Transcriptional regulation I

Genome
- DNA

Epigenome
- DNA methylation
- Histone PTMs
- no-coding RNA

Transcriptome
- RNA

Proteome
- 3D structure
- PTMs
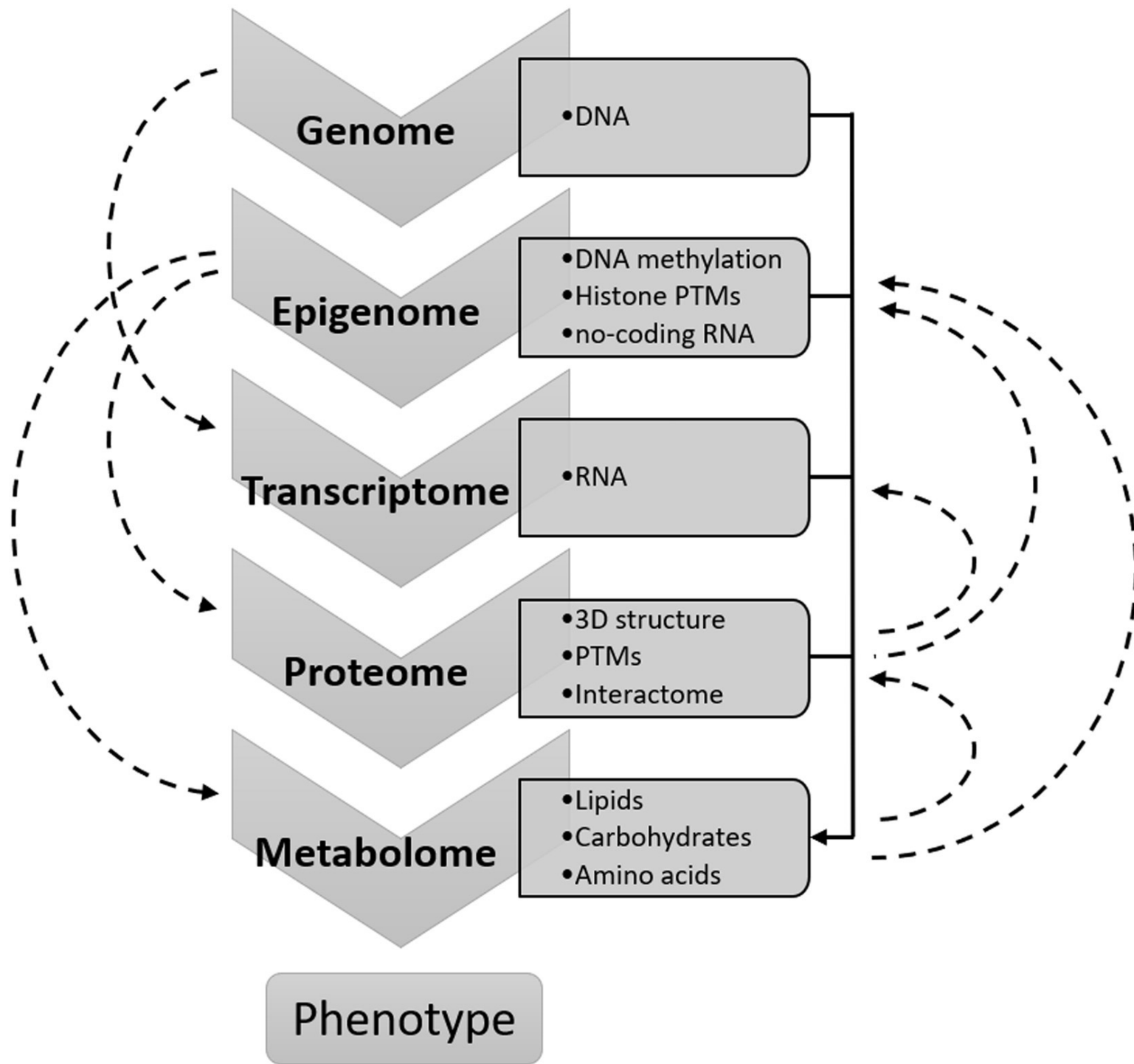- Interactome

Metabolome
- Lipids
- Carbohydrates
- Amino acids

Phenotype

Eukaryotic Transcriptional Regulation

1. Level of Chromatin (DNA accessibility)

- Histone modifications
- Histone modifying enzymes & remodeling complexes
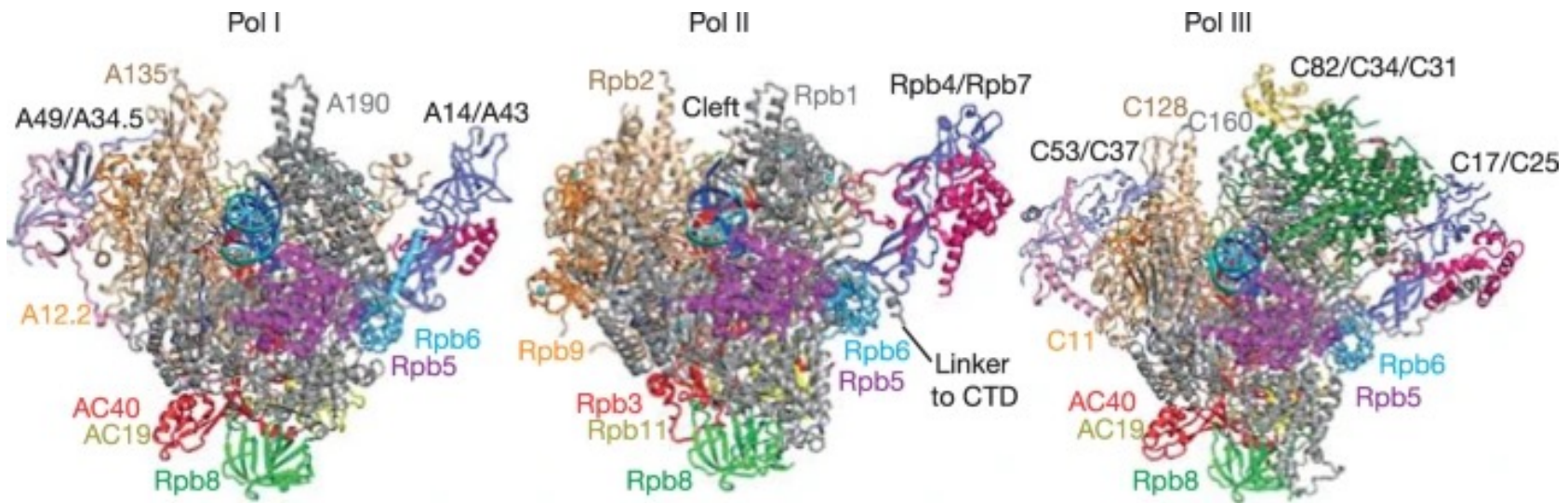- Nucleosome composition
- DNA methylation

2. Level of DNA (Interaction with basal transcription machinery)

- Regulatory sequences (enhancers, silencers)
- Transcription factors (activators, repressors)

3. Level of Regulatory RNA (Interaction with DNA, RNA or protein)
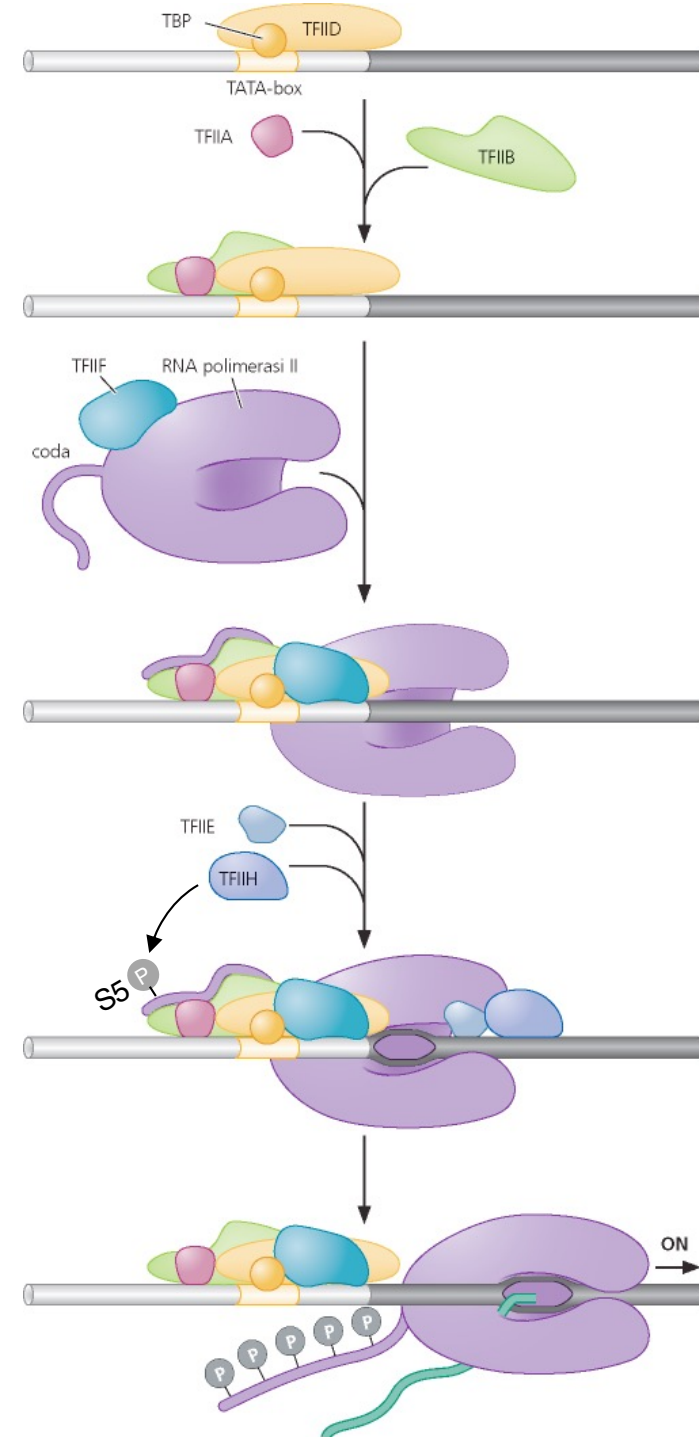
- Small and long non-coding RNAs

The RNA-polimerases

Pol I — rRNA

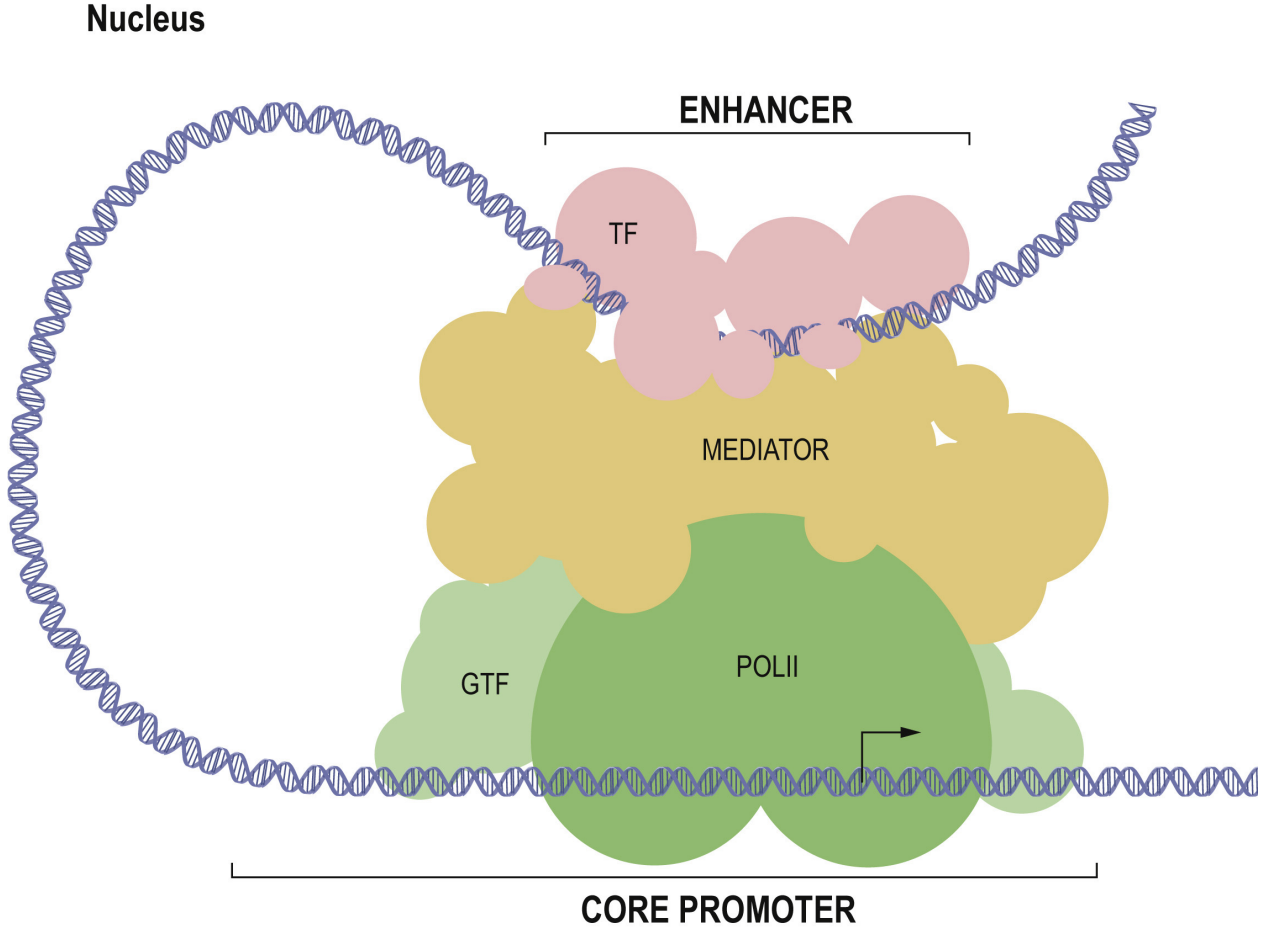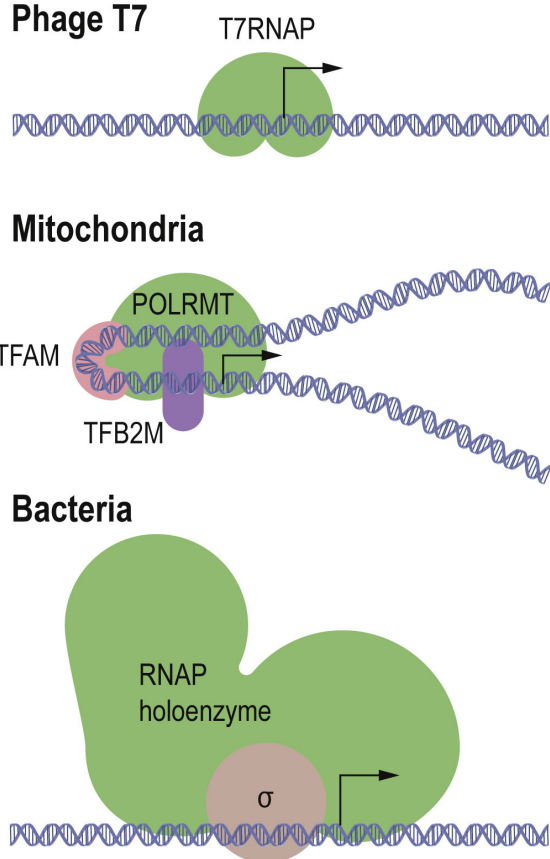Pol II — mRNA, Variety of ncRNA

Pol III — tRNA, Small ribosomal RNA

Transcription initiation in eukaryotes

At the transcription start site, Pol II initiation is regulated by a protein assembly known as the **pre-initiation complex (PIC)** containing **TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, Pol II** and **Mediator**

Schematic comparison of transcription initiation systems

Comparison of transcription activation mechanisms in prokaryotes and eukaryotes

The ultimate purpose of DNA-binding activators is to mediate RNA polymerase recruitment and function. But whereas activators interact directly with RNA polymerase subunits in prokaryotes, the effects of activators on RNA polymerase II in eukaryotes are mostly indirect and involve interactions with diverse cofactors that modify chromatin structure and facilitate recruitment of RNA polymerase II and general initiation factors
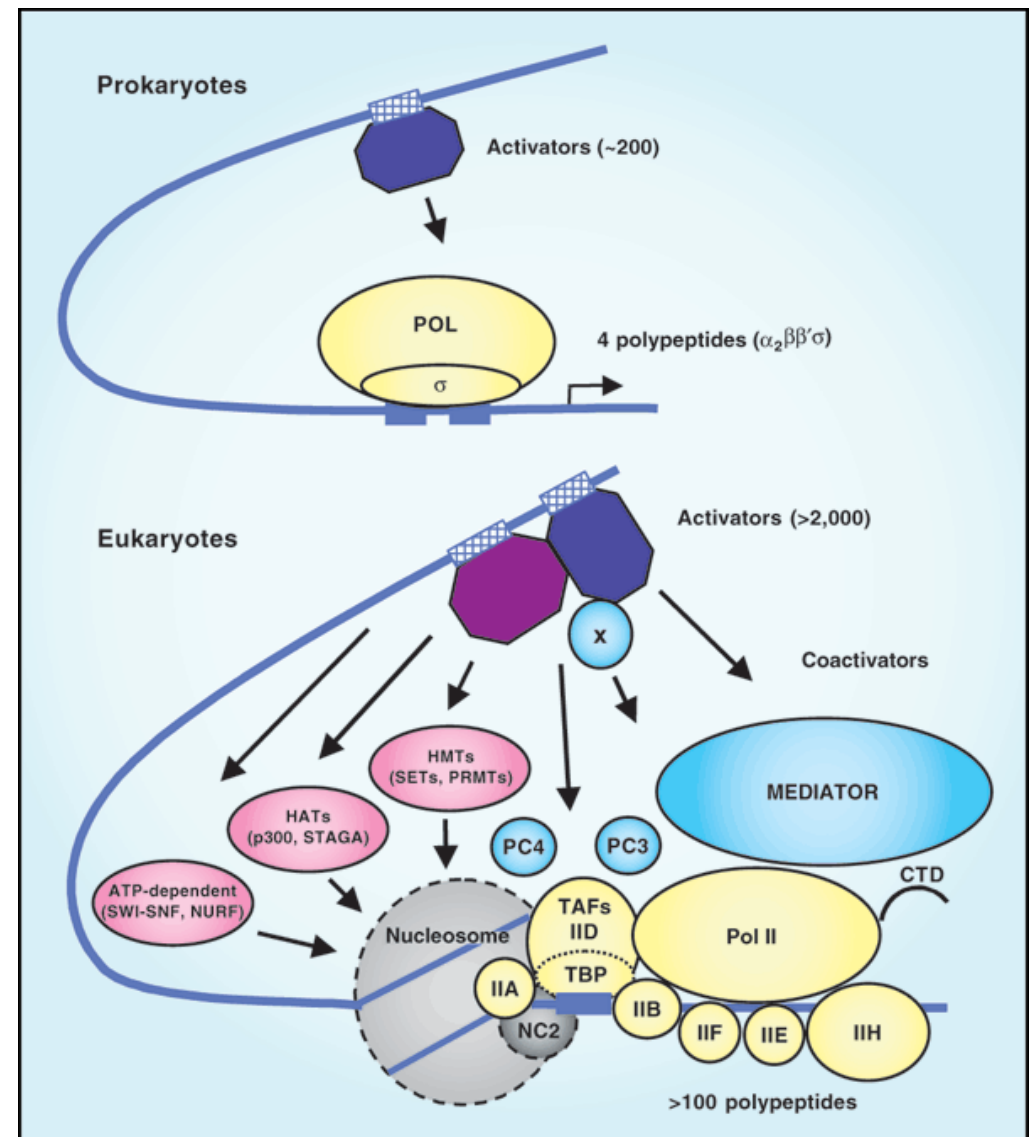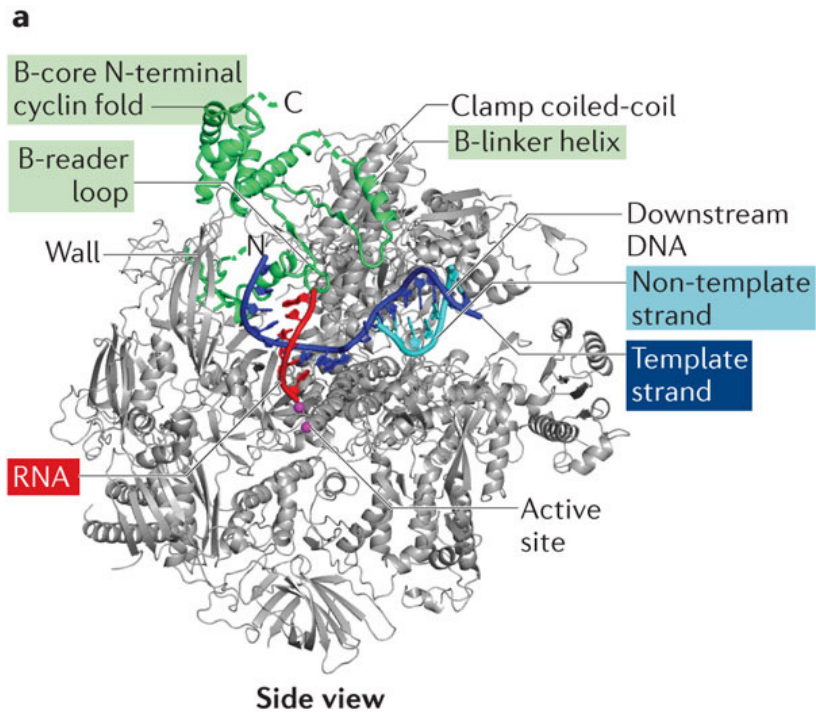
RNA transcription

Types of RNAs transcribed from active regulatory elements

| Type	Acronyms	Estimated number	Half-life	Processing features
pre-messenger RNA	pre-mRNA	19,954	minutes to hours	capped spliced polyA tail
long non-coding RNA	lncRNA lincRNA	30,000	minutes to hours	capped spliced polyA tail
enhancer RNA	eRNA	40,000-65,000	minutes	capped
promoter-associated RNA	uaRNA paRNA PROMPT	Gene and lncRNA promoters	minutes	capped

RNA Pol II

a

B-core N-terminal cyclin fold
B-reader loop
Wall
RNA
C
N
Clamp coiled-coil
B-linker helix
Downstream DNA
Non-template strand
Template strand
Active site

Side view

Factor	Gene name		Mass (kDa)		Uniprot accession number	
	Yeast	Human	Yeast	Human	Yeast	Human
Pol II (RNAP*): transcribing enzyme						
RPB1	*RPO21*	*POLR2A*	191.6	217.2	P04050	P24928
RPB2	*RPB2*	*POLR2B*	138.8	133.9	P08518	P30876
RPB3	*RPB3*	*POLR2C*	35.3	31.4	P16370	P19387
RPB4	*RPB4*	*POLR2D*	25.4	16.3	P20433	O15514
RPB5‡	*RPB5*	*POLR2E*	25.1	24.6	P20434	P19388
RPB6‡	*RPO26*	*POLR2F*	17.9	14.5	P20435	P61218
RPB7	*RPB7*	*POLR2G*	19.1	19.3	P34087	P62487
RPB8‡	*RPB8*	*POLR2H*	16.5	17.1	P20436	P52434
RPB9	*RPB9*	*POLR2I*	14.3	14.5	P27999	P36954
RPB10‡	*RPB10*	*POLR2L*	8.3	7.6	P22139	P62875
RPB11	*RPB11*	*POLR2J*	13.6	13.3	P38902	P52435
RPB12‡	*RPB12*	*POLR2K*	7.7	7.0	P40422	P53803
Total (12 subunits)			513.6	516.7		

The carboxy-terminal domain (CTD)

This domain consists of heptapeptide repeats with the consensus sequence YSPTSPS and forms the **carboxy-terminal domain (CTD)** of the largest subunit of Pol II RNA polymerase B1 (RPB1). The CTD is not required for the catalytic functions of Pol II, but it is crucial for the regulation of multiple steps in transcription, as well as for the coupling of transcription to a number of co- transcriptional processes.
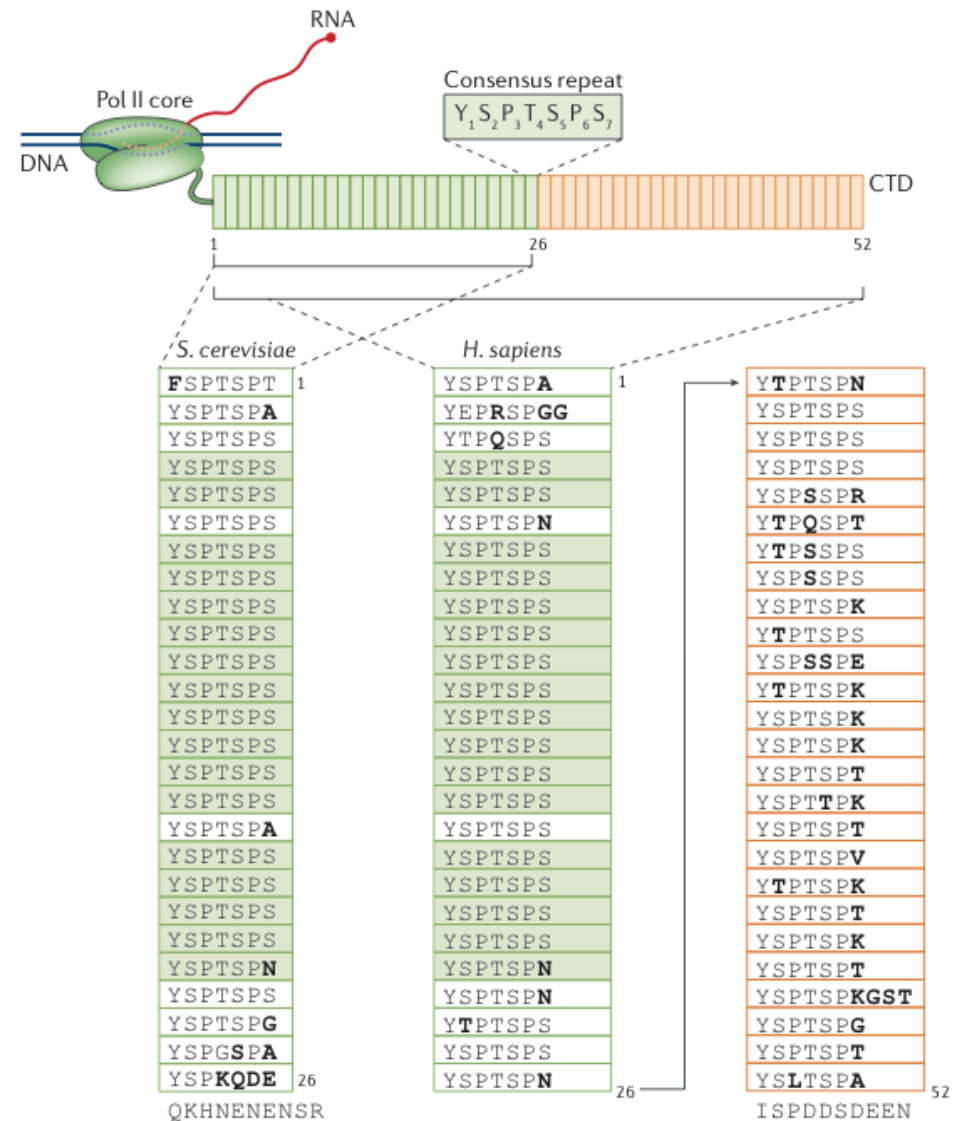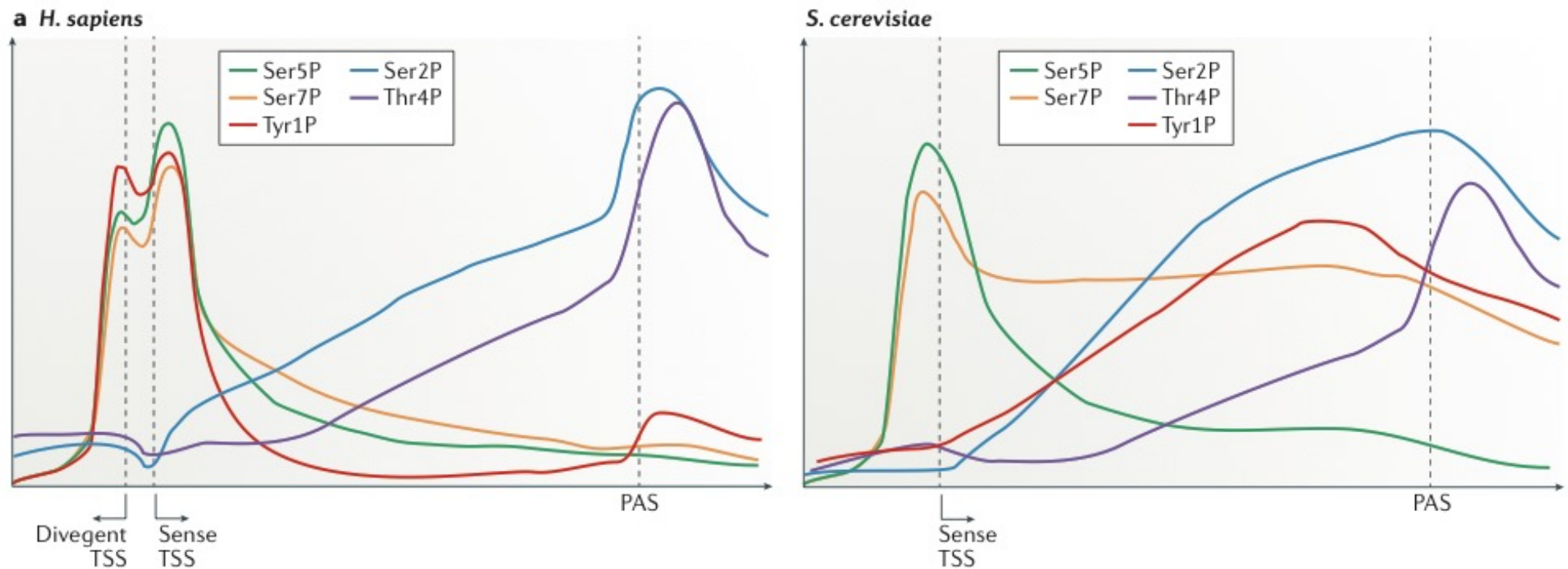
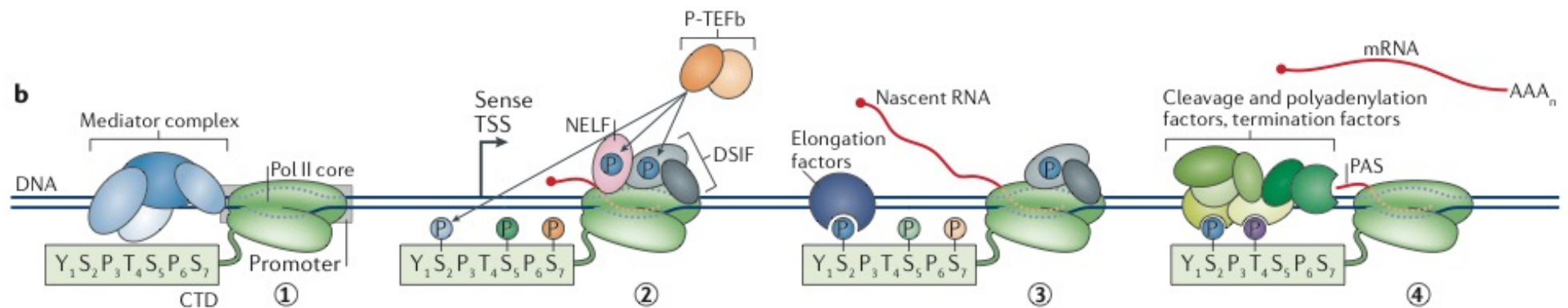# Post-translational modifications of the CTD and their associated processes

Post-translational modification	Position in the CTD	Organisms	Associated process or processes
Ser5 phosphorylation	Multiple repeats	• Saccharomyces cerevisiae • Schizosaccharomyces pombe • Homo sapiens	Transcription initiation, mRNA capping and splicing, non-coding RNA transcription termination and chromatin modification
Ser2 phosphorylation	Multiple repeats	• S. cerevisiae • S. pombe • H. sapiens	Transcription elongation, promoter-proximal pause and release, splicing, transcription termination and DNA topology
Ser7 phosphorylation	Multiple repeats	• S. cerevisiae • S. pombe • H. sapiens	snRNA expression, interaction with the Integrator complex and P-TEFb recognition
Thr4 phosphorylation	Multiple repeats	• S. cerevisiae • S. pombe • Gallus gallus • H. sapiens	Transcription elongation and termination, post-transcriptional splicing, processing of histone mRNA and chromatin remodelling
Tyr1 phosphorylation	Multiple repeats	• S. cerevisiae • S. pombe • G. gallus • H. sapiens	Inhibition of recruitment of transcription termination factors, CTD stability, antisense and enhancer transcription
Arg methylation	Arg1,810 of human RPB1	• Mus musculus • H. sapiens	snRNA and snoRNA regulation, R-loop resolution and transcription termination
Lys methylation	Lys7 in the non-consensus region of human CTD	• H. sapiens • M. musculus • Drosophila melanogaster • Caenorhabditis elegans	Supports nucleosome occupancy at promoters; negatively regulates gene expression
Lys acetylation	Lys7 in the non-consensus region of murine CTD; Lys7 in repeats 39, 42, 47 and 49 of human CTD	• H. sapiens • M. musculus	Induction of growth-factor response genes, transcription elongation; maintains balance between Lys methylation and acetylation and affects mRNA expression levels
O-GlcNAcylation	Ser5 and/or Ser7 in multiple repeats	H. sapiens	Pre-initiation complex assembly
Ubiquitylation	RPB1 Lys residues 859, 1866, 1873, 1887, 1908, 1922	M. musculus	RPB1 degradation

CTD, carboxy-terminal domain of RNA polymerase II; P-TEFb, positive transcription elongation factor B; RPB1, RNA polymerase B1; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA.

