

Gene Expression and Chromatin: The Epigenetic Gatekeeper

Focus: The structural basis of gene control.

Goal: To understand how DNA is organized and accessed for transcription.

Chapter 2 of Gene Regulation and Epigenetics, Carsten Carlberg (2024)

Nov 09 2025

Giảng viên: TS. Lưu Phúc Lợi

Email: Luu.p.loi@googlemail.com

Zalo: 0901802182

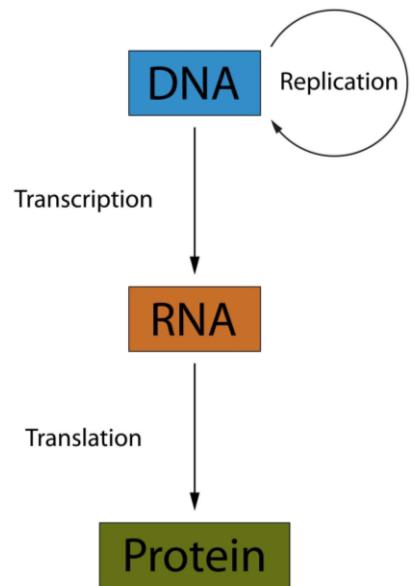
Content

1. Gene Expression defines cell identity and function.
2. Introduction to RNA-seq.
3. The distinct roles of RNA Polymerases I, II, and III.
4. The structure and components of the nucleosome.
5. The functional difference between Euchromatin and Heterochromatin.
6. **Epigenetics** as the heritable, non-sequence layer of information.
7. The composition and purpose of **histone variants** and H1.
8. The significance of **3D Nuclear Architecture** (LADs, loops, TADs).
9. Chromatin's role as the "**Gatekeeper**" of the genome.

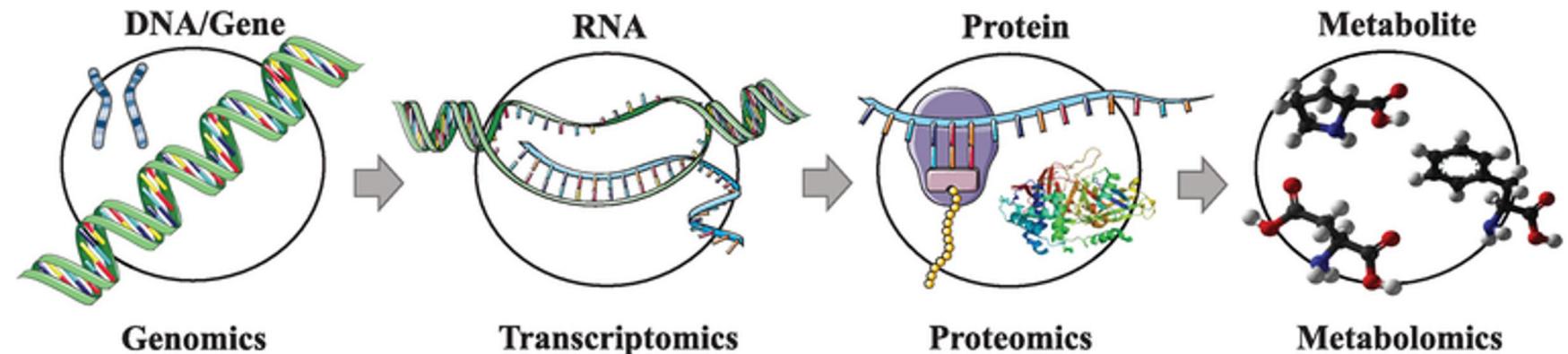
Introduction and Central Dogma

The Central Dogma and Regulation

The central dogma of molecular biology



Central Dogma



PAH gene
Ref ...ATCGAT...
P1 ...AACGAT...

NM_000277.3(PAH):c.971T>A

PAH mRNA
Ref ...AU~~C~~GAU...
P1 ...AA~~C~~GAU...

NM_000277.3(PAH):c.971T>A

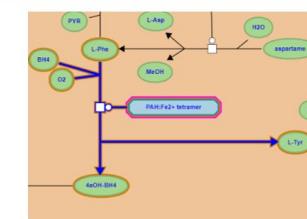
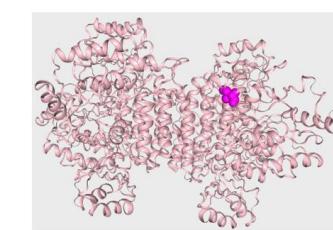
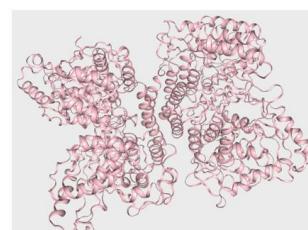
PAH protein
Ref ...Ile-Asp...
P1 ...Asn-Asp...

NM_000277.3(PAH):p.Ile324Asn

PAH
Ref Phe → Tyr

PAH
P1 Phe ~~→~~ Tyr

10.5772/intechopen.71769

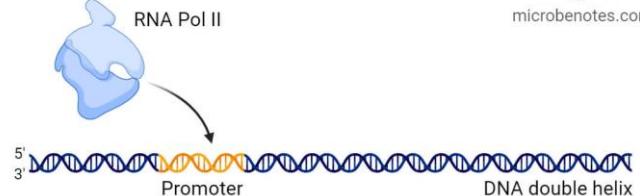


Gene Expression (Transcriptomics): Information from DNA to RNA

- ▶ Gene Expression is the process that converts genetic information into a cellular trait.

Eukaryotic Transcription

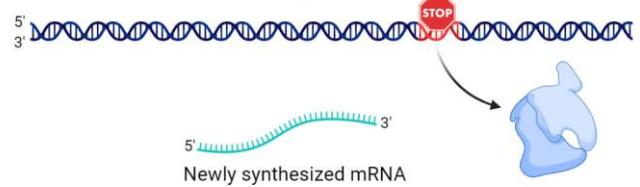
1
Initiation



2
Elongation



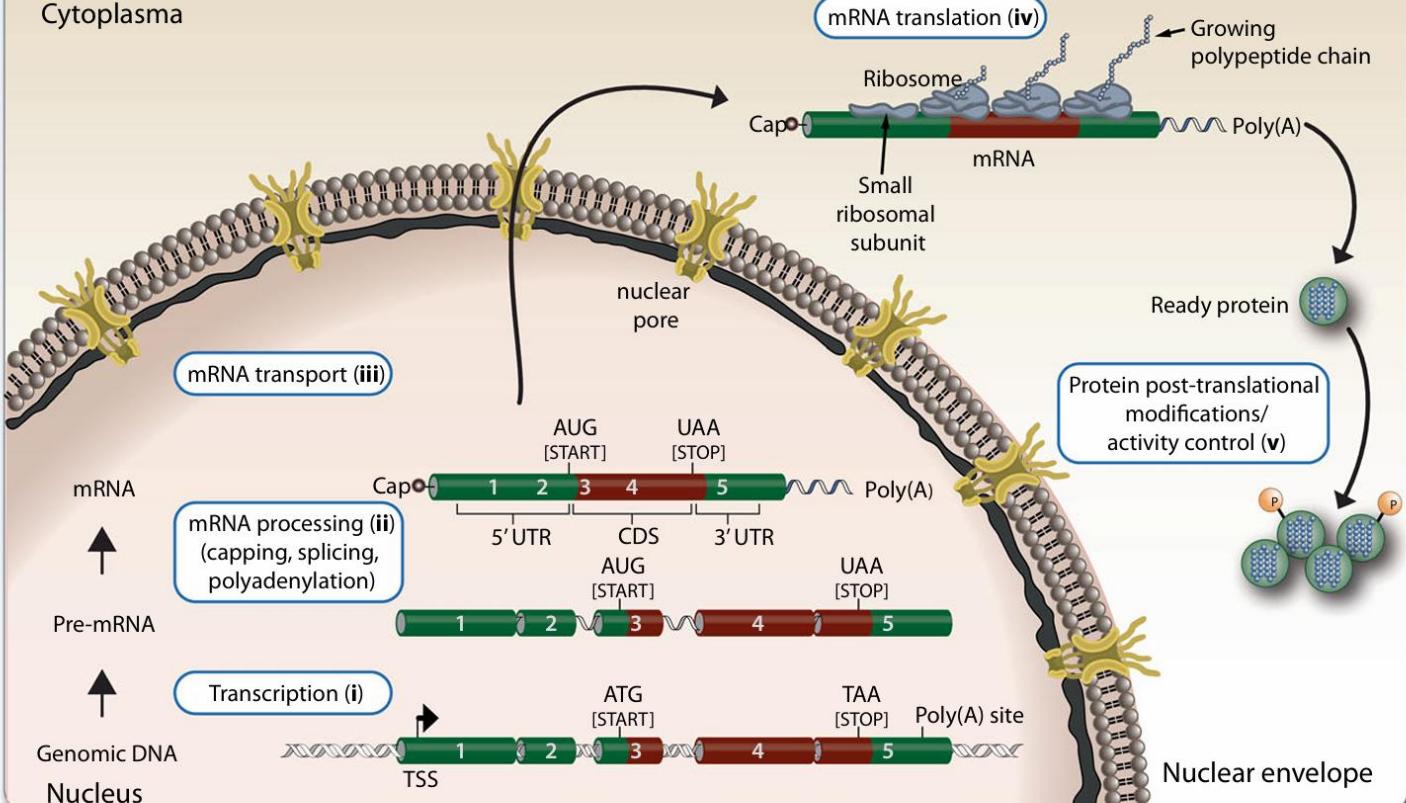
3
Termination



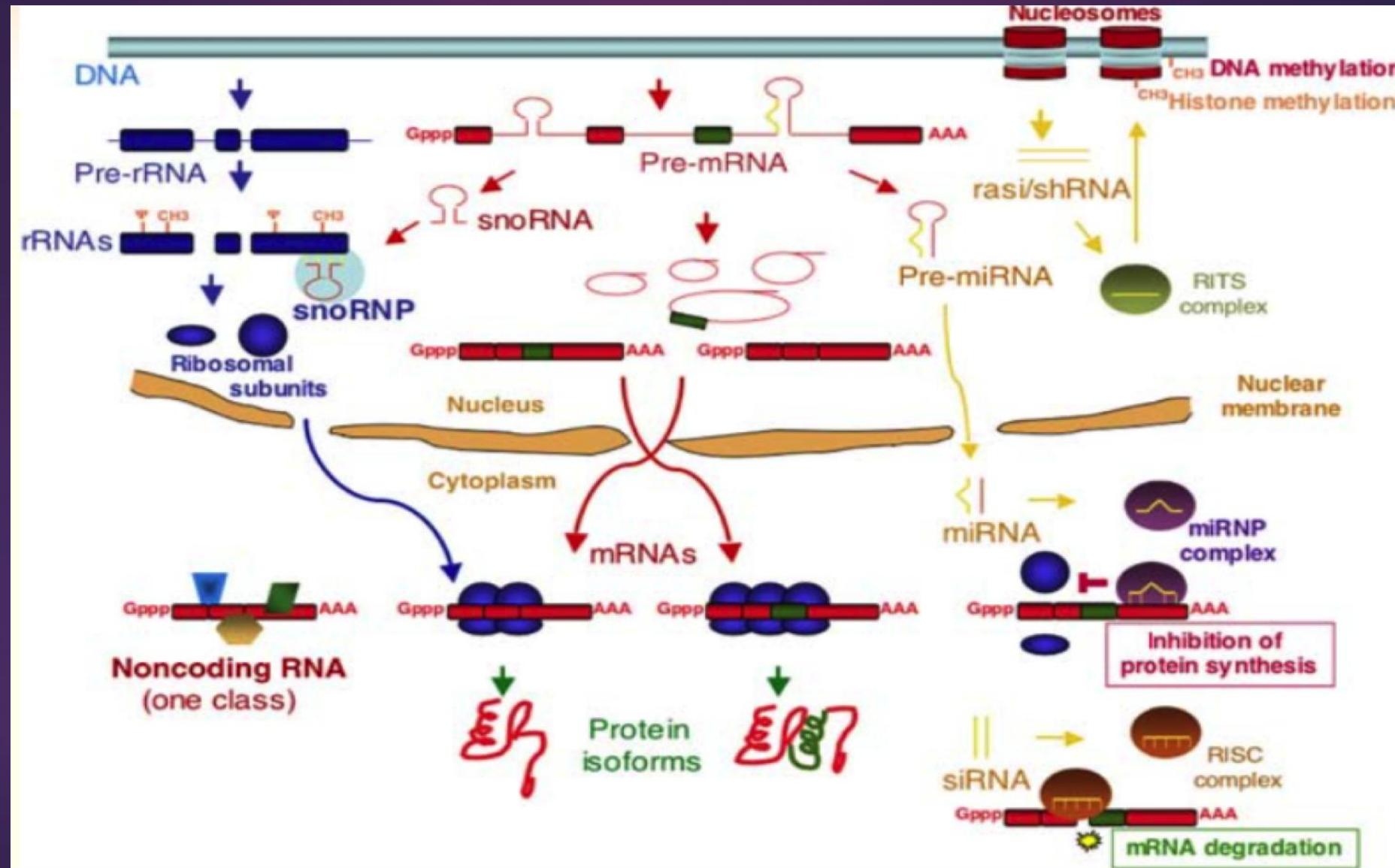
microbenotes.com

Cellular membrane

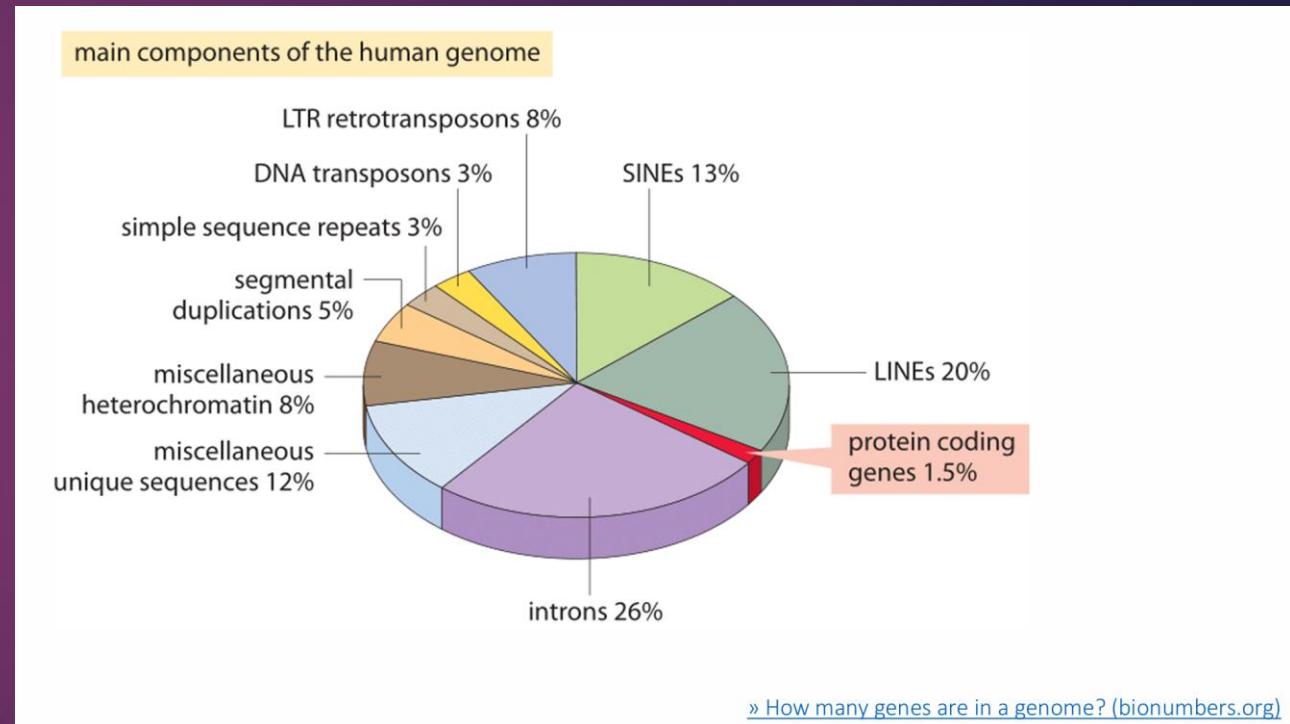
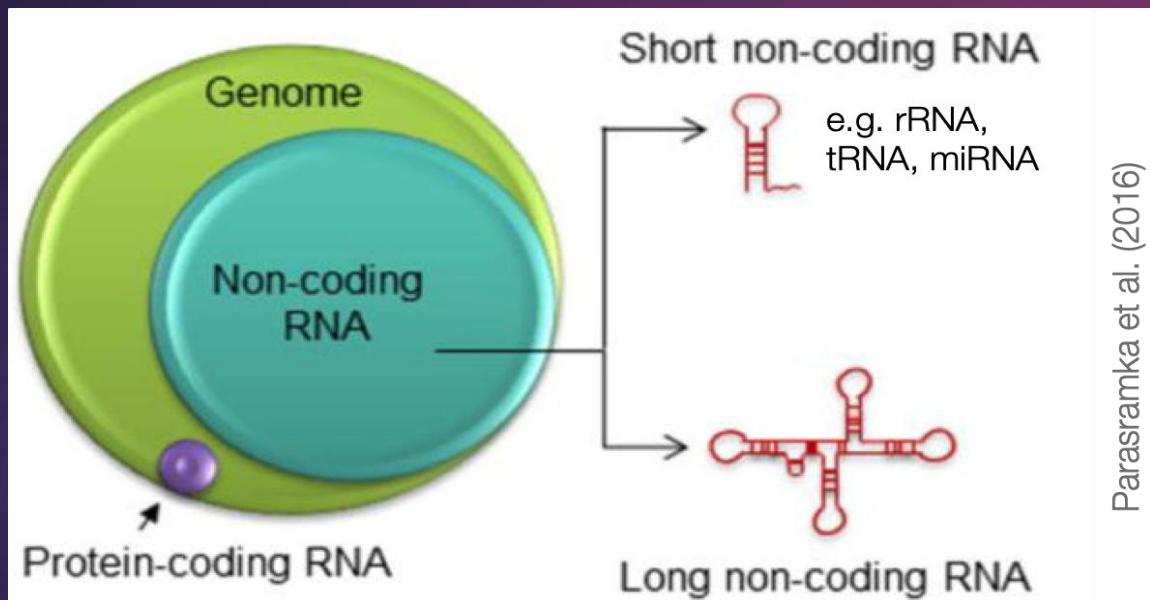
Cytoplasma



