BIO 375: Genetics and Molecular Biology

K. Mark Dewall

Table of Contents

BYU-I Books

URL: https://books.byui.edu/genetics_and_molecul

Dewall, M. (n.d.). *BIO 375: Genetics and Molecular Biology*. BYU-I Books. https://books.byui.edu/genetics_and_molecul

K. Mark Dewall

Biology Department, Brigham Young University-Idaho

1 - Chromosome Structure

The first reading assignment this semester examines the general features of the chromosomes found within viruses, prokaryotes (bacteria), and eukaryotes (fungi, protozoa, algae, plants, and animals).

Viral Chromosomes

The genetic material (**genome**) of viruses can be composed of either RNA or DNA; however, single virus type never has both DNA and RNA in the same virus particle.

The genomes of viruses can be in several forms:

- **Double-stranded DNA (dsDNA)**. A dsDNA genome contains two individual DNA strands held together by hydrogen bonds. Even though you will not be required to know examples of viruses with dsDNA genomes, several human pathogens have dsDNA genomes, including the smallpox virus and the herpes viruses.
- **Single-stranded DNA (ssDNA)**. The genomes of many viruses are composed of a single DNA strand. Parvovirus, which infects dogs and cats, has a single-stranded DNA genome.
- **Double-stranded RNA (dsRNA)**. A dsRNA genome contains two individual RNA strands held together by hydrogen bonds. Rotavirus, which causes severe diarrhea in humans, has a double-stranded RNA genome.
- **Single-stranded RNA (ssRNA)**. The genomes of many viruses are composed of a single RNA strand. Many disease-causing viruses, such as poliovirus, influenza virus, SARS-CoV-2 (causes COVID-19), and the human immunodeficiency virus (HIV) contain single-stranded RNA genomes.

The genomes of viruses can also be **circular** or **linear**. One way to determine if a viral genome is circular or linear is to isolate the viral genome and treat the genome with **nucleases**, enzymes that digest (cut) DNA or RNA. **Exonucleases** digest nucleic acids into nucleotides only if there is a free end; **endonucleases** cut DNA or RNA in the middle of a nucleic acid molecule. As a result, circular genomes are sensitive to endonucleases, while linear genomes are sensitive to both exonucleases and endonucleases.

The virus genome can be contained within one continuous nucleic acid molecule, or the viral genome can be divided into **segments**. The genome of the influenza virus, for example, contains eight linear ssRNA segments.

When the genome of a virus is located within a virus particle, the genome is inert, meaning that the genome is not copied and viral genes are not transcribed. A virus genome is copied and viral genes are transcribed only during the infection of a host cell.

Viral genomes can range from a few thousand base pairs to 250,000 base pairs in length. For comparison, the genome of the bacterium *E. coli* is 4 million base pairs in length, while the haploid human genome is 3 billion base pairs in length.

Key Questions

- What are the four major criteria used for classifying viral genomes?
- How can a scientist determine if a viral genome is linear or circular?

Bacterial Chromosomes

The genome within a bacterial cell is typically composed of a single **chromosome**. Bacteria are prokaryotic, and since prokaryotes do not contain nuclei, the bacterial chromosome is not contained within a nuclear membrane. Instead the bacterial chromosome is found in a region of the bacterial cytoplasm called the **nucleoid** (see **Figure 1.1**).

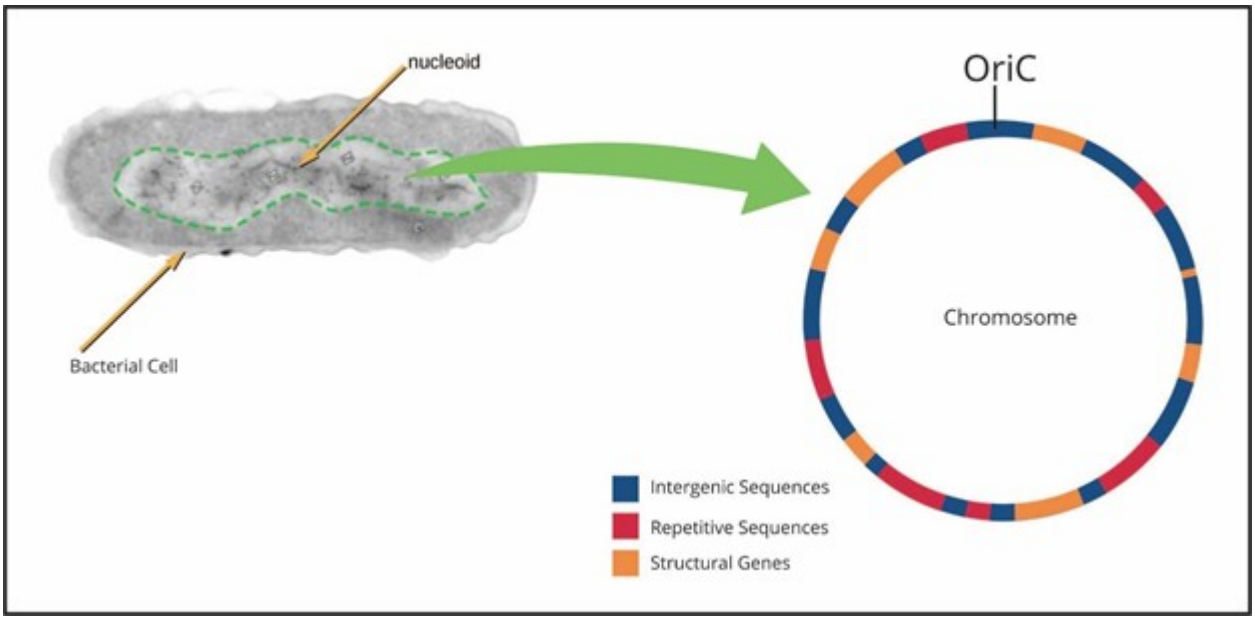


Figure 1.1 ***Bacterial Chromosome Structure*** *--- Prokaryote cell image adapted from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction) --- Image created by SL*

