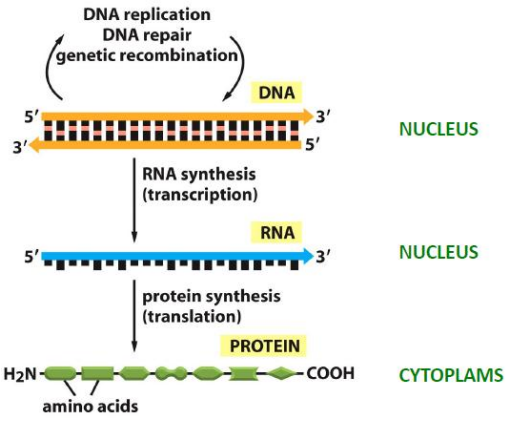
RNA and Transcription

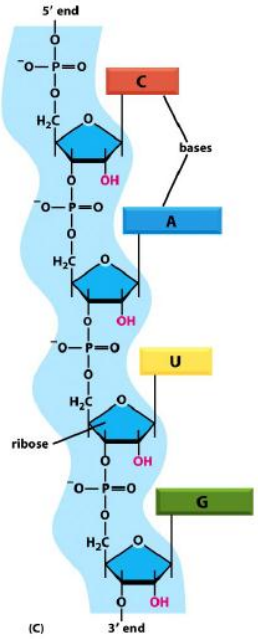


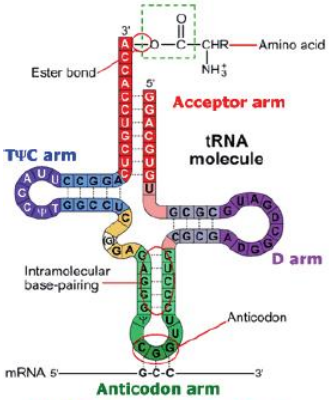
Central dogma of molecular biology: All living cells are built up from genetic information stored in the DNA. **DNA** is **transcribed** intro **RNA** molecules, which in turn are **translated** into **proteins**.

DNA replication is a 1:1 process, one template of DNA gives rise to one newly synthetized DNA molecule. Transcription and translation can occur with different efficiencies (one gene might be transcribed many times whereas another one is only transcribed a few times, different number of copies).

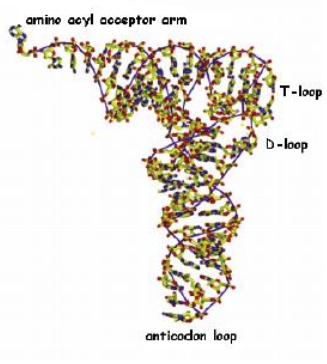
RNA: Similar to DNA, but:

- RNA is composed of **ribonucleotides** (i.e. nucleotides containing ribose, and not deoxyribose sugar)
- RNA is made up of four bases, adenine, guanine, cytosine and uracil (not thymine); uracil is able to base pair with adenine
- RNA is single stranded and generally not assembled in double stranded molecules in living cells
- RNA can fold into **secondary structures** through conventional (e.g. A/U) and non-conventional (e.g. G/U) interactions occurring at bases within the same RNA molecule
- tRNAs are 73 to 93 nucleotide long RNA molecules assembling in a cloverleaf secondary structure containing four base-paired stems and three non-base-paired loops; tRNAs are then stabilized into a compact "L" shaped tertiary structure by base pairing and base stacking





tRNA secondary structure

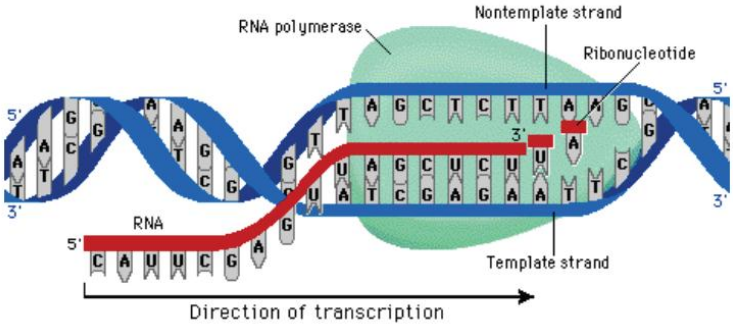


tRNA tertiary structure

Classes of RNA molecules:

- mRNAs: messenger RNAs, code for proteins
- rRNAs: ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
- tRNAs: transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
- snRNAs: small nuclear RNAs, functions in a variety of nuclear processes, including the splicing of pre-mRNA
- snoRNAs: small nucleolar RNAs, used to process and chemically modify rRNAs
- scaRNAs: small cajal RNAs, used to modify snoRNAs and snRNAs
- miRNAs: micro RNAs, regulate gene expression typically by blocking translation of selective mRNAs
- siRNAs: small interfering RNAs, turn off gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures
- other noncoding RNA: function in diverse cellular processes, including telomere synthesis, X.chromosome inactivation, and the transport of proteins into the ER

Transcription: Biosynthesis of RNA is an enzymatic process similar to DNA replication, where a ssDNA strand acts as a template for the synthesis of an RNA molecule. Transcription requires unwinding of the DNA double helix and RNA polymerase activities. Newly synthesized RNA does not remain associated to DNA but after the synthesis is complete, the RNA separates from the DNA and the DNA recoils into its helix. The direction of RNA polymerase is 5' to 3'.



