

Functional Genomics 3: Gene Regulation

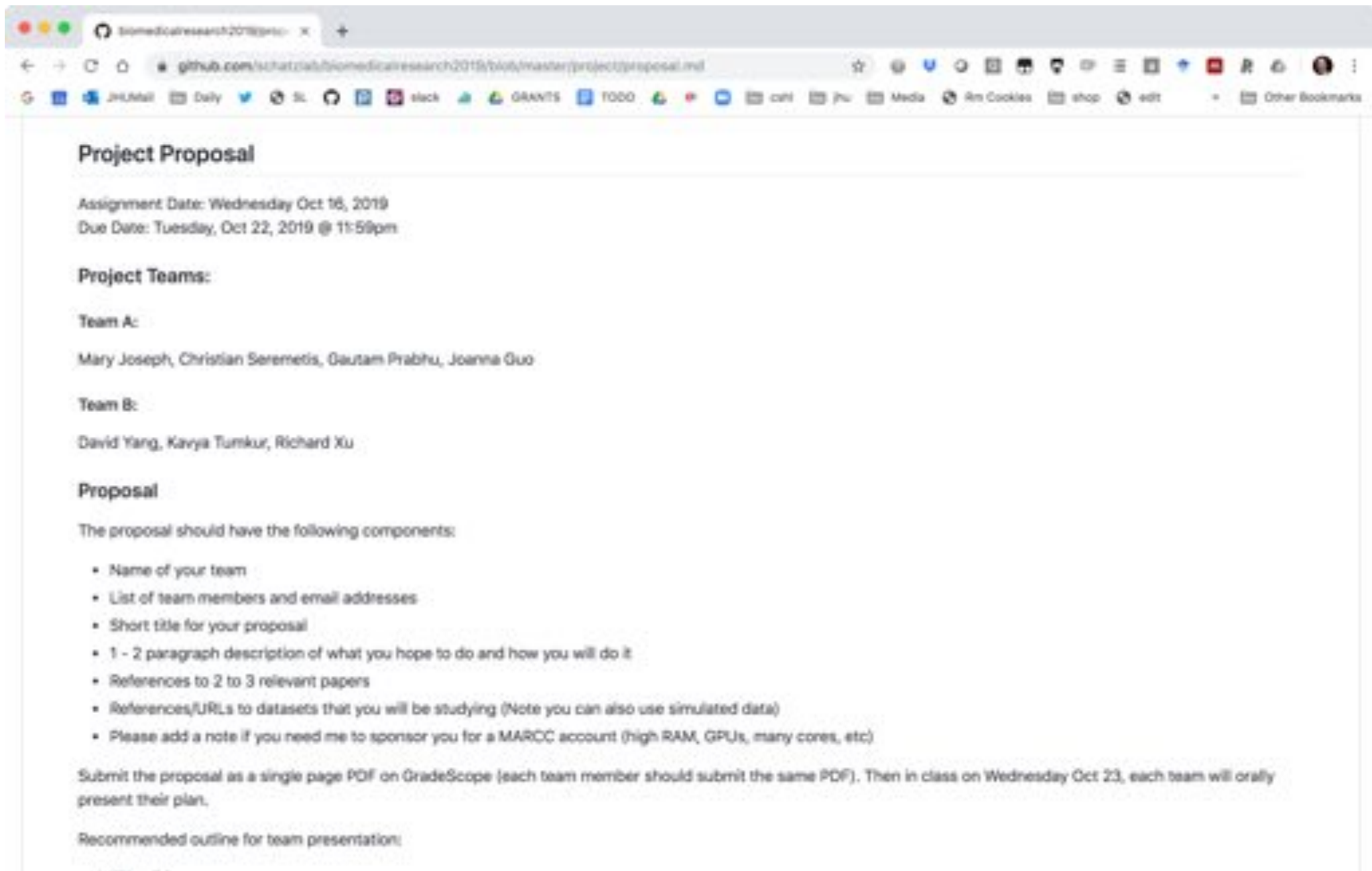
Michael Schatz

Oct 22, 2019

Lecture 15: Computational Biomedical Research



Project Proposal



The screenshot shows a web browser window with a single tab titled 'biomedicalresearch2019@princ...'. The address bar displays the URL 'github.com/schatzlab/biomedicalresearch2019/blob/master/project/proposal.md'. The browser's toolbar includes various icons for navigation, search, and social media. Below the browser window, the content of the GitHub page is visible. It features a heading 'Project Proposal' followed by assignment dates, project teams, and a list of proposal requirements.

Project Proposal

Assignment Date: Wednesday Oct 16, 2019
Due Date: Tuesday, Oct 22, 2019 @ 11:59pm

Project Teams:

Team A:
Mary Joseph, Christian Seremetis, Gautam Prabhu, Joanna Guo

Team B:
David Yang, Kavya Tumkur, Richard Xu

Proposal

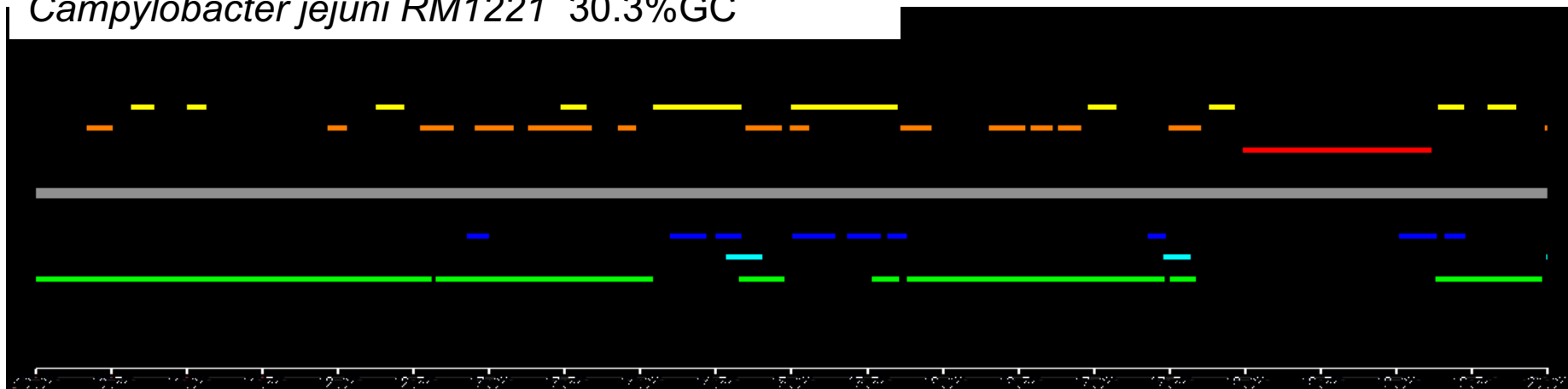
The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 - 2 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)

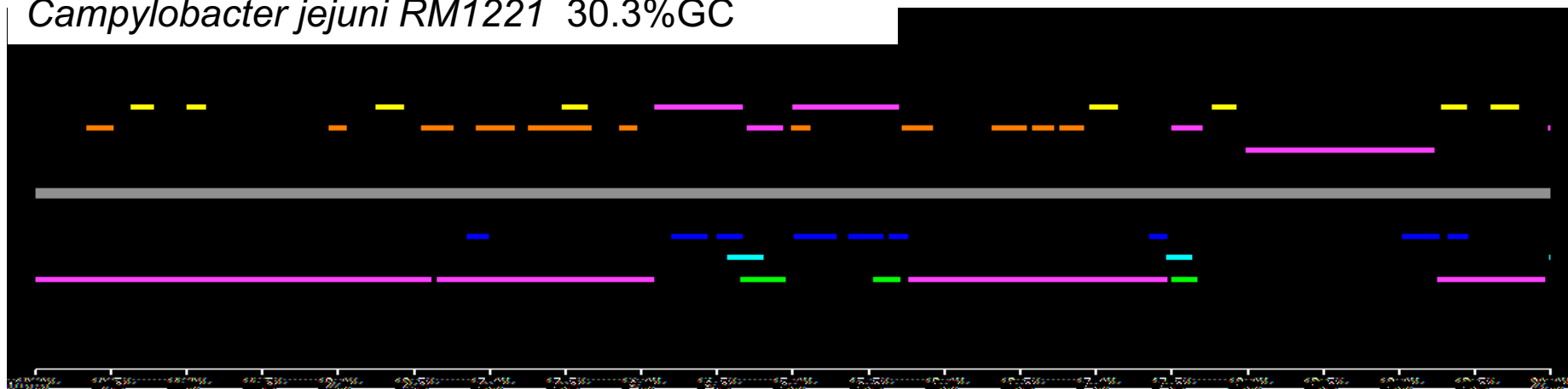
Submit the proposal as a single page PDF on GradeScope (each team member should submit the same PDF). Then in class on Wednesday Oct 23, each team will orally present their plan.

Recommended outline for team presentation:

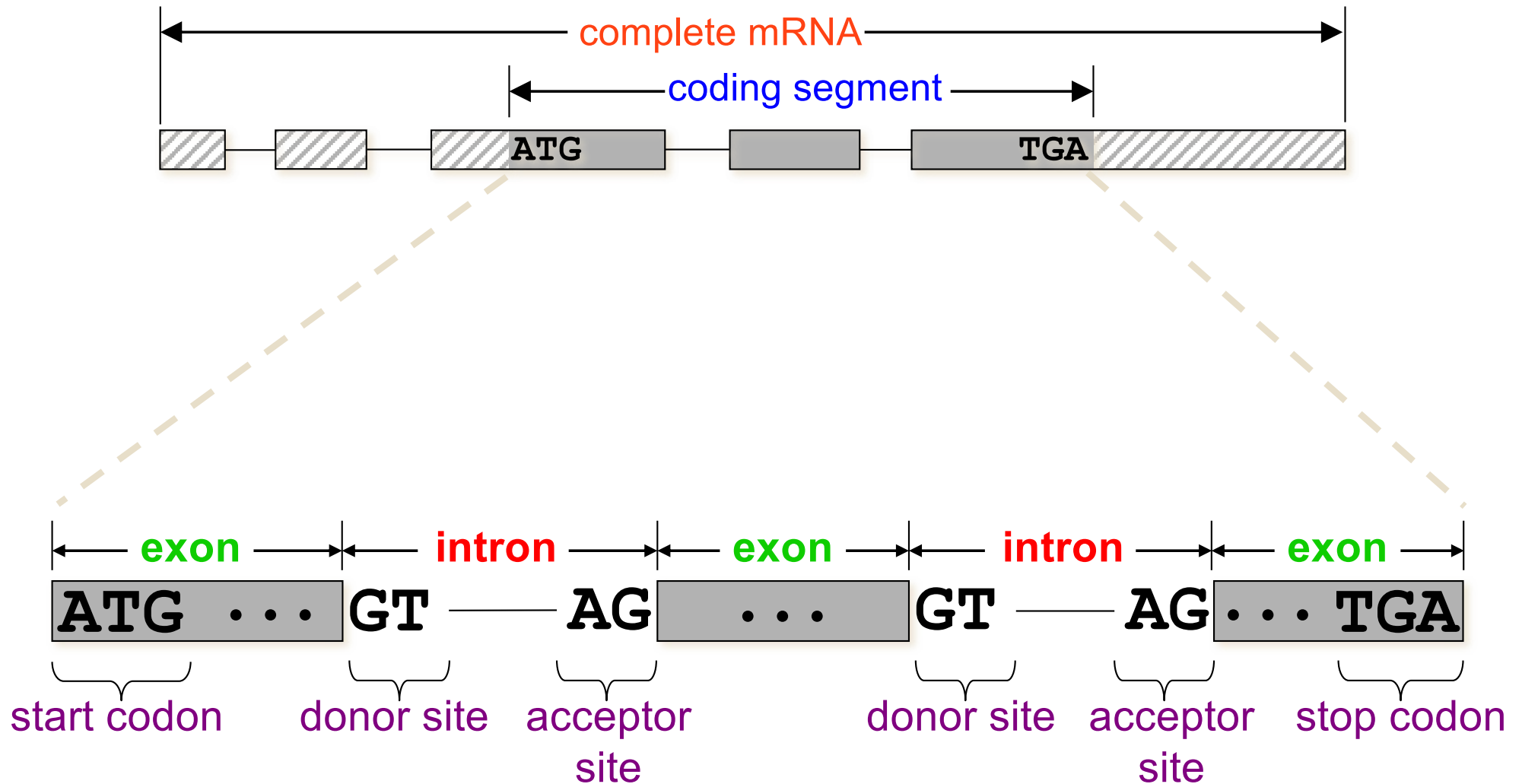
Campylobacter jejuni RM1221 30.3%GC



Campylobacter jejuni RM1221 30.3%GC

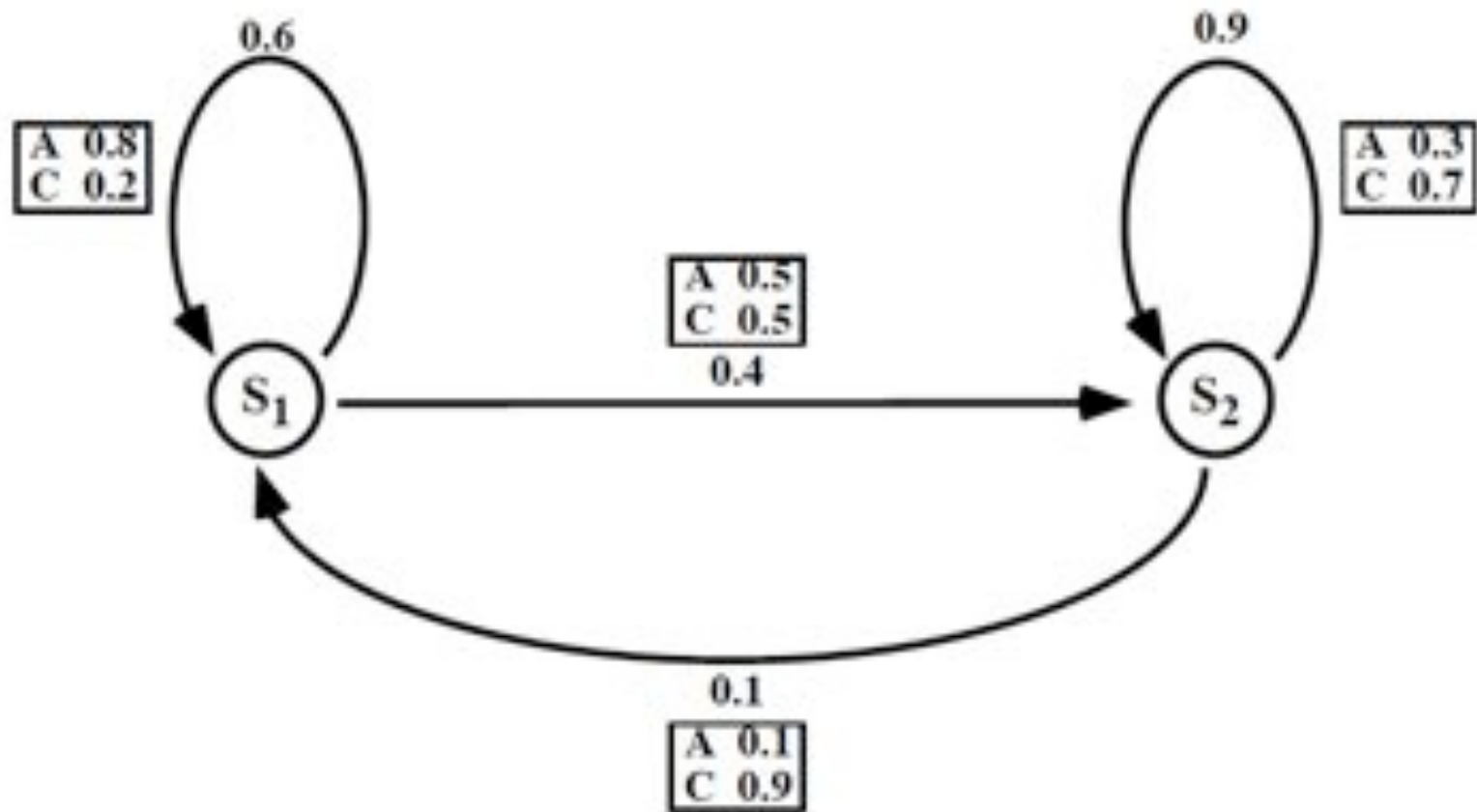


Eukaryotic Gene Syntax



Regions of the gene outside of the CDS are called **UTR**'s (*untranslated regions*), and are mostly ignored by gene finders, though they are important for regulatory functions.

