Lecture 17. Gene Regulation

Michael Schatz

March 24, 2021

Applied Comparative Genomics

Assignment 5: Due Wed Mar 17

140 lines (89 sloc) 8.82 KB Raw Blame

Assignment 5: BWT and RNA-seq

Assignment Date: Wednesday, Mar. 3, 2021
Due Date: Wednesday, Mar. 17 , 2021 @ 11:59pm

Assignment Overview

In this assignment you will write a simple BWT encoder and decoder, and explore a couple of aspects of RNA-seq (with a small introduction to clustering). For this assignment, you will have to generate some visualizations - we recommend R or Python, but use a language you are comfortable with!

Make sure to show your work/code in your writeup!

As a reminder, any questions about the assignment should be posted to Piazza.

Question 1. BWT Encoding [10 pts]

In the language of your choice, implement a BWT encoder and encode the string below. Linear time methods exist for computing the BWT, although for this assignment you can use the simple method based on standard sorting techniques. Your solution does *not* need to be an optimal algorithm and can use O(n^2) space and O(n^2 lg n) time.

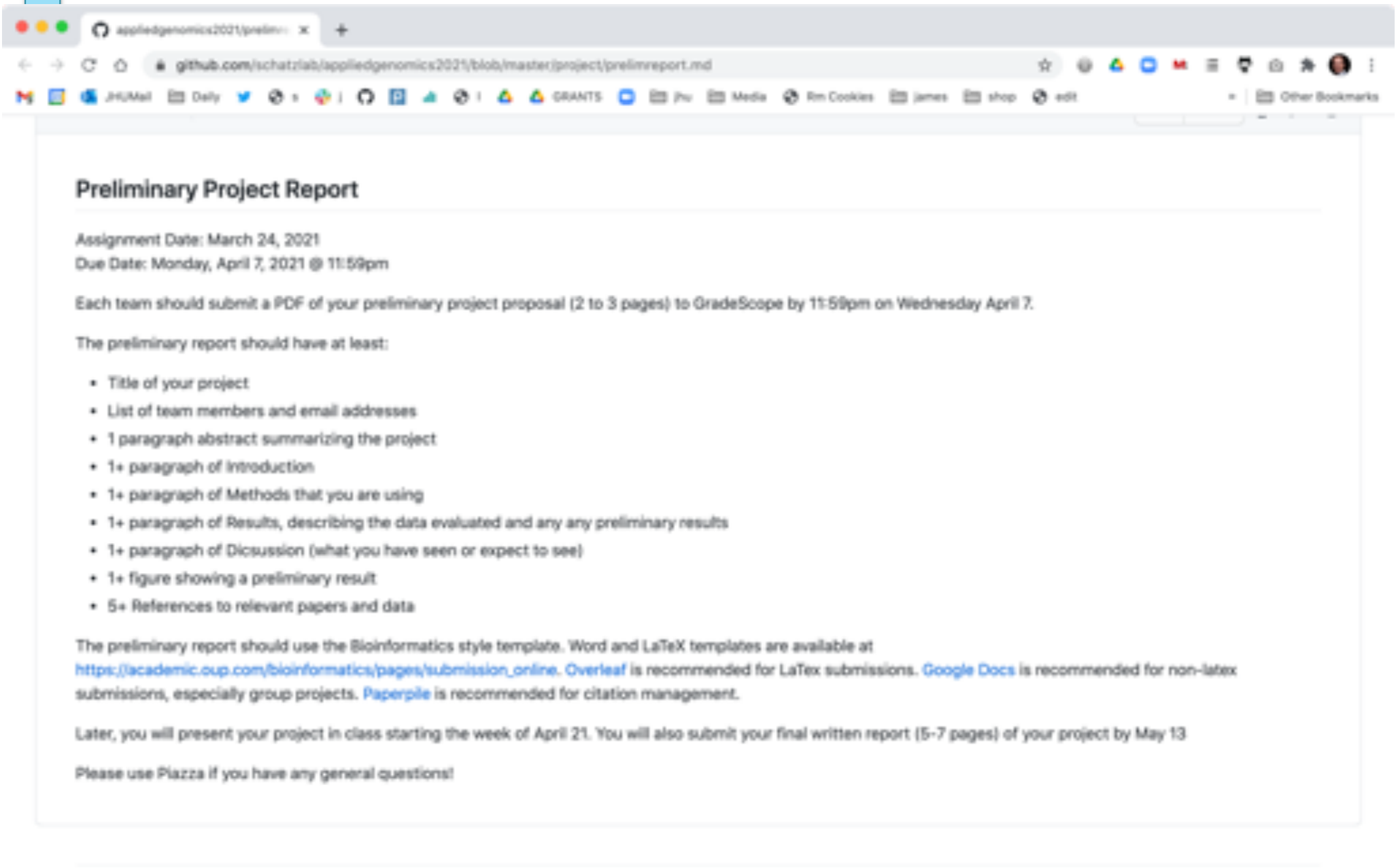
Here is the recommended pseudo code (make sure to submit your code as well as the encoded string):

```
computeBWT(string s)
  ## add the magic end-of-string character
  s = s + "$"

  ## build up the BWM from the cyclic permutations
  ## note the ith cyclic permutation is just "s[i..n] + s[0..i]"
  StringList rows = []
  for (i = 0; i < length(s); i++)
    rows.append(cyclic_permutation(s, i))

  ## just use the builtin sort command
```

Prelim Report: Due Wed April 7



Preliminary Project Report

Assignment Date: March 24, 2021
Due Date: Monday, April 7, 2021 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Wednesday April 7.

The preliminary report should have at least:

- Title of your project
- List of team members and email addresses
- 1 paragraph abstract summarizing the project
- 1+ paragraph of Introduction
- 1+ paragraph of Methods that you are using
- 1+ paragraph of Results, describing the data evaluated and any any preliminary results
- 1+ paragraph of Dicsussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result
- 5+ References to relevant papers and data

The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. Overleaf is recommended for LaTex submissions. Google Docs is recommended for non-latex submissions, especially group projects. Paperpile is recommended for citation management.

Later, you will present your project in class starting the week of April 21. You will also submit your final written report (5-7 pages) of your project by May 13

Please use Piazza if you have any general questions!

Exam Topics

Genomics

- Genomics Technologies
  - Illumina, PacBio, Nanopore
- Genome Assembly
- Whole Genome Alignment
- Read mapping
- Variant Identification
- Gene Finding
- RNA-seq
- Methyl-seq, Chip-Seq, Hi-C
- Genome Annotation
- Single cell vs bulk sequencing

Quantitative Techniques

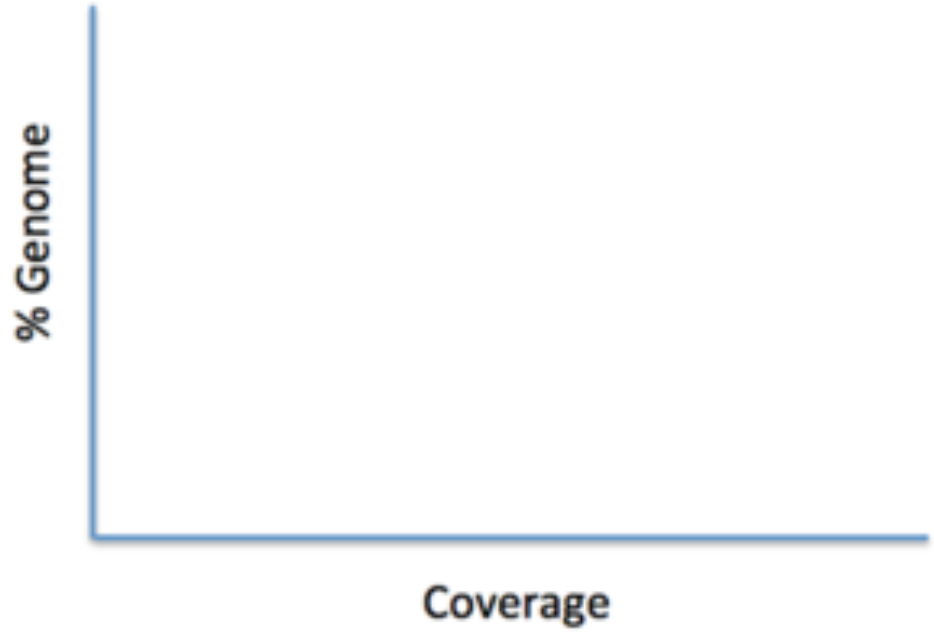
- Normal, Poisson, Binomial, P-value
- de Bruijn and overlap graphs
- Minimizers
- Dot plots
- Quality Values (Phred Scale)
- Full text indexing & BWT
- Seed & Extend
- Hidden Markov Models
- PCA / t-SNE / UMAP
- Differential Expression
- Expectation Maximization

What is the goal? What is the approach? What are the key challenges?

How did we explore these topics in the homeworks and lectures?

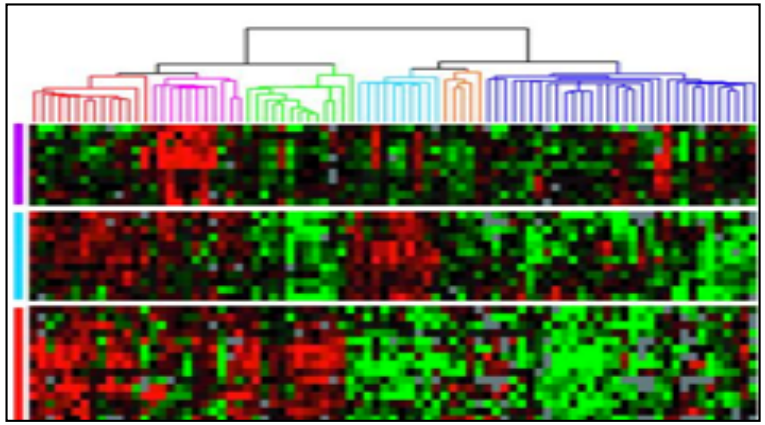
Sample Question

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)

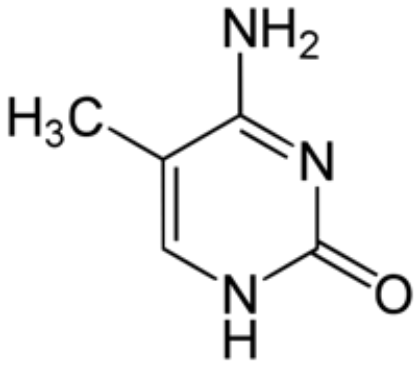


*-seq in 4 short vignettes

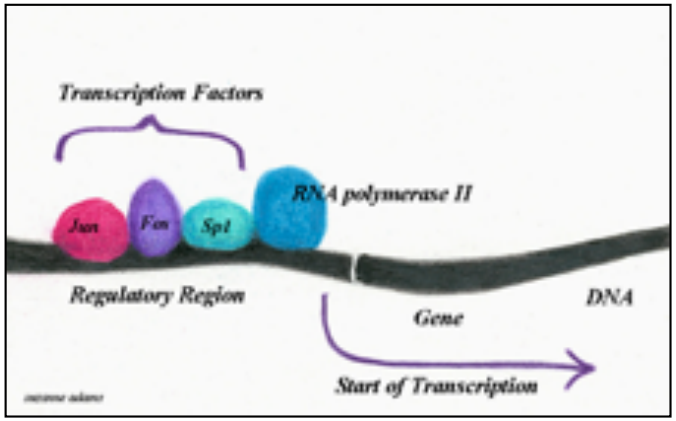
RNA-seq



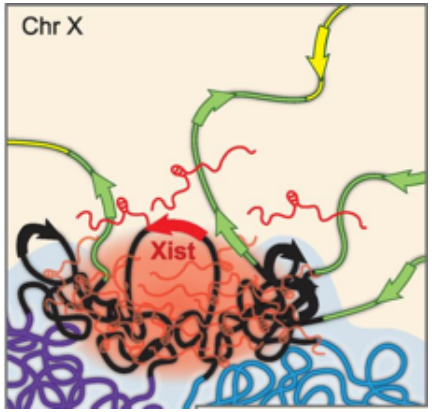
Methyl-seq



ChIP-seq



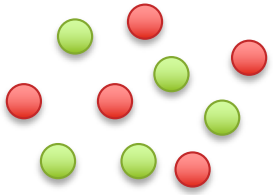
Hi-C



Population Heterogeneity

Red cells express twice the abundance of “brain” genes compared to green cells

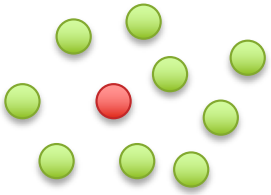
Experiment 1: 50/50



Compared to a control sample of pure green cells, this sample will show:

50% 2x + 50% 1x
= 1.5x over expression of brain genes

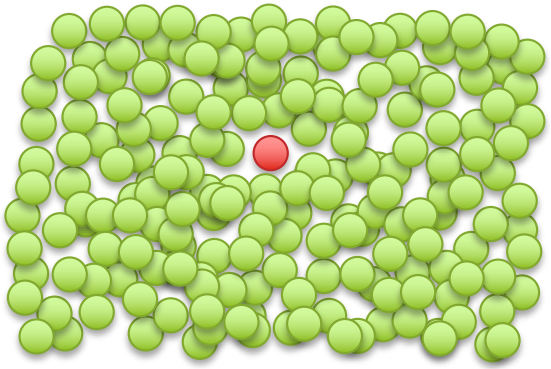
Experiment 2: 1/10



Compared to a control sample of pure green cells, this sample will show:

10% 2x + 90% 1x
= 1.1x over expression of brain genes

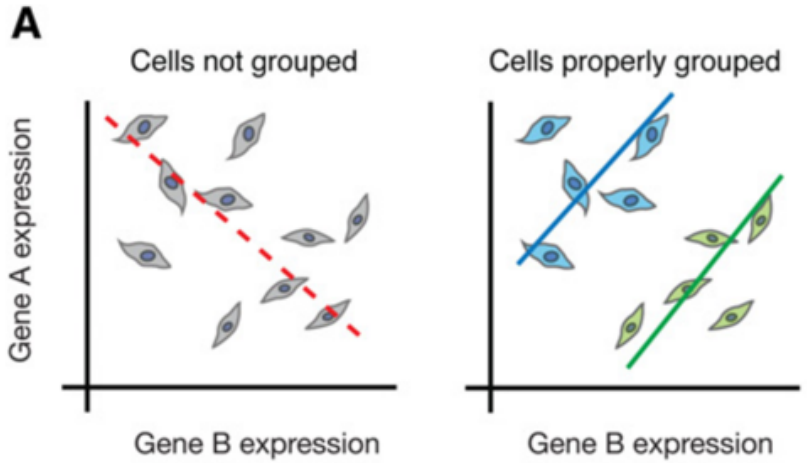
Experiment 3: 1/1000



Compared to a control sample of pure green cells, this sample will show:

0.1% 2x + 99.1% 1x
= 1.001x over expression of brain genes

The paradox of averages



What??? How can the better performing drug depend on if you examine the overall response or separately examine by gender