Prokaryotic vs eukaryotic transcription:

Prokaryotes:

- Transcription occurs in the cytoplasm alongside translation
- Transcription does not need chromatin remodeling factors, the DNA is directly accessible to transcription machineries
- Transcripts are not modified post-transcriptionally

Eukaryotes:

- Transcription occurs mainly in the nucleus; transcripts (mRNAs) are then exported to the cytoplasm to be translated
- Transcription requires a number of factors able to remodel chromatin in order to expose template DNA to be transcribed
- mRNAs are modified once synthesized (RNA splicing, 5' methyl cap addition, 3' polyadenlyation)

What do we need for transcription to occur?

1. A DNA template to be transcribed into RNA
2. An enzymatic machinery able to catalyze the transcription reaction

1. DNA Template

The DNA template to be transcribed into RNA is placed in the genome and bears specific features to make it recognizable. Transcription does not initiate randomly in the cell but at specific places on the DNA template called **promoters**. Promoters contain specific DNA sequences, which provide a binding site for RNA polymerase and/or for proteins that will recruit RNA polymerase. Core promoters are placed immediately upstream of the region of DNA that will be transcribed, typically few tens of nucleotides far away from the transcription start site.

Prokaryotic promoters:

- Relatively easy to isolate due to sequence conservation
- Consist of two short sequences:
    - A sequence at -10 that is called the **Pribnow box (PB)** or **-10 box** and usually consists of the six nucleotides TATAAT; the Pribnow box is absolutely essential to start transctiption
    - A sequence at -35 that is called -35 box and usually consists of the six nucleotides TTGACA; its presence allows a very high transcription rate

Probability of occurrence of each nucleotide

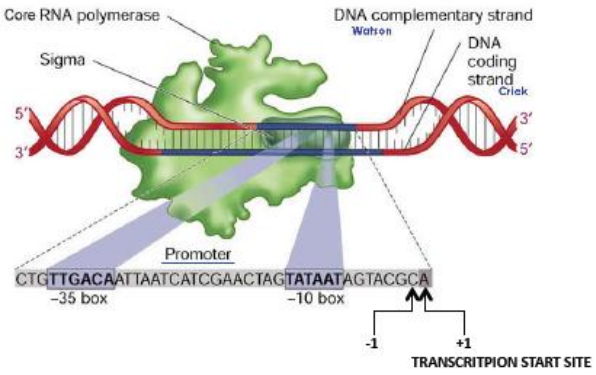
for -10 sequence

T	A	T	A	A	T
77%	76%	60%	61%	56%	82%

for -35 sequence

T	T	G	A	C	A
69%	79%	61%	56%	54%	54%

- Both sequences are not found intact in most promoters but are consensus sequences; on average only 3 of the 6 base pairs in each consensus sequence is found in any given promoter

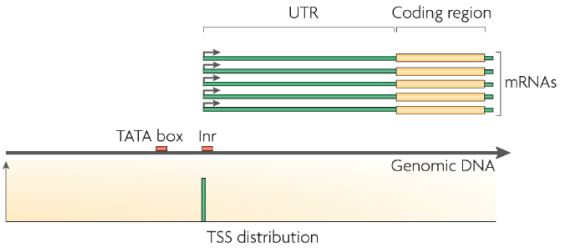


Eukaryotic promoters:

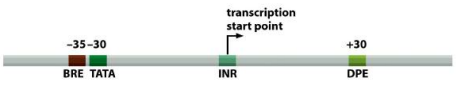
- Extremely diverse and difficult to characterize
- Typically upstream of a gene but can have regulatory elements several kb away from the transcriptional start site (**enhancers**)

Different promoter types:

- **Sharp type promoters** (ex.: human or mouse Syn1 promoter): **TATA box** promoters make up 10 to 20% of human genes



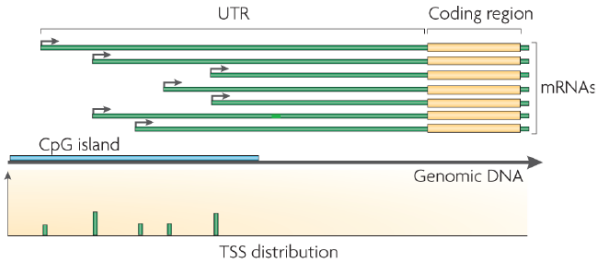
  - AT rich sequences (TATA box, consensus 5' **TATAAA** 3') approximately 28-32 bp upstream (-28/-32) of an **initiator** (Inr) sequence
  - The promoter uses only one or few consecutive nucleotides as TSS (transcription start site)
  - Transcripts are homogenous in length due to the conservation of their 5' ends
  - Generally associated with tissue-specific genes
  - Commonly identified DNA elements (TATA box promoters)
    - **TATA box** at position -30, enforces the ability of the transcription machinery to select a specific TSS in a limited genomic region
    - **TFIIB recognition element (BRE)** at position -35, can either increase or decrease transcription rates
    - **Initiator element (Inr)** surrounding position 1, contains the TSS