Our sample HMM



Let S_1 be initial state, S_2 be final state

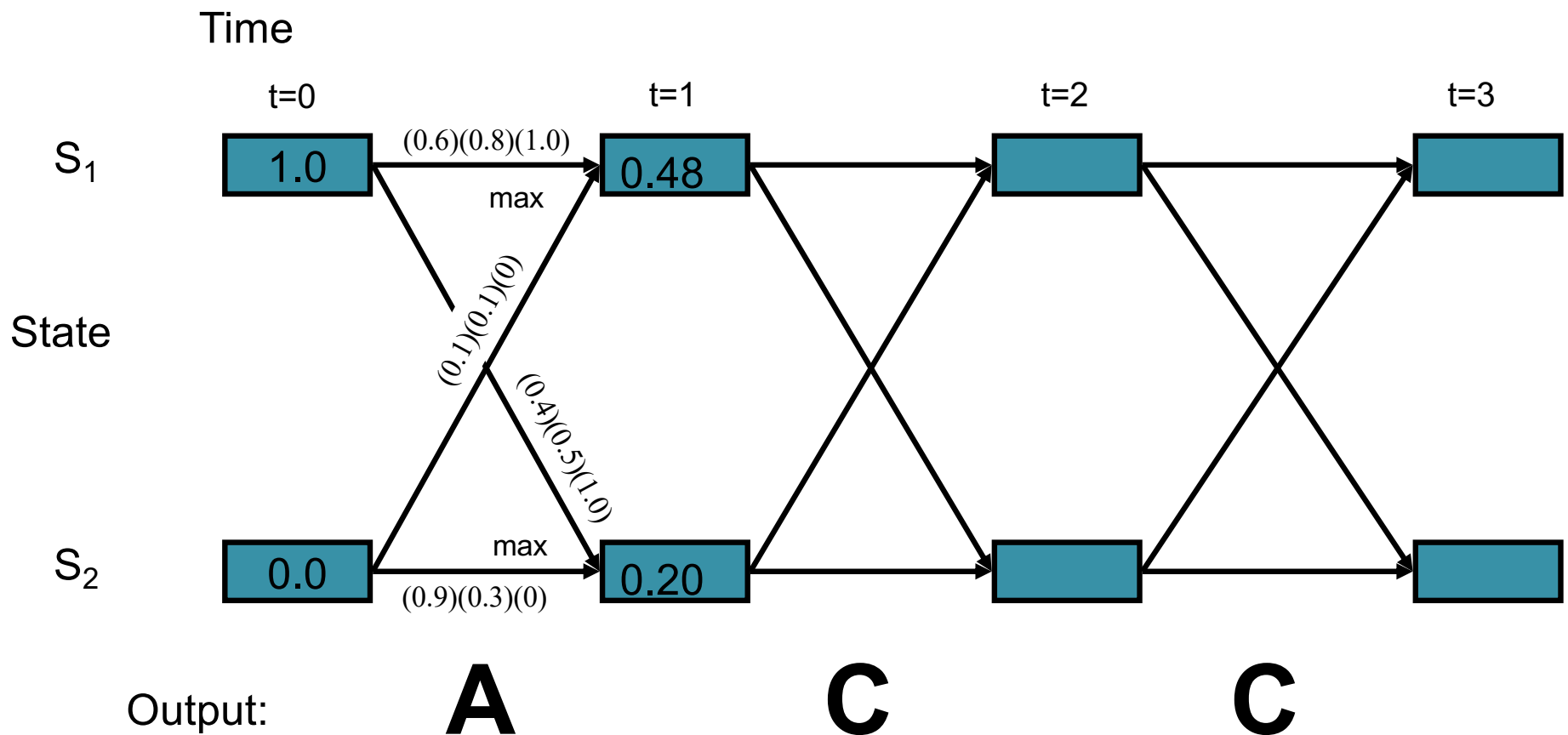
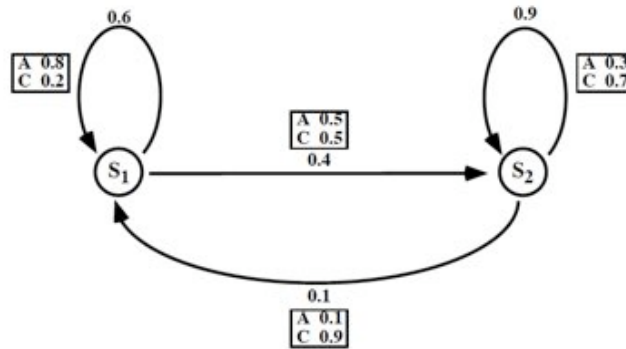
Solving the Decoding Problem: The Viterbi algorithm

- To solve the decoding problem (find the most likely sequence of states), we evaluate the Viterbi algorithm

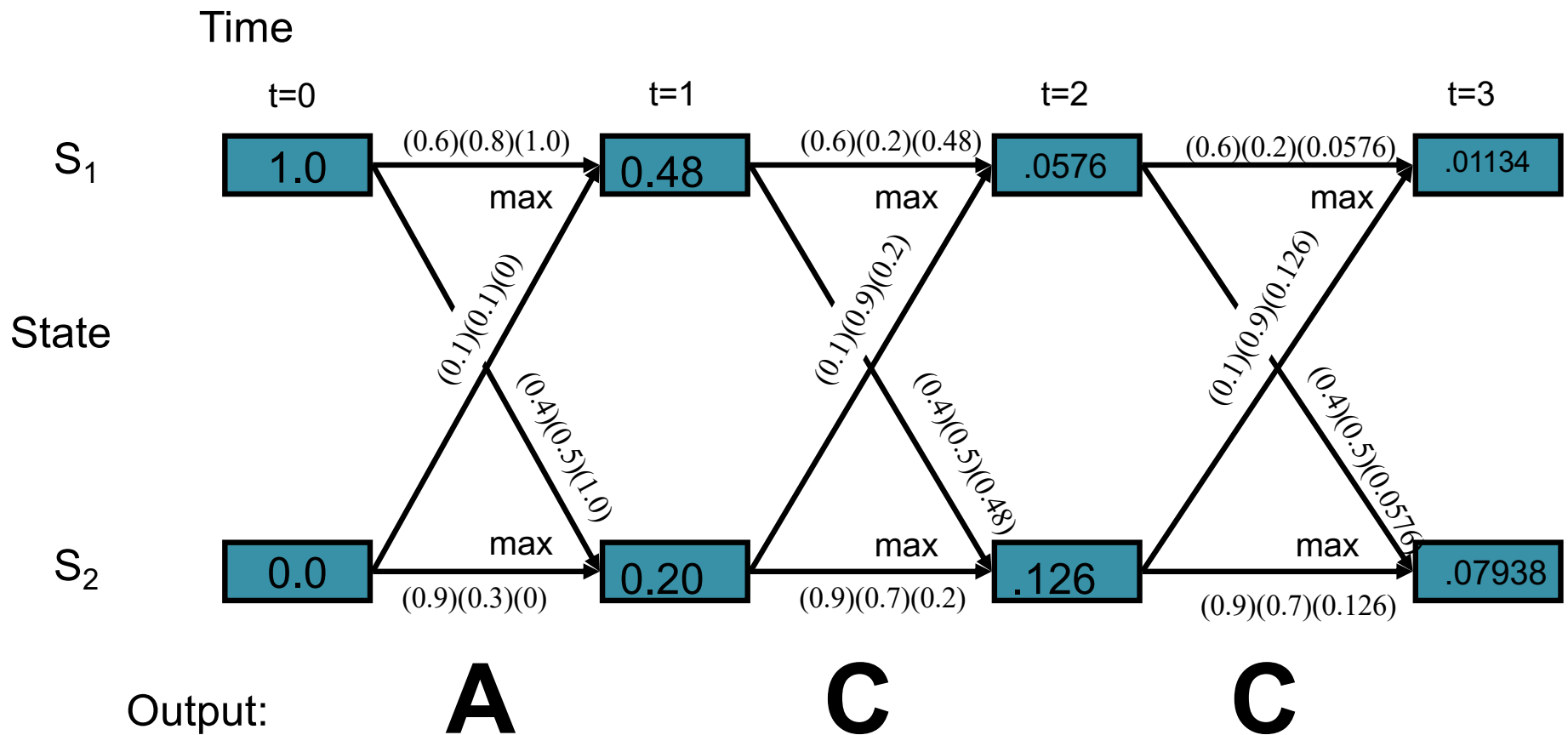
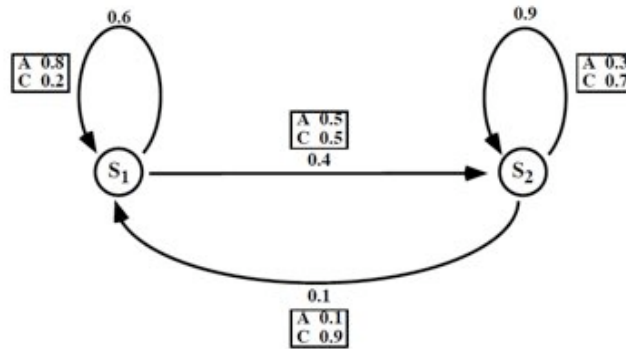
$$V_i(t) = \begin{cases} 0 & : t = 0 \wedge i \neq S_I \\ 1 & : t = 0 \wedge i = S_I \\ \max_j V_j(t-1) a_{ji} b_{ji}(y) & : t > 0 \end{cases}$$

Where $V_i(t)$ is the probability that the HMM is in state i after generating the sequence y_1, y_2, \dots, y_t following the *most probable path* in the HMM

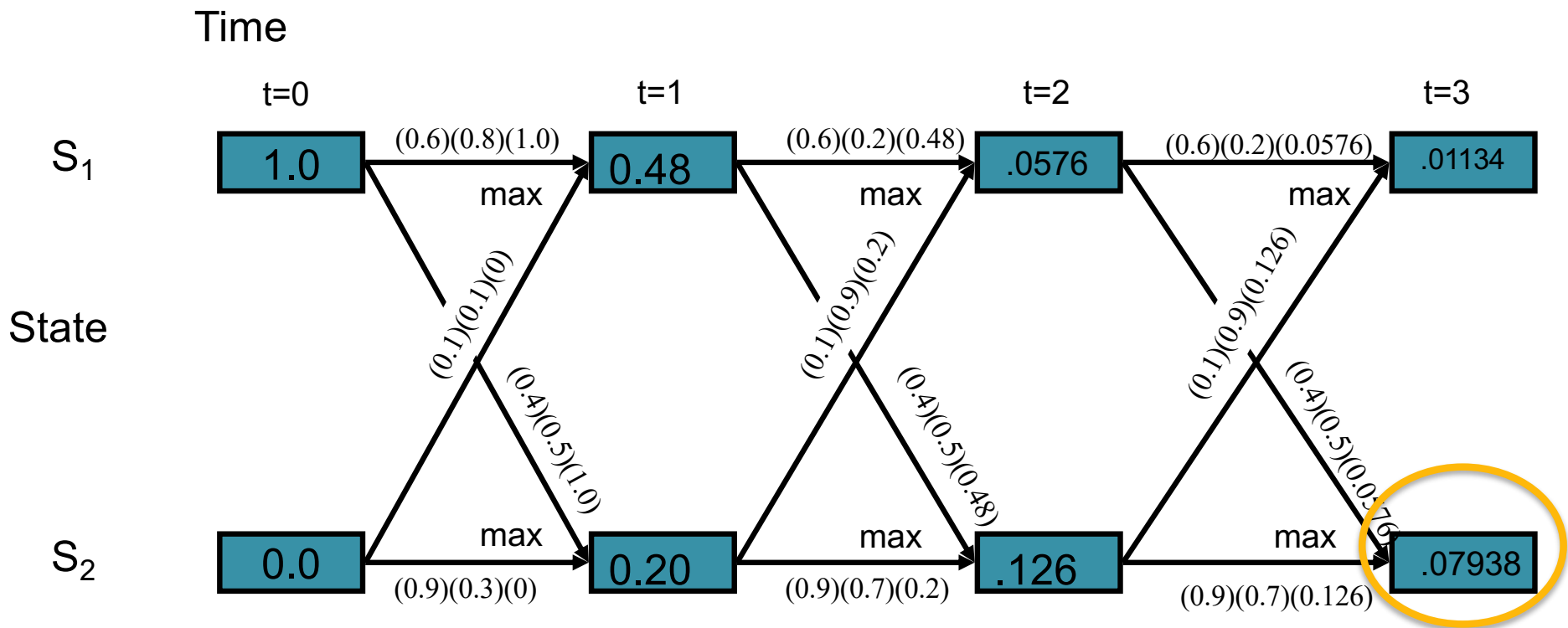
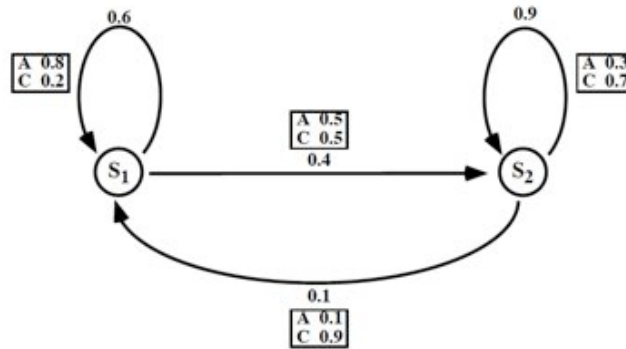
A trellis for the Viterbi Algorithm



A trellis for the Viterbi Algorithm

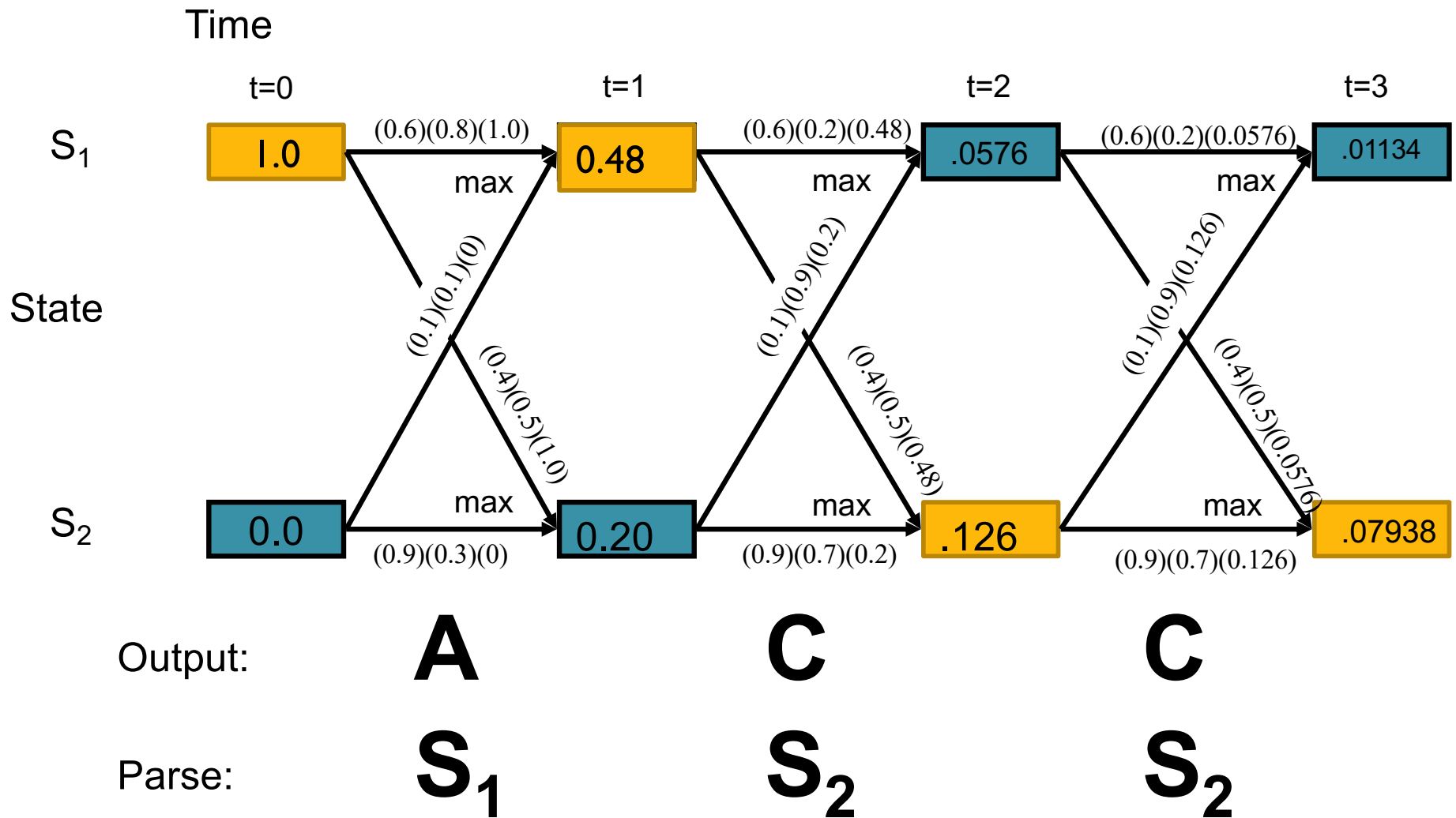


A trellis for the Viterbi Algorithm



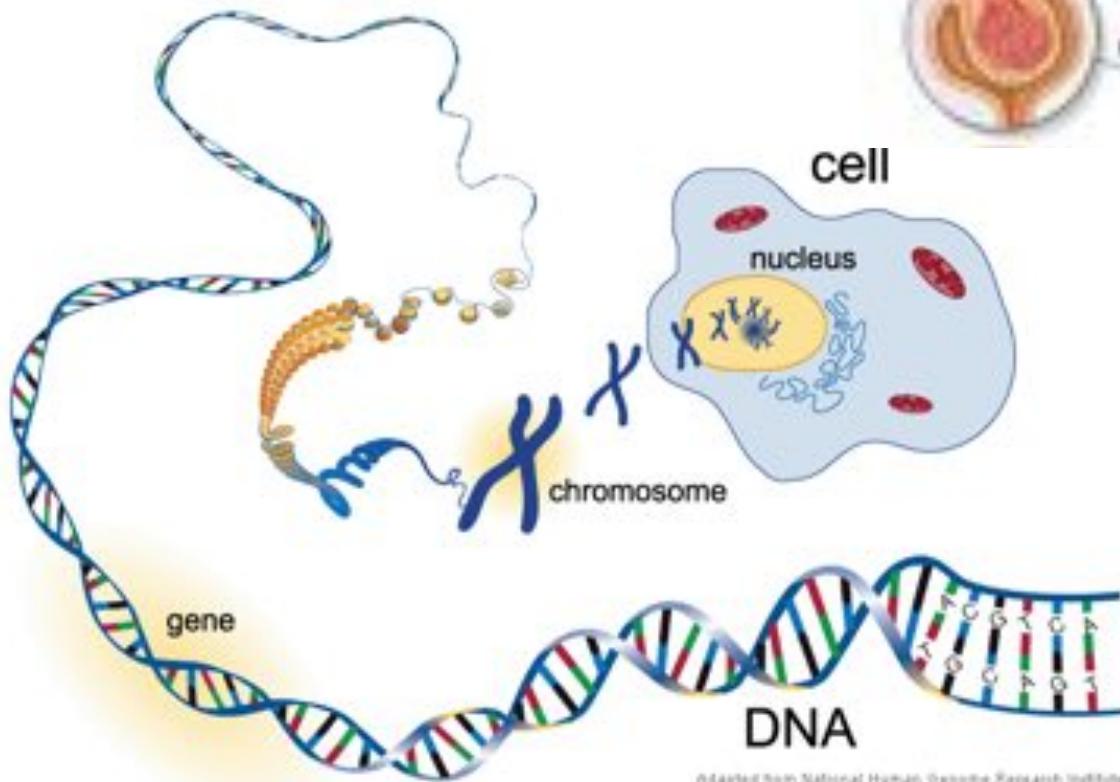
S2 is final state → the most probable sequence of states has a 7.9% probability