Transcription regulation by the CTD code

Average chromatin immunoprecipitation (ChIP) profiles of phosphorylated residues of the carboxy-terminal domain (CTD) of RNA polymerase II (Pol II) across protein-coding genes in humans (*Homo sapiens)* and in budding yeast (*Saccharomyces cerevisiae)* .

b

P-TEFb

mRNA

Mediator complex | Sense TSS | NELF | Nascent RNA | Cleavage and polyadenylation factors, termination factors

AAA$_n$

DNA

Pol II core | DSIF | Elongation factors | PAS

$Y_1S_2P_3T_4S_5P_6S_7$ Promoter CTD ①

$Y_1S_2P_3T_4S_5P_6S_7$ ②

$Y_1S_2P_3T_4S_5P_6S_7$ ③

$Y_1S_2P_3T_4S_5P_6S_7$ ④

1. **Recruitment of the core Pol II enzyme** to the promoter with an unphosphorylated CTD that interacts with the Mediator complex. The Mediator complex has a high affinity for unphosphorylated CTD and upon phosphorylation of Ser5 of the CTD, this affinity is lost and Pol II escapes from the promoter.

2. **Pol II during promoter-proximal pausing**. At this step, Pol II is highly phosphorylated at Ser5 and Ser7, is paused downstream of the transcription start site (TSS) and is bound by **negative elongation factor (NELF)** and **DRB sensitivity-inducing factor (DSIF).** The arrival of **positive elongation factor B (P-TEFb)** leads to the phosphorylation of NELF, DSIF and Ser2. These phosphorylation events are followed by the release of NELF and the transition into productive transcription elongation.

3. **During productive elongation**, the CTD contains lower levels of Ser5P and Ser7P and higher levels of Ser2P, which promotes the recruitment of many transcription elongation, chromatin-modifying and RNA-processing factors that regulate co-transcriptional processes.

4. **Pol II transitions from transcription elongation to termination**. Ser2P and Thr4P levels peak, thereby promoting the recruitment of cleavage and polyadenylation factors, as well as termination factors that release Pol II from the DNA.

General Transcription Factors

| Factor	Gene name		Mass (kDa)		Uniprot accession number		Copies
	Yeast	Human	Yeast	Human	Yeast	Human	
TFIIA[§]: TBP stabilization and counteracts repressive effects of negative co-factors							
Large subunit	TOA1	GTF2A1	32.2	41.5	P32773	P52655	1
Small subunit	TOA2	GTF2A2	13.5	12.5	P32774	P52657	1
Total (2 subunits)			45.7	54.0			
TFIIB: Pol II recruitment, TBP binding and TSS selection							
TFIIB (TFB*)	SUA7	GTF2B	38.2	34.8	P29055	Q00403	1
TFIID: Pol II recruitment and promoter recognition							
TBP (TBP*): recognition of the TATA box	TBP	TBP	27.0	37.7	P13393	P20226	1
TAF1	TAF1	TAF1	120.7	212.7	P46677	P21675	1
TAF2	TAF2	TAF2	161.5	137.0	P23255	Q6P1X5	1
TAF3	TAF3	TAF3	40.3	103.6	Q12297	Q5VWG9	1
TAF4	TAF4	TAF4	42.3	110.1	P50105	O00268	2
TAF5	TAF5	TAF5	89.0	86.			2
TAF6	TAF6	TAF6	57.9	72.			2
TAF7	TAF7	TAF7	67.6	40.			1
TAF8	TAF8	TAF8	58.0	34.			1
TAF9	TAF9	TAF9	17.3	29.0	Q05027	Q16594	2
TAF10	TAF10	TAF10	23.0	21.7	Q12030	Q12962	2
TAF11	TAF11	TAF11	40.6	23.3	Q04226	Q15544	1
TAF12	TAF12	TAF12	61.1	17.9	Q03761	Q16514	2
TAF13	TAF13	TAF13	19.1	14.3	P11747	Q15543	1
TAF14[‖]	TAF14	NA	27.4	NA	P35189	NA	3
Total (14–15 subunits)			1,200[¶]	1,300[¶]			
TFIIE: recruitment of TFIIH and open DNA stabilization							
TFIIEα (TFE*)	TFA1	GTF2E1	54.7	49.5	P36100	P29083	1
TFIIEβ	TFA2	GTF2E2	37.0	33.0	P36145	P29084	1
Total (2 subunits)			91.7	82.5			

TAFs recognize DNA elements in the promoter region
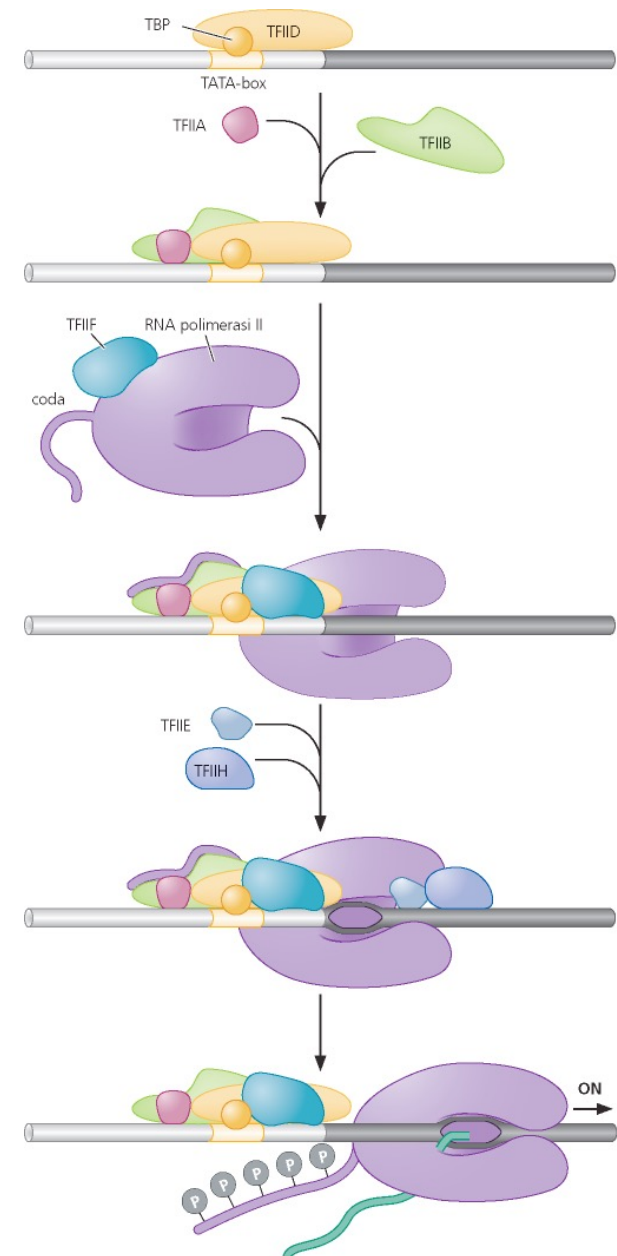
PIC assembly pathway for genes with a TATA-containing core promoter

Assembly of a PIC containing RNA polymerase II and general initiation factors (yellow) is nucleated by binding of TFIID to the core promoter.

A model for the regulation of PIC assembly and function involves, sequentially: (i) **binding of regulatory factors to distal control elements**; (ii) **regulatory factor interactions with cofactors that modify chromatin structure** to facilitate additional factor interactions; and (iii) **regulatory factor interactions with cofactors that act after chromatin remodeling to facilitate**, through direct interactions, recruitment or function of the general transcription machinery.
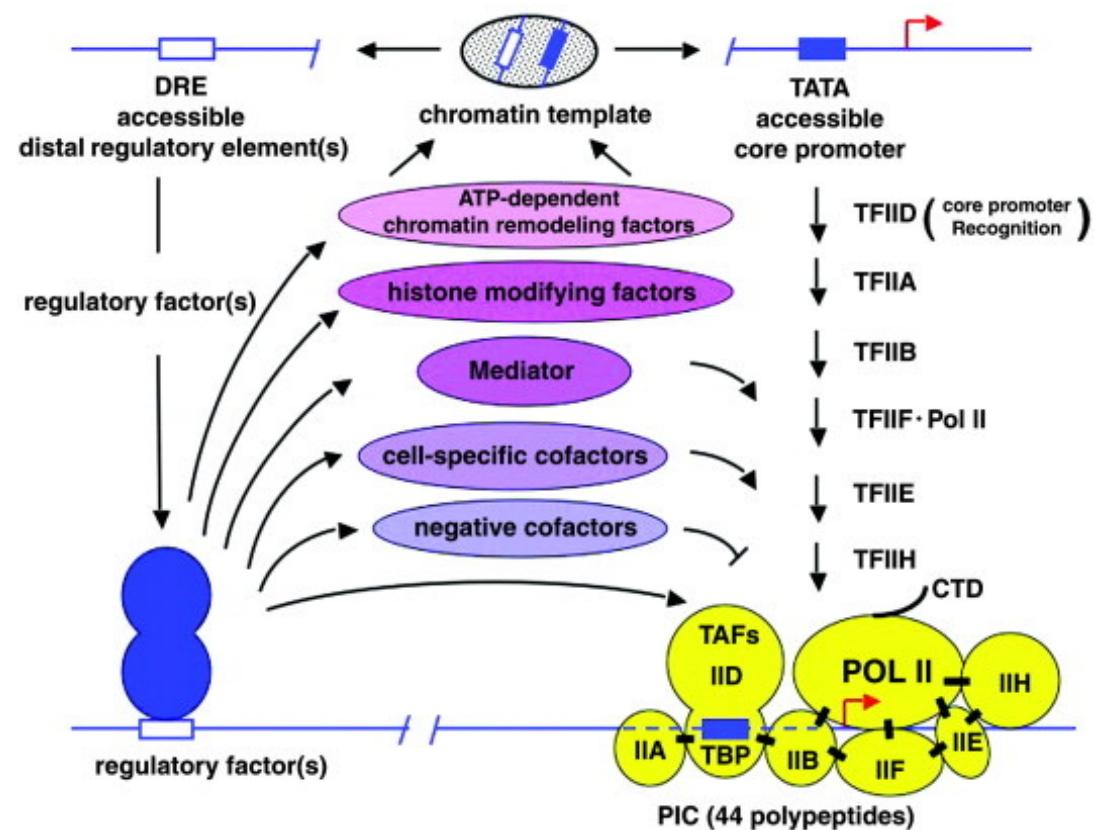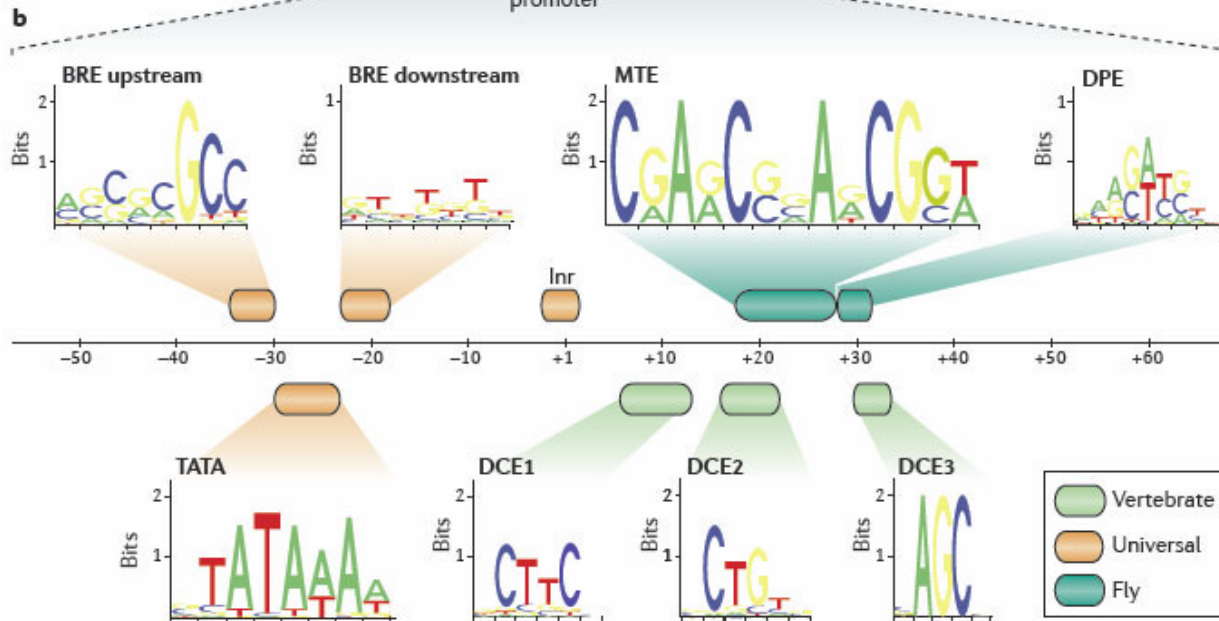
Promoter elements and regulatory signals in Metazoa
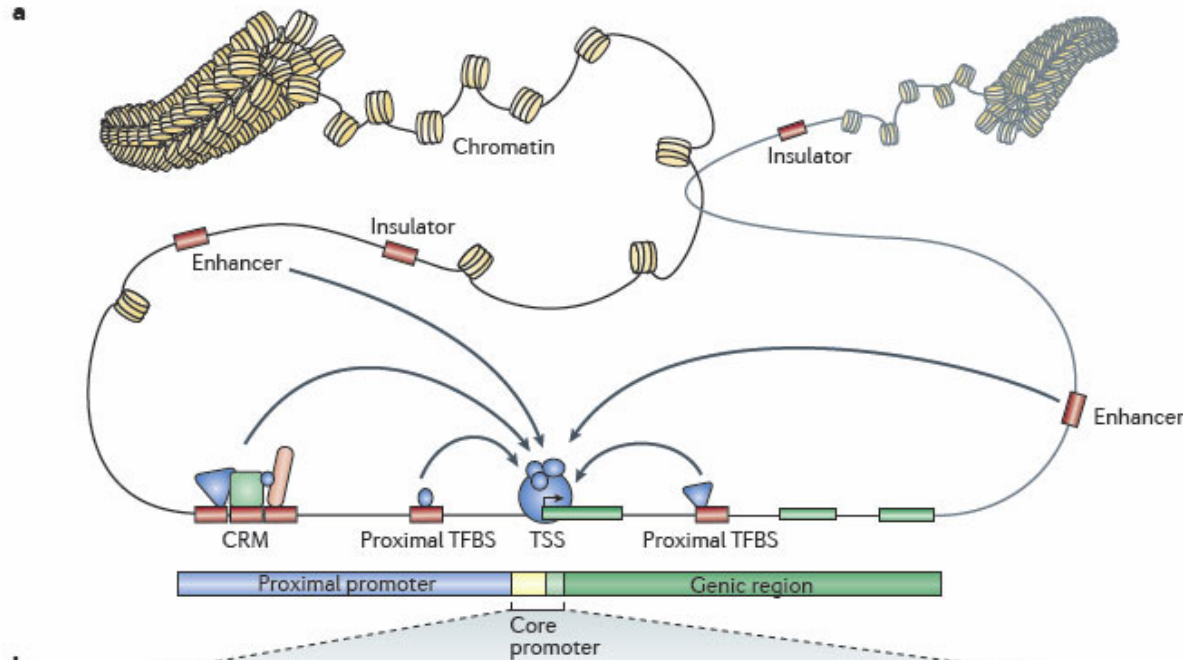
The term '**core promoter**' is often used to focus on the DNA region in the immediate vicinity of the TSS, which is assumed to dock the **pre-initiation complex (PIC)**

The core promoter consists of several interchangeable sequence elements around the TSS, which bind core components of the PIC.

The region around the TSS has several over-represented sequence patterns; the **TATA box** and **initiator (Inr)** are the most studied

BRE, B recognition elements
DCE, downstream core element;
DRE, DNA recognition element;
MTE, motif ten element.

Nature Reviews | Genetics

TFIID and promoter specificity

It was originally thought that TFIID was recruited to Pol II promoters through the TATA-binding activity of TBP. A conserved TATA box is, however, found only in 10%–20% of yeast and human promoters. Analysis of core promoter sequences led to the identification of several additional core promoter elements that are recognized by TAFs. TFIID <u>also binds to transcription factors</u>, which help to recruit TFIID to promoters and may facilitate enhancer–promoter interactions.
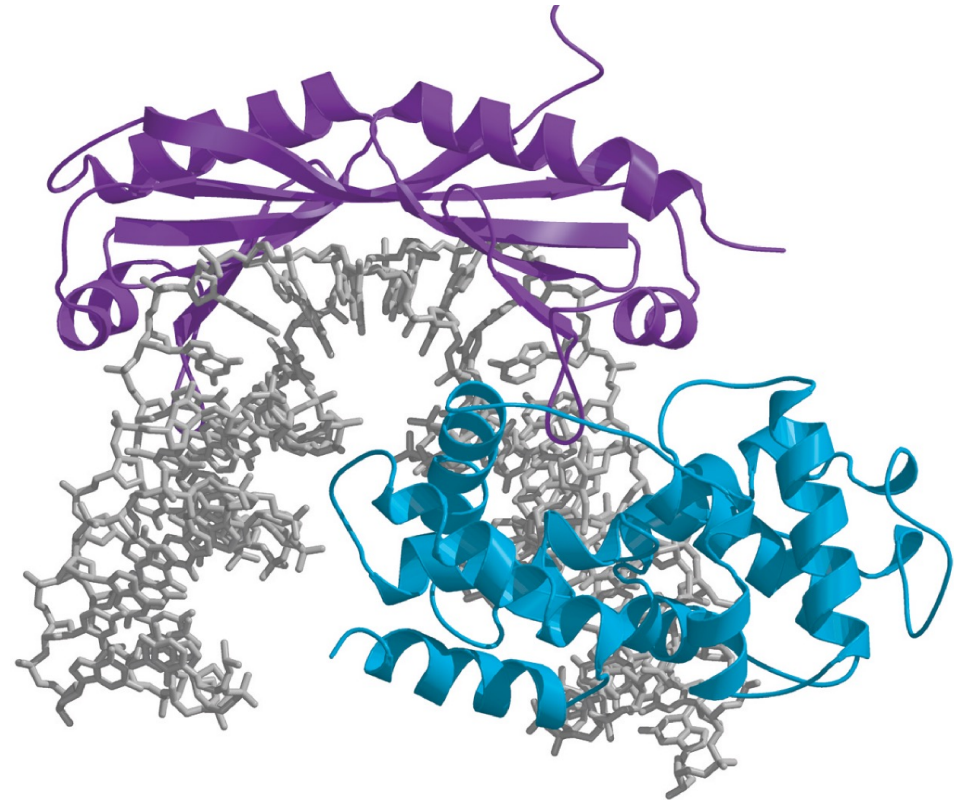

Nature Reviews | Molecular Cell Biology
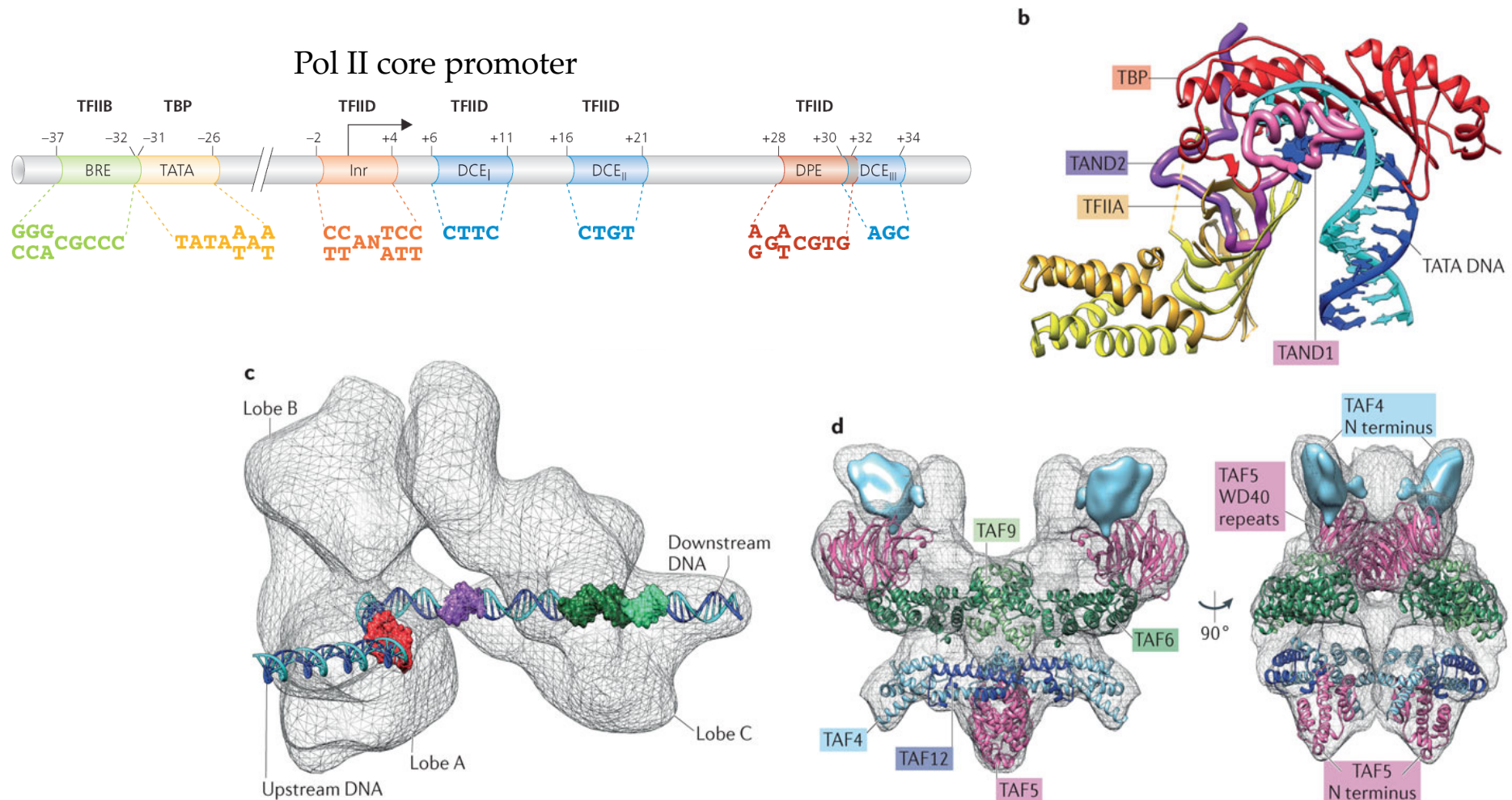
The TBP
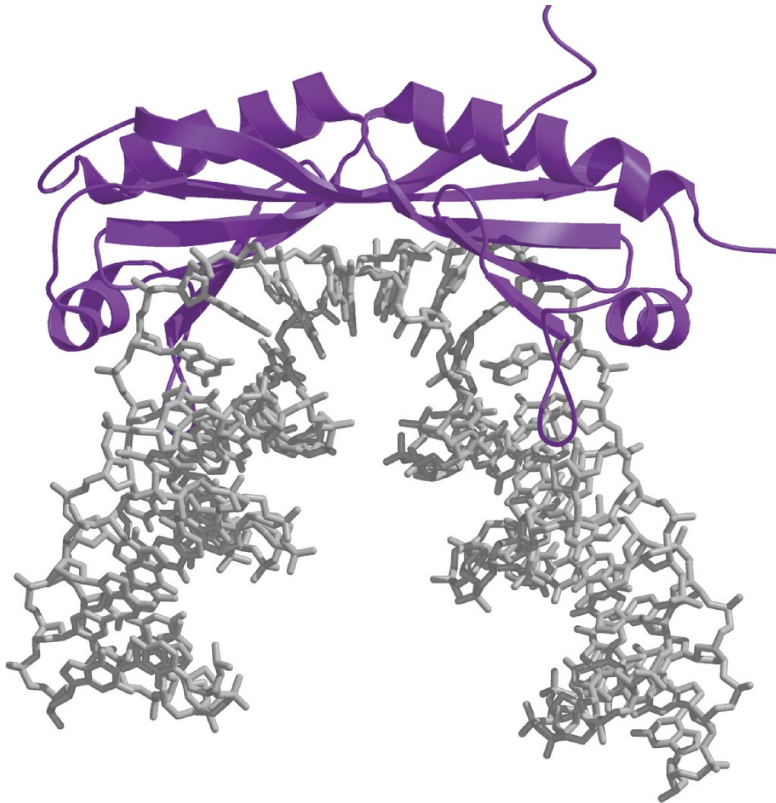
TBP-DNA complex

TBP-DNA + TFIIB complex

The TBP protein is a monomer with a symmetric axis of symmetry. It binds to the minor groove of DNA and has the ability to bend it, **creating an angle of approximately 80°.** This bending facilitates the binding of other factors to the promoter. For example, it allows TFIIB to interact with both DNA and TBP.
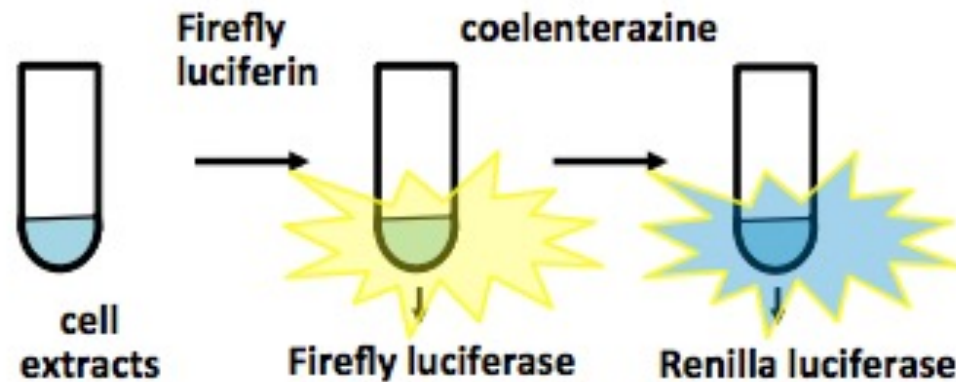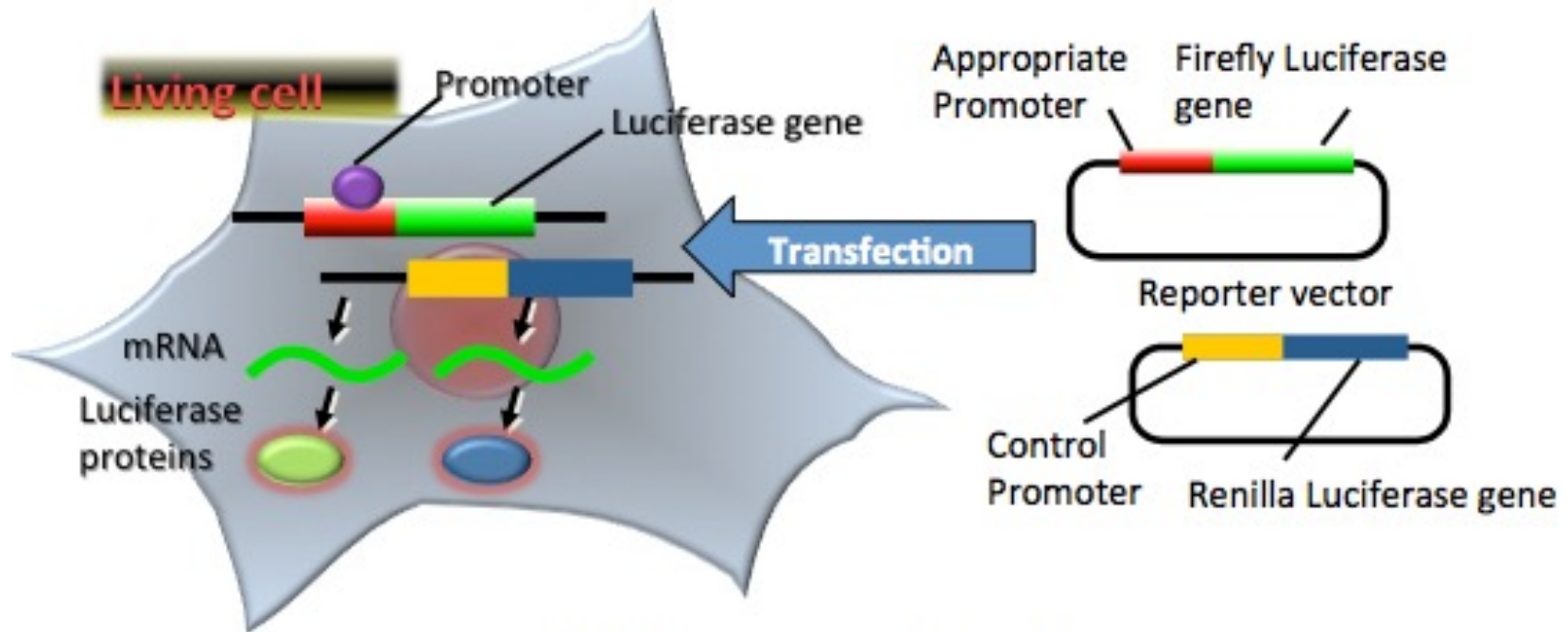
1. How do we identify a promoter?