A bacterial chromosome has the following features:

- The bacterial chromosome is usually a single, circular double-stranded DNA molecule.
- The bacterial chromosome is usually 2 - 5 million base pairs (bp) in length.
- The bacterial chromosome contains 1,000 - 3,000 **structural genes**. These structural genes are transcribed and translated to make protein products.
- The bacterial chromosome has a single **origin of replication**. The origin of replication serves as the binding site for proteins involved in initiating DNA replication. The origin of replication in the bacterium *E. coli* is called ***oriC***.
- The bacterial chromosome includes **intergenic DNA sequences**. Intergenic sequences are located between structural genes and are not transcribed. Intergenic sequences serve as the binding sites for proteins that function to activate or deactivate structural genes.
- The bacterial chromosome contains **repetitive DNA sequences**. Repetitive sequences are repeats of a particular base pair sequence, are often found within the intergenic DNA sequences, and are involved in compacting the chromosome to fit into the nucleoid region of the bacterial cell.

Eukaryotic Chromosomes

The genome within a eukaryotic cell is subdivided into multiple chromosomes. Each eukaryotic chromosome is a single, linear double-stranded DNA molecule that is approximately 10–100 million base pairs (bp) in length (see **Figure 1.2**).

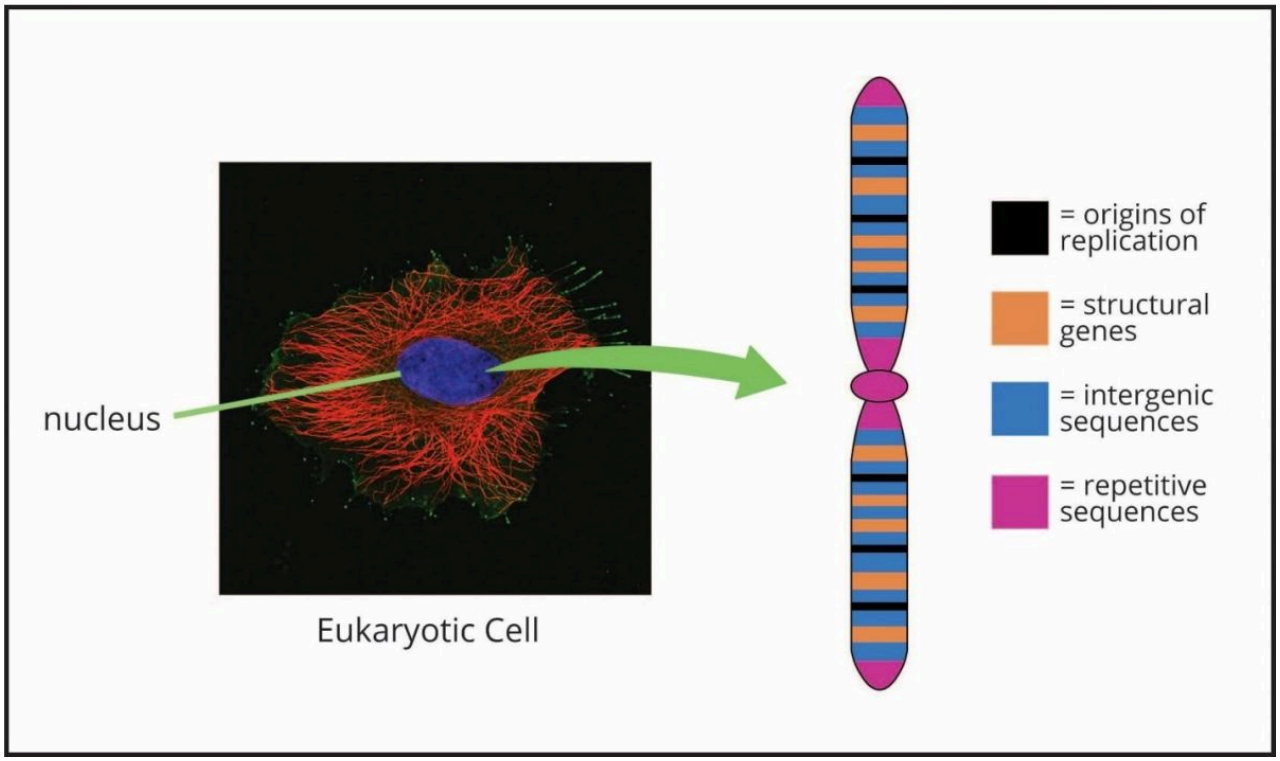


Figure 1.2 ***Eukaryotic Chromosome Structure --- Image created by SL***

A eukaryotic chromosome has several important features:

