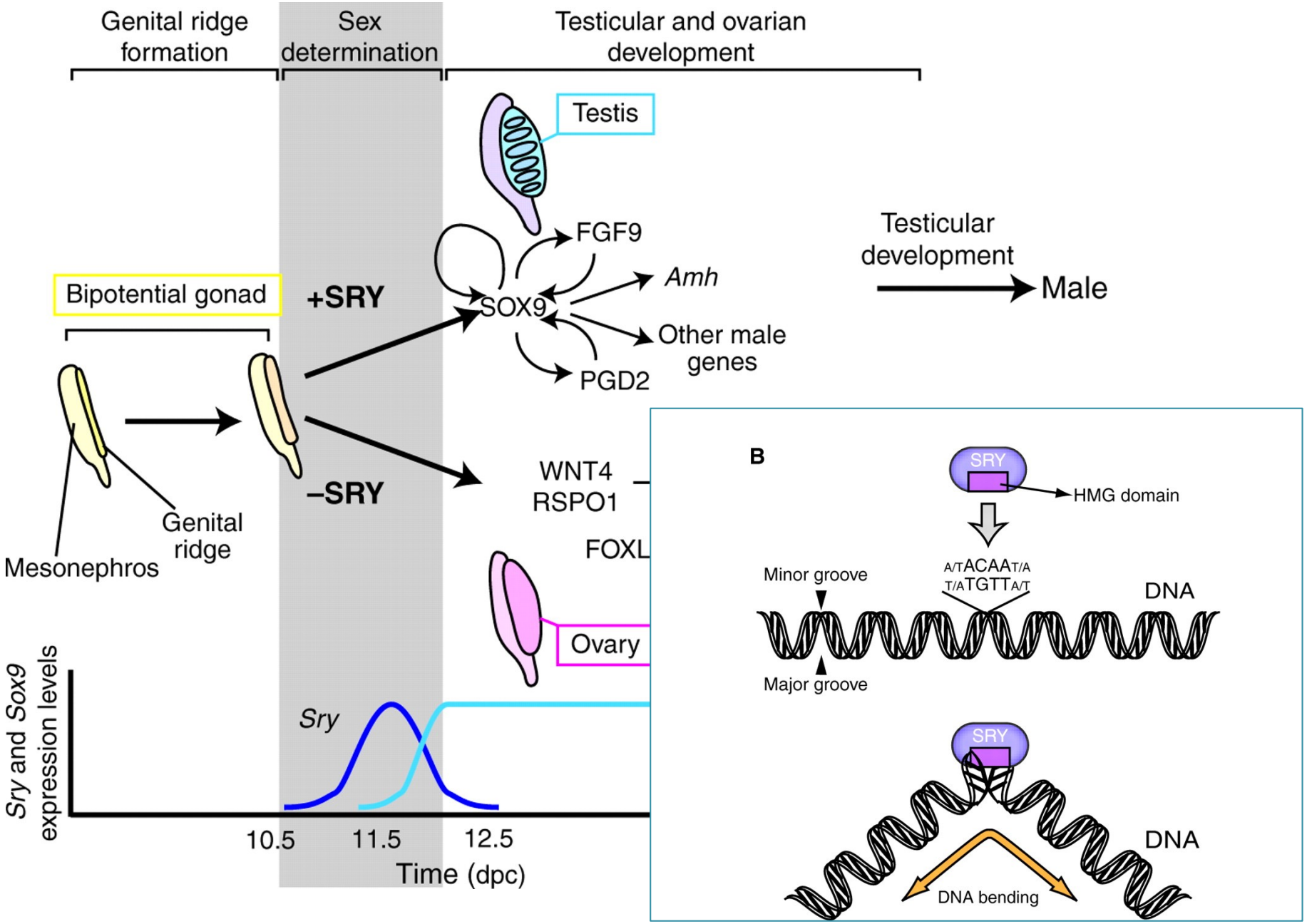
What are DNA sequence motifs?

D'haeseleer (2006) Nature Biotechnology 24, 423 – 425 doi:10.1038/nbt0406-423



SRY: The master switch in mammalian sex determination

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



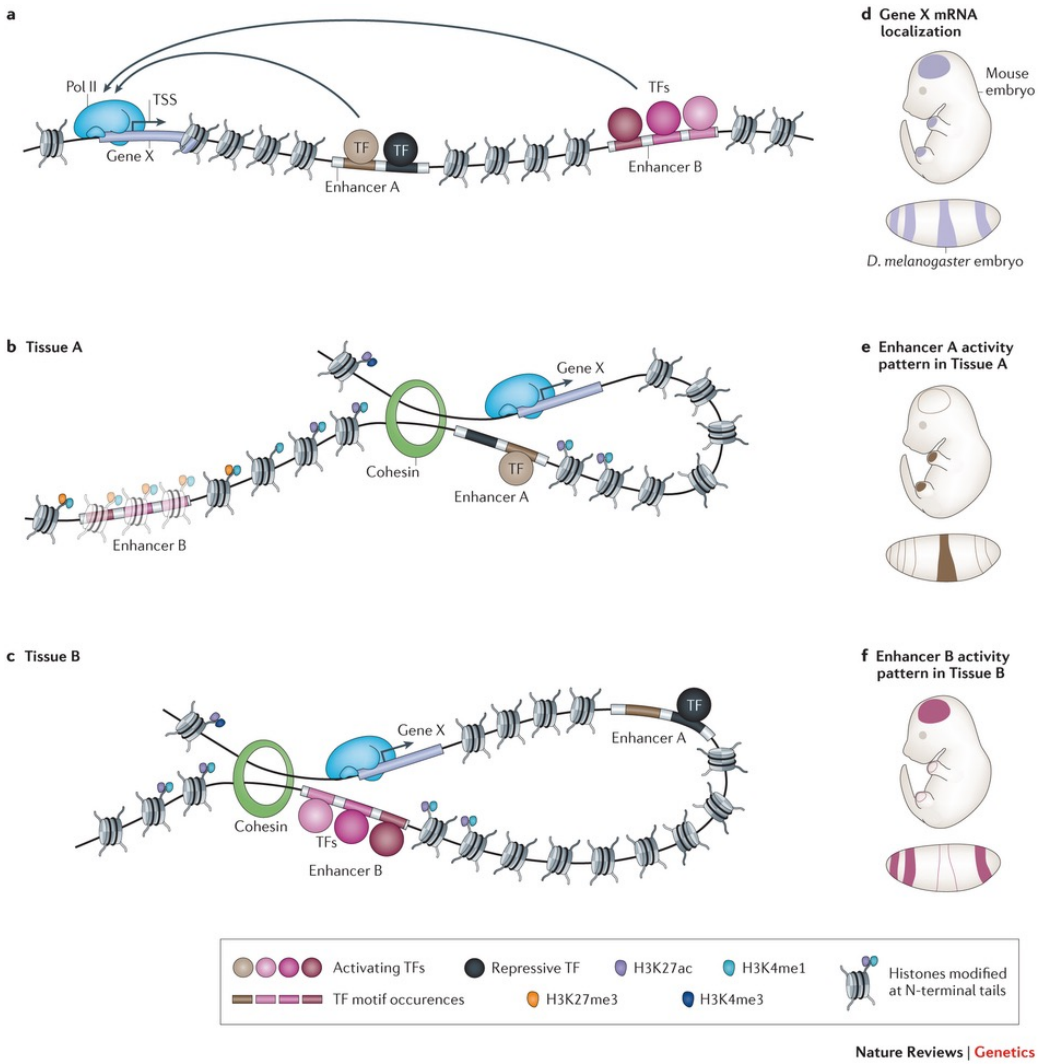
SRY: The master switch in mammalian sex determination