2. How do we study its activity?

The Luciferase reporter assay

The luciferase reporter assay is commonly used as a tool to study gene expression at the transcriptional level. It is widely used because it is convenient, relatively inexpensive, and gives quantitative measurements instantaneously.
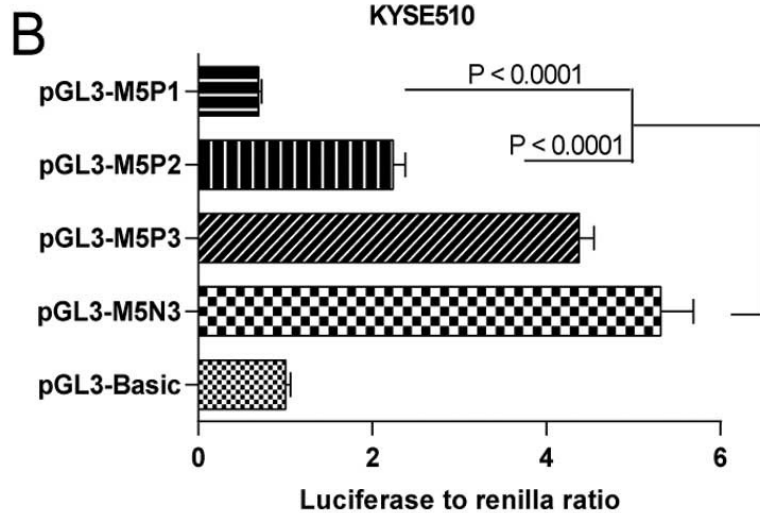
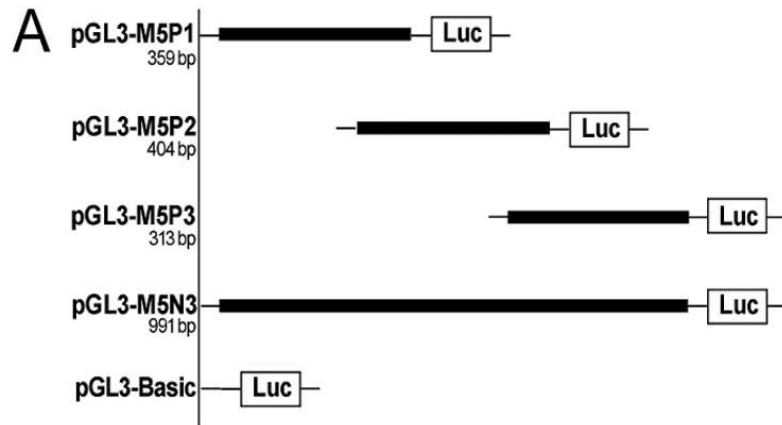$$\text{Luciferin} + O_2 + \text{ATP} \xrightarrow{\text{Luciferase}} \text{Oxyluciferin} + CO_2 + \text{AMP} + \text{PPi} + \text{LIGHT}$$
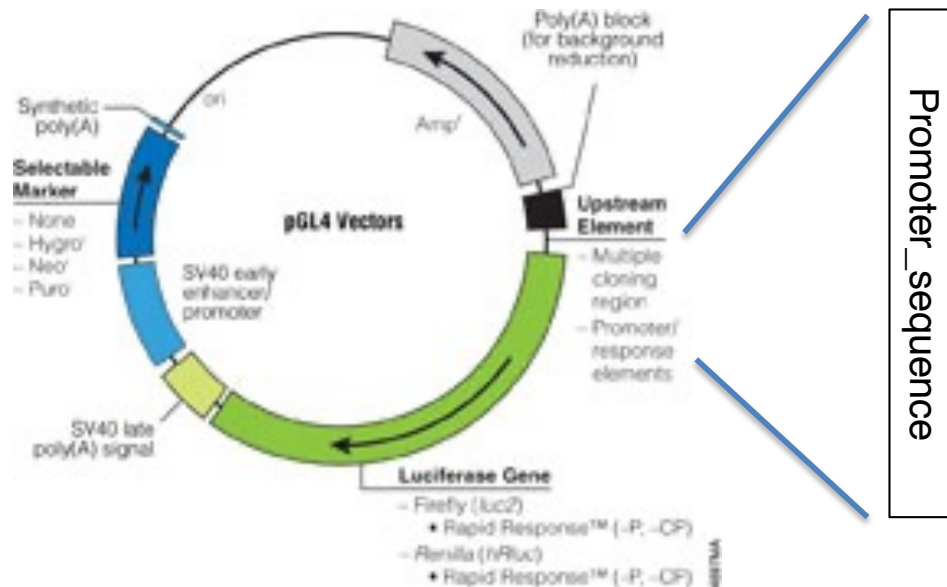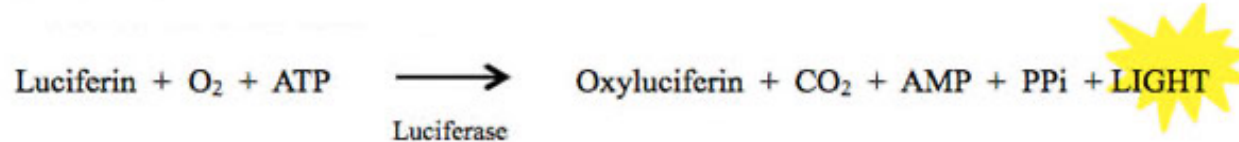
The Luciferase reporter assay

Promoter activity = Firefly luciferase activity / *Renilla* luciferase activity

The Luciferase reporter assay

A

pGL3-M5P1 359 bp — [Luc]

pGL3-M5P2 404 bp — [Luc]

pGL3-M5P3 313 bp — [Luc]

pGL3-M5N3 991 bp — [Luc]

pGL3-Basic [Luc]

B

KYSE510

pGL3-M5P1
pGL3-M5P2
pGL3-M5P3
pGL3-M5N3
pGL3-Basic

P < 0.0001
P < 0.0001

Luciferase to renilla ratio

C M5N3 - GenBank KF765385
chr2:202,122,236-202,123,227 (hg19)

GGGTCTAGGGCTCAGAGCTTTGGAGAACAGACCTCAGTAGCACCAACACTCCAGGAT
CAATGCTACAAAGACACGGGTTACAACTAAACTGGAGAACATGGCCAAGGATGGGAA
CTCAGCCTGAGCAGGGCTGAGCCGAGCAGGGCTAAGCCAAGTAGGGCTGAGCCAGAA
CACTTCCTCCTTTTTTCTGAACAATCTACCTACATTTCAGCTACAGGGCTGGCTTTAC
CCAGTCCGGCGGGAGGGAGGAGAGGGCTGGTCTGTGACTTCAGTGCTGAGGTTTGAT
CAAGGCAAAGGGAAACTTCCTATTCCCAGACCCTTTGCAAGAAAGAATGGCATATTA
CTTGCCACCGACAGGGGTTATTATTACTAAATGGAGTCAGTATAAATGCTTTCCAAT
AAAGCATGTCCAGCGCTCGGGCTTTAGTTTGCACGTCCATGAATTGTCTGCCACATCC
CTCTTCTGAATGGTTGGAAATTGGGCATCTGTTCCTTTAAACAGGAAACATTTCTTG
TTCGAGTGAGTCATCTCTGTTCTGCTTTAGGAGTAAAGTTTACCCTGCAGTTCCTTCT
GTGGTGAAGTTTTCTCTTTCTCTCGGAGACCAGATTCTGCCTTTCTGCTGGAGGGAA
GTGTTTTCACAGGTTCTCCTCCTTTTATCTTTTGTGTTTTTTTTCAAGCCCTGCTGAA
TTTGCTAGTCAACTCAACAGGAAGTGAGGCCATGGAGGGAGGCAGAAGAGCCAGGGT
GGTTATTGAAAGTAGAAGAAACTTCTTCCTGGGAGCCTTTCCCACCCCCTTCCCTGCT
GAGCACGTGGAGTTAGGCAGGTTAGGGGACTCGGAGACTGCGATGGTGCCAGGAAAG
GGTGGAGCGGGTGAGTGCCTGTTGCCAAGGTGGCCTCTTCAACAGGAAACCACAATA
TTTTTGTTTCTTGACTTGCTCTAGAAACAGGGCTGTGGGGGTGGGGAAGCAACTTGG
ATCTGCCCTTCTG

# Any promoter should be study in its genomic context!
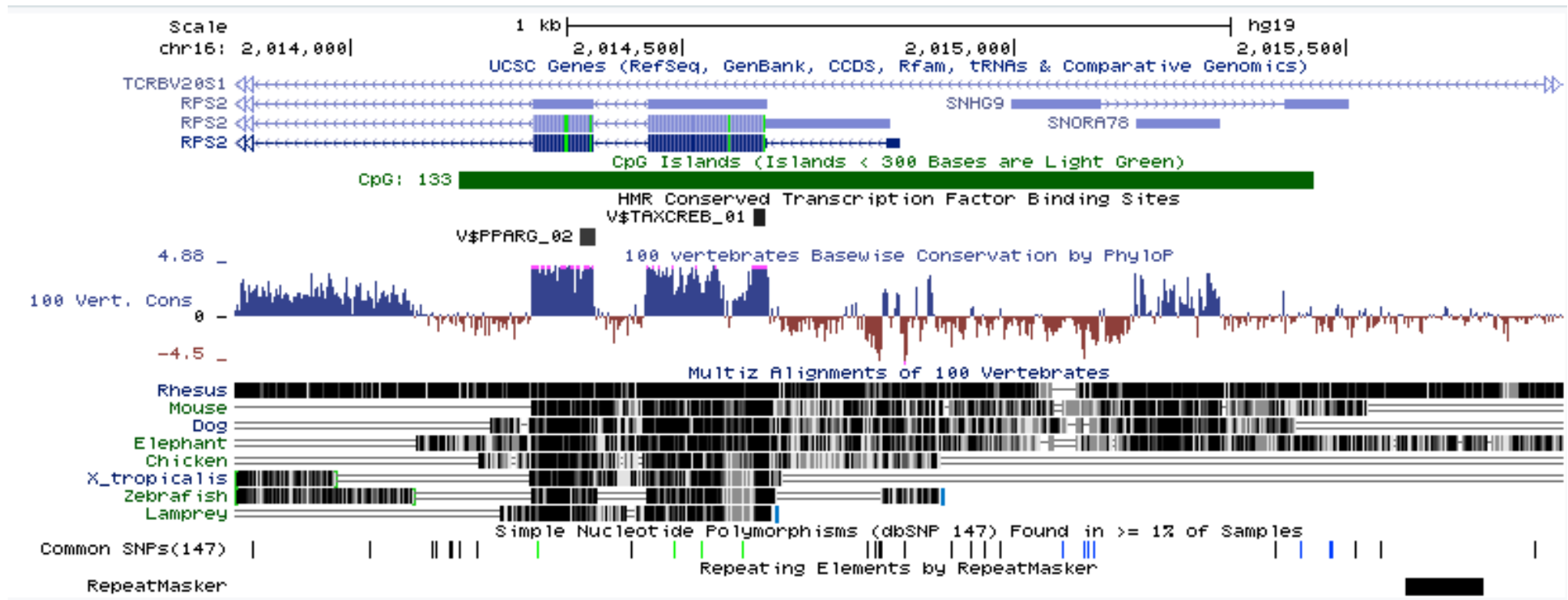
CAGE-seq for TSS indentification

Cap Analysis of Gene Expression (CAGE) is a high-throughput method for transcriptome analysis that utilizes *cap trapping* , a technique based on the biotinylation of the 7-methylguanosine cap of Pol II transcripts, to pulldown the 5'-complete cDNAs reversely transcribed from the captured transcripts.

A linker sequence is ligated to the 3' end of the cDNA and a specific restriction enzyme is used to cleave off a short fragment from the 5' end of the double stranded cDNA. Resulting fragments are then amplified and sequenced using massive parallel high-throughput sequencing technology, which results in a large number of short sequenced tags that can be mapped back to the referent genome to infer the exact position of the transcription start sites (TSSs) used for transcription of captured RNAs
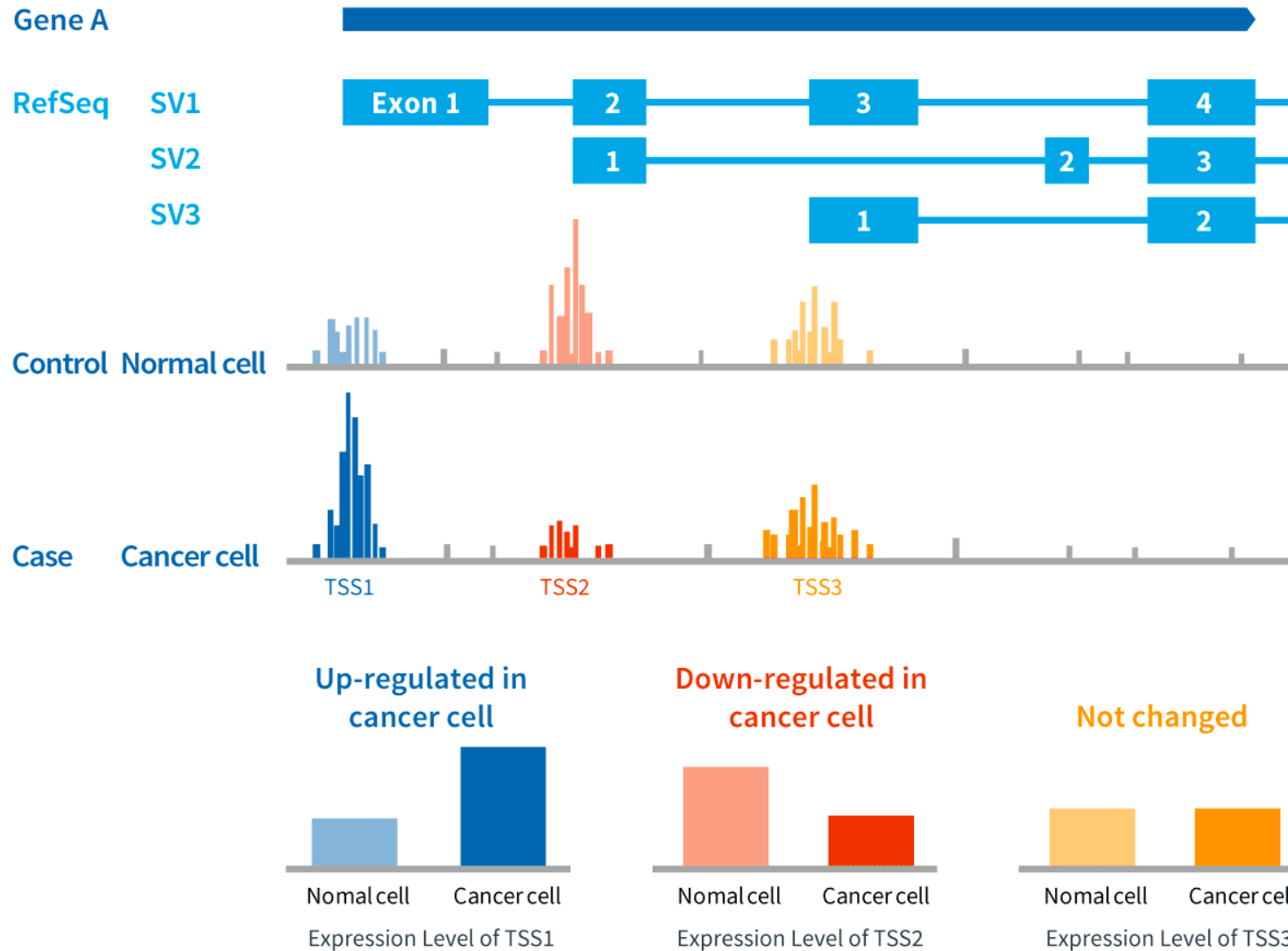
The number of CAGE tags supporting each TSS gives the information on the relative frequency of its usage and can be used as a measure of expression from that specific TSS.
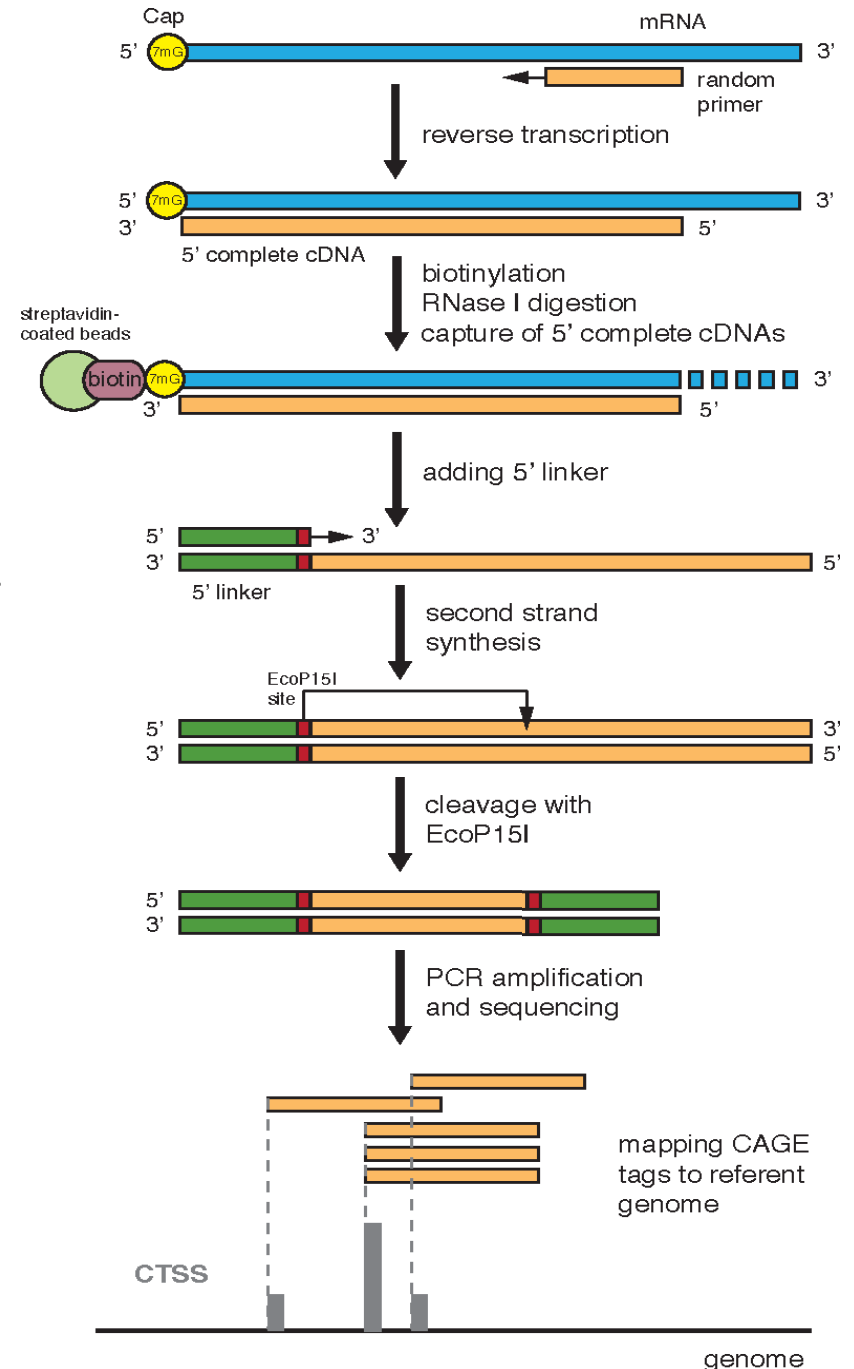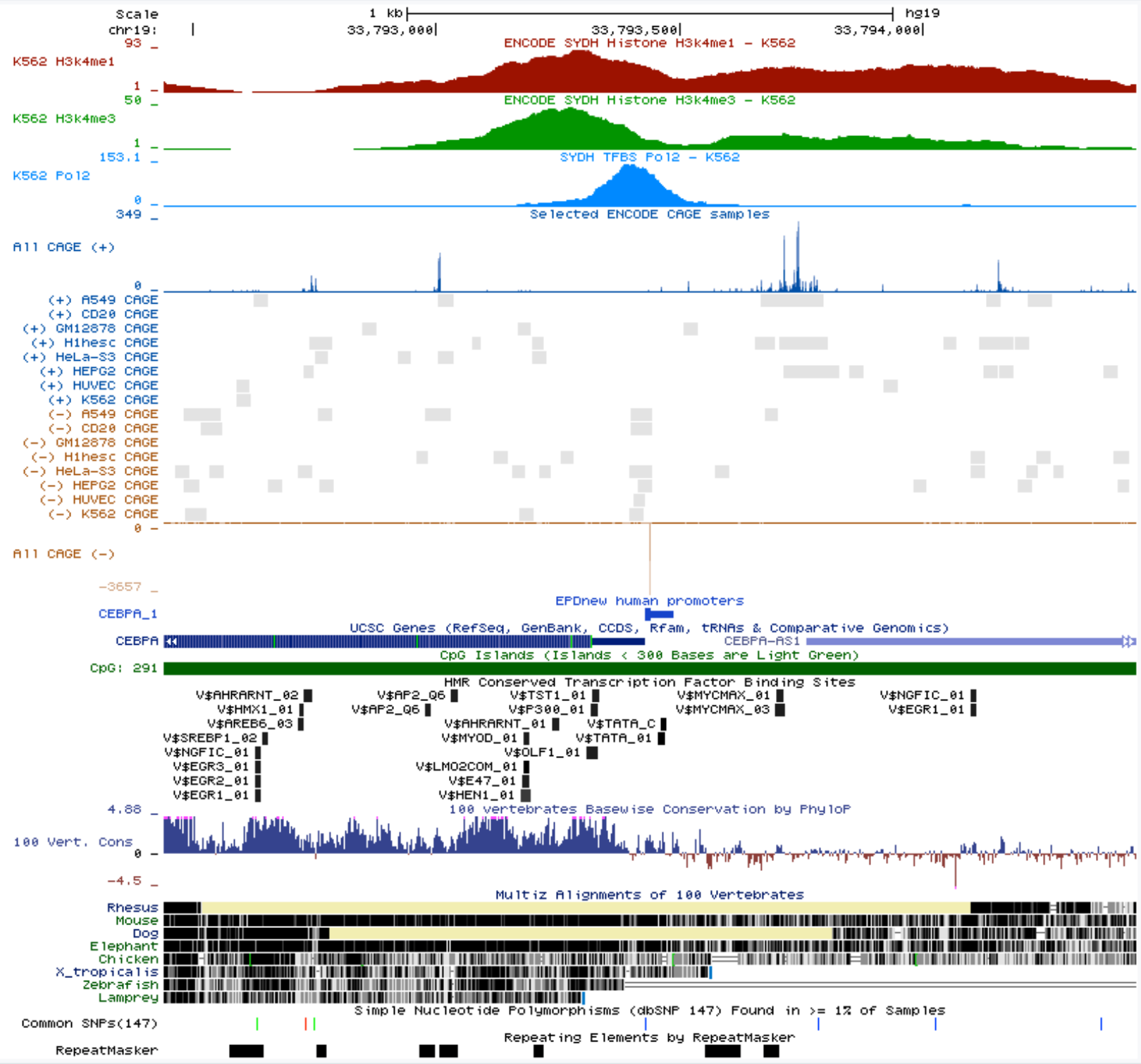
Gene Expression Analysis - Alternative Transcription Start Sites -

e.g.) Detect the expression level of the whole genome of normal cells and cancer cells by CAGE

Comprehensively detect the differentially expressed transcript at each TSS

Gene A

RefSeq

**SV1** — Exon 1 — 2 — 3 — 4

SV2 — 1 — 2 — 3

SV3 — 1 — 2

Control **Normal cell**

Case **Cancer cell**

TSS1   TSS2   TSS3

Up-regulated in cancer cell

Nomal cell   Cancer cell

Expression Level of TSS1

Down-regulated in cancer cell

Nomal cell   Cancer cell

Expression Level of TSS2

Not changed

Nomal cell   Cancer cell

Expression Level of TSS3

Promoter types in Metazoa

Promoter type	Dominant gene function	Common properties	Vertebrate-specific
Major promoters			
Type I ('adult')	Tissue-specific expression in adult peripheral tissues	Sharp ('focused') TSS, TATA-box enrichment, disordered nucleosomes	Mostly no CpG islands
Type II ('ubiquitous')	Broad expression throughout organismal cycle	Broad ('dispersed') TSS, ordered nucleosome configuration	CpG islands, TATA-depleted
Type III ('developmentally regulated')	Differentially regulated genes, often regulators in multicellular development and differentiation	Polycomb repression-regulated genes, broad H3K27me3 marks	Large CpG islands extending into the body of gene
Minor promoters			
TCT promoter	Highly expressed genes of translational apparatus	Sharp, pyrimidine-stretch ('TCT') initiator sequence, often full TATA box, ubiquitous-promoter-like nucleosome configuration	CpG island overlapping

DPE, downstream promoter element; DRE, DNA recognition element; H3K27me3, histone H3 lysine 27 trimethylation; TSS, transcription start site.

Promoter types in Metazoa

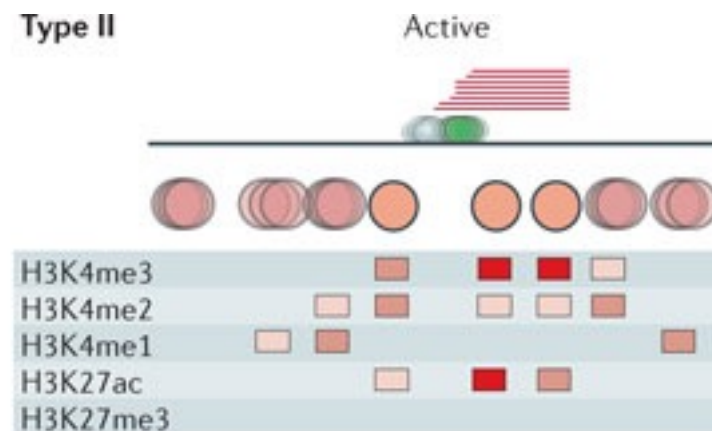| Promoter type	Dominant gene function	Common properties	Vertebrate-specific
Major promoters			
Type I ('adult')	Tissue-specific expression in adult peripheral tissues	Sharp ('focused') TSS, TATA-box enrichment, disordered nucleosomes	Mostly no CpG islands

Genes that are specifically expressed in peripheral terminally differentiated tissues (e.g. liver or skeletal muscle) have **type I promoters** with a pattern of histone modification that is distinct from that of most other genes. H3K4me3 is generally only present downstream of the TSS, and there is no RNAPII binding at these promoters when the genes are not active.
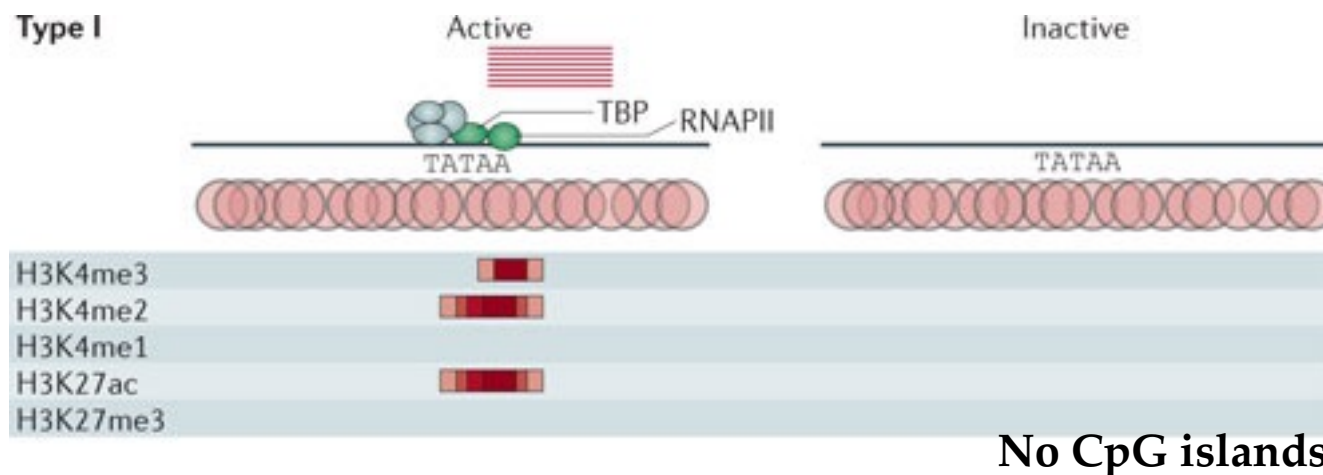
No CpG islands

Promoter types in Metazoa

Promoter type	Dominant gene function	Common properties	Vertebrate-specific
Type II ('ubiquitous')	Broad expression throughout organismal cycle	Broad ('dispersed') TSS, ordered nucleosome configuration	CpG islands, TATA-depleted

Ubiquitously expressed genes have H3K4me3 throughout their **type II promoters** across all tissues. Across classes of genes in vertebrates, H3K4me3 distribution is almost identical with the span of CpG islands. Ubiquitously expressed genes generally have <u>short CpG islands</u>, and the H3K4me3 mark and CpG island typically only overlap the 5′ end of the gene.
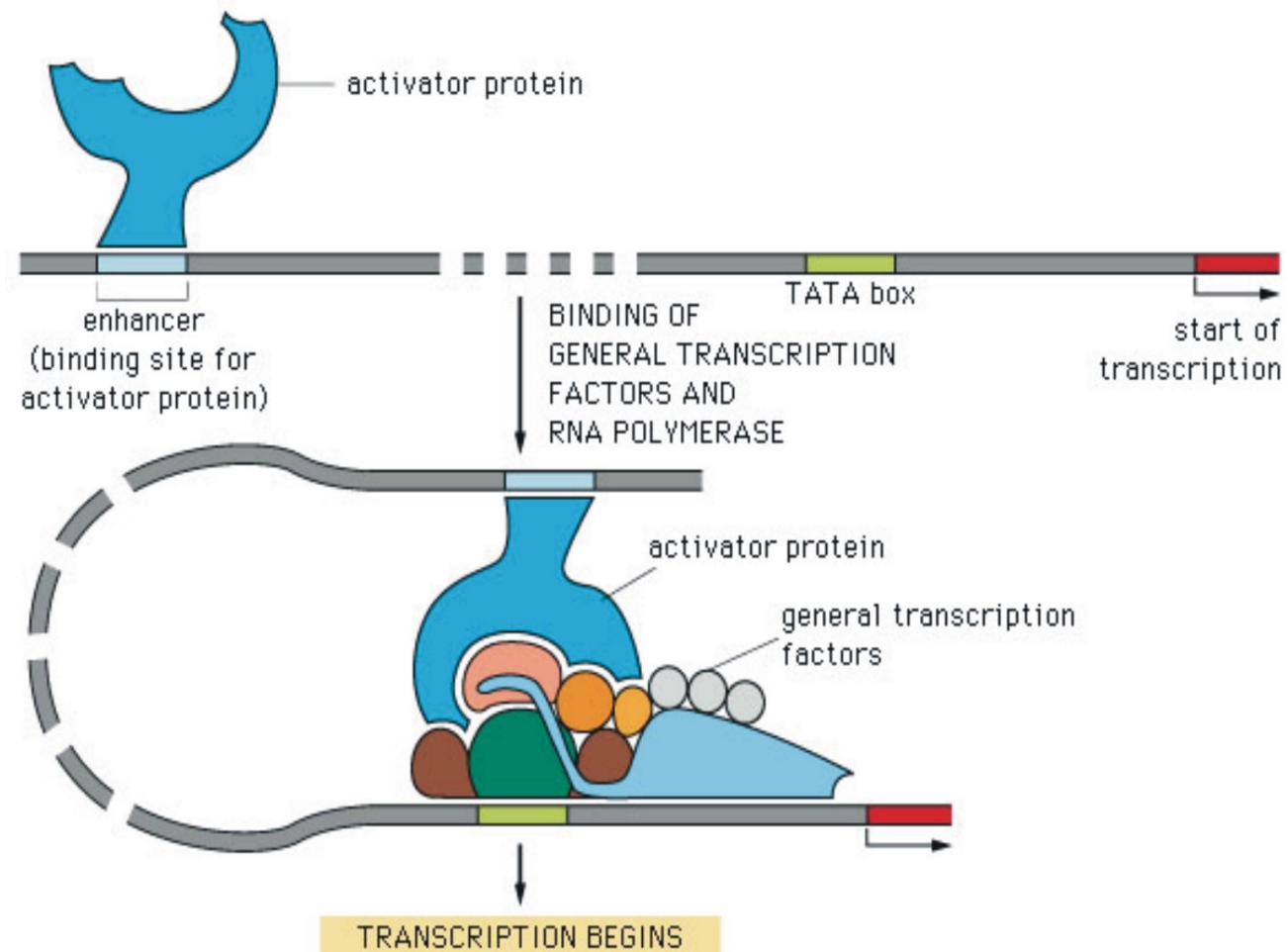

Short CpG islands

Promoter types in Metazoa

Promoter type	Dominant gene function	Common properties	Vertebrate-specific
Type III ('developmentally regulated')	Differentially regulated genes, often regulators in multicellular development and differentiation	Polycomb repression-regulated genes, broad H3K27me3 marks	Large CpG islands extending into the body of gene

Developmentally regulated genes (with **type III promoters**) have a number of features that are associated with repression by **polycomb group proteins (PcG)**. These features include multiple large CpG islands, wide distribution of bound PcG proteins and both H3K27me3 and H3K4me3. Because of the presence of both of these marks, which are associated with repression and activation, respectively, these are described as **bivalent promoters**. The large CpG islands often extend into the body of the genes and are closely tracked by H3K4me3, which is thus not restricted to promoter regions in developmental genes.