- **Origins of replication**. Eukaryotic chromosomes contain many origins of replication, spaced at approximately 100,000 base pair (bp) intervals along the chromosome. In the yeast *Saccharomyces cerevisiae*, each origin is called an **ARS element.**
- **Centromeres**. Each eukaryotic chromosome has a single centromere. Centromeres play a critical role in chromosome separation into daughter cells during mitosis and meiosis. A protein structure called the **kinetochore** covers the centromere DNA sequence. The kinetochore functions to link the centromere DNA to the microtubule spindle of the dividing cell, ensuring proper chromosome movement during mitosis and meiosis.
- **Telomeres**. Telomeres are the ends of eukaryotic chromosomes. Telomeres function to prevent chromosomes from sticking together (i.e., prevent translocations). Telomeres also protect the ends of chromosomes from exonucleases and prevent chromosome shortening during DNA replication.
- **Structural genes**. Several hundred to thousands of structural genes are found within a typical eukaryotic chromosome. Recall that structural genes encode protein products. Eukaryotic structural genes contain two types of DNA sequences: **exons** and **introns**. The exon sequences encode the amino acids within the protein product, while the intron sequences between exons do not code for the protein product.
- **Intergenic DNA sequences.** Intergenic sequences are located between structural genes and are not typically transcribed. Intergenic sequences include DNA sequences that serve as the binding sites for the proteins that function to activate or deactivate genes.
- **Repetitive DNA sequences**. Repetitive DNA sequences are repeats of the same DNA sequence and comprise approximately 60% of the human genome. Most of the repetitive DNA sequences do not encode protein products. Repetitive DNA sequences will be discussed in more detail below.
- **Heterochromatin**. Heterochromatin refers to regions along a chromosome that contain highly condensed DNA. These heterochromatin regions either lack structural genes altogether or contain structural genes that are not actively transcribed. The centromere and telomere regions of chromosomes are composed of heterochromatin.
- **Euchromatin**. Euchromatin refers to the loosely condensed regions along the chromosome. Many structural genes are located within euchromatin.

Key Questions

- How are prokaryotic and eukaryotic chromosomes similar?
- How are prokaryotic and eukaryotic chromosomes different?
- What is meant by the term *structural gene*?
- What is the difference between an exon and an intron?
- What are the functions of centromeres and telomeres?
- What is the difference between heterochromatin and euchromatin?

Repetitive Sequences in Eukaryotes

Some DNA sequences found within eukaryotic chromosomes are **unique DNA sequences.** Keep in mind that most eukaryotes are diploid, having two copies of each chromosome (i.e., a homologous chromosome pair). As a result, eukaryotes typically have two copies of each unique DNA sequence; one copy of the gene on each chromosome within a homologous chromosome pair. Most structural genes are examples of unique DNA sequences.

Eukaryotic genomes also contain **repetitive DNA sequences**. These repetitive DNA sequences include **moderately repetitive DNA sequences** and **highly repetitive DNA sequences.** Moderately repetitive sequences are present in a few hundred to a few thousand copies per genome. Highly repetitive sequences are present in tens of thousands to millions of copies per genome.

DNA Reassociation Experiments

How do we know that eukaryotic genomes have unique, moderately repetitive, and highly repetitive DNA sequences? Before scientists were able to determine the base pair sequence of a DNA molecule, **DNA reassociation experiments** were done to determine the overall composition of the genome, focusing on repetitive DNA sequences. In a typical DNA reassociation experiment, entire chromosomes are isolated and are mechanically sheared into fragments. The chromosome fragments are then denatured into single strands by increasing the temperature of the reaction. The reaction mixture is then cooled. As the reaction cools, single-stranded DNA molecules attempt to find each other and form hydrogen bonds to create double-stranded DNA molecules; different DNA fragments do so at different rates (see **Figure 1.3**). Think of it this way, a single-stranded DNA molecule will move around looking for its complement to reattach to base-for-base to form a double-stranded molecule. For highly or moderately repetitive DNA sequences, there are many single strands in the reaction with a complementary DNA sequence to choose from. As a result, highly and moderately repetitive sequences will find each other more rapidly than unique DNA sequences. The DNA reassociation experiment measures the amount of time it takes for single-stranded DNA to form double strands. DNA reassociation experiments showed that there are three populations of DNA: the DNA sequences that reassociated most rapidly were called highly repetitive, moderately repetitive DNA sequences reassociated next, and finally, unique DNA sequences had the slowest rate of reassociation.