Different types of RNA



Protein coding gene is only 1.5%



Protein coding gene is only 1.5%

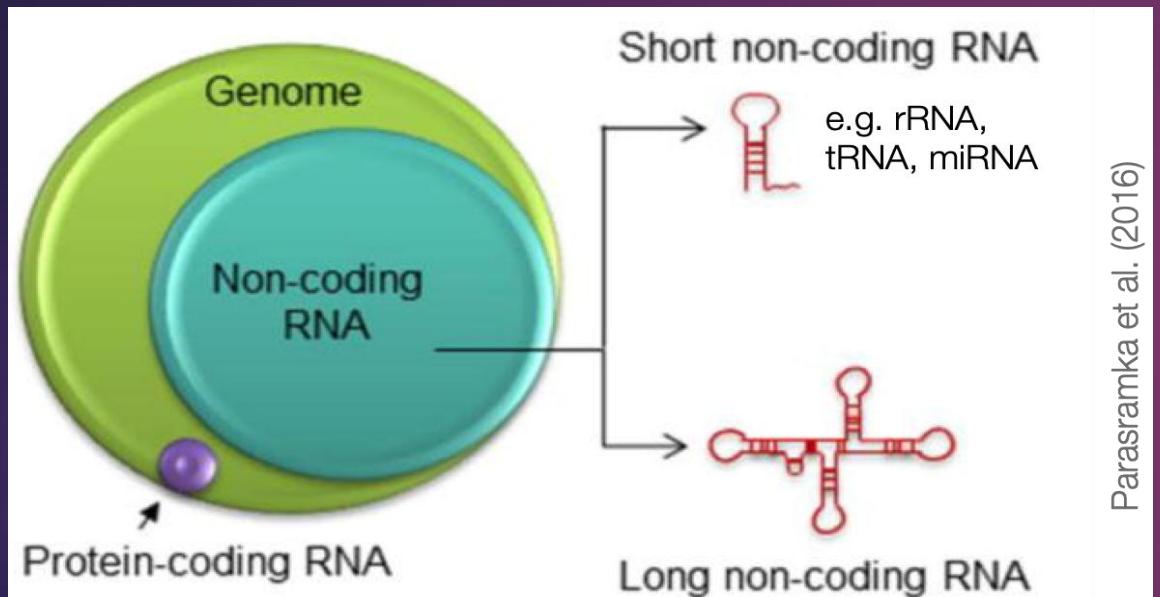


Table 2.1 The human genome in numbers

Parasramka et al. (2016)

Number of chromosomes	22 + X + Y
Genome size (nt)	3,054,815,472
Number of genes	58,037
Number of transcripts	203,835
Number of protein-coding genes	19,901
Number of protein-coding transcripts	80,087
Number of long ncRNA transcripts	15,779
Number of pseudogenes	14,723
Number of small RNAs	7258
Number of miRNAs	2588
Number of tRNAs	631

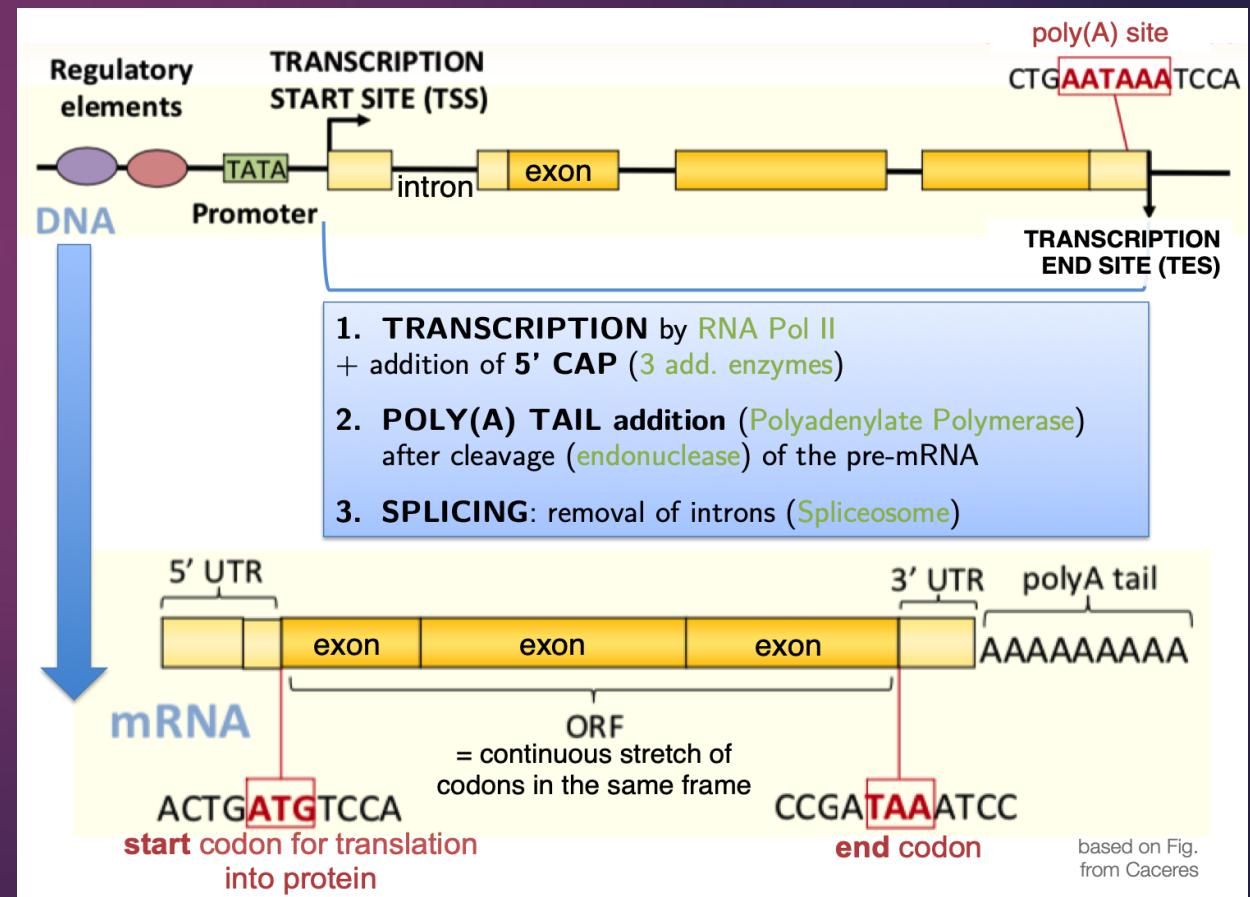
The size, number or genes and transcripts of the latest version of the human genome (hg38) are indicated. Pseudogenes are regions of the genome that contain defective copies of genes. These numbers may still change a bit based on new annotation analyses

Introduction to RNA-seq

Sequencing library prep protocol depends on the RNA properties

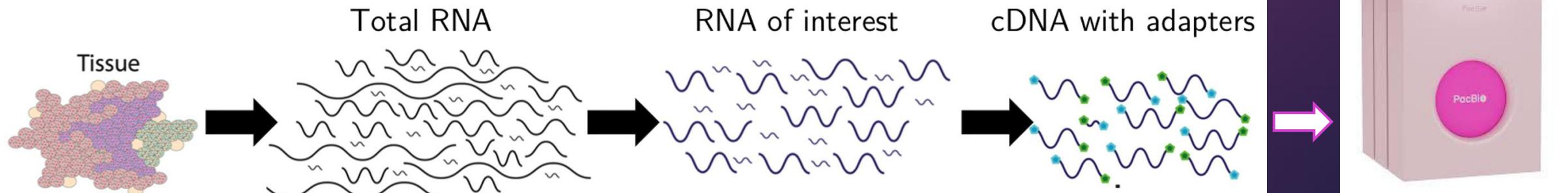
It is not a one-size-fits-all situation!

- Abundance and stability
 - ▶ rRNA: 90-95% (!)
 - ▶ tRNA: 3-5%
 - ▶ mRNA: 2%
 - ▶ all other non-coding RNAs: well below 1%
- Cellular location
 - ▶ most are in the cytoplasm
- Size
 - ▶ miRNAs: 18-23bp
 - ▶ mRNA: several 100 to 1000 bp
- Specific sequences/modifications
 - ▶ poly(A) tails of mRNA
 - ▶ 2D structure
 - ▶ antisense transcripts

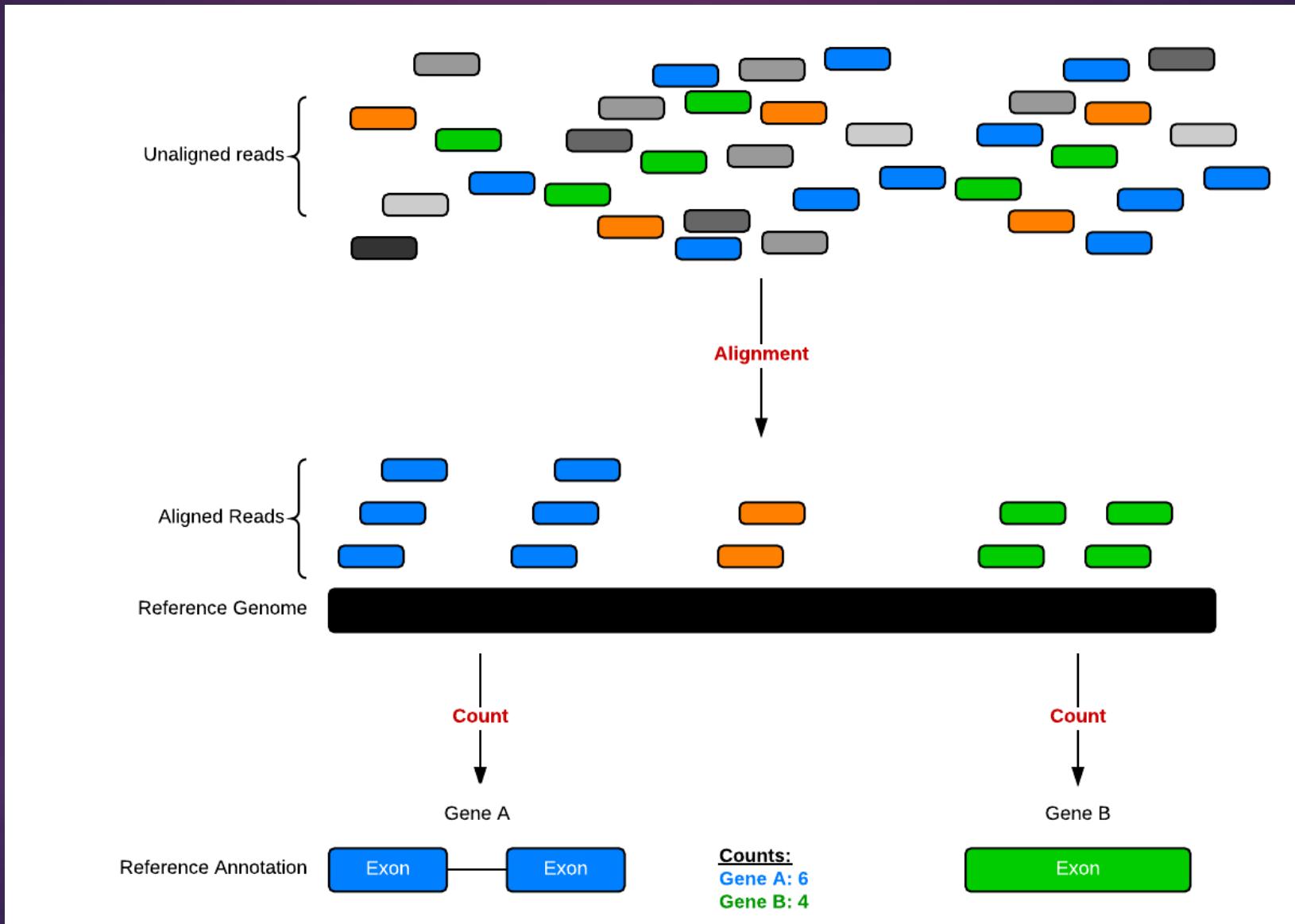


General steps of RNA-seq preparation

- ① RNA extraction (cell lysis, RNA purification)
- ② enrichment of the RNA of interest
 - ▶ mRNA: poly(A) enrichment vs. ribosomal-depletion
 - ▶ small RNAs: size-based enrichment
- ③ fragmentation (ca. 200 bp)
- ④ cDNA synthesis
- ⑤ library prep to obtain cDNA with adapters for sequencing



How to analyse data of RNA-seq?



RNA-seq count table

countData

gene	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
...
...
...

colData

id	treatment	sex
ctrl_1	control	male
ctrl_2	control	female
exp_1	treatment	male
exp_2	treatment	female

Sample names:

ctrl_1, ctrl_2, exp_1, exp_2

countData is the count matrix
(number of reads mapping to each gene for each sample)

colData describes metadata about the *columns* of countData

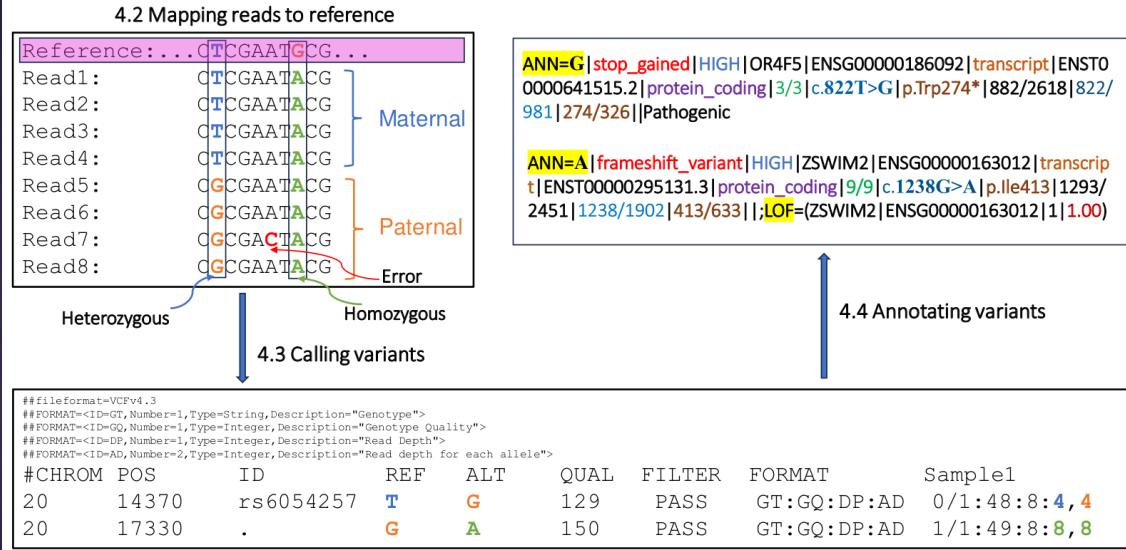
First column of colData must match column names of countData (-1st)

RNA-seq count table

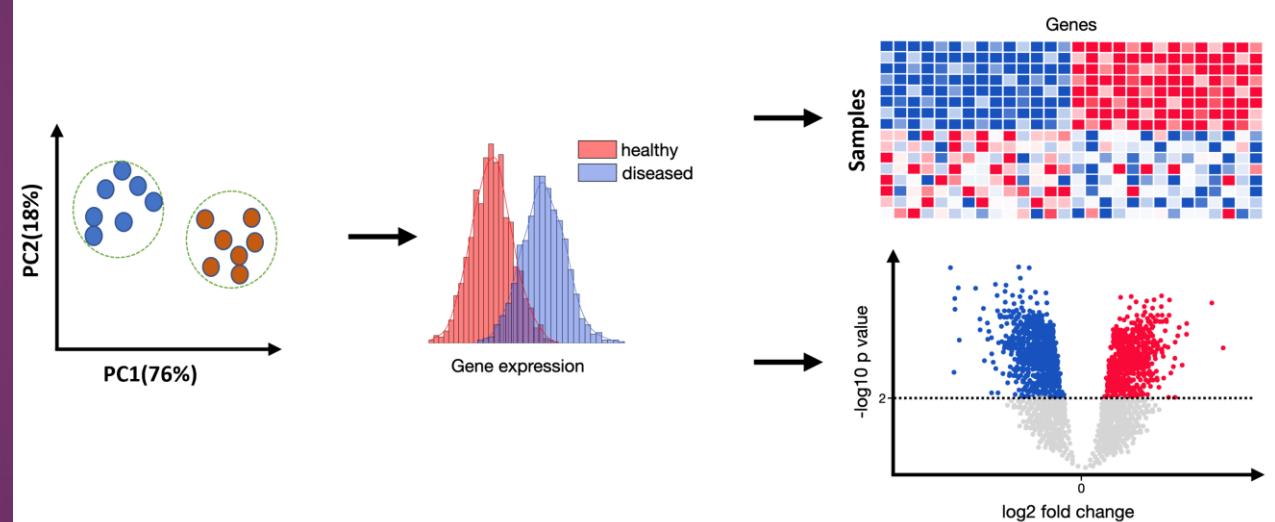
```
## # A tibble: 38,694 x 9
##       ensgene SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 ENSG00000000003     723      486      904      445     1170
## 2 ENSG00000000005      0        0        0        0        0
## 3 ENSG00000000419     467      523      616      371      582
## 4 ENSG00000000457     347      258      364      237      318
## 5 ENSG00000000460      96       81       73       66      118
## 6 ENSG00000000938      0        0        1        0        2
## 7 ENSG00000000971    3413     3916     6000     4308     6424
## 8 ENSG00000001036    2328     1714     2640     1381     2165
## 9 ENSG00000001084    670      372      692      448      917
## 10 ENSG00000001167    426      295      531      178      740
## # ... with 38,684 more rows, and 3 more variables: SRR1039517 <dbl>,
## #   SRR1039520 <dbl>, SRR1039521 <dbl>
```