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

Enhancers

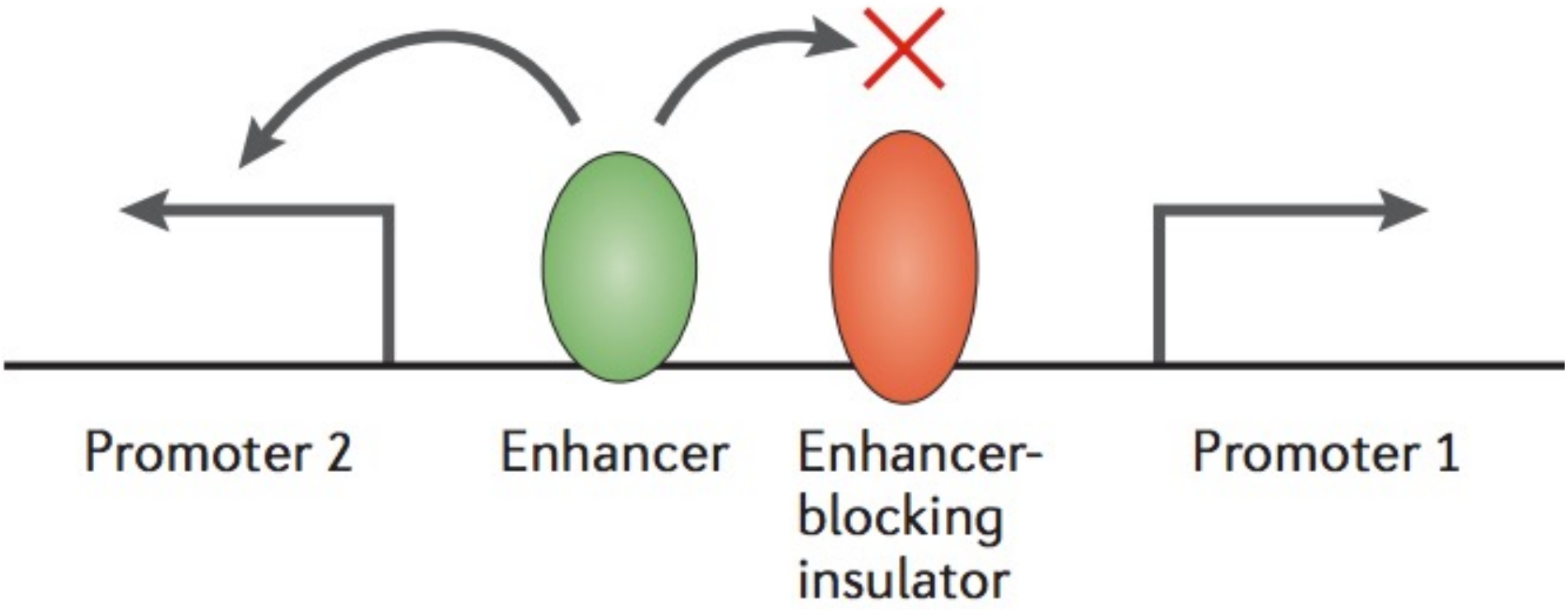
Enhancers are genomic regions that contain binding sites for transcription factors (TFs) and that can upregulate (enhance) the transcription of a target gene.

- 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

Insulators



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

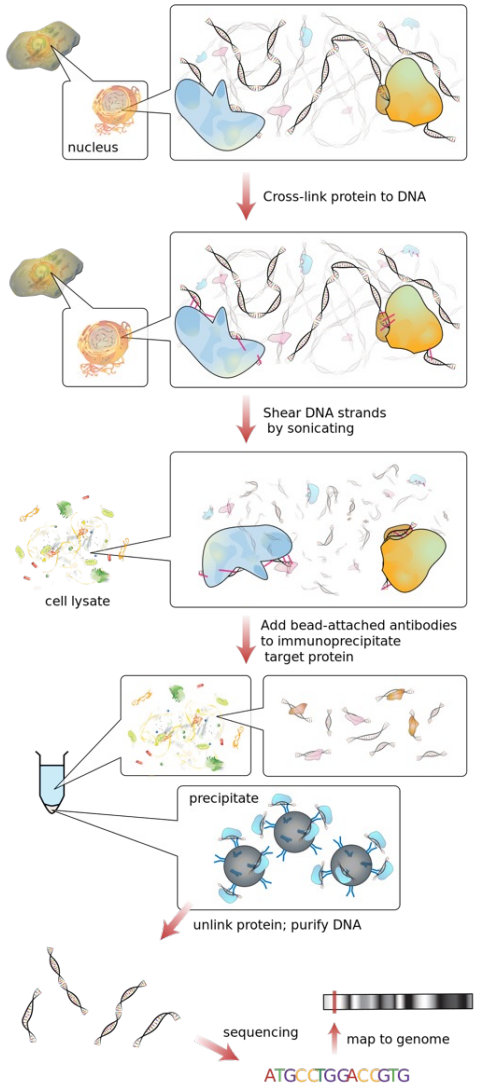
- One type of insulator establishes domains that separate enhancers and promoters to block their interaction,
- Second type creates a barrier against the spread of heterochromatin.

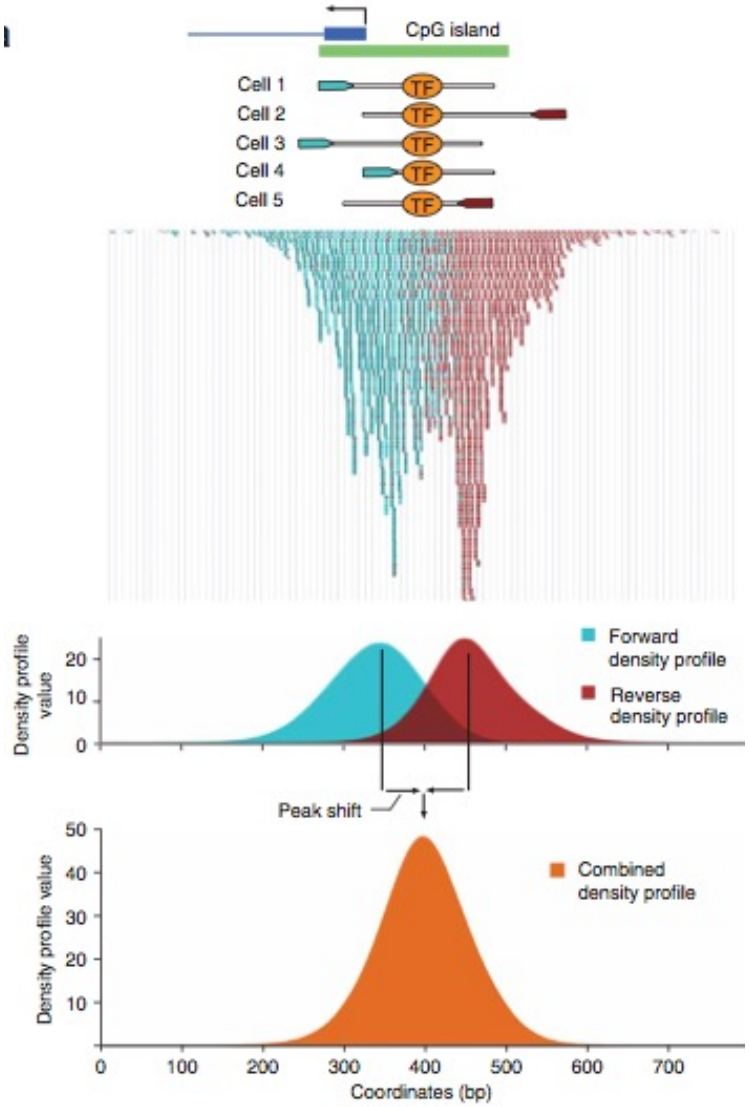
Insulators: exploiting transcriptional and epigenetic mechanisms
Gaszner & Felsenfeld (2006) *Nature Reviews Genetics* 7, 703-713. doi:10.1038/nrg1925

ChIP-seq: TF Binding

Goals:

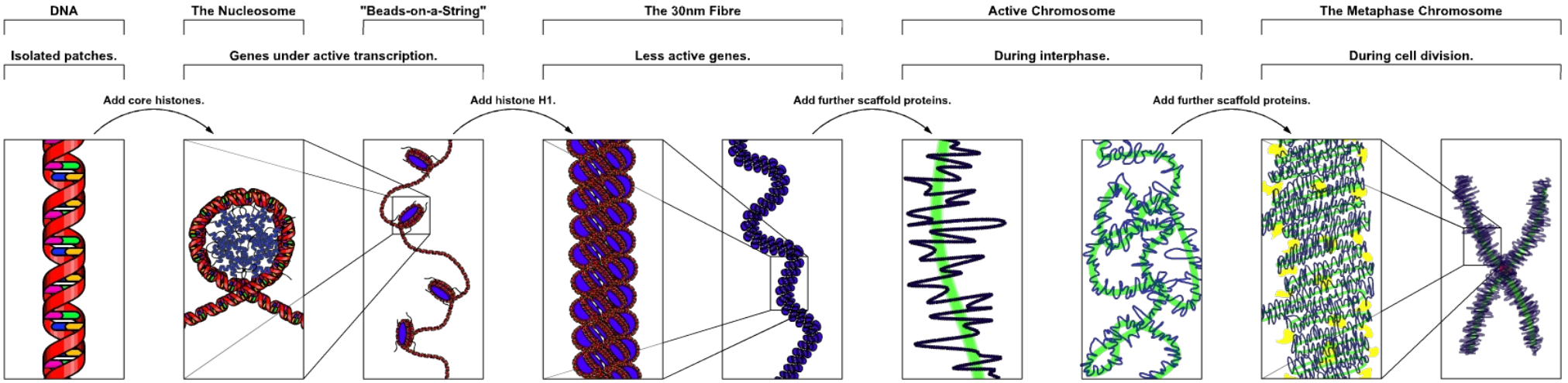
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?





