**Transcription process**

# Transcription

Transcription is the first step in gene expression in which information from a gene is used to construct a functional product such as a protein. The goal of transcription is to make a RNA copy of a gene's DNA sequence. For a protein-coding gene, the RNA copy, or transcript, carries the information needed to build a polypeptide (protein or protein subunit). Eukaryotic transcripts need to go through some processing steps before translation into proteins.

* **Transcription** is the process in which a gene's DNA sequence is copied (transcribed) to make an RNA molecule.
* **RNA polymerase** is the main transcription enzyme.
* Transcription begins when RNA polymerase binds to a **promoter** sequence near the beginning of a gene (directly or through helper proteins).
* RNA polymerase uses one of the DNA strands (the **template strand**) as a template to make a new, complementary RNA molecule.
* Transcription ends in a process called **termination**. Termination depends on sequences in the RNA, which signal that the transcript is finished.

RNA polymerase is crucial because it carries out **transcription**, the process of copying DNA (deoxyribonucleic acid, the genetic material) into RNA (ribonucleic acid, a similar but more short-lived molecule).

Transcription is an essential step in using the information from genes in our DNA to make proteins. Proteins are the key molecules that give cells structure and keep them running. Blocking transcription with mushroom toxin causes liver failure and death, because no new RNAs—and thus, no new proteins—can be made.^22squared

Transcription is essential to life, and understanding how it works is important to human health.

Transcription proceeds in the following general steps:

1. RNA polymerase, together with one or more general transcription factors binds to  promotor DNA.
2. RNA polymerase creates a Transcription bubble, which separates the two strands of the DNA helix. This is done by breaking the hydrogen bonds between complementary DNA nucleotides.9
3. RNA polymerase adds RNA nucleotides (which are complementary to the nucleotides of one DNA strand).
4. RNA sugar-phosphate backbone forms with assistance from RNA polymerase to form an RNA strand.
5. Hydrogen bonds of the RNA–DNA helix break, freeing the newly synthesized RNA strand.
6. If the cell has a nucleus, the RNA may be further processed. This may include polyadenylation, capping, and splicing.
7. The RNA may remain in the nucleus or exit to the cytoplasm through the nuclear pore complex.

The stretch of DNA transcribed into an RNA molecule is called a ***transcription unit*** and encodes at least one gene. If the gene encodes a protein, the transcription produces messenger RNA (mRNA); the **mRNA**, in turn, serves as a template for the protein's synthesis through translation. Alternatively, the transcribed gene may encode for non-coding RNA such as microRNA, ribosomal RNA (rRNA), transfer RNA (tRNA), or enzymatic RNA molecules called ribosomes Overall, RNA helps synthesize, regulate, and process proteins; it therefore plays a fundamental role in performing functions within a cell.

In Virology, the term may also be used when referring to mRNA synthesis from an RNA molecule (i.e., RNA replication). For instance, the genome of a negative-sense single-stranded RNA (ssRNA -) virus may be template for a positive-sense single-stranded RNA (ssRNA +). This is because the positive-sense strand contains the information needed to translate the viral proteins for viral replication afterwards. This process is catalysed by a viral RNA replicase.

In transcription, a region of DNA opens up. One strand, the template strand, serves as a template for synthesis of a complementary RNA transcript. The other strand, the coding strand, is identical to the RNA transcript in sequence, except that it has uracil (U) bases in place of thymine (T) bases.

Example:

Coding strand: 5'-ATGATCTCGTAA-3'
Template strand: 3'-TACTAGAGCATT-5'
RNA transcript: 5'-AUGAUCUCGUAA-3'

For a protein-coding gene, the RNA transcript contains the information needed to synthesize a polypeptide (protein or protein subunit) with a particular amino acid sequence. In this case:

RNA transcript (acting as messenger RNA): 5'-AUGAUCUCGUAA-3'
Polypeptide: Met-Ile-Ser-STOP

**RNA polymerase**

The main enzyme involved in transcription is **RNA polymerase**, which uses a single-stranded DNA template to synthesize a complementary strand of RNA. Specifically, RNA polymerase builds an RNA strand in the 5' to 3' direction, adding each new nucleotide to the 3' end of the Strand.

RNA polymerase synthesizes an RNA strand complementary to a template DNA strand. It synthesizes the RNA strand in the 5' to 3' direction, while reading the template DNA strand in the 3' to 5' direction. The template DNA strand and RNA strand are antiparallel.

RNA transcript: 5'-UGGUAGU...-3' (dots indicate where nucleotides are still being added at 3' end)
DNA template: 3'-ACCATCAGTC-5'

Figure 1RNA polymerase

**Stages of transcription**

Transcription of a gene takes place in three stages: initiation, elongation, and termination. Here, we will briefly see how these steps happen in bacteria.

**Initiation**

 RNA polymerase binds to a sequence of DNA called the **promoter**, found near the beginning of a gene. Each gene (or group of co-transcribed genes, in bacteria) has its own promoter. Once bound, RNA polymerase separates the DNA strands, providing the single-stranded template needed for transcription. Transcription begins with the binding of RNA polymerase, together with one or more general transcription factors to a specific DNA sequence referred to as a "promotor" to form an RNA polymerase-promoter "closed complex". In the "closed complex" the promoter DNA is still fully double-stranded.

RNA polymerase, assisted by one or more general transcription factors, then unwinds approximately 14 base pairs of DNA to form an RNA polymerase-promoter "open complex". In the "open complex" the promoter DNA is partly unwound and single-stranded. The exposed, single-stranded DNA is referred to as the "transcription bubble."

RNA polymerase, assisted by one or more general transcription factors, then selects a **transcription start site** in the transcription bubble, binds to an initiating NTP and an extending NTP (or a short RNA primer and an extending NTP) complementary to the transcription start site sequence, and catalyses bond formation to yield an initial RNA product.

In bacteria, RNA polymerase holoenzyme consists of five subunits: 2 α subunits, 1 β subunit, 1 β' subunit, and 1 ω subunit. In bacteria, there is one general RNA transcription factor known as a sigma factor. RNA polymerase core enzyme binds to the bacterial general transcription (sigma) factor to form RNA polymerase holoenzyme and then binds to a promoter. (RNA polymerase is called a holoenzyme when sigma subunit is attached to the core enzyme which is consist of 2 α subunits, 1 β subunit, 1 β' subunit only).