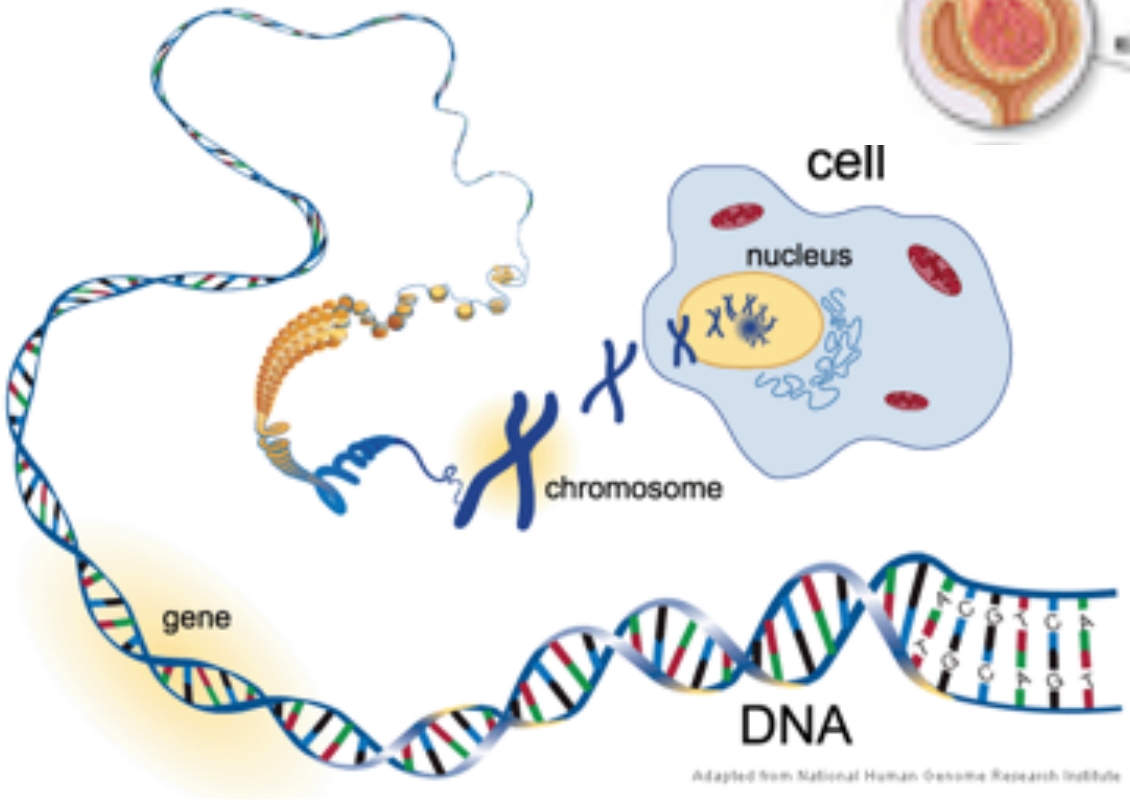
Example of Simpson's paradox:
Trend of the overall average may reverse the trends of each constituent group

In this example, the "lurking" variable (or confounding variable) is the severity of the case (represented by the doctors' treatment decision trend of favoring B for less severe cases), which was not previously known to be important until its effects were included. (Based on real analysis of kidney stone treatments)

(Trapnell, 2015, Genome Research)

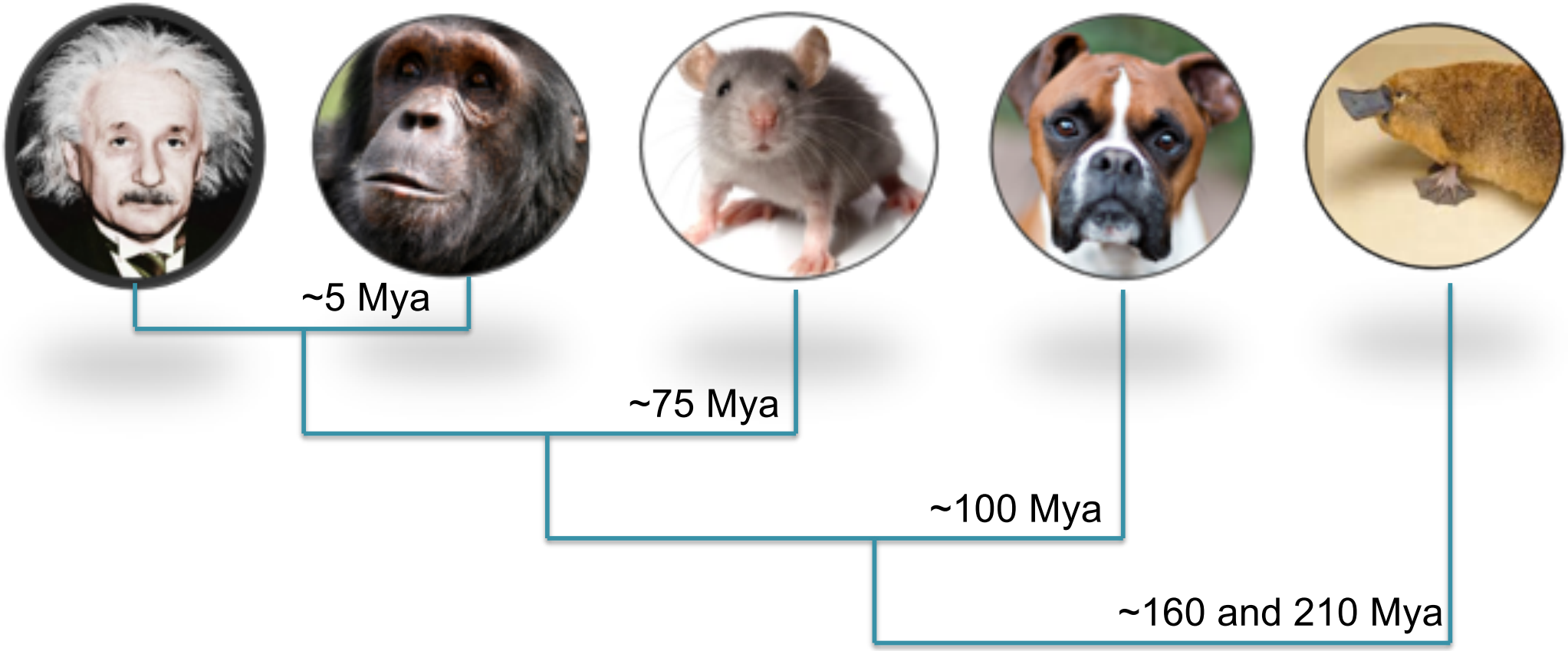
Why Genes?

Each cell of your body contains an exact copy of your 3 billion base pair genome.



Your body has a few hundred (thousands?) major cell types, largely defined by the gene expression patterns

Human Evolution

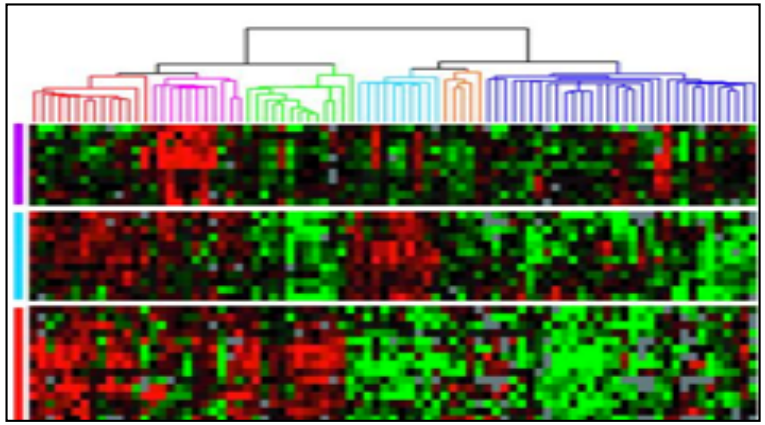


As expected, the majority of platypus genes (82%; 15,312 out of 18,596) have orthologues in these five other amniotes (Supplementary Table 5). The remaining 'orphan' genes are expected to primarily reflect rapidly evolving genes, for which no other homologues are discernible, erroneous predictions, and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

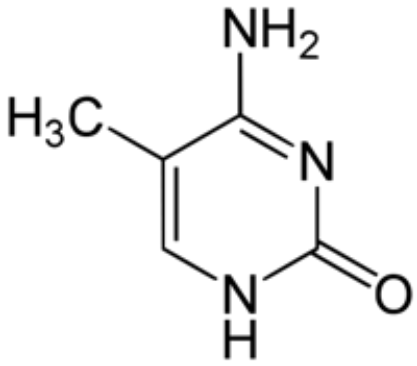
Genome analysis of the platypus reveals unique signatures of evolution
(2008) *Nature.* 453, 175-183 doi:10.1038/nature06936

*-seq in 4 short vignettes

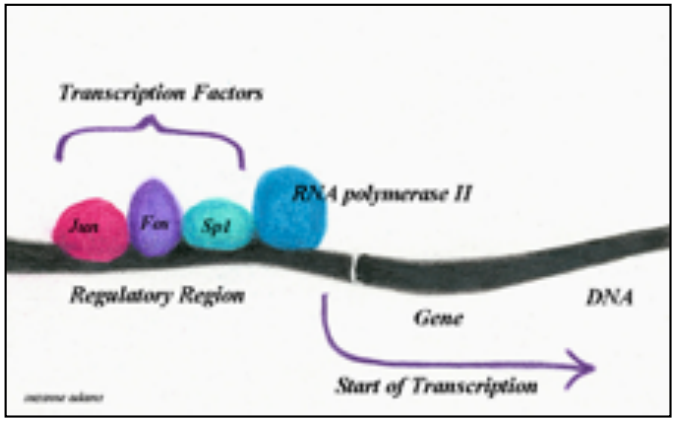
RNA-seq



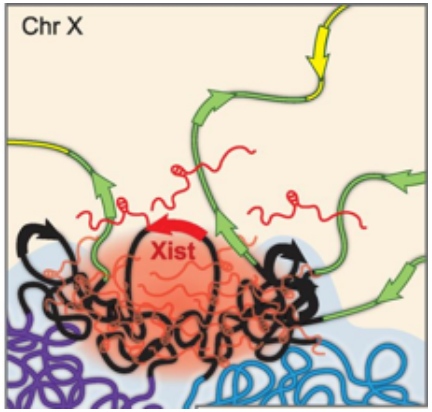
Methyl-seq



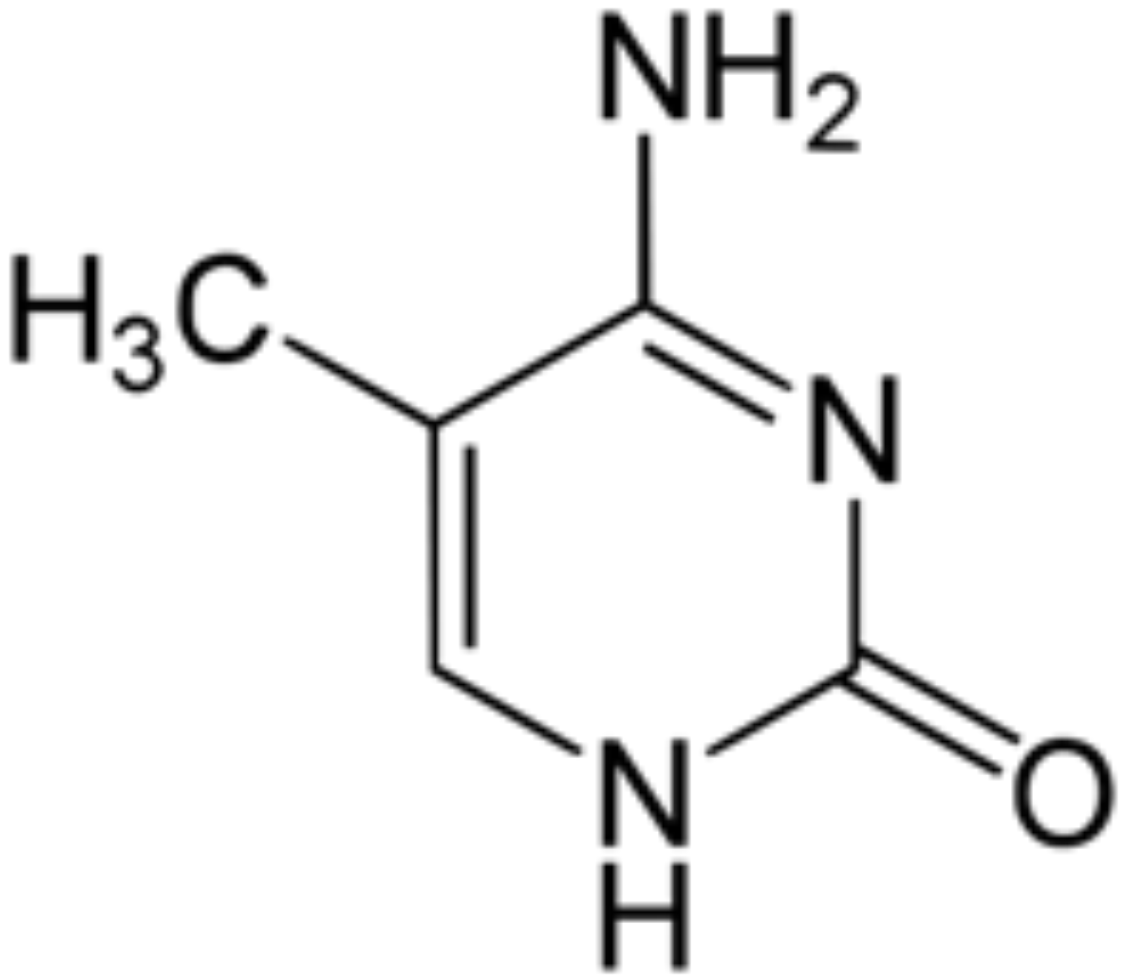
ChIP-seq



Hi-C

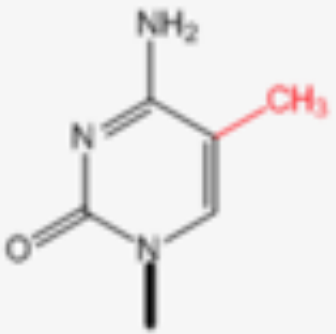


Methyl-seq

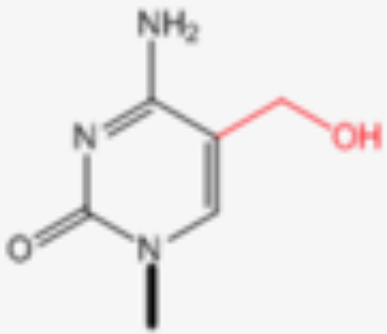


Finding the fifth base: Genome-wide sequencing of cytosine methylation
Lister and Ecker (2009) *Genome Research.* 19: 959-966