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



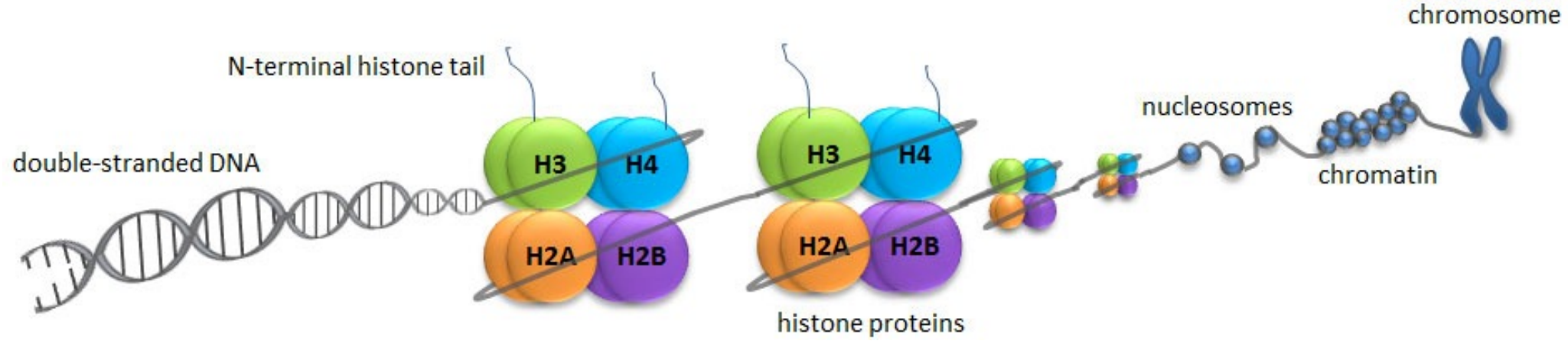
Nucleosome is a basic unit of DNA packaging in eukaryotes

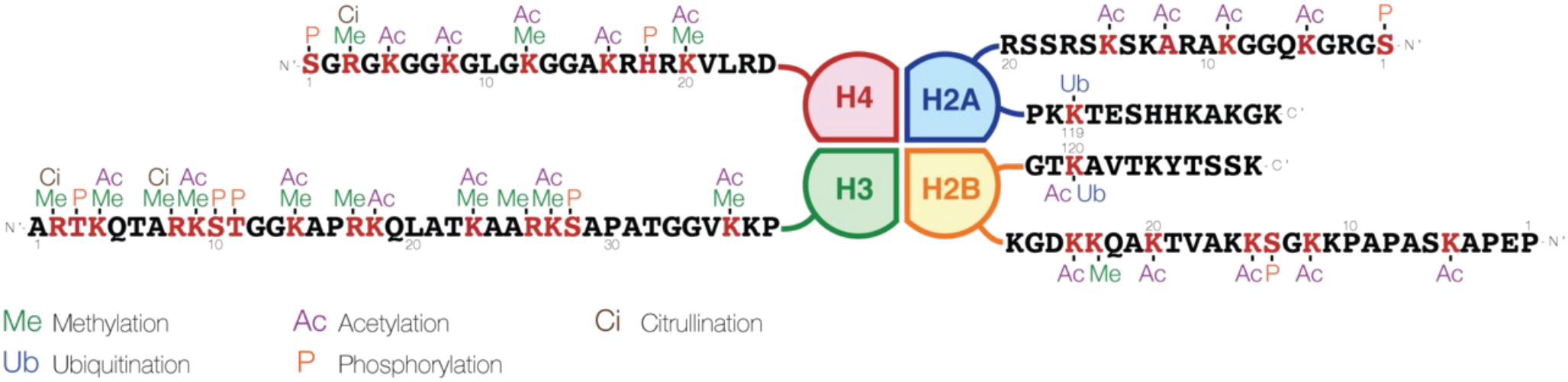
- Consists of a segment of 146bp DNA wound in sequence around eight histone protein cores (thread wrapped around a spool) followed by a ~38bp linker
- Under active transcription, nucleosomes appear as “beads-on-a-string”, but are more densely packed for less active genes

Nucleosomes form the fundamental repeating units of eukaryotic chromatin

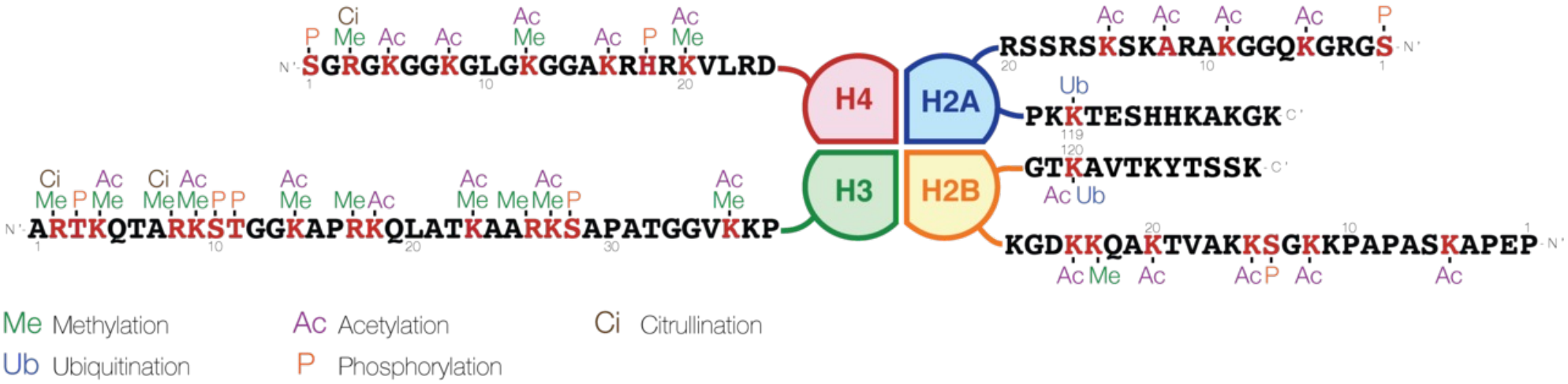
- Used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 μm diameter).

ChIP-seq: Histone Modifications





ChIP-seq: Histone Modifications

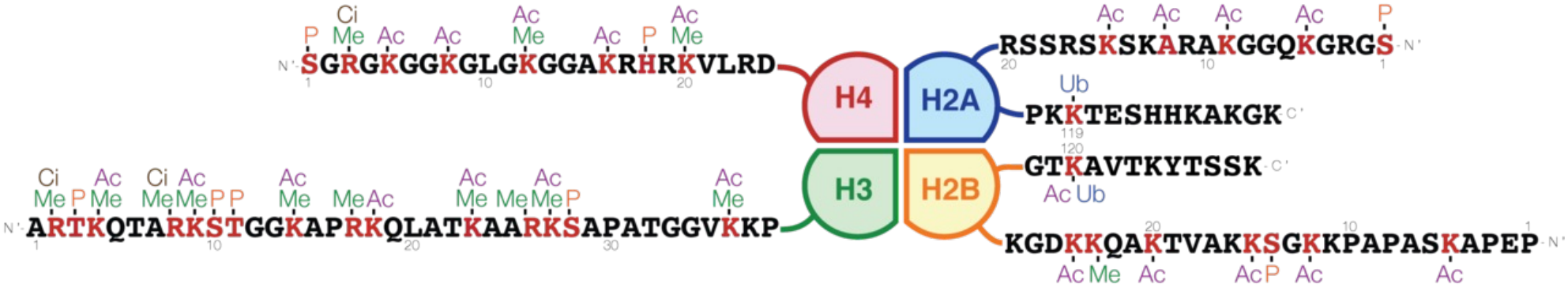


The common nomenclature of histone modifications is:

- 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.