In archaea and eukaryotes, RNA polymerase contains subunits homologous to each of the five RNA polymerase subunits in bacteria and also contains additional subunits. In archaea and eukaryotes, the functions of the bacterial general transcription factor sigma are performed by multiple general transcription factors that work together.[[5]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-MBOG-5) In archaea, there are three general transcription factors: [TBP](https://en.wikipedia.org/wiki/TATA-binding_protein), [TFB](https://en.wikipedia.org/wiki/Archaeal_transcription_factor_B), and [TFE](https://en.wikipedia.org/wiki/2,2,2-Trifluoroethanol). In eukaryotes, in  RNA polymerase 2 -dependent transcription, there are six general transcription factors: [TFIIA](https://en.wikipedia.org/wiki/TFIIA), [TFIIB](https://en.wikipedia.org/wiki/TFIIB) (an [ortholog](https://en.wikipedia.org/wiki/Orthologous) of archaeal TFB), [TFIID](https://en.wikipedia.org/wiki/TFIID) (a multi subunit factor in which the key subunit, [TBP](https://en.wikipedia.org/wiki/TATA-binding_protein), is an [ortholog](https://en.wikipedia.org/wiki/Orthologous) of archaeal TBP), [TFIIE](https://en.wikipedia.org/wiki/TFIIE) (an [ortholog](https://en.wikipedia.org/wiki/Orthologous) of archaeal TFE), [TFIIF](https://en.wikipedia.org/wiki/TFIIF), and [TFIIH](https://en.wikipedia.org/wiki/TFIIH). The TFIID is the first component to bind to DNA due to binding of TBP, while TFIIH is the last component to be recruited. In archaea and eukaryotes, the RNA polymerase-promoter closed complex is usually referred to as the "[preinitiation complex](https://en.wikipedia.org/wiki/Transcription_preinitiation_complex)."[[6]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-Roeder1991-6)

Transcription initiation is regulated by additional proteins, known as [activators](https://en.wikipedia.org/wiki/Activator_(genetics)) and [repressors](https://en.wikipedia.org/wiki/Repressor), and, in some cases, associated [coactivators](https://en.wikipedia.org/wiki/Coactivator) or [corepressors](https://en.wikipedia.org/wiki/Coactivator), which modulate formation and function of the transcription initiation complex.[[5]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-MBOG-5)

### **Promoter escape**

After the first bond is synthesized, the RNA polymerase must escape the promoter. During this time there is a tendency to release the RNA transcript and produce truncated transcripts. This is called [abortive initiation](https://en.wikipedia.org/wiki/Abortive_initiation), and is common for both eukaryotes and prokaryotes.[[7]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-7) Abortive initiation continues to occur until an RNA product of a threshold length of approximately 10 nucleotides is synthesized, at which point promoter escape occurs and a transcription elongation complex is formed.

Mechanistically, promoter escape occurs through [DNA scrunching](https://en.wikipedia.org/wiki/DNA_scrunching), providing the energy needed to break interactions between RNA polymerase holoenzyme and the promoter.[[8]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-8)

In bacteria, it was historically thought that the [sigma factor](https://en.wikipedia.org/wiki/Sigma_factor) is definitely released after promoter clearance occurs. This theory had been known as the *obligate release model* however later data showed that upon and following promoter clearance, the sigma factor is released according to a [stochastic model](https://en.wikipedia.org/wiki/Stochastic_process) known as the *stochastic release model*.[[9]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-9)

In eukaryotes, at an RNA polymerase II-dependent promoter, upon promoter clearance, TFIIH phosphorylates serine 5 on the carboxy terminal domain of RNA polymerase II, leading to the recruitment of capping enzyme (CE).[[10]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-10)[[11]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-11) The exact mechanism of how CE induces promoter clearance in eukaryotes is not yet known.