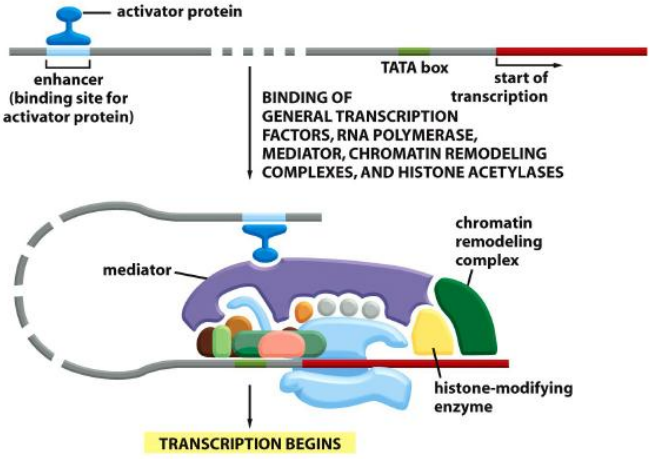
element	consensus sequence	general transcription factor
BRE	G/C G/C G/A C G C C	TFIIB
TATA	T A T A A/T A A/T	TBP
INR	C/T C/T A N T/A C/T C/T	TFIID
DPE	A/G G A/T C G T G	TFIID

- **Downstream promoter element (DPE)** at position +28 to +32, similarly to the TATA box, directs the transcription machinery to a nearby TSS

- **Broad type promoters** (ex.: human/mouse PURA promoter): CpG island promoters make up approximately 50-70% of human genes
  - Promoter sequences lie within long stretches of DNA with an unusually high content in CpG dinucleotides (70-80%)
  - The promoter directs transcription from several TSS within a 100 bp region
  - Transcripts are heterogeneous in length due to differences in their 5' ends

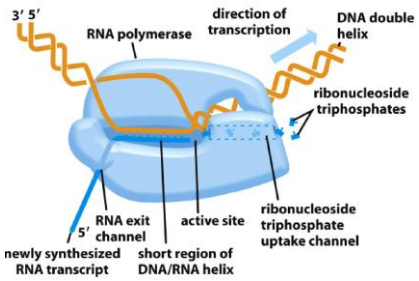


  - Although the RNA 5' end varies, the final protein product is often the same
  - Generally associated with house-keeping or ubiquitously expressed genes



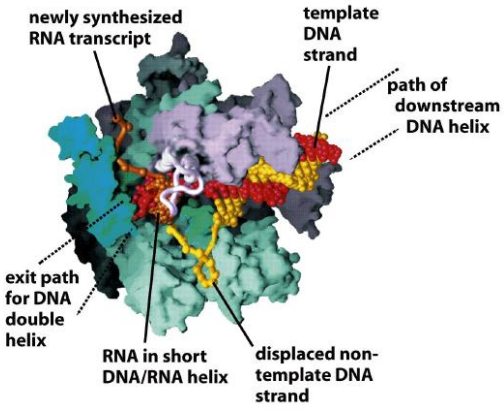
Enhancers: Eukaryotic promoter regulatory elements that can lay several kb away from the transcriptional start site. Transcriptional complex can cause the DNA to bend back on itself, allowing for placement of regulatory sequences far from the actual site of transcription.

2. Enzymatic Machinery



RNA polymerases are multiprotein complexes catalyzing transcription by unwinding DNA ahead of their active site and incorporating ribonucleoside triphosphates into newborn RNA molecules complementary in sequence to the template DNA. RNA polymerases transcribe unidirectionally (5' -3'). The RNA polymerase core enzyme consists of five subunits: 2 α subunits, 1 β subunit, 1 β' subunit and 1 ω subunit.

Multiple copies of RNA molecules can be synthesized simultaneously from one DNA molecule (and pre-ribosomes might pre-assemble at the 5' end of the nascent RNA molecule).



DNA polymerases: Catalyze the formation of phosphodiester bonds between two adjacent deoxyribonucleotides in a **DNA** molecule, need an **RNA primer** to start the reaction and copy DNA with an error rate of about **1 in 10^7** nucleotides.

RNA polymerases: Catalyze the formation of phosphodieser bonds between two adjacent ribonucleotides in **RNA**, do not need an RNA primer to start the reaction and copy with an error rate of about **1 in 10^4** nucleotides.

Bacteria contain only one RNA polymerase while eukaryotic cells have three RNA polymerases sharing a subset of common subunits and transcribing different genes:

TYPE OF POLYMERASE	GENES TRANSCRIBED
RNA polymerase I	5.8S, 18S, and 28S rRNA genes
RNA polymerase II	all protein-coding genes, plus snoRNA genes, miRNA genes, siRNA genes, and most snRNA genes
RNA polymerase III	tRNA genes, 5S rRNA genes, some snRNA genes and genes for other small RNAs

Transcription steps:

- Initiation:

  RNA polymerase recognizes and binds a core promoter sequence in the DNA; in prokaryotes, RNA polymerases bind directly to core promoter DNA with the help of **sigma factor**; in eukaryotes, RNA polymerase binding to promoters is mediated by accessory **transcription factors**; only a basal rate of transcription is driven by this **preinitiation complex**

- Promoter clearance:

  RNA polymerase synthesizes and releases immediately short truncated transcripts both in eukaryotes and prokaryotes (**abortive initiation**); once the transcript reaches approximately 23 nucleotides it no longer slips and elongation can occur