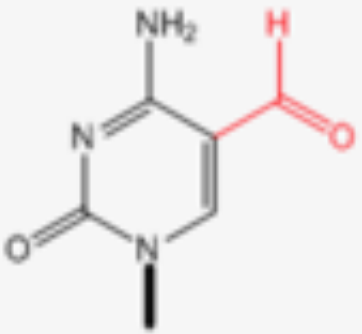
Epigenetic Modifications to DNA



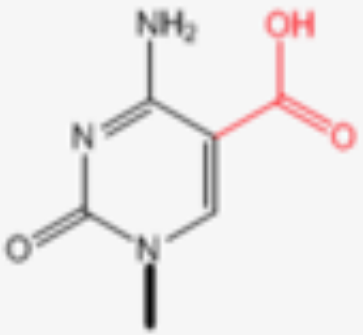
5-mC



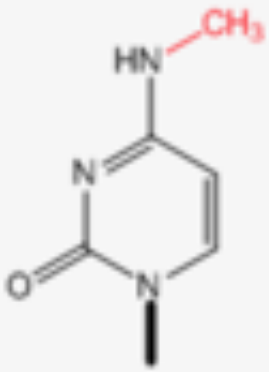
5-hmC



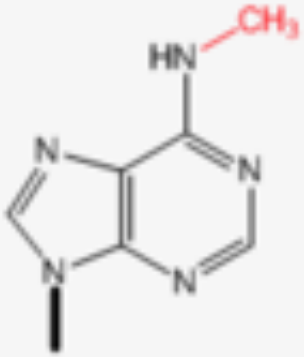
5-fC



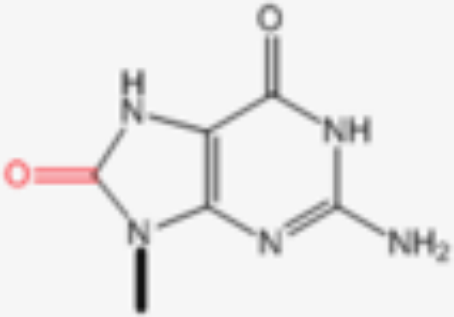
5-caC



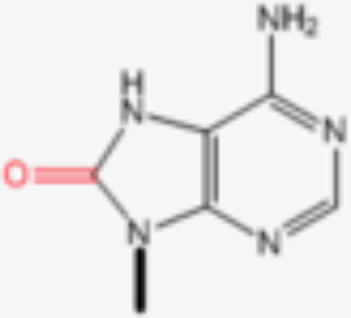
4-mC



6-mA

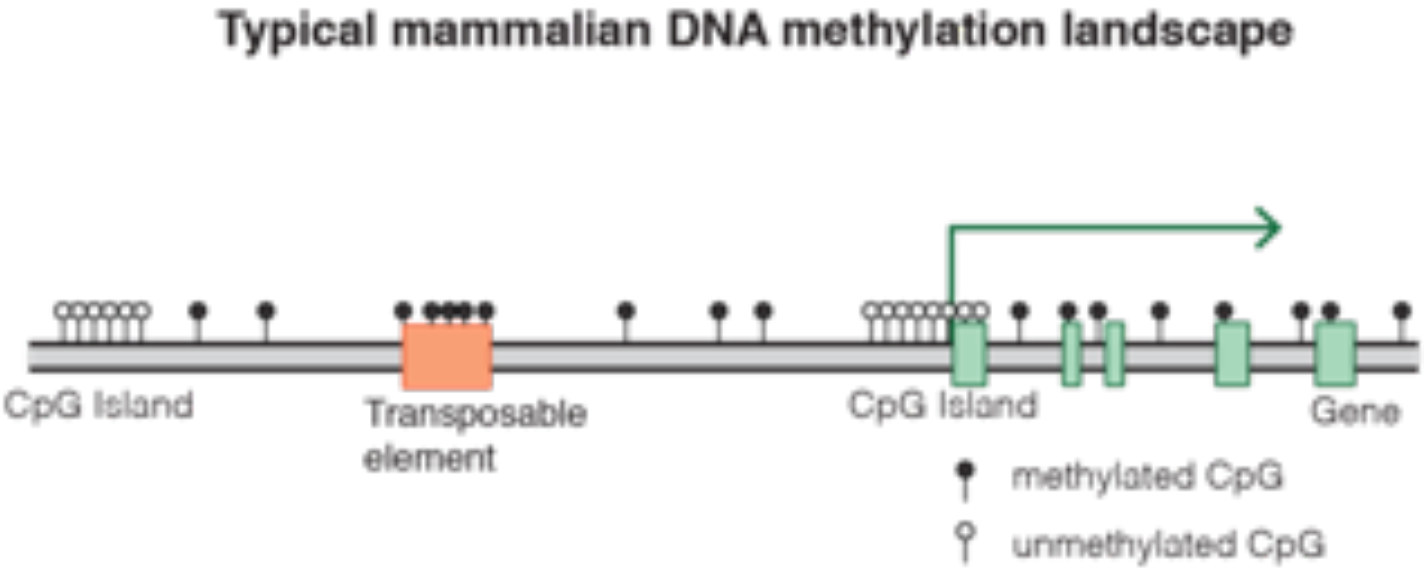


8-oxoG



8-oxoA

Methylation of CpG Islands



CpG islands are (usually) defined as regions with

1) a length greater than 200bp,
2) a G+C content greater than 50%,
3) a ratio of observed to expected CpG greater than 0.6

Methylation in promoter regions correlates negatively with gene expression.

- CpG-dense promoters of actively transcribed genes are never methylated
- In mouse and human, around 60-70% of genes have a CpG island in their promoter region and most of these CpG islands remain unmethylated independently of the transcriptional activity of the gene
- Methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene
- Methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs) that can modify histones, thereby forming compact, inactive chromatin, termed heterochromatin.

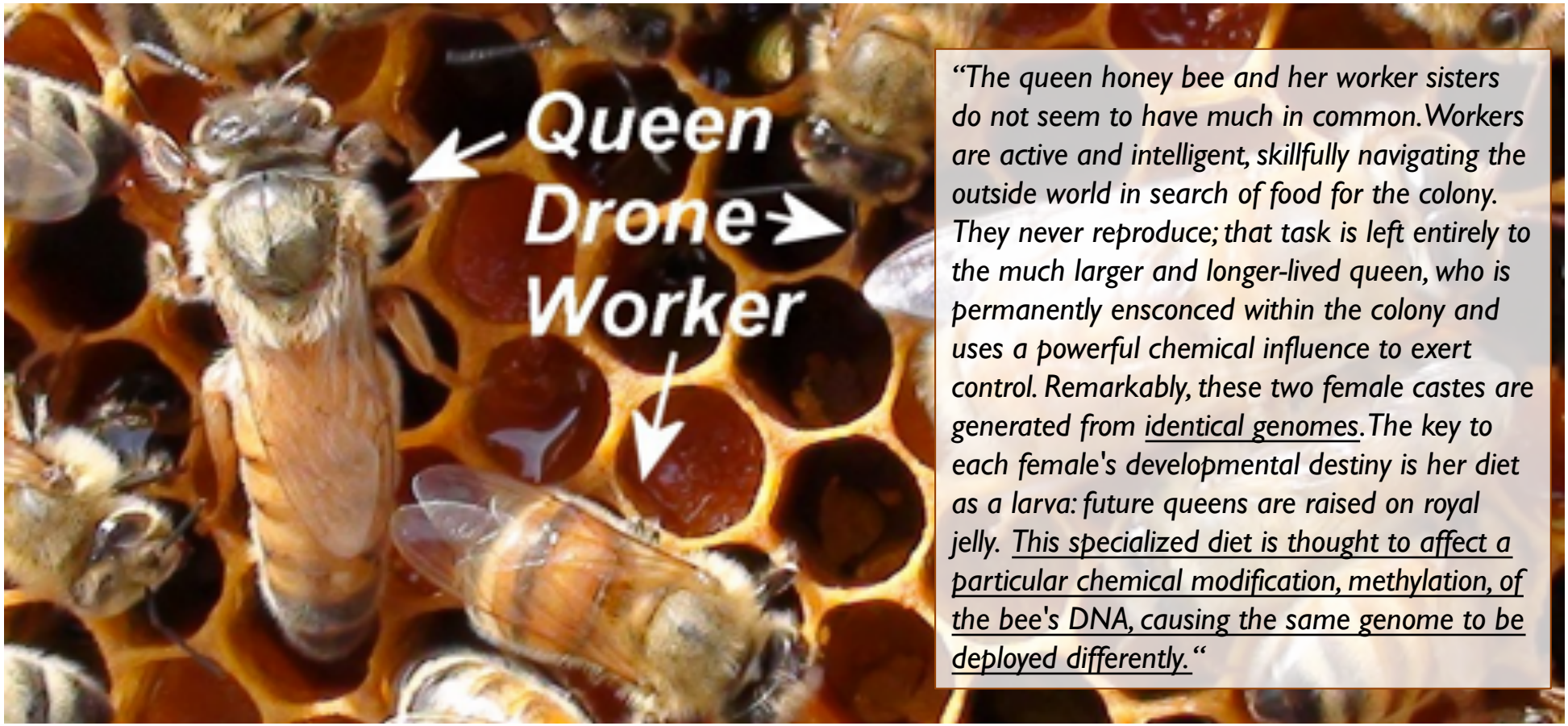
OPEN ACCESS Freely available online

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."

Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm
Ong-Abdullah, et al (2015) *Nature. doi:10.1038/nature15365*

Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm
Ong-Abdullah, et al (2015) *Nature. doi:10.1038/nature15365*

Somaclonal variation arises in plants and animals when differentiated somatic cells are induced into a pluripotent state, but the resulting clones differ from each other and from their parents. In agriculture, somaclonal variation has hindered the micropropagation of elite hybrids and genetically modified crops, but the mechanism responsible remains unknown. The oil palm fruit 'mantled' abnormality is a somaclonal variant arising from tissue culture that drastically reduces yield, and has largely halted efforts to clone elite hybrids for oil production. Widely regarded as an epigenetic phenomenon, 'mantling' has defied explanation, but here we identify the MANTLED locus using epigenome-wide association studies of the African oil palm Elaeis guineensis. DNA hypomethylation of a LINE retrotransposon related to rice Karma, in the intron of the homeotic gene DEFICIENS, is common to all mantled clones and is associated with alternative splicing and premature termination. **Dense methylation near the Karma splice site (termed the Good Karma epiallele) predicts normal fruit set, whereas hypomethylation (the Bad Karma epiallele) predicts homeotic transformation, parthenocarpy and marked loss of yield.** *Loss of Karma methylation and of small RNA in tissue culture contributes to the origin of mantled, while restoration in spontaneous revertants accounts for non-Mendelian inheritance. The ability to predict and cull mantling at the plantlet stage will facilitate the introduction of higher performing clones and optimize environmentally sensitive land resources.*

Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm
Ong-Abdullah, et al (2015) *Nature. doi:10.1038/nature15365*

Hypomethylation distinguishes genes of some human cancers from their normal counterparts

Andrew P. Feinberg & Bert Vogelstein

Cell Structure and Function Laboratory, The Oncology Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

It has been suggested that cancer represents an alteration in DNA, heritable by progeny cells, that leads to abnormally regulated expression of normal cellular genes; DNA alterations such as mutations[1,2], rearrangements[3–5] and changes in methylation[6–8] have been proposed to have such a role. Because of increasing evidence that DNA methylation is important in gene expression (for review see refs 7, 9–11), several investigators have studied DNA methylation in animal tumours, transformed cells and leukaemia cells in culture[8,12–30]. The results of these studies have varied; depending on the techniques and systems used, an increase[12–19], decrease[20–24], or no change[25–29] in the degree of methylation has been reported. To our knowledge, however, primary human tumour tissues have not been used in such studies. We have now examined DNA methylation in human cancer with three considerations in mind: (1) the methylation pattern of specific genes, rather than total levels of methylation, was determined; (2) human cancers and adjacent analogous normal tissues, unconditioned by culture media, were analysed; and (3) the cancers were taken from patients who had received neither radiation nor chemotherapy. In four of five patients studied, representing two histological types of cancer, substantial hypomethylation was found in genes of cancer cells compared with their normal counterparts. This hypomethylation was progressive in a metastasis from one of the patients.

and (3) *Hpa*II and *Hha*I cleavage sites should be present in the regions of the genes.

The first cancer studied was a grade D (ref. 43), moderately well differentiated adenocarcinoma of the colon from a 67-yr-old male. Tissue was obtained from the cancer itself and also from colonic mucosa stripped from the colon at a site just outside the histologically proven tumour margin. Figure 1 shows the pattern of methylation of the studied genes. Before digestion with restriction enzymes, all DNA samples used in the study had a size >25,000 base pairs (bp). After *Hpa*II cleavage, hybridization with a probe made from a cDNA clone of human growth hormone (HGH) showed that significantly more of the DNA was digested to low-molecular weight fragments in DNA from the cancer (labelled C in Fig. 1) than in DNA from the normal colonic mucosa (labelled N). In the hybridization conditions used, the HGH probe detected the human growth hormone genes as well as the related chorionic somatotropin

Table 1 Quantitation of methylation of specific genes in human cancers and adjacent analogous normal tissues

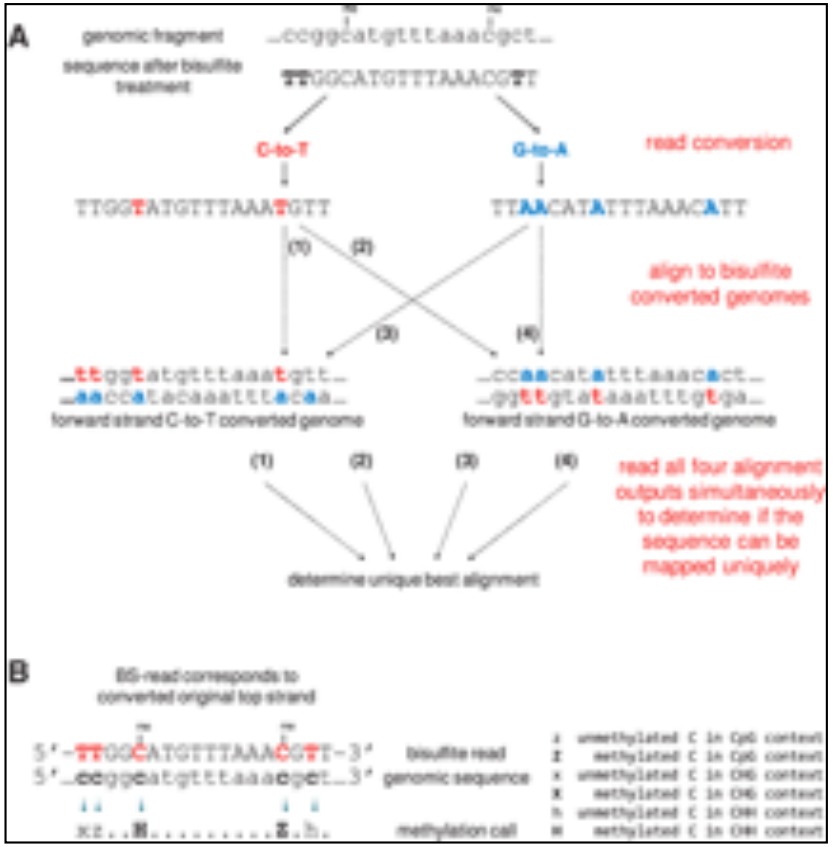
Patient	Carcinoma	Probe	Enzyme	% Hypomethylated fragments		
				N	C	M
1	Colon	HGH	*Hpa*II	<10	35	—
			*Hha*I	<10	39	—
		γ-Globin	*Hpa*II	<10	52	—
			*Hha*I	<10	39	—
		α-Globin	*Hpa*II	<10	<10	—
			*Hha*I	<10	<10	—
2	Colon	HGH	*Hpa*II	<10	76	—
			*Hha*I	<10	85	—
		γ-Globin	*Hpa*II	<10	58	—
			*Hha*I	<10	23	—
		α-Globin	*Hpa*II	<10	<10	—
			*Hha*I	<10	<10	—
3	Colon	HGH	*Hpa*II	<10	41	—
			*Hha*I	<10	38	—
		γ-Globin	*Hpa*II	<10	50	—
			*Hha*I	<10	22	—

Nature 301, 89 - 92 (06 January 1983); doi:10.1038/301089a0

Bisulfite Conversion

Treating DNA with sodium bisulfite will convert unmethylated C to T

- 5-MethylC will be protected and not change, so can look for differences when mapping
- Requires great care when analyzing reads, since the complementary strand will also be converted (G to A)
- Typically analyzed by mapping to a "reduced alphabet" where we assume all Cs are converted to Ts once on the forward strand and once on the reverse