Aim of the methods

Bulk DNA-seq



Bulk RNA-seq



Aim of the methods

Bulk DNA-seq

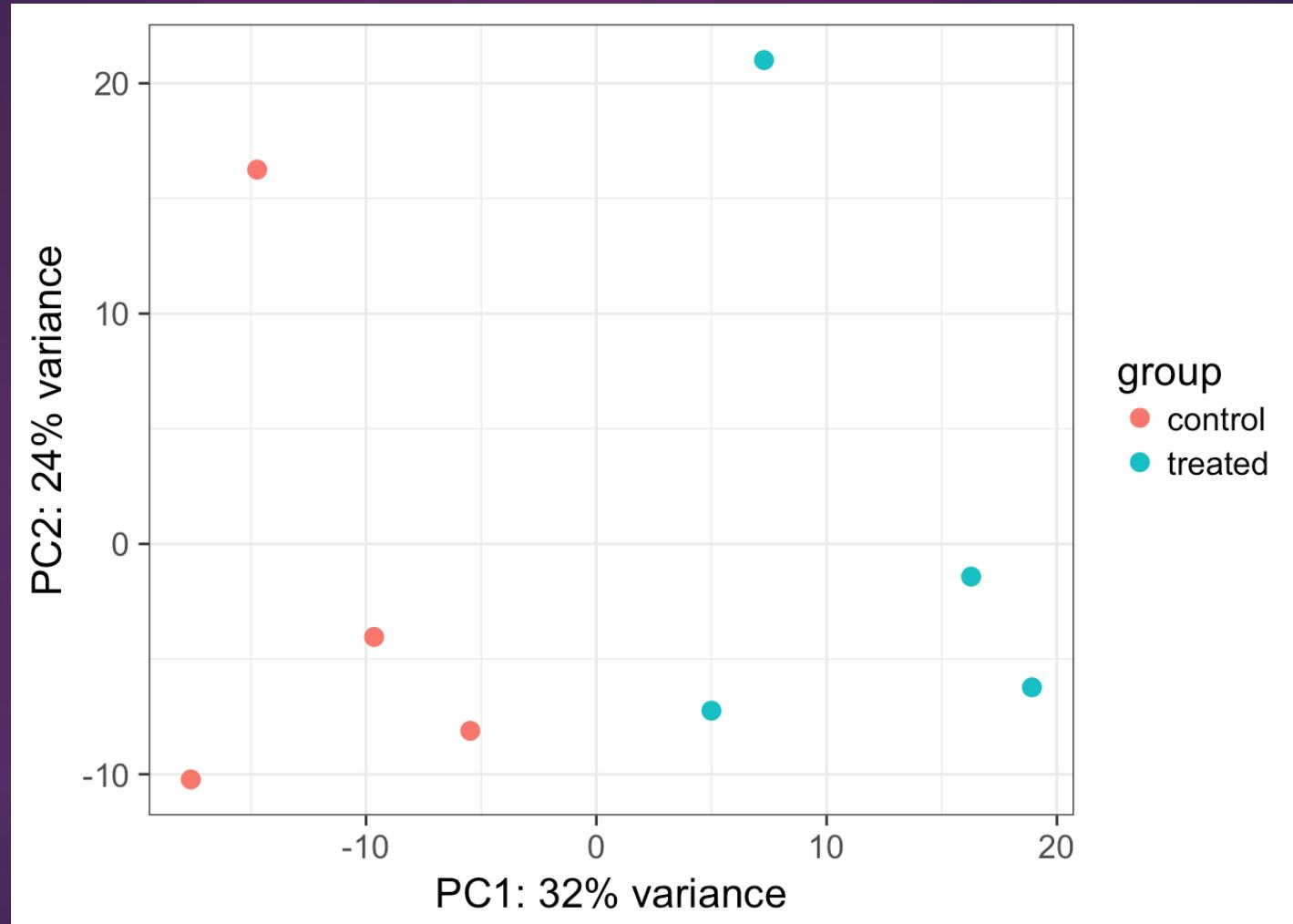
ANN=G|stop_gained|HIGH|OR4F5|ENSG00000186092|transcript|ENST0000641515.2|protein_coding|3/3|c.822T>G|p.Trp274*|882/2618|822/981|274/326||Pathogenic

ANN=A|frameshift_variant|HIGH|ZSWIM2|ENSG00000163012|transcrip
t|ENST00000295131.3|protein_coding|9/9|c.1238G>A|p.Ile413|1293/
2451|1238/1902|413/633||;LOF=(ZSWIM2|ENSG00000163012|1|1.00)

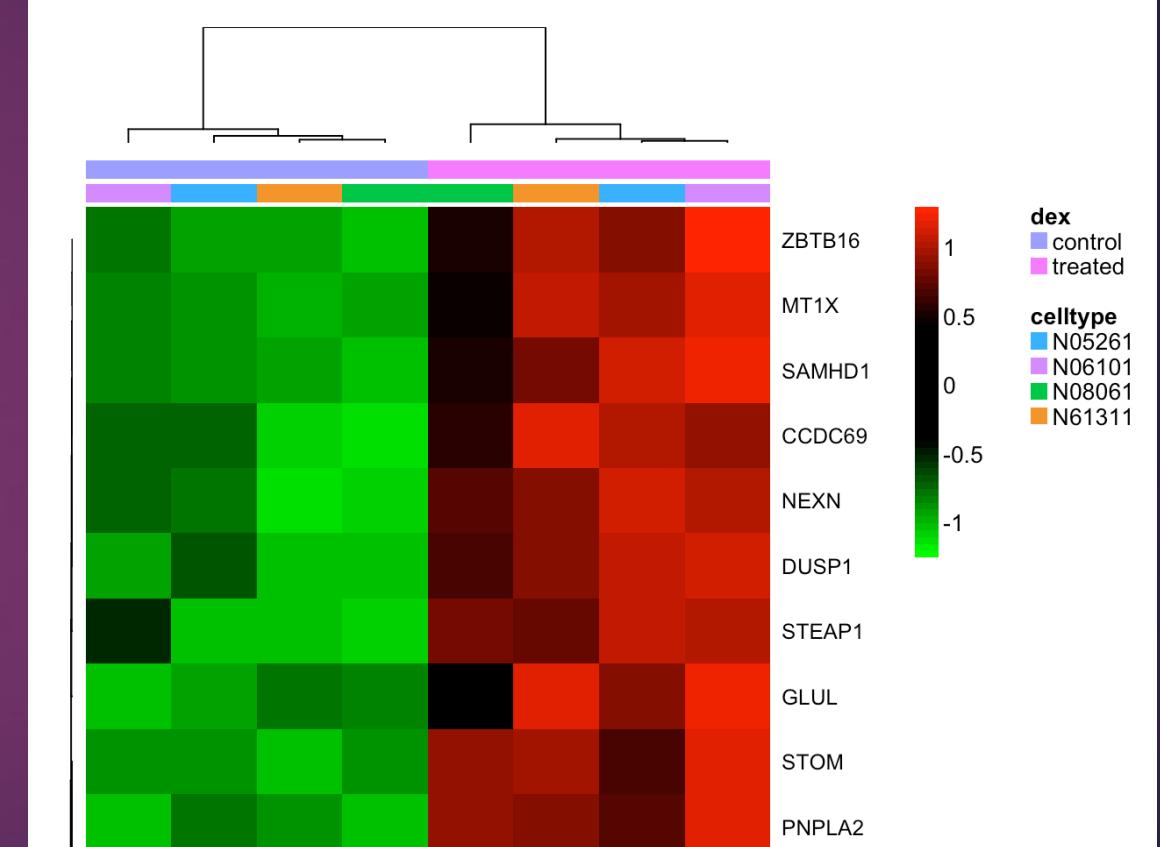
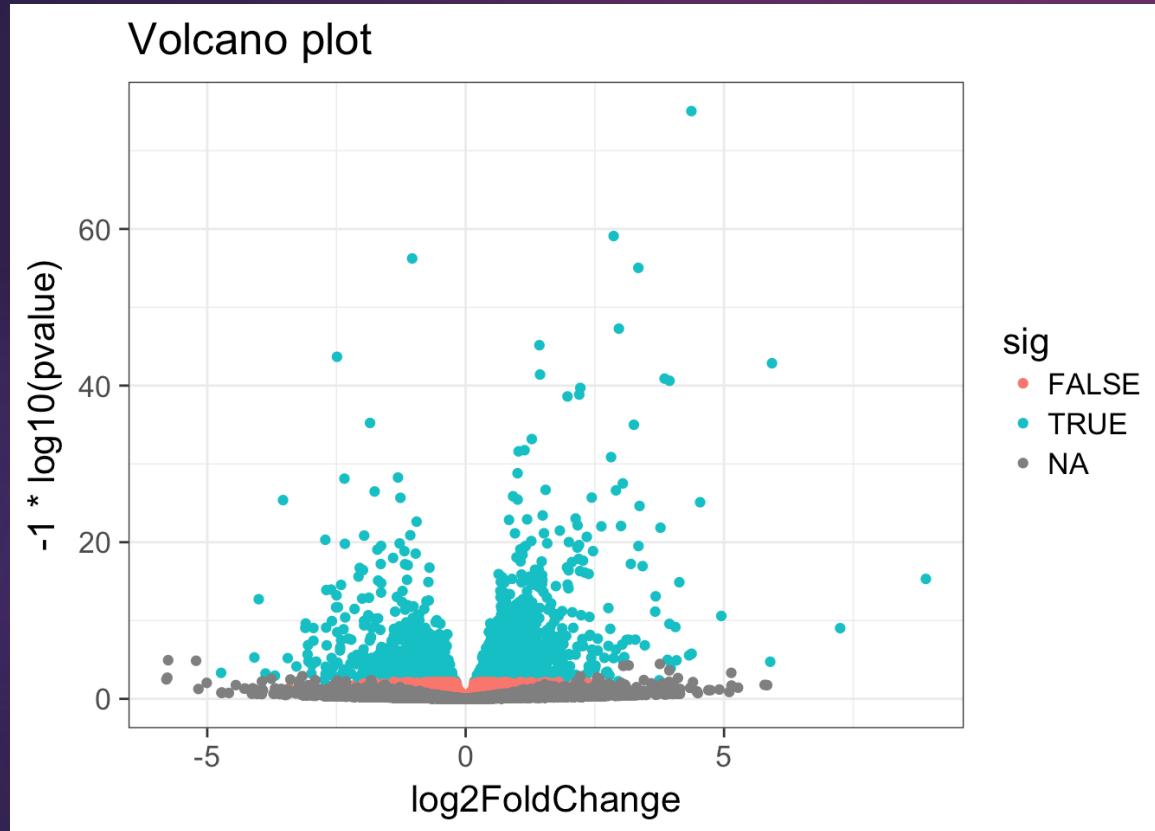
Bulk RNA-seq

Gene	Sample 1 Healthy	Sample 2 Healthy	Sample 3 Healthy	Sample 4 Tumor	Sample 5 Tumor	Sample 6 Tumor
A	100	93	87	160	163	154
B	100	80	92	1	2	3
C	45	45	45	146	146	146
D	111	121	134	0.0012	0.0014	0.0013