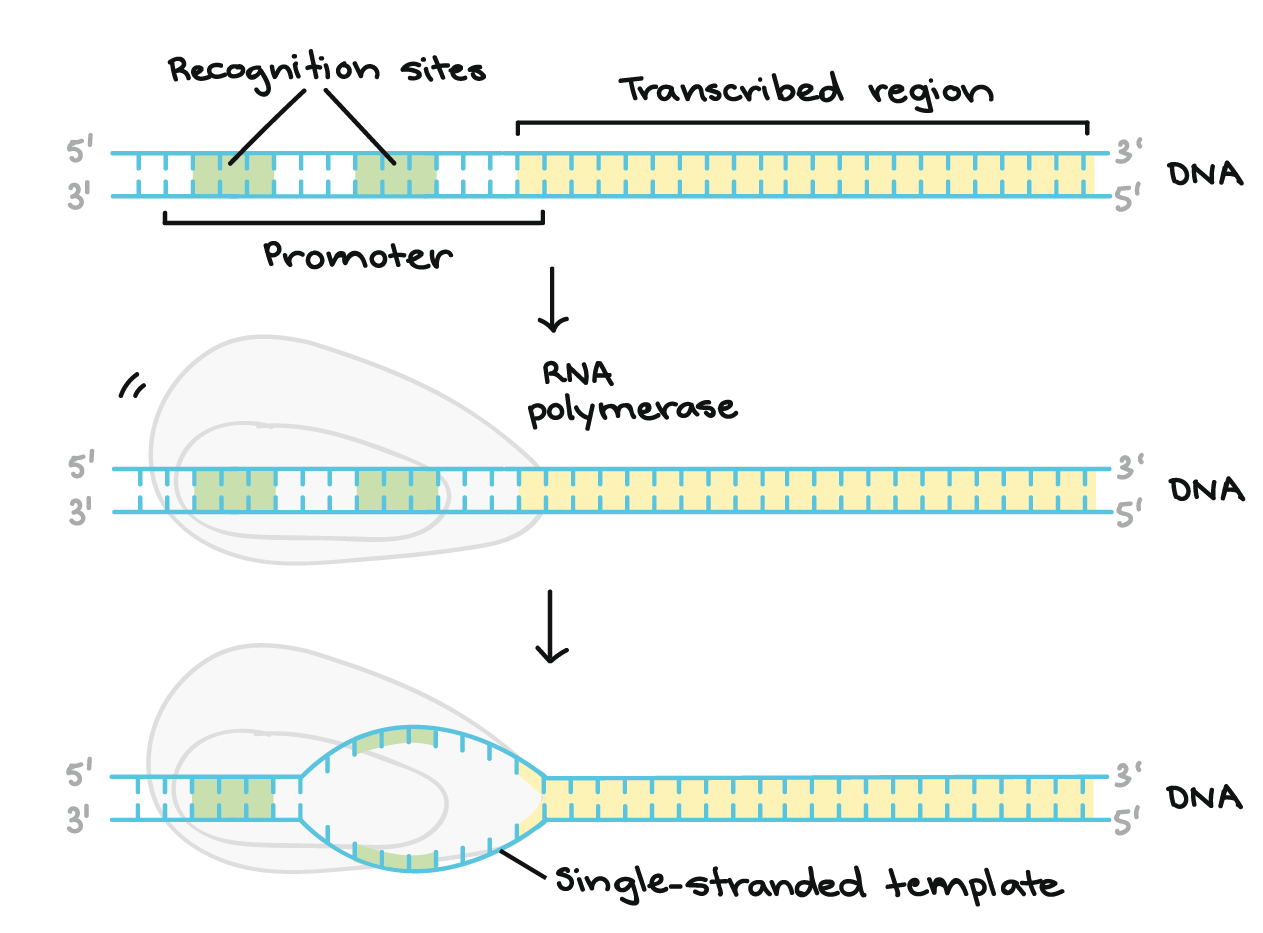


Figure promoter escape

The promoter region comes before (and slightly overlaps with) the transcribed region whose transcription it specifies. It contains recognition sites for RNA polymerase or its helper proteins to bind to. The DNA opens up in the promoter region so that RNA polymerase can begin transcription.

**2.Elongation**

One strand of the DNA, the *template strand* (or noncoding strand), is used as a template for RNA synthesis. As transcription proceeds, RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy (which elongates during the traversal). Although RNA polymerase traverses the template strand from 3' → 5', the coding (non-template) strand and newly formed RNA can also be used as reference points, so transcription can be described as occurring 5' → 3'.

This produces an RNA molecule from 5' → 3', an exact copy of the coding strand (except that [thymine](https://en.wikipedia.org/wiki/Thymine) are replaced with [uracil](https://en.wikipedia.org/wiki/Uracil), and the nucleotides are composed of a ribose (5-carbon) sugar where DNA has deoxyribose (one fewer oxygen atom) in its sugar-phosphate backbone mRNA transcription can involve multiple RNA polymerases on a single DNA template and multiple rounds of transcription (amplification of particular mRNA), so many mRNA molecules can be rapidly produced from a single copy of a gene. The characteristic elongation rates in prokaryotes and eukaryotes are about 10-100 nts/sec.[[12]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-12) In eukaryotes, however, [nucleosomes](https://en.wikipedia.org/wiki/Nucleosome) act as major barriers to transcribing polymerases during transcription elongation.[[13]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-13)[[14]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-:0-14) In these organisms, the pausing induced by nucleosomes can be regulated by transcription elongation factors such as TFIIS.[[14]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-:0-14)

Elongation also involves a proofreading mechanism that can replace incorrectly incorporated bases. In eukaryotes, this may correspond with short pauses during transcription that allow appropriate RNA editing factors to bind. These pauses may be intrinsic to the RNA polymerase or due to chromatin structure.

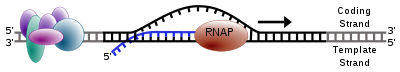
[](https://en.wikipedia.org/wiki/File:Simple_transcription_elongation1.svg)

Figure 3diagram of transcription

RNA polymerase synthesizes an RNA transcript complementary to the DNA template strand in the 5' to 3' direction. It moves forward along the template strand in the 3' to 5' direction, opening the DNA double helix as it goes. The synthesized RNA only remains bound to the template strand for a short while, then exits the polymerase as a dangling string, allowing the DNA to close back up and form a double helix.

In this example, the sequences of the coding strand, template strand, and RNA transcript are:

Coding strand: 5' - ATGATCTCGTAA-3'

Template strand: 3'-TACTAGAGCATT-5'

RNA: 5'-AUGAUC...-3' (the dots indicate where nucleotides are still being added to the RNA strand at its 3' end)

Figure 4elongation step

RNA polymerase synthesizes an RNA transcript complementary to the DNA template strand in the 5' to 3' direction. It moves forward along the template strand in the 3' to 5' direction, opening the DNA double helix as it goes. The synthesized RNA only remains bound to the template strand for a short while, then exits the polymerase as a dangling string, allowing the DNA to close back up and form a double helix.

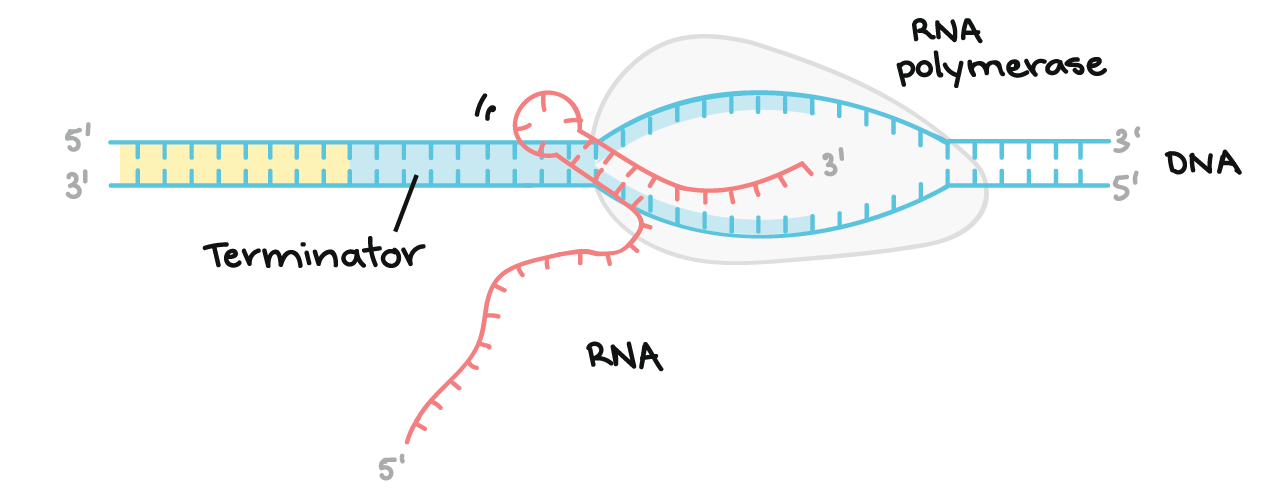
In this example, the sequences of the coding strand, template strand, and RNA transcript are:

Coding strand: 5' - ATGATCTCGTAA-3'

Template strand: 3'-TACTAGAGCATT-5'

RNA: 5'-AUGAUC...-3' (the dots indicate where nucleotides are still being added to the RNA strand at its 3' end)

1. **Termination.** Sequences called **terminators** signal that the RNA transcript is complete. Once they are transcribed, they cause the transcript to be released from the RNA polymerase. An example of a termination mechanism involving formation of a hairpin in the RNA is shown below.



Bacteria use two different strategies for transcription termination – Rho-independent termination and Rho-dependent termination. In [Rho-independent transcription termination](https://en.wikipedia.org/wiki/Rho-independent_transcription_termination), RNA transcription stops when the newly synthesized RNA molecule forms a G-C-rich [hairpin loop](https://en.wikipedia.org/wiki/Hairpin_loop) followed by a run of Us. When the hairpin forms, the mechanical stress breaks the weak rU -dA bonds, now filling the DNA–RNA hybrid. This pulls the poly-U transcript out of the active site of the RNA polymerase, terminating transcription. In the "Rho-dependent" type of termination, a protein factor called "[Rho](https://en.wikipedia.org/wiki/Rho_factor)" destabilizes the interaction between the template and the mRNA, thus releasing the newly synthesized mRNA from the elongation complex.[[15]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-15)

Transcription termination in eukaryotes is less well understood than in bacteria, but involves cleavage of the new transcript followed by template-independent addition of adenines at its new 3' end, in a process called [polyadenylation](https://en.wikipedia.org/wiki/Polyadenylation).[[16]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-16)

## Inhibitors

Transcription inhibitors can be used as [antibiotics](https://en.wikipedia.org/wiki/Antibiotic) against, for example, [pathogenic bacteria](https://en.wikipedia.org/wiki/Pathogenic_bacteria) ([anti bacterials](https://en.wikipedia.org/wiki/Antibacterial)) and [fungi](https://en.wikipedia.org/wiki/Fungus) ([antifungals](https://en.wikipedia.org/wiki/Antifungal_medication)). An example of such an antibacterial is [rifampicin](https://en.wikipedia.org/wiki/Rifampicin), which inhibits [bacterial transcription](https://en.wikipedia.org/wiki/Bacterial_transcription) of DNA into mRNA by inhibiting DNA-dependent [RNA polymerase](https://en.wikipedia.org/wiki/RNA_polymerase) by binding its beta-subunit, while [8-hydroxyquinoline](https://en.wikipedia.org/wiki/8-Hydroxyquinoline) is an antifungal transcription inhibitor.[[17]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-17) The effects of [histone methylation](https://en.wikipedia.org/wiki/Histone_methylation) may also work to inhibit the action of transcription.

## Endogenous inhibitors

In vertebrates, the majority of gene [promoters](https://en.wikipedia.org/wiki/Promoter_(genetics)#CpG_islands_in_promoters) contain a [CpG island](https://en.wikipedia.org/wiki/CpG_site#CpG_islands) with numerous [CpG sites](https://en.wikipedia.org/wiki/CpG_site).[[18]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-pmid16432200-18) When many of a gene's promoter CpG sites are [methylated](https://en.wikipedia.org/wiki/DNA_methylation) the gene becomes inhibited (silenced).[[19]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-Bird-19) Colorectal cancers typically have 3 to 6 [driver](https://en.wikipedia.org/wiki/Somatic_evolution_in_cancer#Glossary) mutations and 33 to 66 [hitchhiker](https://en.wikipedia.org/wiki/Genetic_hitchhiking) or passenger mutations.[[20]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-pmid23539594-20) However, transcriptional inhibition (silencing) may be of more importance than mutation in causing progression to cancer. For example, in colorectal cancers about 600 to 800 genes are transcriptionally inhibited by CpG island methylation (see [regulation of transcription in cancer](https://en.wikipedia.org/wiki/Regulation_of_transcription_in_cancer)). Transcriptional repression in cancer can also occur by other [epigenetic](https://en.wikipedia.org/wiki/Cancer_epigenetics) mechanisms, such as altered expression of [microRNAs](https://en.wikipedia.org/wiki/MicroRNA#DNA_repair_and_cancer).[[21]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-pmid24616890-21) In breast cancer, transcriptional repression of [BRCA1](https://en.wikipedia.org/wiki/BRCA1) may occur more frequently by over-expressed microRNA-182 than by hypermethylation of the BRCA1 promoter (see [Low expression of BRCA1 in breast and ovarian cancers](https://en.wikipedia.org/wiki/BRCA1#Low_expression_of_BRCA1_in_breast_and_ovarian_cancers)).

## Transcription factories

Active transcription units are clustered in the nucleus, in discrete sites called [transcription factories](https://en.wikipedia.org/wiki/Transcription_factories) or [euchromatin](https://en.wikipedia.org/wiki/Euchromatin). Such sites can be visualized by allowing engaged polymerases to extend their transcripts in tagged precursors (Br-UTP or Br-U) and immuno-labelling the tagged nascent RNA. Transcription factories can also be localized using fluorescence in situ hybridization or marked by antibodies directed against polymerases. There are ~10,000 factories in the nucleoplasm of a [HeLa cell](https://en.wikipedia.org/wiki/HeLa_cell), among which are ~8,000 polymerase II factories and ~2,000 polymerase III factories. Each polymerase II factory contains ~8 polymerases. As most active transcription units are associated with only one polymerase, each factory usually contains ~8 different transcription units. These units might be associated through promoters and/or enhancers, with loops forming a "cloud" around the factor.[[22]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-22)

**Eukaryotic RNA modifications**

In bacteria, RNA transcripts can act as **messenger RNAs** (**mRNAs**) right away. In eukaryotes, the transcript of a protein-coding gene is called a **pre-mRNA** and must go through extra processing before it can direct translation.