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

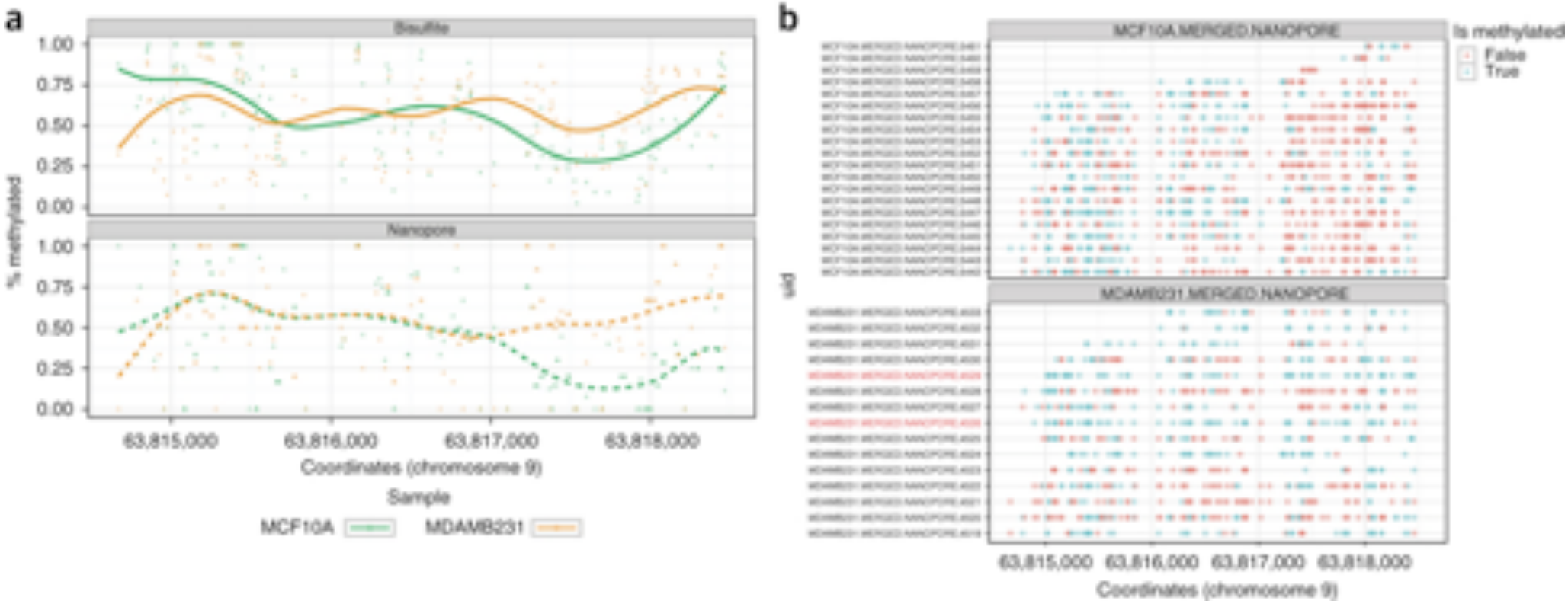
Bisulfite Conversion

T
w

-
-
-

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

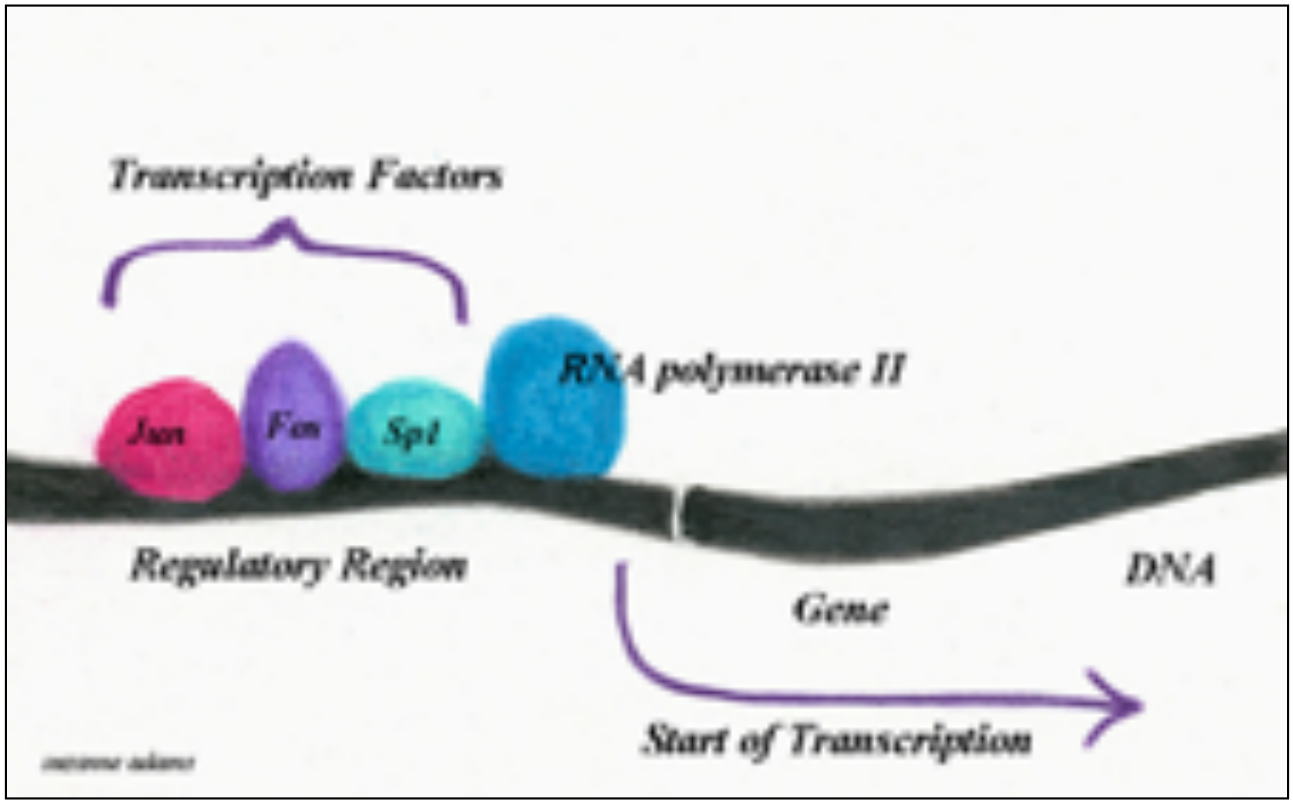
Methylation changes in cancer detected by Nanopore Sequencing



Comparison of bisulfite sequencing and nanopore-based R7.3 data in reduced representation data sets from cancer and normal cells. (a) Raw data (points) and smoothed data (lines) for methylation, as determined by bisulfite sequencing (top) and nanopore-based sequencing using an R7.3 pore (bottom), in a genomic region from the human mammary epithelial cell line MCF10A (green) and metastatic mammary epithelial cell line MDA-MB-231 (orange). (b) Same region as in a but with individual nanopore reads plotted separately. Each CpG that can be called is a point. Blue indicates methylated; red indicates unmethylated.

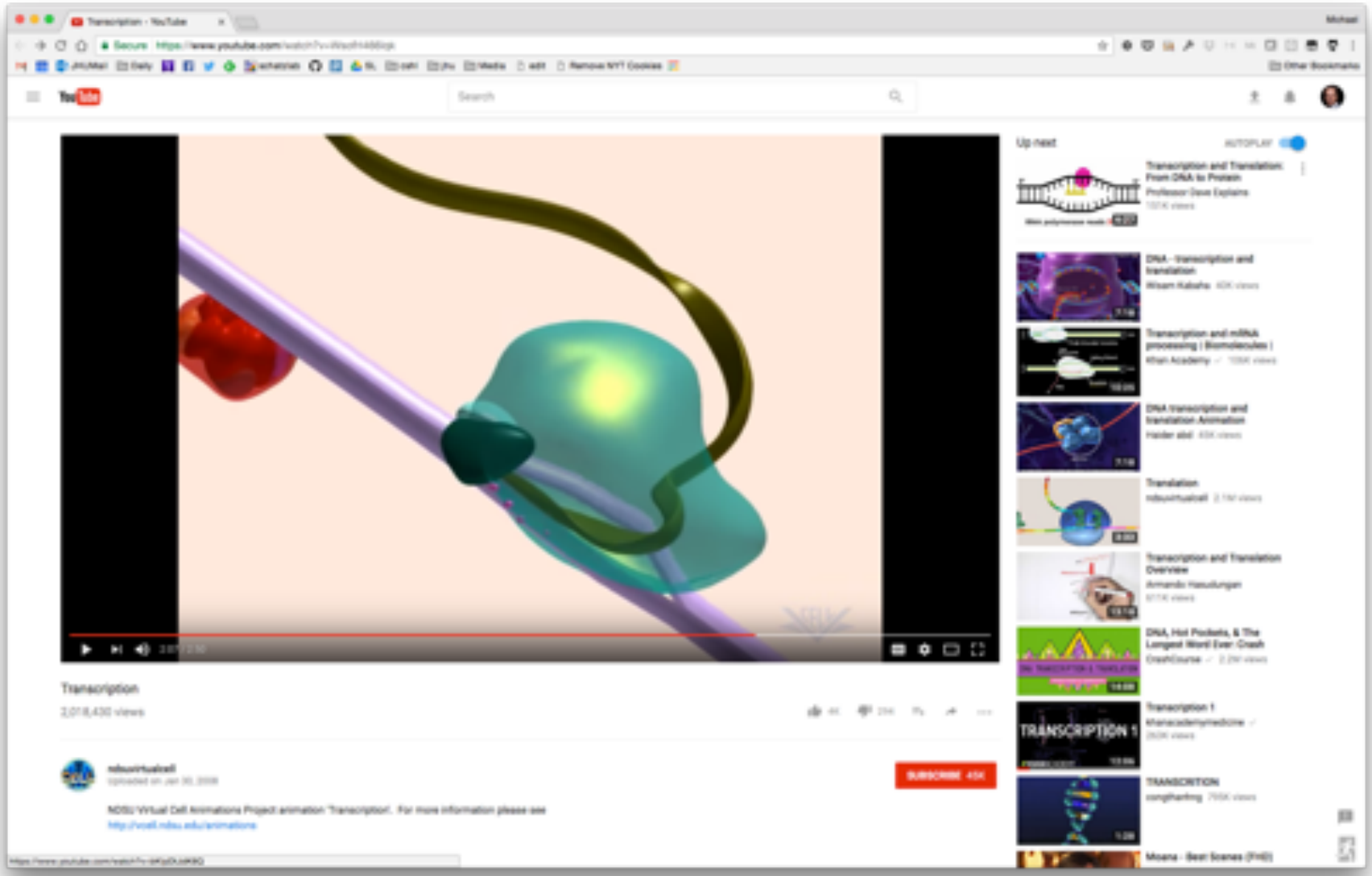
Detecting DNA cytosine methylation using nanopore sequencing
Simpson, Workman, Zuzarte, David, Dursi, Timp (2017) Nature Methods. doi:10.1038/nmeth.4184

ChIP-seq



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

Transcription

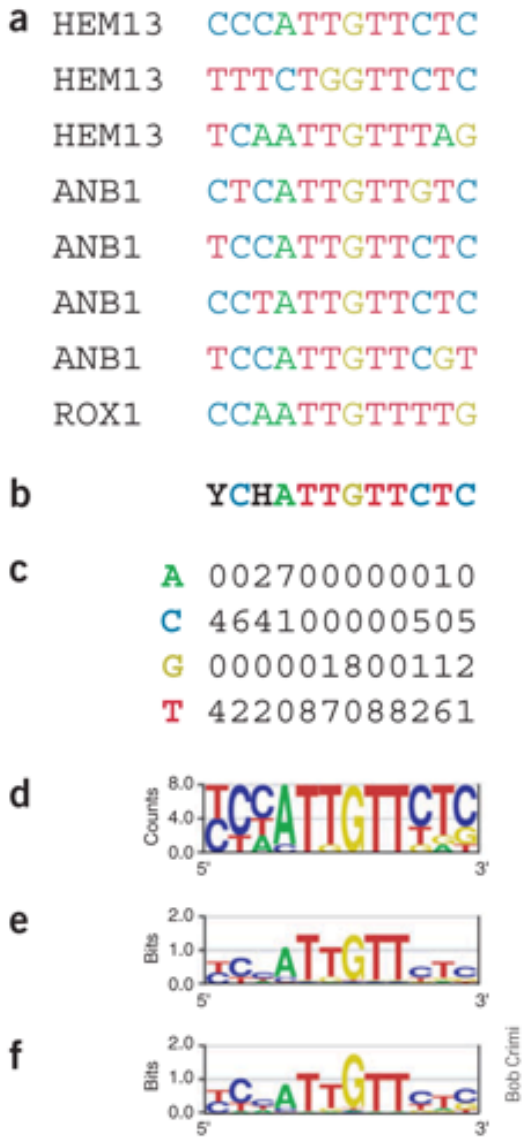


https://www.youtube.com/watch?v=WsofH466lqk

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.

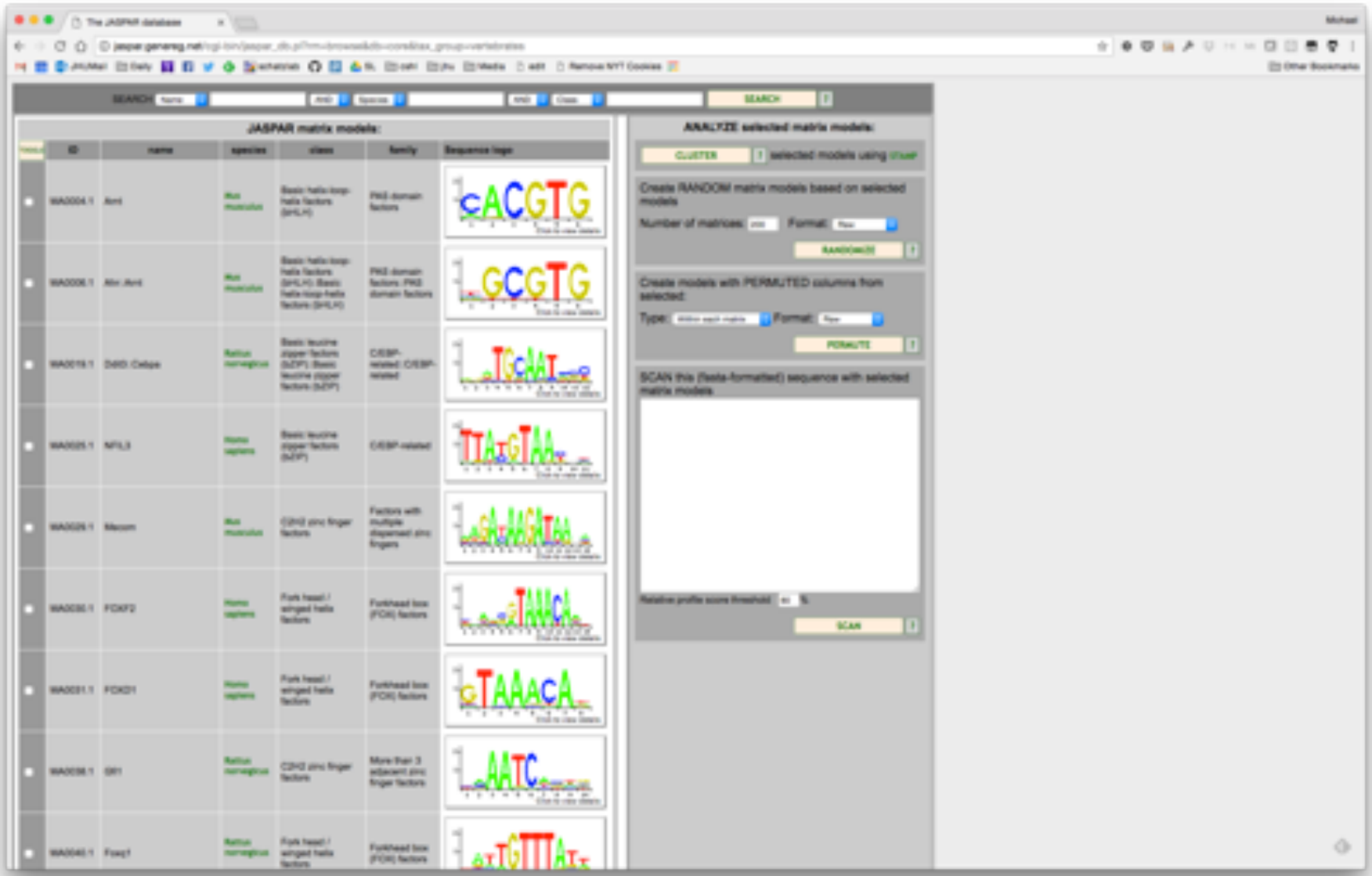
- Transcription factors work alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase to specific genes.
- A defining feature of transcription factors is that they contain at least one DNA-binding domain (DBD)
- Figure (a) 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.