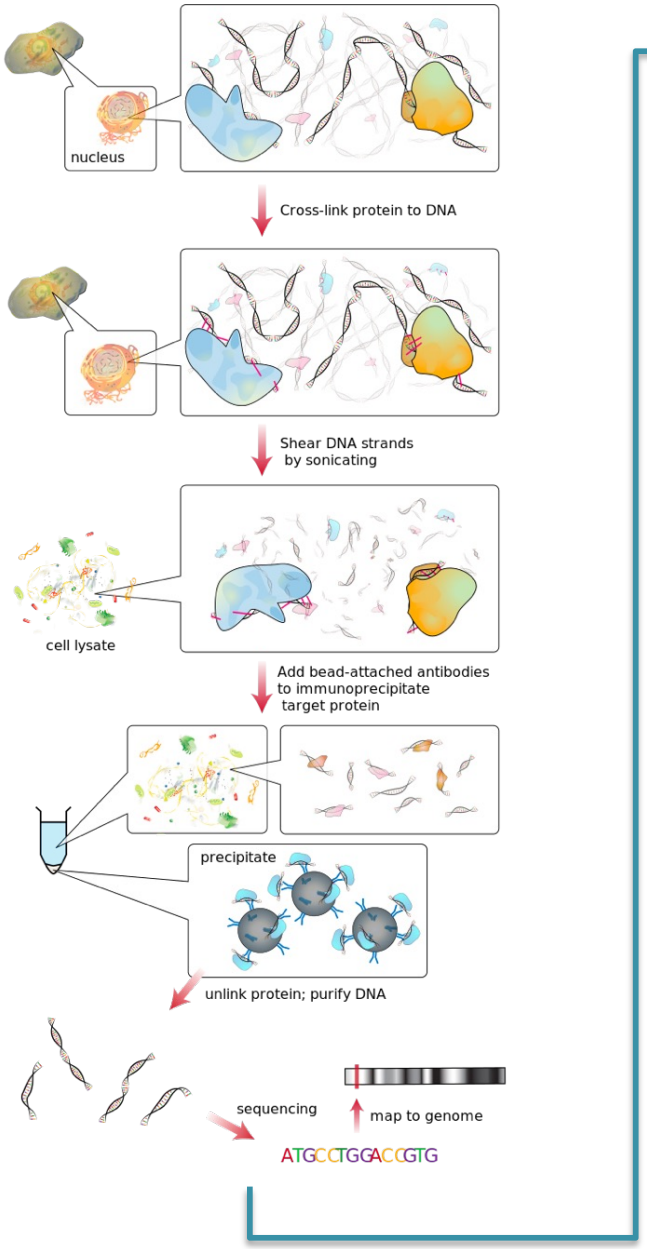
ChIP-seq: Histone Modifications



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

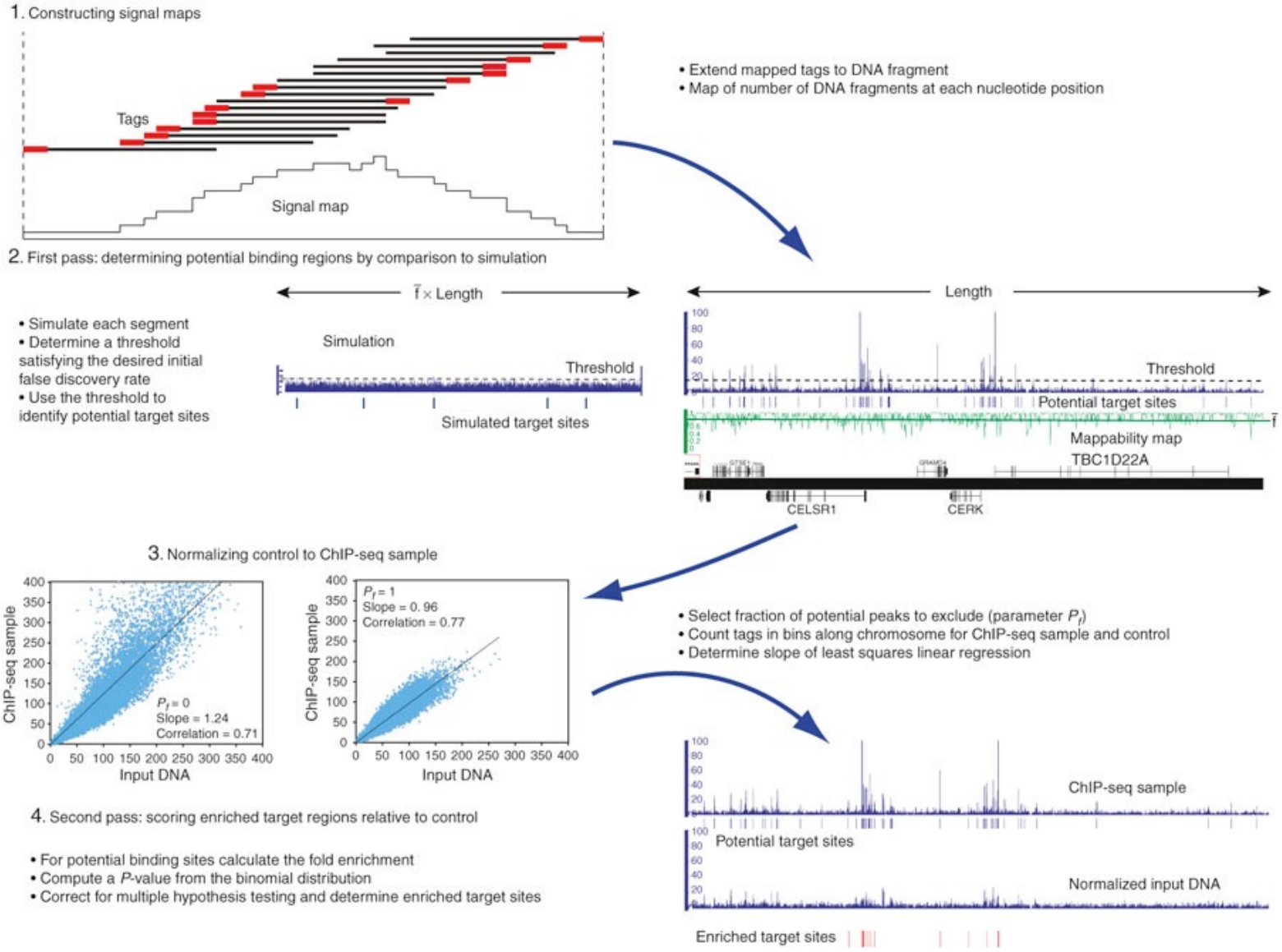
Analyze Enriched Regions

Visualization

Peak-peak comparisons

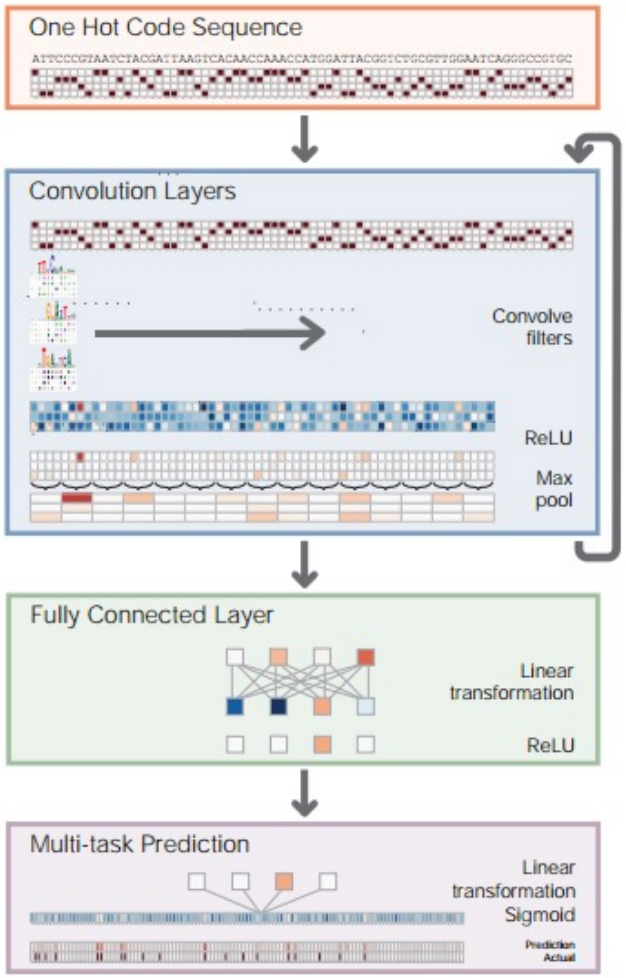
Annotation / Interpretation