* Eukaryotic pre-mRNAs must have their ends modified, by addition of a **5' cap** (at the beginning) and **3' poly-A tail** (at the end).
* Many eukaryotic pre-mRNAs undergo **splicing**. In this process, parts of the pre-mRNA (called **introns**) are chopped out, and the remaining pieces (called **exons**) are stuck back together.

Top of image: Diagram of a pre-mRNA with a 5' cap and 3' poly-A tail. The 5' cap is on the 5' end of the pre-mRNA and is a modified G nucleotide. The poly-A tail is on the 3' end of the pre-mRNA and consists of a long string of A nucleotides (only a few of which are shown).

The pre-mRNA still contains both exons and introns. Along the length of the mRNA, there is an alternating pattern of exons and introns: Exon 1 - Intron 1 - Exon 2 - Intron 2 - Exon 3. Each consists of a stretch of RNA nucleotides.

During splicing, the introns are removed from the pre-mRNA, and the exons are stuck together to form a mature mRNA.

Bottom of image:  Mature mRNA that does not contain the intron sequences (Exon 1 - Exon 2 - Exon 3 only).

Above is shown Diagram of a pre-mRNA with a 5' cap and 3' poly-A tail. The 5' cap is on the 5' end of the pre-mRNA and is a modified G nucleotide. The poly-A tail is on the 3' end of the pre-mRNA and consists of a long string of A nucleotides (only a few of which are shown).

The pre-mRNA still contains both exons and introns. Along the length of the mRNA, there is an alternating pattern of exons and introns: Exon 1 - Intron 1 - Exon 2 - Intron 2 - Exon 3. Each consists of a stretch of RNA nucleotides.

During splicing, the introns are removed from the pre-mRNA, and the exons are stuck together to form a mature mRNA.

Bottom of image: Mature mRNA that does not contain the intron sequences (Exon 1 - Exon 2 - Exon 3 only).

End modifications increase the stability of the mRNA, while splicing gives the mRNA its correct sequence. (If the introns are not removed, they'll be translated along with the exons, producing a "gibberish" polypeptide.)

To learn more about pre-mRNA modifications in eukaryotes, check out the article on [pre-mRNA processing](https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/eukaryotic-pre-mrna-processing).

**Transcription happens for individual genes**

Not all genes are transcribed all the time. Instead, transcription is controlled individually for each gene (or, in bacteria, for small groups of genes that are transcribed together). Cells carefully regulate transcription, transcribing just the genes whose products are needed at a particular moment.

For example, the diagram below shows a "snapshot" of an imaginary cell's RNAs at a given moment in time. In this cell, genes 1, 2 and 3, are transcribed, while gene 4 is not. Also, genes 1, 2, and 3 are transcribed at different levels, meaning that different numbers of RNA molecules are made for each.

Diagram showing that individual genes are transcribed in different amounts.

A region of DNA containing four genes is shown, with the transcribed region of each gene highlighted in dark blue. The number of transcripts of each gene is indicated above the DNA (on a Y- axis). There are six transcripts of gene 1, one transcript of gene 2, twelve transcripts of gene 3, and no transcripts of gene 4.

This is not an illustration of any actual set of genes and their transcription levels, but rather, illustrates that transcription is controlled individually for genes and other transcription units.

Figure 5individual genes are transcribed in different amount

A region of DNA containing four genes is shown, with the transcribed region of each gene highlighted in dark blue. The number of transcripts of each gene is indicated above the DNA (on a Y- axis). There are six transcripts of gene 1, one transcript of gene 2, twelve transcripts of gene 3, and no transcripts of gene 4.

This is not an illustration of any actual set of genes and their transcription levels, but rather, illustrates that transcription is controlled individually for genes and other transcription units.

In the following articles, we'll take a more in-depth look at RNA polymerase, the stages of transcription, and the process of RNA modification in eukaryotes. We'll also consider some important differences between bacterial and eukaryotic transcription.

## Key points:

* When an RNA transcript is first made in a eukaryotic cell, it is considered a **pre-mRNA** and must be processed into a **messenger RNA**(**mRNA)**.
* A **5' cap** is added to the beginning of the RNA transcript, and a 3' **poly-A tail** is added to the end.
* In **splicing**, some sections of the RNA transcript (**introns**) are removed, and the remaining sections (**exons**) are stuck back together.
* Some genes can be **alternatively spliced**, leading to the production of different mature mRNA molecules from the same initial transcript.

. In humans and other eukaryotes, a freshly made RNA transcript (hot off the RNA polymerase "presses") is not quite ready to go. Instead, it's called a **pre-mRNA** and has to go through some processing steps to become a mature **messenger RNA** (**mRNA**) that can be translated into a protein. These include:

* Addition of cap and tail molecules to the two ends of the transcript. These play a protective role, like a book's front and back covers.
* Removal of "junk" sequences called **introns**. Introns are sort of like blank or messed-up pages made during a book's printing, which have to be removed in order for the book to be readable .

In this article, we'll take a closer look at the cap, tail, and splicing modifications that eukaryotic RNA transcripts receive, seeing how they're carried out and why they are important for making sure we get the right protein from our RNA.

## Overview of pre-mRNA processing in eukaryotes

As a quick review, gene expression (the "reading out" of a gene to make a protein, or chunk of a protein) happens a little bit differently in bacteria and eukaryotes such as humans.

Left panel: eukaryotic cell. In the nucleus, a pre-mRNA is produced through transcription of a region of DNA from a linear chromosome. This transcript must undergo processing (splicing and addition of 5' cap and poly-A tail) while it is still in the nucleus in order to become a mature mRNA. The mature mRNA is exported from the nucleus to the cytosol, where it is translated at a ribosome to make a polypeptide.