A trellis for the Viterbi Algorithm



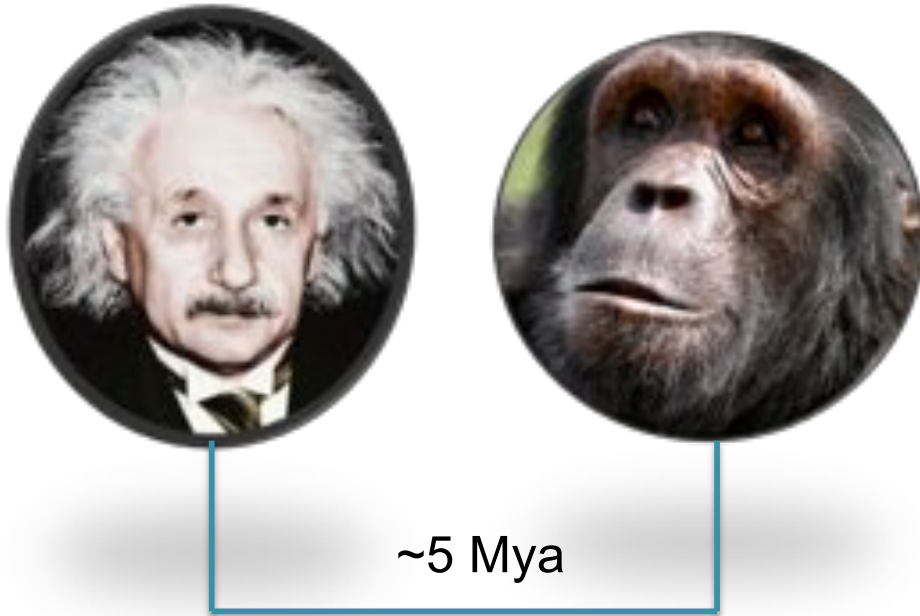
Why Genes?

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

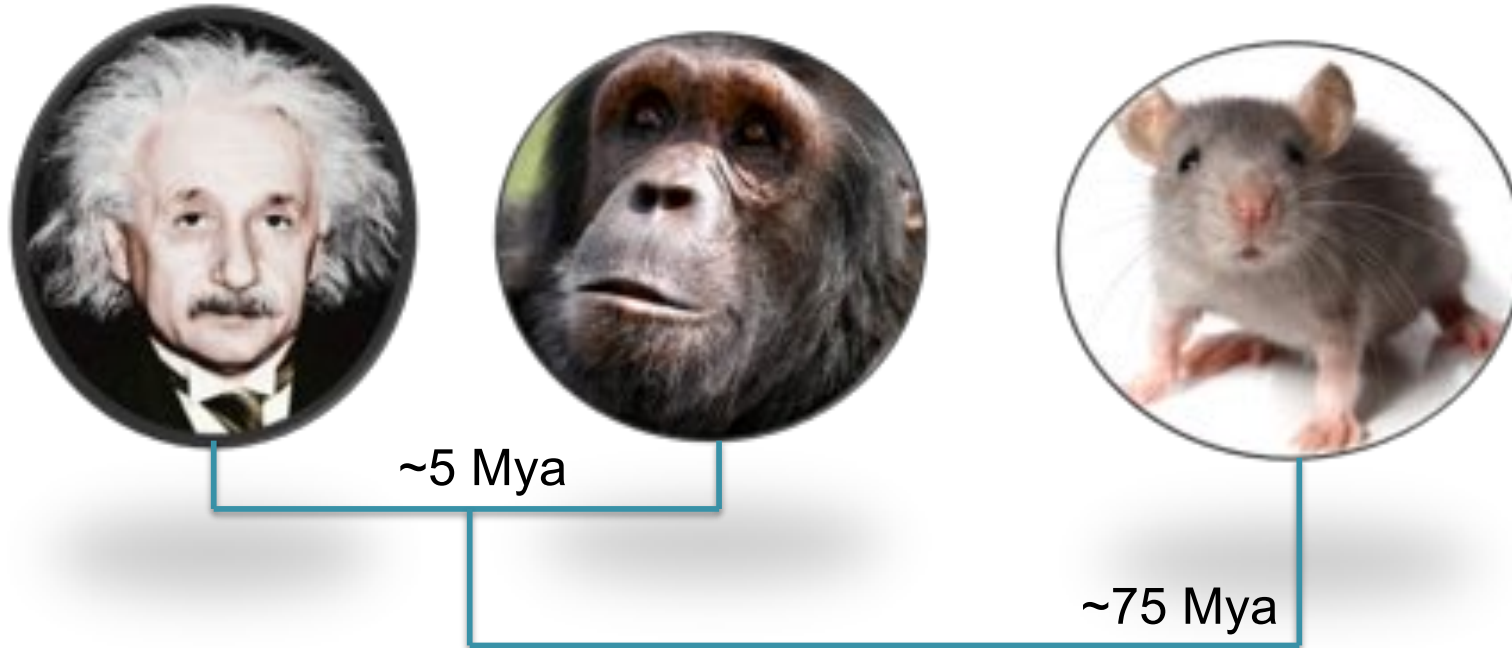
Human Evolution



- Humans and chimpanzees shared a common ancestor ~5-7 million years ago (Mya)
- Single-nucleotide substitutions occur at a mean rate of 1.23% but ~4% overall rate of mutation: comprising ~35 million single nucleotide differences and ~90 Mb of insertions and deletions
- Orthologous proteins in human and chimpanzee are extremely similar, with ~29% being identical and the typical orthologue differing by only two amino acids, one per lineage

Initial sequence of the chimpanzee genome and comparison with the human genome
(2005) *Nature* 437, 69-87 doi:10.1038/nature04072

Human Evolution



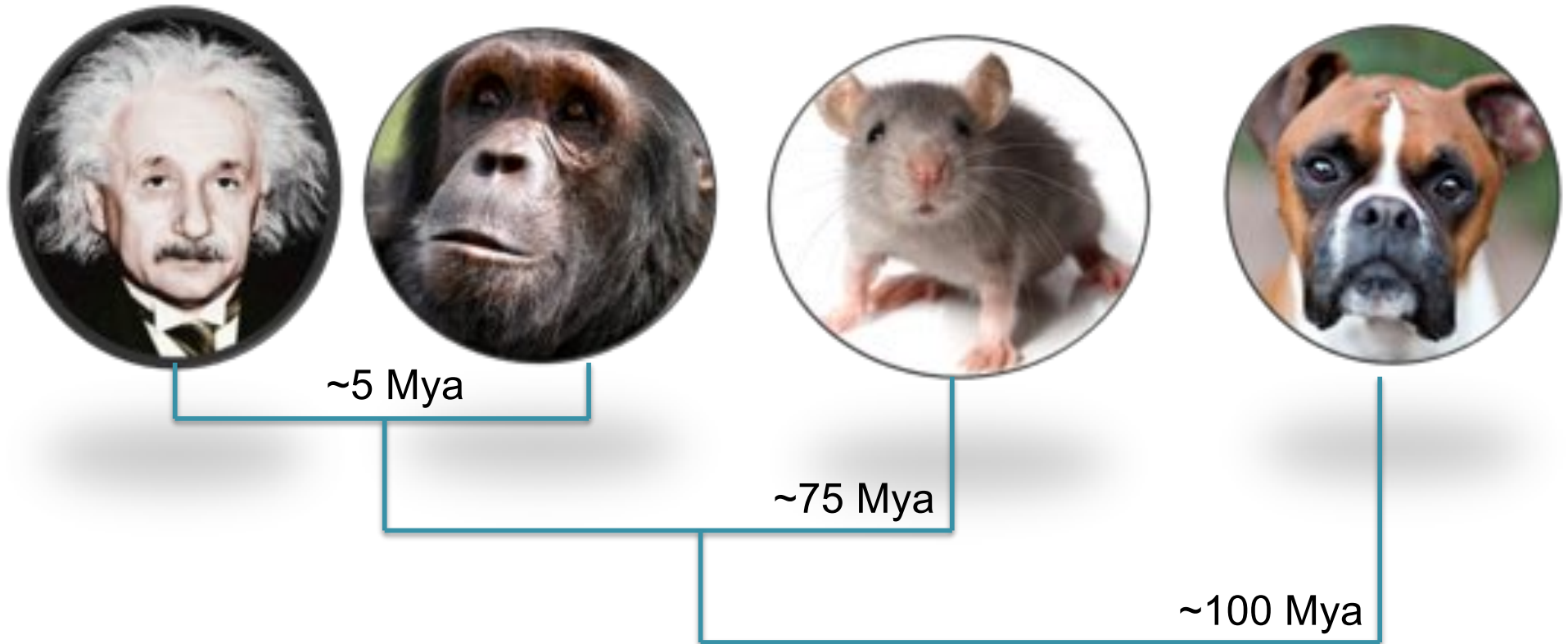
“In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by ***nearly one substitution for every two nucleotides***”

“The mouse and human genomes each seem to contain about 30,000 protein-coding genes. These refined estimates have been derived from both new evidence-based analyses that produce larger and more complete sets of gene predictions, and new de novo gene predictions that do not rely on previous evidence of transcription or homology. The proportion of mouse genes with a single identifiable orthologue in the human genome seems to be approximately 80%. ***The proportion of mouse genes without any homologue currently detectable in the human genome (and vice versa) seems to be less than 1%.***”

Initial sequencing and comparative analysis of the mouse genome

Chinwalla et al (2002) *Nature*. 420, 520-562 doi:10.1038/nature01262

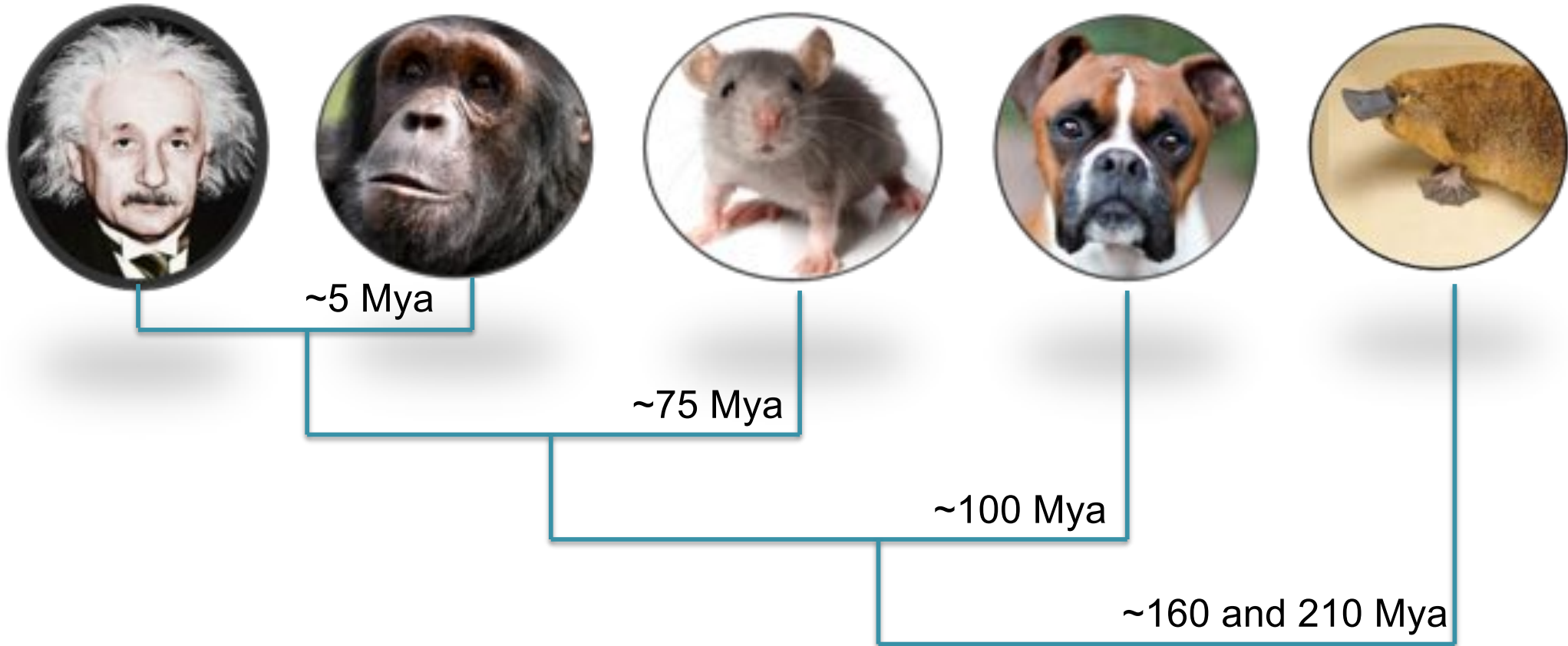
Human Evolution



“We generated gene predictions for the dog genome using an evidence-based method (see Supplementary Information). The resulting collection contains **19,300 dog gene predictions, with nearly all being clear homologues of known human genes**. The dog gene count is substantially lower than the ~22,000-gene models in the current human gene catalogue (Ensembl build 26). For many predicted human genes, we find no convincing evidence of a corresponding dog gene. Much of the excess in the human gene count is attributable **to spurious gene predictions in the human genome**”

Genome sequence, comparative analysis and haplotype structure of the domestic dog
Lindblad-Toh et al (2005) *Nature*. 438, 803-819 doi:10.1038/nature04338