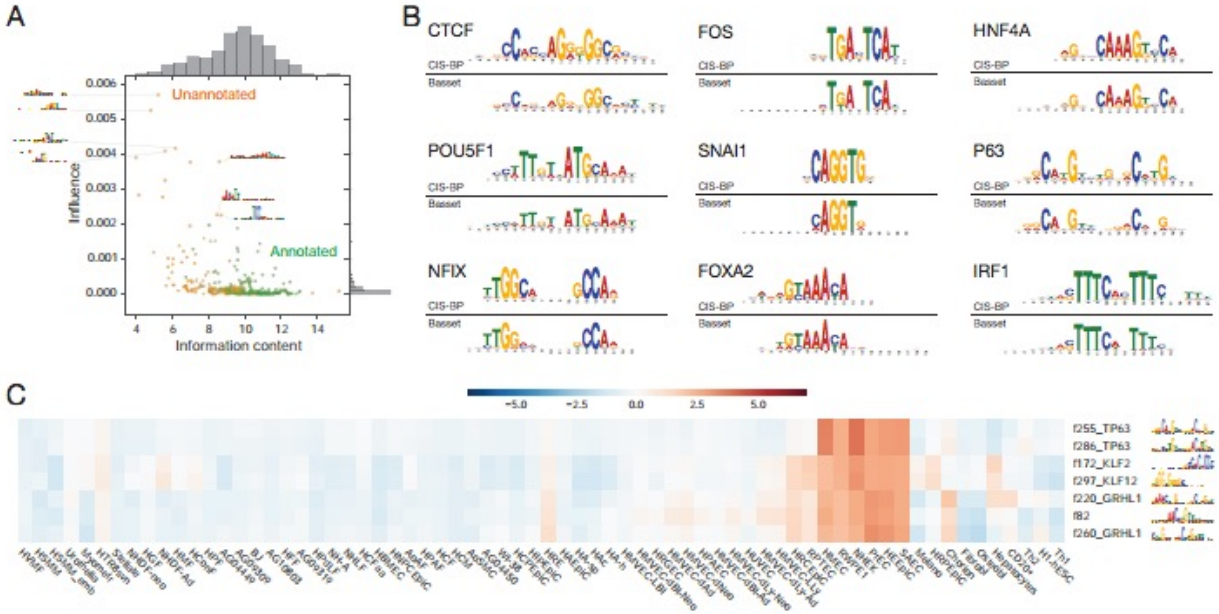
PeakSeq



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

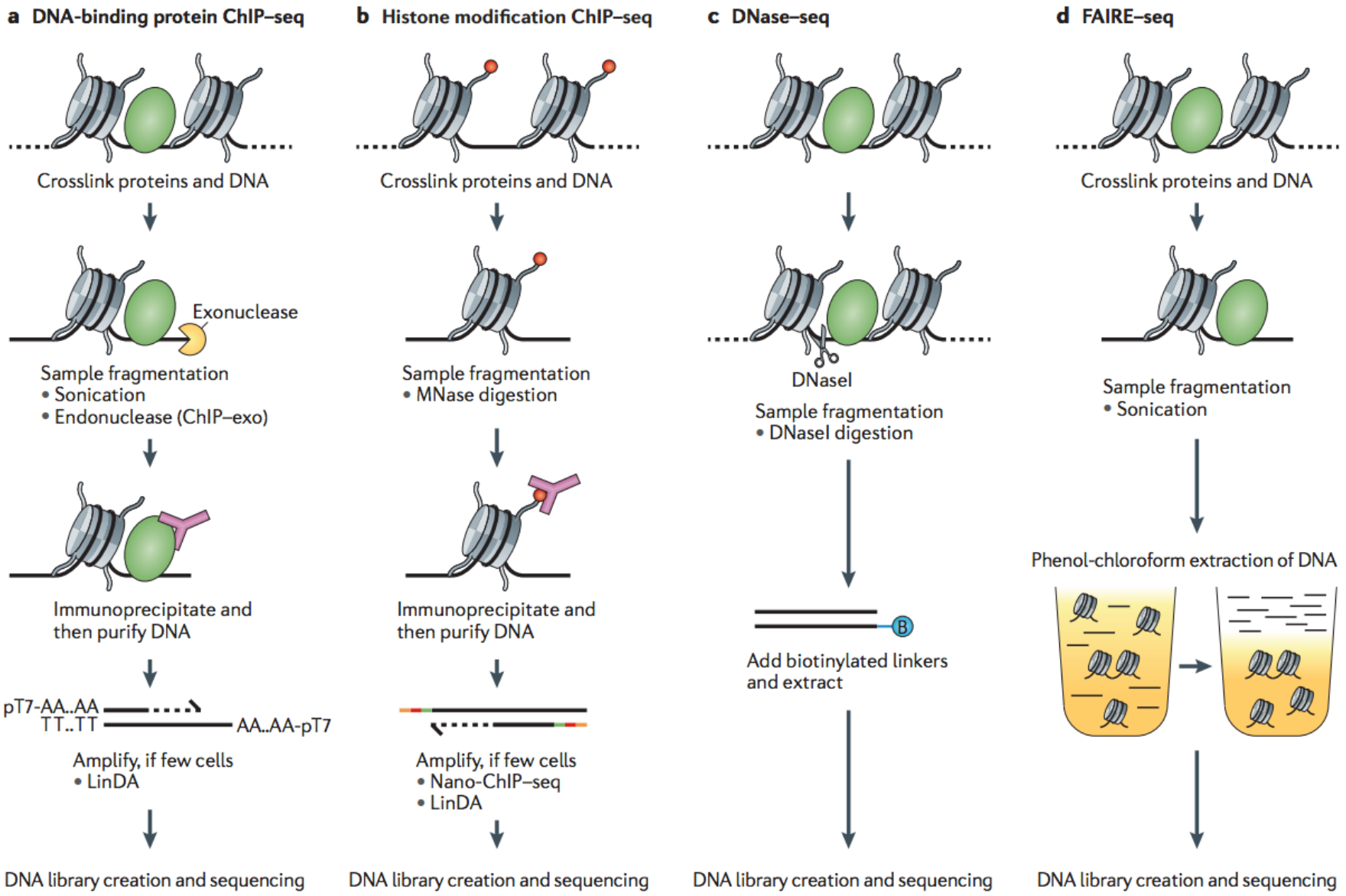
Basset





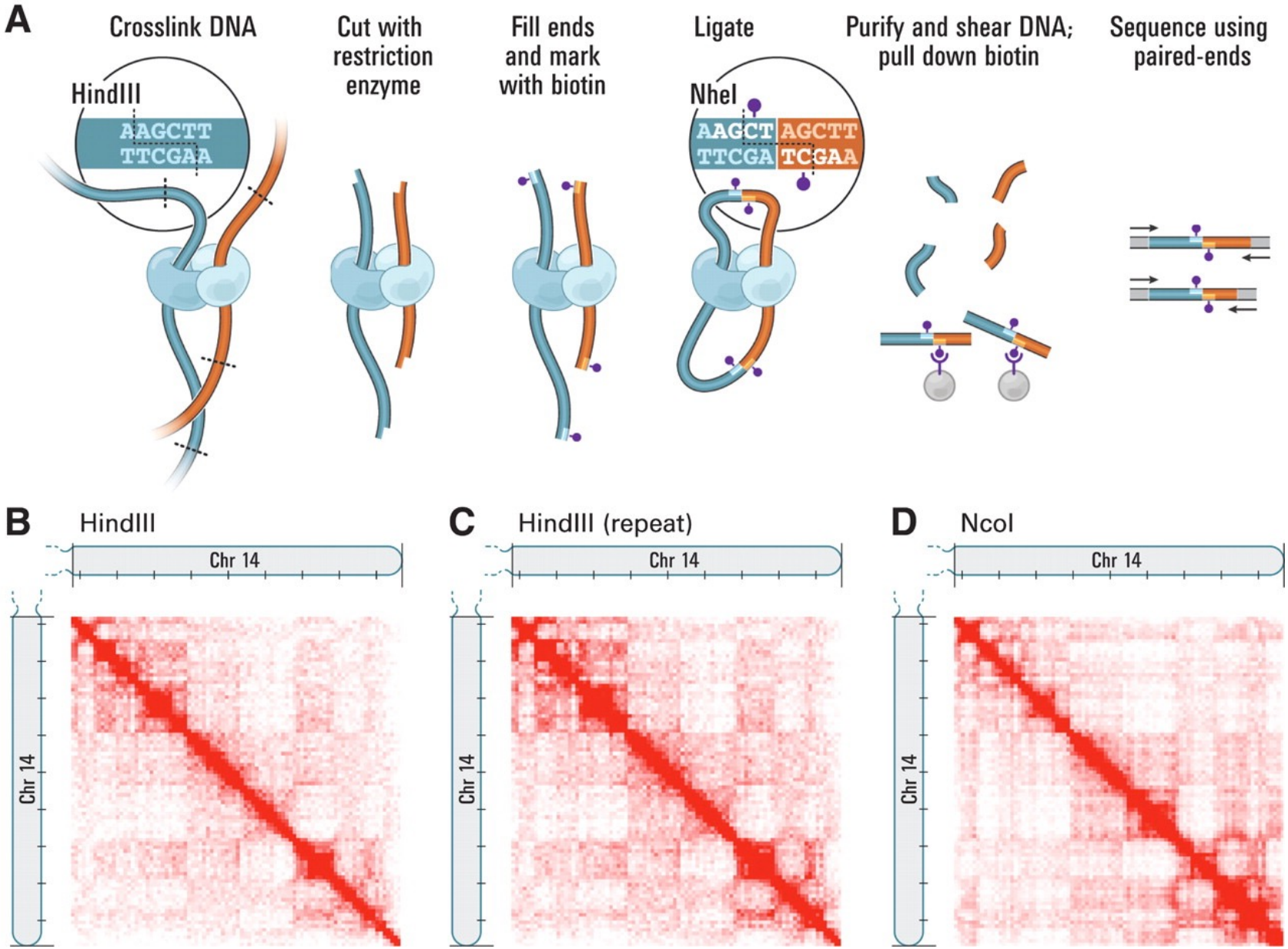
Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks
Kelley et al. (2016) Genome Research doi: 10.1101/gr.200535.115