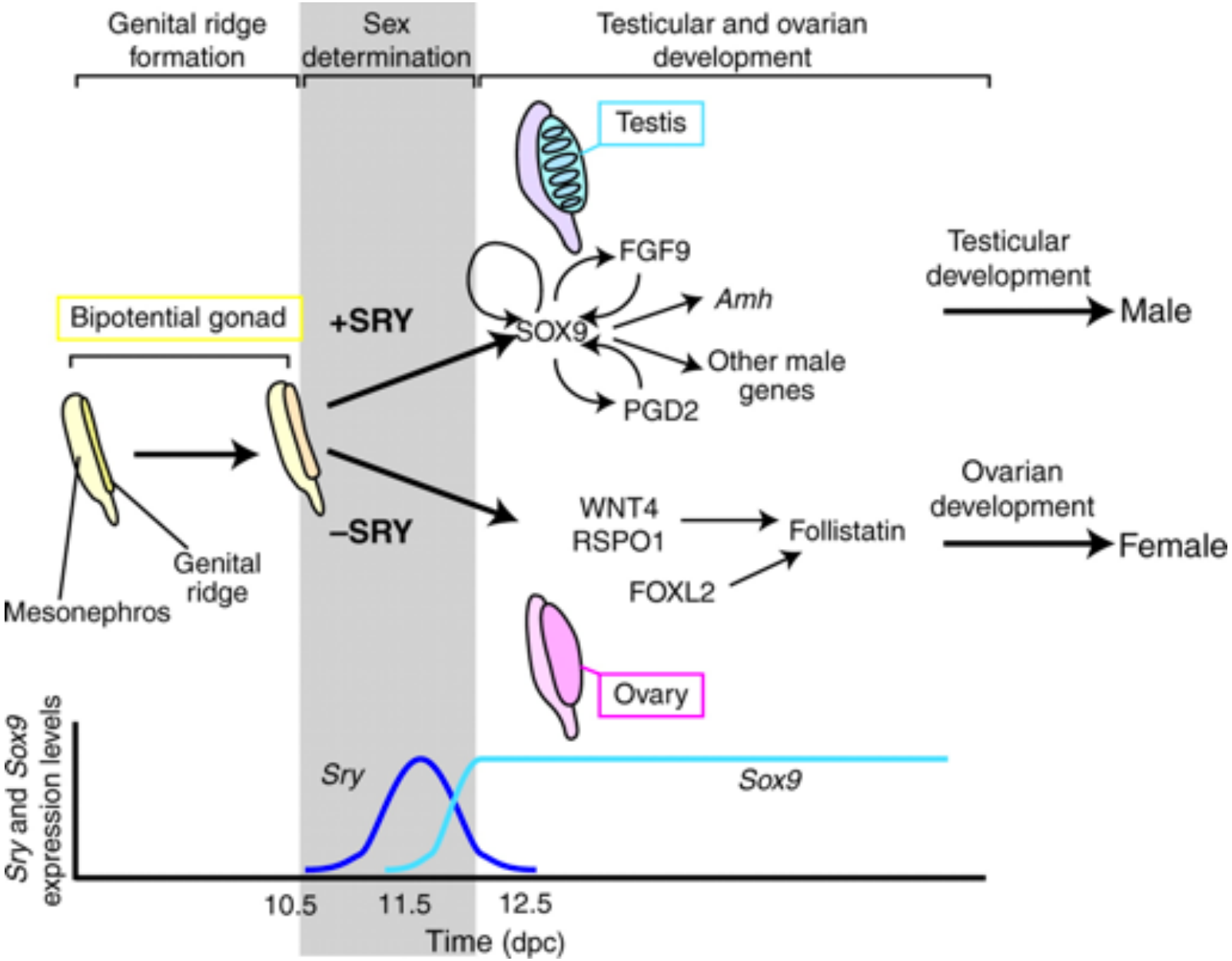
What are DNA sequence motifs?

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

Transcription Factors Database

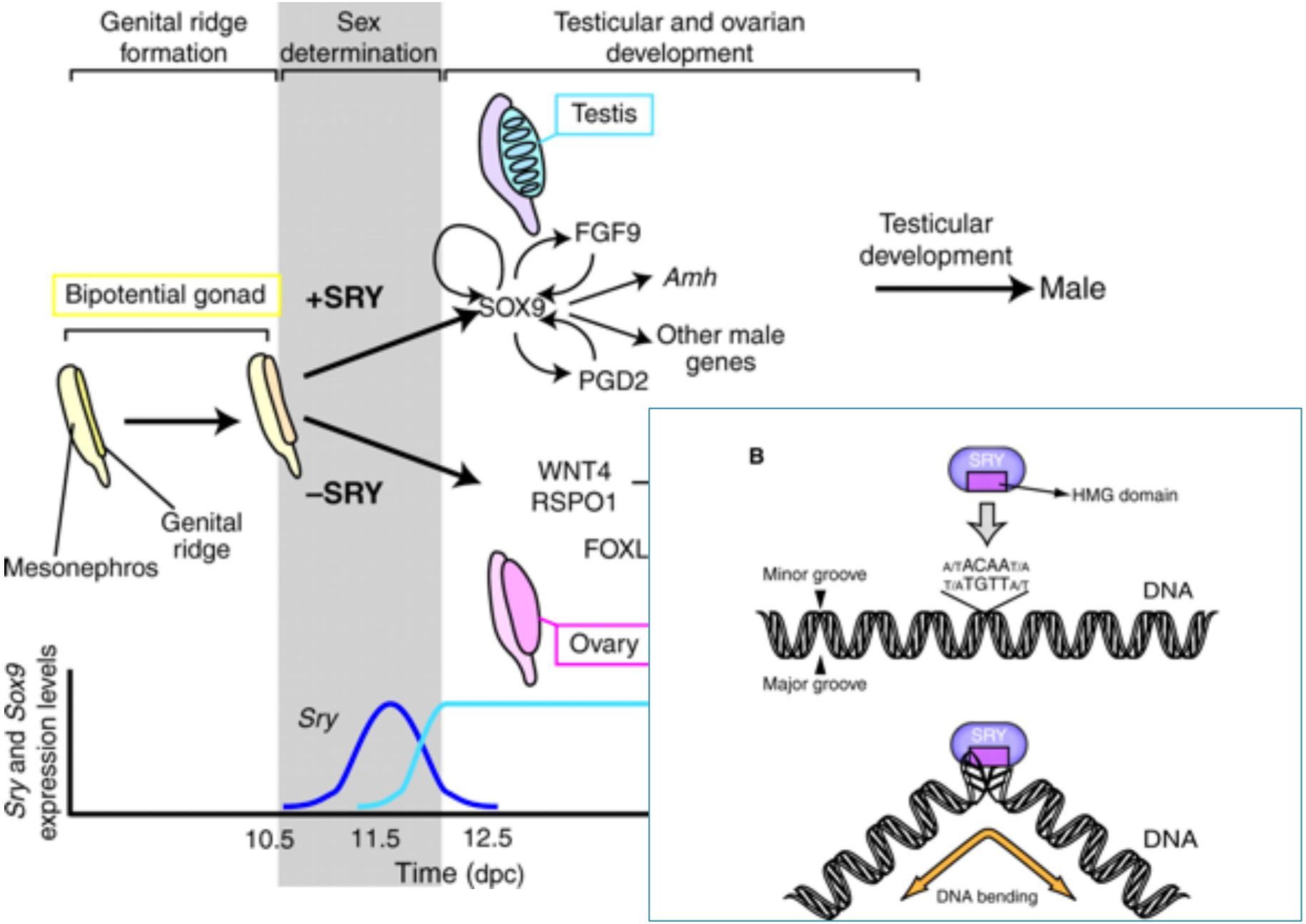


JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles
Anthony Mathelier (2014) Nucleic Acids Res. 42 (D1): D142-D147. DOI: https://doi.org/10.1093/nar/gkt997



SRY: The master switch in mammalian sex determination

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



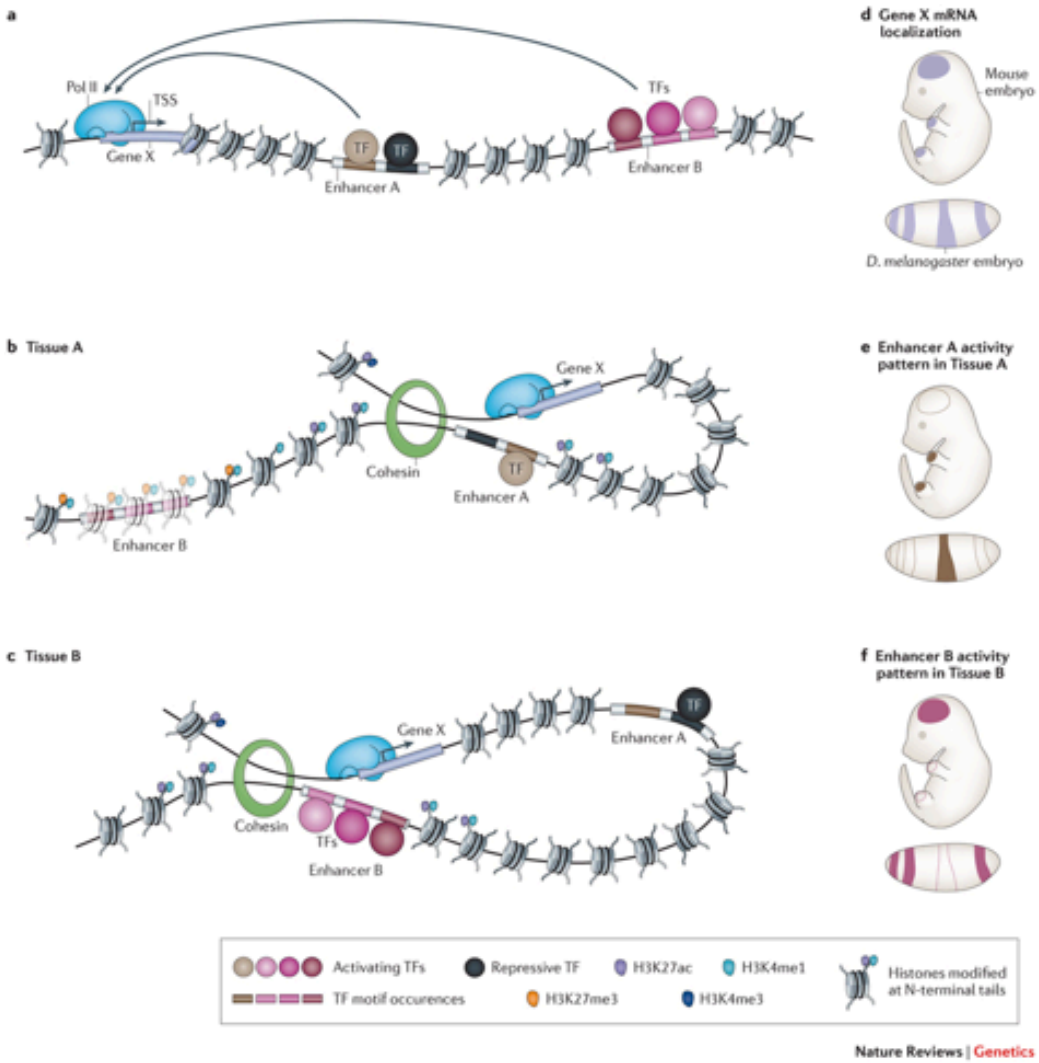
SRY: The master switch in mammalian sex determination

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

Enhancers

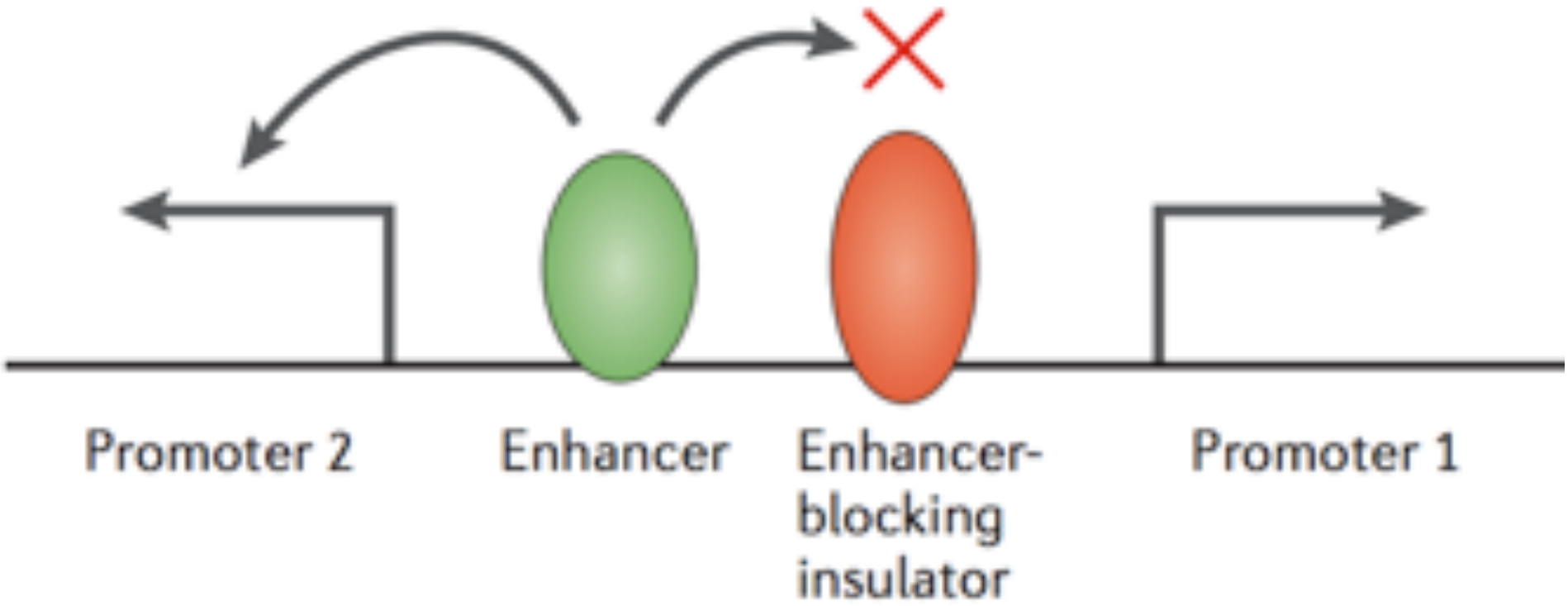
Enhancers are genomic regions that contain binding sites for transcription factors (TFs) and that can upregulate (enhance) the transcription of a target gene.

- 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.

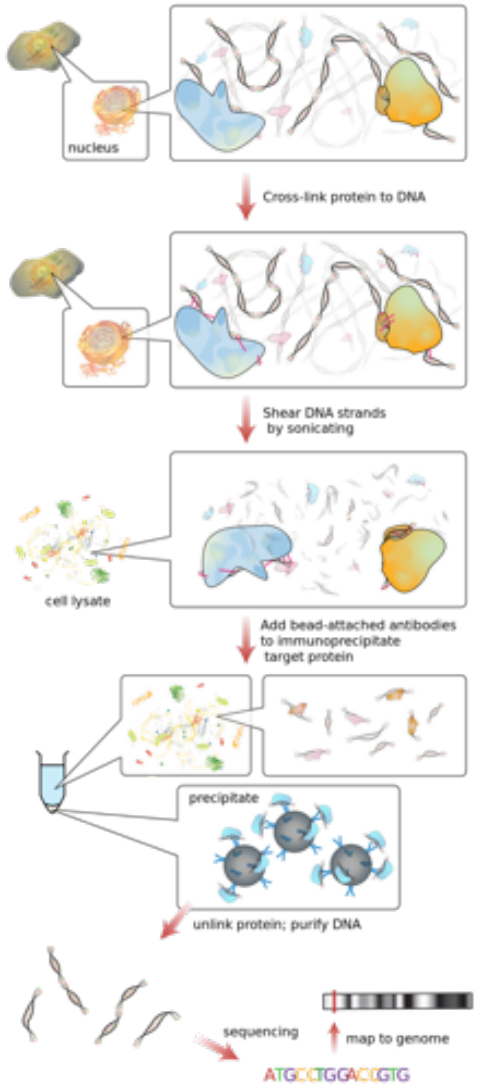
- One type of insulator establishes domains that separate enhancers and promoters to block their interaction,
- Second type creates a barrier against the spread of heterochromatin.