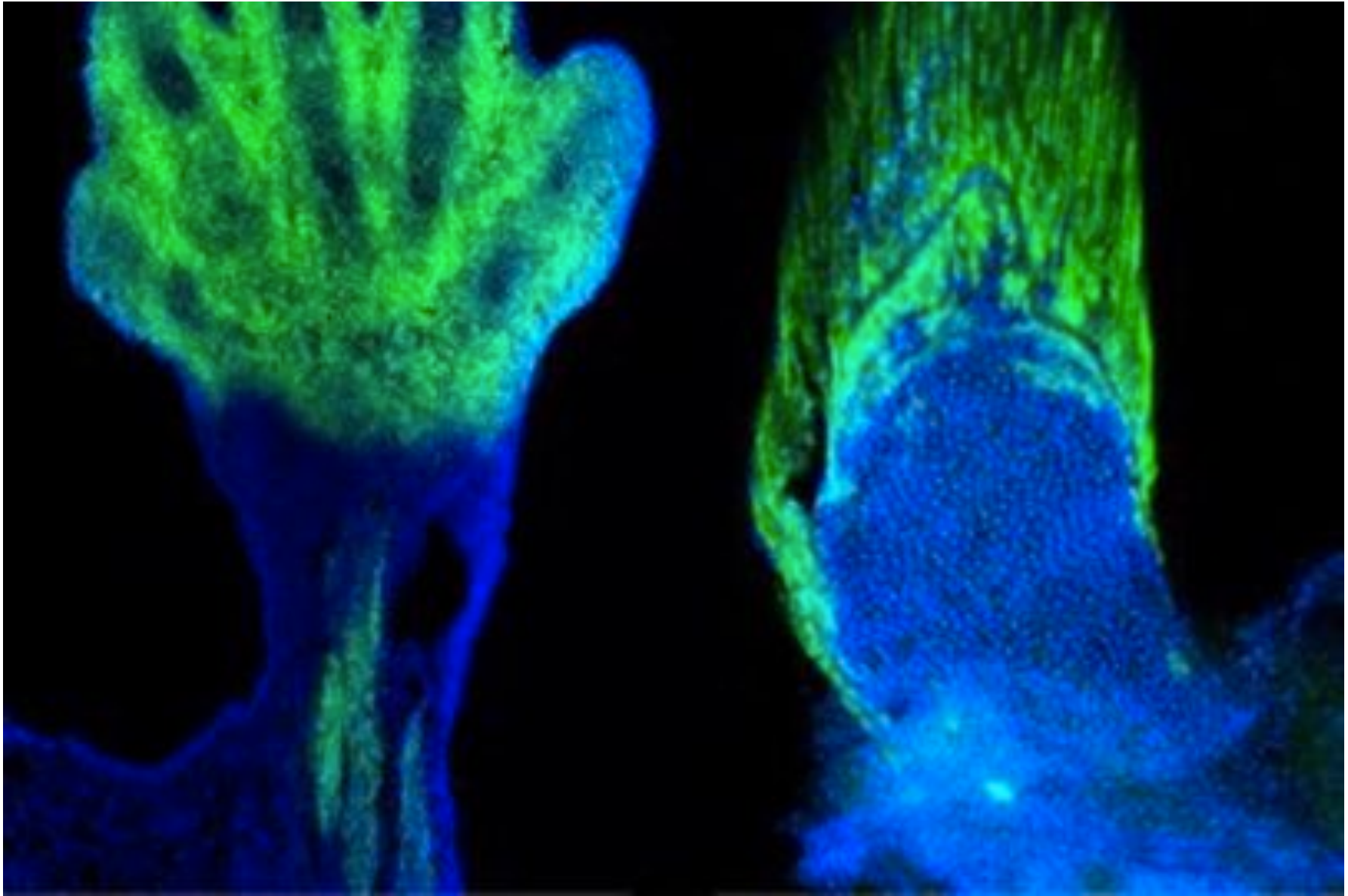
Human Evolution



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

Human Evolution



Digits and fin rays share common developmental histories

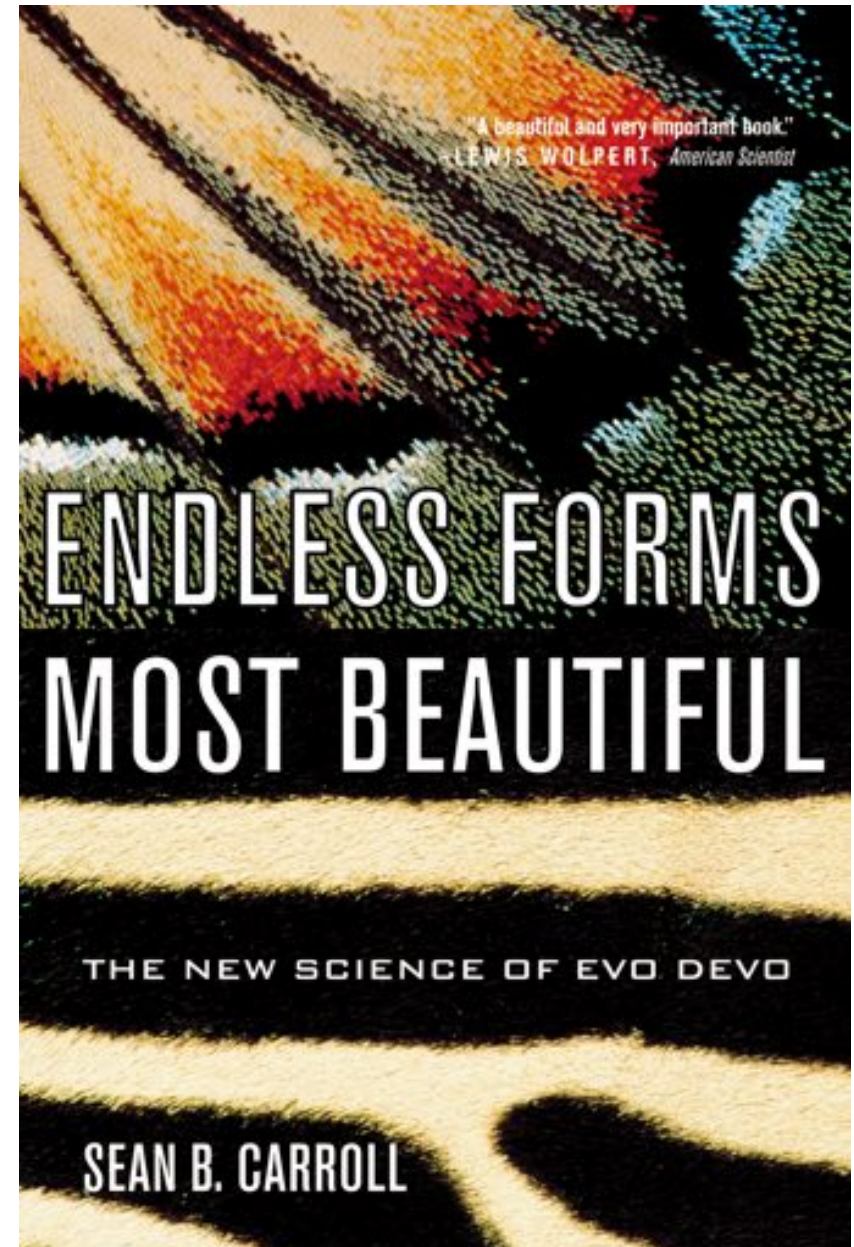
Nakamura et al (2016) *Nature*. 537, 225–228. doi:10.1038/nature19322

More Information

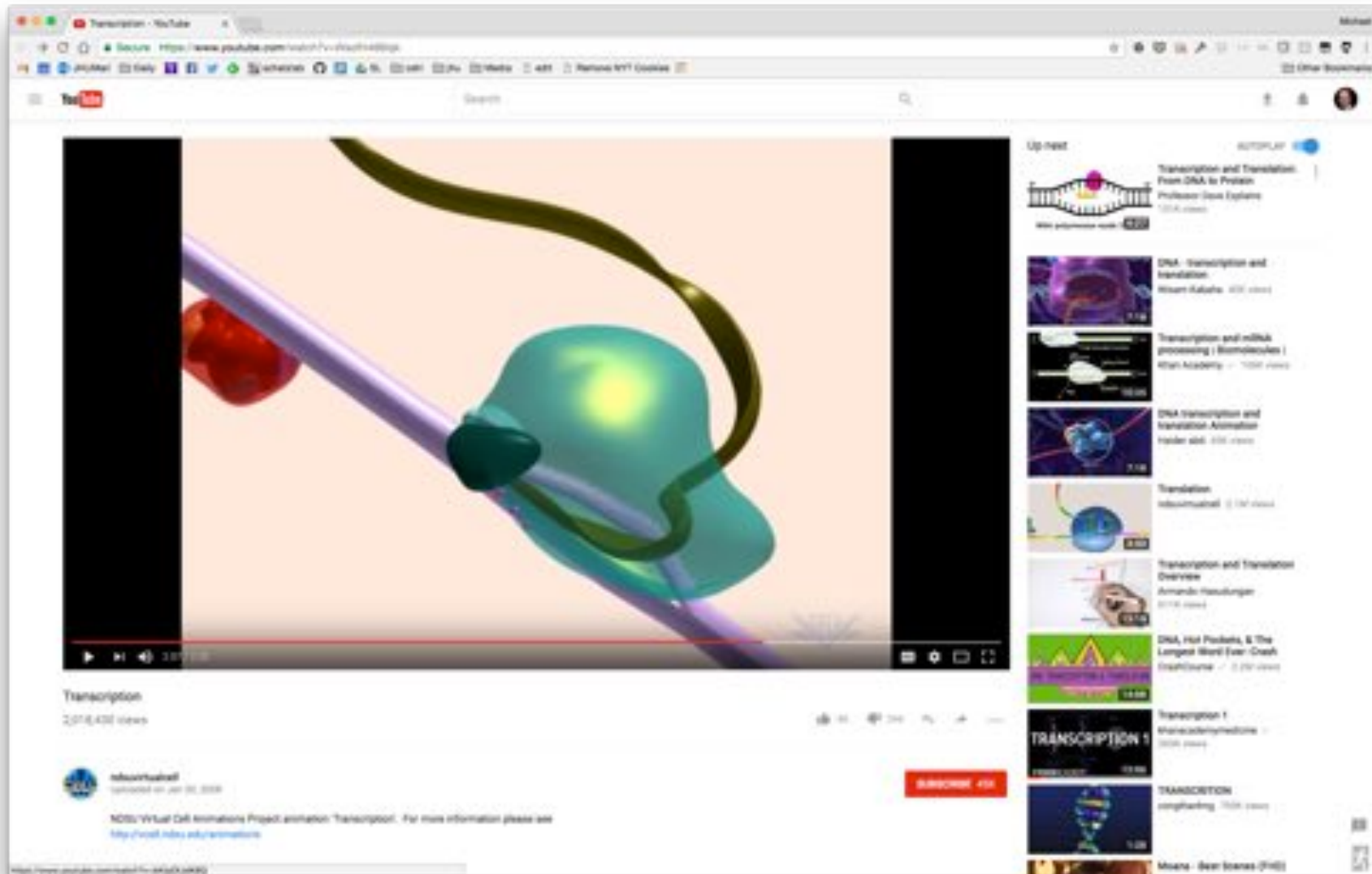


*“Anything found to be true of
E. coli must also be true of
elephants”*

-Jacques Monod

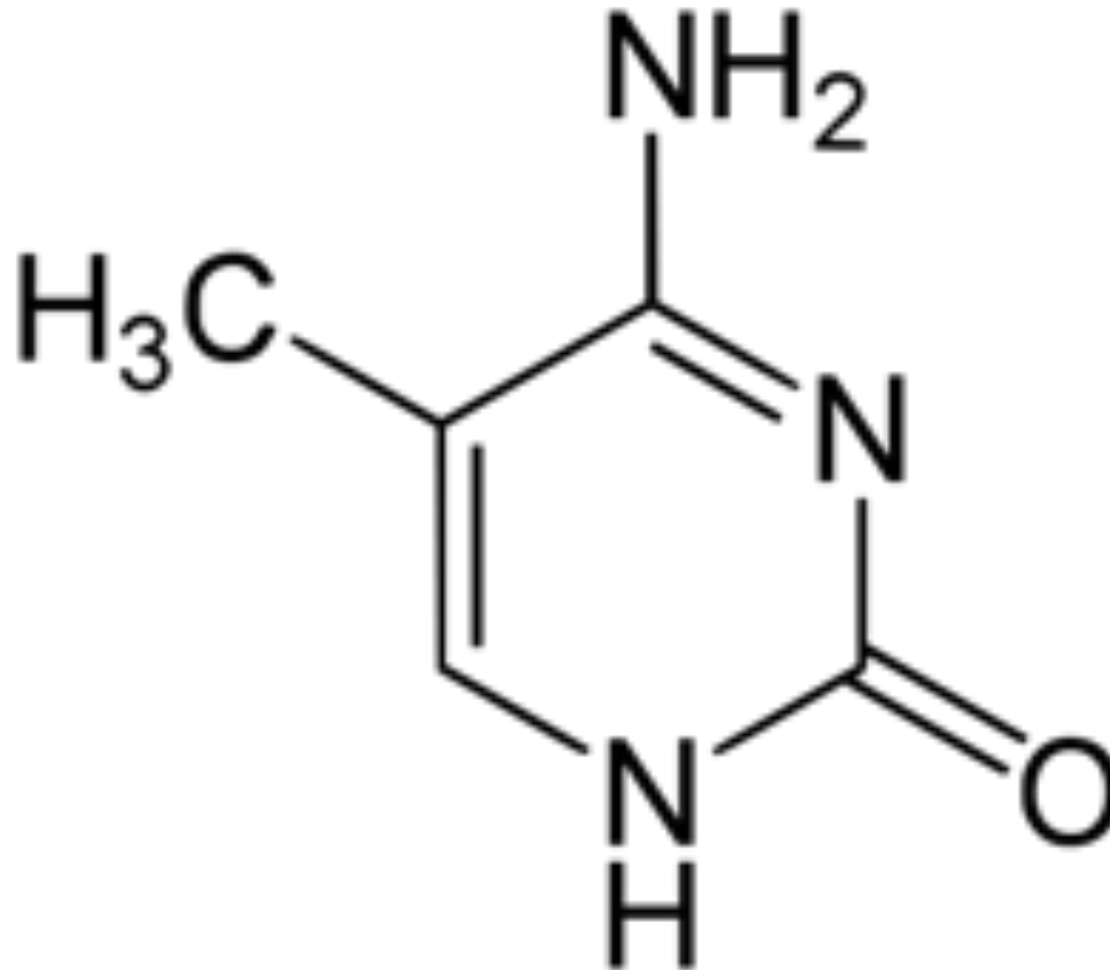


Transcription



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

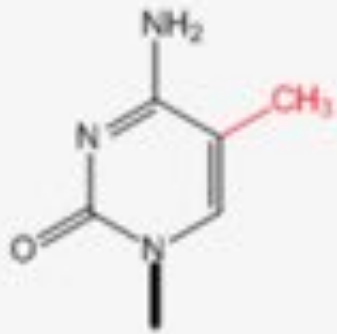
Methyl-seq



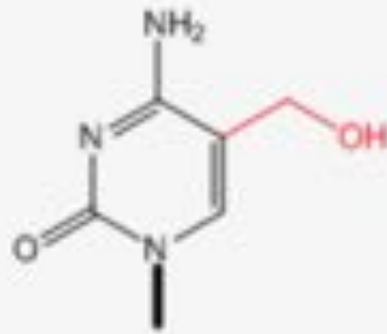
Finding the fifth base: Genome-wide sequencing of cytosine methylation

Lister and Ecker (2009) *Genome Research*. 19: 959-966

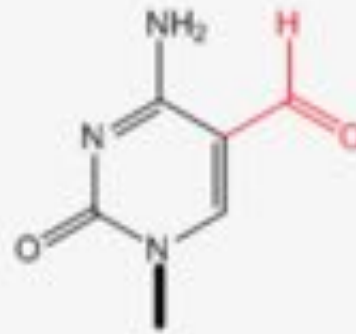
Epigenetic Modifications to DNA



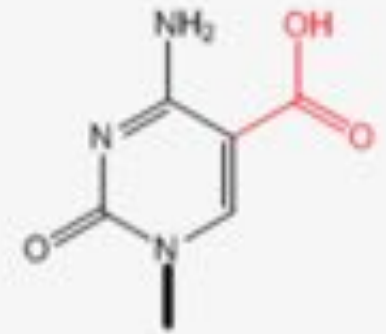
5-mC



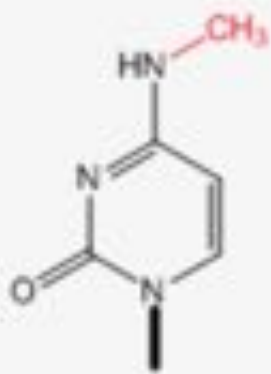
5-hmC



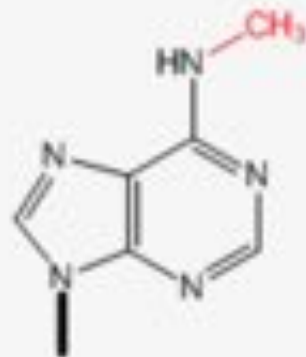
5-fC



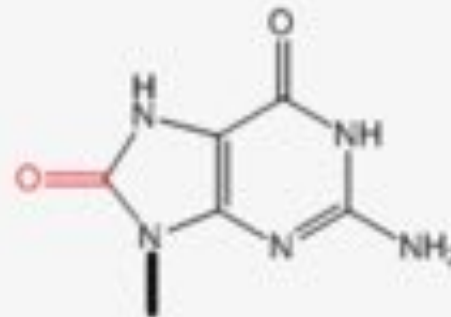
5-caC



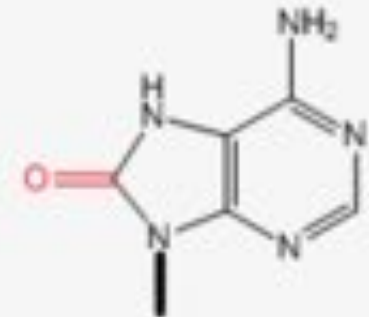
4-mC



6-mA



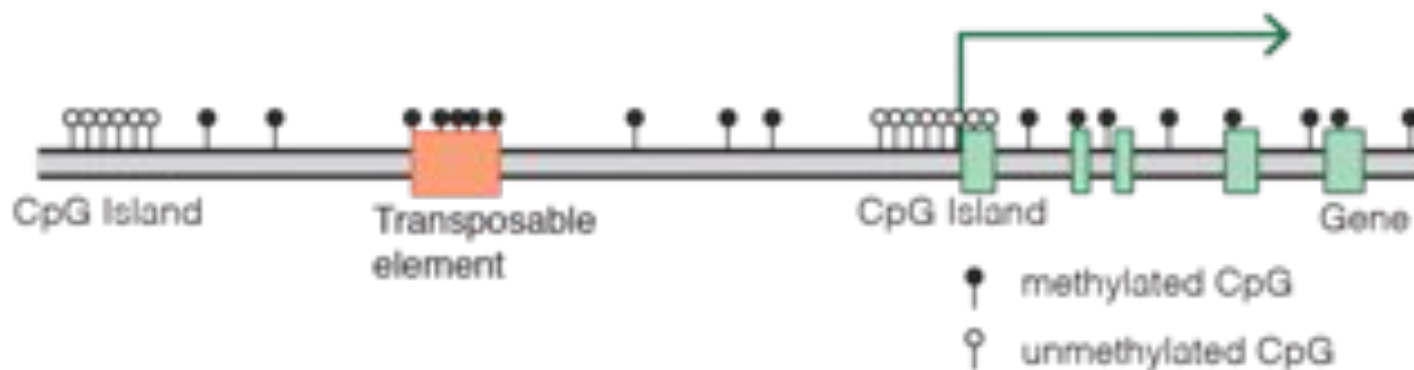
8-oxoG



8-oxoA

Methylation of CpG Islands

Typical mammalian DNA methylation landscape



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.