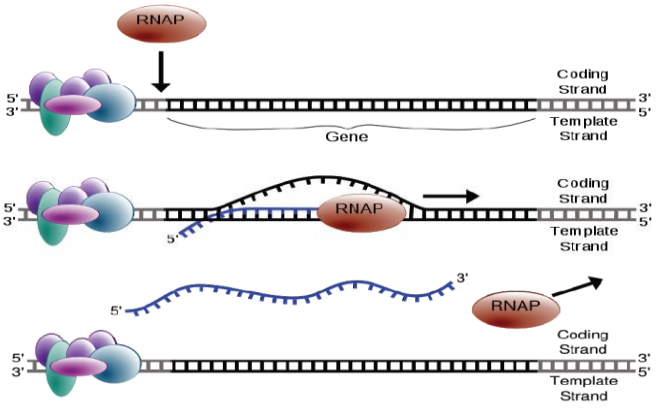
- Elongation:

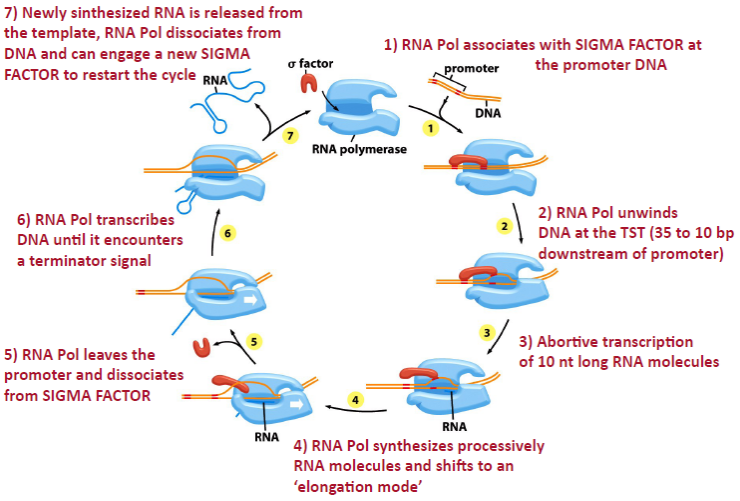
  One strand of DNA, the template strand, is used as a template for RNA synthesis; RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA

template to create a RNA copy; RNA polymerase traverses the template strand in a 3' - 5' direction producing RNA molecules in a 5' - 3' direction; this step also involves proofreading mechanisms that can replace incorrectly incorporated bases

- Termination:

  Polymerase terminates the transcription of the template strand and abandons the DNA; termination results in the release of the newly synthesized mRNA from the elongation complex





In bacteria, transcription termination is either

- Rho-dependent: Protein factor called Rho destabilizes the interaction between the template and the mRNA, thus releasing the newly synthesized mRNA from the elongation complex (50%)
- Rho-independent: The newly synthesized RNA molecule forms a GC rich hairpin loop, followed by a run of U's, which makes it detach from the DNA template (50%)

In eukaryotes, transcription termination is more complex and less understood than in bacteria. It involves cleavage of the newly synthesized transcript, followed by template-independent addition of A's at its new 3' end (polyadenylation).

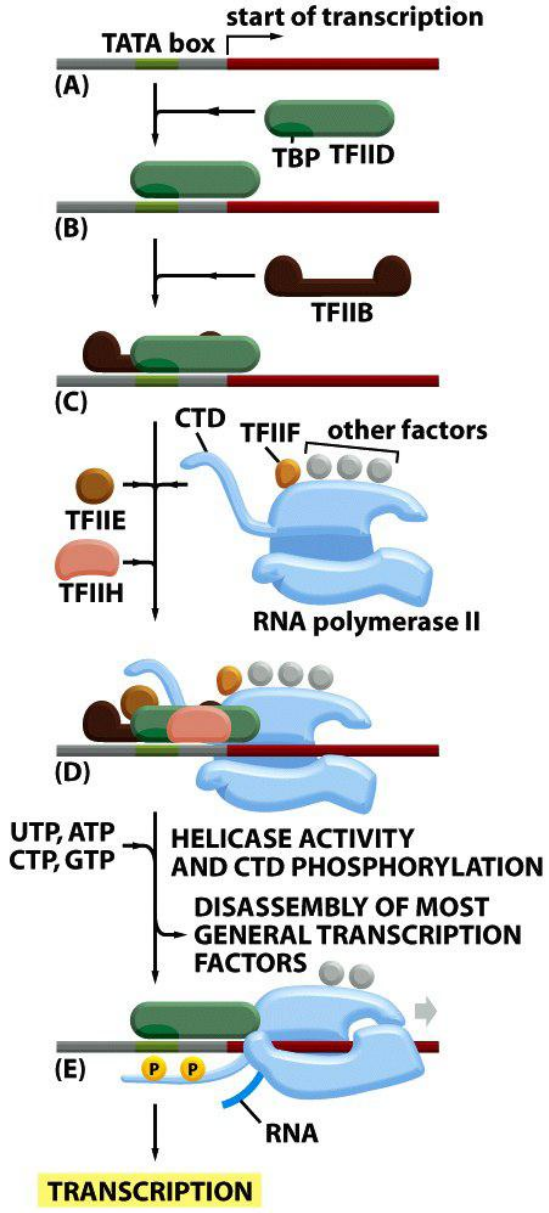
NAME	NUMBER OF SUBUNITS	ROLES IN TRANSITION INITIATION
TFIID		
TBP subunit	1	recognizes TATA box
TAF subunits	~11	recognizes other DNA sequences near the transcription start point; regulates DNA-binding by TBP
TFIIB	1	recognizes BRE element in promoters; accurately positions RNA polymerase at the start site of transcription
TFIIF	3	stabilizes RNA polymerase interaction with TBP and TFIIB; helps attract TFIIE and TFIIH
TFIIE	2	attracts and regulates TFIIH
TFIIH	9	unwinds DNA at the transcription start point, phosphorylates Ser5 of the RNA polymerase CTD; releases RNA polymerase from the promoter