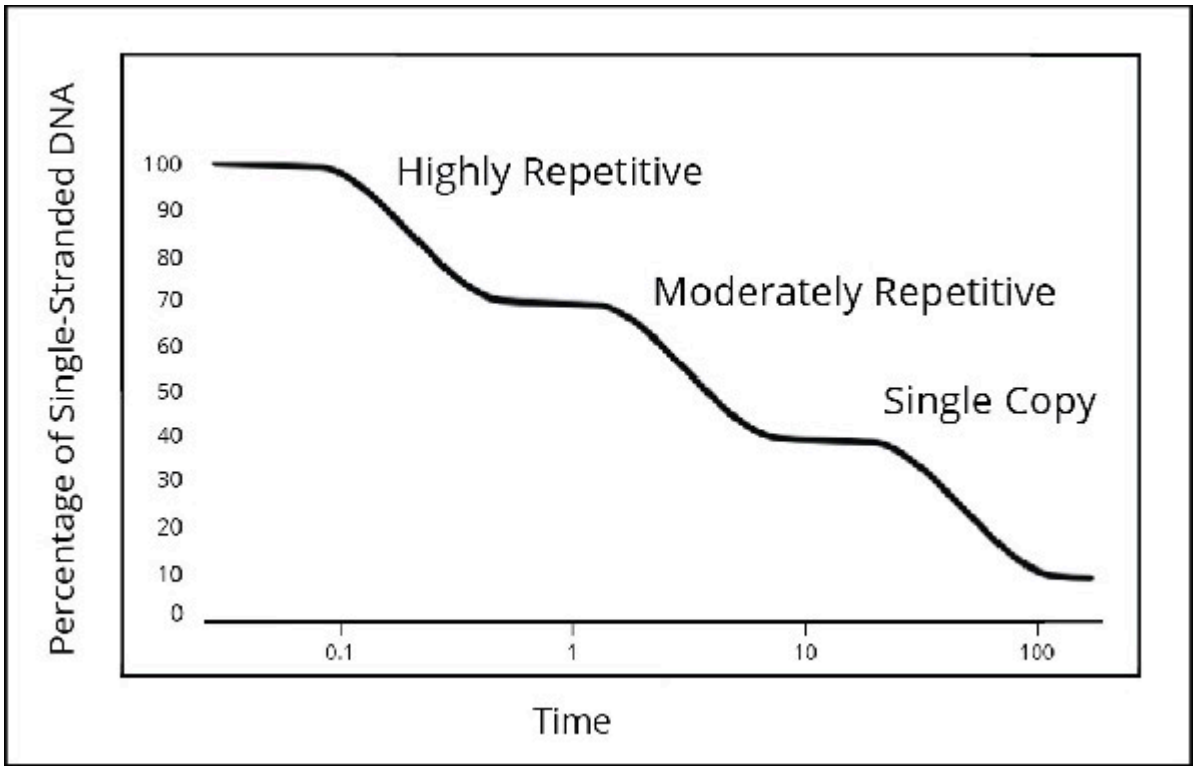


Figure 1.3 ***DNA Reassociation Experiment*** *--- Image created by SL*

Key Questions

- Describe how highly repetitive, moderately repetitive, and unique DNA sequences behave in a DNA reassociation experiment.

Moderately Repetitive Sequences

Moderately repetitive DNA sequences include some genes that produce products. For example, the genes that produce the **ribosomal RNA (rRNA)** components of ribosomes (see Part 11) and the genes that make **histone** proteins (see Part 2) are considered moderately repetitive DNA sequences.

Moderately repetitive DNA sequences also include sequences of unknown function. A good example of this type of moderately repetitive sequence is the **variable number tandem repeat (VNTR)** sequences. VNTRs are typically 15 to 100 base pairs long, are often located between genes, and are present in multiple copies repeated along the length of the chromosome. The number of VNTR repeats on each chromosome is unique to each individual. As a result, this variation in VNTR repeats is the basis of the forensics technique **DNA fingerprinting** (see **Figure 1.4**).

The **telomere repeat** sequences (see **figure 1.6**) are also moderately repetitive DNA sequences.