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

Frank Lyko^{1,3}, Sylvain Foret^{2,3}, Robert Kucharski³, Stephan Wolf⁴, Cassandra Falckenhayn¹, 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





Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm

Ong-Abdullah, et al (2015) *Nature*. doi:[10.1038/nature15365](https://doi.org/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^{3,4} and changes in methylation⁵⁻⁸ 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^{8,12-20}. The results of these studies have varied; depending on the techniques and systems used, an increase¹²⁻¹⁹, decrease²⁰⁻²⁴, or no change²⁵⁻²⁹ 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

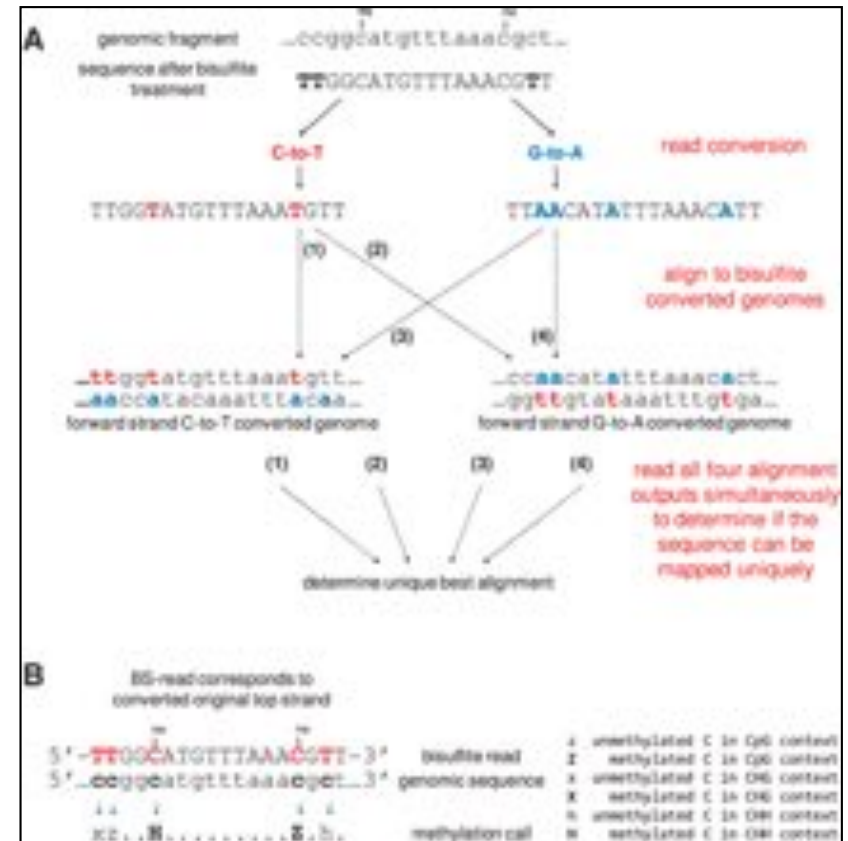
Table 1 Quantitation of methylation of specific genes in human cancers and adjacent analogous normal tissues

Patient	Carcinoma	Probe	Enzyme	% Hypomethylated fragments		
				N	C	M
1	Colon	HGH	<i>fHpa</i> II	<10	35	—
			<i>fHha</i> I	<10	39	—
		γ -Globin	<i>fHpa</i> II	<10	52	—
			<i>fHha</i> I	<10	39	—
		α -Globin	<i>fHpa</i> II	<10	<10	—
			<i>fHha</i> I	<10	<10	—
2	Colon	HGH	<i>fHpa</i> II	<10	76	—
			<i>fHha</i> I	<10	85	—
		γ -Globin	<i>fHpa</i> II	<10	58	—
			<i>fHha</i> I	<10	23	—
		α -Globin	<i>fHpa</i> II	<10	<10	—
			<i>fHha</i> I	<10	<10	—
3	Colon	HGH	<i>fHpa</i> II	<10	41	—
			<i>fHha</i> I	<10	38	—
		γ -Globin	<i>fHpa</i> II	<10	50	—

Bisulfite Conversion

Treating DNA with sodium bisulfite will convert unmethyated 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

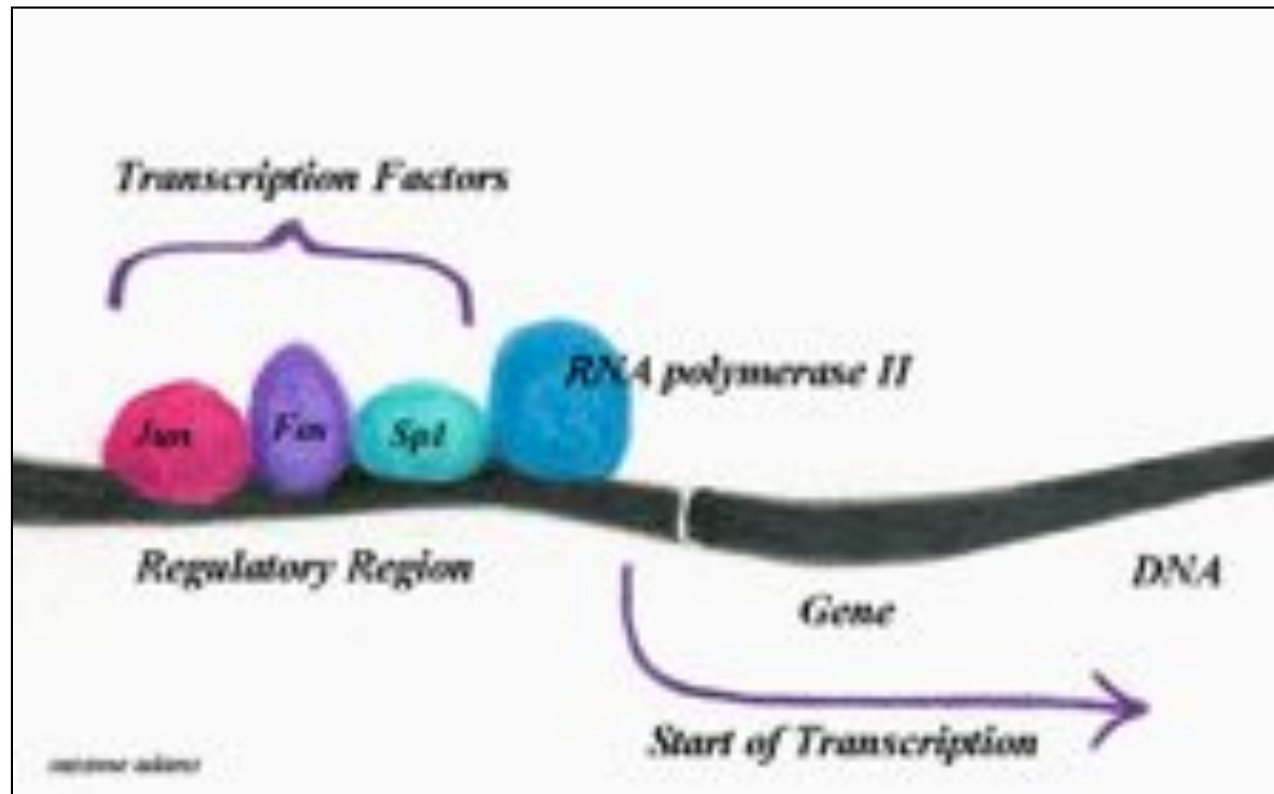
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

-
-
-



Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

ChIP-seq



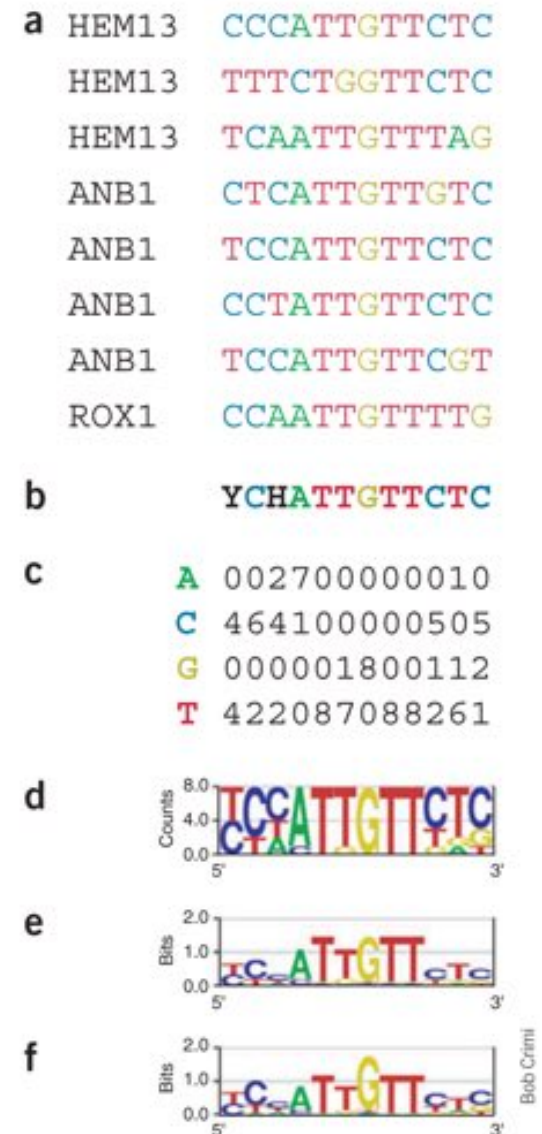
Genome-wide mapping of in vivo protein-DNA interactions.

Johnson et al (2007) *Science*. 316(5830):1497-502

Transcription Factors

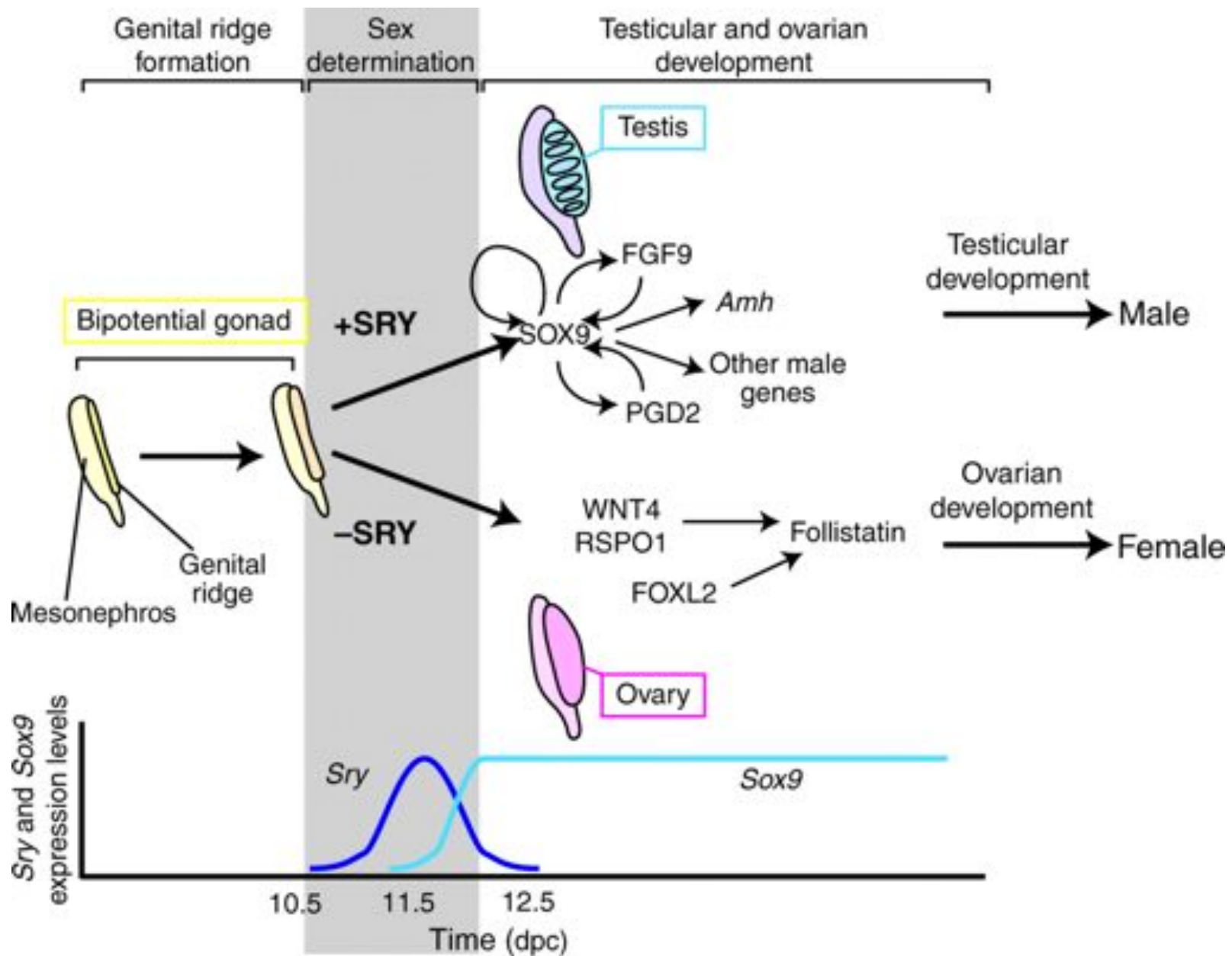
A transcription factor (or sequence-specific DNA-binding factor) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.

- 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*.



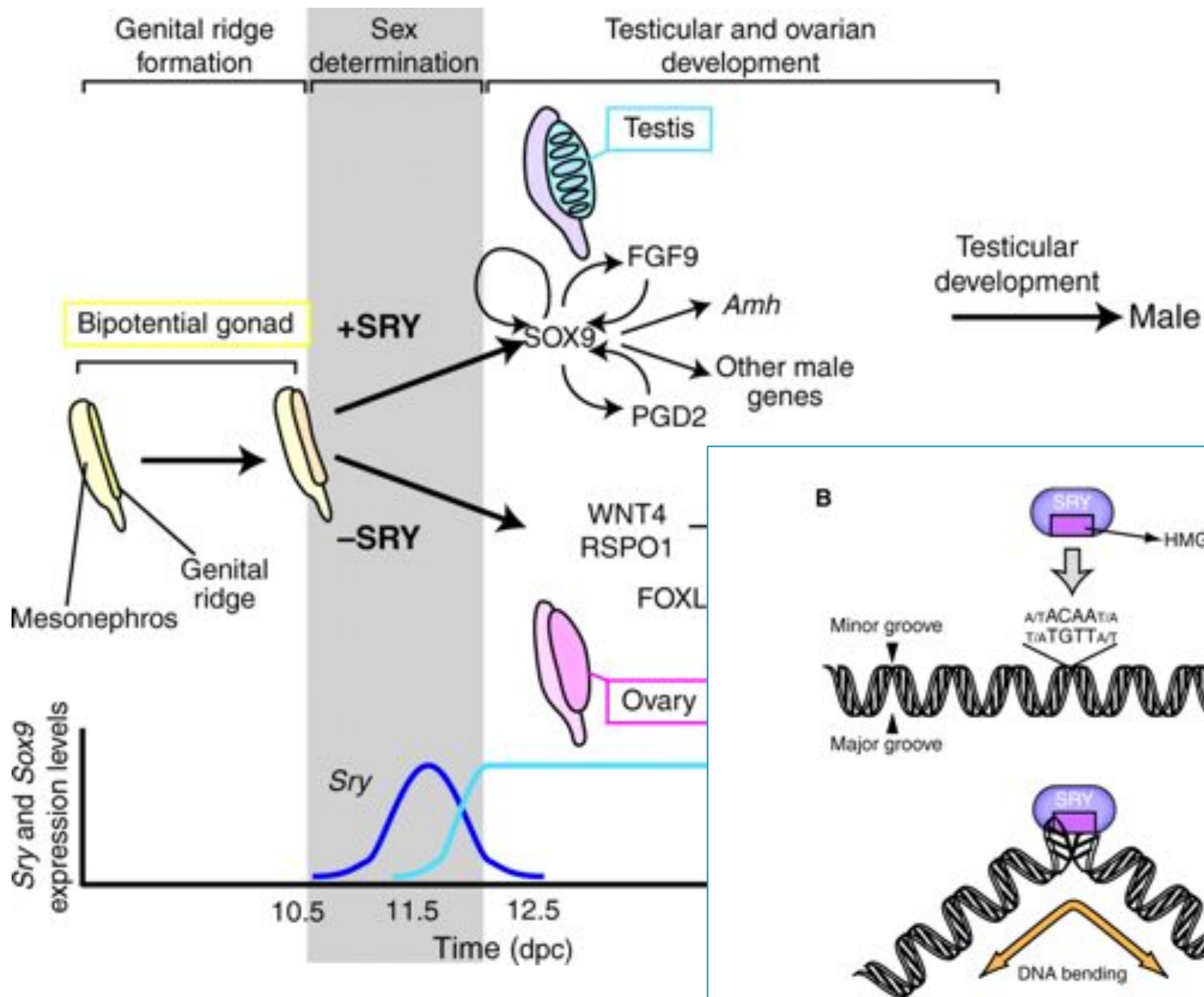
What are DNA sequence motifs?