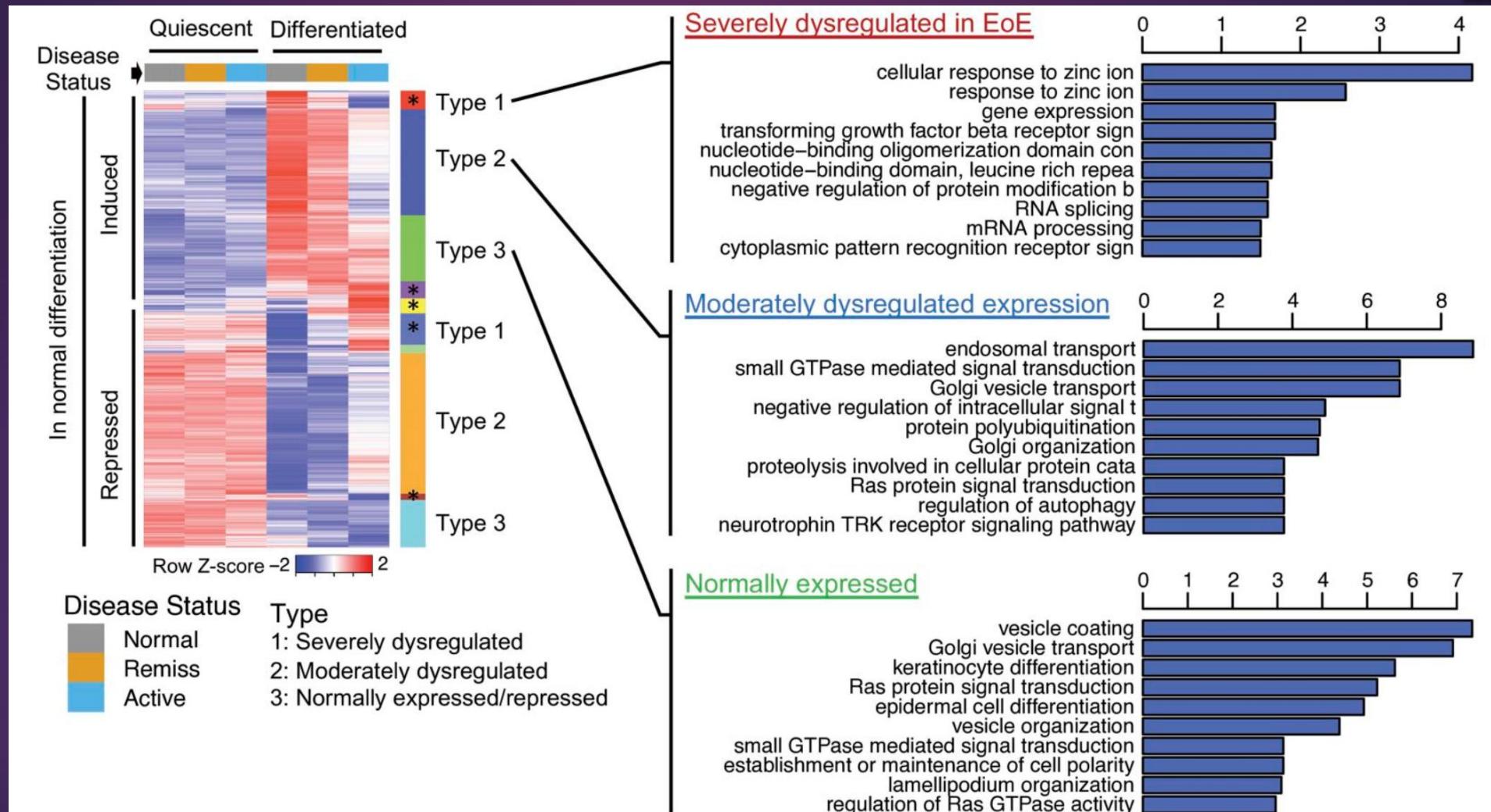
RNA-seq Downstream Analysis



RNA-seq Downstream Analysis



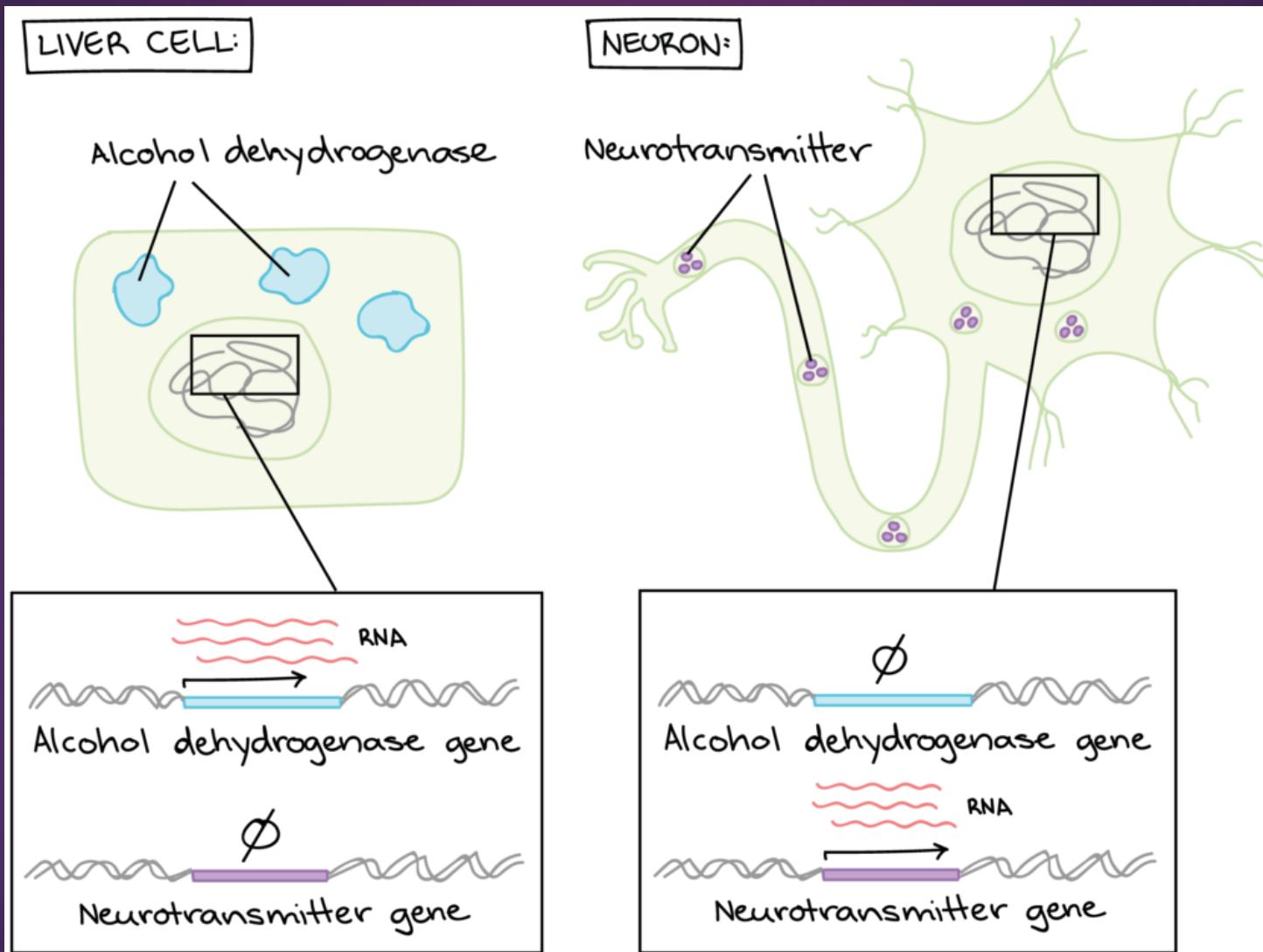
RNA-seq Downstream Analysis



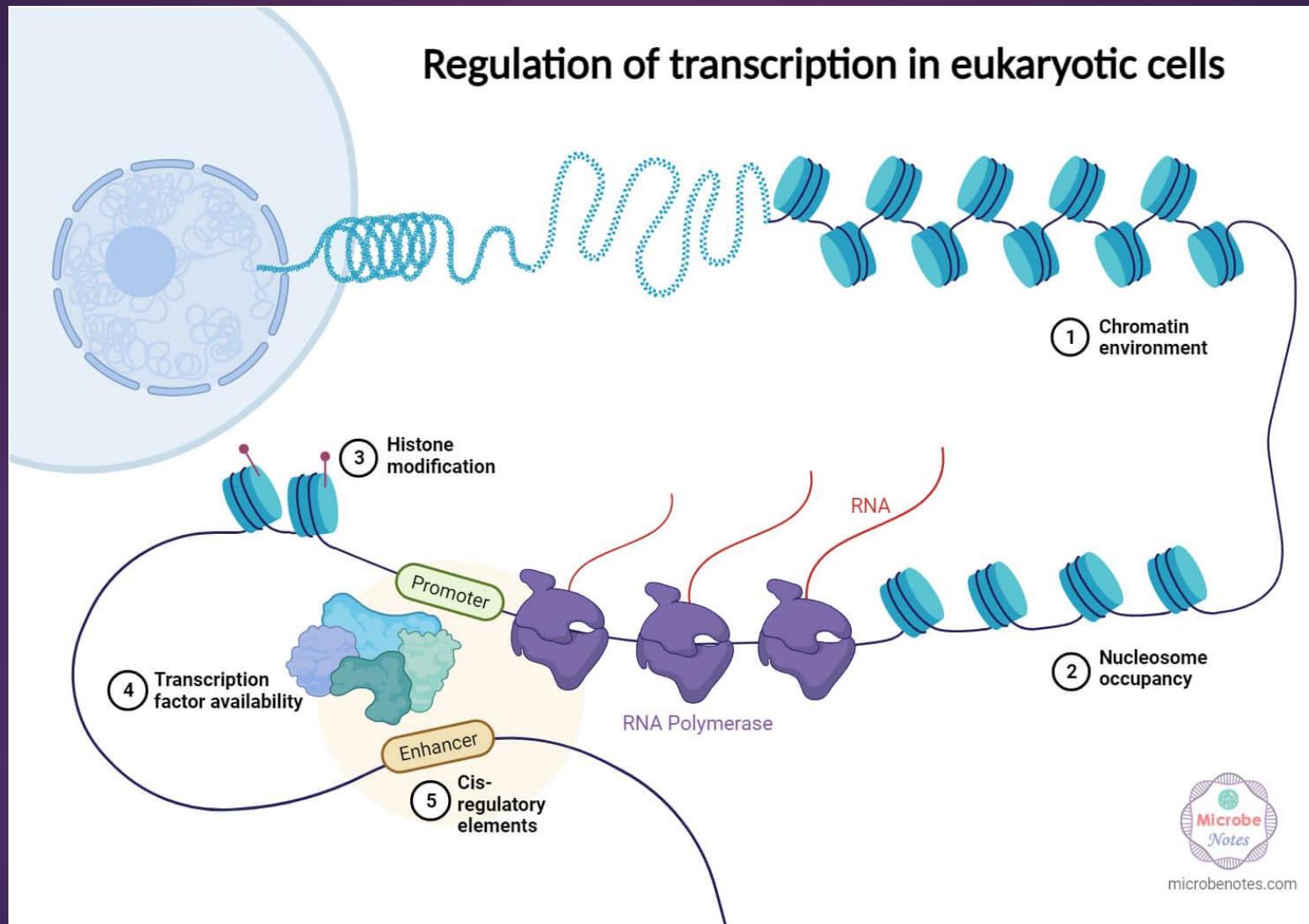
The Outcome of Gene Expression

- Gene Expression is the process that converts genetic information into a cellular trait.
- **Result:** Production of proteins (enzymes, structural elements, signals) that define the phenotype, function, and developmental state of a cell.
- **Key Regulatory Point:** Gene regulation in eukaryotes occurs primarily at the level of **Transcription Initiation**—controlling whether a gene is accessed by the RNA polymerase machinery.
- **Dynamic Nature:** Expression patterns are highly cell-specific but are constantly adjusted by internal and external signals (e.g., hormones, nutrients, stress).

The Outcome of Gene Expression



The Outcome of Gene Expression



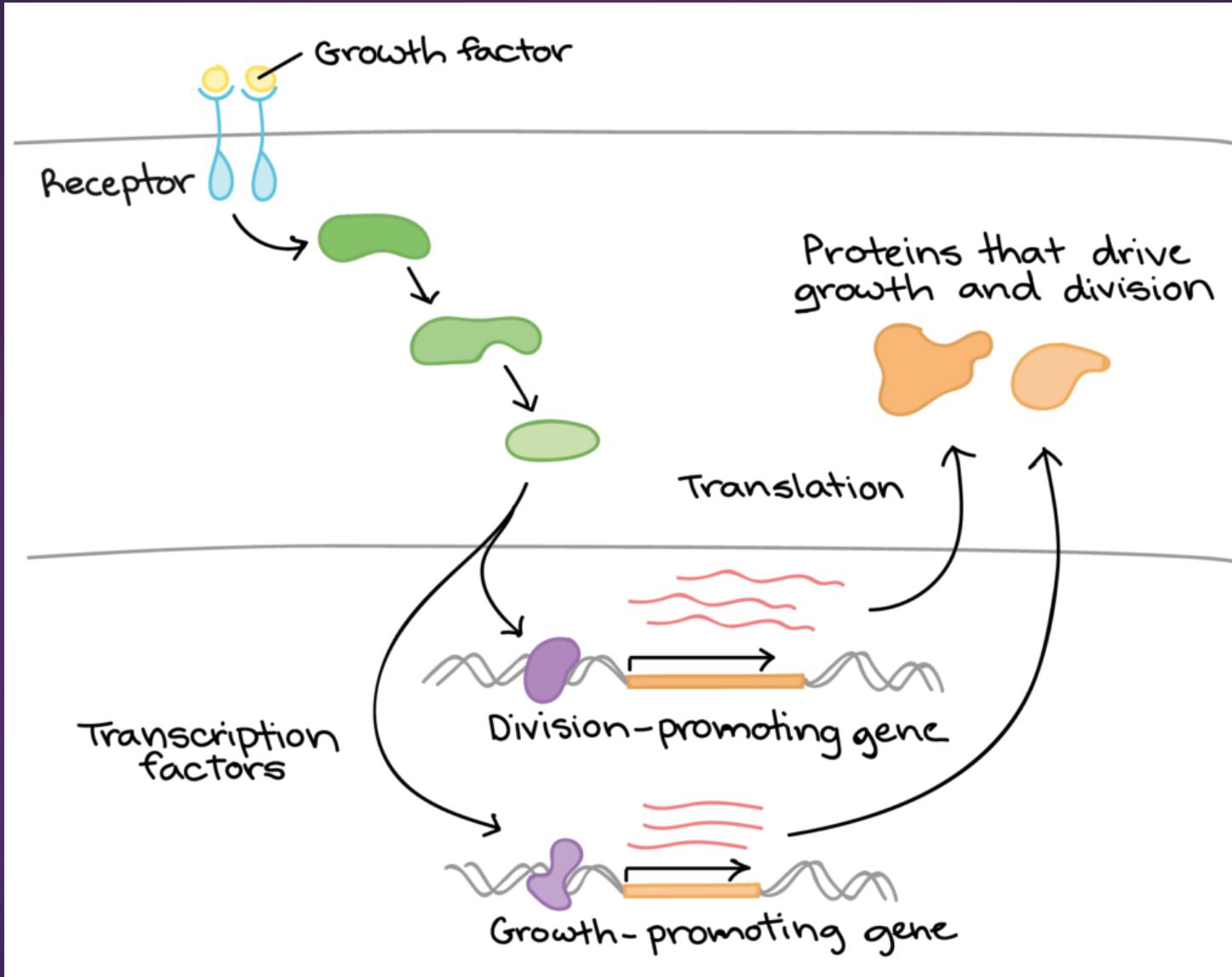
microbenotes.com

RNA Polymerases and Signaling

Signal Transduction Cascades (Input)

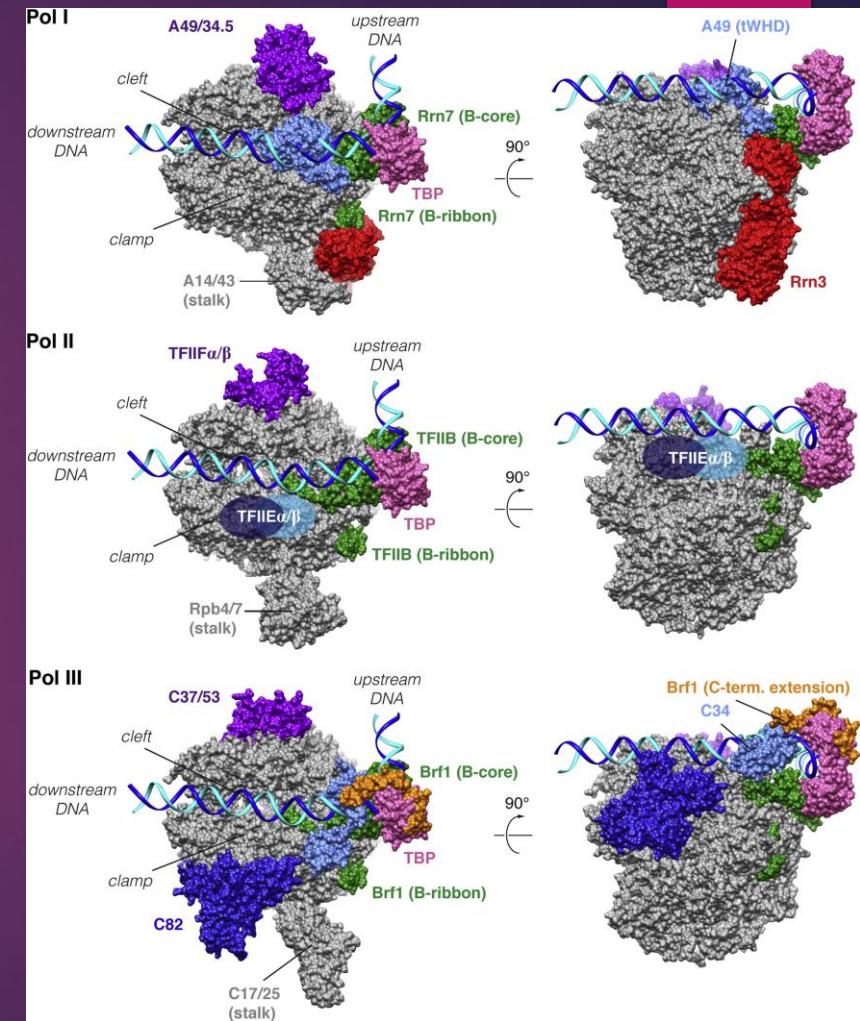
- Gene expression is triggered by a signal from the cell environment.
- Process:
 1. An Extracellular Signal (ligand) binds to a cell-surface or intracellular Receptor.
 2. This activates a downstream cascade, often involving sequential activation of Protein Kinases.
 3. The final step modifies a nuclear protein: a Transcription Factor (TF) or a Chromatin Modifying Enzyme.
- This input orchestrates gene expression changes in the nucleus.

Gene expression is triggered by a signal from the cell environment



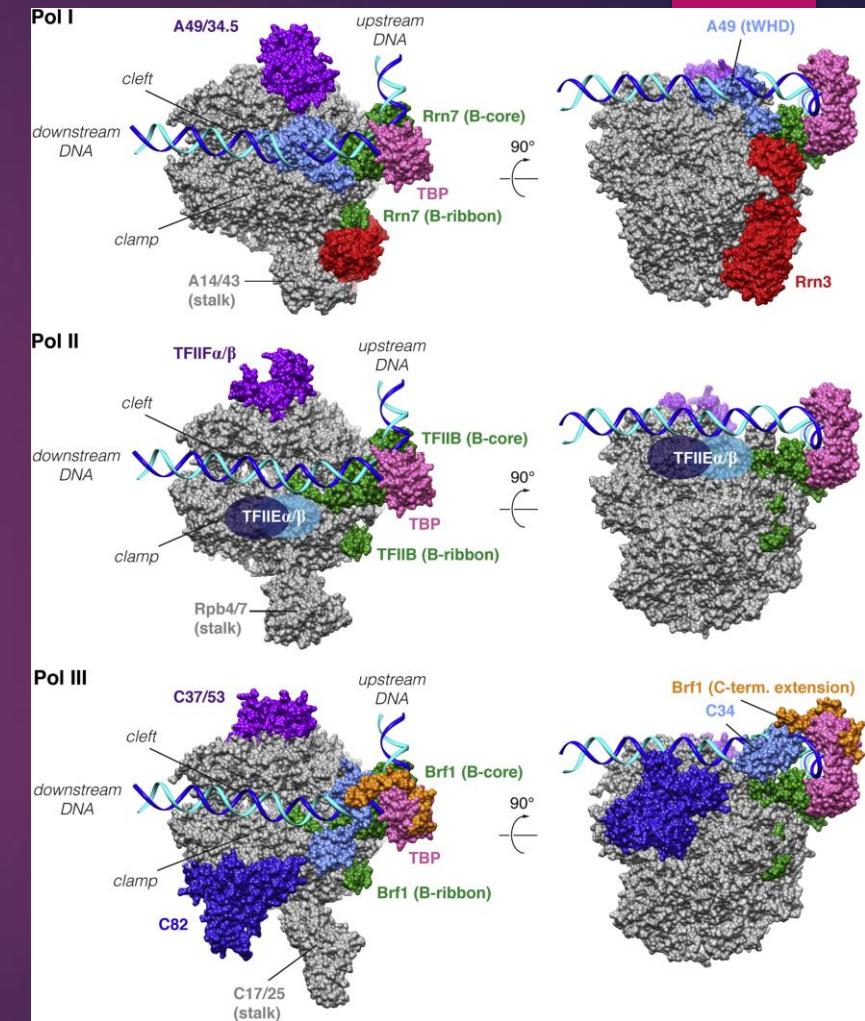
RNA Polymerase I and III (The Housekeepers)

- The three RNA polymerases distribute the workload for RNA synthesis.
- RNA Pol I:
 - Transcribes the large ribosomal RNA genes (rRNAs: 5.8S, 18S, 28S).
 - Constitutes over 80% of total cellular RNA.
 - Essential for ribosome structure and basic protein synthesis (Ubiquitously expressed).
- RNA Pol III:
 - Transcribes small RNAs (tRNAs, 5S rRNA, small nuclear RNAs like U6).
 - Required for basic functions like translation (tRNAs) and splicing (snRNAs) (Ubiquitously expressed).



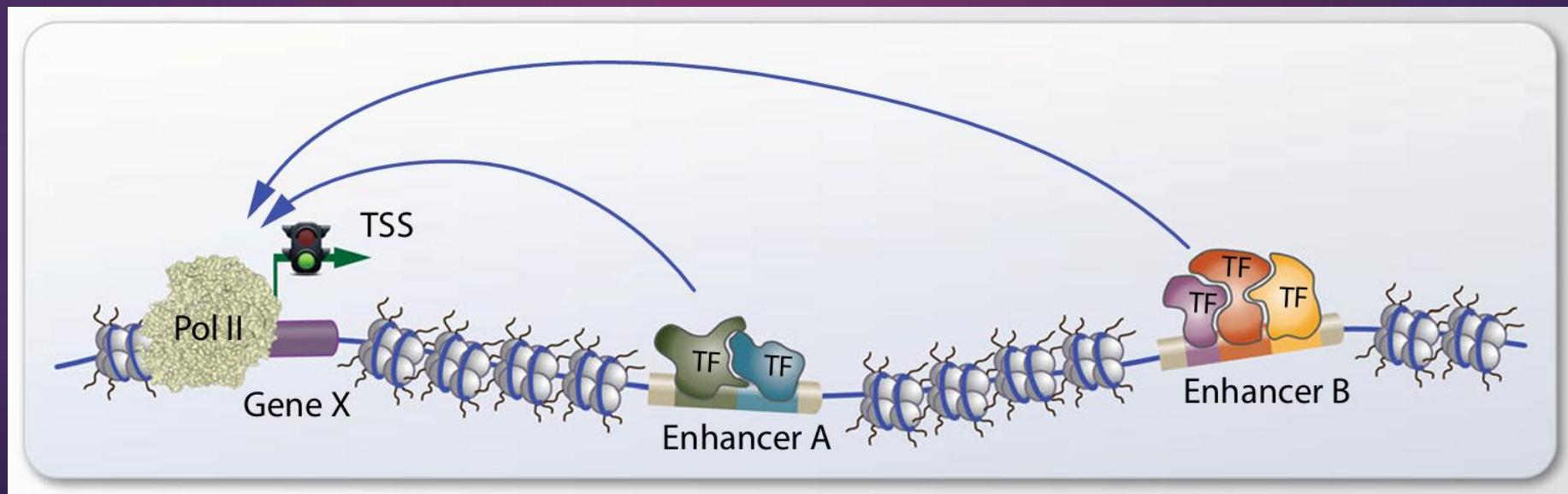
RNA Polymerase II (The Regulator)

- RNA Pol II:
 - Transcribes all protein-coding genes (mRNA).
 - Transcribes the majority of non-coding RNAs (miRNAs, lncRNAs).
- Function: It is responsible for the synthesis of transcripts that define the cell's specialized function and its response to external stimuli.
- Regulation: Its activity is the most highly regulated and cell-specific of the three polymerases.