Related Assays



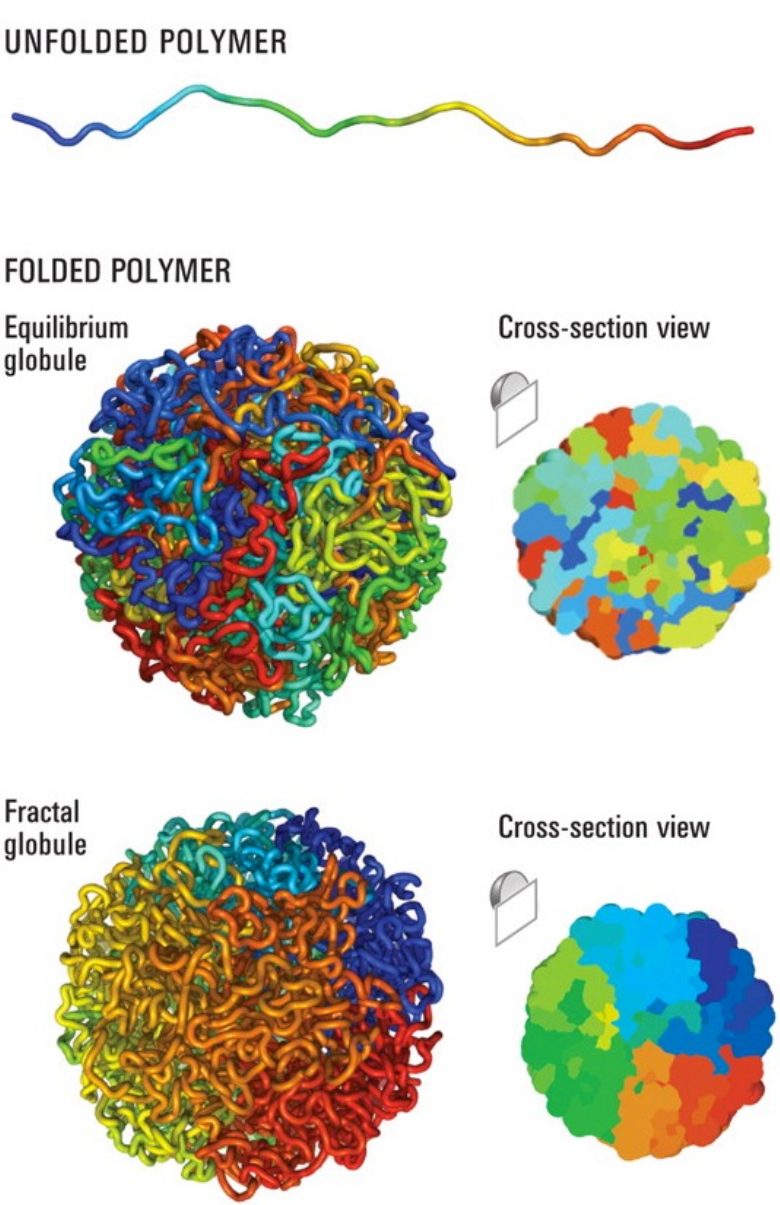
ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions
Furey (2012) *Nature Reviews Genetics*. 13, 840-852