Right panel: bacterium. The DNA takes the form of a circular chromosome and is located in the cytosol. While the DNA is being transcribed to make an RNA, the RNA (which is already considered a mRNA at this point) can associate with a ribosome and start being translated to make a polypeptide.

Left panel: eukaryotic cell. In the nucleus, a pre-mRNA is produced through transcription of a region of DNA from a linear chromosome. This transcript must undergo processing (splicing and addition of 5' cap and poly-A tail) while it is still in the nucleus in order to become a mature mRNA. The mature mRNA is exported from the nucleus to the cytosol, where it is translated at a ribosome to make a polypeptide.

Right panel: bacterium. The DNA takes the form of a circular chromosome and is located in the cytosol. While the DNA is being transcribed to make an RNA, the RNA (which is already considered a mRNA at this point) can associate with a ribosome and start being translated to make a polypeptide.

In bacteria, RNA transcripts are ready to act as messenger RNAs and get translated into proteins right away. In eukaryotes, things are a little more complex, though in an pretty interesting way. The molecule that's directly made by transcription in one of your (eukaryotic) cells is called a **pre-mRNA**, reflecting that it needs to go through a few more steps to become an actual messenger RNA (mRNA). These are:

* Addition of a **5' cap** to the beginning of the RNA
* Addition of a **poly-A tail** (tail of A nucleotides) to the end of the RNA
* Chopping out of **introns**, or "junk" sequences, and pasting together of the remaining, good sequences (**exons**)

Once it's completed these steps, the RNA is a mature mRNA. It can travel out of the nucleus and be used to make a protein.

## 5' cap and poly-A tail

Both ends of a pre-mRNA are modified by the addition of chemical groups. The group at the beginning (5' end) is called a cap, while the group at the end (3' end) is called a tail. Both the cap and the tail protect the transcript and help it get exported from the nucleus and translated on the ribosomes (protein-making "machines") found in the cytosol^11start superscript, 1, end superscript.

The **5’ cap** is added to the first nucleotide in the transcript during transcription. The cap is a modified guanine (G) nucleotide, and it protects the transcript from being broken down. It also helps the ribosome attach to the mRNA and start reading it to make a protein.

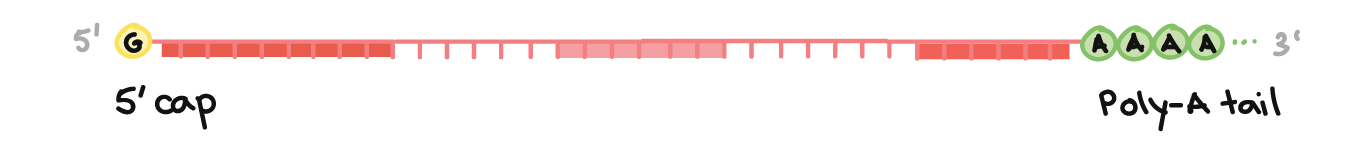


Image of a pre-mRNA with a 5' cap and 3' poly-A tail. The 5' cap is on the 5' end of the pre-mRNA and is a modified G nucleotide. The poly-A tail is on the 3' end of the pre-mRNA and consists of a long string of A nucleotides (only a few of which are shown).

How is the poly-A tail added? The 3' end of the RNA forms in kind of a bizarre way. When a sequence called a **polyadenylation signal** shows up in an RNA molecule during transcription, an enzyme chops the RNA in two at that site. Another enzyme adds about 100100100 - 200200200 adenine (A) nucleotides to the cut end, forming a **poly-A tail**. The tail makes the transcript more stable and helps it get exported from the nucleus to the cytosol.

RNA processing The mRNA which has been transcribed up to this point is referred to as pre-mRNA. Processing must occur to convert this into mature mRNA. The processing

to convert this into mature mRNA. The processing includes:

Capping Addition of a methylated guanine cap confers protection to the mRNA. This is necessary as RNA is much more unstable than DNA. It involves:

Addition of a methylated guanine Occurs at the 5’ end of the mRNA transcript Protects the mRNA from degradation Polyadenylation/tailing Addition of a poly A tail confers protection to the mRNA. This is necessary as RNA is much more unstable than DNA. It involves:

Endonucleases\* recognise a speciﬁc sequence along the mRNA transcript and cleaves it there Several adenine nucleotides are added (approximately 200) to the transcript by the enzyme poly A polymerase Occurs at the 3’ end, protecting the mRNA from degradation \* Endonucleases break within the polynucleotide. Exonucleases degrade the polynucleotide from the ends (either the 5’ or 3’ end)

## RNA splicing

The third big RNA processing event that happens in your cells is **RNA splicing**. In RNA splicing, specific parts of the pre-mRNA, called **introns** are recognized and removed by a protein-and-RNA complex called the **spliceosome**. Introns can be viewed as "junk" sequences that must be cut out so the "good parts version" of the RNA molecule can be assembled.

What are the "good parts"? The pieces of the RNA that are not chopped out are called **exons**. The exons are pasted together by the spliceosome to make the final, mature mRNA that is shipped out of the nucleus.