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



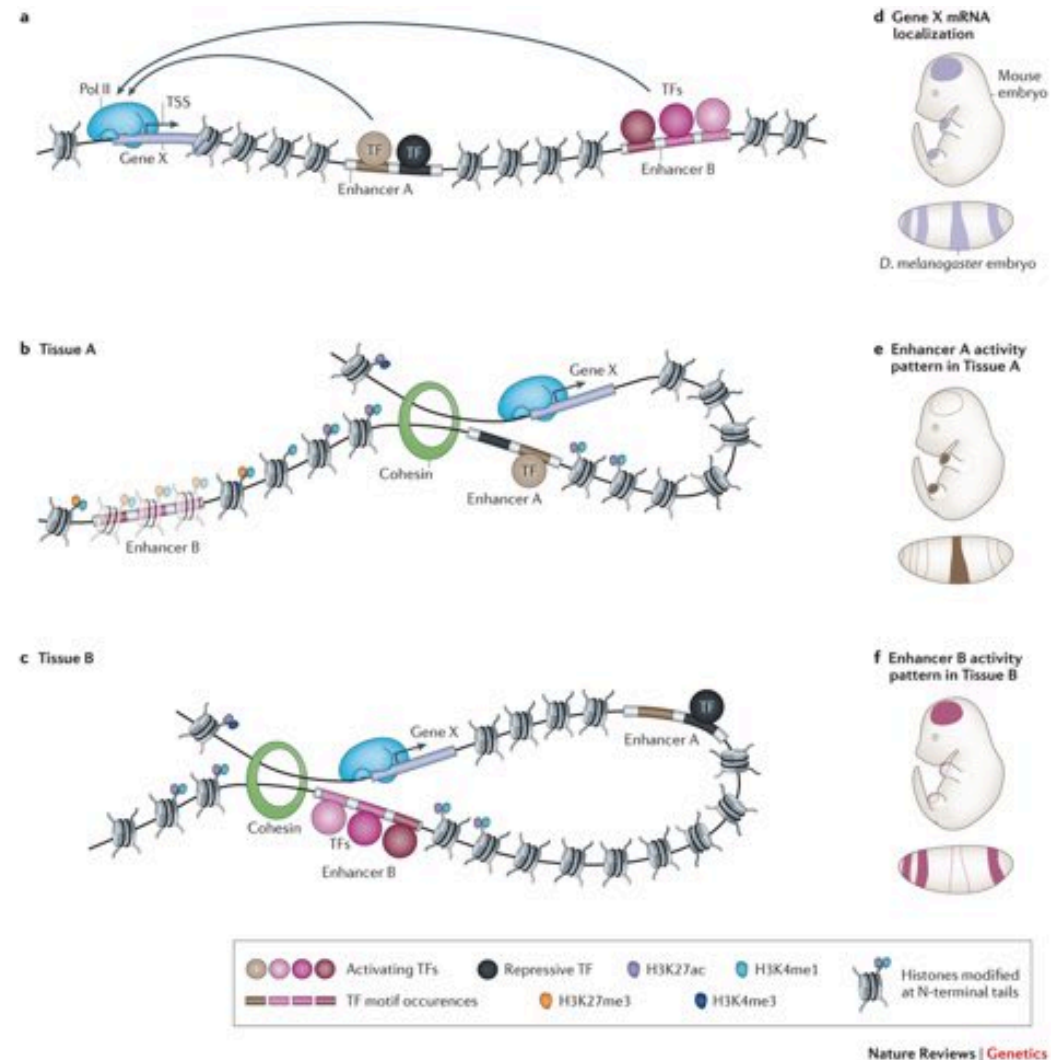
Sry: the master switch in mammalian sex determination

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.

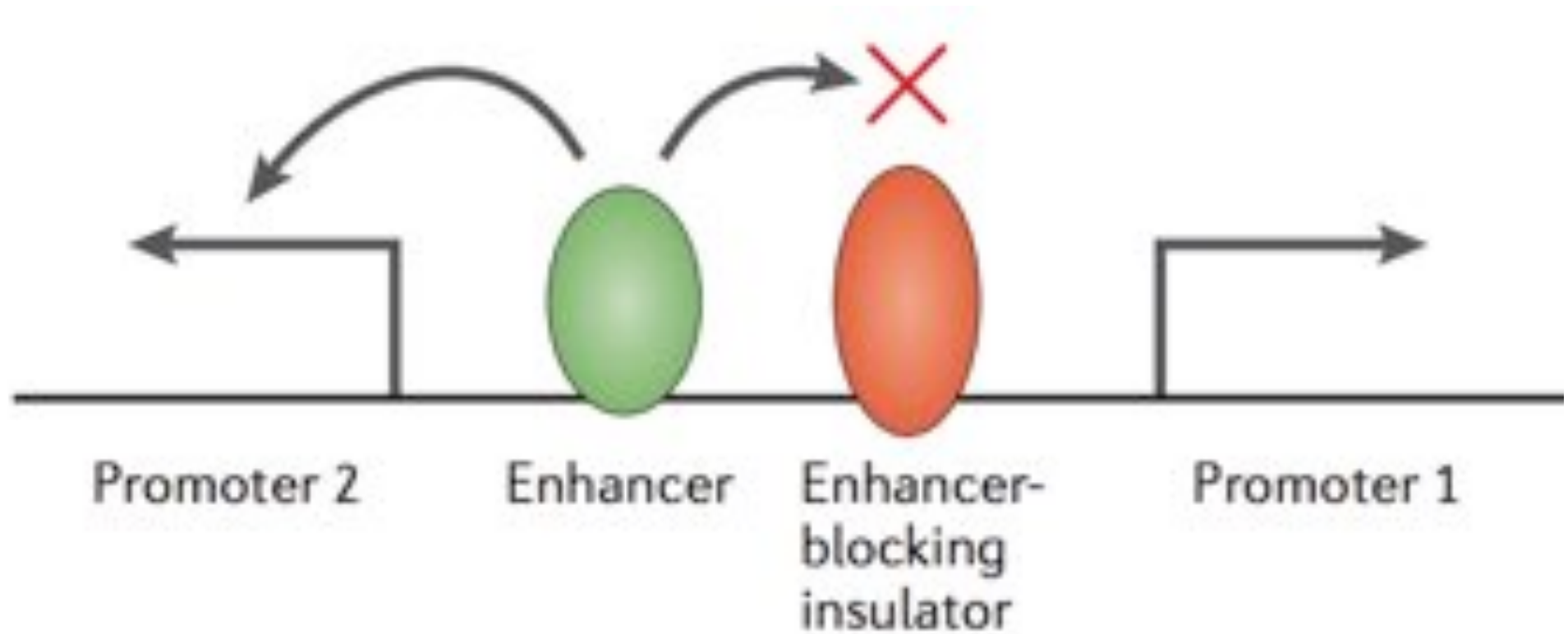
- Enhancers can be located at any distance from their target genes (up to ~1Mbp)
- In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping
- Active and inactive gene regulatory 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.

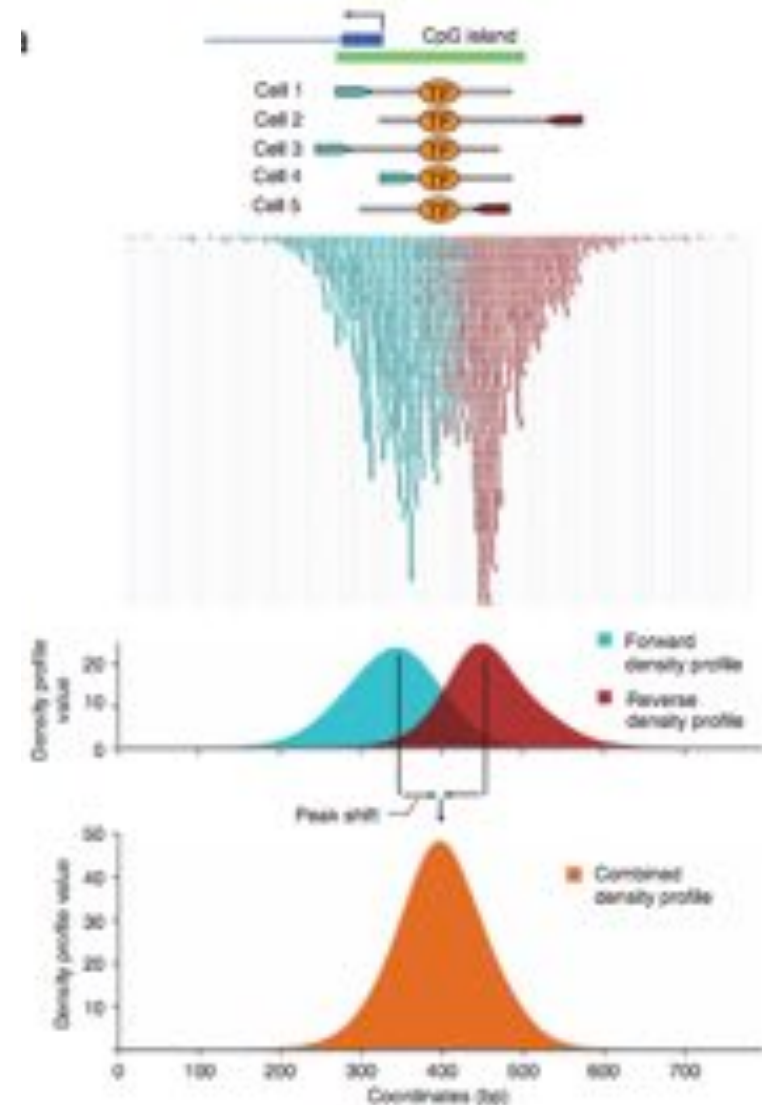
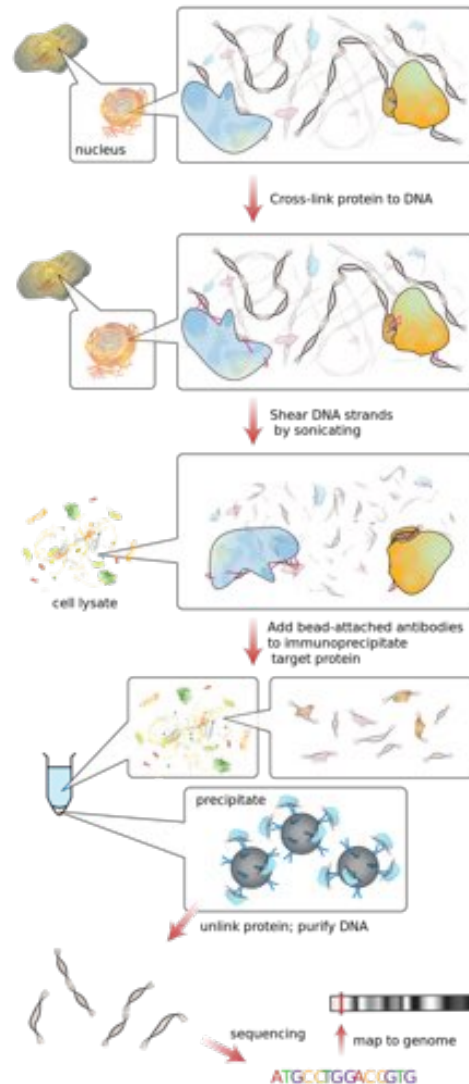
Insulators: exploiting transcriptional and epigenetic mechanisms

Gaszner & Felsenfeld (2006) *Nature Reviews Genetics* 7, 703-713. doi:10.1038/nrg1925

ChIP-seq: TF Binding

Goals:

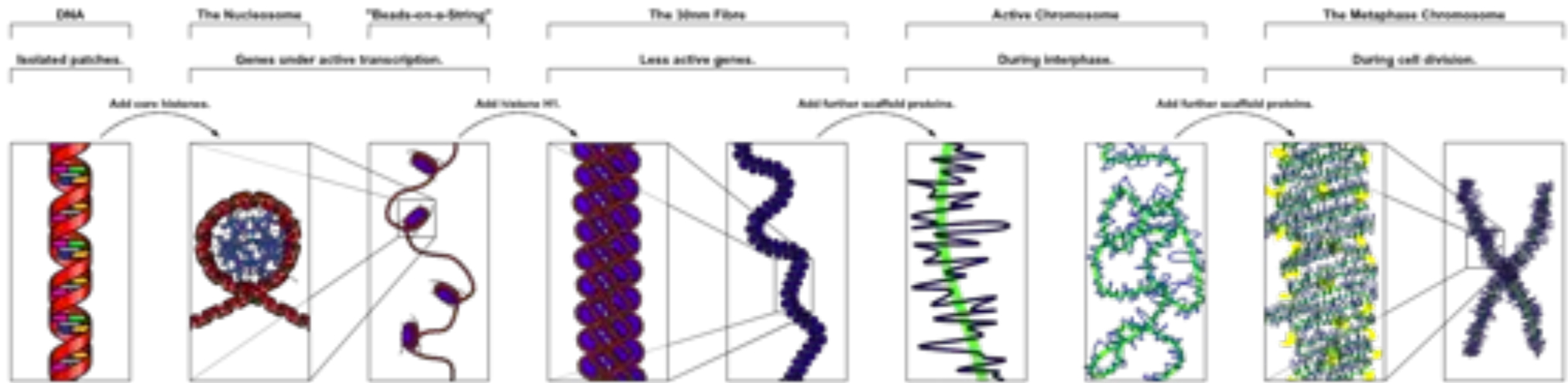
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?



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



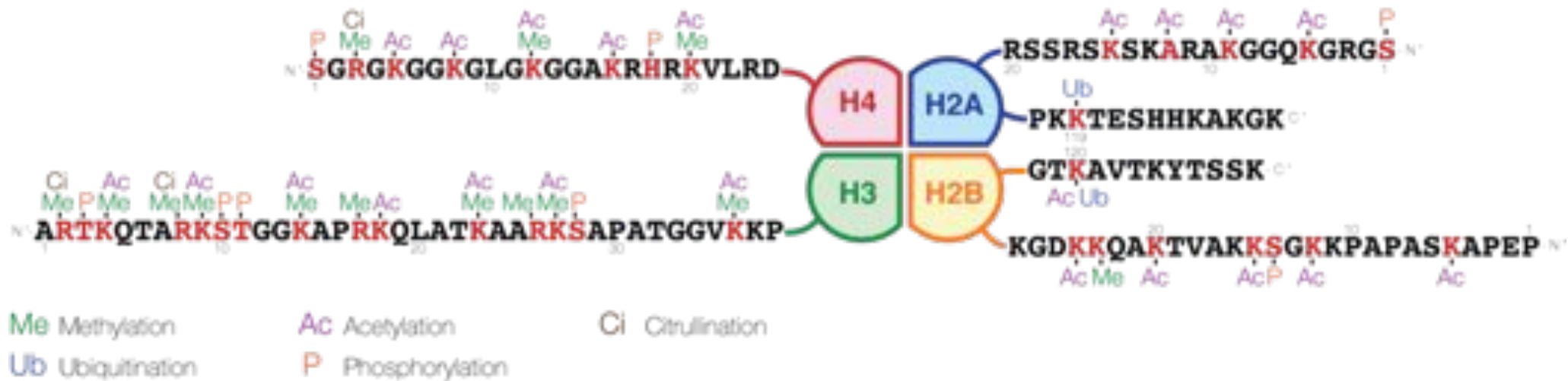
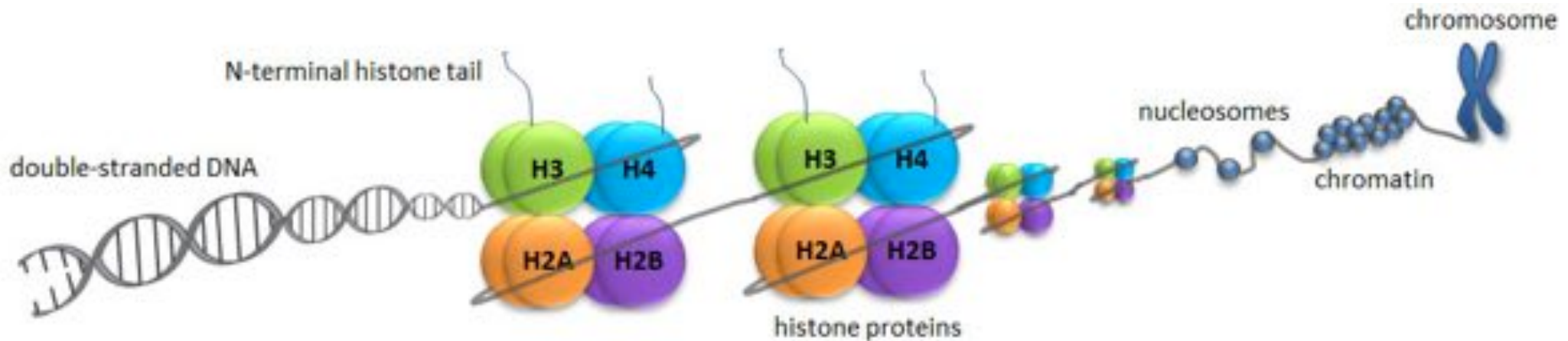
Nucleosome is a basic unit of DNA packaging in eukaryotes