Large CpG islands
Long range enhancer

Nature Reviews | Genetics
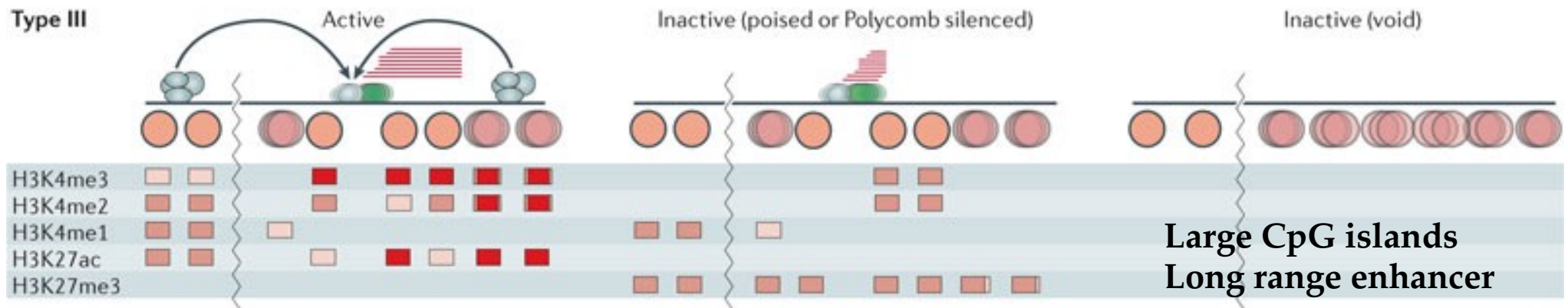
Enhancers

Enhancer is a short (50-1500 bp) region of DNA that can be bound by proteins (transcriptional activators) to increase transcription of a particular gene.

activator protein

enhancer
(binding site for
activator protein)

BINDING OF
GENERAL TRANSCRIPTION
FACTORS AND
RNA POLYMERASE

TATA box

start of
transcription

activator protein

general transcription
factors

TRANSCRIPTION BEGINS

©1998 GARLAND PUBLISHING

Enhancers

Genomic regions that function as transcriptional enhancers are enriched in closely spaced recognition motifs for sequence-specific transcription factors. Cell type-specific enhancers are marked by specific epigenomic features and chromatin accessibility. In particular, enhancers display <u>enrichment of H3K4me1 or H3K4me2</u> and depletion of H3K4me3 compared with promoters
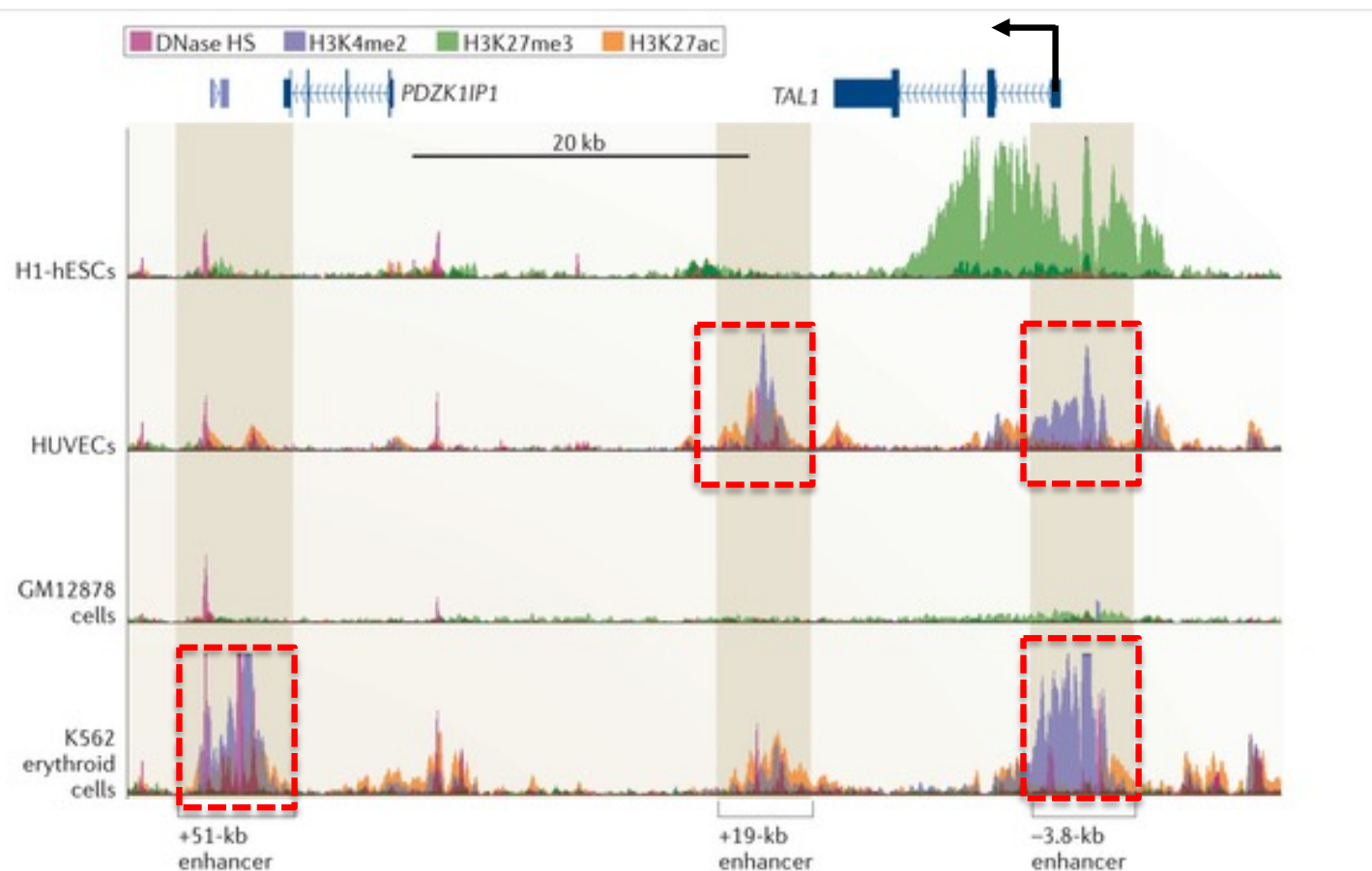
It is estimated that the human genome contains approximately 150,000 enhancers, but 'only' about 10,000–50,000 of them are active in any given cell type

Enhancers

Enhancer activation begins with the binding of transcription factors and local nucleosome remodelling. Enhancer states can broadly be classified as:
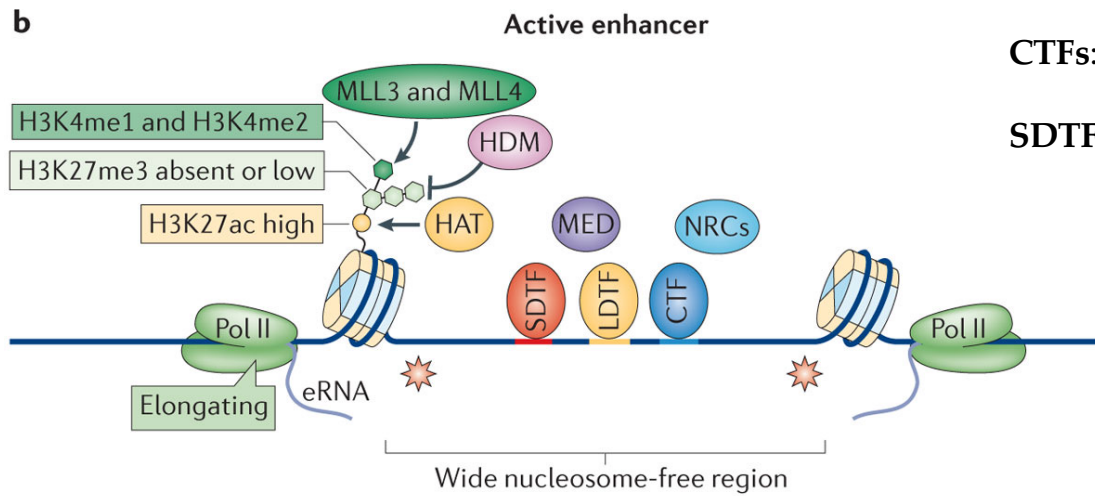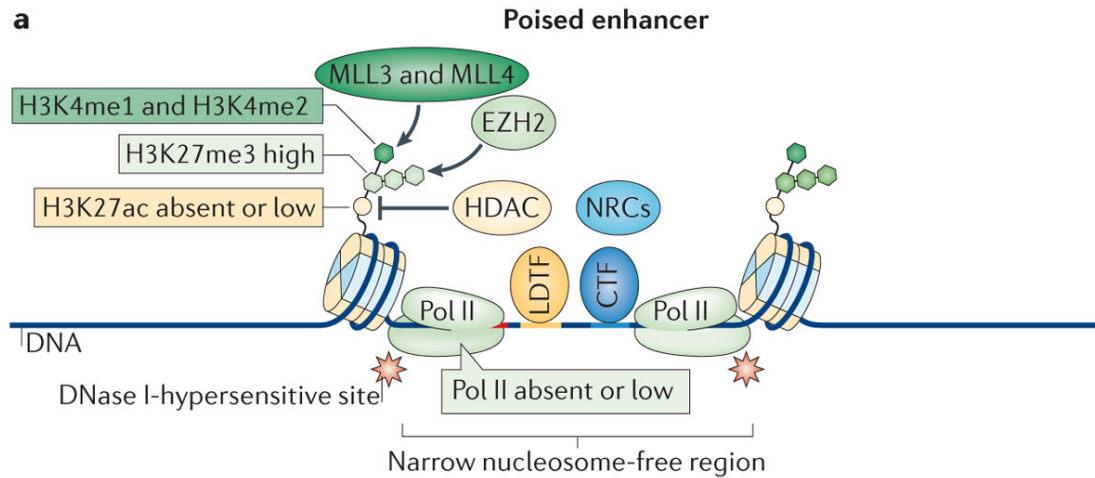
Inactive enhancers: are essentially buried in <u>compact chromatin</u> and are <u>devoid of transcription factor binding</u> and <u>histone modifications</u>.

Primed/Poised enhancers: Primed enhancer are characterized by <u>closely bound sequence-specific transcription factors</u> that establish a nucleosome-free region of <u>open chromatin</u>. However, they may require additional cues to accomplish their function, which may include signal-dependent activation, the recruitment of additional transcription factors and the eventual recruitment of co-activators that lead to enhancer activation. **Poised enhancers** can be defined as primed enhancers that also contain <u>repressive epigenetic chromatin marks</u>

Active enhancers: present co-activator complexes containing histone demethylase (HDM) complexes that remove H3K27me3 marks, histone acetyltransferases (HATs) that deposit H3K27 acetylation (H3K27ac) marks, and the **Mediator complex (MED)**. The transformation to elongating Pol II results in bidirectional transcription — a hallmark of active enhancers — and the generation of **enhancer RNAs (eRNAs)**, which is closely coupled to enhancer activity.

Importantly, enhancers are mobile in 3D space, and therefore have the potential to interact with multiple promoters over time

Enhancers

a **Poised enhancer**

MLL3 and MLL4

H3K4me1 and H3K4me2

EZH2

H3K27me3 high

H3K27ac absent or low

HDAC

NRCs

LDTF

CTF

Pol II

Pol II

DNA

DNase I-hypersensitive site

Pol II absent or low

Narrow nucleosome-free region

b **Active enhancer**

MLL3 and MLL4

H3K4me1 and H3K4me2

HDM

H3K27me3 absent or low

H3K27ac high

HAT

MED

NRCs

SDTF

LDTF

CTF

Pol II

Pol II

Elongating

eRNA

Wide nucleosome-free region

LDTFs: lineage determining transcription factors

CTFs: collaborating transcription factors

SDTFs: signal-dependent transcription factors

Cell type-specific enhancer selection and activation

lineage-determining transcription factors (LDTFs)
collaborating transcription factors (CTFs)
signal-dependent transcription factors (SDTFs).
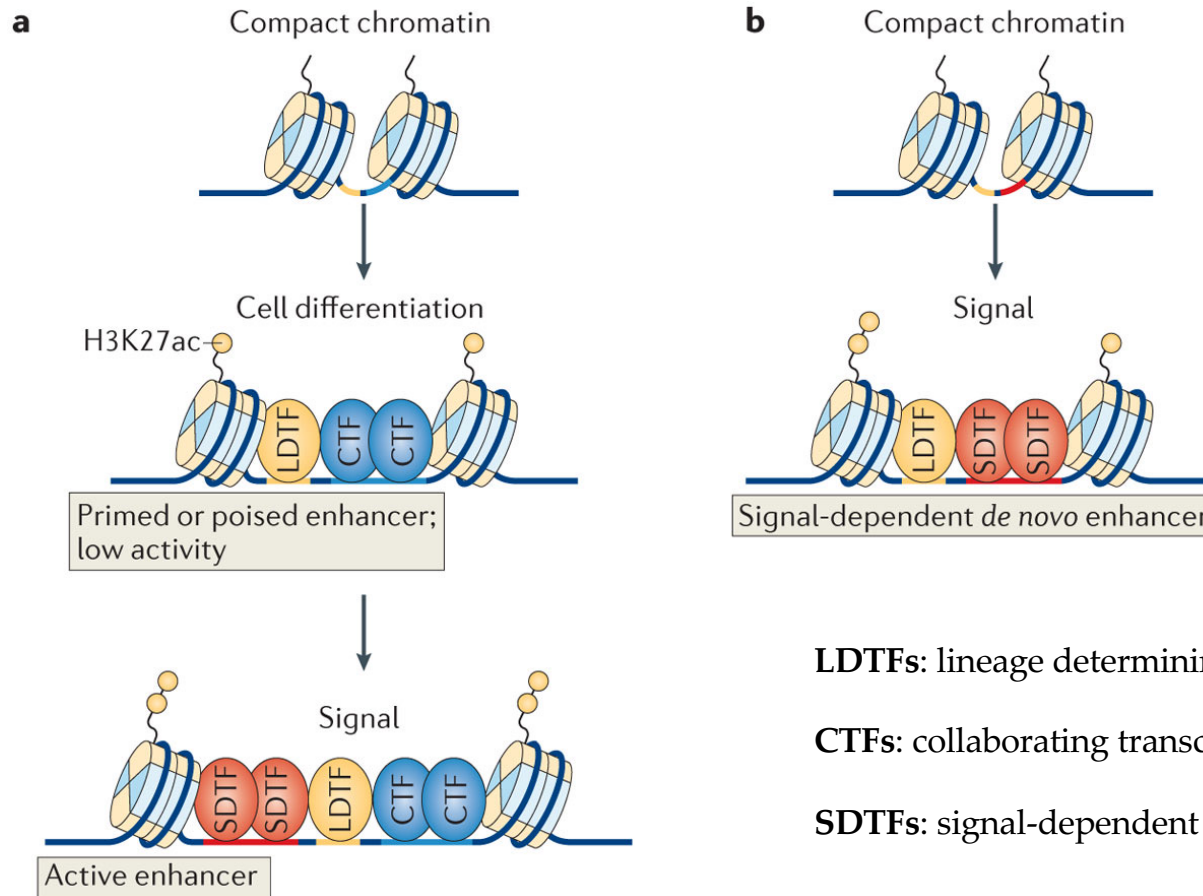
LDTFs: lineage determining transcription factors

CTFs: collaborating transcription factors

SDTFs: signal-dependent transcription factors

Widespread transcription of mammalian genome


The traditional view of transcription initiation

The traditional view of gene expression and the role of enhancers and core promoters has been challenged by the observation that many genomic positions outside annotated gene starts initiate transcription, including positions within enhancers.
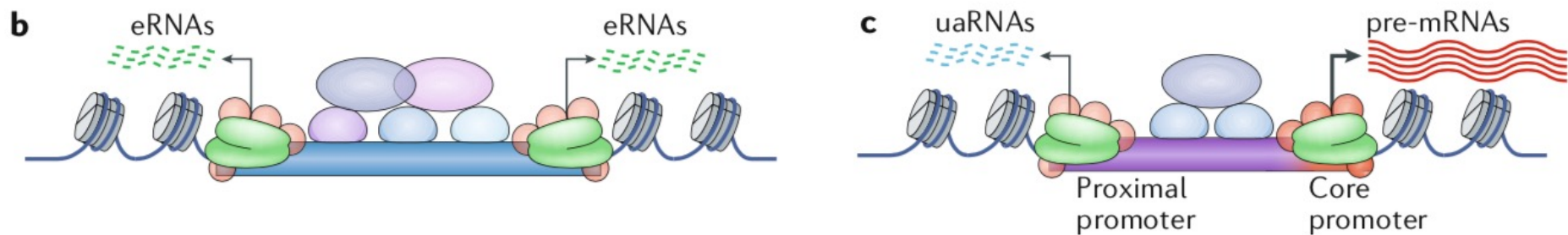
**The modern view of transcription initiation**

Divergent transcription from promoter and enhancer produces non coding RNAs

Most mammalian promoters lack a TATA element (TATA-less). For these promoters, TBP is recruited through sequence-specific transcription factors (such as Sp1) and components of the TFIID complex that have little sequence specificity. Thus, in the absence of strong TATA elements, TBP-complexes are recruited on both sides of the transcription factors to form preinitiation complexes in both orientations. Since transcription factors with transcription activation domains also bind at <u>enhancer sites</u>, these are also sites of divergent transcription.

Wu and Sharp, 2013 Cell
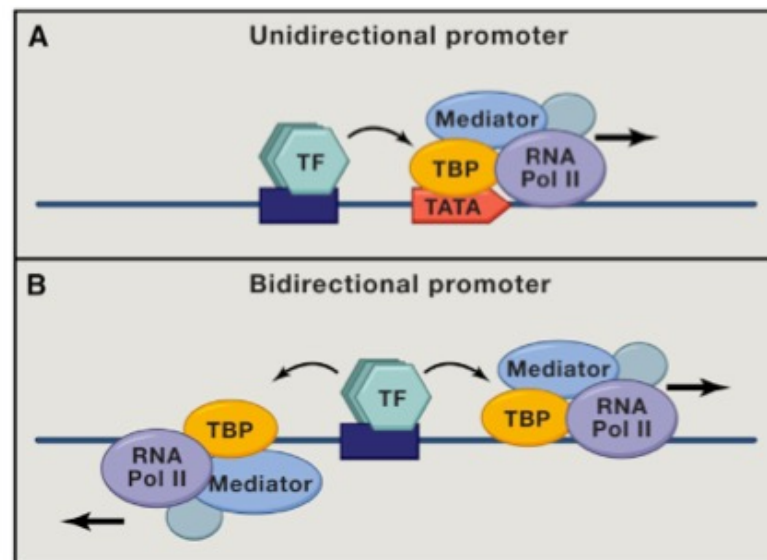
Enhancer RNAs (eRNAs)

- eRNAs are produced from transcriptional enhancers

- eRNA transcription correlates with the activation of the neighboring protein-coding genes (*cis* –acting)

- The majority of eRNAs are reported to be monoexonic and not polyadenylated (2–5 kb in length), but few cases of spliced and polyadenylated eRNAs have been described.
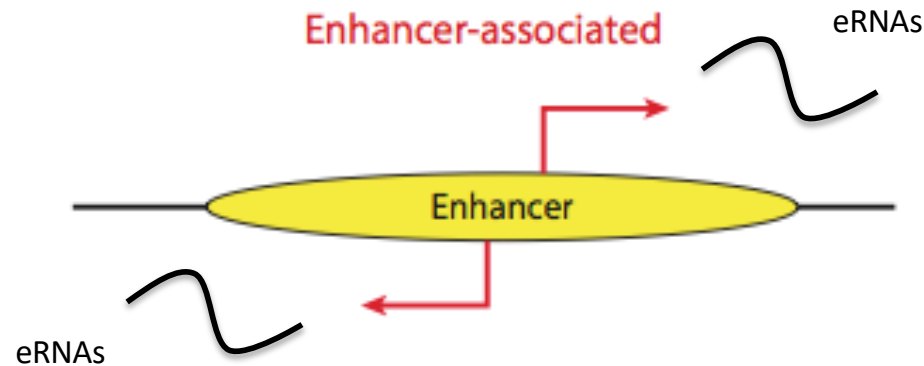
Enhancer RNAs

eRNAs **cooperate with the the Mediator and Cohesin complexes (A)** to promote enhancer–promoter interactions (chromatin looping) or they **recruit chromatin remodeling (B) or chromatin modifying complexes (C)** to the targeted promoters.
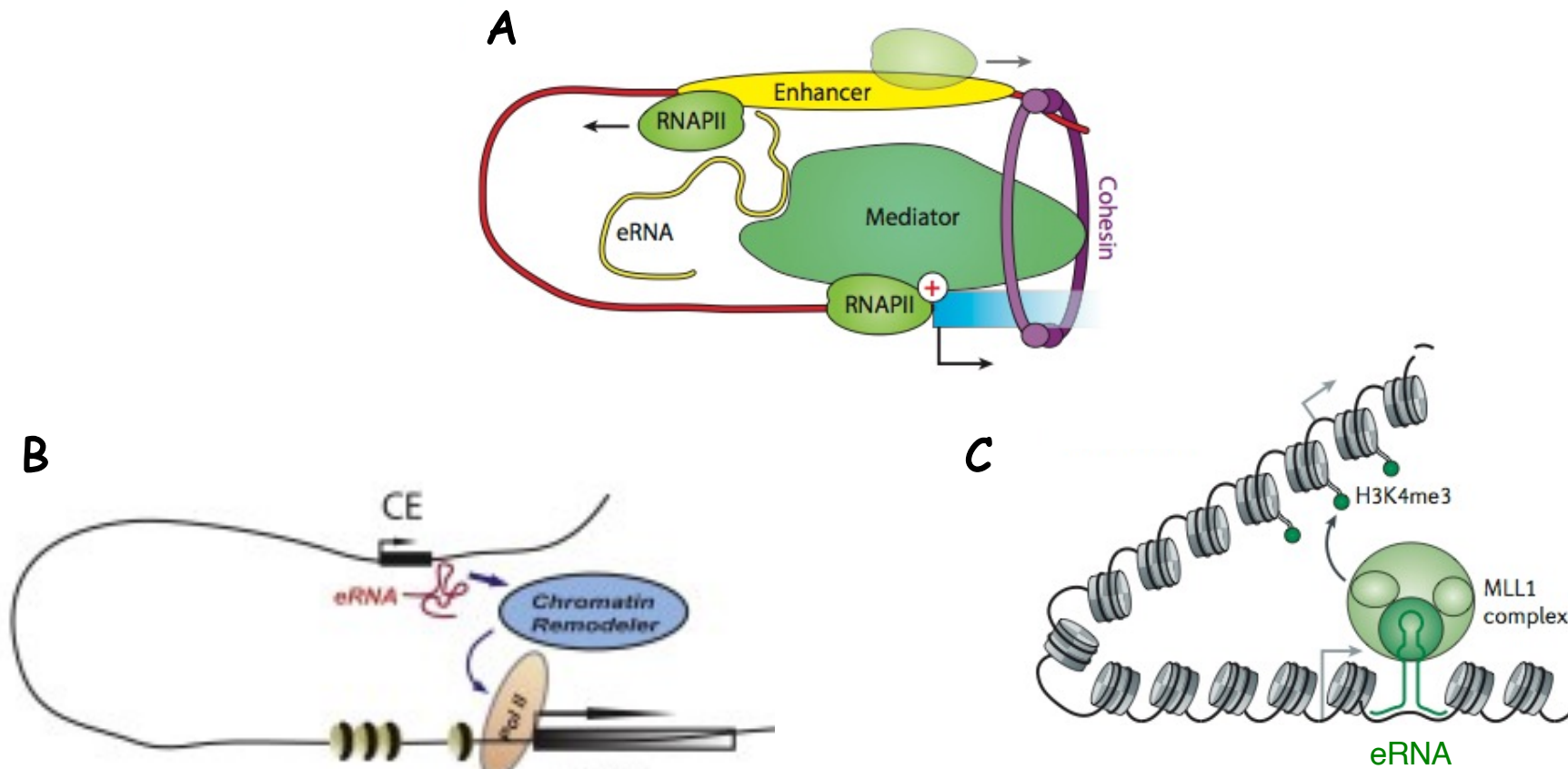
eRNA Binding to CBP Stimulates Histone Acetylation and Transcription

CBP/p300 are histone acetyltransferase (HAT) whose binding is a signature of enhancers.
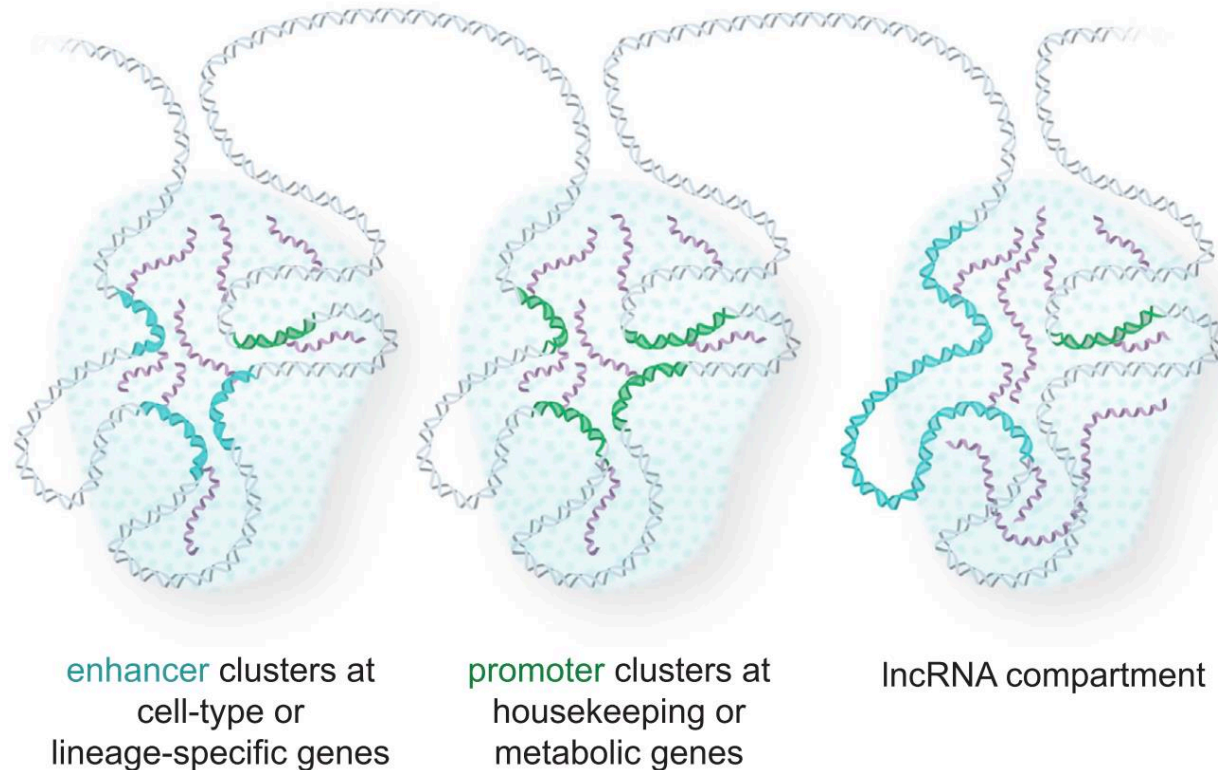
eRNA binding region in the HAT domain of CBP allows RNA to stimulate CBP's HAT activity.

At enhancers where CBP interacts with eRNAs, stimulation manifests in RNA-dependent changes in the histone acetylation mediated by CBP, such as H3K27ac, and by corresponding changes in gene expression.

By interacting directly with CBP, eRNAs contribute to the unique chromatin structure at active enhancers, which, in turn, is required for regulation of target genes.

RNA binding stimulates CBP HAT activity, histone acetylation and transcription

Bose et al., 2017. Cell
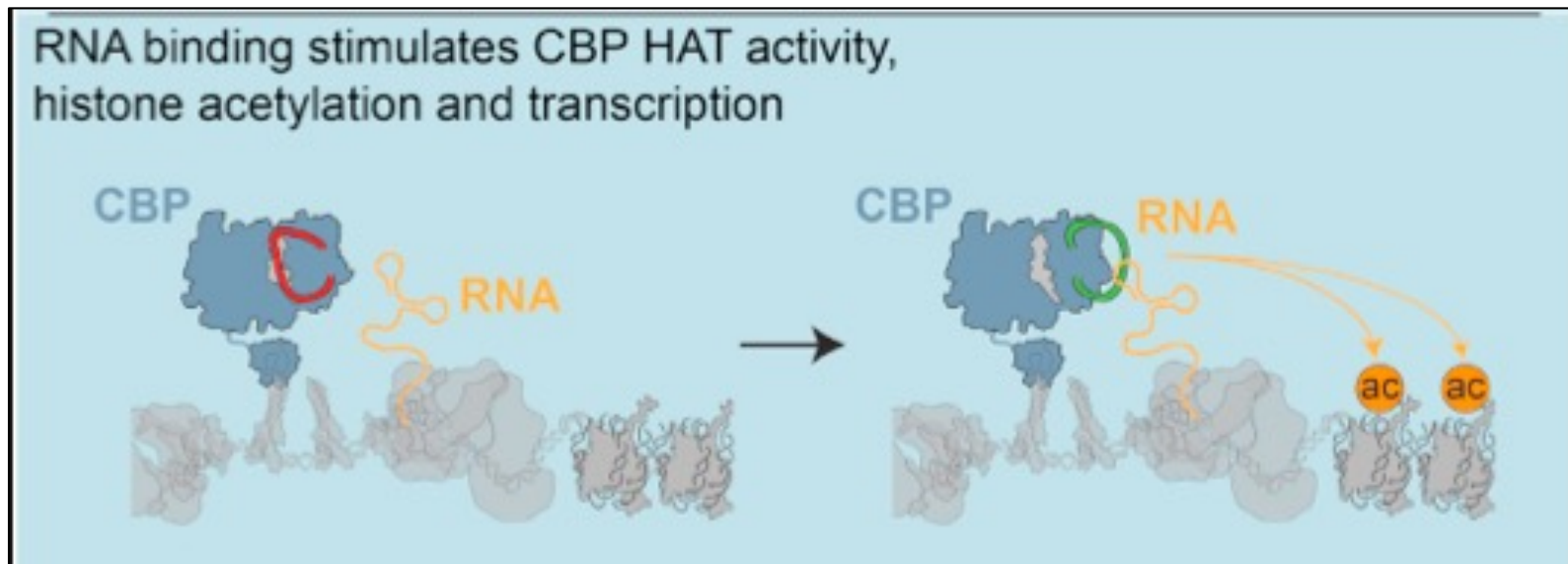
Three-dimesional spatial compartments in transcription

Active enhancers and promoters assemble large amounts of transcription apparatus, forming transcriptional compartments that have been called **condensates**. Many condensates harbor RNA components and these can influence the formation and stability of the assembly.

enhancer clusters at cell-type or lineage-specific genes

promoter clusters at housekeeping or metabolic genes

lncRNA compartment

The mediator complex

Mediator is a multiprotein complex that is implicated in regulating many fundamental processes involved in transcription including:

- **transcription initiation**,

- **transcription elongation**,

- **chromatin architecture**,

- **enhancer-promoter gene looping**

The mediator complex

The human Mediator complex has <u>30 subunits and is 1.2 MDa in size.</u>

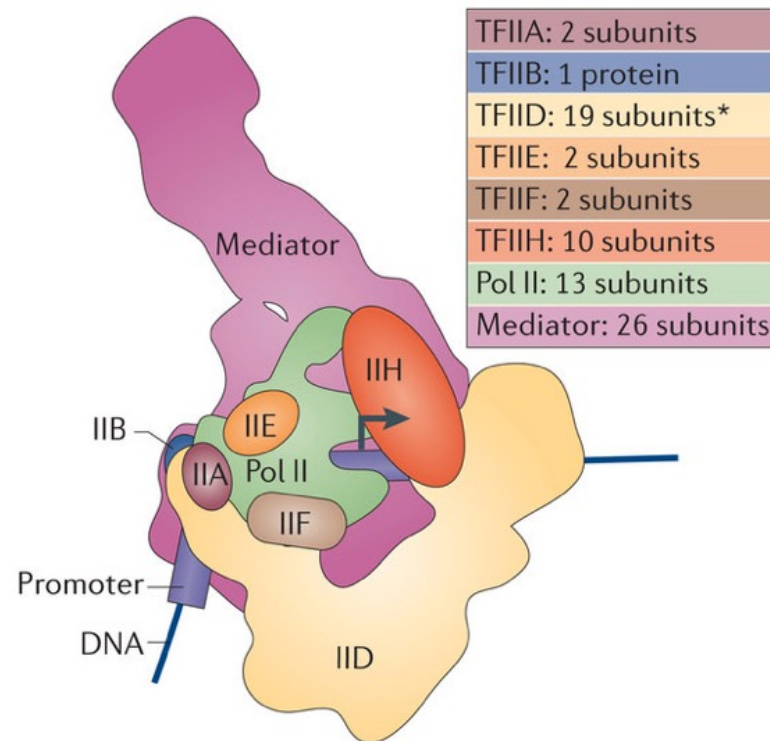Mediator is universally required for the expression of almost all genes

A fundamentally important characteristic of Mediator is that its subunit composition can change; subunits can be lost or added to affect its biological function. individual subunits reside in one of four distinct modules designated '**head**', '**middle**', '**tail** and '**kinase'**

Whereas the head, middle and tail modules form a relatively stable 'core' structure, the <u>kinase module associates reversibly </u>with the Mediator complex

Head
Middle
MED1 or MED26
Tail
Kinase

MED11
MED22
MED20
MED17
MED13
MED12
MED18
MED30
MED19
MED28
MED8
MED6
CycC
MED21
MED4
MED9
CDK8
MED7
MED31
MED10
MED1
MED2 (MED29)
MED3 (MED27)
MED26
MED25
MED15
MED5 (MED24)
MED23
MED14
MED16

VP16, p53, HSF
REST, NANOG, β–catenin
Nuclear receptors: TR, VDR, PPARγ, HNF4, ER, GR
SREBP1, NHR49, TGFβ, nodal
Nuclear receptors: GR, HNF4, PPARγ
VP16
E1A, ELK1

Nature Reviews | Genetics

Mediator has key roles in the assembly of the PIC

For gene-specific transcription, Pol II must be recruited to specific sites on the genome. This is generally controlled by sequence-specific, DNA-binding TFs. Although <u>TFs do not directly bind to Pol II, one mechanism by which they can promote Pol II recruitment is by binding to the Mediator complex</u>. Mediator enables Pol II recruitment via interaction with the CTD of the Pol II. The large size of Mediator is likely to promote stable PIC formation by allowing the complex to directly interact with multiple PIC factors. Moreover, Mediator helps to regulate the recruitment and/or the activity of the PIC components.