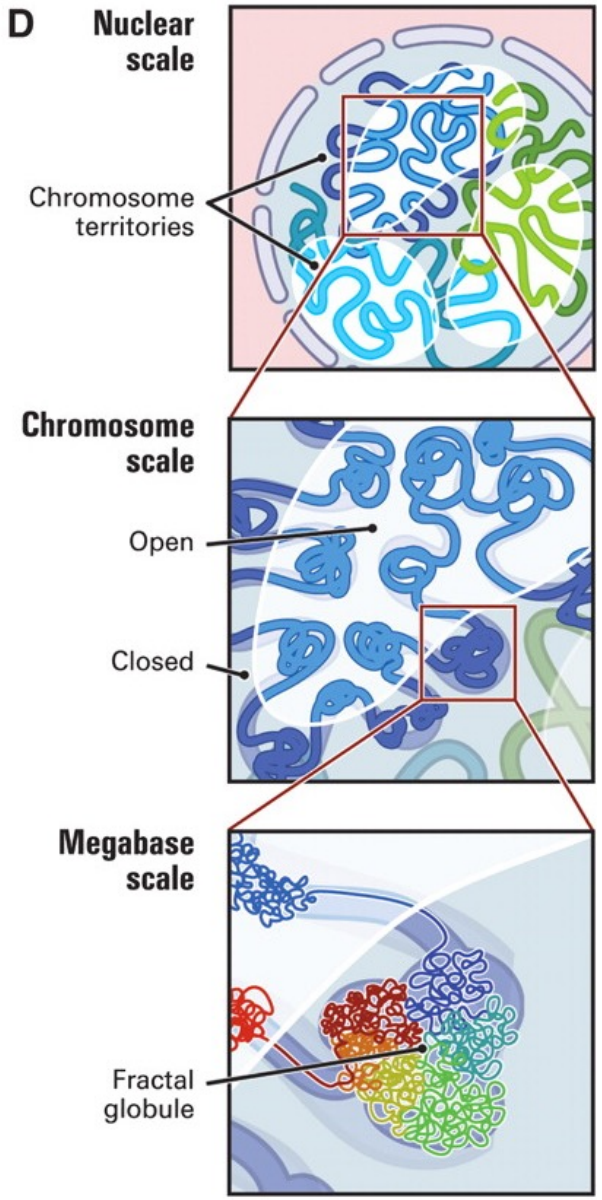
HI-C: Mapping the folding of DNA



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

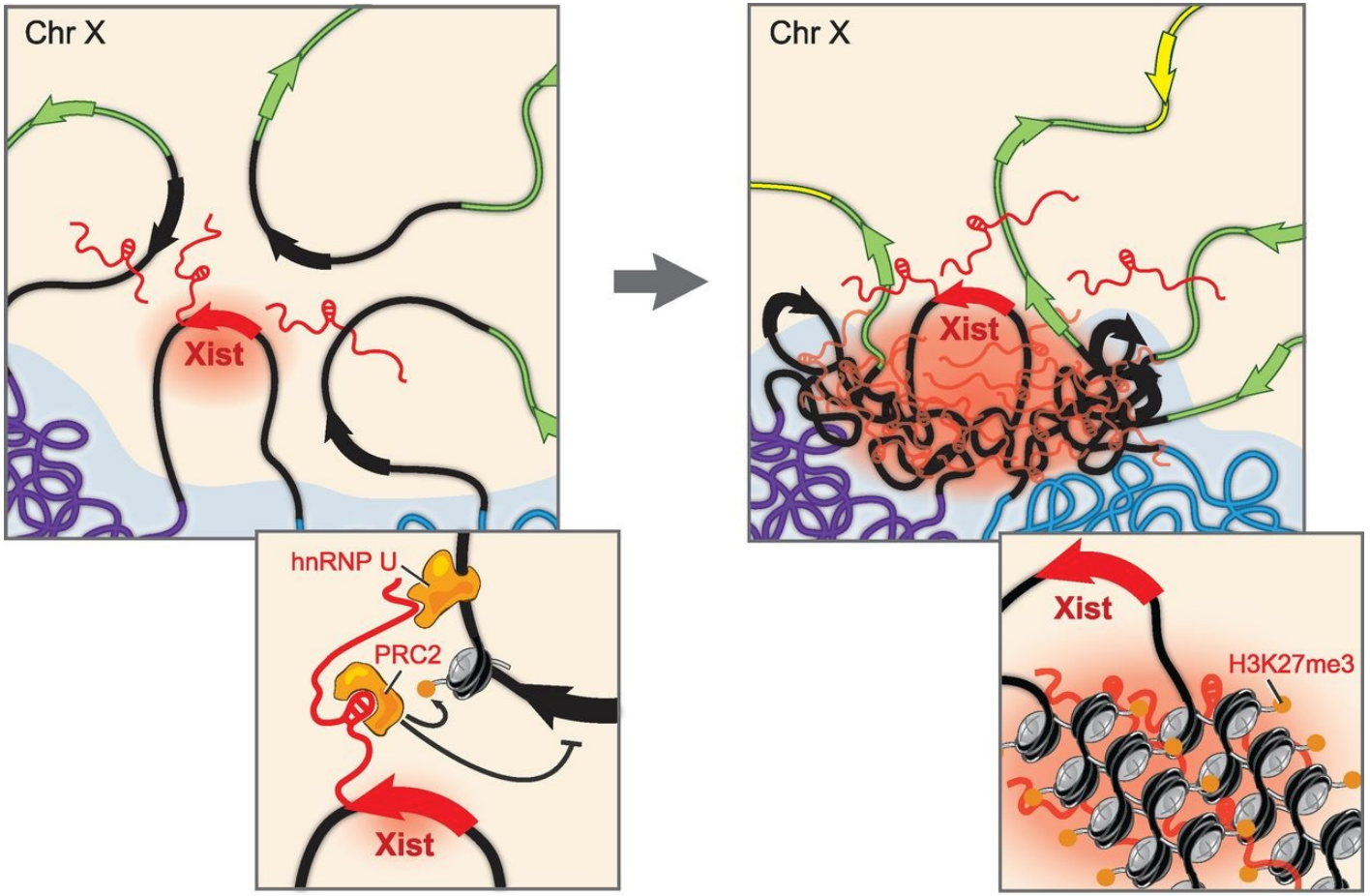
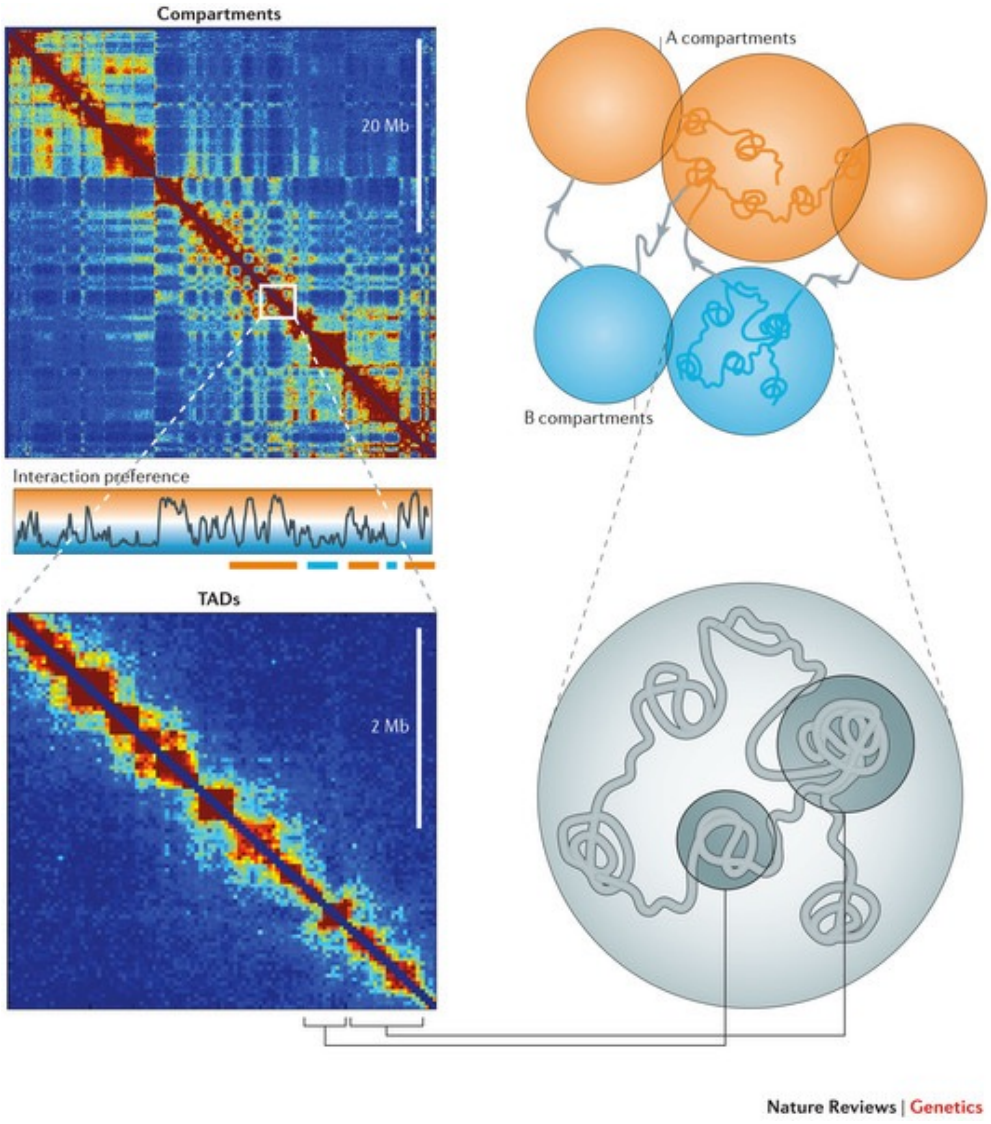


Fig 6. A model for how Xist exploits and alters three-dimensional genome architecture to spread across the X chromosome.

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

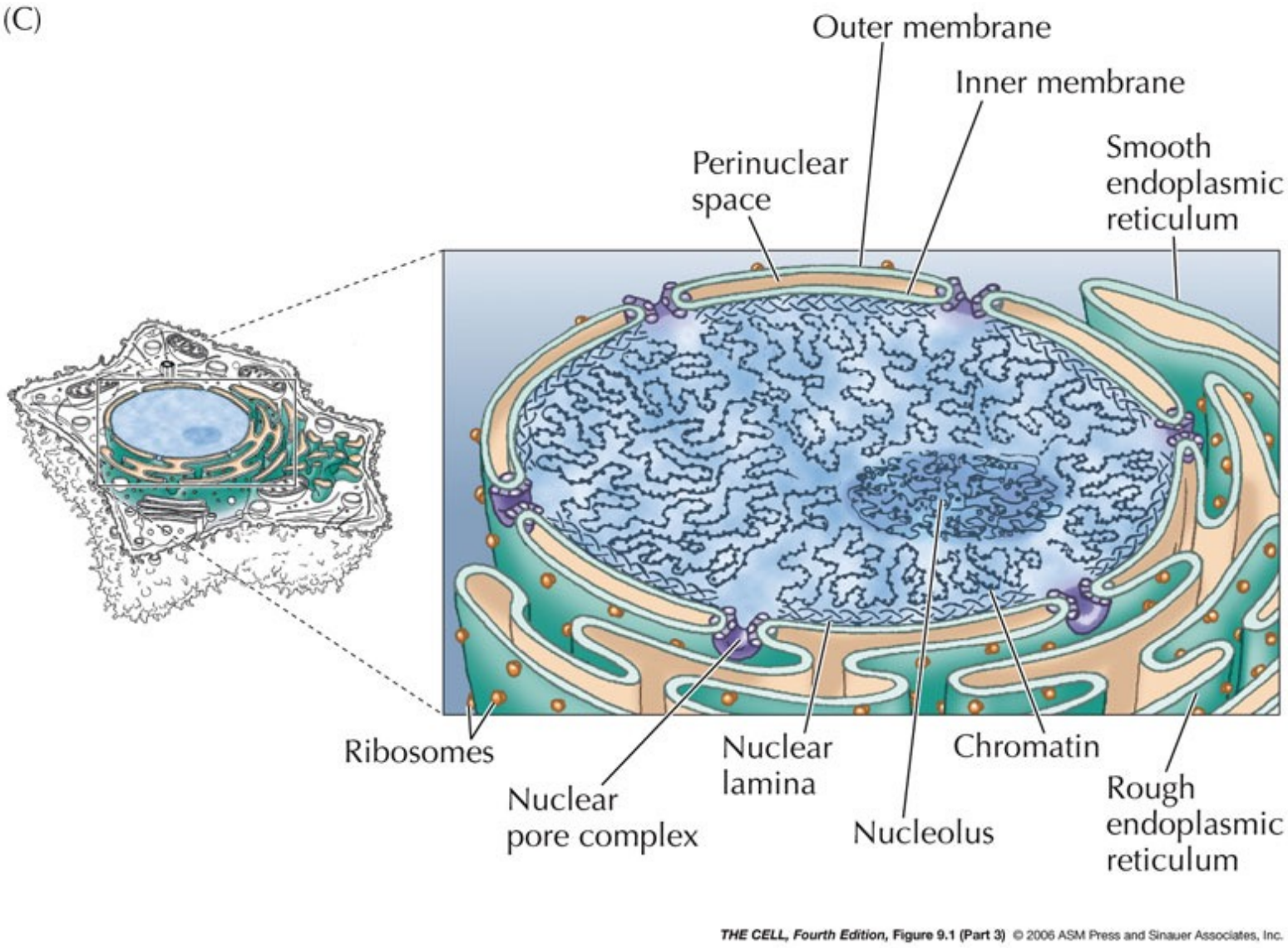
- alternate along chromosomes and have a characteristic size of ~5 Mb each.
- A compartments (orange) preferentially interact with other A compartments; B compartments (blue) associate with other B compartments.
- A compartments are largely euchromatic, transcriptionally active regions.