- 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

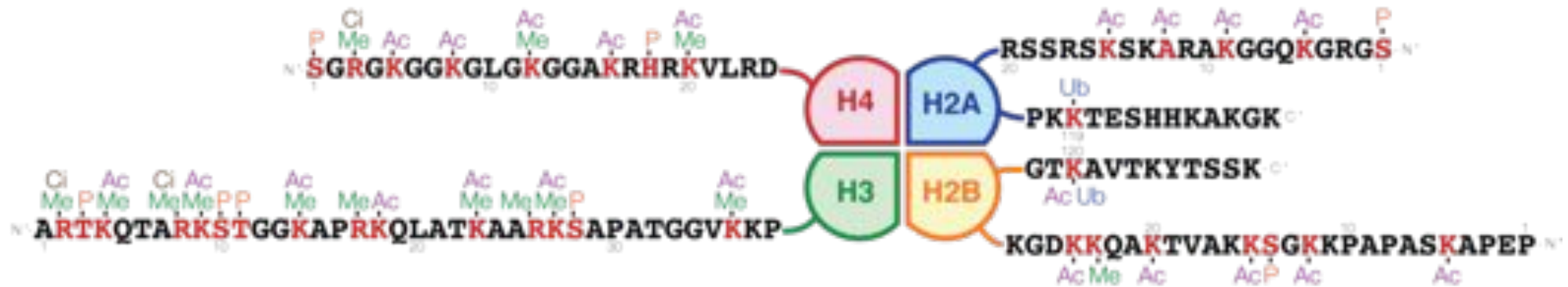
Nucleosomes form the fundamental repeating units of eukaryotic chromatin

- 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



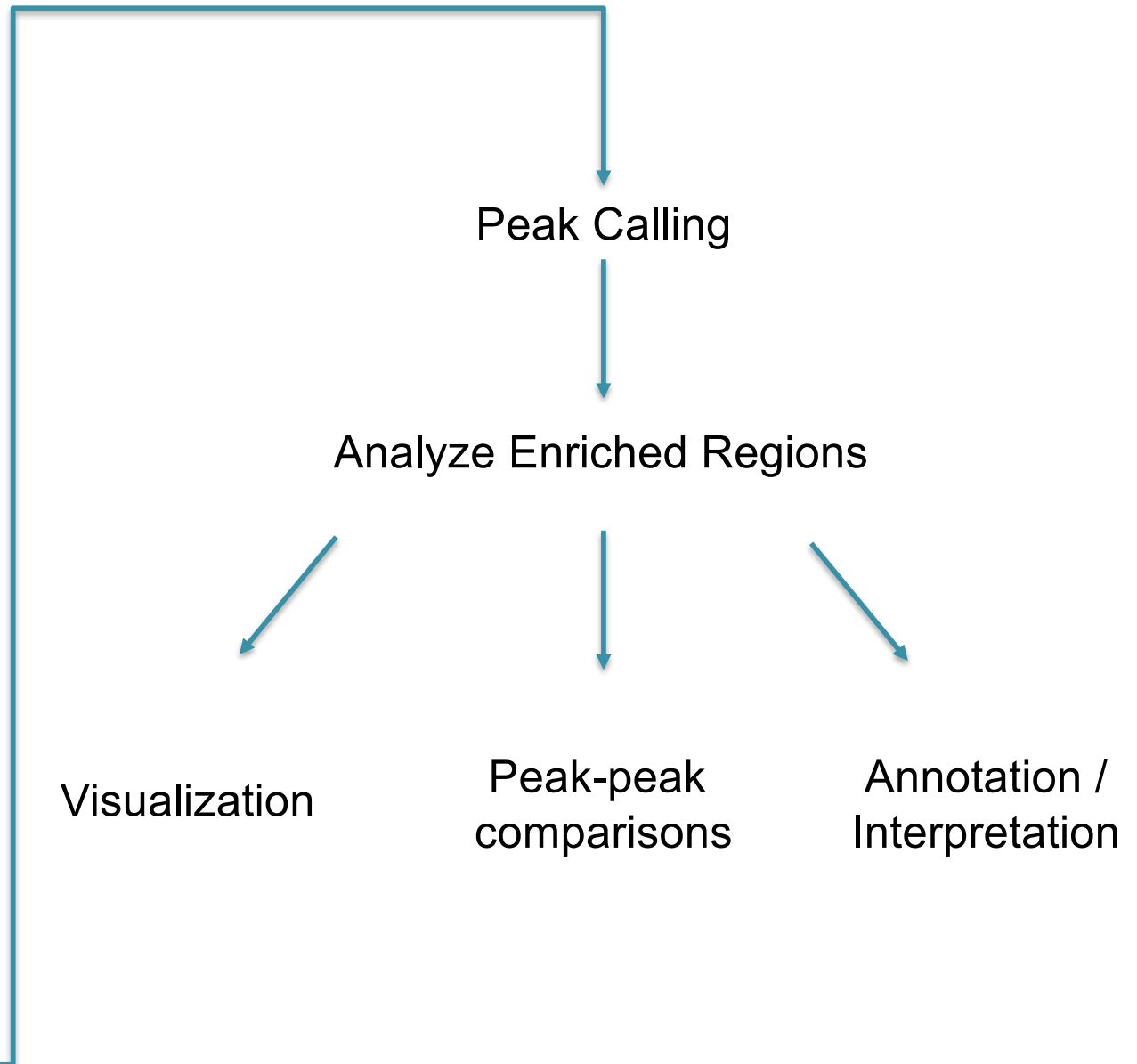
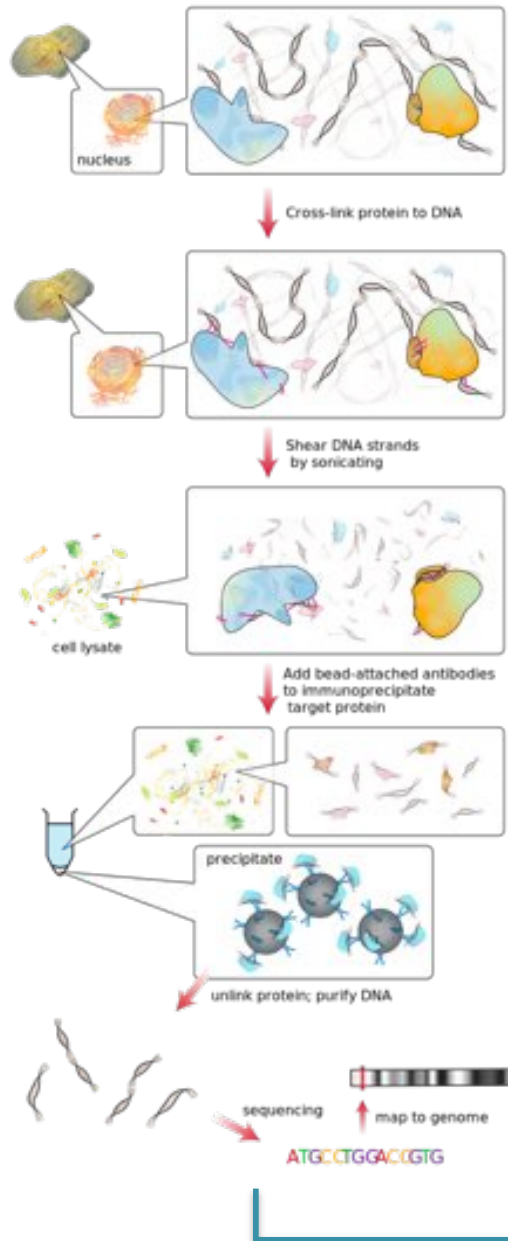
ChIP-seq: Histone Modifications



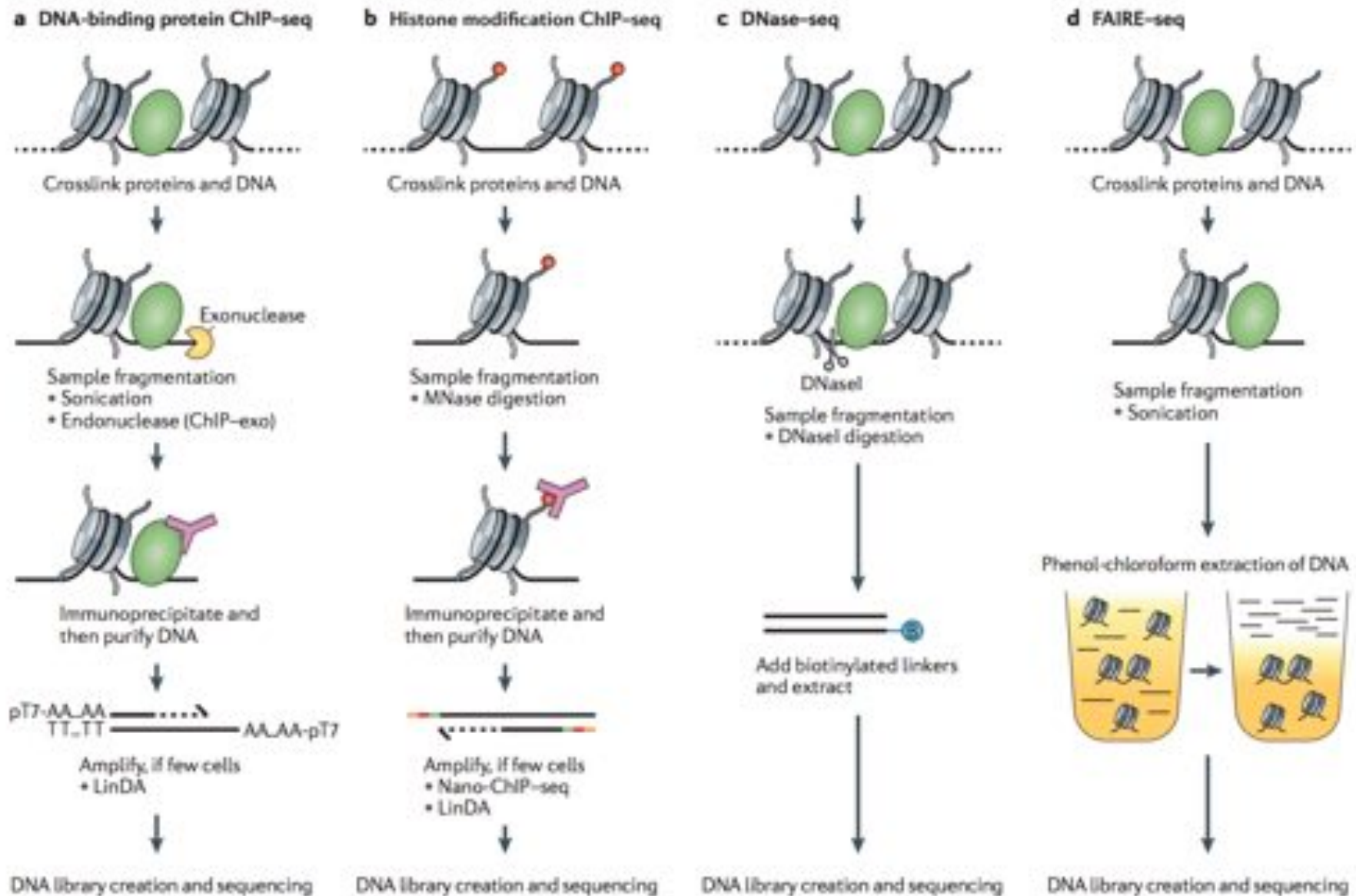
Type of modification	Histone							
	H3K4	H3K9	H3K14	H3K27	H3K79	H3K122	H4K20	H2BK5
mono-methylation	activation ^[6]	activation ^[7]		activation ^[7]	activation ^{[7][8]}		activation ^[7]	activation ^[7]
di-methylation	activation	repression ^[3]		repression ^[3]	activation ^[8]			
tri-methylation	activation ^[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