TFIIA: 2 subunits	
TFIIB: 1 protein	
TFIID: 19 subunits*	
TFIIE: 2 subunits	
TFIIF: 2 subunits	
TFIIH: 10 subunits	
Pol II: 13 subunits	
Mediator: 26 subunits	

Model for the Mediator-dependent transcriptional activation pathway

Mediator function is proposed to be manifested at two distinct levels. (1) First, in concert with the activators, **Mediator** might promote PIC assembly (**recruitment**). (2) Second, following structural or topological conversion, Mediator could modulate the (basal) function of pol II in the PIC (**post-recruitment**).
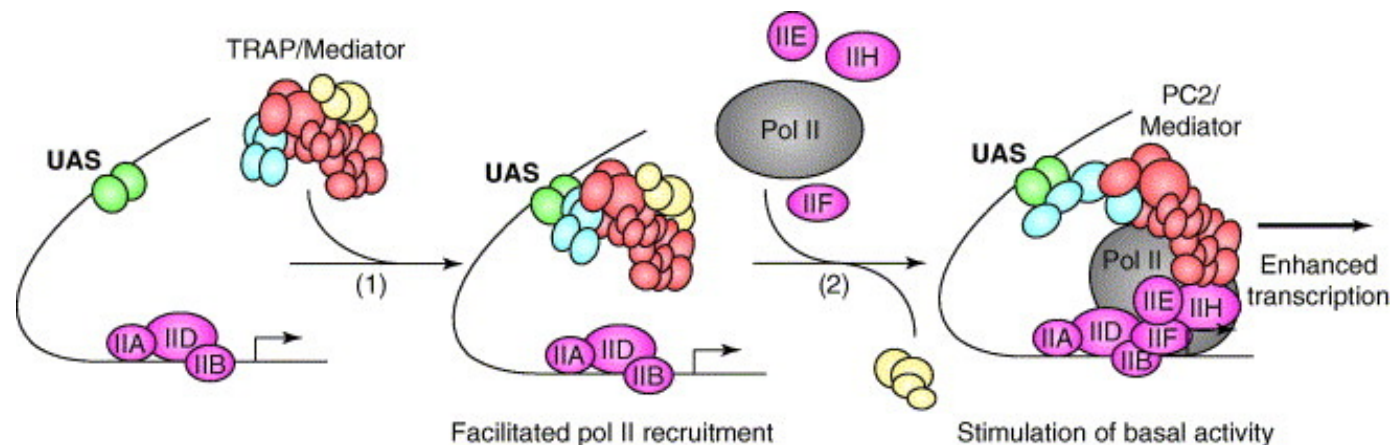
By inserting itself between the activators and the general transcription factors, the Mediator affords additional opportunities to fine-tune the diverse regulatory inputs received both from the DNA-binding factors and, most likely, from other signals and to present an <u>appropriately calibrated output to the pol II machinery</u>.

Mediator and chromatin architecture

The 3D organization of the genome has basic roles in coordinating gene expression programmes. A common theme is the formation of **DNA loops** that enable interaction between linearly separated DNA sequences. <u>Mediator has a central role in the formation and/or stabilization of looped DNA structures in eukaryotes.</u> In mammalian cells, long-range looping interactions between enhancer and promoter sequences appear to be important for driving high-level and cell type-specific gene expression.
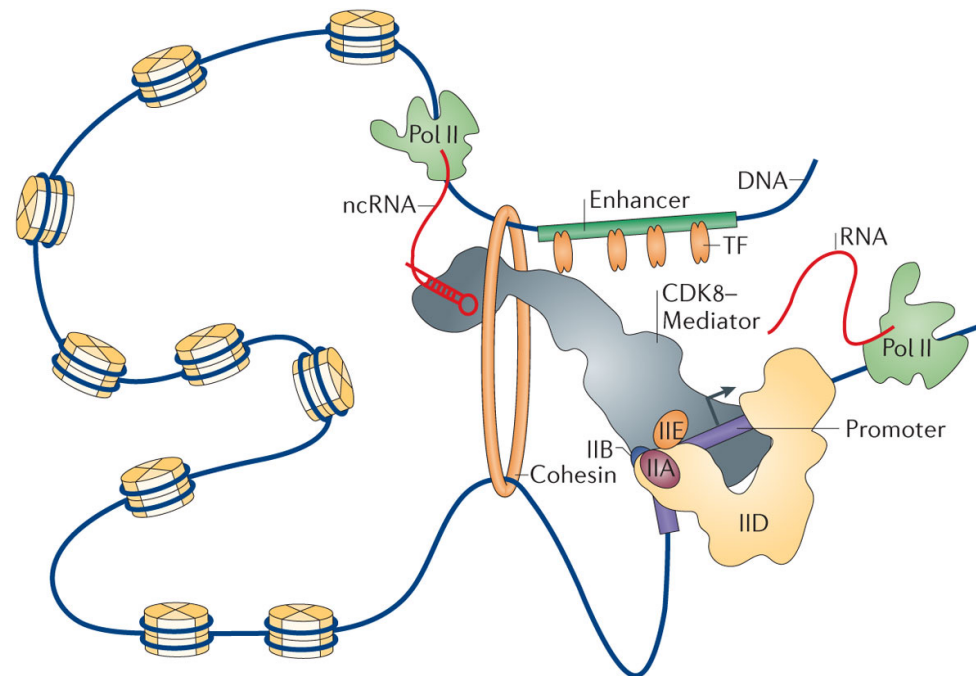
Transcription bursting

A basic requirement for bursting is reinitiation, which may be facilitated by a PIC scaffold complex that remains at the promoter after the initial round of Pol II transcription.
The scaffold PIC retained **TFIIA, TFIID, TFIIE, TFIIH and Mediator** — that is, Pol II, TFIIF and TFIIB were missing. Furthermore, the PIC scaffold was stabilized by a TF, and TF binding resulted in a higher rate of transcription reinitiation
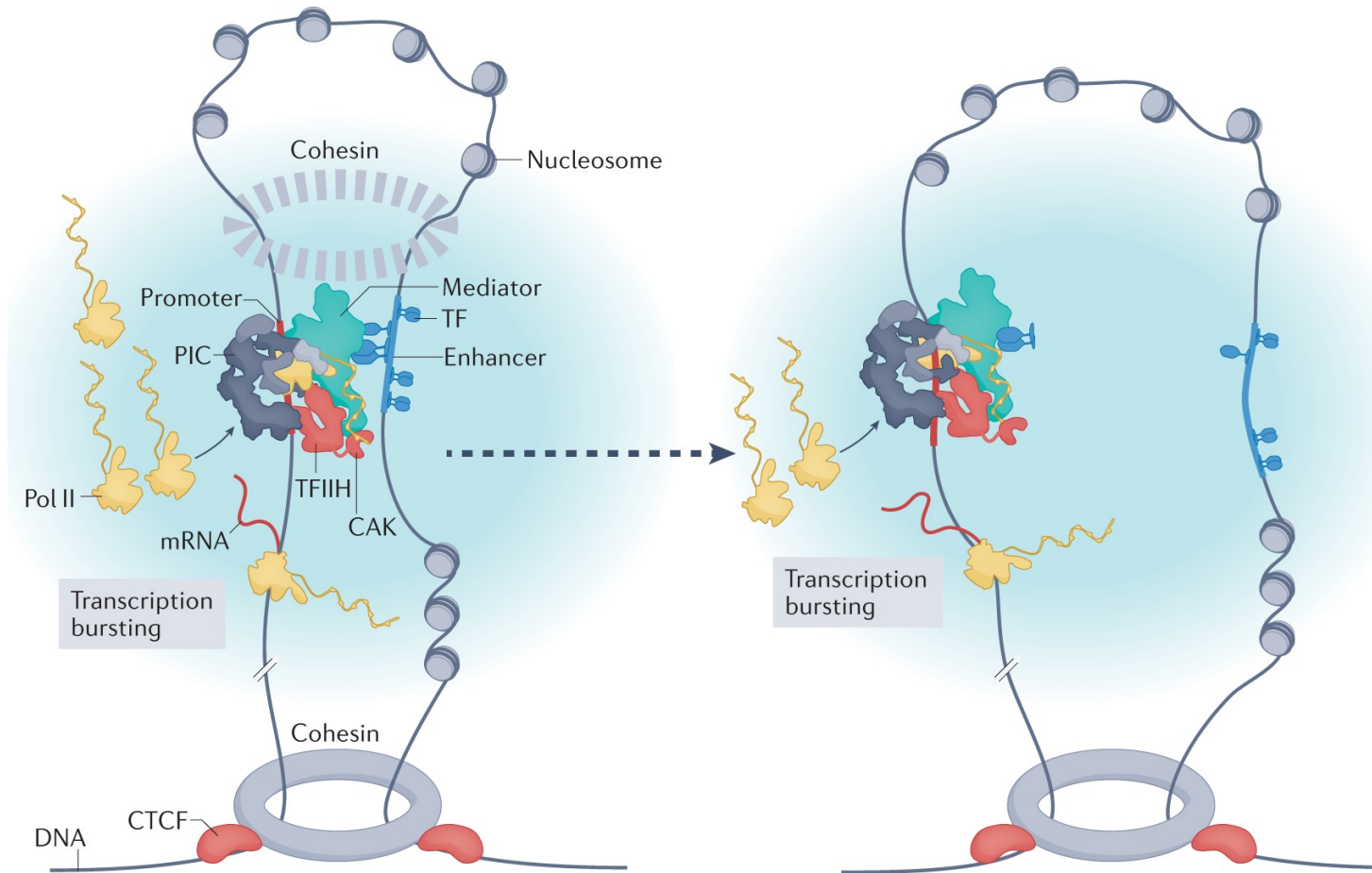
In mammalian cells, genomic regions of 1 Mb or more in size are organized into topologically associating domains (TADs), which are formed and maintained by CTCF and cohesin

The mediator complex can also contribute to epigenetic silencing

Mediator can contribute to **long-term silencing** of neuronal genes in extraneuronal cells. Starting with an active **PIC**, **REST** binding to the cognate site commits the gene to a heterochromatin fate as development proceeds. A ternary complex containing REST, the histone methyltransferase G9a and the intact Mediator is first assembled. The multipartite interactions are anchored through **the MED12 subunit of the kinase module**. Following G9a action and dimethylation of histone H3 lysine 9 (H3K9) to H3K9me2, heterochromatin protein 1 (HP1) and DNA methyltransferase 1 (DNMT1) are recruited to the site. Ultimately, the gene is embedded in transcriptionally inert heterochromatin marked by H3K9me2 and methylated DNA (.
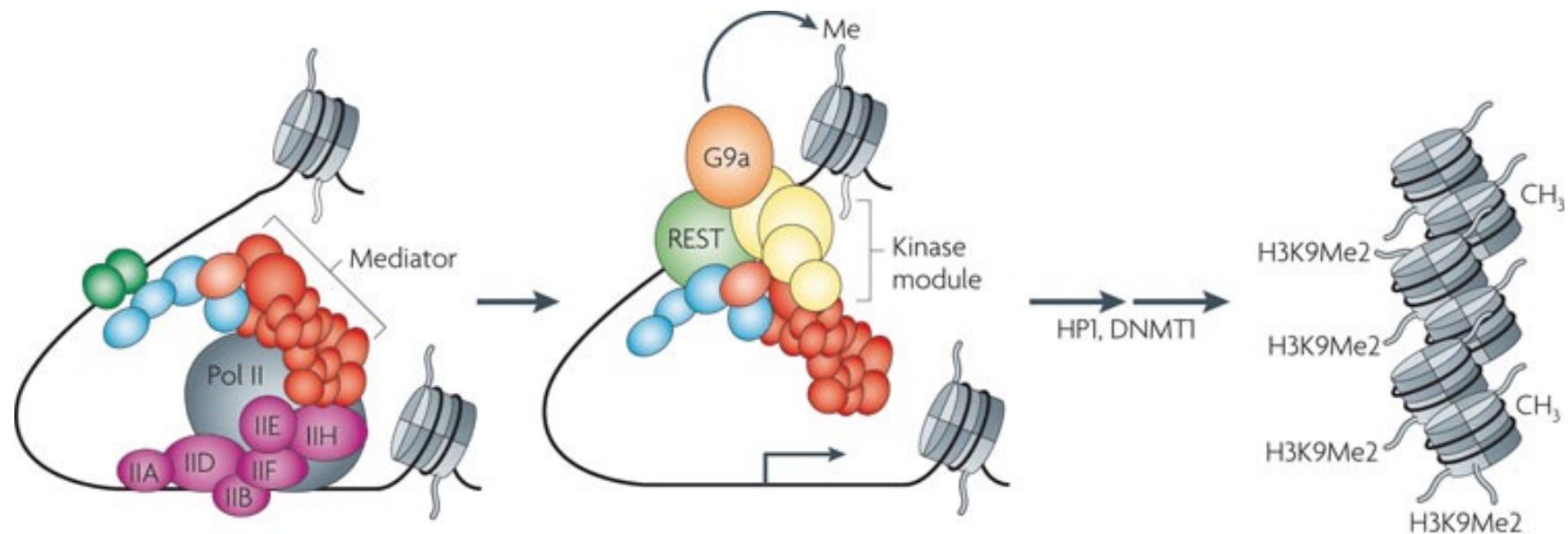
MED12 and G9a Interaction
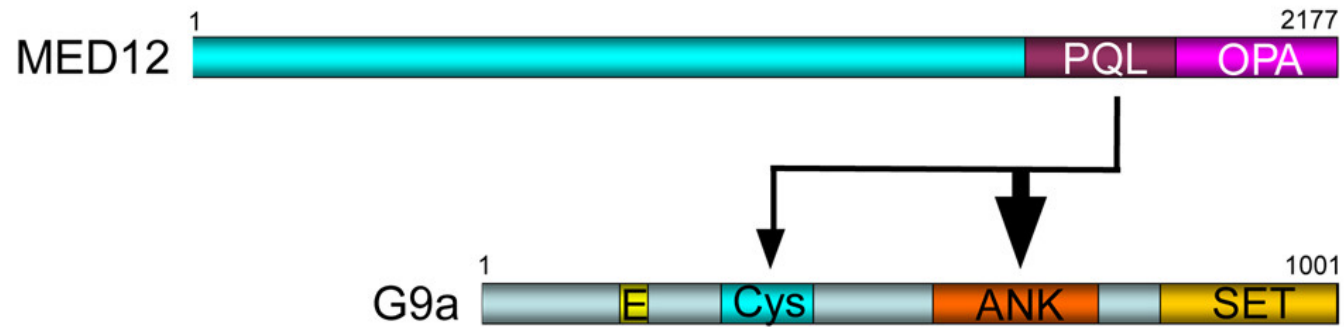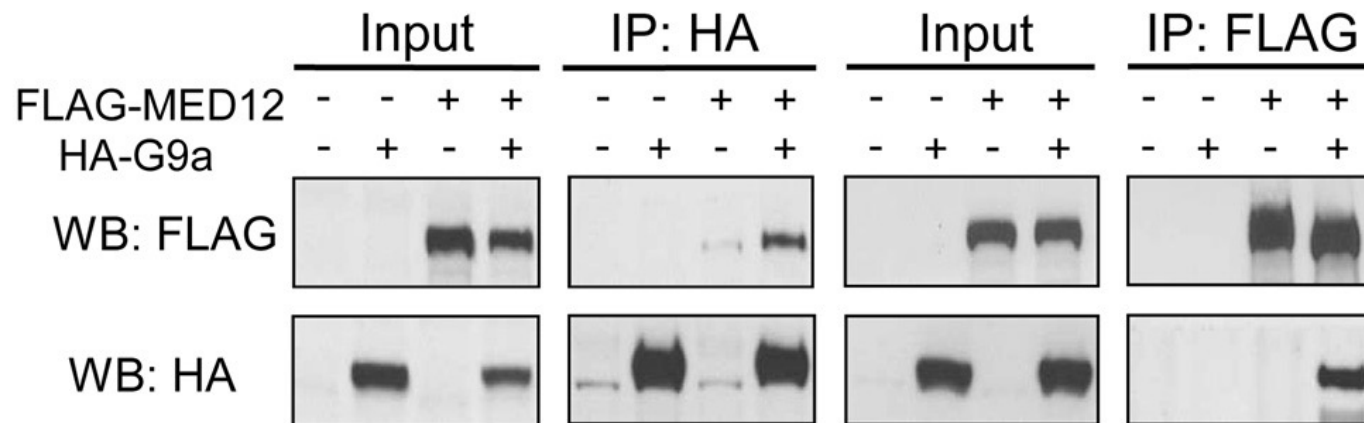
Transcription hubs or condensates

Enhancer sequences are characterized by clusters of TF-binding sites, and, consequently, active enhancers will be bound by an array of TFs. Because typical mammalian enhancers encompass several hundred base pairs, and because TFs recruit Mediator to enhancers it is plausible that multiple Mediator complexes could simultaneously occupy an active enhancer
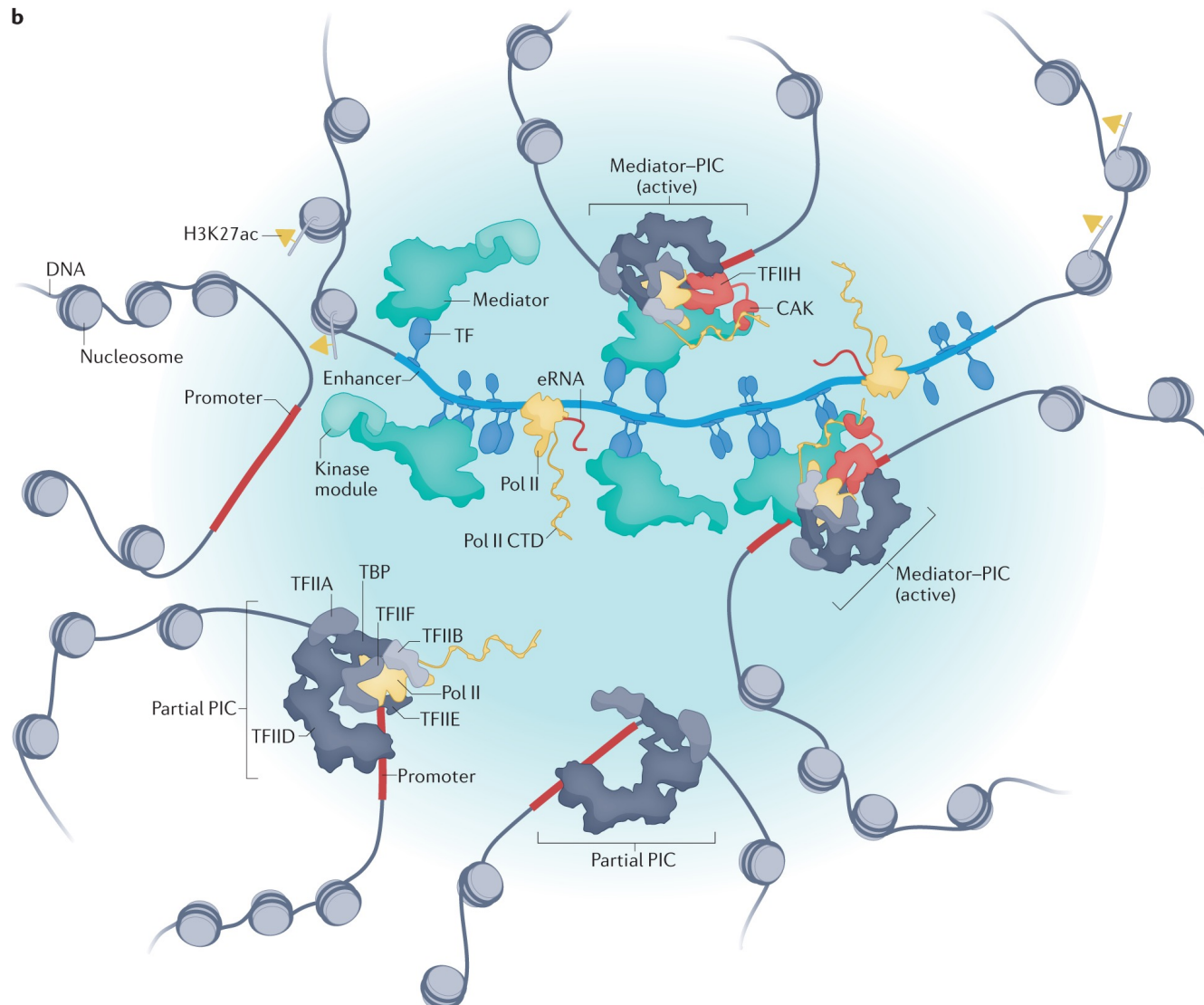

a model has emerged in which active enhancers and promoters are juxtaposed in cell nuclei, whereas transcriptionally silent regions are sequestered elsewhere.

Clustering of enhancers and promoters through spatial proximity will result in high local concentrations of bound factors such as chromatin remodellers, TFs, Mediator and Pol II

Cohesin connects gene expression and chromatin architecture

Recruitment of **Mediator** to the enhancer, or to other upstream elements, results in the formation of a <u>chromatin loop</u> that brings together the enhancer and the promoter of the gene to be transcribed. Subsequent recruitment of **cohesin** could potentially stabilize enhancer–promoter interactions by embracing the base of the chromatin loop.

Although **cohesin** is recruited to active promoters, it also becomes associated with the DNA-binding factor **CTCF**, which has been implicated in the formation of insulator elements. Thus, <u>cohesin is thought to have roles in transcription activation at some genes and in silencing at others</u>.
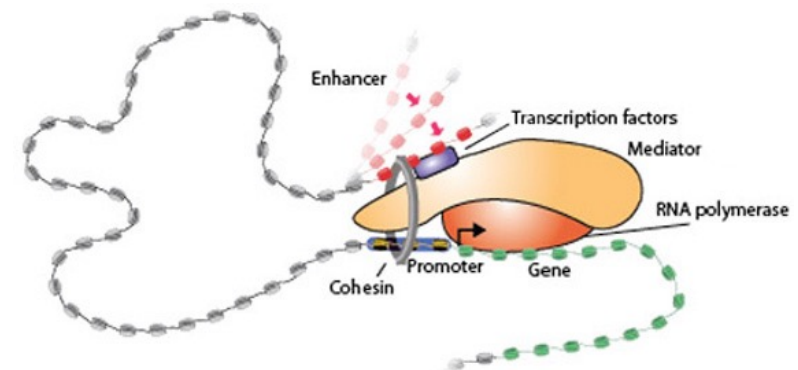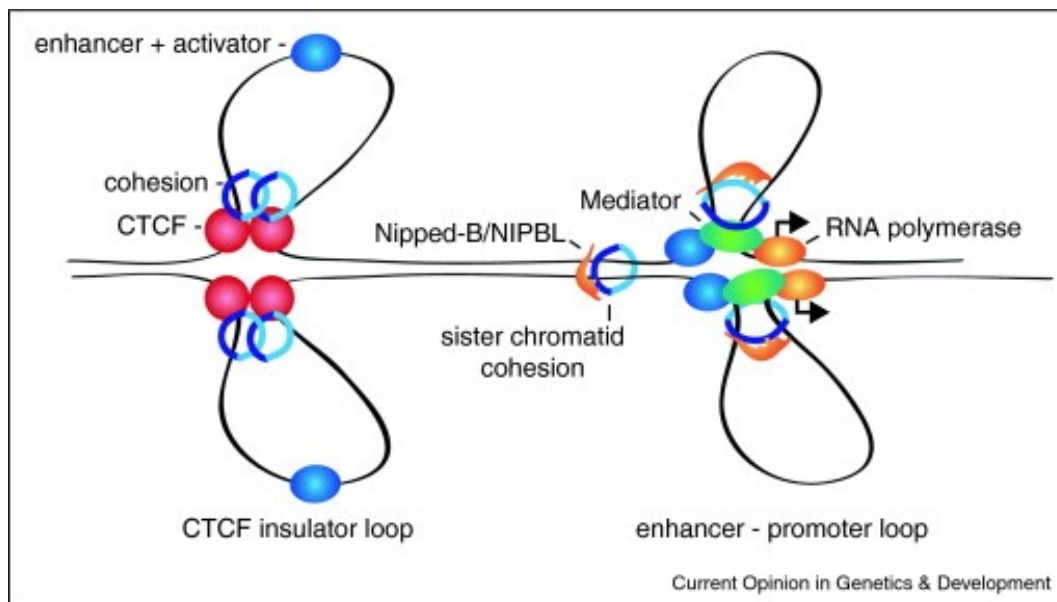
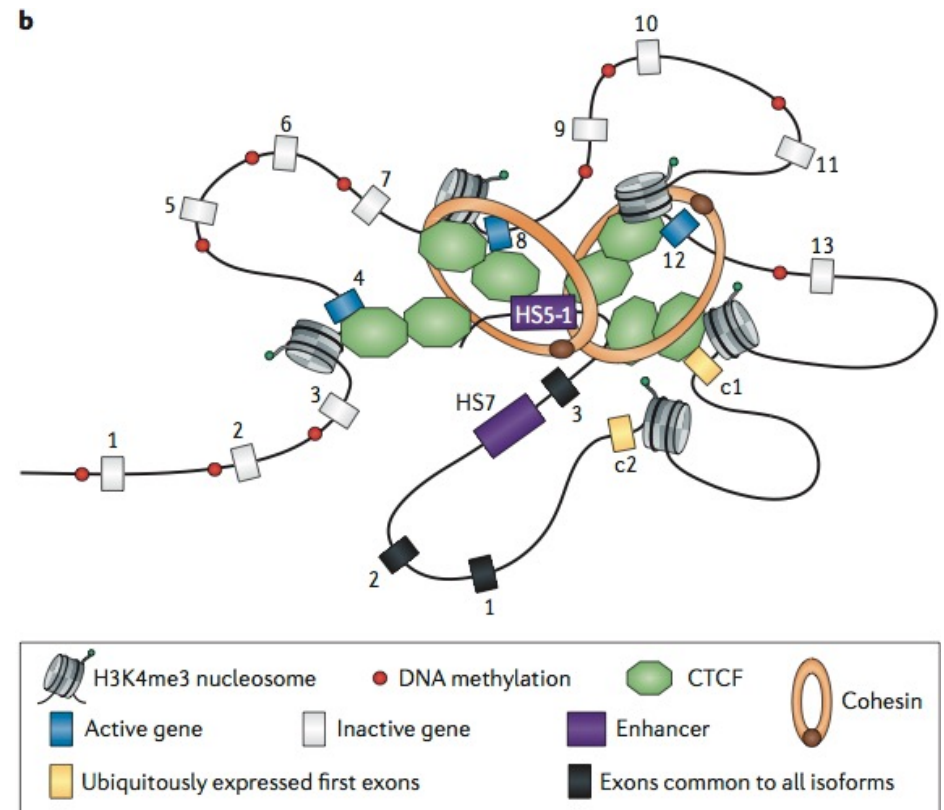# CTCF links the architecture of the genome to its function

The eukaryotic genome is organized in the three-dimensional nuclear space in a specific manner that is both a cause and a consequence of its function. This organization is partly established by a special class of <u>architectural proteins</u>, of which **CCCTC-binding factor (CTCF)** is the best characterized.

CTCF creates boundaries between topologically associating domains in chromosomes and, within these domains, facilitates interactions between transcription regulatory sequences.

CTCF links the architecture of the genome to its function

CTCF was initially classified as **insulators** -that is, a protein with the ability to interfere with enhancer–promoter communication.