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*

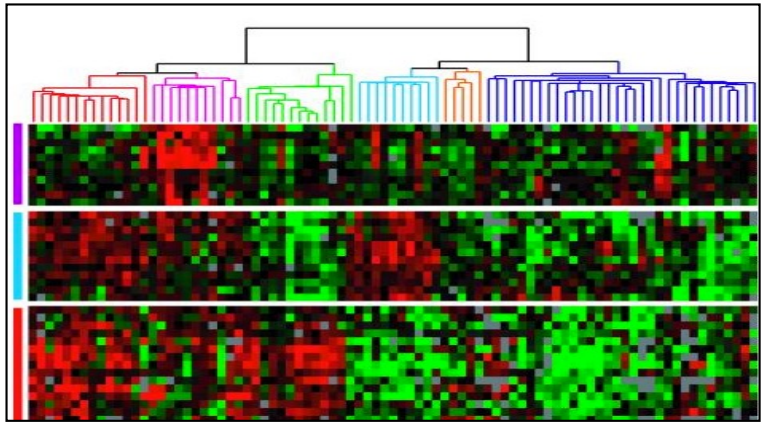
“Lamina-Associated Domains are the B compartment”



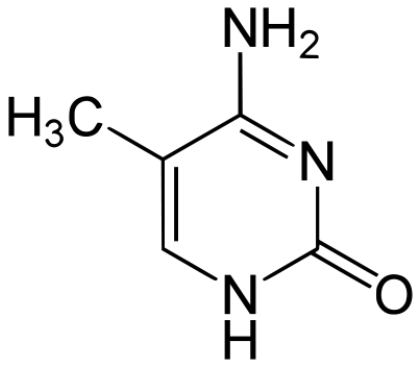
Chromosome Conformation Paints Reveal the Role of Lamina Association in Genome Organization and Regulation
Luperchio et al. (2017) bioRxiv. *doi: https://doi.org/10.1101/122226*

Putting it all together!

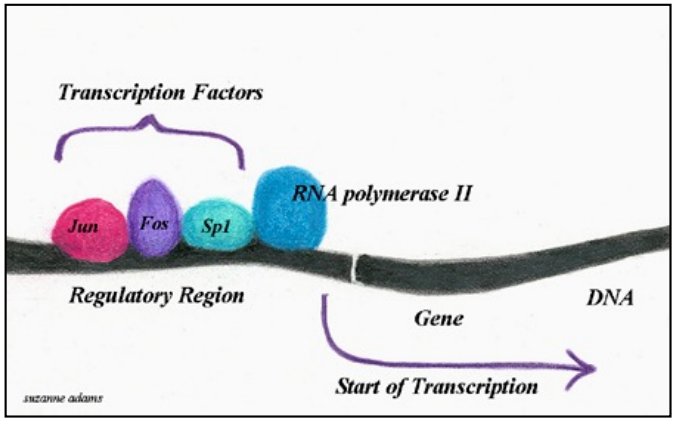
RNA-seq



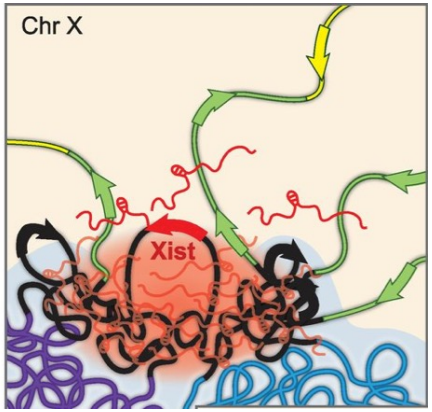
Methyl-seq



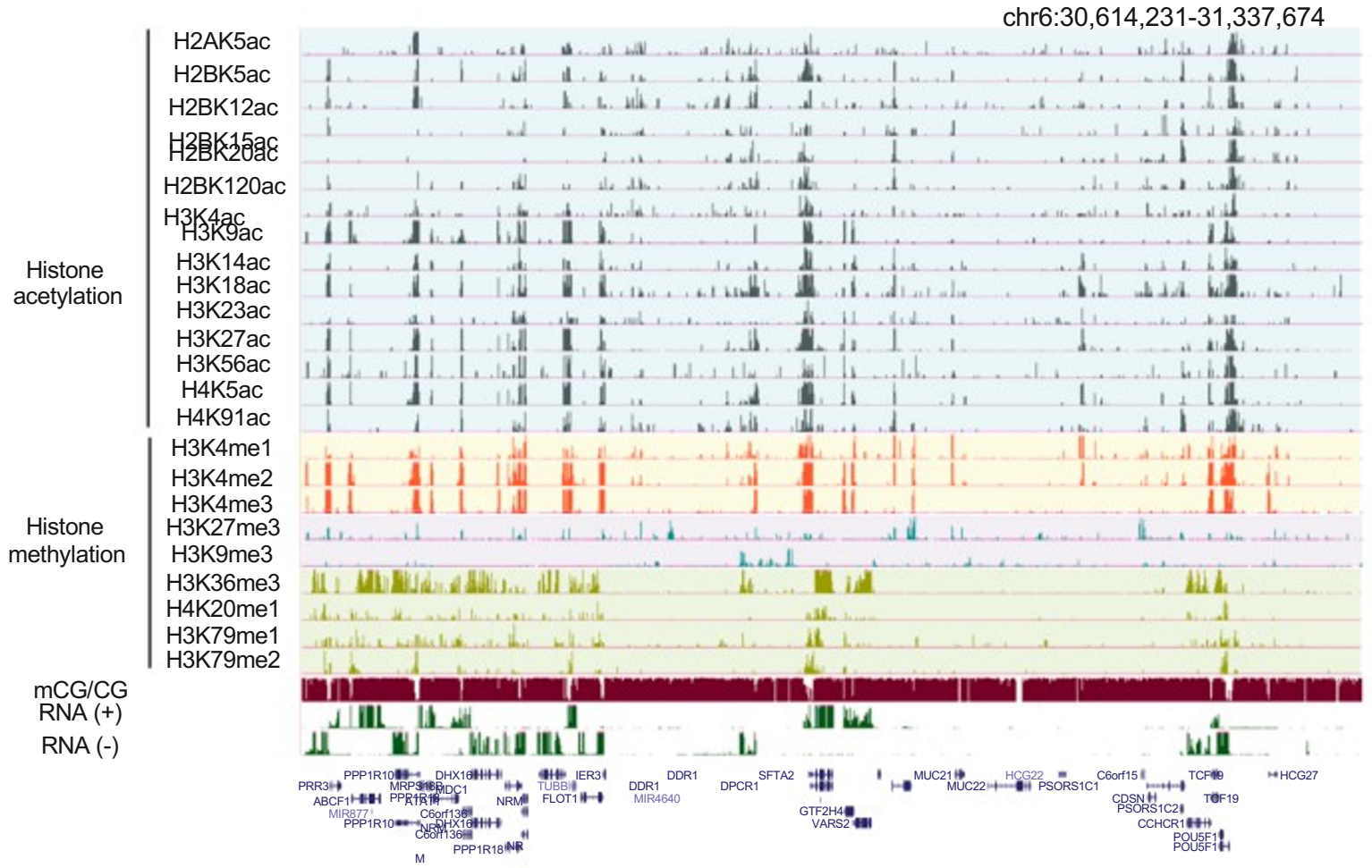
ChIP-seq



Hi-C

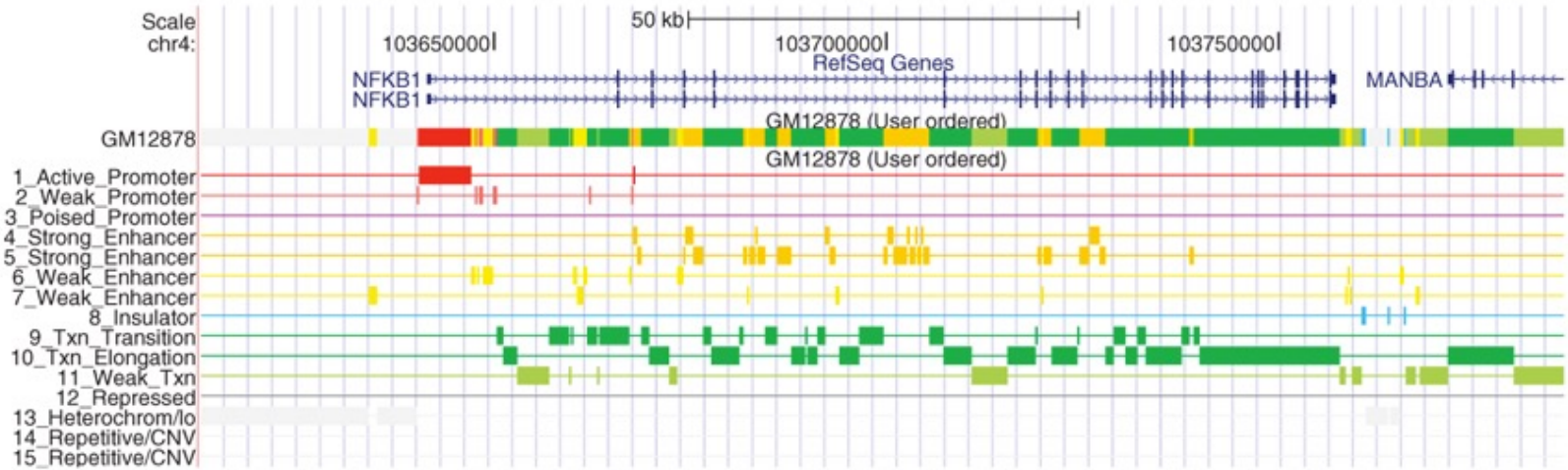


We can call peaks, but…



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

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.