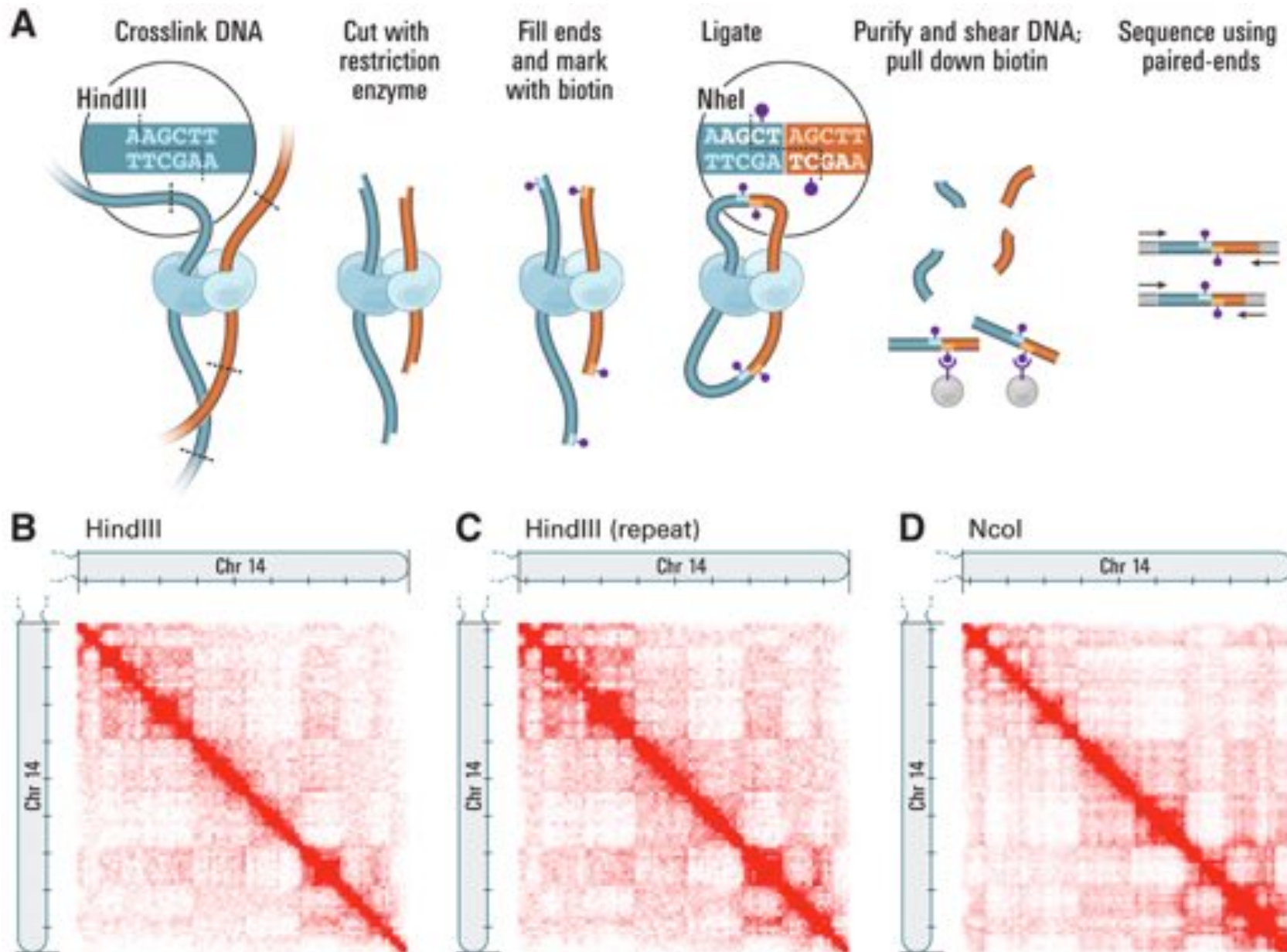
Related Assays



ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions

Furey (2012) *Nature Reviews Genetics*. 13, 840-852

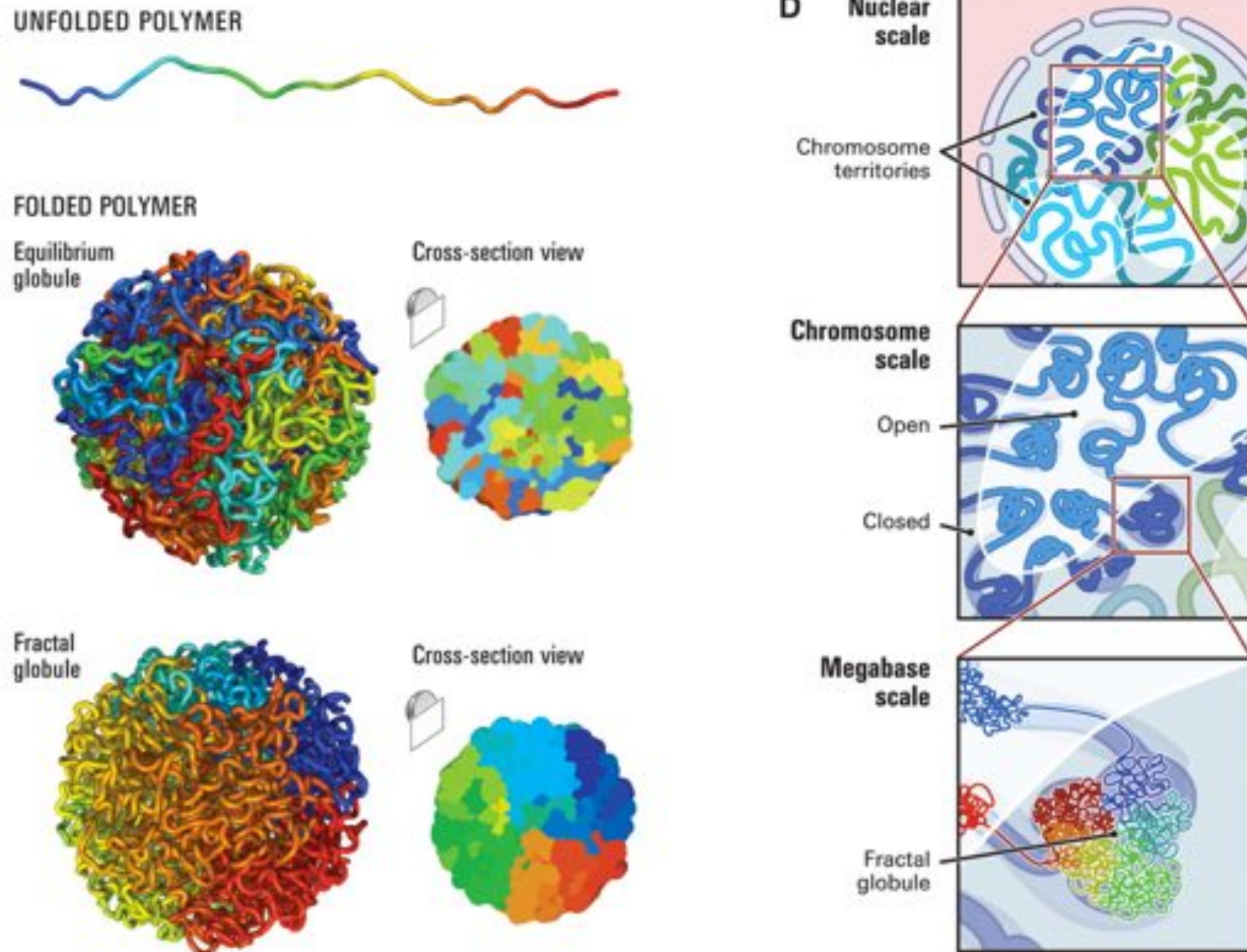
Hi-C: Mapping the folding of DNA



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

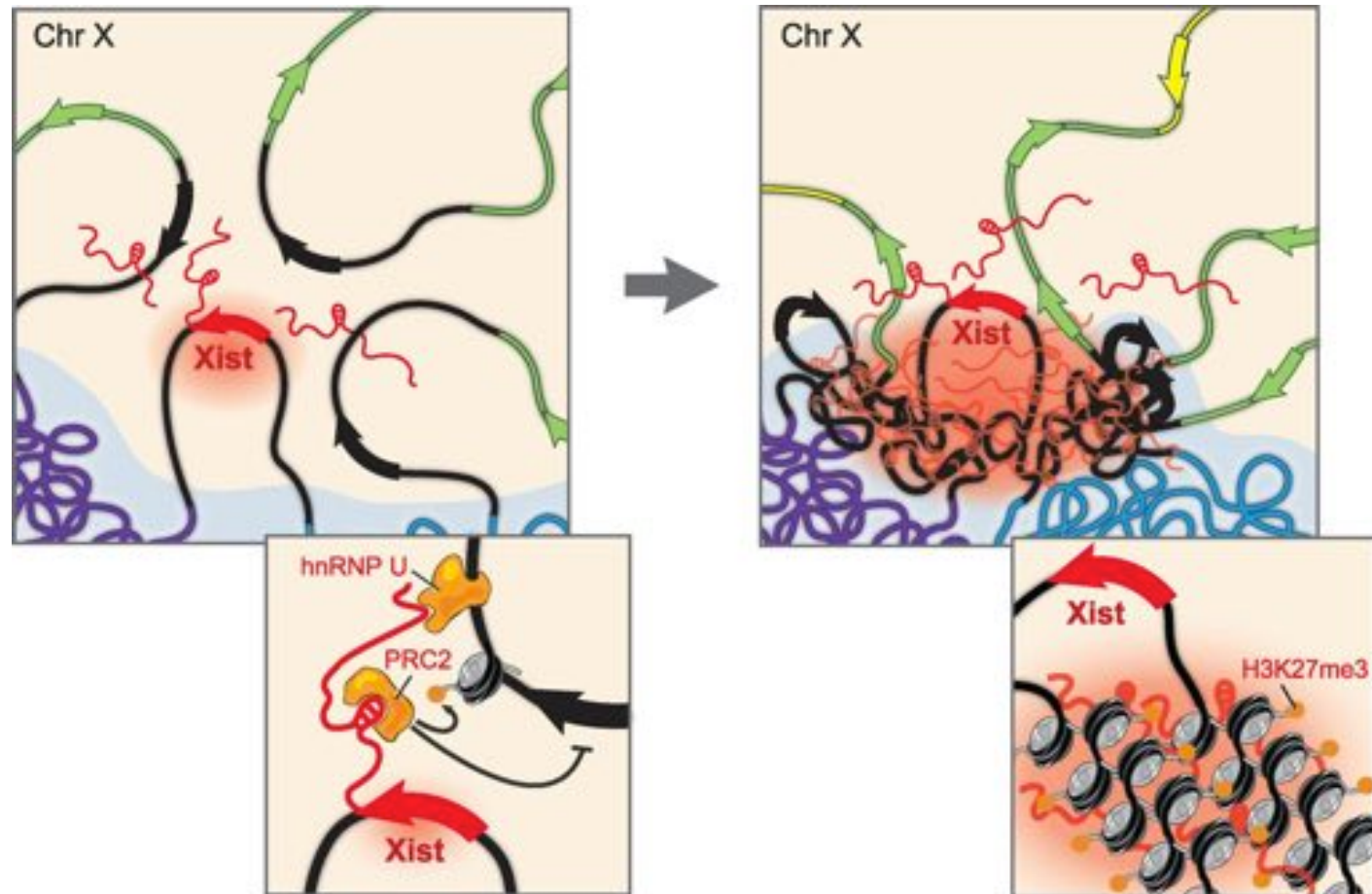
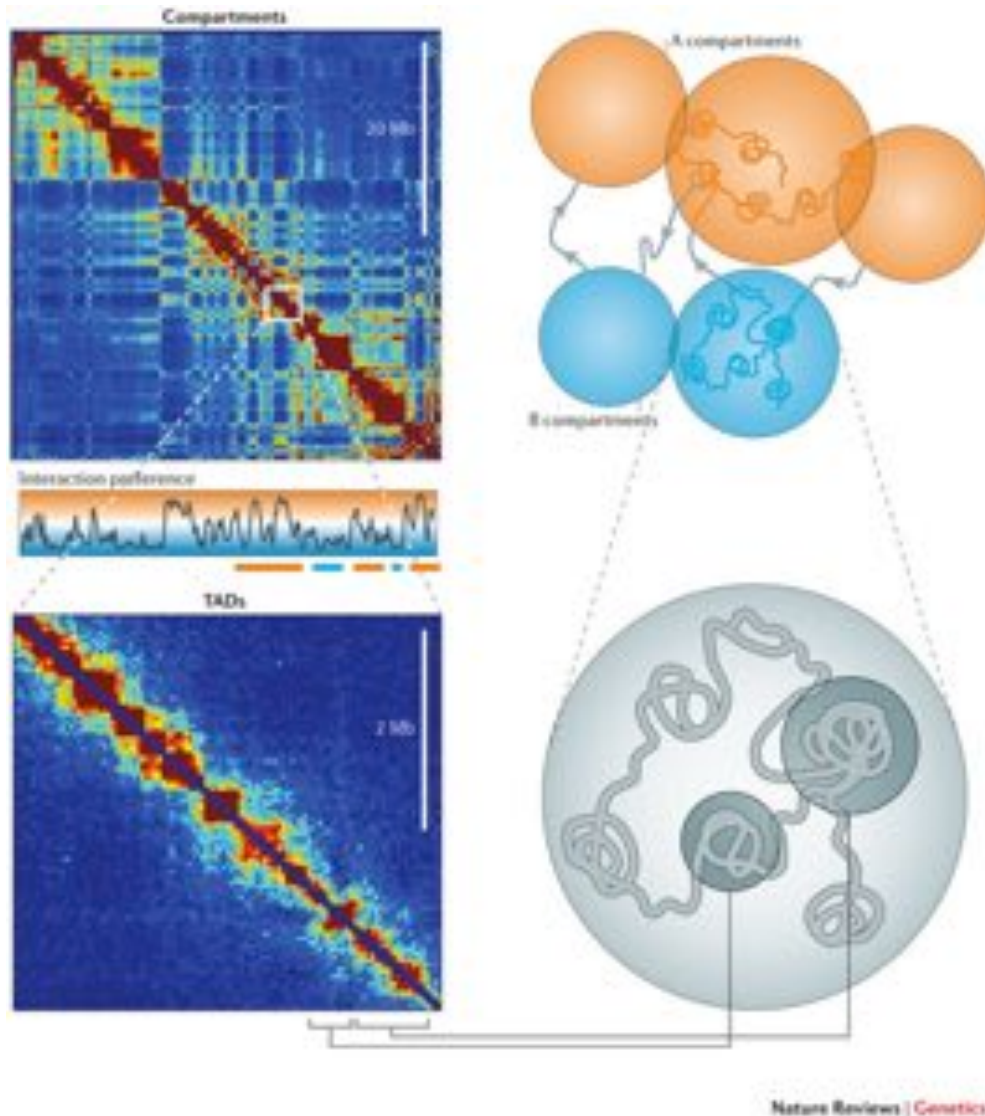


Fig 6. A model for how Xist exploits and alters three-dimensional genome architecture to spread across the X chromosome.

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.

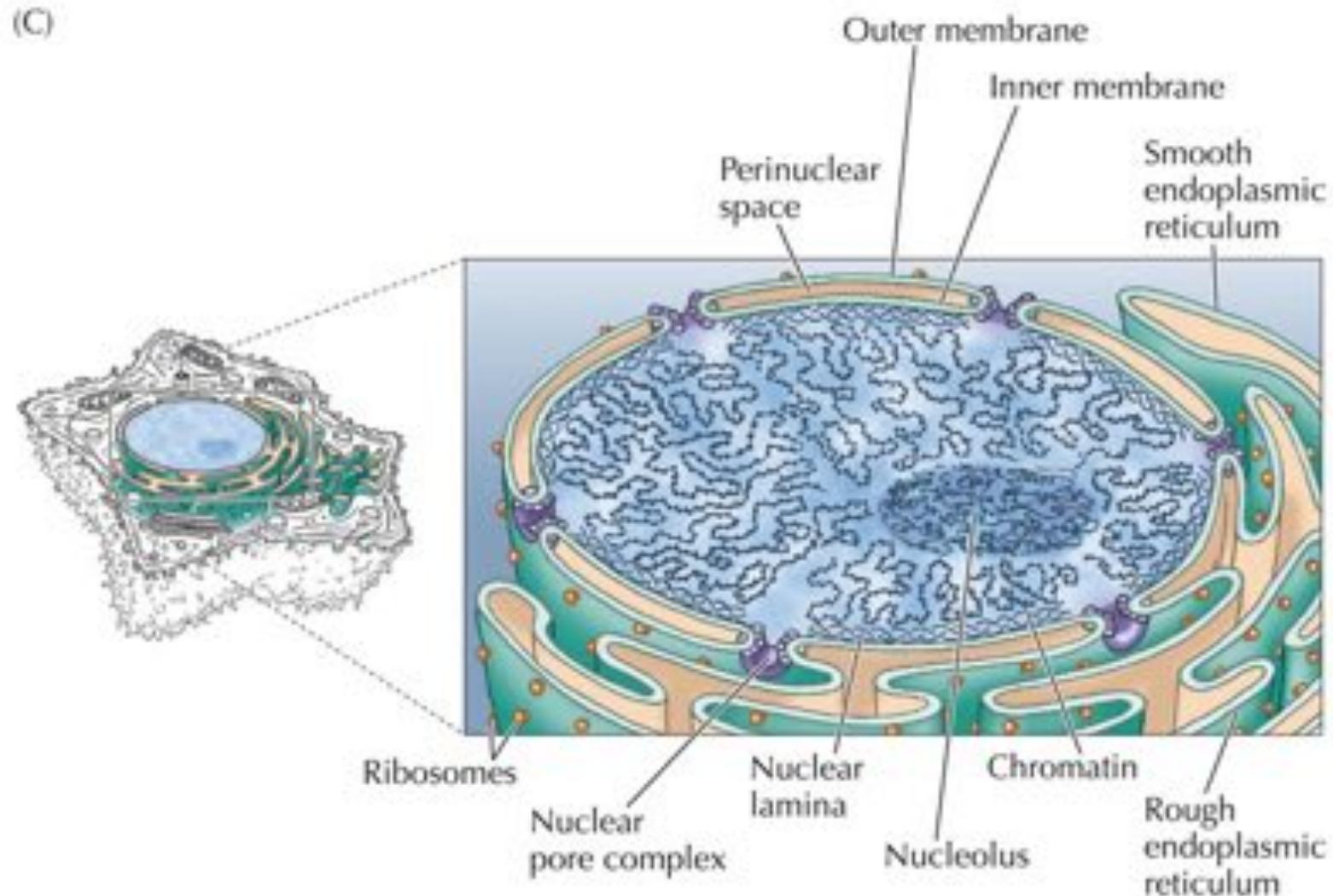
Topologically associating domains (TADs)

- 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”



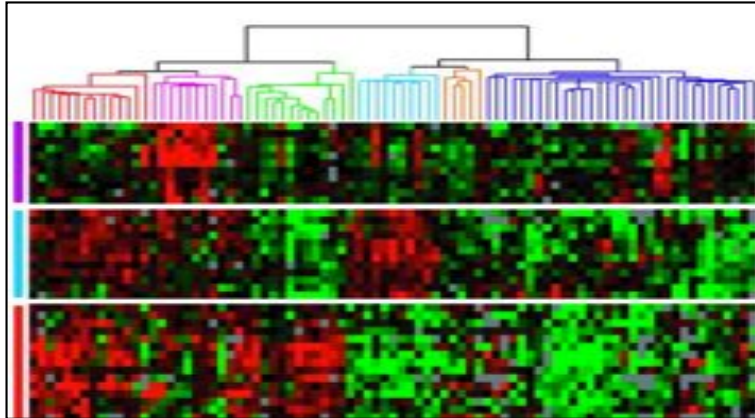
THE CELL, Fourth Edition, Figure 9.1 (Part B) © 2006 ASM Press and Garland Science, Inc.

Chromosome Conformation Paints Reveal the Role of Lamina Association in Genome Organization and Regulation

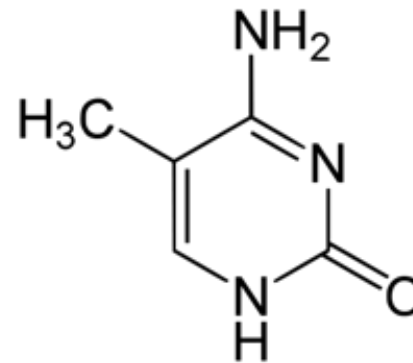
Luperchio et al. (2017) bioRxiv. doi: <https://doi.org/10.1101/122226>

Putting it all together!

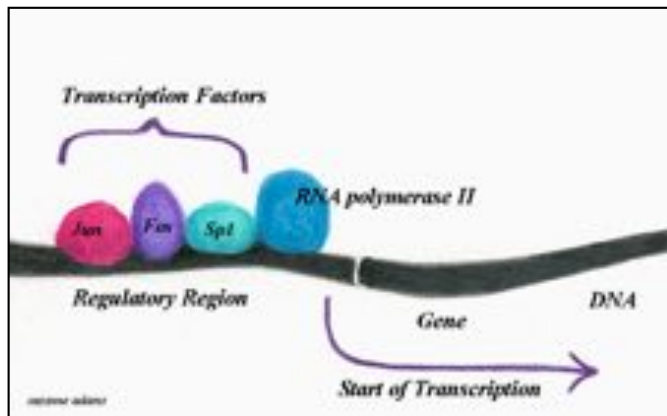
RNA-seq



Methyl-seq



ChIP-seq



Hi-C