CTCF mediates both inter- and intra-chromosomal interactions between distant sites in the genome. As a result of these interactions, CTCF elicits specific functional outcomes that are <u>context dependent</u>, determined by the nature of the two sequences brought together and by the proteins with which they interact ('architectural' protein)

H3K4me3 nucleosome	● DNA methylation	🟢 CTCF
		Cohesin
🟦 Active gene	⬜ Inactive gene	🟪 Enhancer
🟨 Ubiquitously expressed first exons	⬛ Exons common to all isoforms	

Regulation of CTCF binding to DNA

Constitutive binding sites of CTCF are present <u>in non-methylated and nucleosome-free regions</u>. **Cell-type-specific CTCF binding is partly regulated by differential DNA methylation and nucleosome occupancy across different cell types**. This suggests that cells can use ATP-dependent chromatin remodelling complexes to regulate nucleosome occupancy at specific CTCF-binding sites and control the interaction of this protein with DNA. In addition, the methylation status of cell-type-specific CTCF-binding sites may be determined by a combination of activities of *de novo* methyltransferases and TET enzymes that regulate the presence and levels of 5mC at specific sites. Immortalized cancer cell lines contain high levels of 5mC at CTCF-binding sites, which correlates with the low CTCF occupancy in these cells.
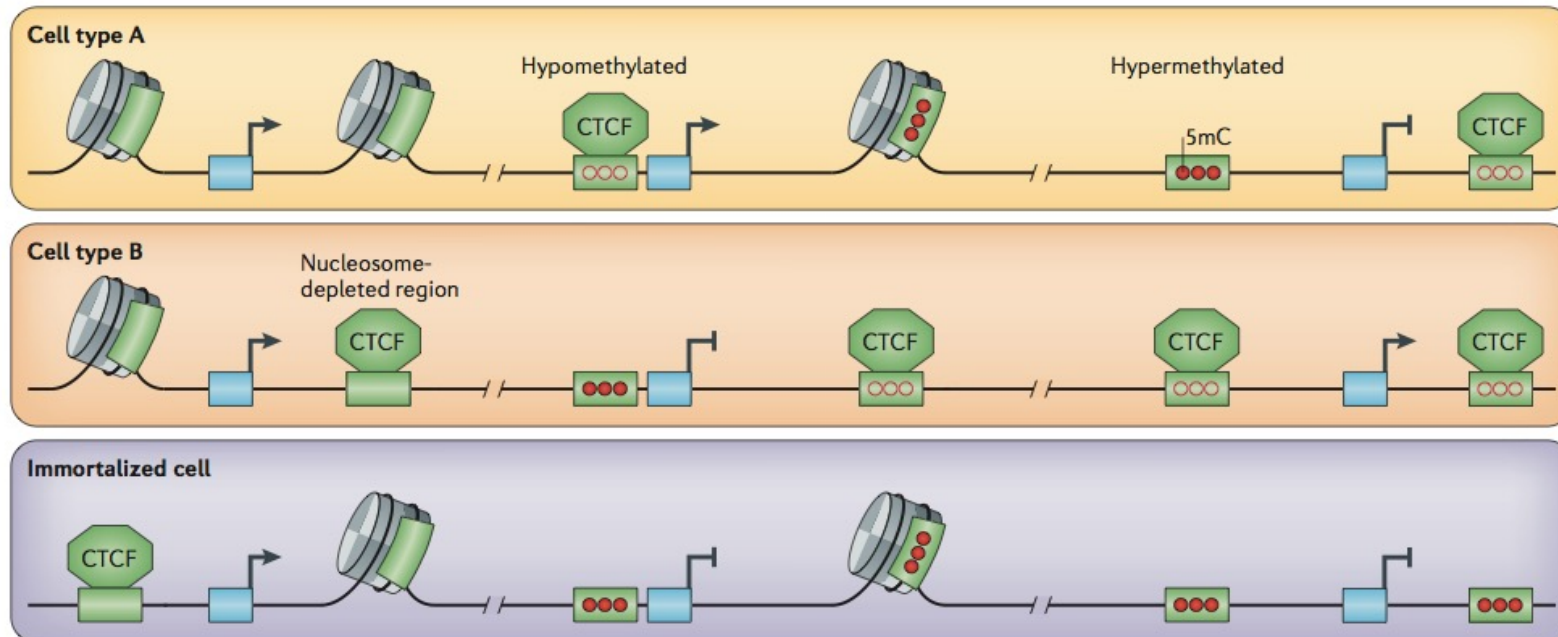
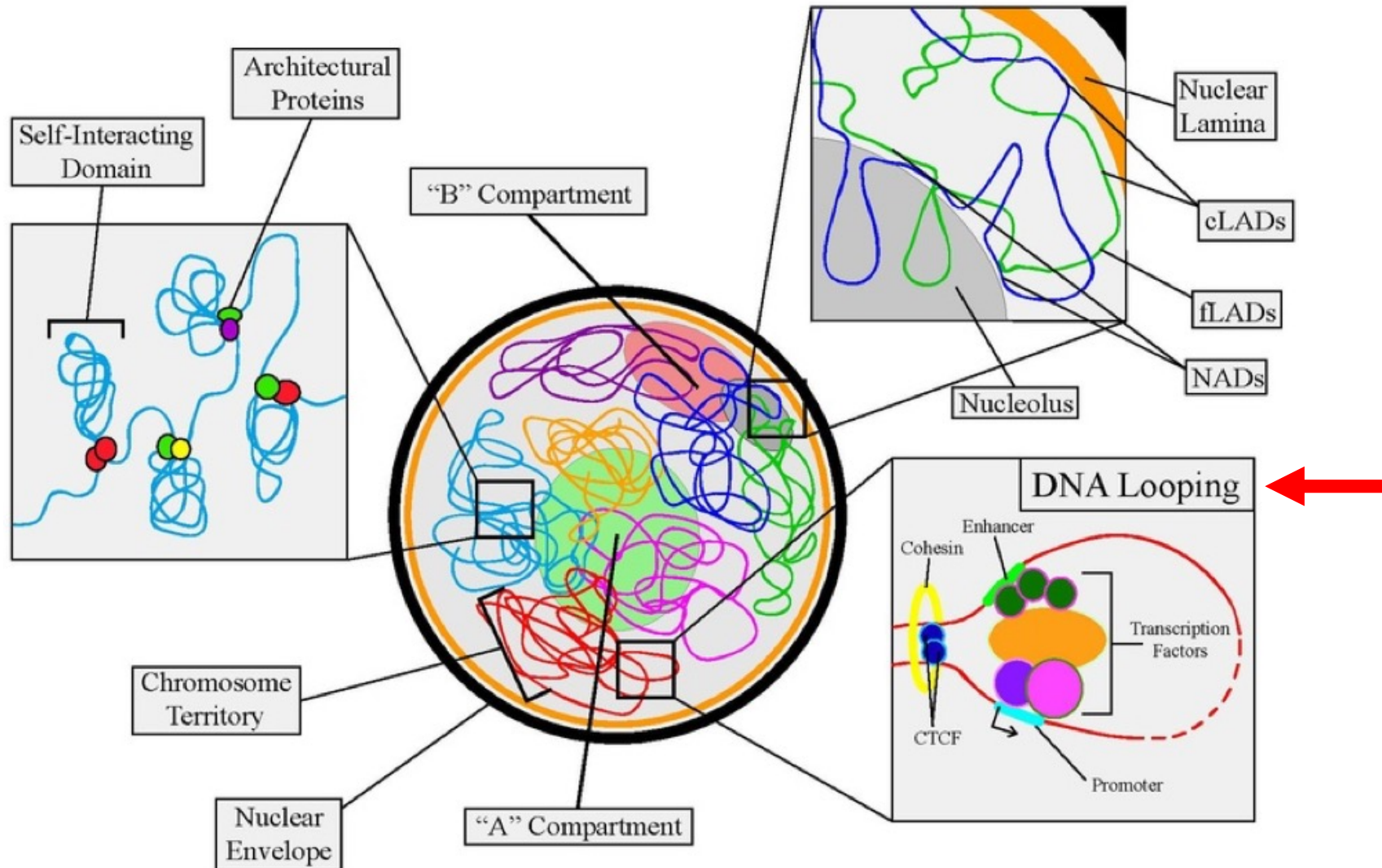# Transcriptional Control of Genes

Transcriptional regulation occurs at two interconnected levels: the first involves **transcription factors and the transcription apparatus (A-C)**, and the second involves **chromatin and its regulators (D-E).**
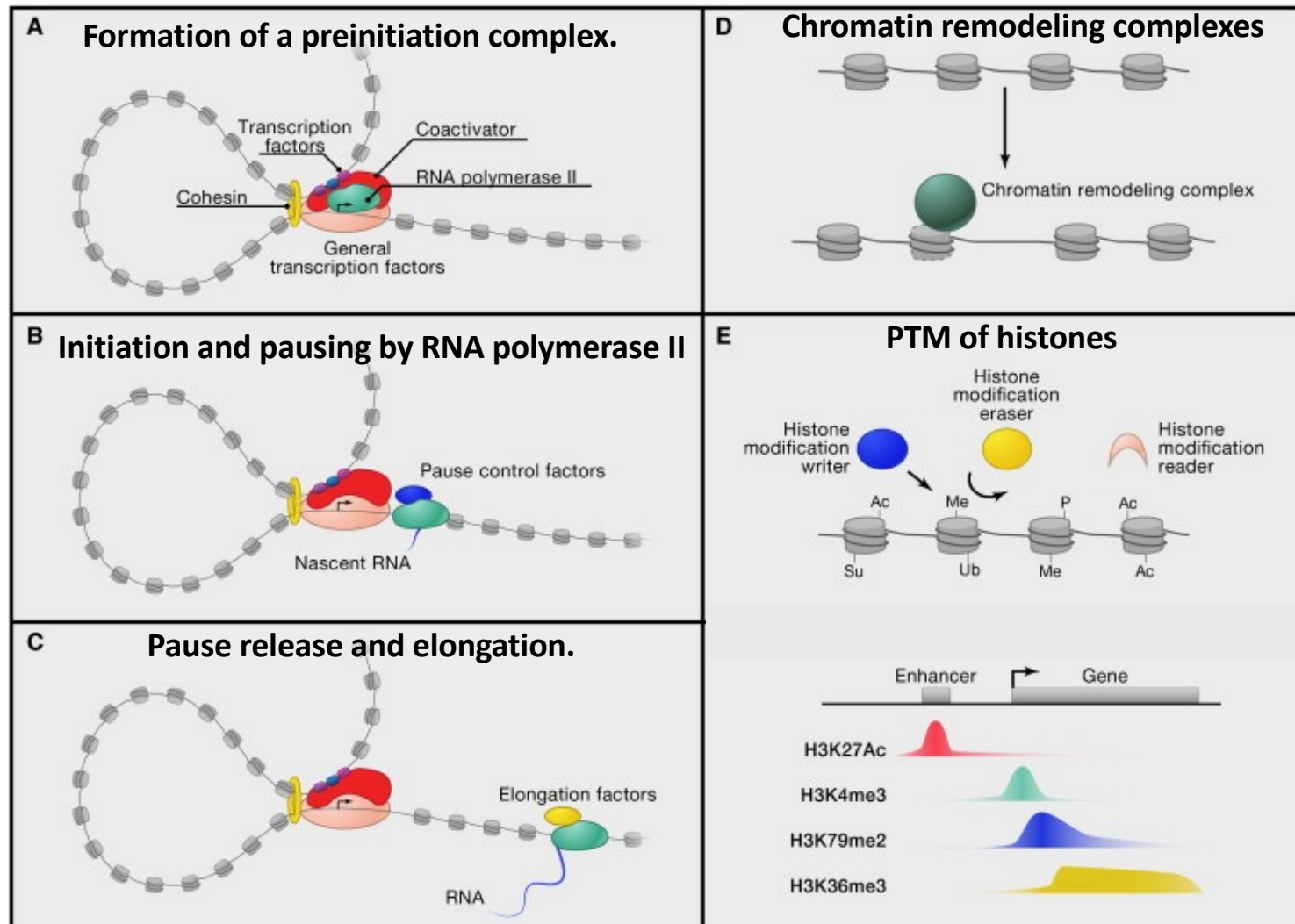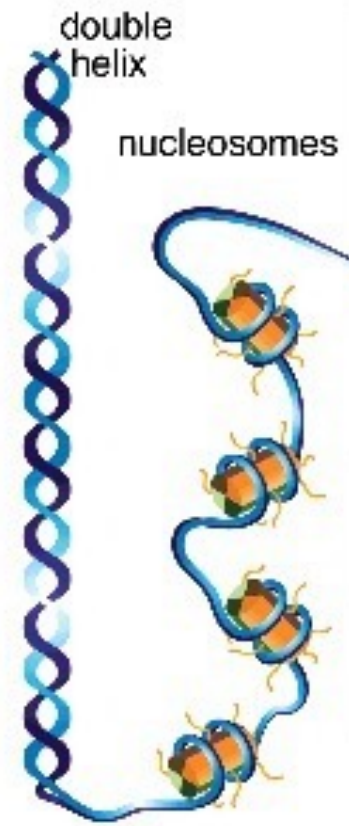
Transcriptional regulation II

Functional elements are physically connected to each other

small scale
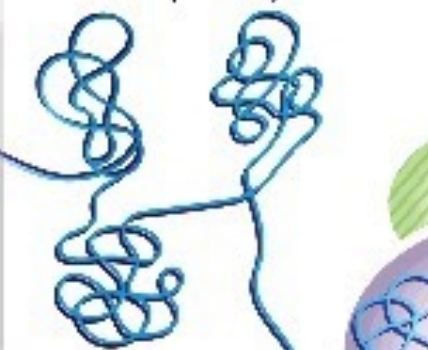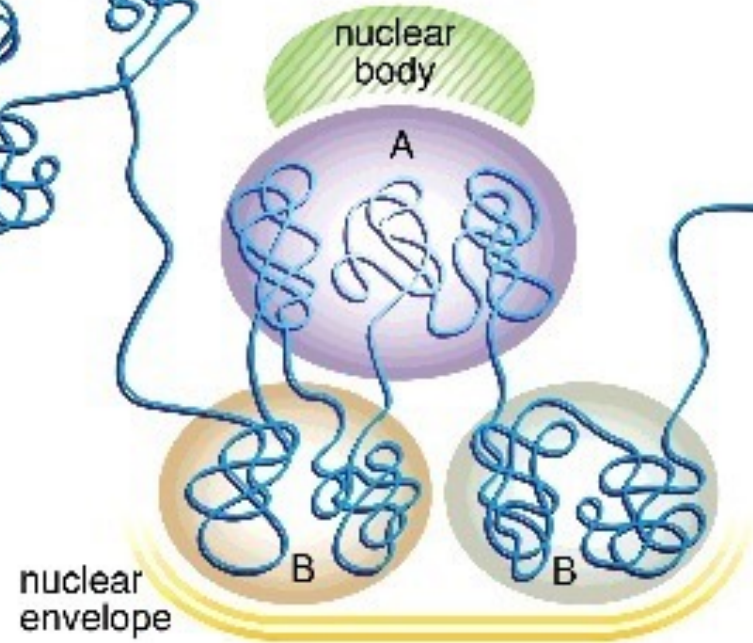10 bp - few kb

compaction of DNA into chromatin

double helix

nucleosomes

intermediate scale
~10 kb - few Mb

local regulation of contact frequency

intra- and interchromosomal regulation of contact frequency

chromatin loops and topologically associating domains (TADs)

epigenomic compartments ("active" vs. "inactive")

nuclear body

A

B

B

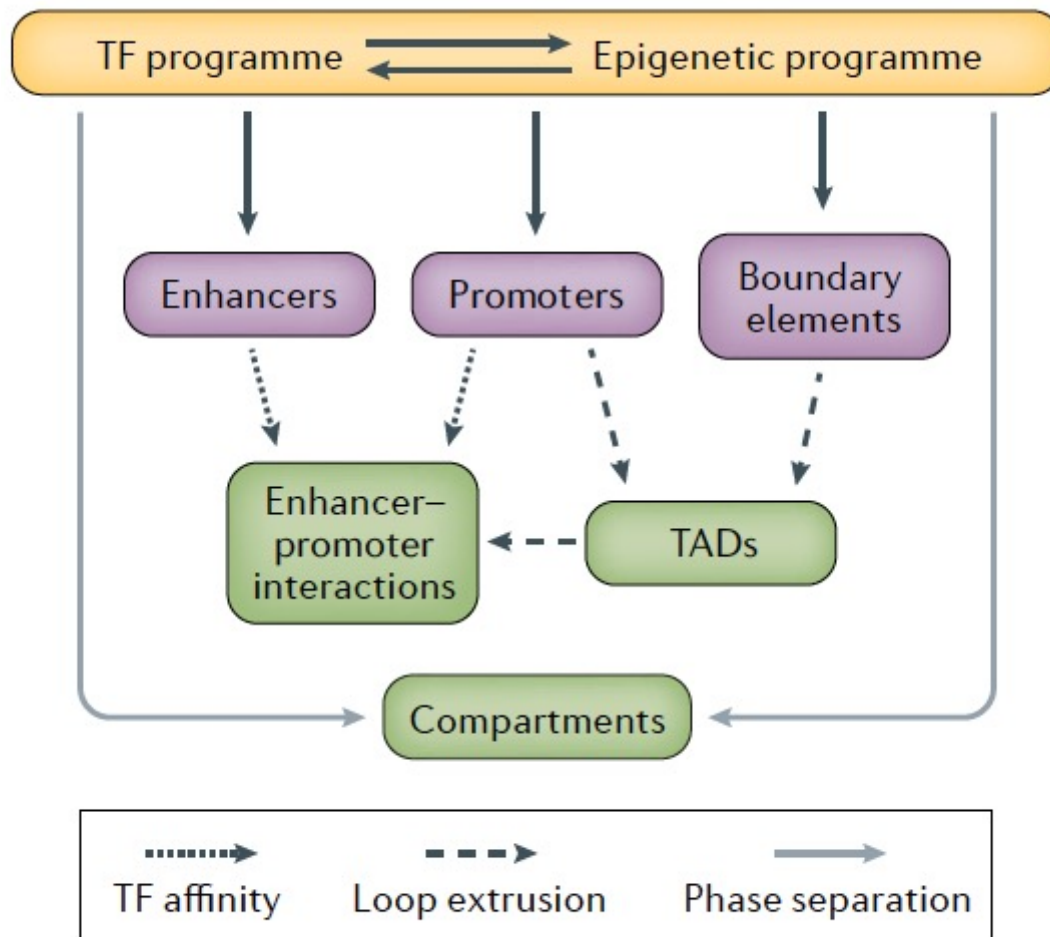nuclear envelope

large scale
~100 Mb - ~3000 Mb

regulation of contact frequency between individual chromosomes
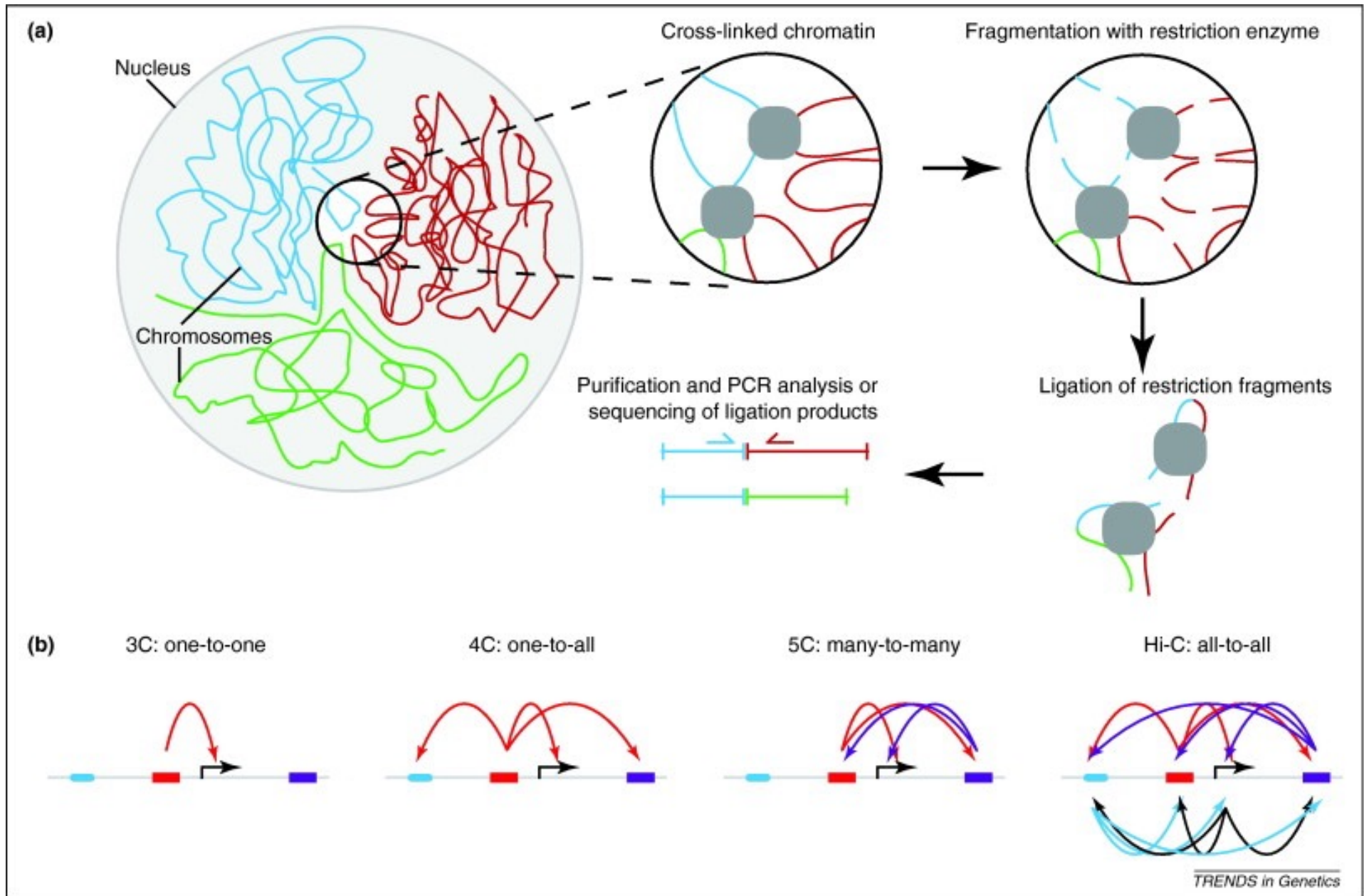
chromosome territories
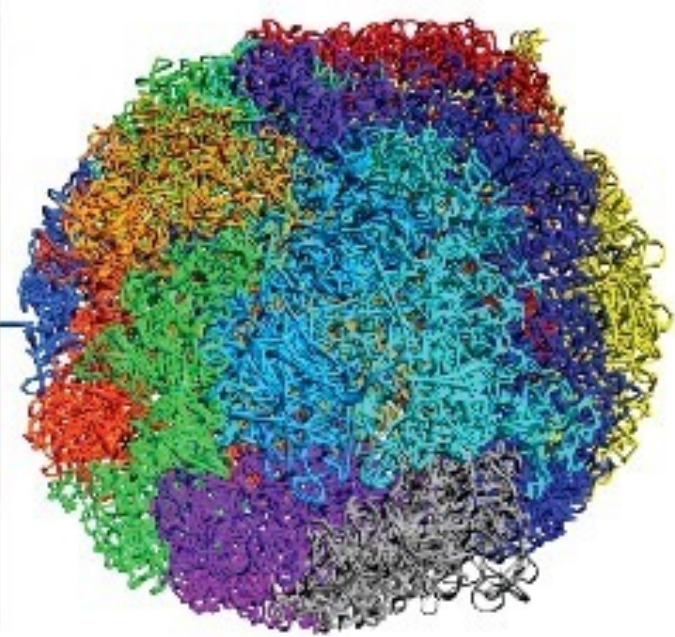
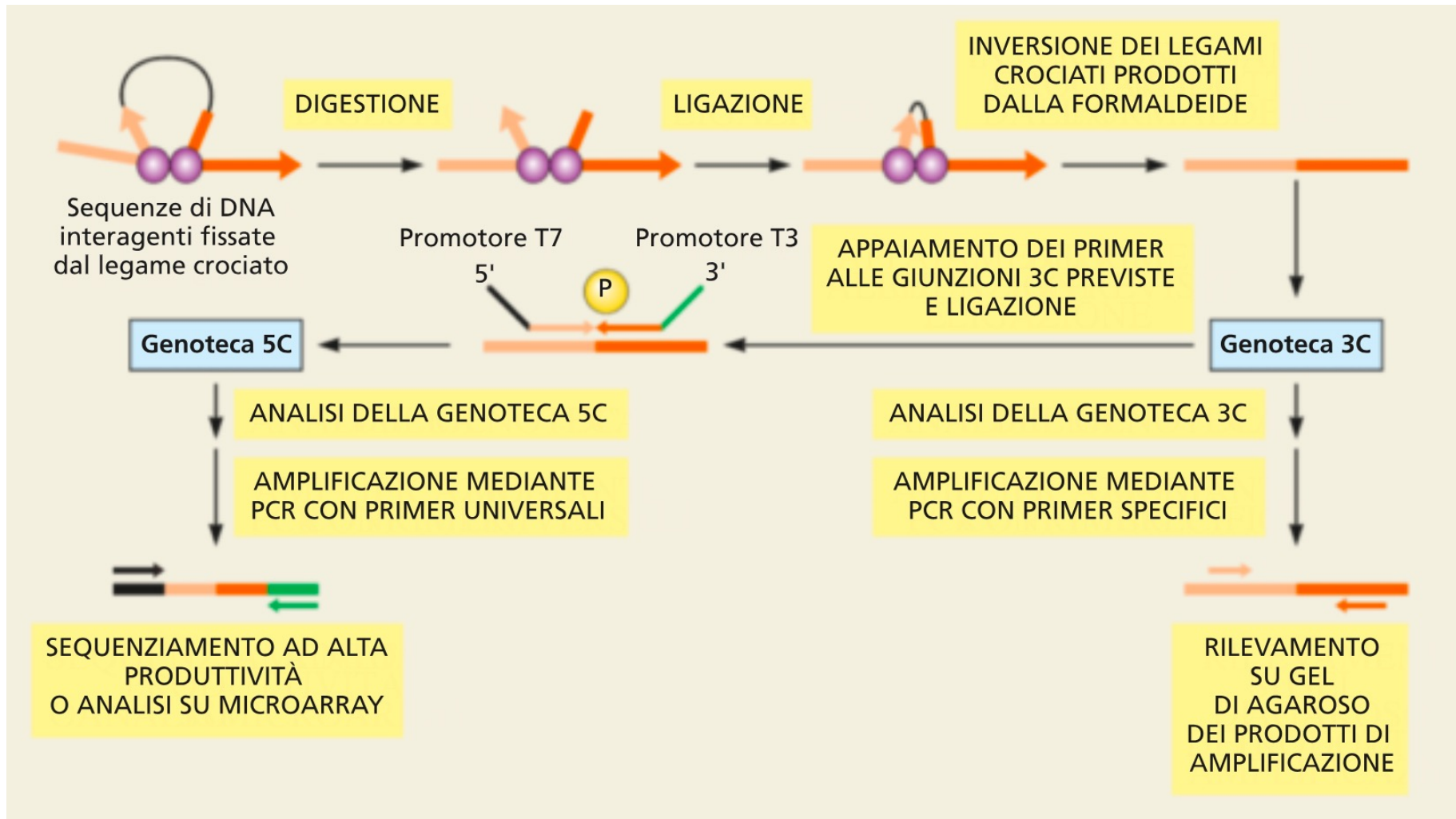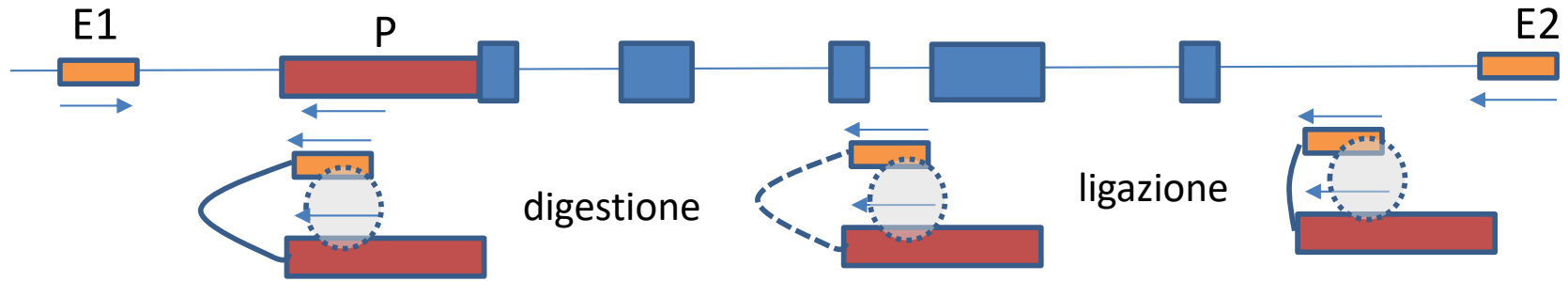# The relationship between genome structure and function

The activity of the regulatory elements of the genome is related to the TF programme and the epigenetic programme in the cell nucleus. Together, the sequence, order and activity of these regulatory elements instruct the processes underlying higher-order chromatin structures. CTCF-binding boundary elements demarcate the boundaries of cohesin mediated loop extrusion and thereby control the positions of TAD borders; active gene promoters might act in a similar way. Within TADs, the process of loop extrusion brings enhancers and promoters into close physical proximity; this may facilitate specific enhancer–promoter interactions, which are stabilized by affinity between their bound TFs and cofactors. Dependent on their chromatin state, regions of euchromatin and heterochromatin form spatially separated compartments via a process of phase separation.