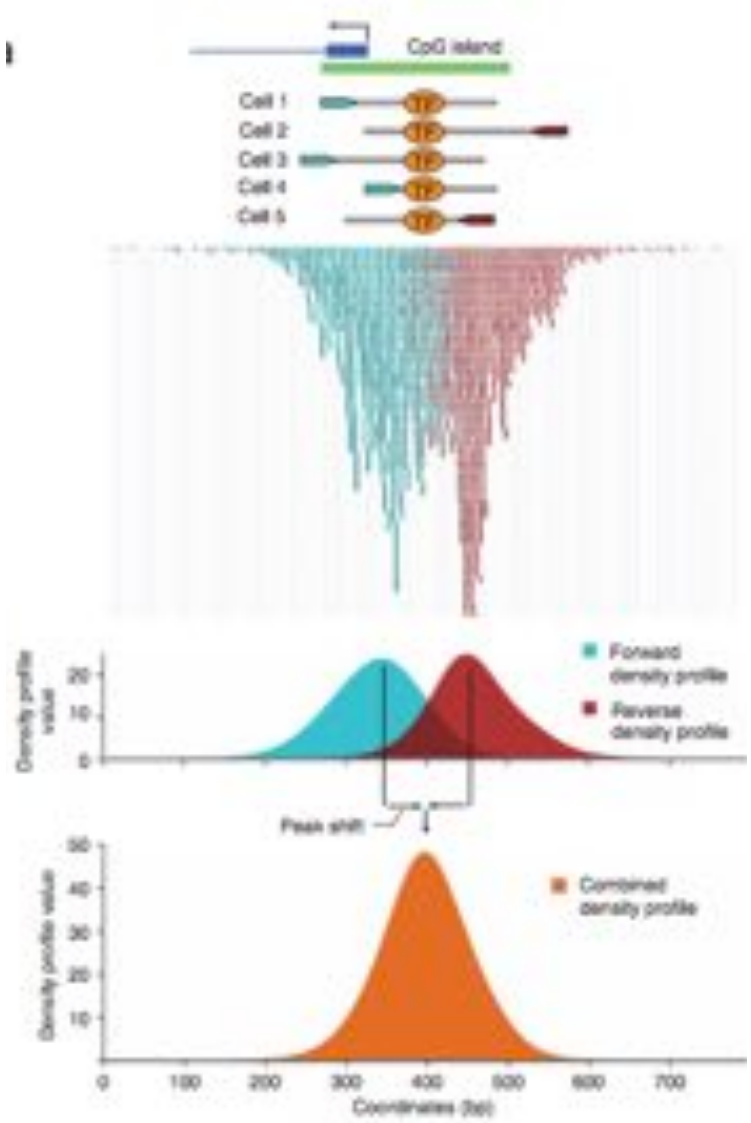
Insulators: exploiting transcriptional and epigenetic mechanisms
Gaszner & Felsenfeld (2006) *Nature Reviews Genetics* 7, 703-713. doi:10.1038/nrg1925

ChIP-seq: TF Binding

Goals:

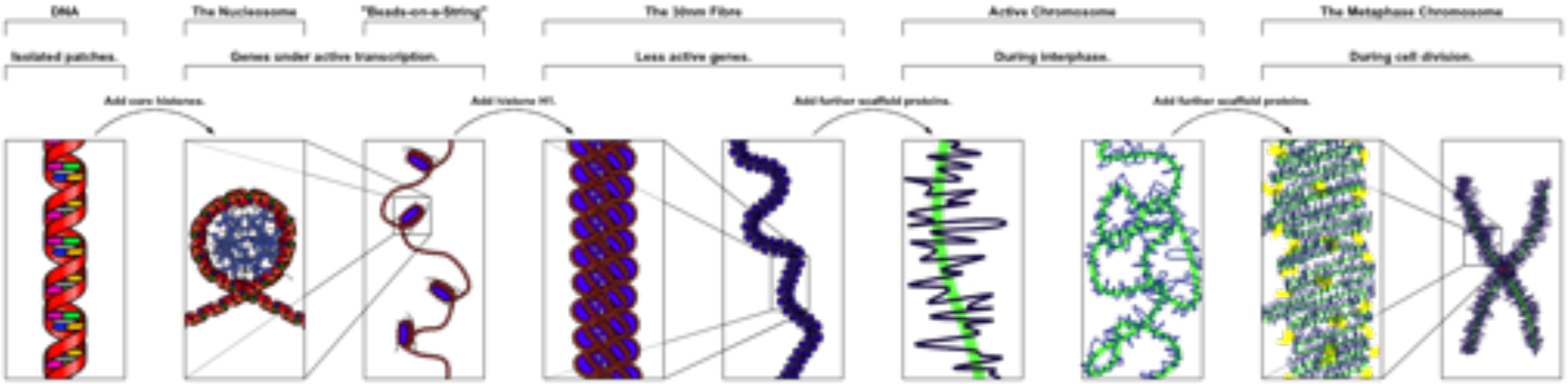
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?





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



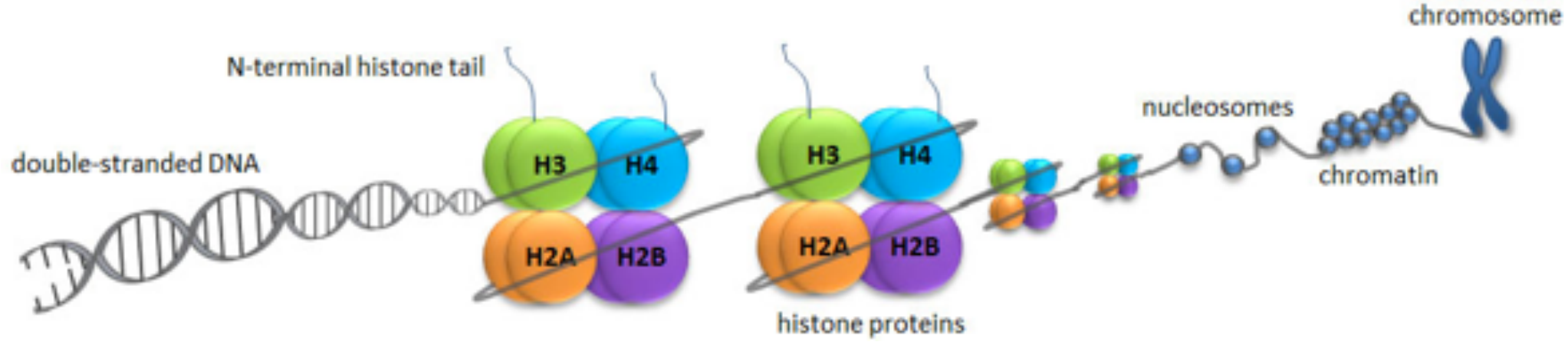
Nucleosome is a basic unit of DNA packaging in eukaryotes

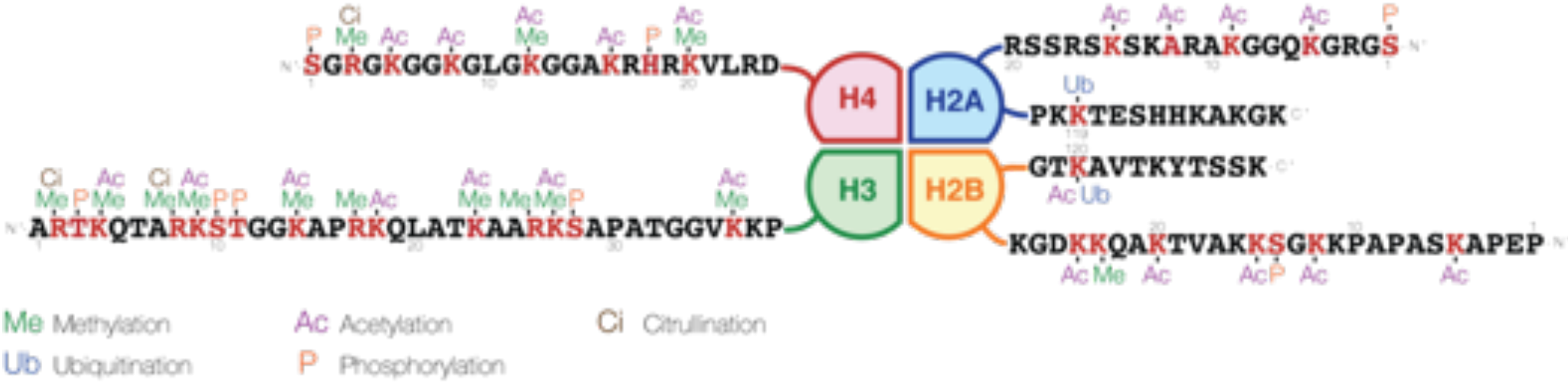
- Consists of a segment of 146bp DNA wound in sequence around eight histone protein cores (thread wrapped around a spool) followed by a ~38bp linker
- Under active transcription, nucleosomes appear as “beads-on-a-string”, but are more densely packed for less active genes

Nucleosomes form the fundamental repeating units of eukaryotic chromatin

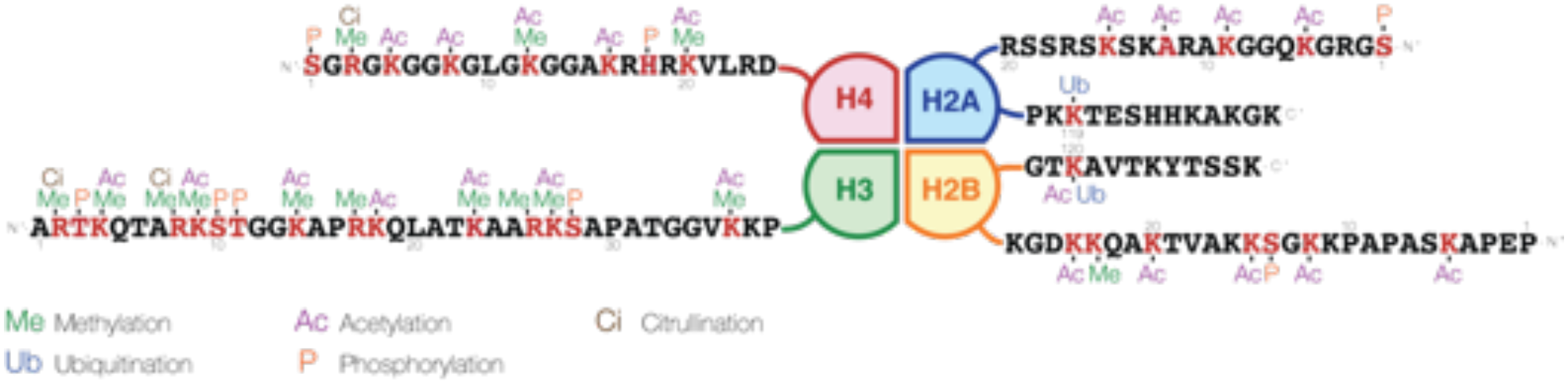
- Used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 μm diameter).

ChIP-seq: Histone Modifications





ChIP-seq: Histone Modifications

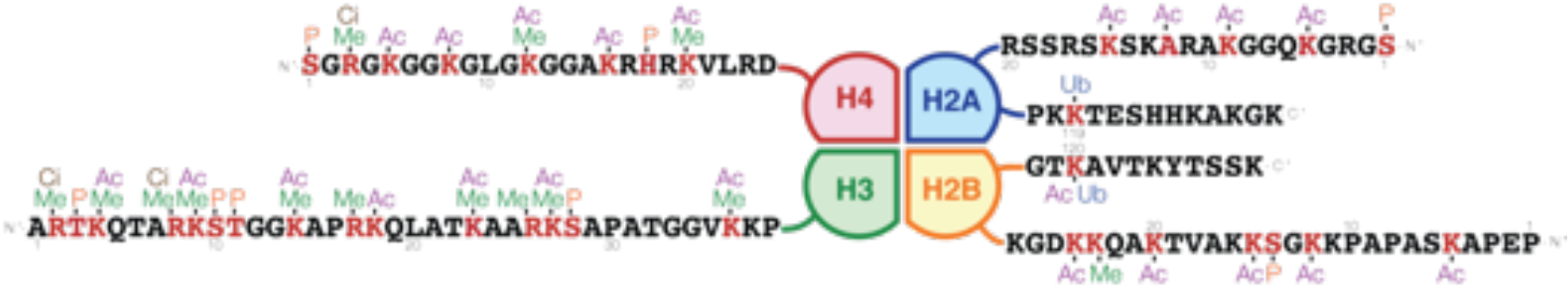


The common nomenclature of histone modifications is:

- 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.