The Promoter and Gene Structure

- **Promoter:** The DNA sequence immediately upstream of the gene where RNA Pol II and its general transcription factors bind to initiate transcription.
- **Enhancer:** Distal DNA sequences that significantly boost transcription by recruiting activators and looping to contact the promoter.
- **Chromatin's Role:** All of these regulatory sequences must be accessible (in euchromatin) for Pol II to initiate transcription.



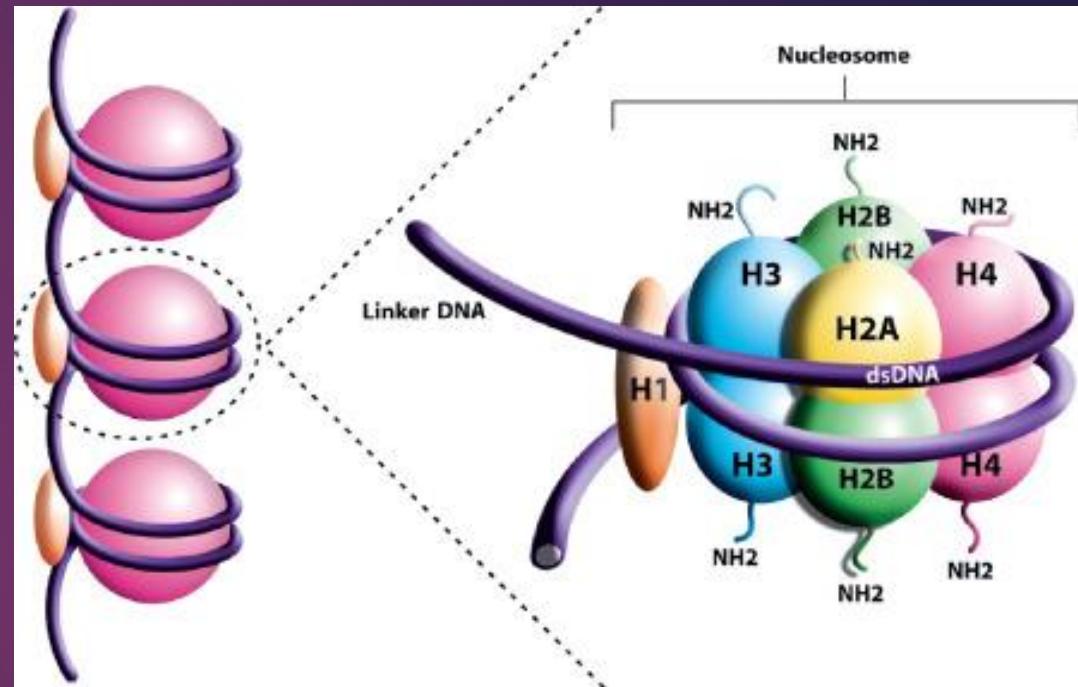
Chromatin Structure: The Nucleosome

The Packaging Problem

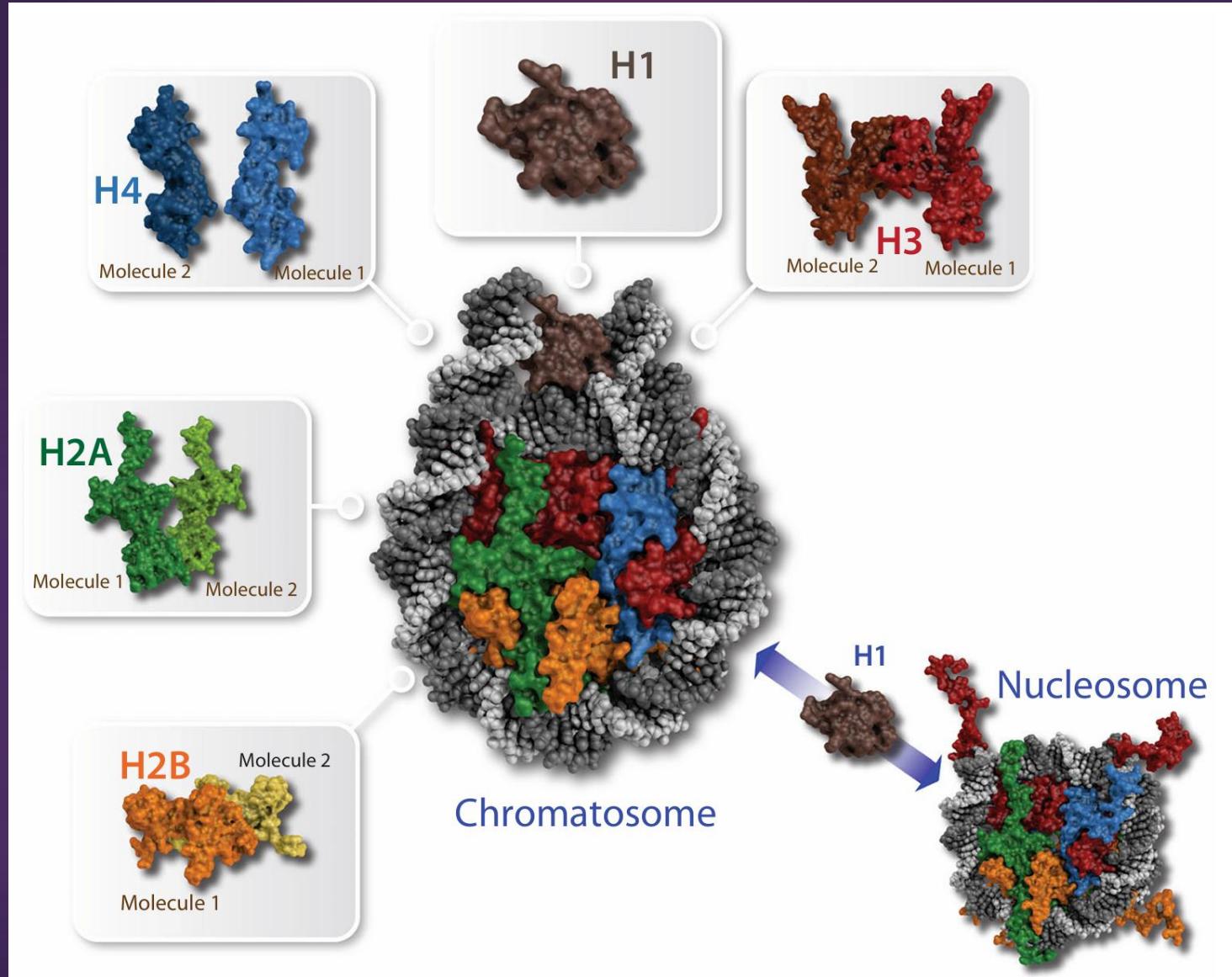
- The genome (2 meters of DNA) must fit inside a 5–10 μm nucleus.
- **Issue:** DNA is a large polymer with a strong negative charge (phosphate backbone).
- **Necessity:** Tight condensation is required, but it must remain reversibly accessible for replication and transcription.

Introducing the Nucleosome

- The Nucleosome is the fundamental, repeating unit of DNA packaging.
- It represents the first level of compaction (up to 7-fold).
- Visually, it creates the "beads-on-a-string" structure (the 11 nm fiber).

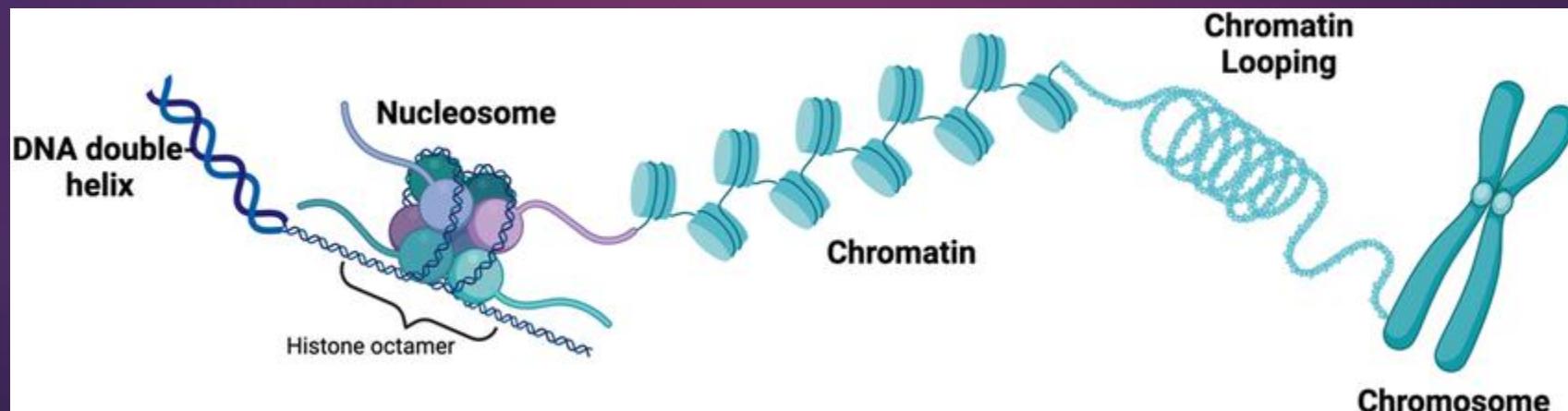


The Histone Octamer Core



The Histone Octamer Core

- The central structure of the nucleosome is the Histone Octamer.
- It is a complex assembled from two copies of each of the four Core Histones:
 - H2A and H2B (form a dimer).
 - H3 and H4 (form a tetramer).
 - The H3-H4 tetramer forms the core scaffold, and two H2A-H2B dimers cap it.



The Histone Octamer Core

Table 2.2 Types and properties of human histones

Histone	Molecular weight [kDa]	Number of amino acids	Content of basic amino acids	
			Lys [%]	Arg [%]
H1	22.1	223	29.5	1.3
H2A	14.0	129	10.9	9.3
H2B	13.8	125	16.0	6.4
H3	15.3	135	9.6	13.3
H4	11.2	102	10.8	13.7

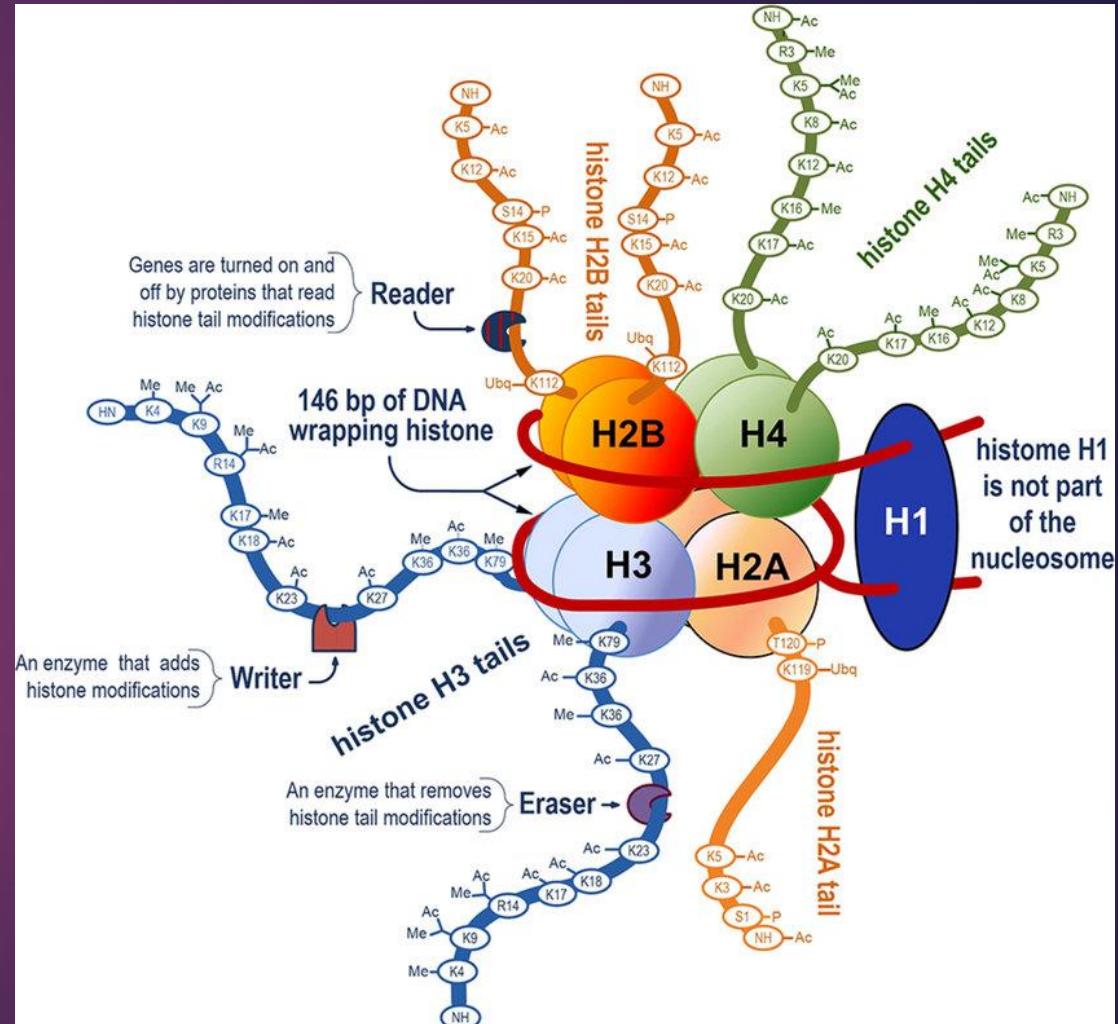
Histone H1 binds to linker DNA, while histones each a pair of H2A, H2B, H3 and H4 form the nucleosome core

Histone Composition and DNA Binding

- Histone Proteins are small, highly conserved proteins.
- They are rich in positively charged amino acids: Lysine (K) and Arginine (R).
- **Mechanism:** The strong electrostatic attraction between the positive histones and the negative DNA backbone drives the DNA to wrap nearly twice (1.67 turns) around the octamer, spanning 147 base pairs.

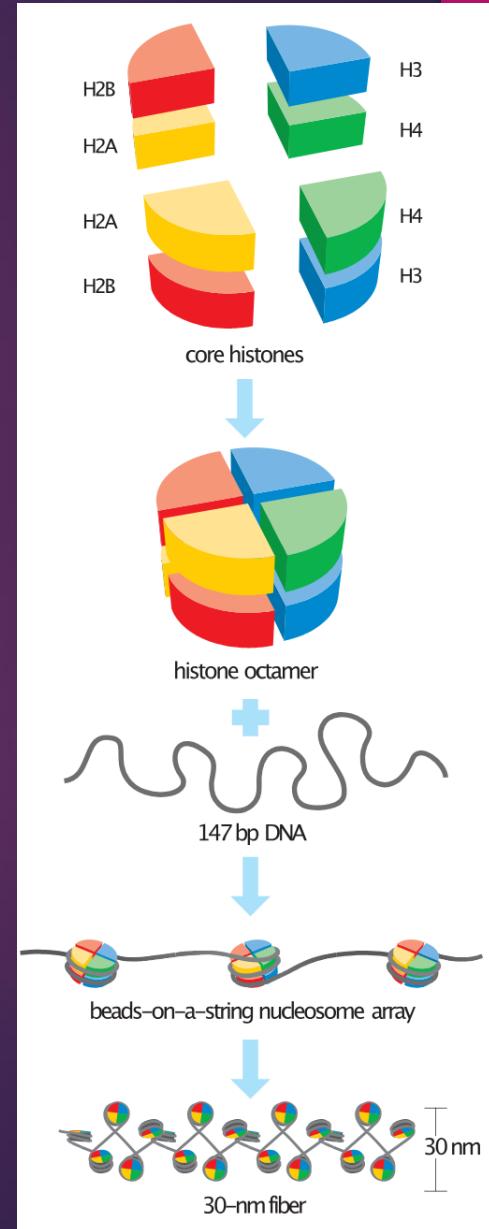
The Histone Tails: PTM Substrates

- All four core histones have flexible Amino-Termini (tails) that protrude from the nucleosome core.
- These tails are the primary docking sites for enzymes that catalyze Post-Translational Modifications (PTMs).
- PTM Examples: Acetylation, Methylation, Phosphorylation.
- Function: The PTMs recruit specific "reader" proteins, essentially forming the basis of the Histone Code.



Linker Histone H1 and the Chromatosome

- Linker DNA: The stretch of DNA connecting adjacent nucleosomes (10-80 bp).
- Histone H1: The Linker Histone that binds to the linker DNA region, stabilizing the structure where DNA enters and exits the core.
- Chromatosome: The term for a nucleosome plus the associated H1.
- H1 Function: H1 is essential for promoting the formation of the higher-order 30 nm chromatin fiber by pulling adjacent nucleosomes together.