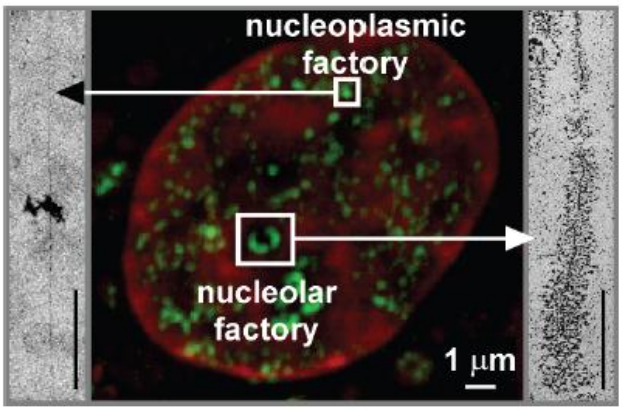
TFIID is composed of TBP and ~11 additional subunits called TAFs (TBP-associated factors); CTD, C-terminal domain.

Eukaryotic RNA polymerases need a plethora of accessory factors in order to initiate and complete transcription. Transcription factors helping RNA polymerases II to perform transcription are designated TFII B, D...)

RNA Pol II: Transcription initiation at TATA box containing promoters:

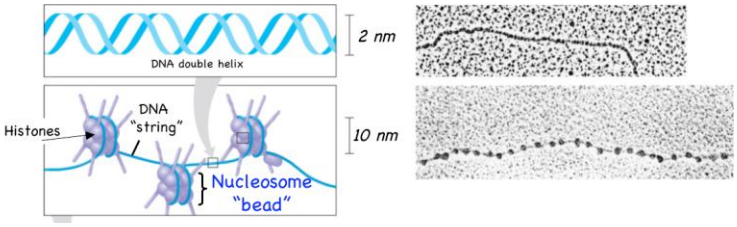
- TFIID binds to the TATA box in the promoter through its TBP subunit and DNA is distorted; thereby, signaling the presence of an active promoter
- Additional TFs (TFIIB) assemble at the promoter
- RNA pol II core is recruited to the promoter together with additional TFs (TFIIE, TFIIH)
- A complete transcription initiation complex is assembled at the promoter
- The helicase activity of TFIIH unwinds a short stretch of promoter DNA allowing RNA pol II to start transcribing short abortive RNA molecules; THIIH phosphorylates the repeat containing C-terminal tail (CTD) of RNA pol II at serine 5
- RNA pol II dissociates from the cluster of general transcription factors and undergoes conformational changes that reinforce association of the polymerase with DNA -> transcription can start





Visualization of transcription: Active transcription units are clustered in the nucleus of eukaryotic cells in discrete sites called transcription factories. Such sites can be visualized using labeled ribonucleotides (Br-UTP, visualization of nascent RNA) or using antibodies directed against polymerases. There are about 10'000 factories in the nucleus of one human HeLa cell. In green, extended nascent RNA, nuclear DNA stained in red. Factories in the nucleoplasm contain RNA pol II and III. Factories in the nucleoli are made by RNA polymerase I and factories in the cytoplasm are made by mitochondrial polymerases.

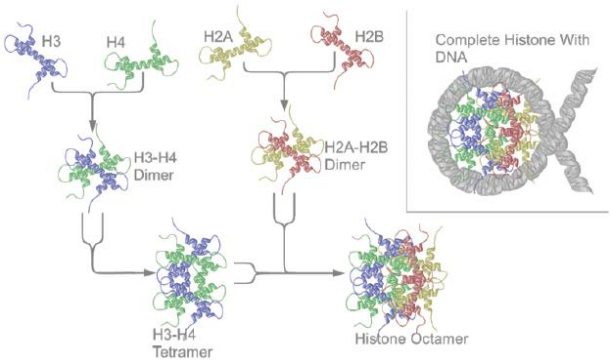
Chromatin and Transcription Regulation



Chromatin

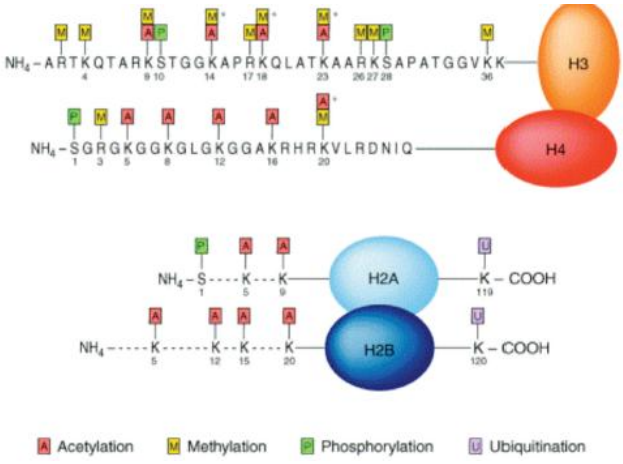
Genomes are enormous and to be assembled in a cell nucleus several layers of structural organization are needed. DNA within a cell (2nm) is packaged into chromatin, which consists of **DNA**, **histone proteins** and **non-histone proteins**. Within chromatin, the basic repeat unit is the **nucleosome**, which is composed of 146bp of two superhelical turns of DNA wrapped around the core of 8 histones (10nm). In **beads on a string**, about 50bp of DNA separate two adjacent nucleosomes. There are a total of six classes of histones: **H1, H2A, H2B, H3, H4 and H5**. Histone classes are organized in super classes, the **core histones** (H2A, H2B, H3 and H4) and **linker histones** (H1 and H5).