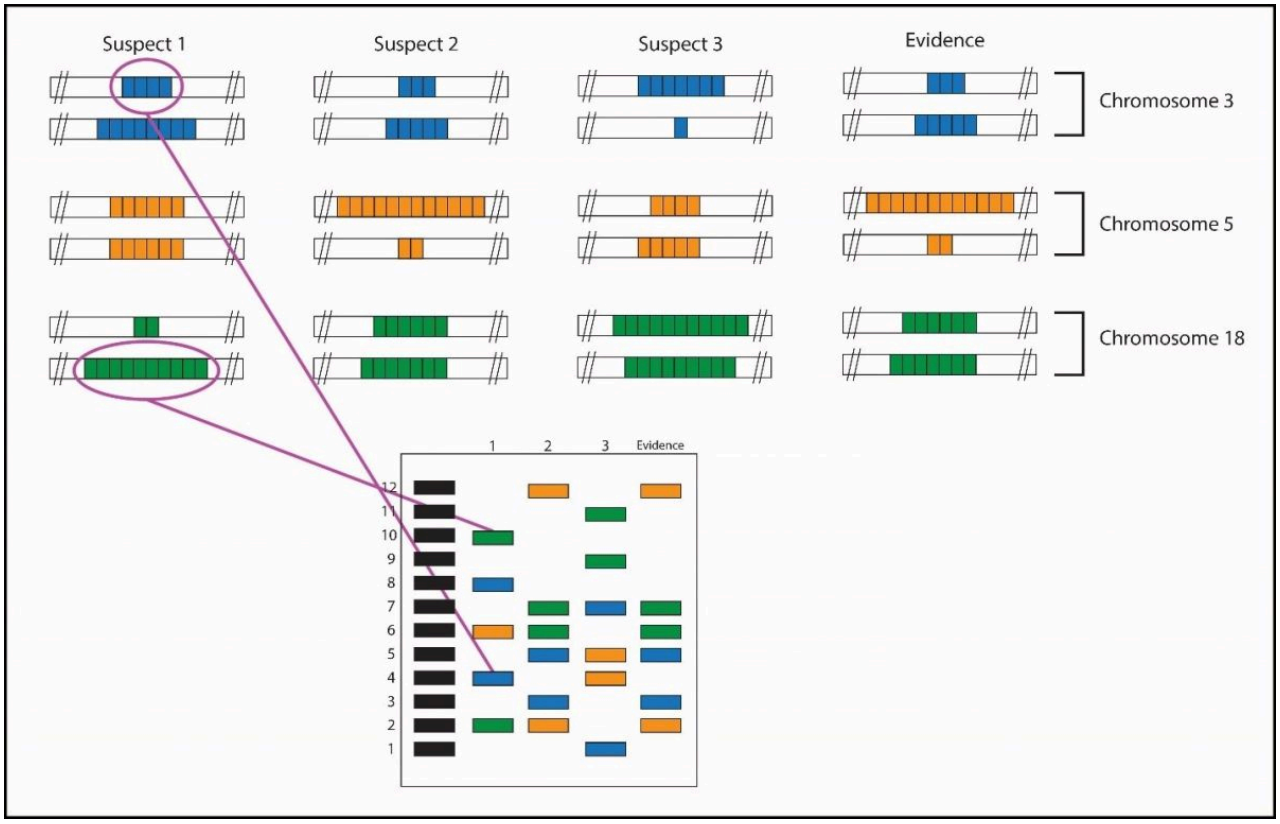


Figure 1.4 ***VNTRs can be used in forensics.*** *The VNTR repeats of three suspect individuals and a sample left at a crime scene (evidence) are shown. The number of boxes on each chromosome indicate the number of VNTR repeats. Recall that each person has two copies of each chromosome (i.e., a homologous chromosome pair). The bottom of the image represents the agarose gel electrophoresis technique (see Part 8) that separates DNA molecules by size; VNTRs with fewer repeats migrate farther through the gel (i.e., towards the bottom of the gel) than VNTRs with more repeats. The expected migration pattern of DNA molecules with 1-12 VNTR repeats is indicated on the left side of the gel in black. The results presented above shows that Suspect 2 had VNTR repeats that match the evidence left at a crime scene. --- Image created by SL*

Key Questions

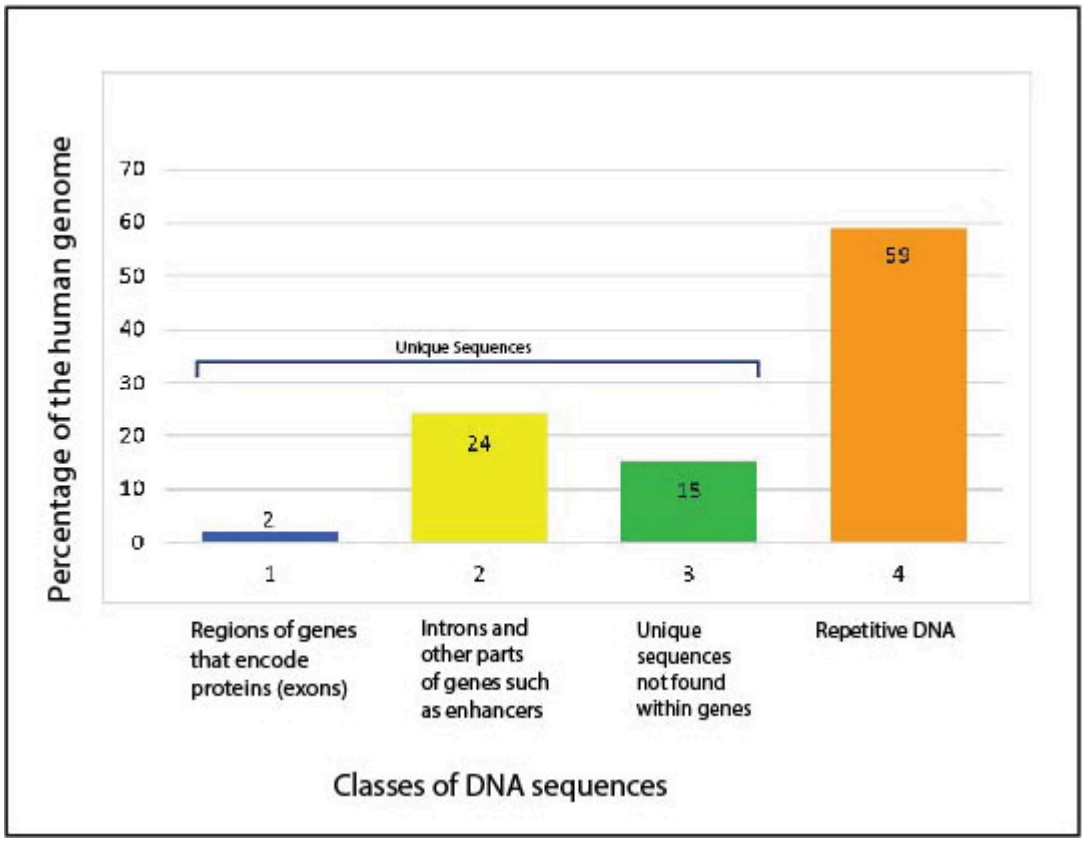
- What are four examples of moderately repetitive DNA sequences?
- Why are VNTRs well suited for forensics?

Highly Repetitive Sequence

The centromere region (**CEN** region) of the chromosome contains highly repetitive DNA sequences. In humans, the CEN region is approximately 10^6 base pairs (bp) long, consisting of a 170 bp **tandem repeat** (i.e., copies of the same 170 bp DNA sequence repeated many times in a row).