Histone Variants: Altering Function

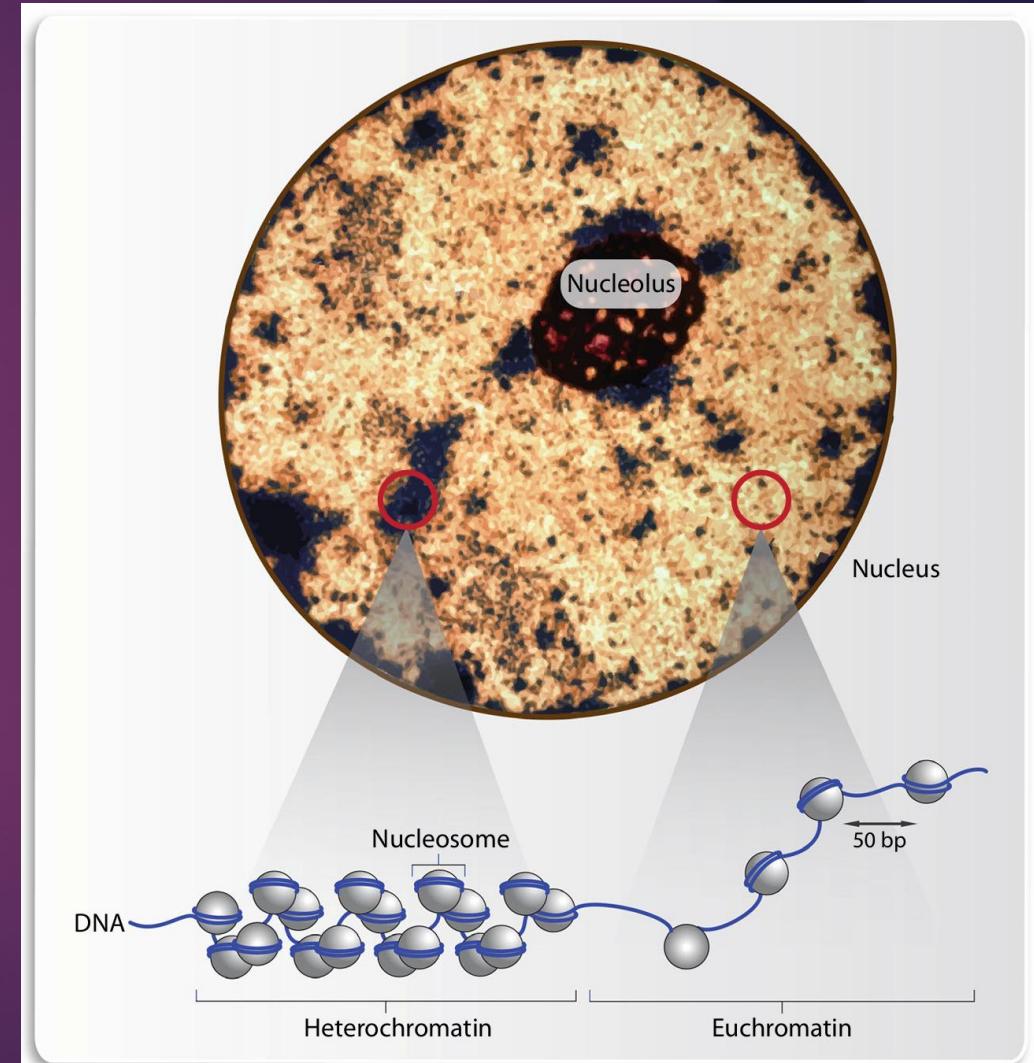
- Histone Variants are specialized forms of core histones that replace the standard histones at specific genomic loci.
- Incorporation: These variants are typically incorporated into chromatin independent of DNA synthesis.
- Example 1 (CENP-A): Replaces H3 at centromeres, establishing the kinetochore attachment site.
- Example 2 (H2A.Z): Often deposited at the edges of active promoters to make them more pliable and accessible.
- Variants enable the creation of specialized, functional chromatin domains.

Chromatin States and Epigenetics



Chromatin in Interphase

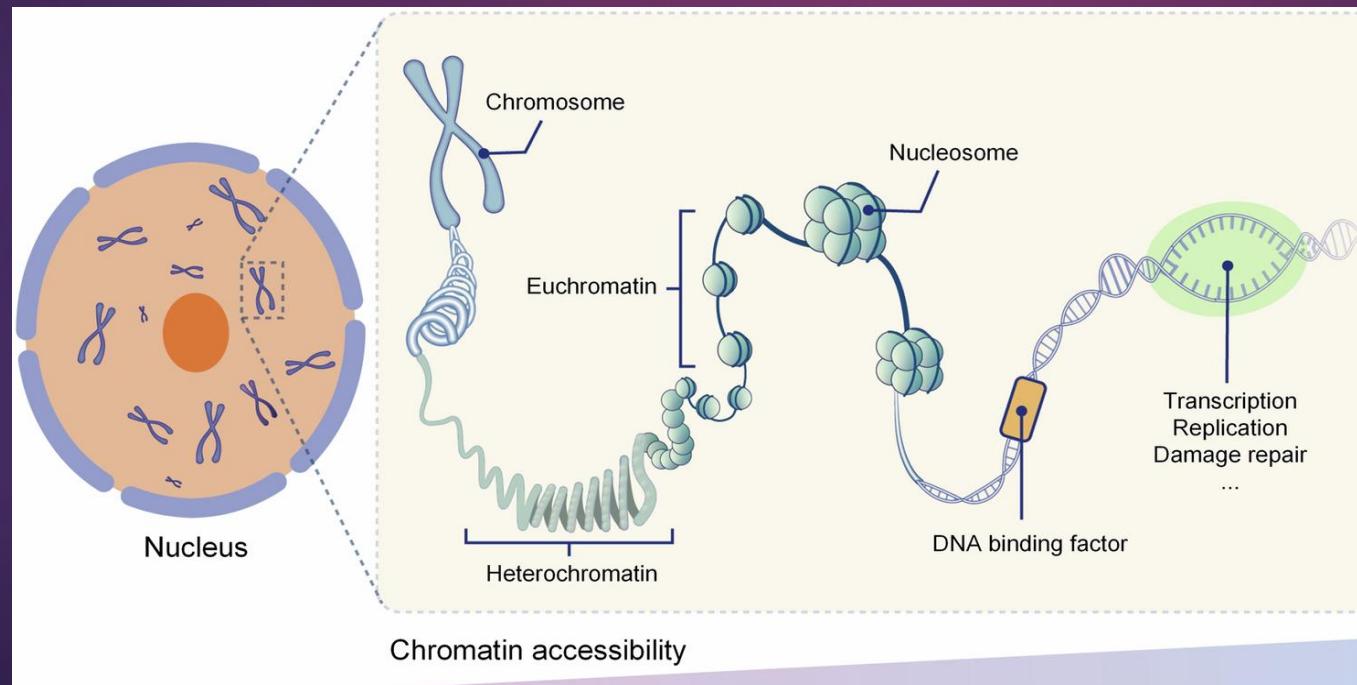
- During the non-dividing state (Interphase), the nucleus exhibits two main chromatin states based on condensation:
 - Euchromatin (loosely packed, active)
 - Heterochromatin (tightly packed, silenced)
- These states are visible via microscopy and are the physical manifestation of epigenetic status.



Eu-and heterochromatin

Euchromatin: Structure and Function

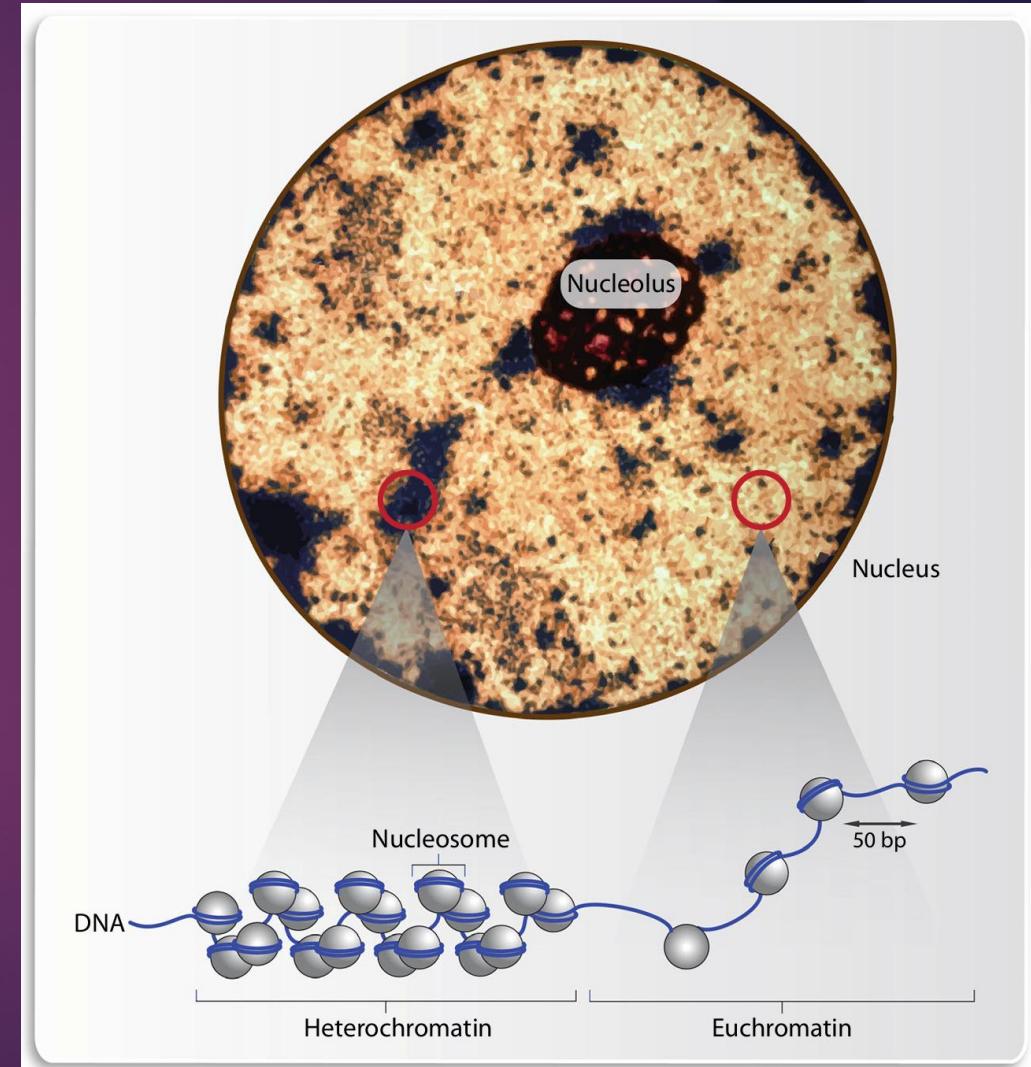
- Structure: Decondensed, often seen as the 11 nm "beads-on-a-string" fiber.
- Accessibility: High.
- The DNA is readily accessible to transcriptional machinery.
- Gene Activity: Home to the vast majority of actively transcribed genes.
- Regulatory Status: Marked by activating histone modifications (e.g., acetylation).



<https://www.nature.com/articles/s41392-024-02030-9>

Heterochromatin: Structure and Function

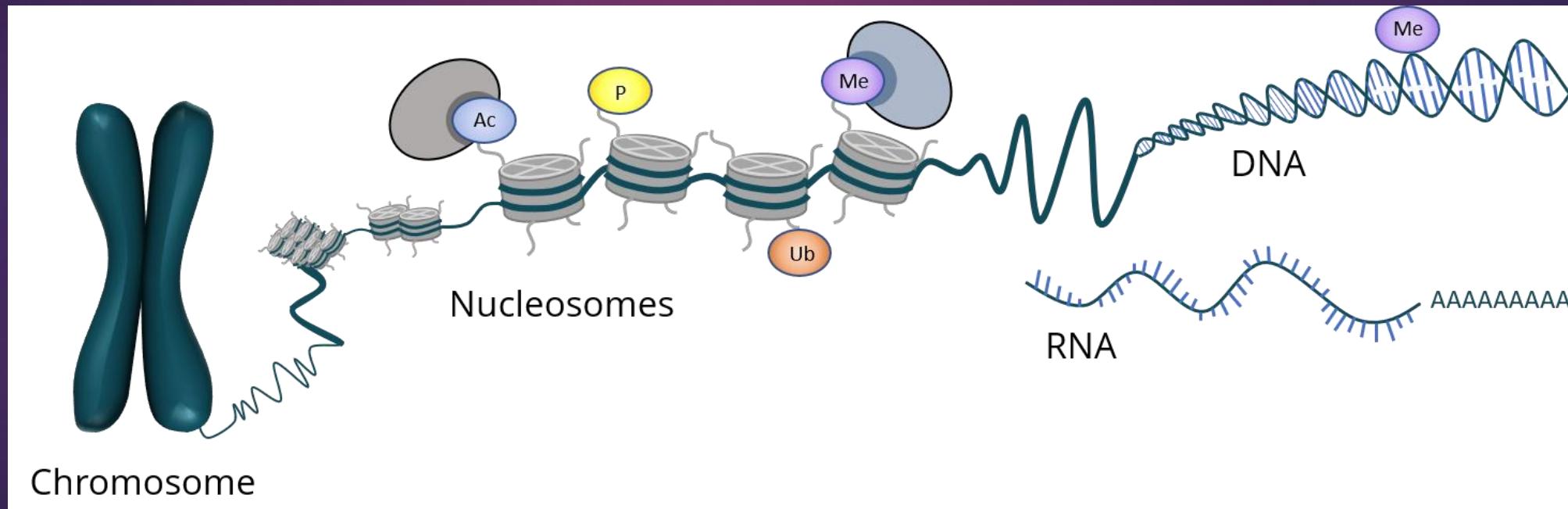
- Structure: Highly condensed, organized into the 30 nm fiber and higher-order structures.
- Accessibility: Low/None. The DNA is inaccessible and hidden from factors.
- Gene Activity: Home to silenced genes and large tracts of repetitive/junk DNA.
 - Types: Constitutive: Always heterochromatin (e.g., centromeres).
 - Facultative: Can be reversibly converted to euchromatin (e.g., developmental genes).



Eu-and heterochromatin

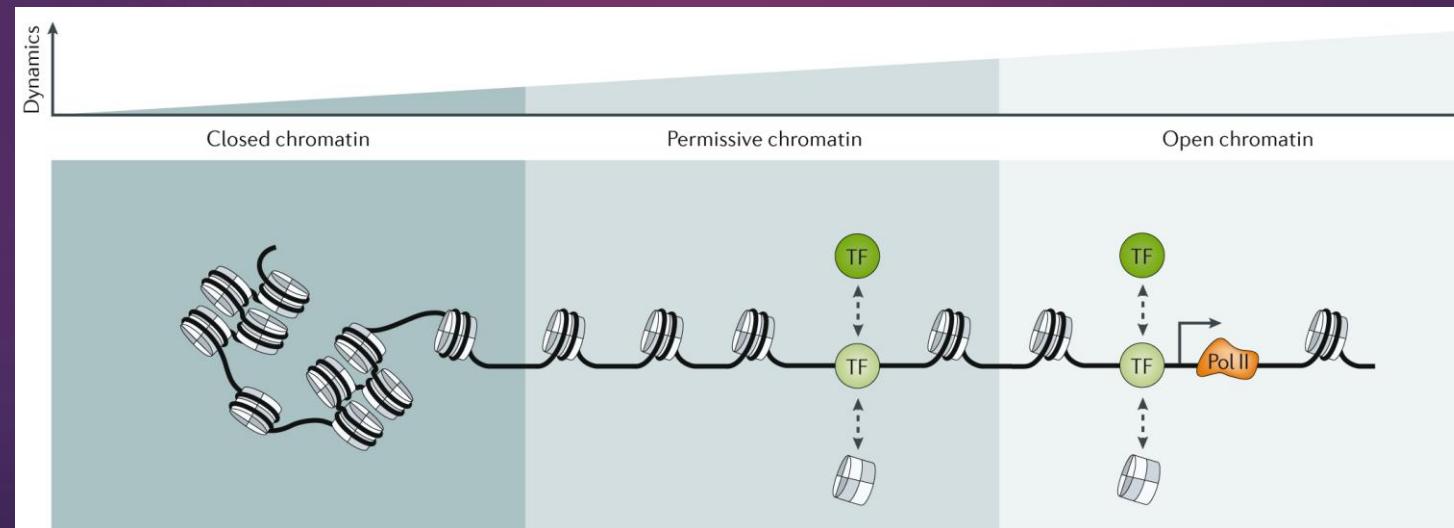
Defining Epigenetics Precisely

- Etymology: Epi- (Greek: "above," "on top of") + Genetics.
- Definition: "Changes in gene function that are heritable (mitotic and/or meiotic) and do not involve a change in the DNA sequence."
- Molecular Components: DNA methylation, histone modifications, and non-coding RNAs



Chromatin as the Epigenetic Gatekeeper

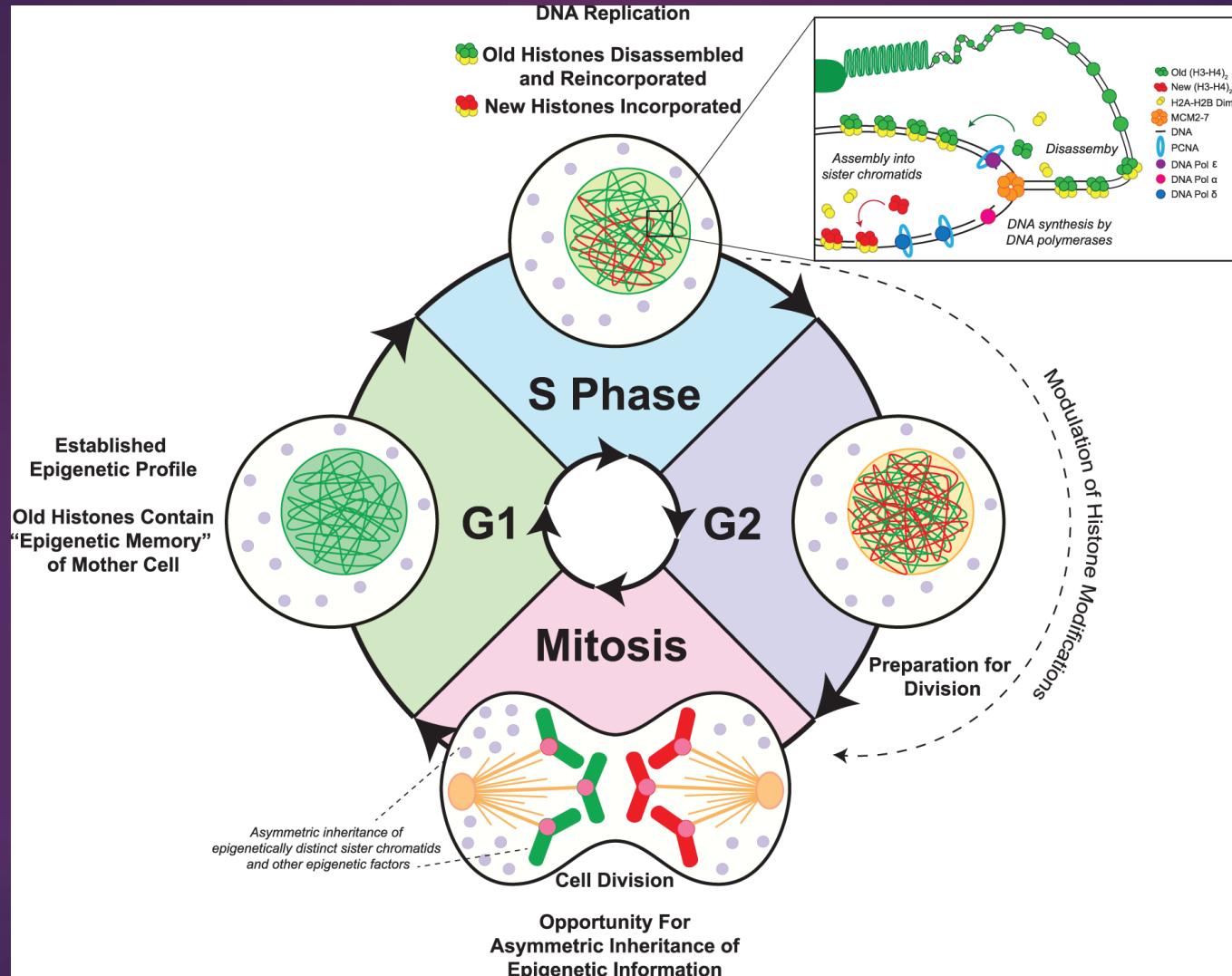
- Chromatin serves as the epigenetic gatekeeper for gene expression.
- **The Problem:** The cell must selectively express 10% of its genes while silencing the other 90%.
- **The Solution:** Chromatin structure physically hides the 90% in inaccessible heterochromatin, preventing accidental activation.
- **Action:** Only when a gene's regulatory region is moved into euchromatin (the gate is opened) can transcription factors bind and initiate transcription.