Methodologies for studying 3D genome organization
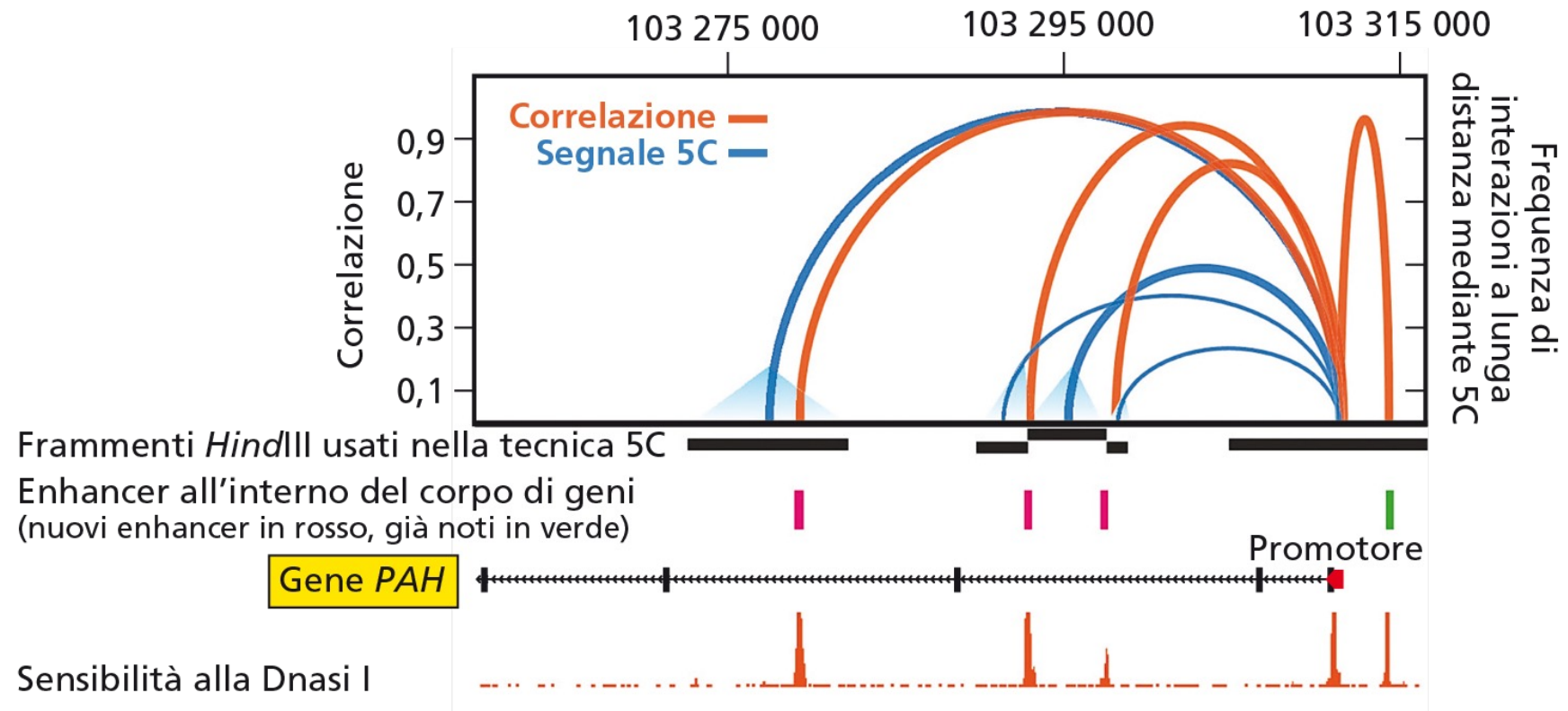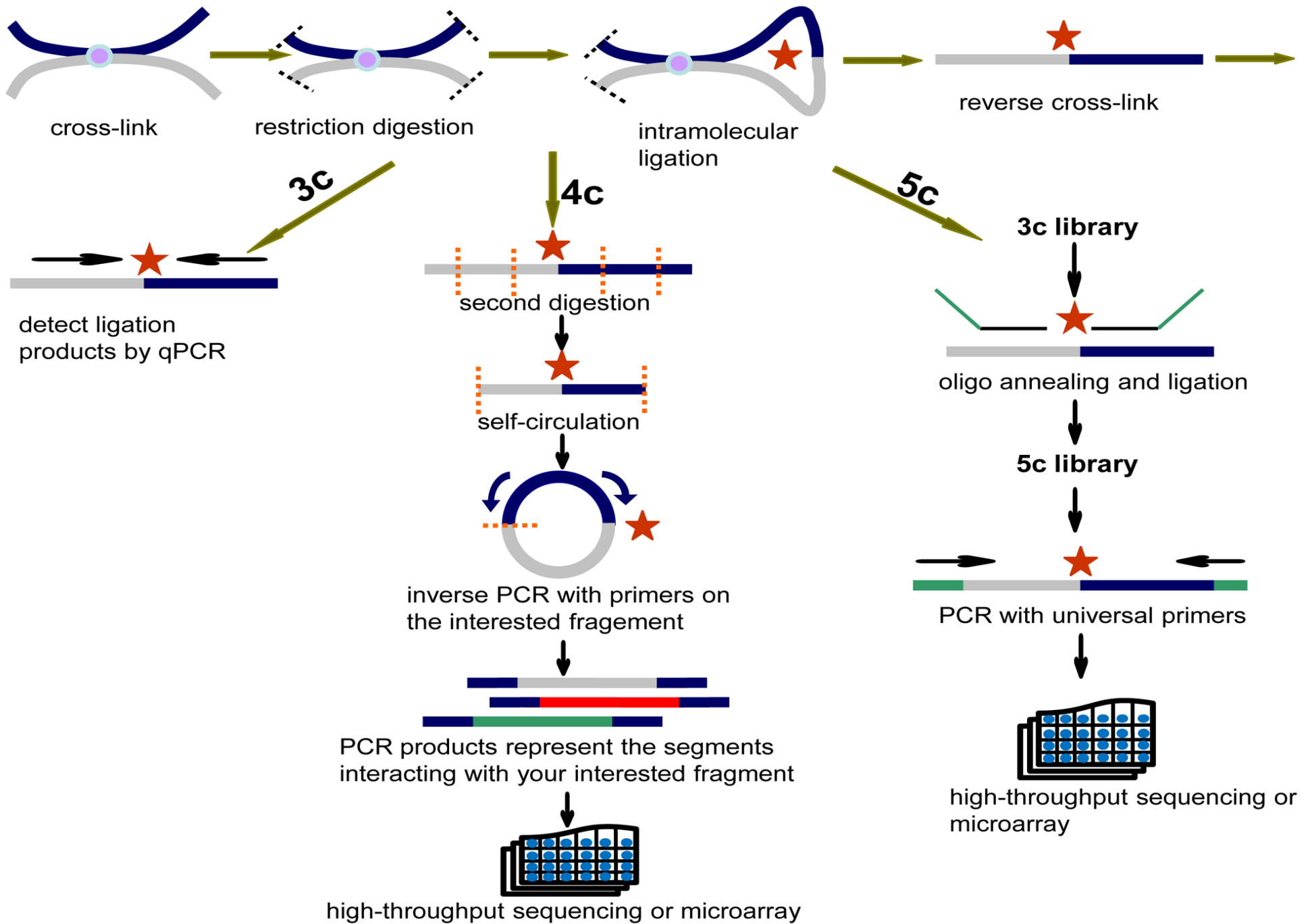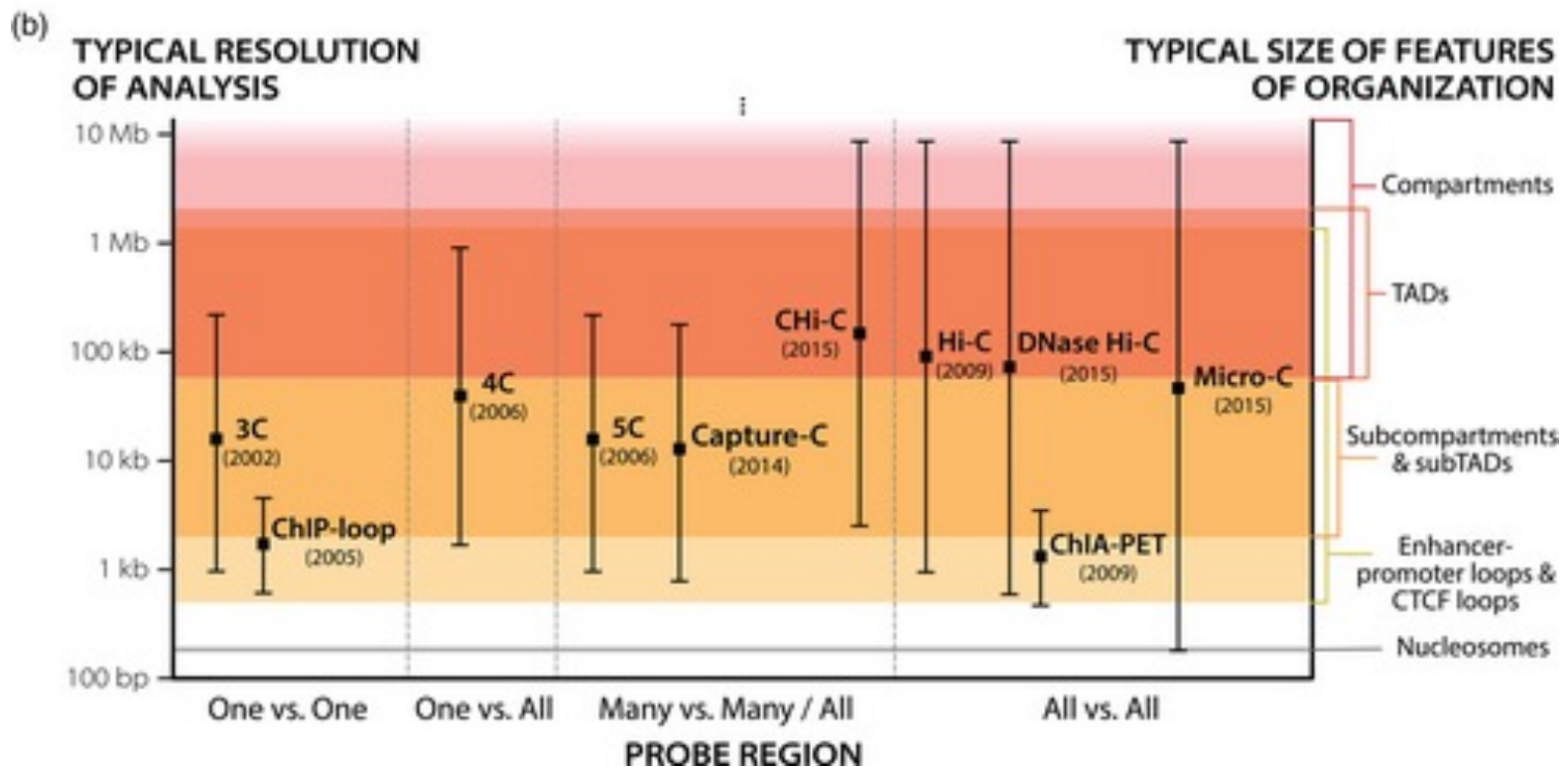
Chromosome conformation capture

TRENDS in Genetics

Chromosome Conformation Capture

Connessioni tra i DHS e il promotore del gene PAH

cross-link

restriction digestion

intramolecular ligation

reverse cross-link

3c

4c

5c

detect ligation products by qPCR

second digestion

self-circulation

inverse PCR with primers on the interested fragment

PCR products represent the segments interacting with your interested fragment

high-throughput sequencing or microarray

3c library

oligo annealing and ligation

5c library

PCR with universal primers

high-throughput sequencing or microarray

Timeline and comparison of major chromosome conformation capture techniques.
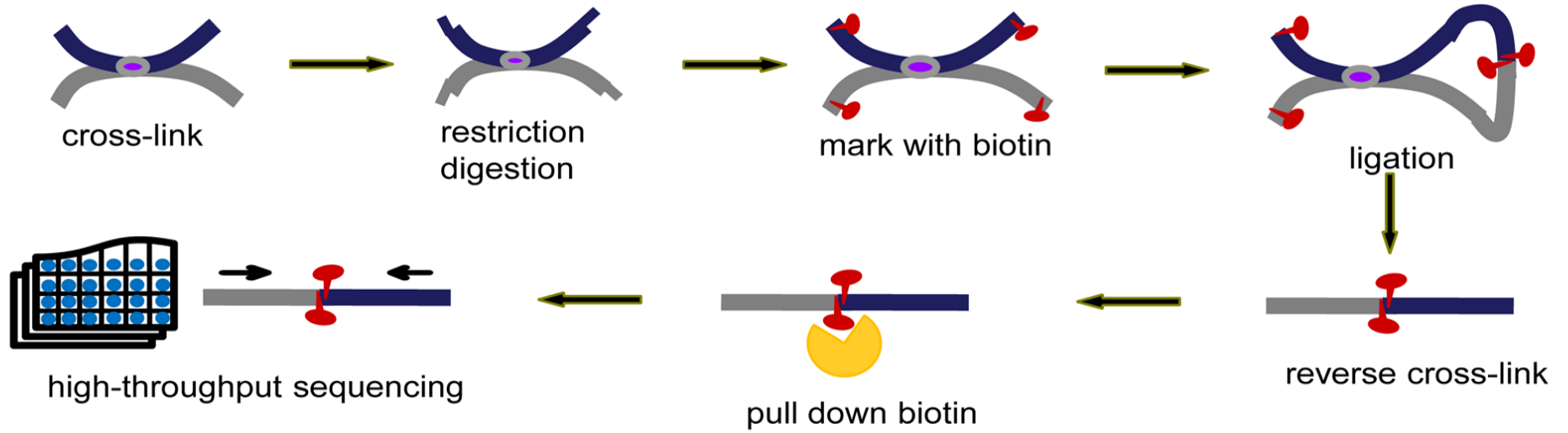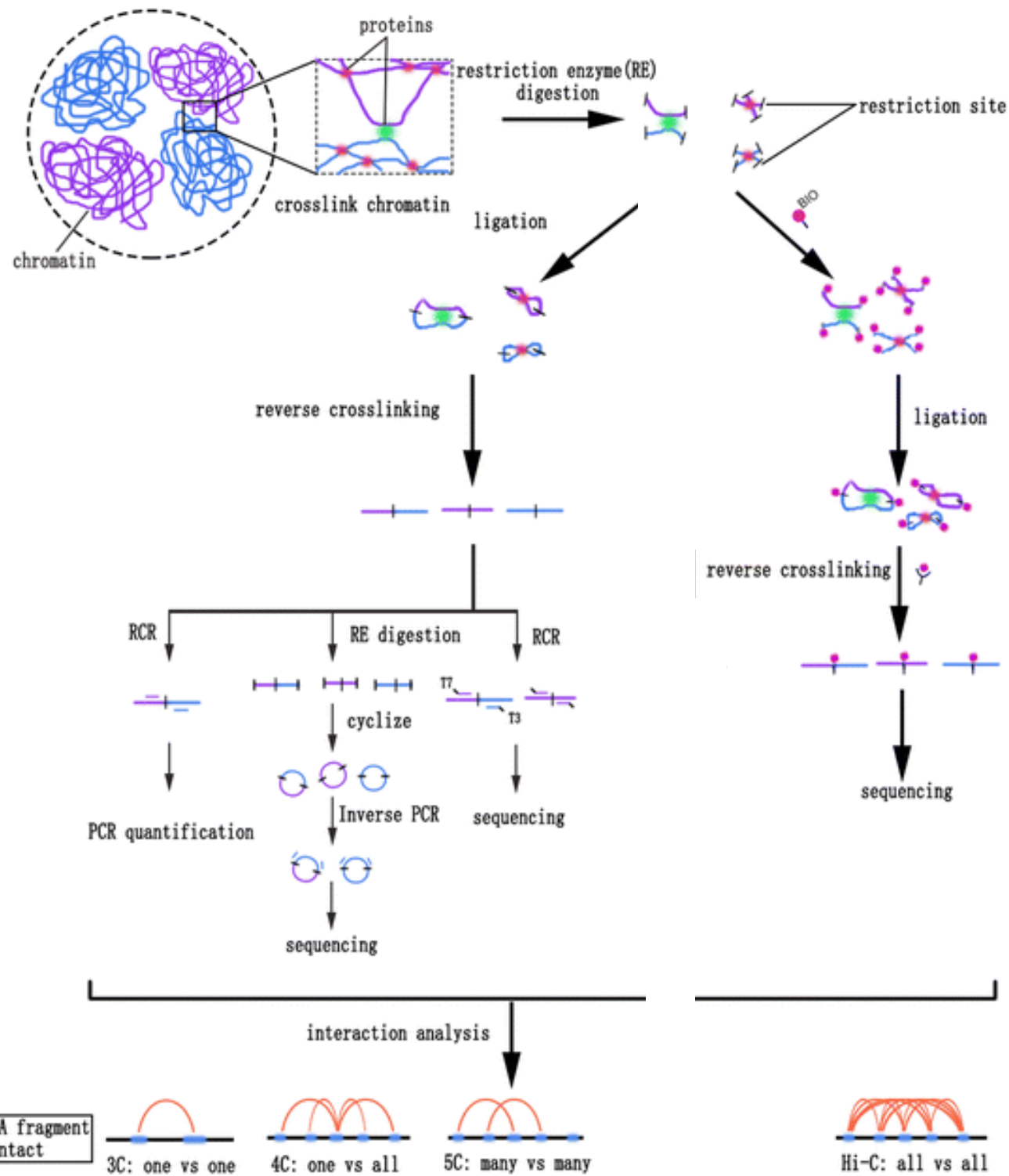
Hi-C

cross-link

restriction digestion

mark with biotin

ligation

reverse cross-link

pull down biotin

high-throughput sequencing

proteins

restriction enzyme(RE) digestion

restriction site

crosslink chromatin

chromatin

ligation

reverse crosslinking

ligation

RCR   RE digestion   RCR

reverse crosslinking

cyclize

Inverse PCR   sequencing

PCR quantification

sequencing

sequencing

interaction analysis

— DNA fragment
⌒ contact

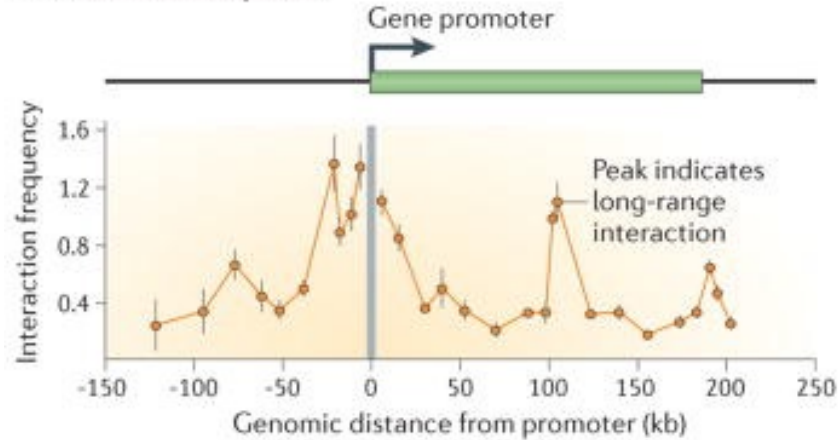3C: one vs one   4C: one vs all   5C: many vs many

Hi-C: all vs all

3C, 4C, 5C and Hi-C datasets

a 3C interaction profile

Gene promoter

Interaction frequency

Peak indicates long-range interaction

Genomic distance from promoter (kb)

b 4C interaction profile

20 Mb

4C anchor

4C signal

Number of reads

DNAseI hypersensitive sites

Number of reads

Correlation between presence of DHSs and long-range interactions
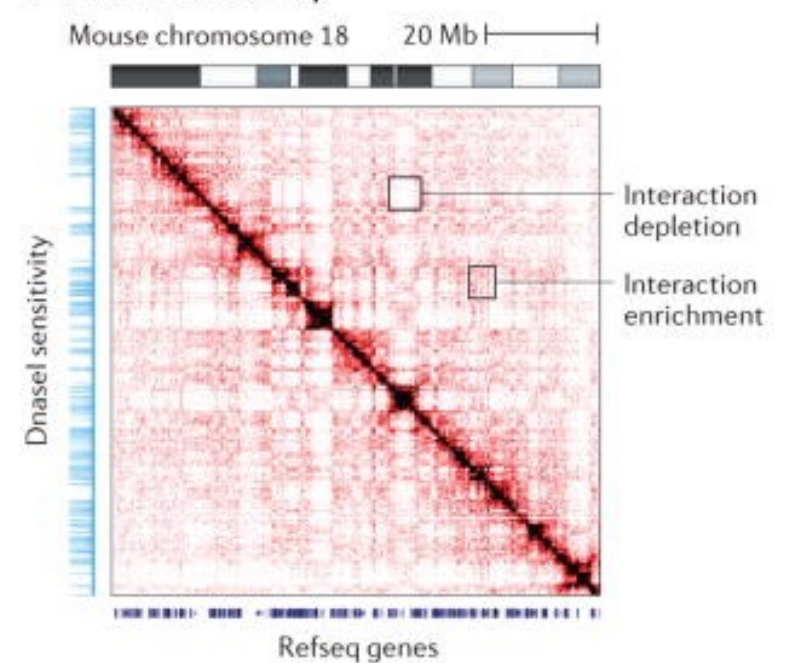
c 5C interaction map

HBB HBD HBG

ENm009 (1 Mb)

TSS

γ-δ globin

Distal fragments

0   350

d Hi-C interaction map

Mouse chromosome 18

20 Mb

Dnase I sensitivity

Interaction depletion

Interaction enrichment

Refseq genes
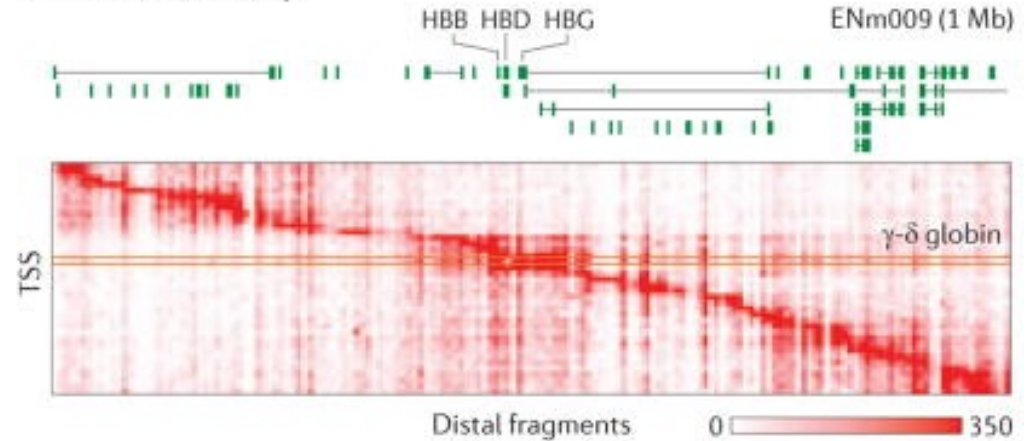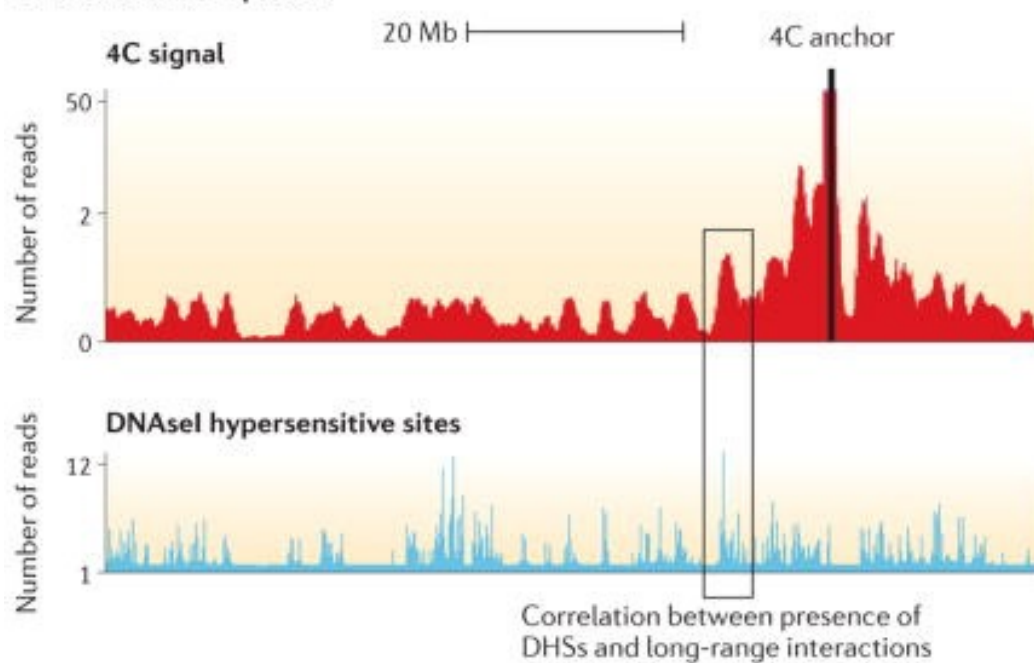
Chromosome Conformation Capture Methods

Common steps:
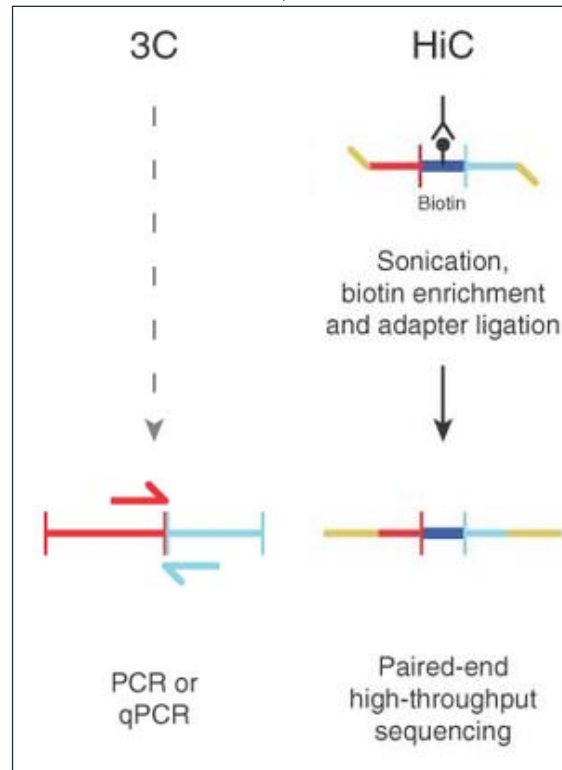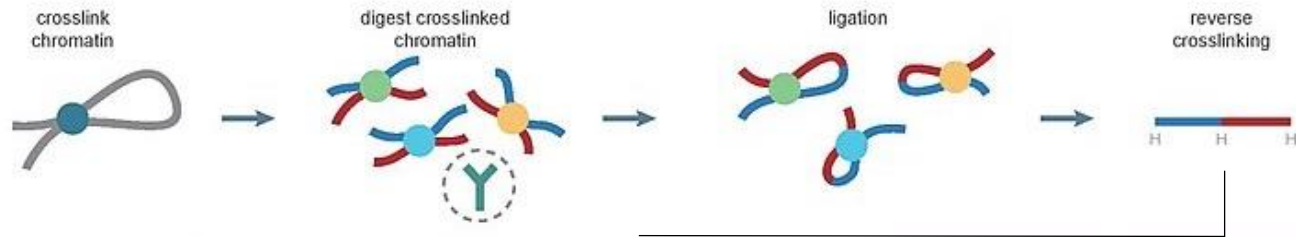
Hi-C Method

Hi-C protocol involves crosslinking cells with formaldehyde, permeabilizing them with nuclei intact, digesting DNA with a suitable 4-cutter restriction enzyme filling the 5'-overhangs while incorporating a biotinylated nucleotide, ligating the resulting blunt-end fragments, shearing the DNA, capturing the biotinylated ligation junctions with streptavidin beads, and analyzing the resulting fragments with sequencing.

Visualization of genome interactions

Sequencing

Map creation
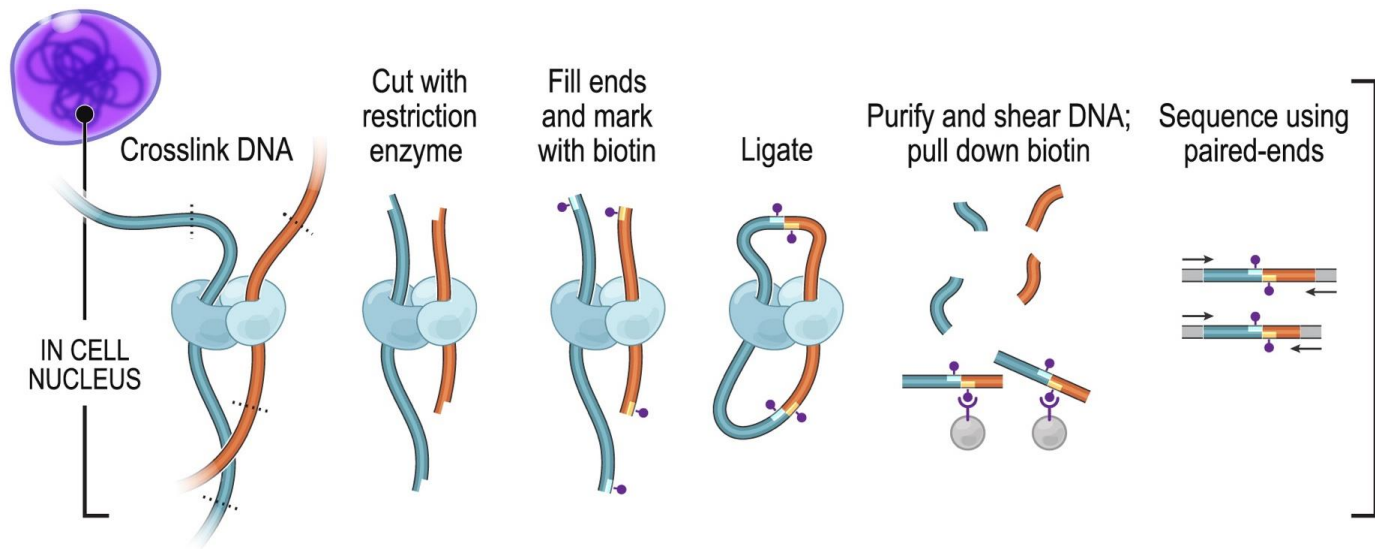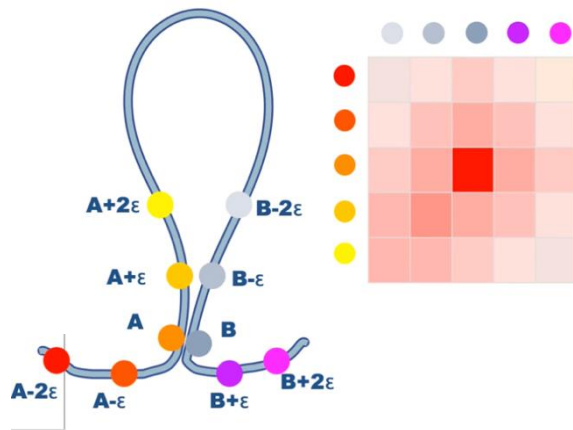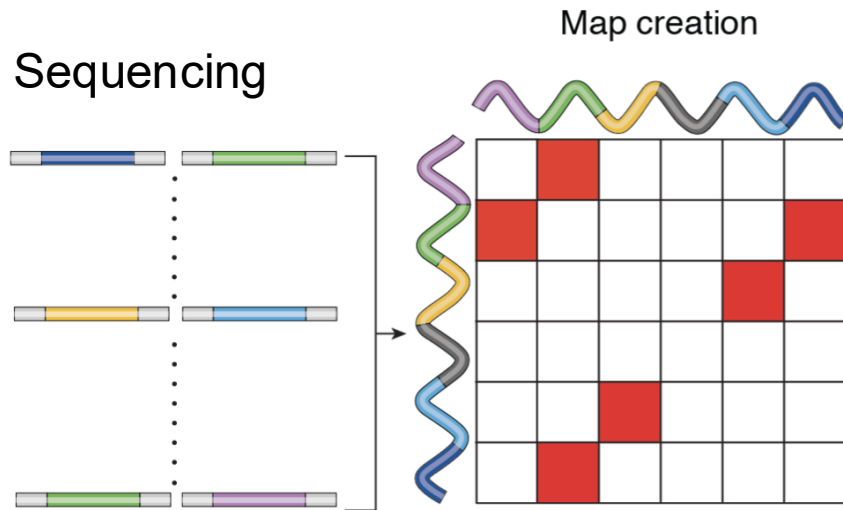
Hi-C Matrices and Models
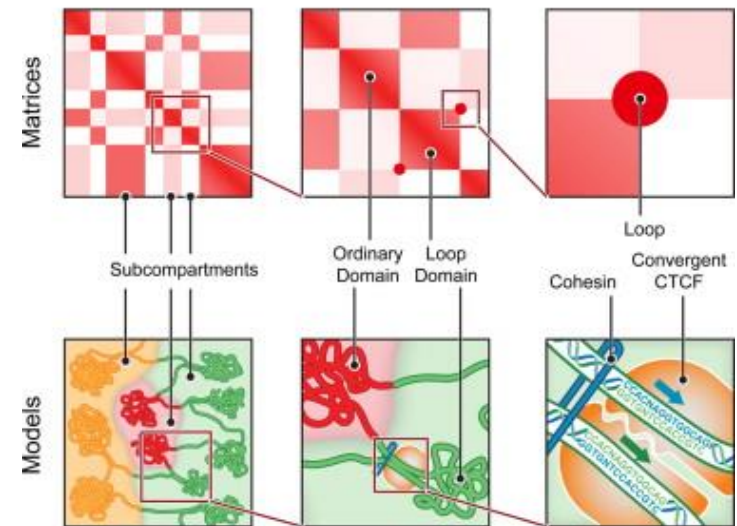
small scale
10 bp - few kb
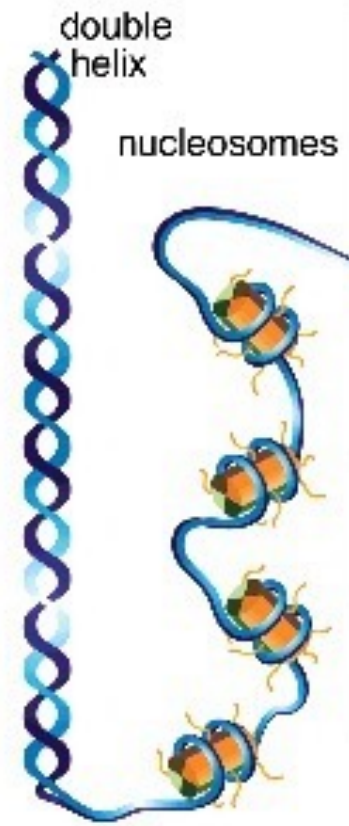
compaction of DNA into chromatin

double helix

nucleosomes

intermediate scale
~10 kb - few Mb

local regulation of contact frequency

intra- and interchromosomal regulation of contact frequency

chromatin loops and topologically associating domains (TADs)

epigenomic compartments ("active" vs. "inactive")

nuclear body

A

B   B

nuclear envelope
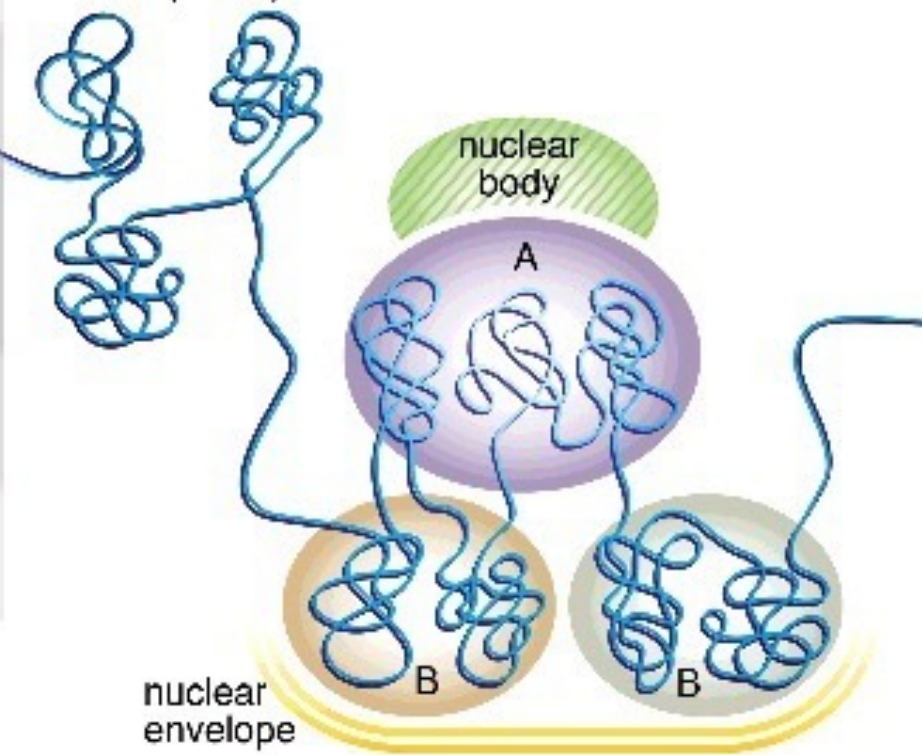
large scale
~100 Mb - ~3000 Mb

regulation of contact frequency between individual chromosomes

chromosome territories
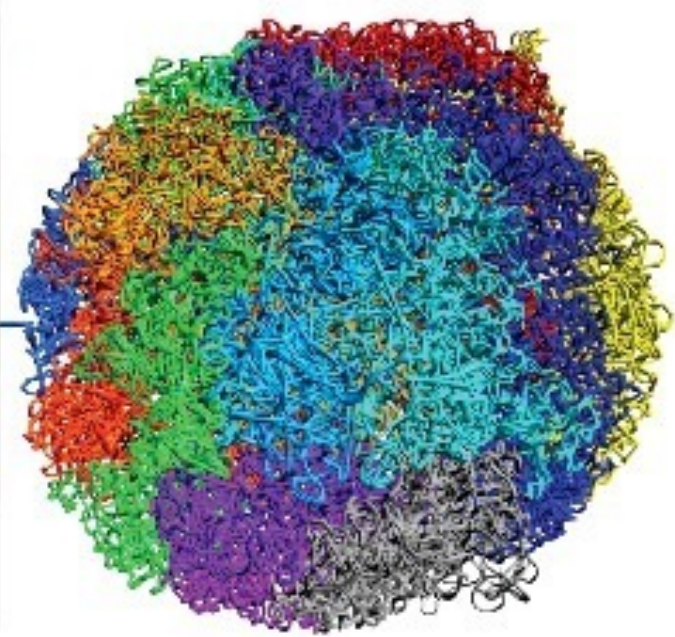
Visualization of genome interactions
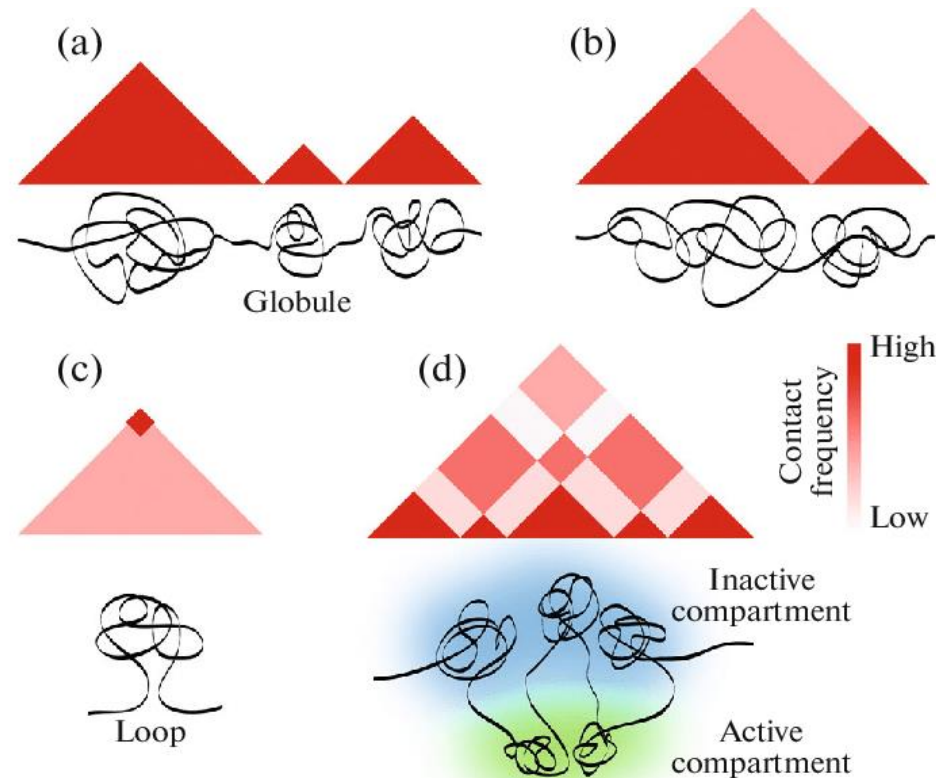
Hi-C map patterns (at the top) and their interpretation (at the bottom). (a) Isolated triangles on a heat map are commonly interpreted as chromatin globules. (b) A combination of small triangles into larger ones may reflect the hierarchic organization of globules (small globules fuse into larger ones). (c) An intense signal at the apex of a triangle suggests the interaction of TAD boundaries and the formation of a chromatin loop, internal regions of which may similarly interact with each other to form a globule. (d) A checkerboard pattern suggests the formation of chromatin compartments.