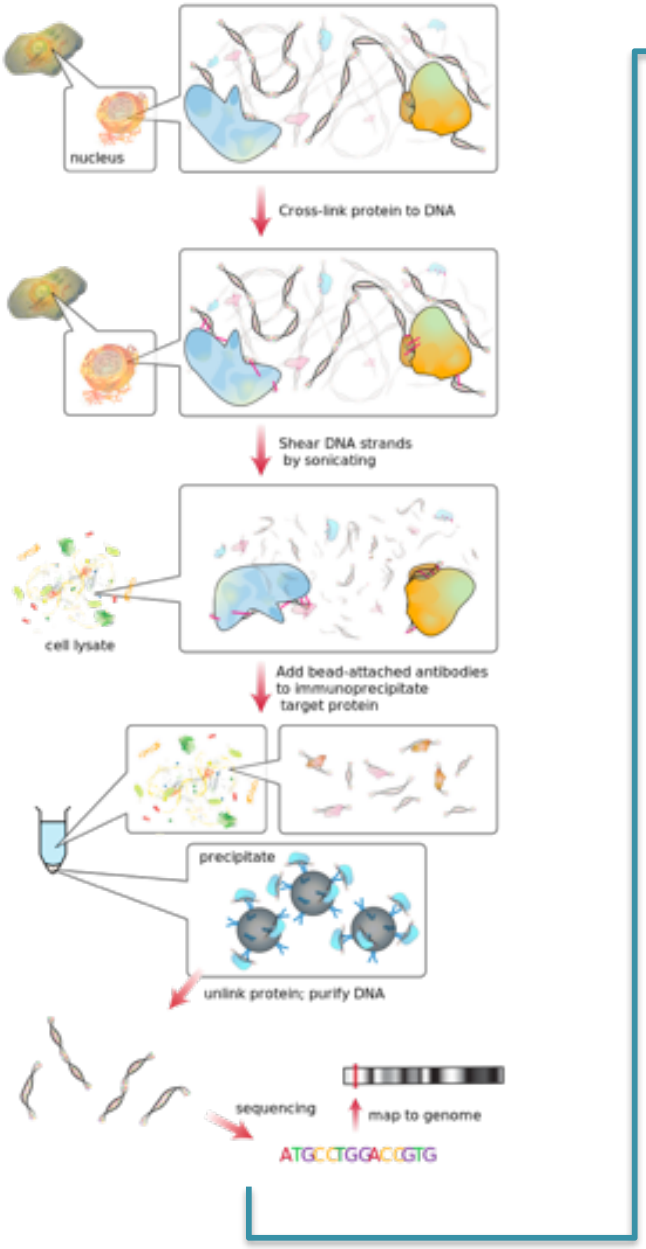
ChIP-seq: Histone Modifications



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

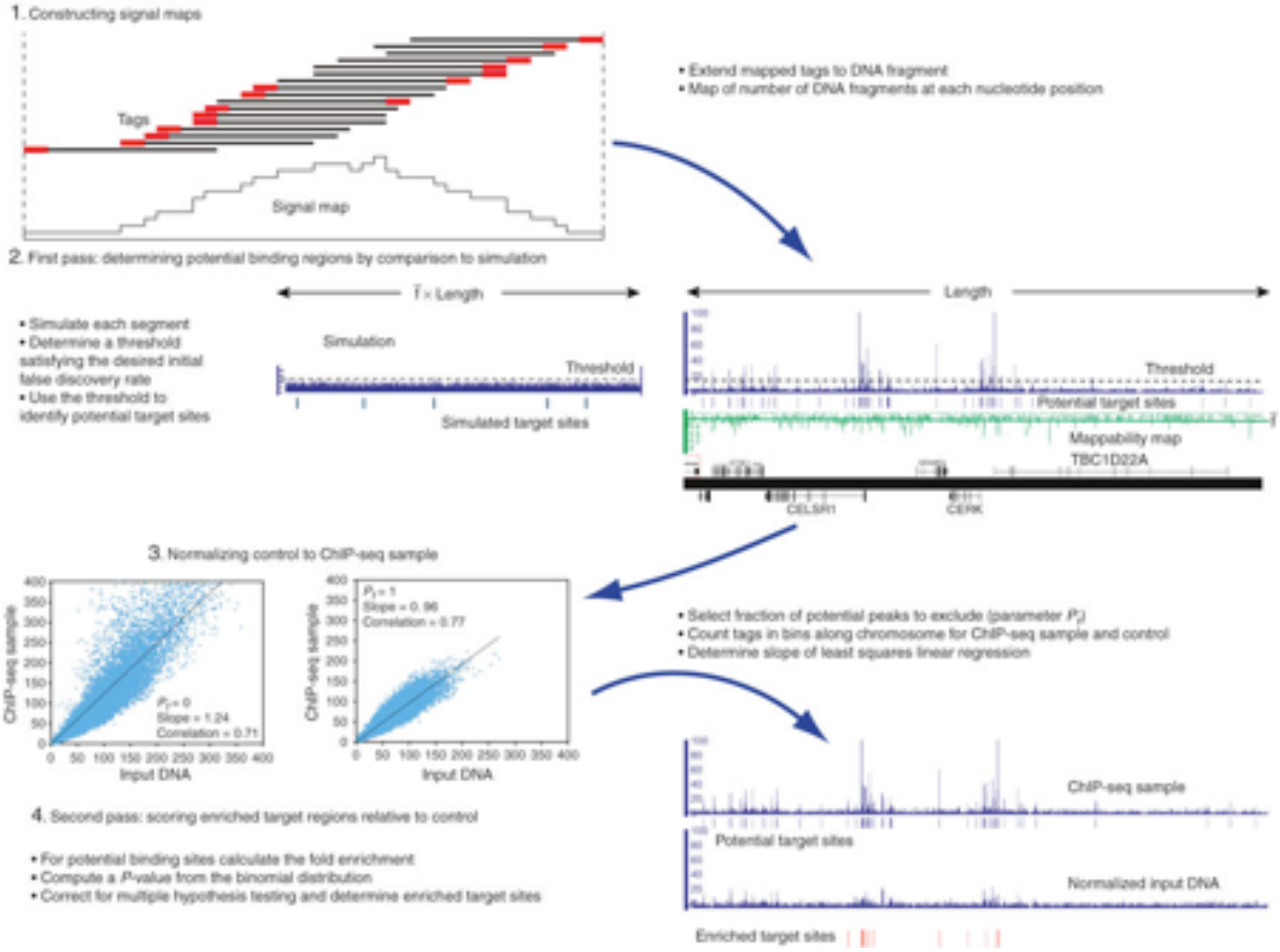
Analyze Enriched Regions

Visualization

Peak-peak comparisons

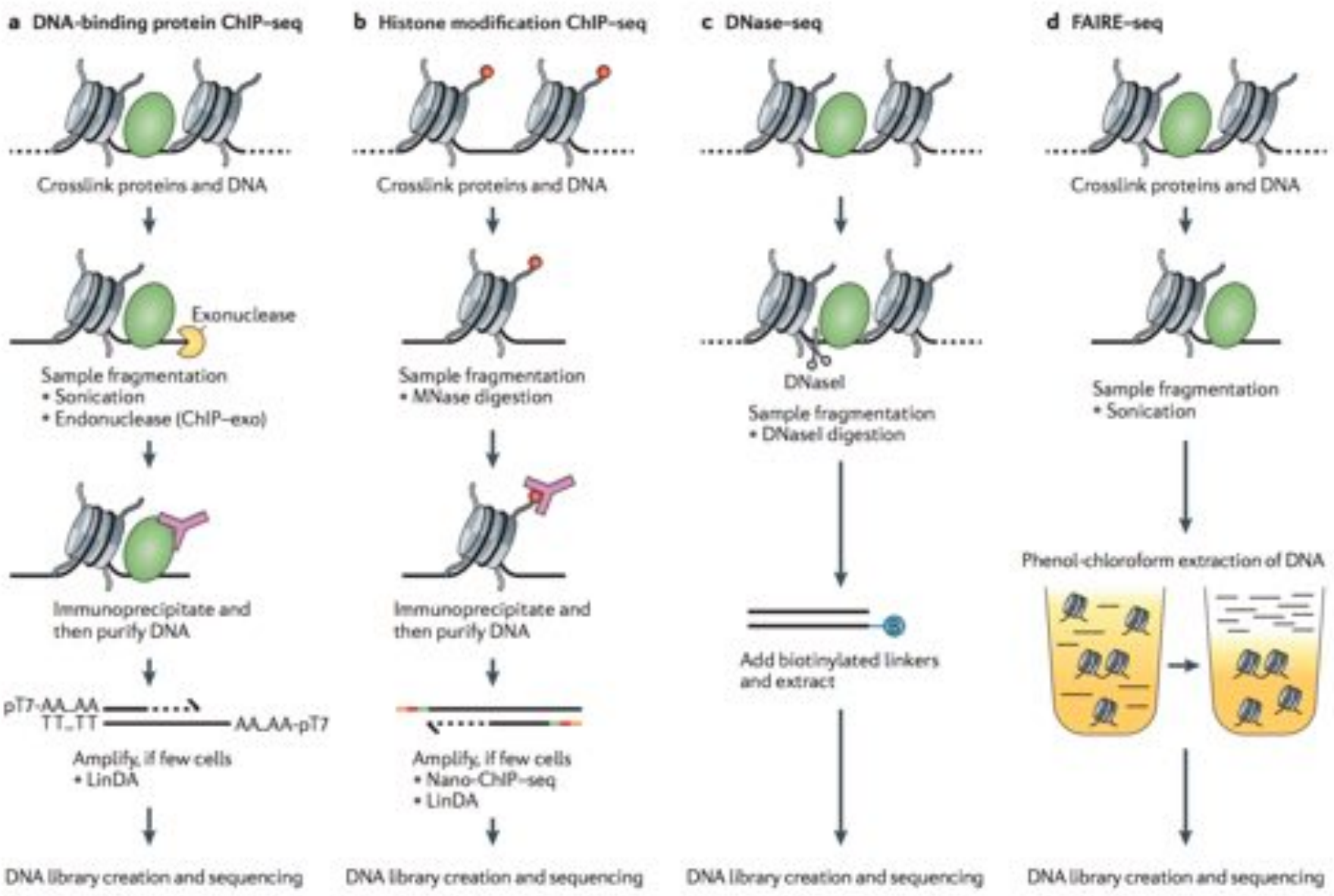
Annotation / Interpretation

PeakSeq



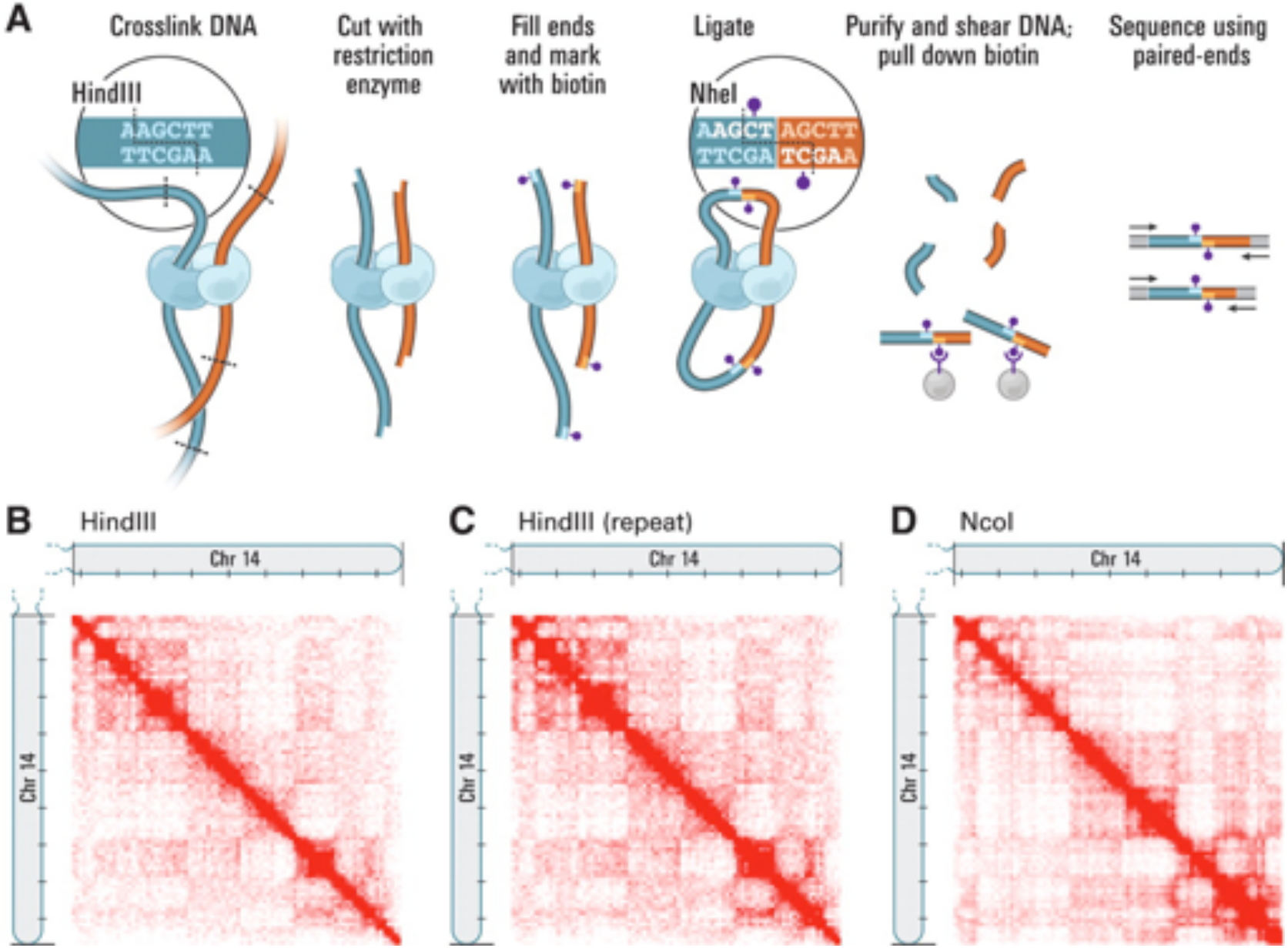
PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls
Rozowsky et al (2009) Nature Biotechnology 27, 66 - 75

Related Assays



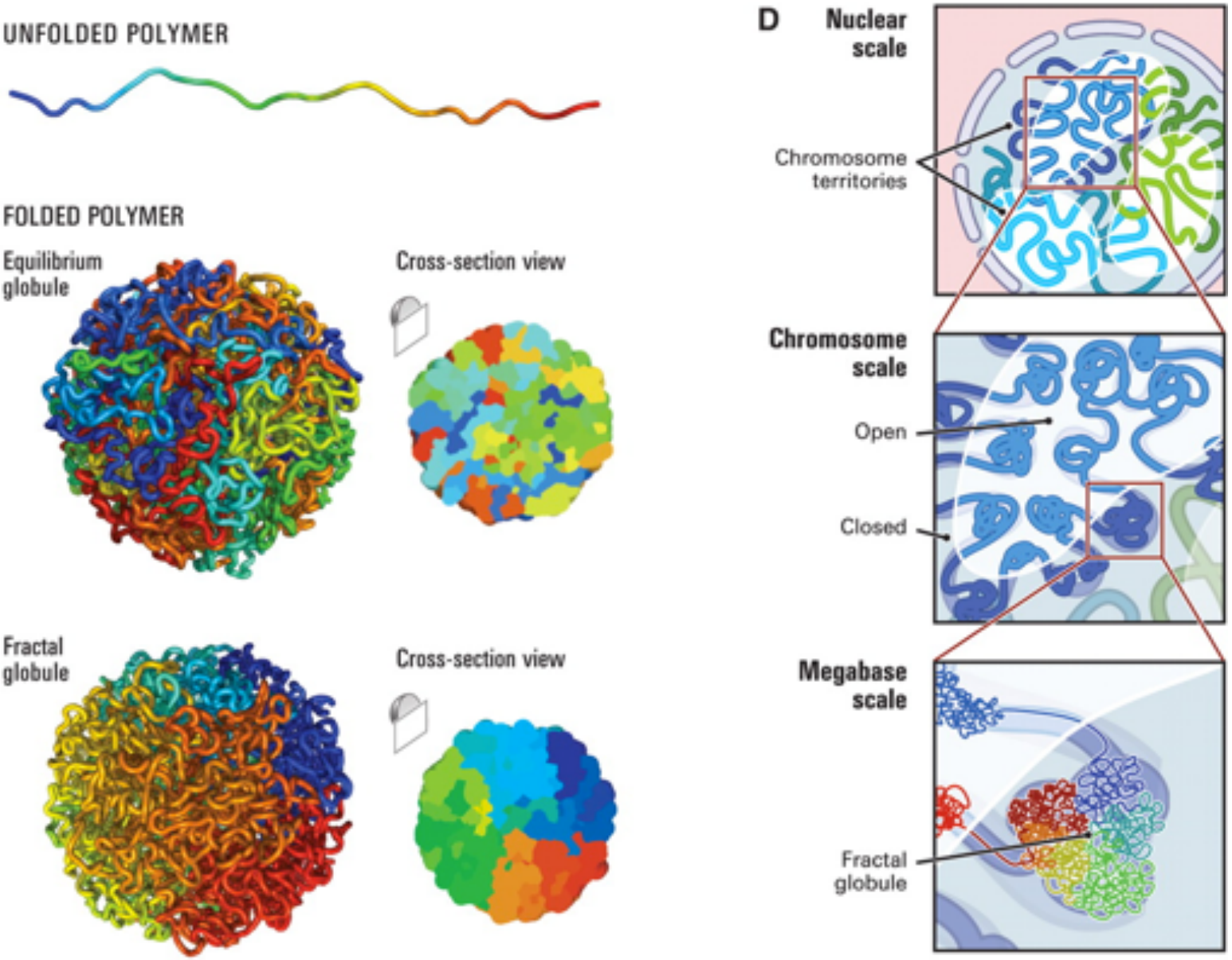
ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions
Furey (2012) *Nature Reviews Genetics*. 13, 840-852