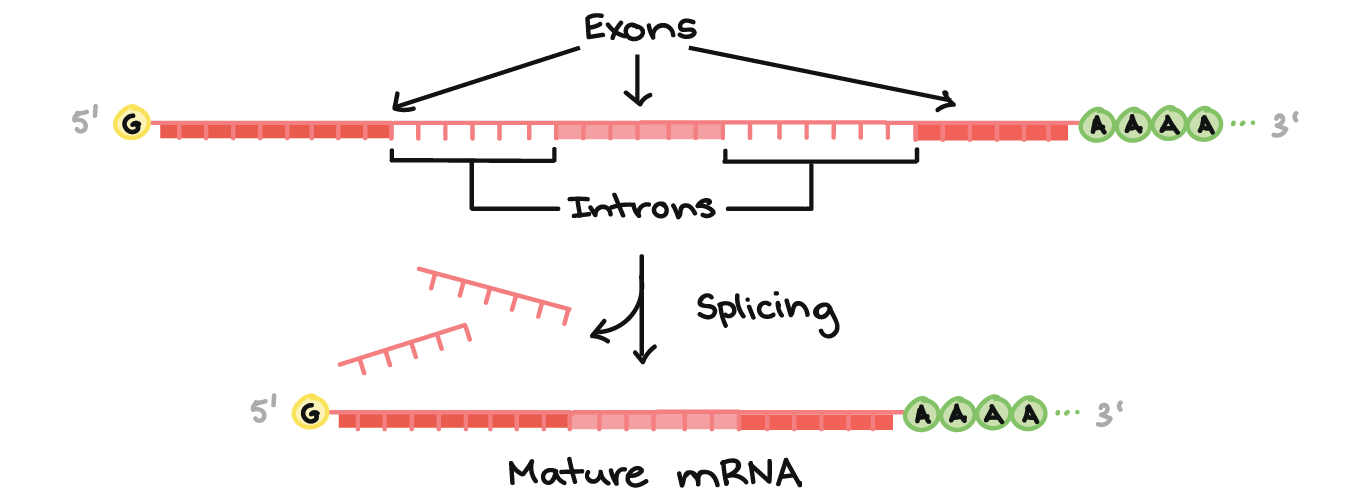


Figure RNA splicing

Diagram of a pre-mRNA showing exons and introns. Along the length of the mRNA, there is an alternating pattern of exons and introns: Exon 1 - Intron 1 - Exon 2 - Intron 2 - Exon 3. Each consists of a stretch of RNA nucleotides. During splicing, the introns are removed from the pre-mRNA, and the exons are stuck together to form a mature mRNA that does not contain the intron sequences.

A key point here is that it's only the exons of a gene that encode a protein. Not only do the introns not carry information to build a protein, they actually have to be removed in order for the mRNA to encode a protein with the right sequence. If the spliceosome fails to remove an intron, an mRNA with extra "junk" in it will be made, and a wrong protein will get produced during translation.

### **Alternative splicing**

Why splice? We don't know for sure why splicing exists, and in some ways, it seems like a wasteful system. However, splicing does allow for a process called **alternative splicing**, in which more than one mRNA can be made from the same gene. Through alternative splicing, we (and other eukaryotes) can sneakily encode more different proteins than we have genes in our DNA.

In alternative splicing, one pre-mRNA may be spliced in either of two (or sometimes many more than two!) different ways. For example, in the diagram below, the same pre-mRNA can be spliced in three different ways, depending on which exons are kept. This results in three different mature mRNAs, each of which translates into a protein with a different structure.

Diagram of alternative splicing.

A sequence of DNA encodes a pre-mRNA transcript that contains five regions that may potentially be used as exons: Exon 1, Exon 2, Exon 3, Exon 4, and Exon 5. The exons are arranged in linear order along the pre-mRNA and have introns in between them.

In splicing event #1, all five exons are retained in the mature mRNA. It consists of Exon 1 - Exon 2 - Exon 3 - Exon 4 - Exon 5. When it is translated, it specifies Protein A, a protein with five domains: Coil 1 (specified by Exon 1), Coil 2 (specified by Exon 2), Loop 3 (specified by Exon 3), Loop 4 (specified by Exon 4), and Coil 5 (specified by Exon 5).

In splicing event #2, Exon 3 is not included in the mature mRNA. It consists of Exon 1 - Exon 2 - Exon 4 - Exon 5. When it is translated, it specifies Protein, B a protein with four domains: Coil 1 (specified by Exon 1), Coil 2 (specified by Exon 2), Loop 4 (specified by Exon 4), and Coil 5 (specified by Exon 5). It does not contain Loop 3 because Exon 3 is not present in the mRNA.

In splicing event #3, Exon 4 is not included in the mature mRNA. It consists of Exon 1 - Exon 2 - Exon 3 - Exon 5. When it is translated, it specifies Protein C, a protein with four domains: Coil 1 (specified by Exon 1), Coil 2 (specified by Exon 2), Loop 3 (specified by Exon 3), and Coil 5 (specified by Exon 5). It does not contain Loop 4 because Exon 4 is not present in the mRNA.

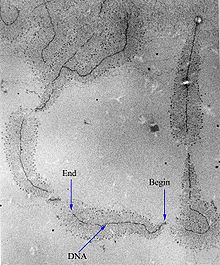
## History

A molecule that allows the genetic material to be realized as a protein was first hypothesized by [François Jacob](https://en.wikipedia.org/wiki/Fran%C3%A7ois_Jacob) and [Jacques Monod](https://en.wikipedia.org/wiki/Jacques_Monod). [Severo Ochoa](https://en.wikipedia.org/wiki/Severo_Ochoa) won a [Nobel Prize in Physiology or Medicine](https://en.wikipedia.org/wiki/Nobel_Prize_in_Physiology_or_Medicine) in 1959 for developing a process for synthesizing RNA [*in vitro*](https://en.wikipedia.org/wiki/In_vitro) with [polynucleotide phosphorylase](https://en.wikipedia.org/wiki/Polynucleotide_phosphorylase), which was useful for cracking the [genetic code](https://en.wikipedia.org/wiki/Genetic_code). RNA synthesis by [RNA polymerase](https://en.wikipedia.org/wiki/RNA_polymerase) was established *in vitro* by several laboratories by 1965; however, the RNA synthesized by these enzymes had properties that suggested the existence of an additional factor needed to terminate transcription correctly.

In 1972, [Walter Fiers](https://en.wikipedia.org/wiki/Walter_Fiers) became the first person to actually prove the existence of the terminating enzyme.