Nucleosome assembly: 2 H3-H4 dimers associate in a tetramer structure and associate with 2 H2A-H2B dimers, thereby forming a histone octamer able to organize 146bp of dsDNA. The linker histone H1 binds the nucleosome at the entry and exit sites of the DNA, thus locking the DNA into place and allowing the formation of higher order structures.

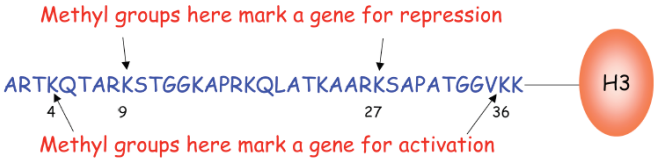


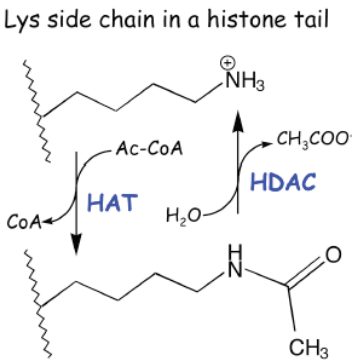
Regulation of chromatin structure: Two main chromatin states are present in interphase cells: **euchromatin** which is an open and highly extended state, readily accessible to transcription and **heterochromatin** which is a highly condensed state, refractory (but not completely resistant) to transcription.

Histone modifications: Each core histone is composed of a globular structured domain and an unstructured amino-terminal tail of 25-40 residues. This unstructured tail extends through the DNA gyres and into the space surrounding the nucleosomes. Histone tials provide sites for a variety of post-translational modifications, inclding **acetylation**, **phosphorylation** and **methylation**. Histone tail modifications determine the interactions of histones with DNA and other proteins, which in turn regulates chromatin structure.



The histone code hypothesis: The constellation of histone post-translational modifications associated with any gene is a code dictating the activity state of that gene. Lysine acetylation and methylation are of primary importance for the histone code. Lysine acetylation generally leads to transcriptional activation (with exceptions) and lysine methylation can lead to activation or repression depending on the site.





Histone acetylation is a covalent transfer of an acetyl group (-$COCH_3$) from acetyl-coenzyme a to lysine residues in core histone tails. Acetylation removes positive charges from histones, thereby reducing the affinity between histones and DNA and generating a more relaxed chromatin state. RNA polymerase and transcription factors can access promoter regions more easily. Therefore, in most cases, histone acetylation enhances transcription while histone deacetylation represses transcription. Histone acetylation is catalyzed by **histone acetyltransferases (HATs)** and histone deacetylation is catalyzed by **histone deacetylases (HDACs)**. Several different forms of HATs and HDACs have been identified. Among them, CBP/p300 is probably the most important HAT, since it can interact with numerous transcription regulators.

Histone methylation is a covalent transfer of a methyl group (-CH_3) from the cofactor S-adenosyl methionine to lysine or arginine residues in core histone tails. Mehylation is generally associated with transcriptional repression. However, methylation of some lysine and arginine residues of histones result in transcriptional activation. Examples include methylation of lysine 4 of histone 3 (H3K4) and arginine residues on H3 and H4. One single histone lysine can be mono-, di- or tri-methylated by histone **methyltransferase** enzymes. For many years histone methylation was thought to be a permanent modification. Very recently, two families of histone demethylating enzymes were discovered:

- **Lysine specific demethylase 1 (LSD1),** a flavin-dependent monoamine oxidase which can demethylate mono- and di-methylated lysines, specifically H3K4 and H3K9
- **Jumonji domain-contianing (JMJC) histone demethylases** , which are able to demethylate mono-, di- or tri-methylated lysines thereby disproving the theory that histone methylation is permanent once and for all