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

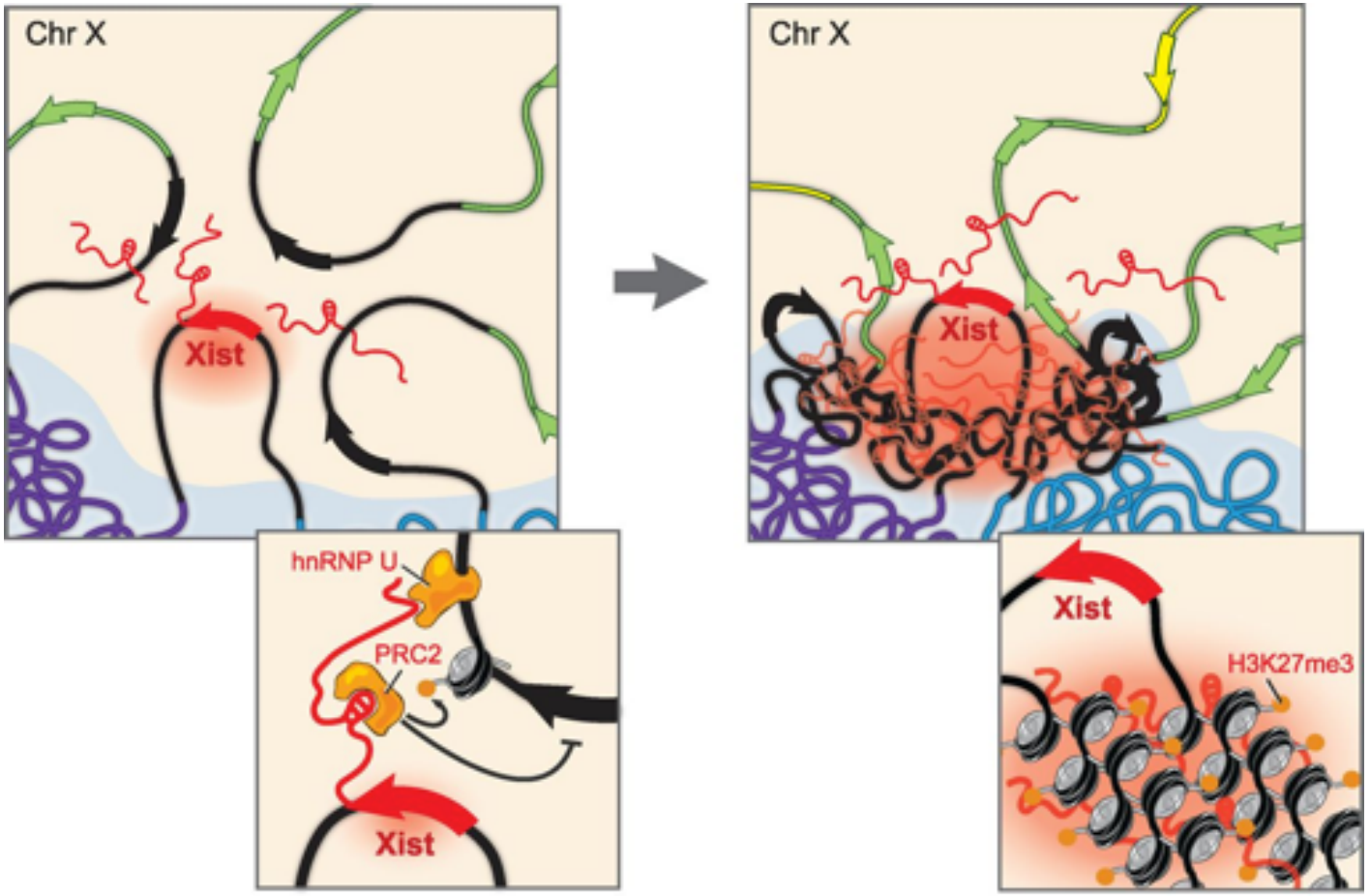
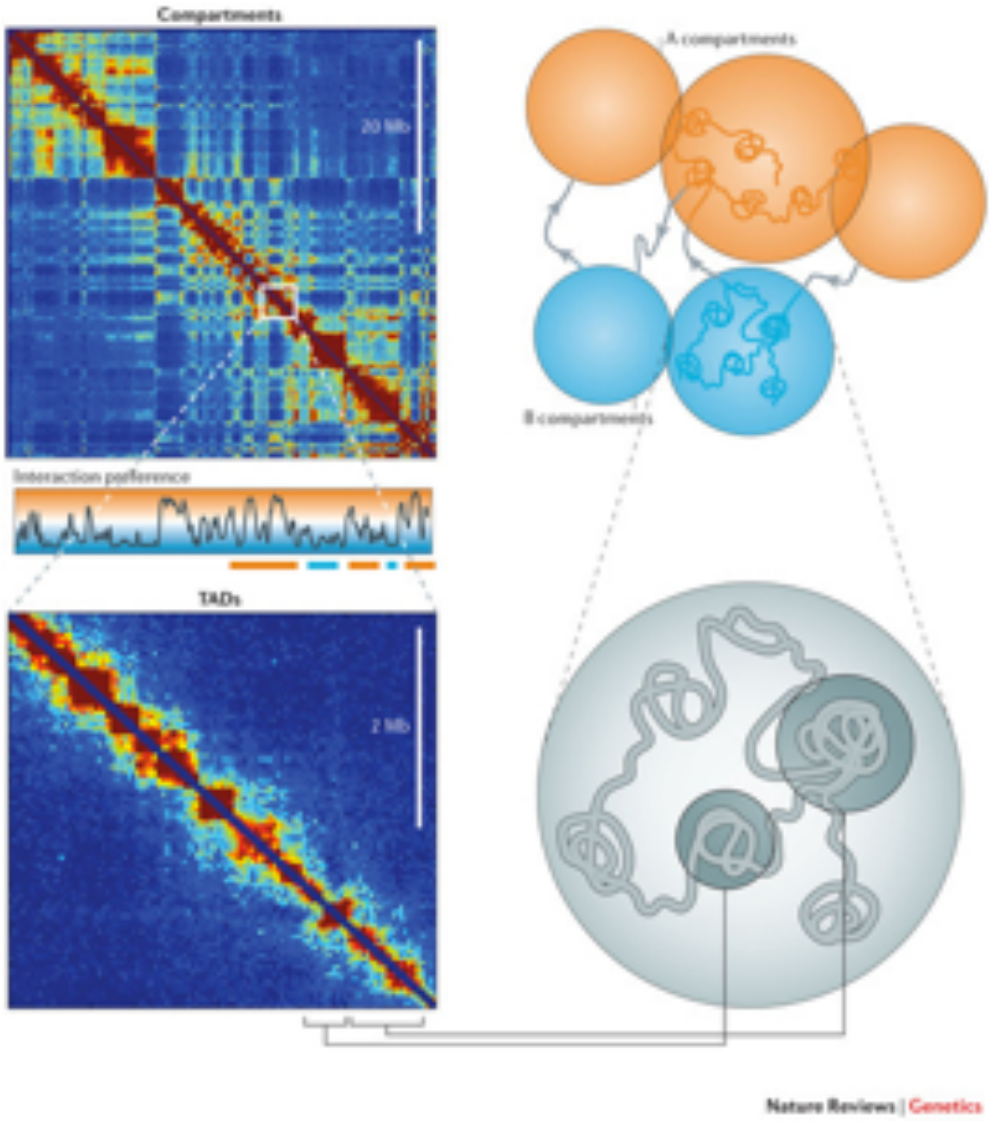


Fig 6. A model for how Xist exploits and alters three-dimensional genome architecture to spread across the X chromosome.

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

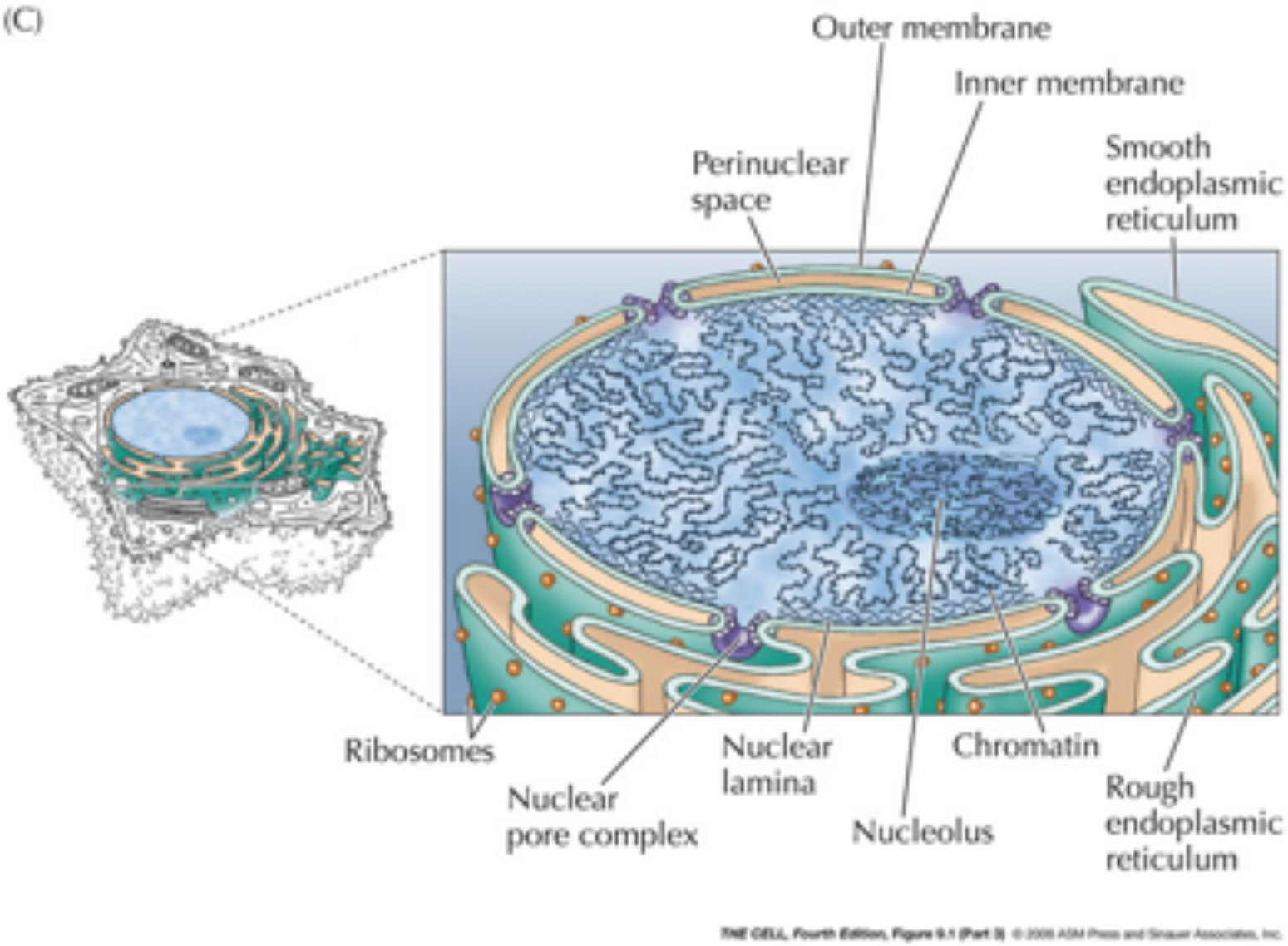
- alternate along chromosomes and have a characteristic size of ~5 Mb each.
- A compartments (orange) preferentially interact with other A compartments; B compartments (blue) associate with other B compartments.
- A compartments are largely euchromatic, transcriptionally active regions.

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

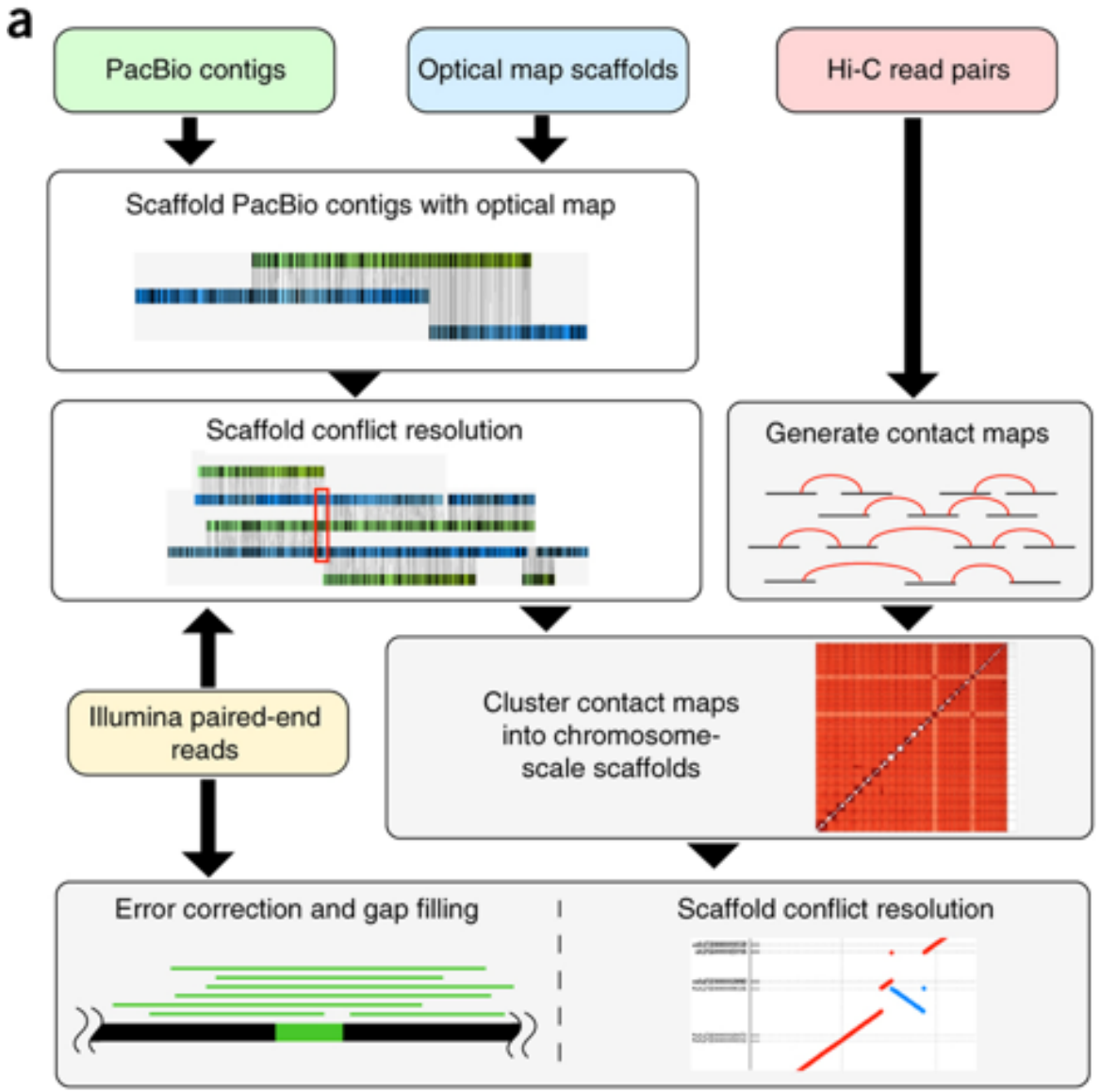
Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data
Dekker et al. (2013) *Nature Reviews Genetics 14, 390–403*

“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*

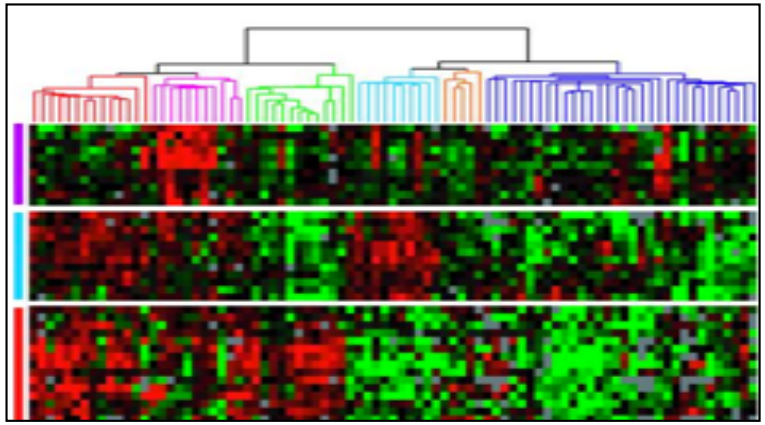
Scaffolding with Hi-C



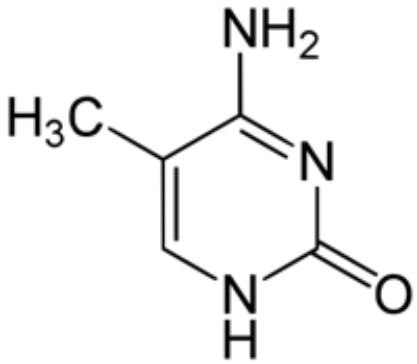
Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome
Bickhart et al (2017) Nature Genetics (2017) doi:10.1038/ng.3802

Putting it all together!

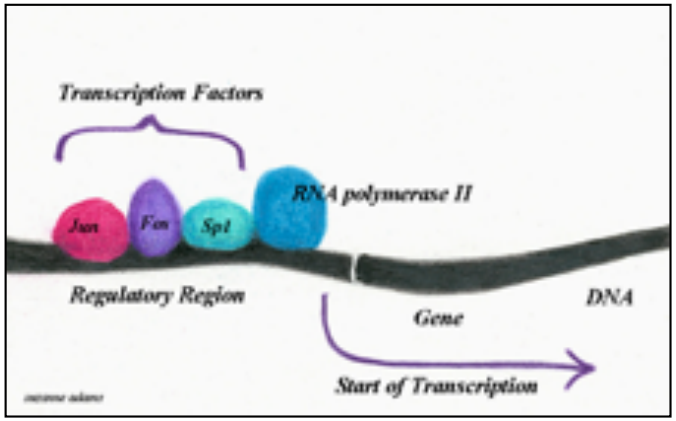
RNA-seq



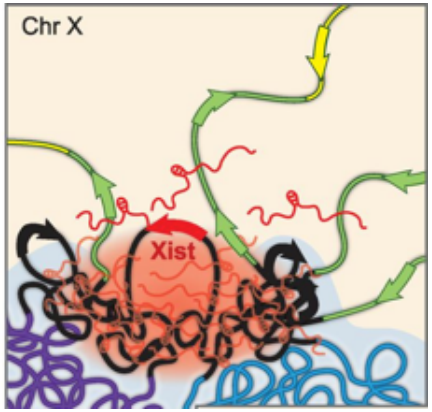
Methyl-seq



ChIP-seq



Hi-C



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.