Visualization of genome interactions
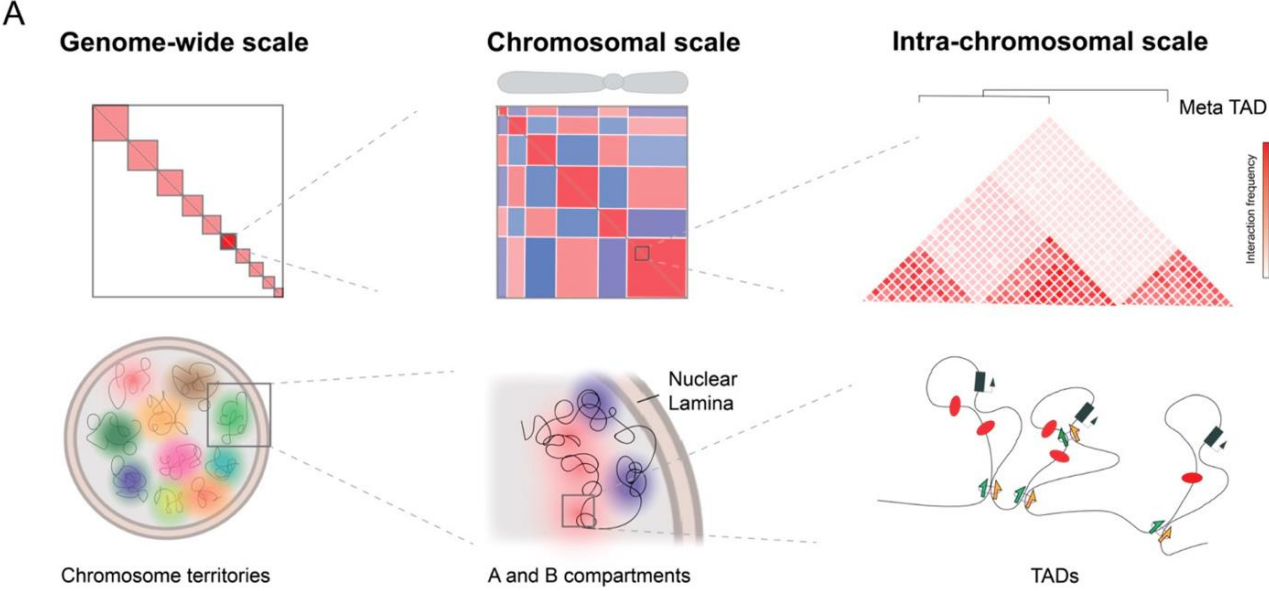
TADs: Topologically associating domains
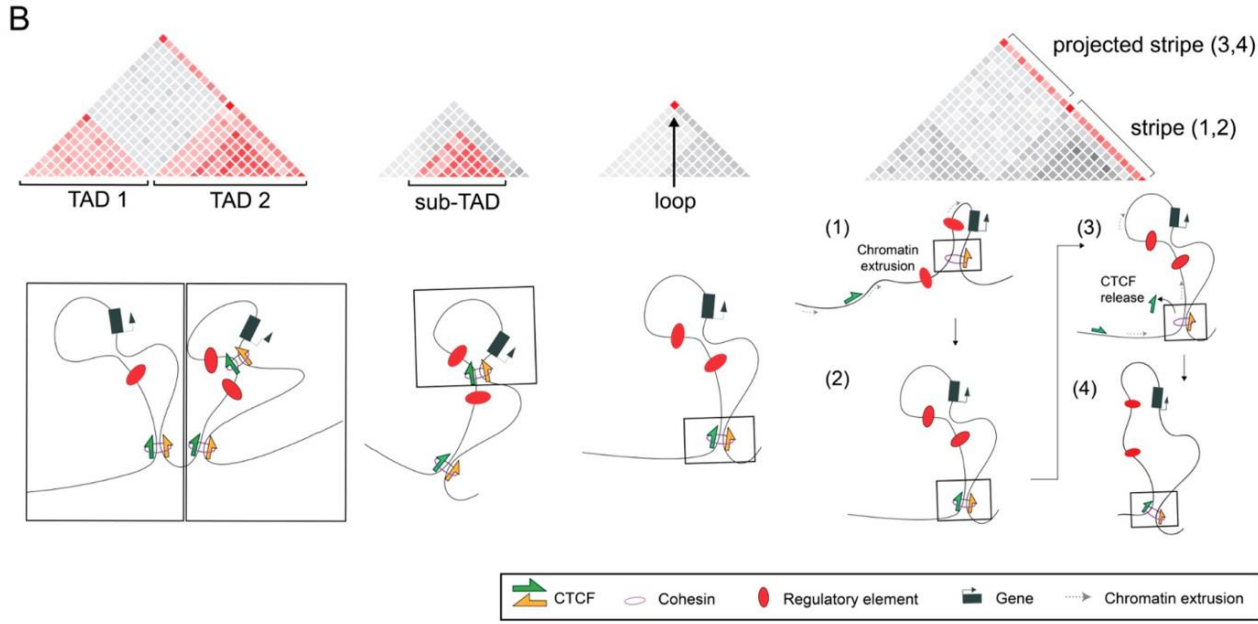
Hi-C: A comprehensive technique to capture the conformation of genomes

a Wild type

Contact frequency

Gene (active/inactive)
Enhancer (active/inactive)
CTCF sites

b Boundary deletion

c Boundary insertion

d Inversion
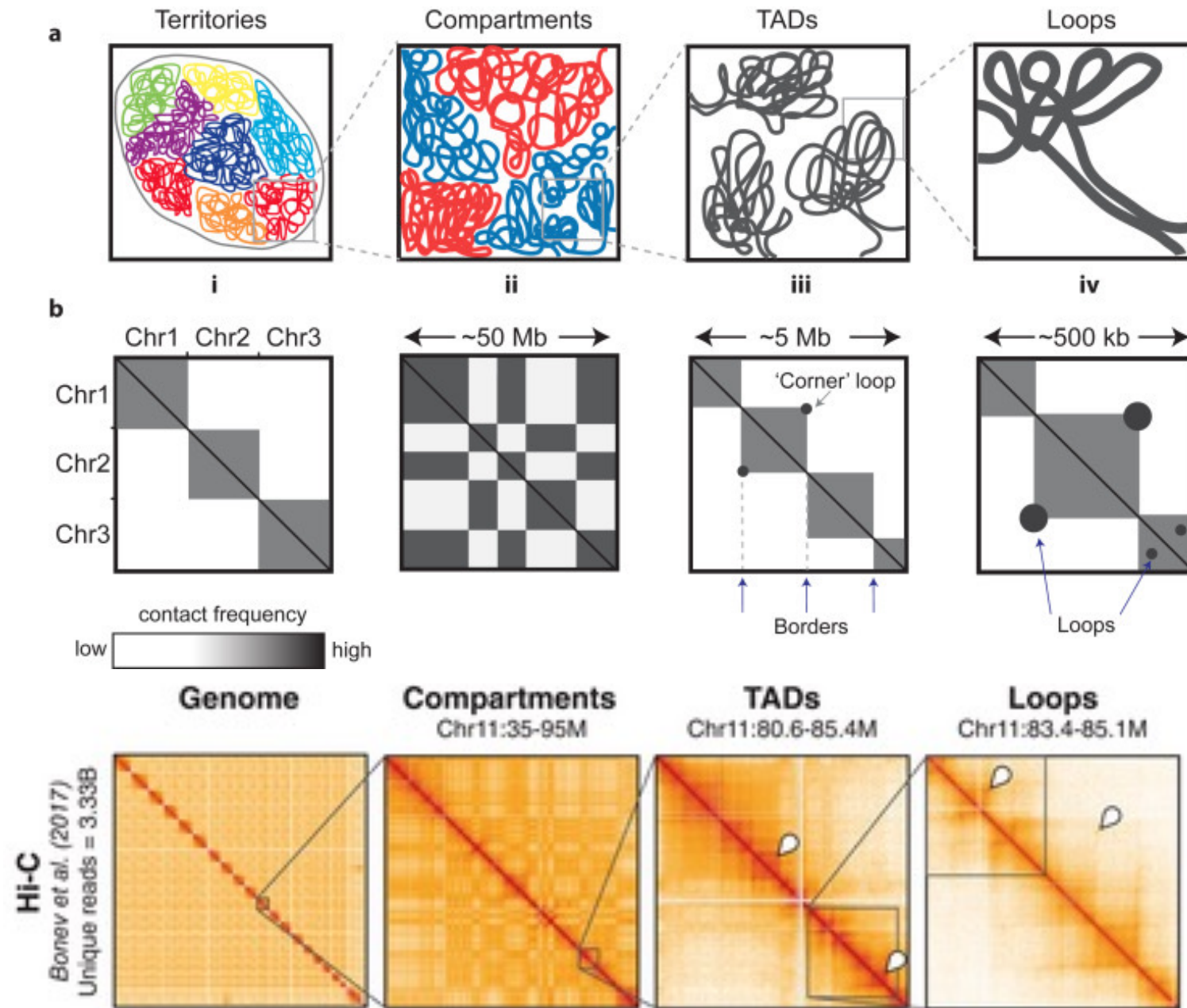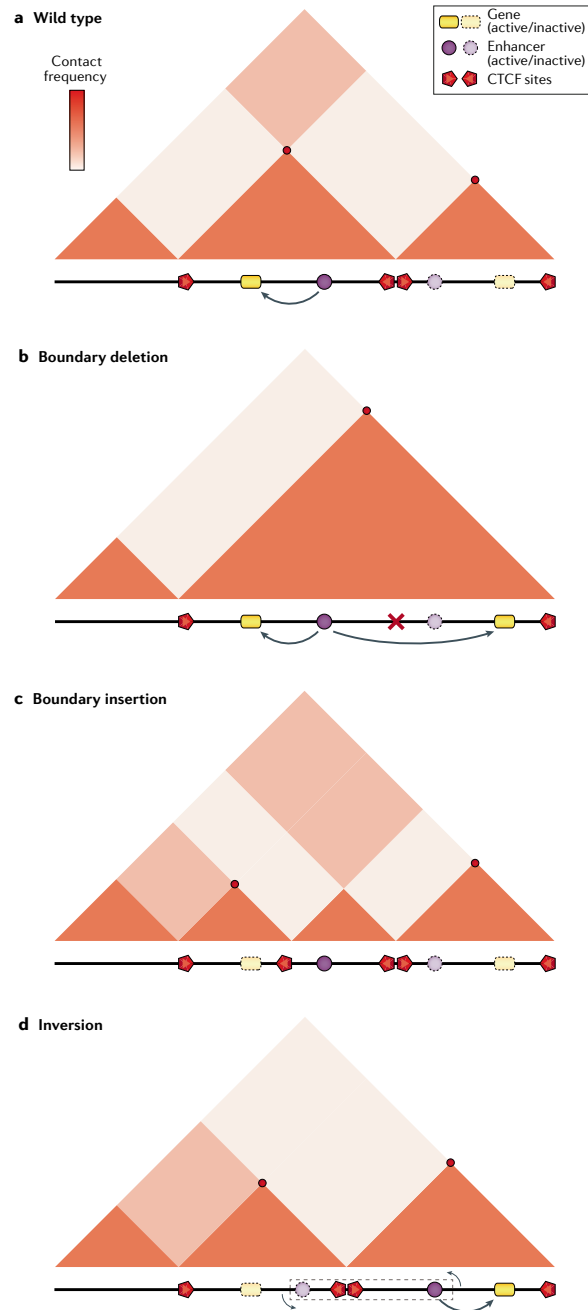
The relationship between the linear order of the regulatory elements and the organization of the genome.

In a hypothetical region of the genome, an active enhancer activates an upstream gene in the middle topologically associating domain (TAD). **b** | Deletion of the downstream boundary of the middle TAD (red cross) causes two TADs to merge; the active enhancer now also activates the downstream gene. **c** | Insertion of a new boundary element within the middle TAD causes the formation of an extra TAD, with the active enhancer and the upstream gene now located in two separate TADs. The enhancer can no longer interact with and activate the upstream gene promoter. **d** | Inversion of the region highlighted by the dashed box repositions the active enhancer in a new TAD. It can no longer interact with the upstream gene promoter, but instead activates the downstream promoter. Note that in each scenario the changes in activity within the domains alter their long-range compartmentalization. CTCF, CCCTC-binding factor.

Master Transcriptional Regulators

The set of genes that are transcribed largely defines the cell. The gene expression program of a specific cell type includes RNA species from genes that are active in most cells (**housekeeping genes**) and genes that are active predominantly in one or a limited number of cell types (**cell-type-specific genes**).

Transcription factors that have dominant roles in the control of specific cell states are <u>capable of reprogramming cell states when ectopically expressed in various cell types</u>.

Fibroblast	B-cell	Fibroblast	Pancreatic exocrine	Fibroblast	Fibroblast	Fibroblast	Fibroblast	Fibroblast	Hepatocyte	Fibroblast
MyoD	C/EBPα or C/EBPβ	OCT4 SOX2 KLF4 MYC	PDX1 NGN3 MAFA	PRDM16 C/EBPβ	BRN2 ASCL1 MYTL1 (NEUROD1)	GATA4 TBX5 MEF2C	GATA4 HNF1A FOXA3 or HNF4A FOXA1/2/3	BRN2 ASCL1 MYTL1	NR5A1 WT1 DMRT1 GATA4 SOX9
Muscle	Macrophage	Pluripotent stem cell	Islet β-cell	Brown fat	Neuron	Cardio-myocyte	Hepatocyte	Neuron	Sertoli

Transcription factors

- Transcription factors can be separated into two classes based on their regulatory responsibilities: **control of initiation** versus **control of elongation**. This distinction is not absolute, as some transcription factors may contribute to control of both initiation and elongation.

- Transcription factors typically bind cofactors, which are protein complexes that contribute to activation (**coactivators**) and repression (**corepressors**).

- These coactivators include **histone modifiers** (e.g. HAT) and the **Mediator complex**, among others.

Transcription factors form a vast network

In addition to contributing to the annotation of functional DNA elements, the ENCODE project has enabled analysis of the connections between these linear elements, such as the interaction network among transcription factors and long-range chromatin interactions.

The analysis of target sites for 119 transcription factors examined by the ENCODE project using ChIP-seq has revealed the presence of a hierarchical interaction network. At the top of the pyramid are the most powerful transcription factors, while further down are factors that are more regulated than regulatory.

[www.regulatorynetworks.org](http:/ /www.regulatorynetworks.org): links between specific transcription factors in certain cell types

Connection networks among transcription factors show high cell specificity

(A) Network di regolazione trascrizionale completo (per 475 fattori)

Collegamenti cellula-specifici

Collegamenti non unici

Collegamenti comuni

Esterno

475 TF distribuiti in modo identico

TF con basso grado di connettività nelle ESC

TF con alto grado di connettività nelle ESC

Interno

Epitelio corticale renale
Collegamenti unici: 789
Collegamenti totali: 7586

Cellule embrionali staminali
Collegamenti unici: 3366
Collegamenti totali: 13 176

Cellule staminali ematopoietiche
Collegamenti unici: 2146
Collegamenti totali: 13 240

Mioblasti scheletrici
Collegamenti unici: 1269
Collegamenti totali: 10 969

Endotelio dei vasi del derma neonatale
Collegamenti unici: 1764
Collegamenti totali: 13 311

Cervello fetale
Collegamenti unici: 831
Collegamenti totali: 9293

475 transcription factors in 6 connection networks of different cell types.
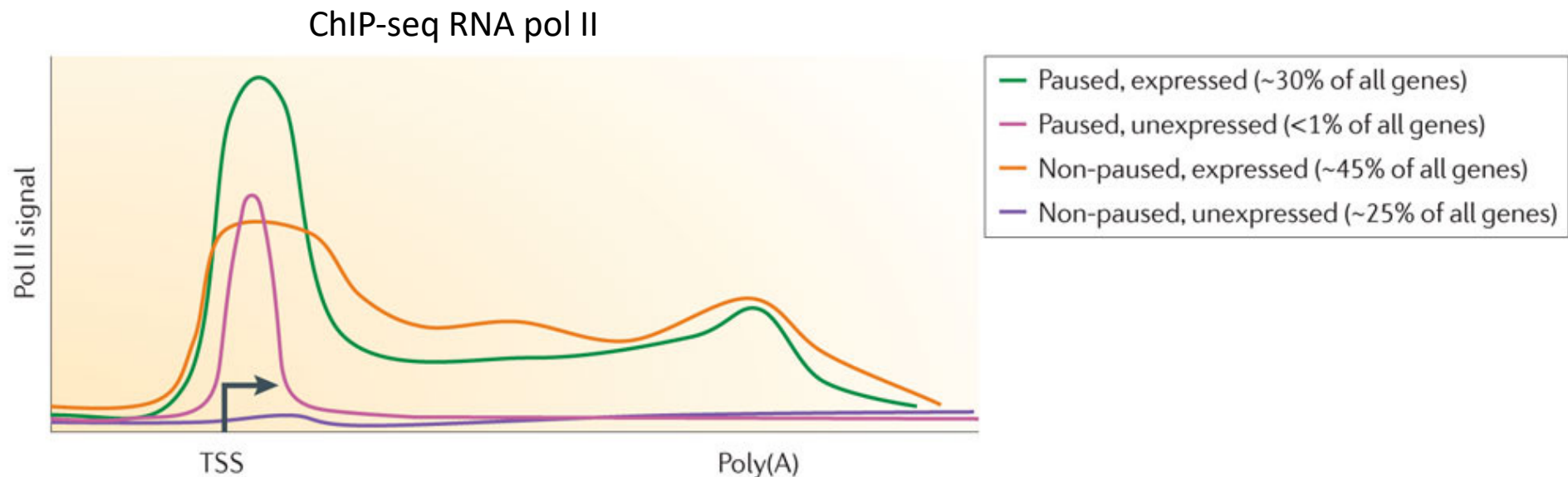
Connection networks among transcription factors show high cell specificity

(B) Network di livello successivo relativo a cellule ematopoietiche e a fattori di pluripotenza (8 fattori)

Cellule embrionali staminali

Interno — Esterno

Mioblasti scheletrici

Esterno

KLF4
NANOG
OCT4
SOX2
} Fattori di pluripotenza

PU.1
TAL1
ELF1
GATA2
} Fattori ematopoietici

Interno
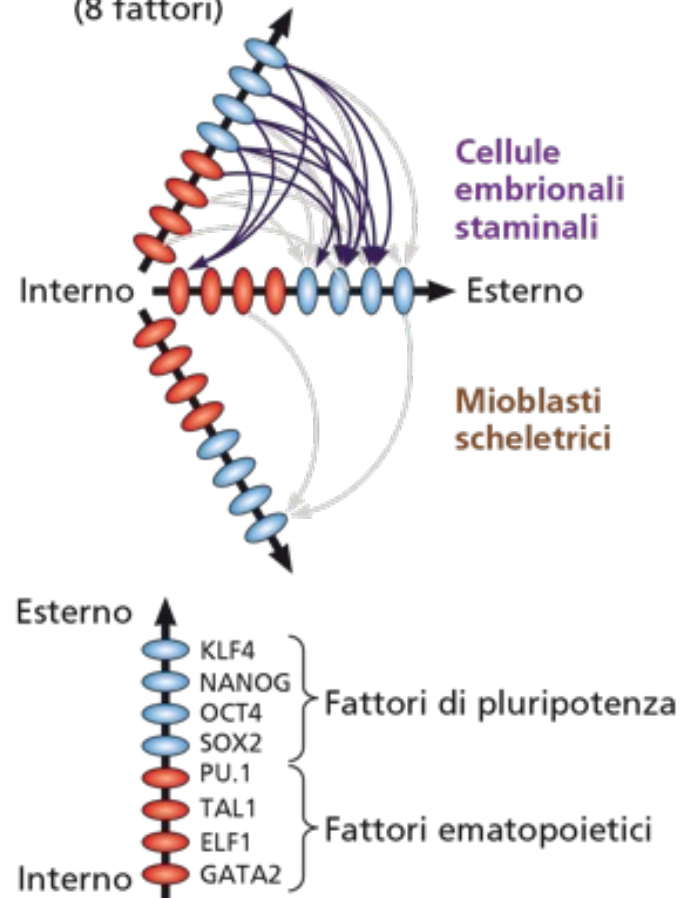
Transcriptional control by RNA pol II pausing

Whereas traditional models focused solely on the events that brought RNA Pol II to a gene promoter to initiate RNA synthesis, emerging evidence points to the <u>pausing of Pol II during early elongation as a widespread regulatory mechanism in higher eukaryotes</u>.

ChIP-seq RNA pol II

Legend:
— Paused, expressed (~30% of all genes)
— Paused, unexpressed (<1% of all genes)
— Non-paused, expressed (~45% of all genes)
— Non-paused, unexpressed (~25% of all genes)

Pol II signal

TSS   Poly(A)

Nature Reviews | Genetics

Transcriptional control by RNA pol II pausing

Once the recruited RNA Pol II molecules initiate transcription, they generally transcribe a short distance, typically 20–80 bp, and then pause. This process is controlled by the factors **DSIF** and **NELF**, which are physically associated with the paused RNA Pol II. The paused polymerases may transit to active elongation through pause release, or may ultimately terminate transcription with release of the small RNA species.

Pause release and subsequent elongation occur through recruitment and activation of **positive transcription elongation factor b (P-TEFb)**, which phosphorylates the Ser2 of paused polymerase and its associated pause control factors.

Transcription factors may stimulate P-TEFb-mediated release of RNA polymerase II from these pause sites and thus contribute to the control of transcription elongation
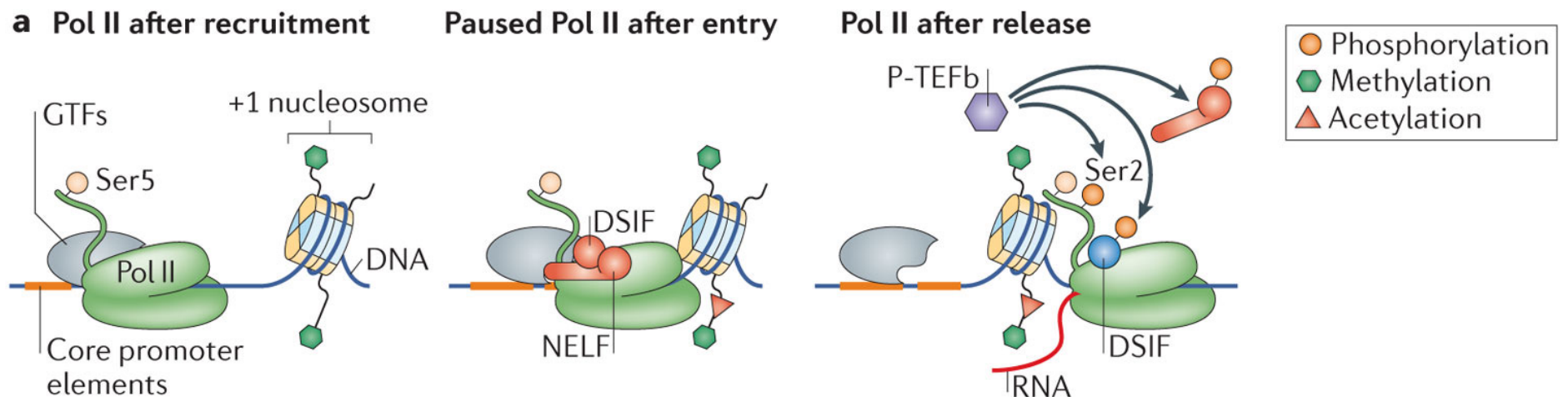
Establishment and release of paused Pol II

a. Promoter opening often involves binding of transcription factor (TF) that brings in chromatin remodellers to remove nucleosomes from around the TSS and to render the promoter accessible for recruitment of the transcription machinery.

b. Pre-initiation complex formation involves the recruitment of a set of general transcription factors and Pol II. This step precedes the initiation of RNA synthesis. **c. Pol II pausing** occurs shortly after transcription initiation and involves the association of pausing factors DSIF and NELF. The paused Pol II is phosphorylated on its CTD domain. **d. Pause release** is triggered by the recruitment of the P-TEFb kinase, either directly or indirectly by a transcription factor (TF2). P-TEFb kinase phosphorylates the DSIF–NELF complex to release paused Pol II. **e. Phosphorylation of DSIF–NELF** dissociates NELF from the elongation complex and transforms DSIF into a positive elongation factor that associates with Pol II throughout the gene.
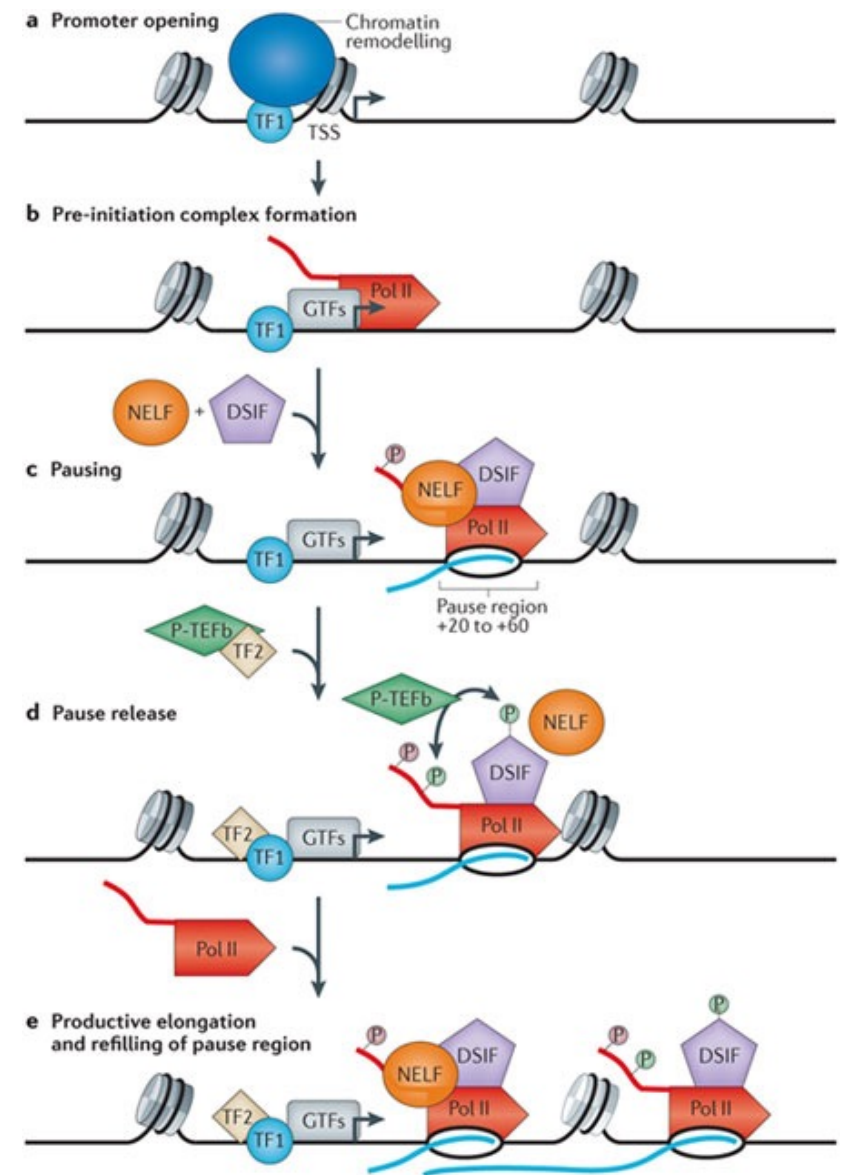
7SK RNA regulated p-TEFb activity

7SK RNA forms the core of a small nuclear ribonucleoprotein (snRNP) complex that minimally contains P-TEFb, and the proteins HEXIM1/2 and LARP7, which control the activity of P-TEFb and stabilize 7SK RNA. The P-TEFb kinase is inactive when present in the 7SK snRNP complex and becomes active when it dissociates from the complex.