The ***Alu*** family of DNA sequences in humans is another example of a highly repetitive sequence. An individual *Alu* sequence within the human genome is only 300 bp long; however, there are so many copies of this *Alu* sequence scattered throughout the human genome that approximately 10% of the human genome is thought to be composed strictly of *Alu* sequences (see **Figure 1.5**). To put this into perspective, only 2% of the human genome is composed of structural genes that produce protein products. Some of these *Alu* sequences are particularly interesting because they have the potential to move from one location in the genome to another. DNA sequences that can move within the genome are called **transposable elements**.

Finally, the **heterochromatin** regions of a chromosome often contain highly repetitive DNA sequences.



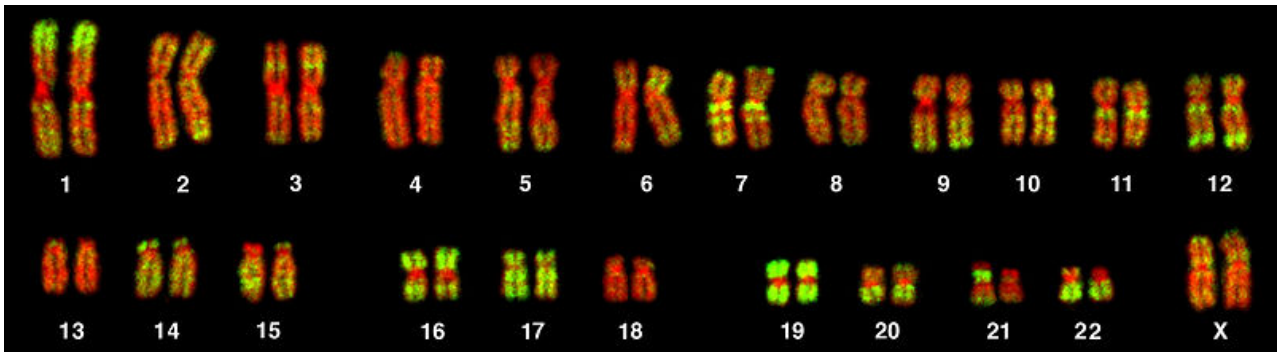


Figure 1.5 ***Top) Classes of DNA Sequence in the Human Genome.*** *Repetitive DNA sequences make up approximately 60% of the human genome, while the exon regions within structural genes compose only 2% of the genome. --- Image created by SL* ***Bottom) Alu Elements.*** *The Alu sequence elements are labeled with a fluorescent green tag. Notice that Alu sequences are found on all 46 chromosomes. --- Chromosomes Alu Fish by Andreas Bolzer is licensed under CC BY 2.5*

Key Questions

- What are three examples of highly repetitive DNA sequences?
- What makes transposable elements unique?

Telomeres

The telomeres of eukaryotic chromosomes have the following features:

- **Telomeres contain tandem repeat DNA sequences.** The tandem repeat DNA sequences within telomeres are 6–8 base pairs (bp) long; each tandem repeat contains multiple G and T nucleotides. For example, the telomere repeat sequence in humans is 5'-TTAGGG-3'. Depending on the eukaryotic species, each telomere may contain several hundred to several thousand tandem repeats of the same telomere DNA sequence (see **Figure 1.6**). Thus, the telomere repeat sequences are moderately repetitive.
- **Telomeres contain 3' single-stranded overhangs.** The 3' overhang is a single-stranded DNA sequence, containing multiple copies of the telomere repeat. The 3' overhang is typically 12–16 bp in length.

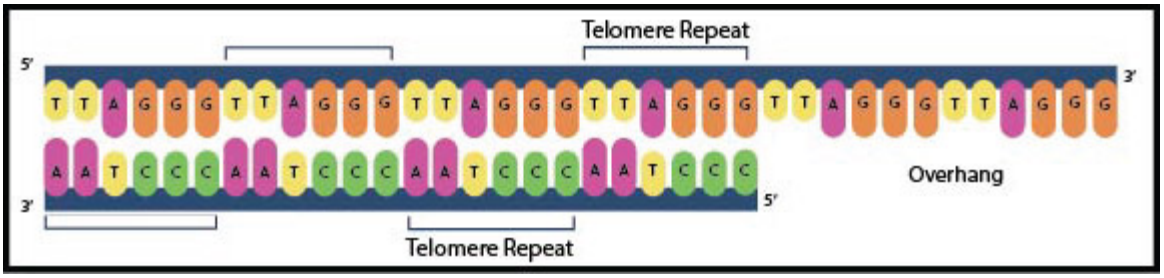


Figure 1.6 ***Telomere Structure ---*** *Image created by SL*

- **Telomere 3' single-stranded overhangs form loops**. The 3' single-stranded overhang within the telomere can turn back on itself to form a **t-loop** (see **Figure 1.7**). Within the t-loop, the 3' single-stranded overhang of the telomere invades another portion of the same chromosome and forms unusual hydrogen bonds between guanine (G) nitrogenous bases. These unusual hydrogen bonds involve four G bases, producing a **G quartet** structure. The t-loop is thought to be the actual structure that protects the eukaryotic chromosome from exonucleases.