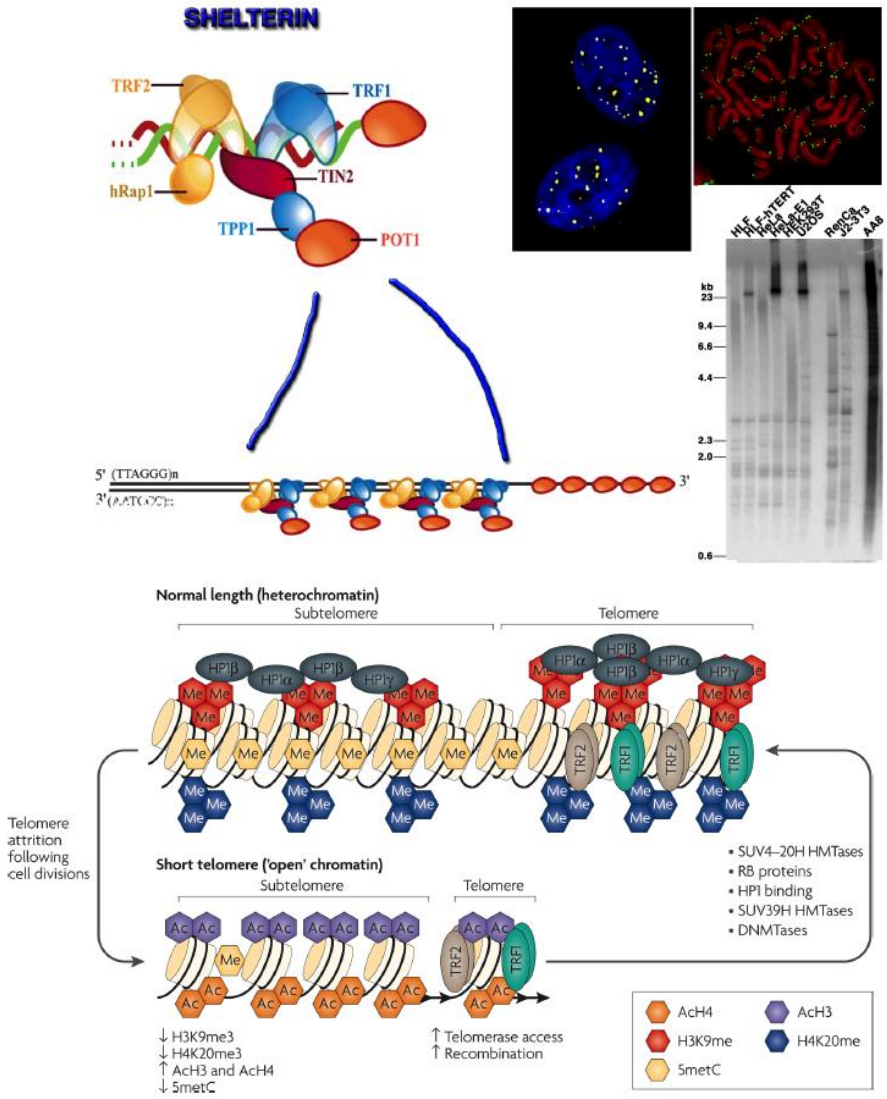
Type of modification						
	H3K4	H3K9	H3K27	H3K79	H4K20	H2BK5
monomethylation	activation	activation	activation	activation	activation	activation
dimethylation				activation		
trimethylations	activation	repression	repression	repression activation		
	H3K9	H3K14				
acetylation	activation	activation				

DNA methylation is a covalent chemical modification of DNA that can be inherited or removed without changing the original DNA sequence. As such, it is part of the epigenetic code and is also the best characterized epigenetic mechanism. DNA methylation in humans regulates epigenetic memory storage. DNA methylation involves the addition of a methyl group (CH_3) to DNA, e.g., a methyl group can be added to the number 5 carbon of the cytosine pyrimidine ring. Cytosine-methylation has been found in every vertebrate examined. In adult somatic tissues, DNA methylation typically occurs at CpG dinucleotides while non-CpG methylation is prevalent in embryonic stem cells.

In mammals, DNA methylation is carried out by two general classes of enzymatic activities: **maintenance methylation** and **de novo methylation DNA methyltransferases**. Maintenance methylation activity is necessary to preserve DNA methylation after every cellular DNA replication cycle and is essentially carried out by **DNMT1**. DNMT1 copies DNA methylation patterns to the daughter strands during DNA replication. De novo methylation is essentially carried out by **DNMT3a** and **DNMT3b**, that set up DNA

methylation patterns early in development. DNA methylation at CpG island broad type promoters negatively regulates gene expression.

Telomeres

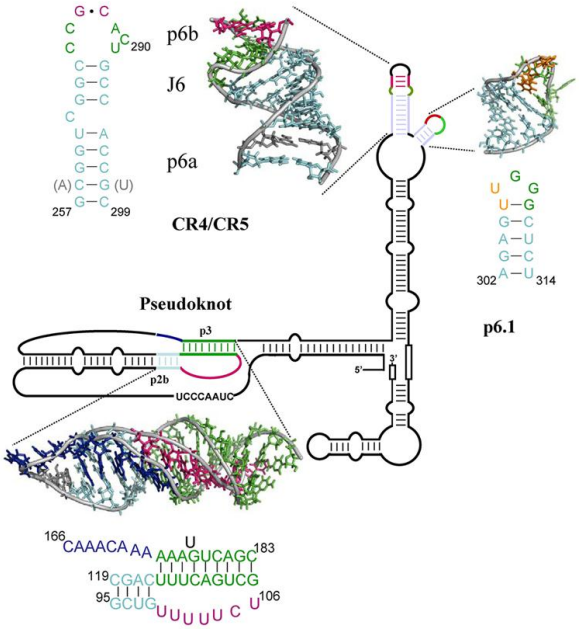


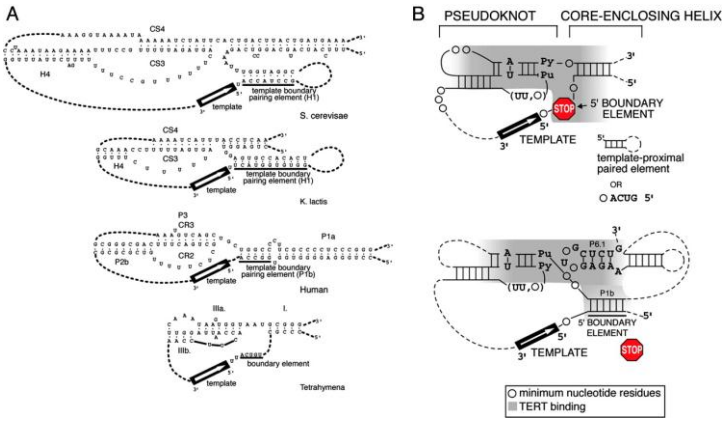
Telomeres are nucleoprotein structures at the end of linear eukaryotic chromosomes. They possess heterochromatic features and the heterochromatic state of telomeres directly correlates with telomere length. Telomeres protect from end fusions (telomeres are not breaks). Moreover, they limit proliferation (clocks). In normal somatic cells, telomerase is absent which leads to cellular senescence, whereas in immortal or cancerous cells, telomerase is active and unlimited division can occur.