Epigenetic Memory and Inheritance

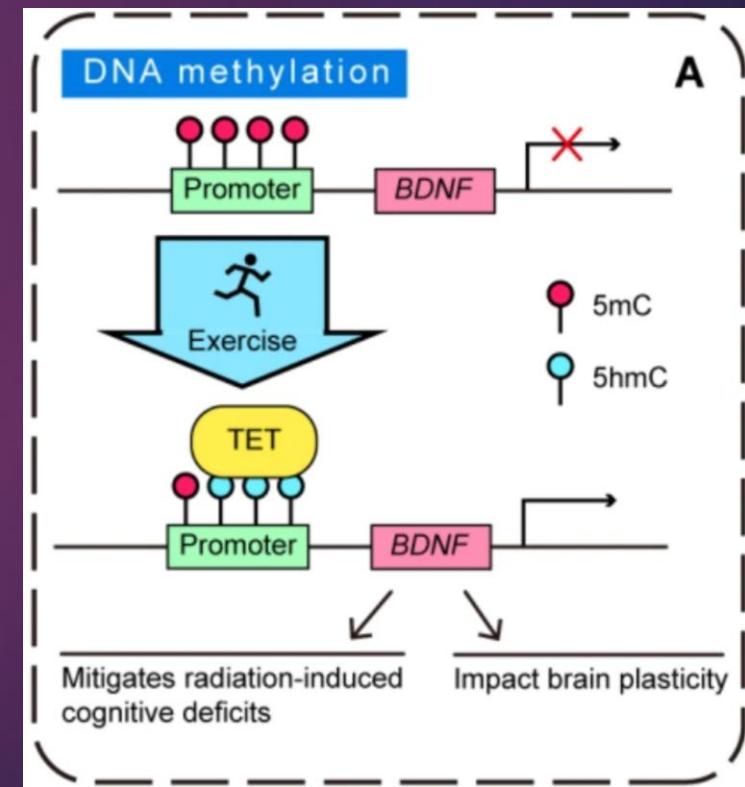
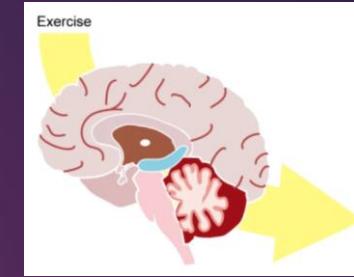
- **Cellular Identity:** Epigenetic marks establish and maintain the identity of a specialized cell (e.g., once a liver cell, always a liver cell). This is **Mitotic Inheritance**.
- **Meiotic Inheritance:** Some epigenetic marks (though less common and highly dynamic) can potentially be passed down through the gametes to the next generation.
- **Implications:** Epigenetics links cell fate and environmental exposure to gene function.

Epigenetic inheritance in the context of the cell cycle



Epigenetic Dynamics and Reversibility

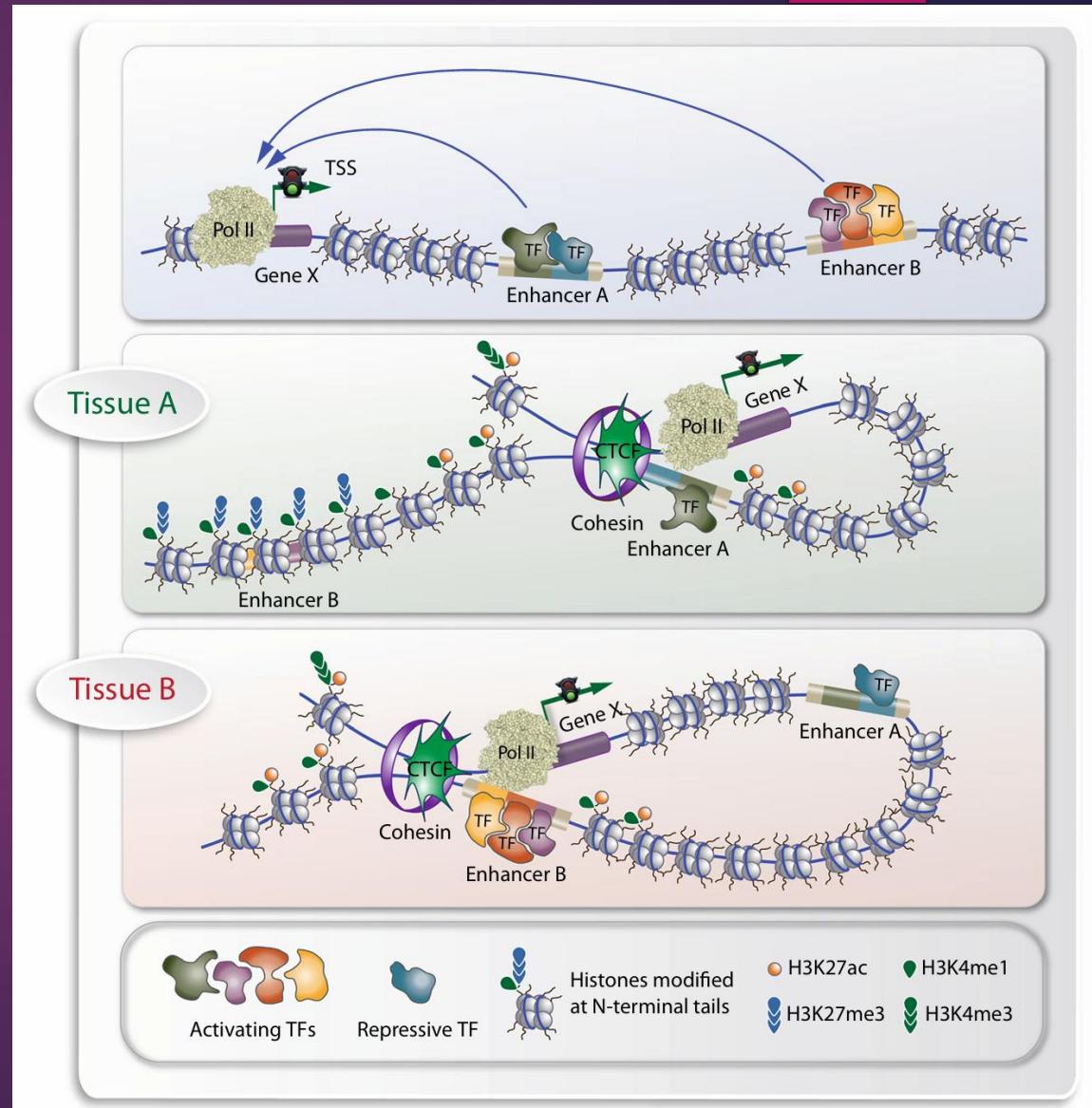
- Epigenetic states are not permanent. They are **Dynamic**.
- **Reversibility:** Most marks are reversible. The addition of a mark (e.g., Acetylation) is balanced by its removal (Deacetylation).
- **Example:** Lifestyle changes (diet, exercise) can rapidly reprogram the epigenetic landscape in metabolic tissues, potentially reversing conditions like insulin resistance.



3D Nuclear Architecture

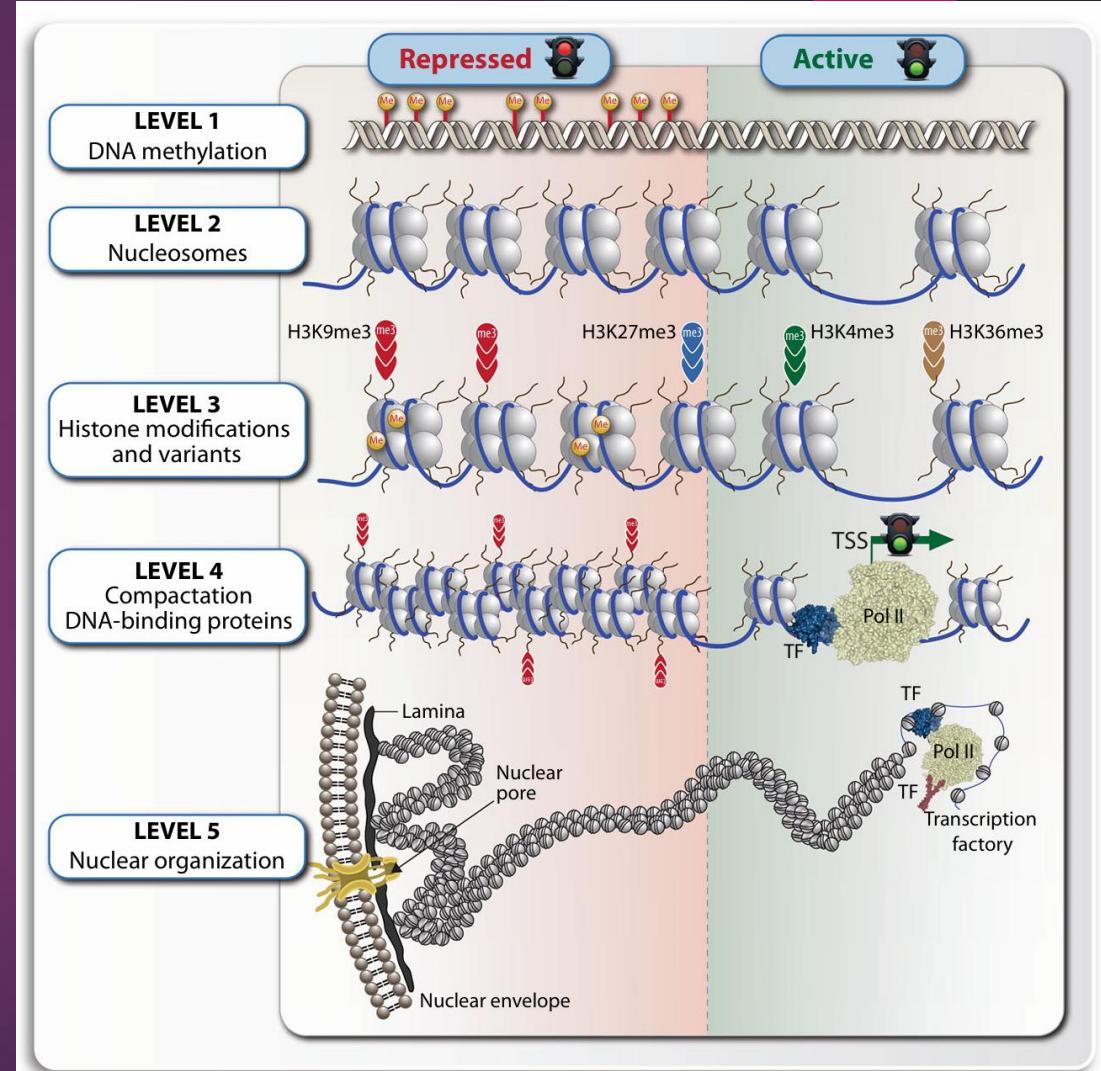
Chromatin Organization in 3D Space

- The nucleus is not a simple bag; it has a **sophisticated** 3D architecture that aids regulation.
- The **spatial organization** is dictated by various structural proteins (e.g., Lamins, CTCF, Cohesin).
- **Purpose:** To bring distant regulatory elements together and segregate active regions from silent ones.



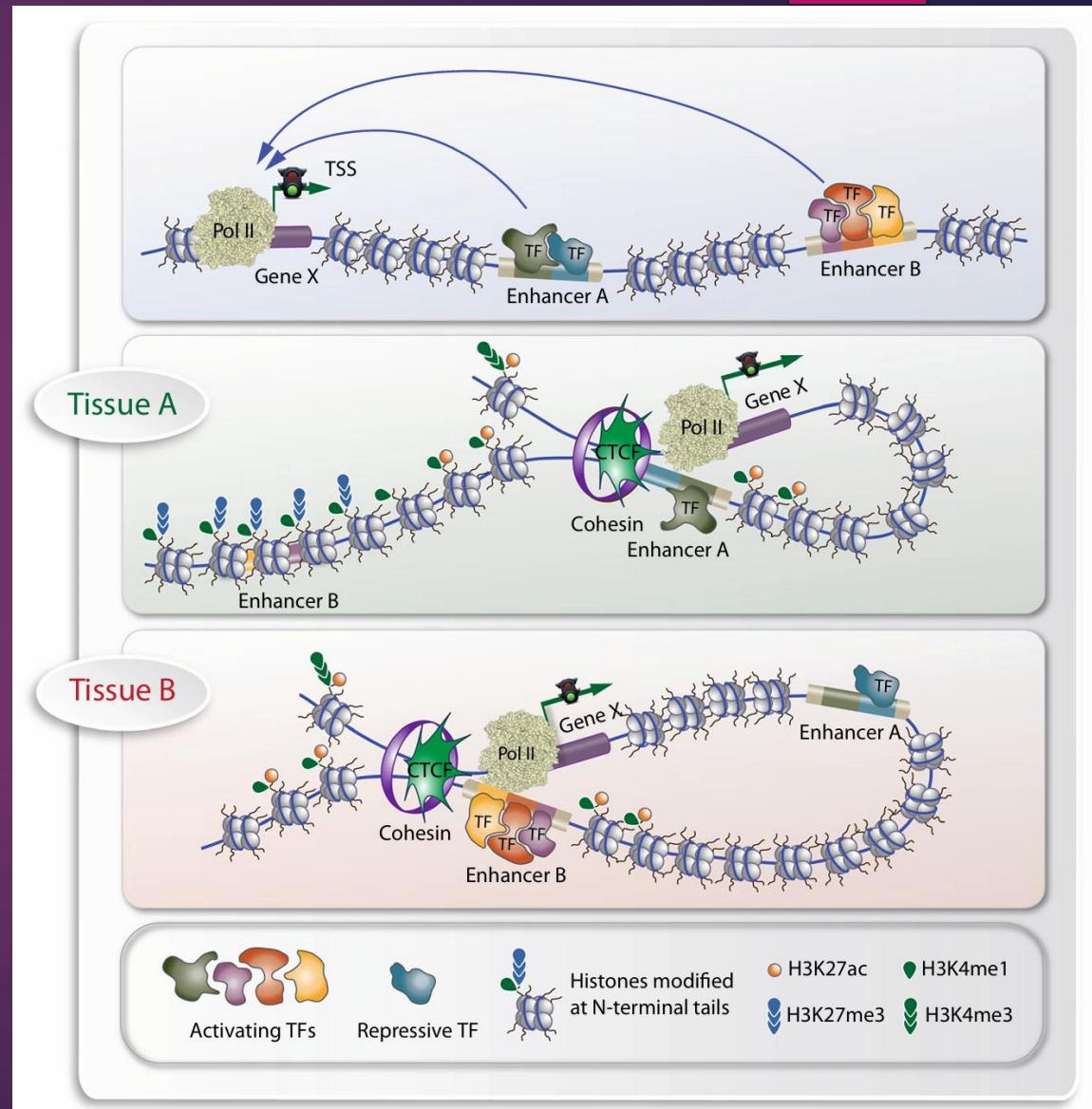
Nuclear Periphery and Silencing

- **Nuclear Lamina:** A protein meshwork beneath the inner nuclear membrane.
- **Lamin-Associated Domains (LADs):** Large regions of heterochromatin that are physically anchored to the Nuclear Lamina.
- **Consequence:** Tethering to the periphery is strongly associated with stable, long-term gene silencing.