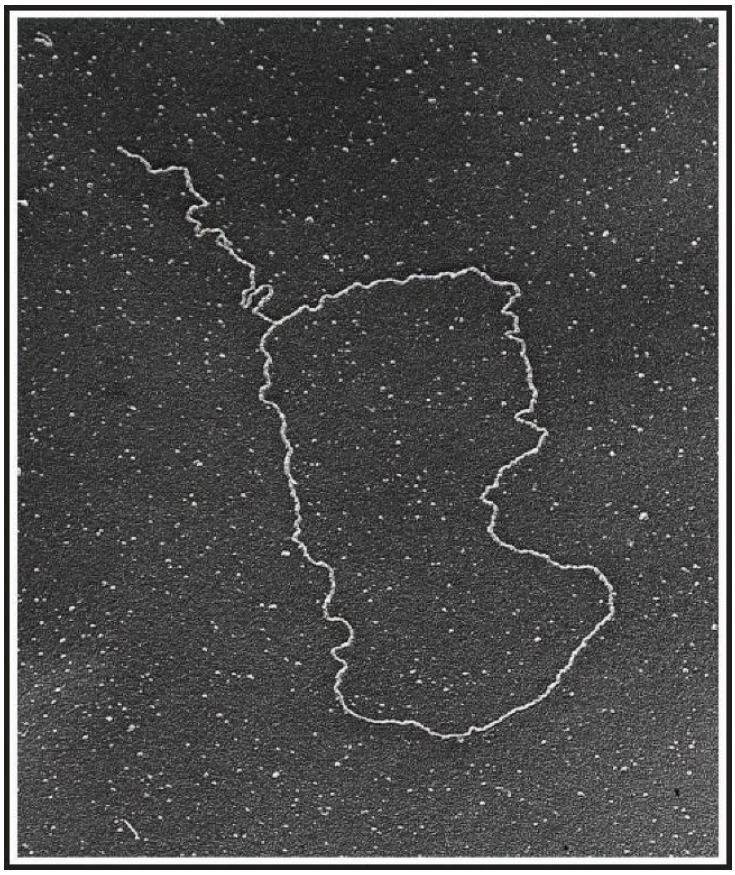


Figure 1.7 ***A telomere loop (t-loop) as visualized in the electron microscope.*** *--- Image courtesy of Dr. Jack Griffith*

Key Questions

- At what locations on a chromosome are you likely to find repetitive DNA sequences?
- What is the function of a t-loop?

Classifying Chromosomes

Eukaryotic chromosomes can be distinguished from each other in the microscope by the location of the centromere (see **Figure 1.8**), the size of the chromosome, and the banding patterns produced along the chromosome after staining with certain chemical dyes. The centromere separates the chromosome into halves (each half is called an arm); the shorter of the two chromosome arms is designated ***p***, while the longer arm is designated ***q***. In terms of centromere location, chromosomes are classified as follows:

- **Metacentric**. The centromere of a metacentric chromosome is located near the center of the chromosome.
- **Submetacentric**. The centromere of a submetacentric chromosome is located slightly off center.
- **Acrocentric**. The centromere of an acrocentric chromosome is located significantly off center. In humans, there are five pairs of acrocentric chromosomes: 13, 14, 15, 21, and 22. These five pairs of chromosomes contain short *p* arms having multiple copies of the same **ribosomal RNA (rRNA)** genes. Having many copies of the ribosomal RNA genes ensures that the cell is able to produce enough ribosome components for translation (see Part 11). The number of rRNA gene copies varies among individuals, but the average is 100 copies per genome. Thus, the rRNA genes are considered moderately repetitive.
- **Telocentric**. The centromere of a telocentric chromosome is located near the end of the chromosome. The human genome does not contain telocentric chromosomes; however, the mouse genome contains telocentric chromosomes.

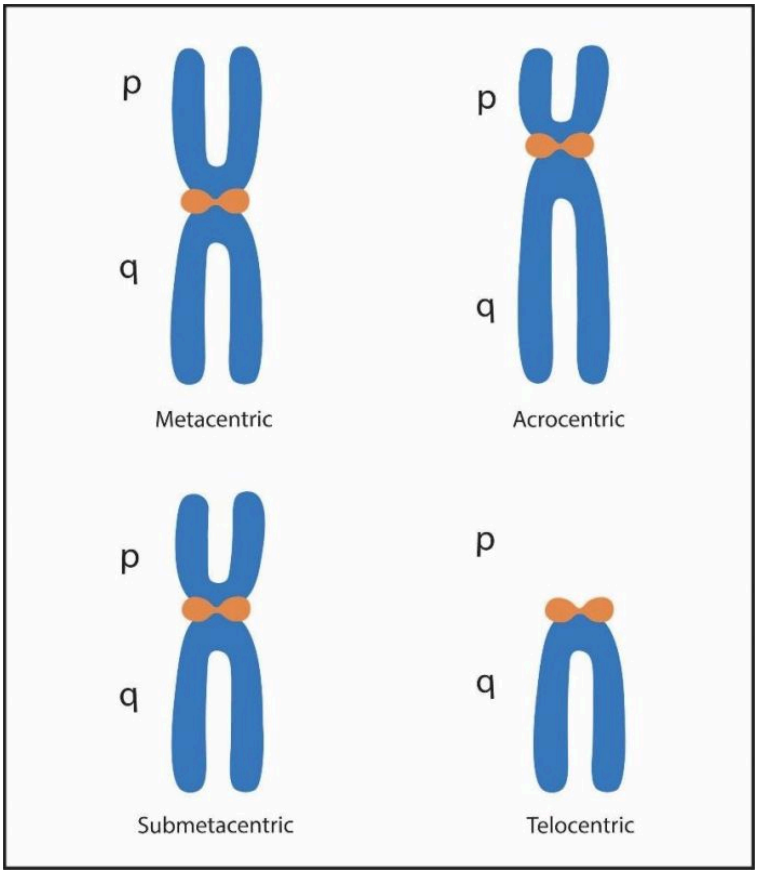


Figure 1.8 Centromere Location. *Note that each chromosome has gone through DNA replication, forming two sister chromatids per chromosome.--- Image created by SL*

Key Questions

- What is meant by the terms metacentric, submetacentric, acrocentric, and telocentric?
- What family of genes is found on the short arms of the five acrocentric human chromosomes?