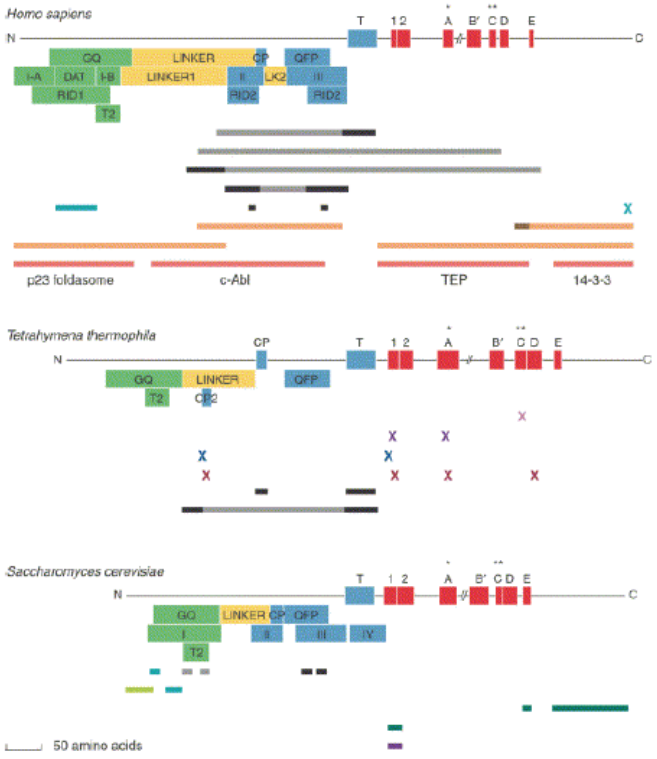
Shelterin complex, consisting of six proteins in humans, TRF1/2, bind DNA-sequence, POT1 binds ss part of telomeres and shelterin.

Telomerase was first discovered in *Tetrahymena thermophila* by Elizabeth Blackburn and Carol Greider. Telomerase is a specialized **reverse transcriptase** that adds newly synthesized telomeric repeats to the 3' end of a telomeric overhang. Telomerase carried its own RNA molecule, which is used as a template when it elongates telomeres. Telomerase core enzymes comprise a catalytic subunit (the **telomerase reverse transcriptase**, TERT in humans, Est2p in yeast), and an RNA moiety containing the template sequence used during reverse transcription (TR in humans or TLC1 in yeast).

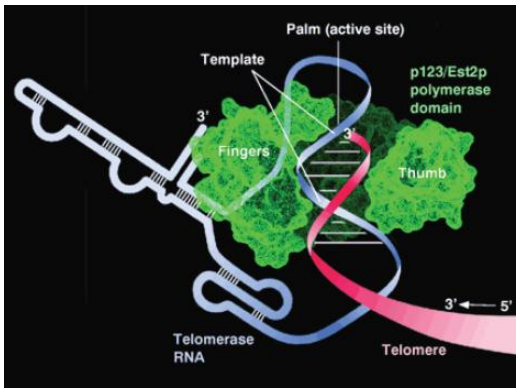
Telomerase RNA folds into a precise secondary structure leaving the template region exposed. Human telomerase RNA is 451nt long and an RNA pol II transcript, in mouse 397nt pol II, budding yeast 1300nt pol II and protozoan 160-190nt and pol III. Although different in size and sequence, the telomerase RNA secondary structure is conserved throughout different eukaryotes (N-terminal domain green, RNA binding domain blue, reverse transcriptase domain red).







Telomerase Structure: The high resolution structure of the *T. castaneum* TERT subunit reveals that the protein consists of four conserved domains (TRBD, finders, pal, and thumb), organized into a ring configuration that shares common features with retroviral reverse transcriptases, viral RNA polymerases and bacteriophage B-family DNA polymerase.



Telomerase action *in vitro* requires only TERT/Est2p and TR/TLC1 subunits, while *in vivo* it requires a number of accessory factors, e.g. Cdc13 and Est1p in yeast and POT1 and TPP1 in humans.

Telomere position effect (TPE) is a cellular phenomenon observed in eukaryotic cells resulting in the reversible silencing of genes near telomeres. TPE was originally discovered in *Drosophila* but is has been extensively characterized in the budding yeast *Saccharomyces cerevisiae*. More recently, TPE was demonstrated also in human cells. TPE originates in the telomeric tract and spreads towards centromeres to modulate expression of genes placed subtelomerically. In cultured cells, TPE silences the expression of luciferase reporter genes inserted in close proximity of a telomere. TPE is revesres by Trichostatin A treatment. TPE strength directly correlates with telomere length (forced elongation of telomeres induces increased TPE).