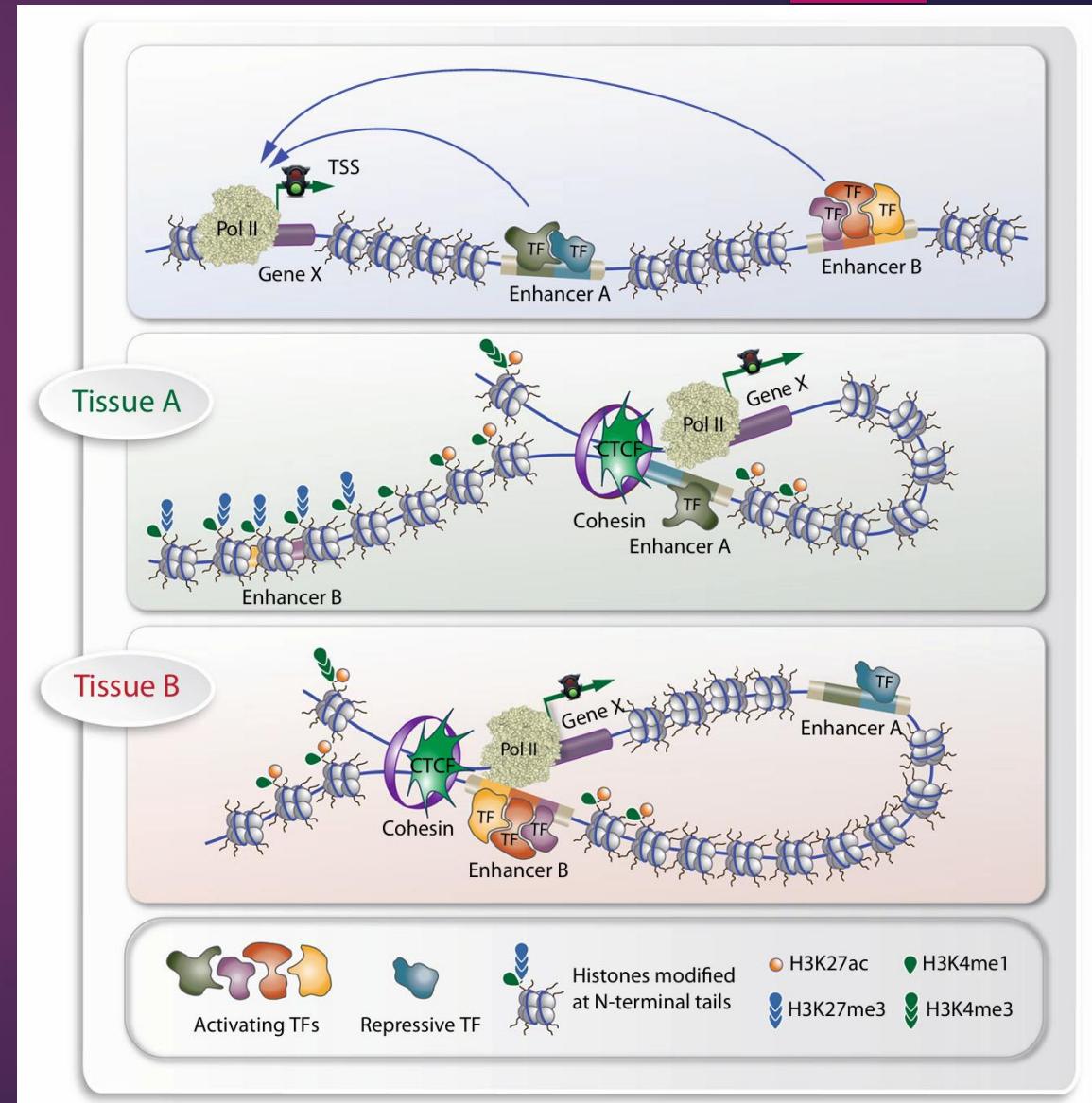
The Importance of DNA Looping

- **Mechanism:** Regulatory elements (like Enhancers) can be tens or even hundreds of kilobases away from their target Promoters.
- **Solution:** DNA Looping brings these elements into physical contact in 3D space.
- **Function:** This contact is necessary for the Enhancer-bound activator proteins to efficiently communicate with the basal transcriptional machinery at the Promoter.



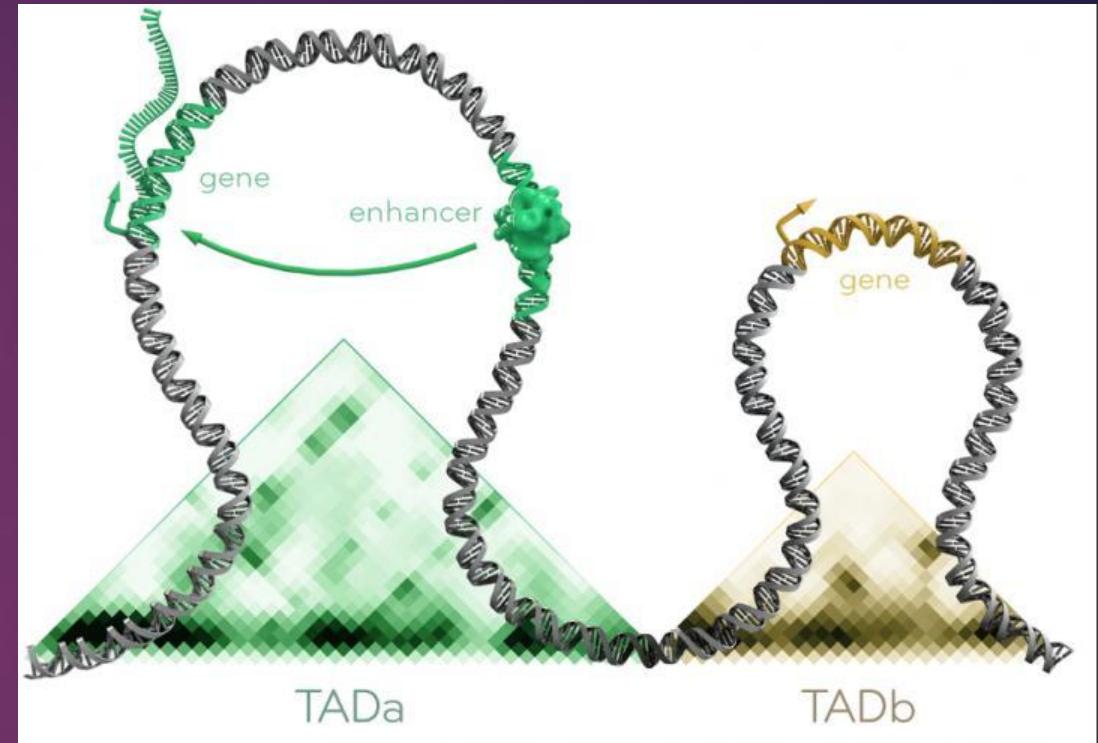
CTCF and Cohesin: The Loop Architects

- The formation of these crucial regulatory loops is dependent on two key protein components: CTCF (CCCTC-binding factor): A sequence-specific DNA-binding protein.
- Cohesin Complex: A ring-shaped protein complex that holds two segments of DNA together.
- Role: CTCF often defines the anchors (boundaries) of the loops, while Cohesin mediates the physical connection.

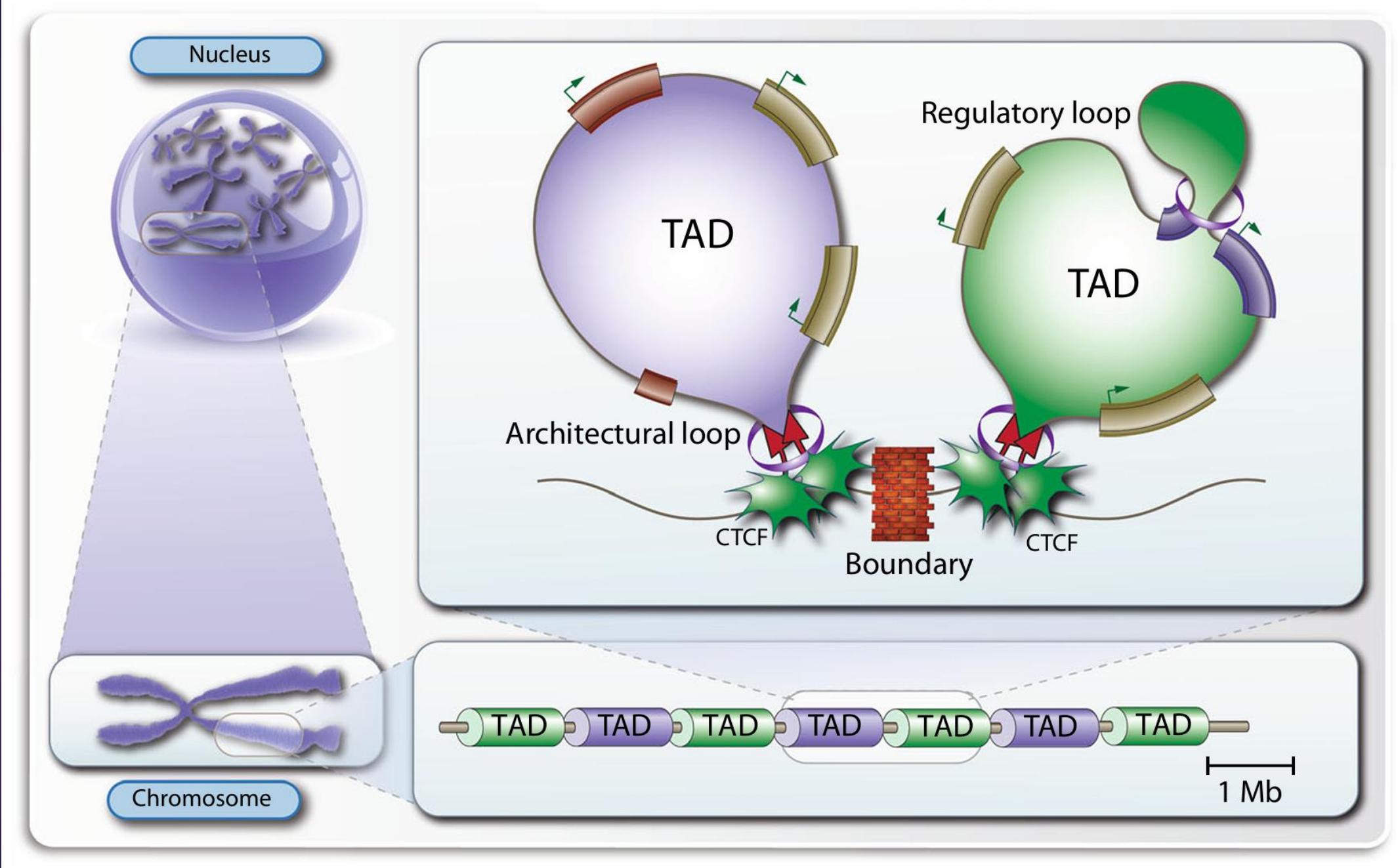


Topologically Associated Domains (TADs)

- TADs are large, stable genomic regions (hundreds of kb to megabases) that partition the genome.
- Definition: DNA sequences within a TAD interact frequently with each other, but interactions across the TAD boundary are rare.
- Function: They act as regulatory neighborhoods, ensuring that an Enhancer only activates a gene within its own TAD, preventing regulatory crosstalk.

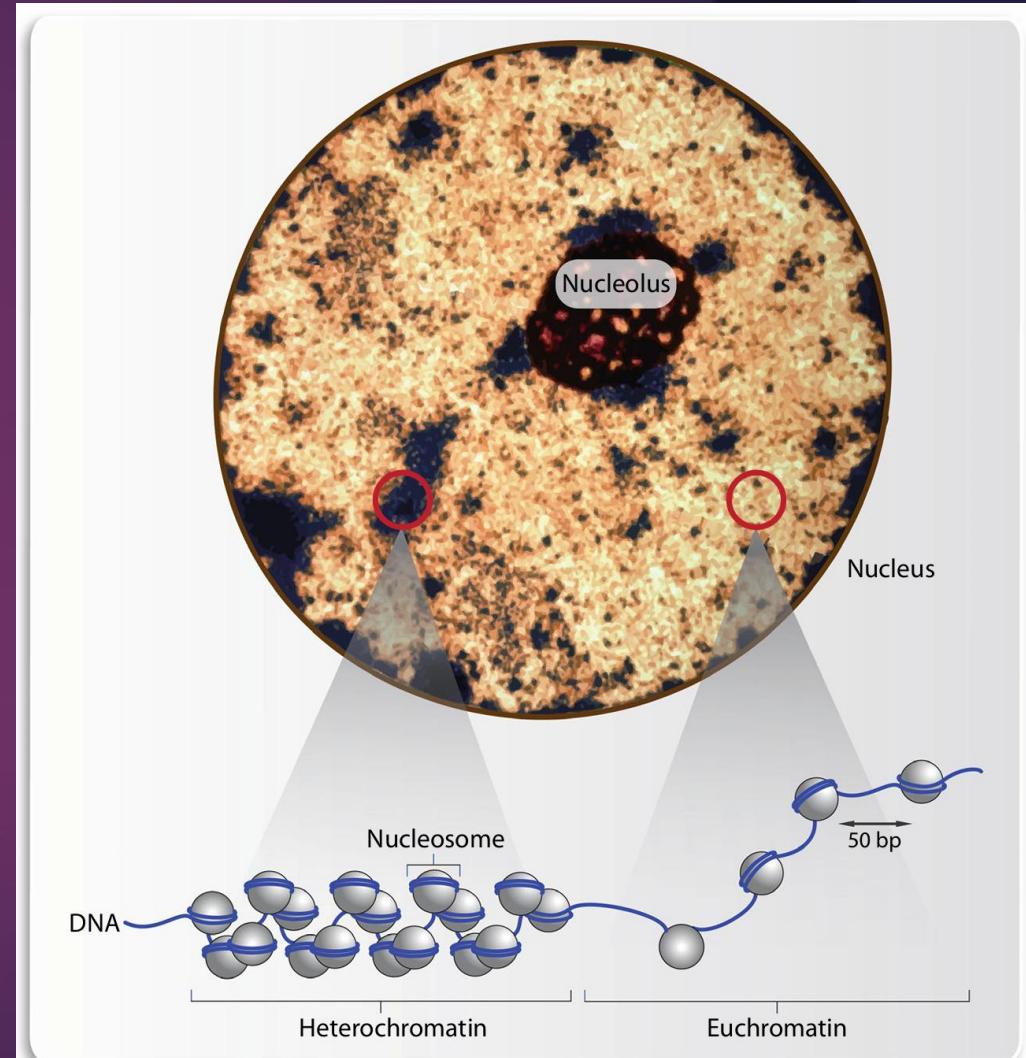


<https://www.eurekalert.org/multimedia/588589>



Transcriptional Hubs (Factories)

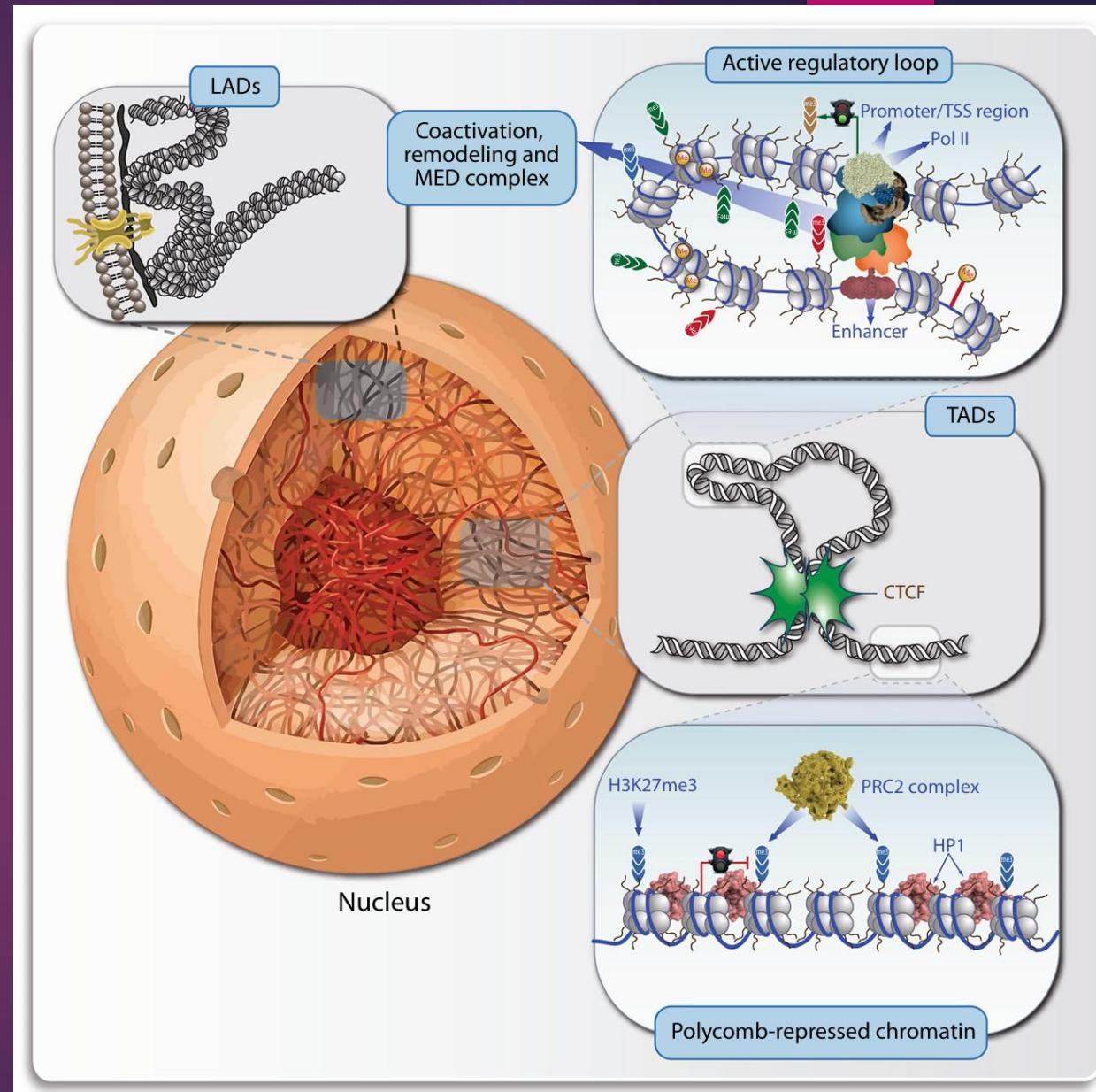
- Transcription is **not random** in the nucleus.
- **Transcriptional Hubs:** Active regions tend to cluster together in the nuclear interior, forming factories where high concentrations of RNA Pol II, TFs, chromatin remodelers, and the Mediator (MED) complex are concentrated.
- **Efficiency:** The 3D architecture enhances transcription efficiency by spatially concentrating all necessary components.



Eu-and heterochromatin

Summary

- Chromatin structure is the primary layer of gene regulation.
- The Nucleosome is the histone-DNA complex that determines accessibility.
- Euchromatin is accessible and active.
- Heterochromatin is inaccessible and silent.
- 3D Architecture (TADs, loops) organizes the genome for precise control.



Chromatin architecture



Xin chân thành cảm ơn!

LUU PHUC LOI, PHD

ZALO: 0901802182

LUU.P.LOI@GOOGLEMAIL.COM