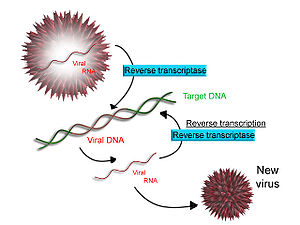
[Roger D. Kornberg](https://en.wikipedia.org/wiki/Roger_D._Kornberg) won the 2006 [Nobel Prize in Chemistry](https://en.wikipedia.org/wiki/Nobel_Prize_in_Chemistry) "for his studies of the molecular basis of [eukaryotic transcription](https://en.wikipedia.org/wiki/Eukaryotic_transcription)".[[23]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-23)

## Measuring and detecting

[](https://en.wikipedia.org/wiki/File:Transcription_label_en.jpg)

[Electron micrograph](https://en.wikipedia.org/wiki/Electron_micrograph) of transcription of ribosomal RNA. The forming [ribosomal RNA](https://en.wikipedia.org/wiki/Ribosomal_RNA) strands are visible as branches from the main DNA strand.[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]

**Transcription can be measured and detected in a variety of ways:**

* [G-Less Cassette](https://en.wikipedia.org/wiki/G-Less_Cassette) transcription assay: measures promoter strength
* [Run-off transcription](https://en.wikipedia.org/wiki/Run-off_transcription) assay: identifies transcription start sites (TSS)
* [Nuclear run-on](https://en.wikipedia.org/wiki/Nuclear_run-on) assay: measures the relative abundance of newly formed transcripts
* [RNase protection assay](https://en.wikipedia.org/wiki/RNase_protection_assay) and [ChIP-Chip](https://en.wikipedia.org/wiki/ChIP-Chip) of [RNAP](https://en.wikipedia.org/wiki/RNAP): detect active transcription sites
* [RT-PCR](https://en.wikipedia.org/wiki/RT-PCR): measures the absolute abundance of total or nuclear RNA levels, which may however differ from transcription rates
* [DNA microarrays](https://en.wikipedia.org/wiki/DNA_microarrays): measures the relative abundance of the global total or nuclear RNA levels; however, these may differ from transcription rates
* [In situ hybridization](https://en.wikipedia.org/wiki/In_situ_hybridization): detects the presence of a transcript
* [MS2 tagging](https://en.wikipedia.org/wiki/MS2_tagging): by incorporating RNA [stem loops](https://en.wikipedia.org/wiki/Stem-loop), such as MS2, into a gene, these become incorporated into newly synthesized RNA. The stem loops can then be detected using a fusion of GFP and the MS2 coat protein, which has a high affinity, sequence-specific interaction with the MS2 stem loops. The recruitment of GFP to the site of transcription is visualized as a single fluorescent spot. This new approach has revealed that transcription occurs in discontinuous bursts, or pulses (see [Transcriptional bursting](https://en.wikipedia.org/wiki/Transcriptional_bursting)). With the notable exception of in situ techniques, most other methods provide cell population averages, and are not capable of detecting this fundamental property of genes.[[24]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-24)
* [Northern blot](https://en.wikipedia.org/wiki/Northern_blot): the traditional method, and until the advent of [RNA-Seq](https://en.wikipedia.org/wiki/RNA-Seq), the most quantitative
* [RNA-Seq](https://en.wikipedia.org/wiki/RNA-Seq): applies next-generation sequencing techniques to sequence whole [transcriptomes](https://en.wikipedia.org/wiki/Transcriptome), which allows the measurement of relative abundance of RNA, as well as the detection of additional variations such as fusion genes, post-transcriptional edits and novel splice sites
* [Single cell RNA-Seq](https://en.wikipedia.org/wiki/Single-cell_RNA-sequencing): amplifies and reads partial transcriptomes from isolated cells, allowing for detailed analyses of RNA in tissues, embryos, and cancers
* Reverse transcription
  + [](https://en.wikipedia.org/wiki/File:RetroTranscription.jpg)
* Scheme of [reverse transcription](https://en.wikipedia.org/wiki/Reverse_transcription)
* Some [viruses](https://en.wikipedia.org/wiki/Viruses) (such as [HIV](https://en.wikipedia.org/wiki/HIV), the cause of [AIDS](https://en.wikipedia.org/wiki/AIDS)), have the ability to transcribe RNA into DNA. HIV has an RNA genome that is *reverse transcribed* into DNA. The resulting DNA can be merged with the DNA genome of the host cell. The main enzyme responsible for synthesis of DNA from an RNA template is called [reverse transcriptase](https://en.wikipedia.org/wiki/Reverse_transcriptase).
* In the case of HIV, reverse transcriptase is responsible for synthesizing a [complementary DNA](https://en.wikipedia.org/wiki/Complementary_DNA) strand (cDNA) to the viral RNA genome. The enzyme [ribonuclease H](https://en.wikipedia.org/wiki/Ribonuclease_H) then digests the RNA strand, and reverse transcriptase synthesises a complementary strand of DNA to form a double helix DNA structure ("cDNA"). The cDNA is integrated into the host cell's genome by the enzyme [integrase](https://en.wikipedia.org/wiki/Integrase), which causes the host cell to generate viral proteins that reassemble into new viral particles. In HIV, subsequent to this, the host cell undergoes programmed cell death, or [apoptosis](https://en.wikipedia.org/wiki/Apoptosis) of [T cells](https://en.wikipedia.org/wiki/T_cell).[[25]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-25) However, in other retroviruses, the host cell remains intact as the virus buds out of the cell.
* Some eukaryotic cells contain an enzyme with reverse transcription activity called [telomerase](https://en.wikipedia.org/wiki/Telomerase). Telomerase is a reverse transcriptase that lengthens the ends of linear chromosomes. Telomerase carries an RNA template from which it synthesizes a repeating sequence of DNA, or "junk" DNA. This repeated sequence of DNA is called a [telomere](https://en.wikipedia.org/wiki/Telomere) and can be thought of as a "cap" for a chromosome. It is important because every time a linear chromosome is duplicated, it is shortened. With this "junk" DNA or "cap" at the ends of chromosomes, the shortening eliminates some of the non-essential, repeated sequence rather than the protein-encoding DNA sequence, that is farther away from the chromosome end.
* Telomerase is often activated in cancer cells to enable cancer cells to duplicate their genomes indefinitely without losing important protein-coding DNA sequence. Activation of telomerase could be part of the process that allows cancer cells to become *immortal*. The immortalizing factor of cancer via telomere lengthening due to telomerase has been proven to occur in 90% of all carcinogenic tumours [*in vivo*](https://en.wikipedia.org/wiki/In_vivo) with the remaining 10% using an alternative telomere maintenance route called ALT or Alternative Lengthening of Telomeres.[[26]](https://en.wikipedia.org/wiki/Transcription_(biology)#cite_note-pmid20351727-26)

## 