Human Karyotype and Staining

A **karyotype** is an image of all of the chromosomes within a dividing cell, in which the homologous chromosomes (recall that one chromosome in a homologous pair is inherited from mom; the other chromosome is inherited from dad)

are arranged in pairs (see **Figure 1.9**). The chromosomes are aligned so that their *p* arms are above the centromere and the *q* arms are located below the centromere. Human **autosomes** (non-sex chromosomes) are numbered from the largest to the smallest chromosome, 1 to 22. The **sex chromosomes** are labeled X and Y.

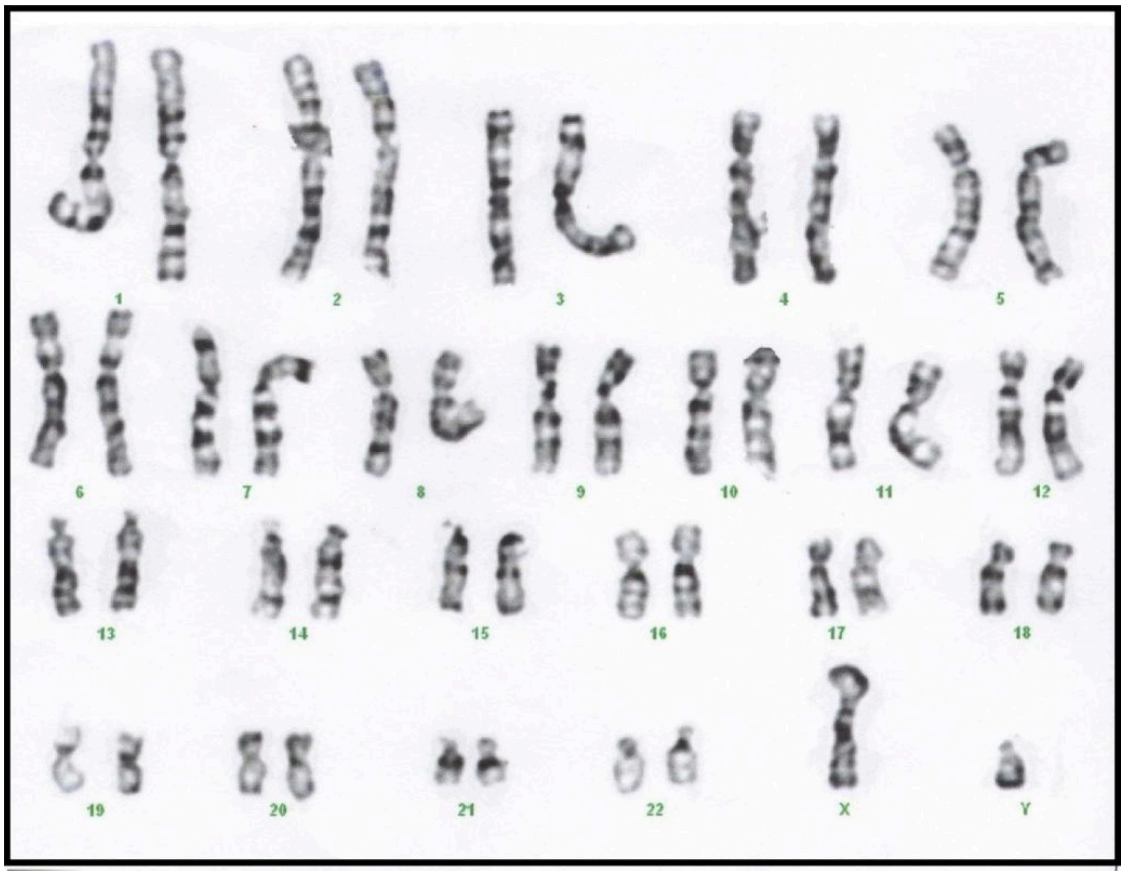


Figure 1.9 ***Karyotype of Human Male.*** *The chromosomes have been stained with the chemical dye Giemsa.* **---** *Karyotype by Can H. is licensed under CC BY 2.0*

Some chromosomes are similar in size and in centromere location. As a result, these chromosomes are difficult to distinguish from each other in the microscope, unless the chromosomes are stained with dyes to produce banding patterns that are unique to each chromosome. A common staining procedure involves the chemical dye **Giemsa**, which produces a unique pattern of light and dark bands (G banding) on each chromosome. Dark bands on the chromosomes represent areas of the DNA that are tightly compacted (heterochromatin); light bands represent areas of the DNA that are loosely compacted (euchromatin).

Key Questions

- What are three ways that scientists can distinguish chromosomes from each other?

Chromosome Nomenclature

A numbering system has been established to describe human chromosomes based on the size, centromere location, and banding pattern. This numbering system assists in determining where chromosome mutations (deletions, duplications, etc.) occur and helps to delineate the exact location of the abnormality. For example, band 22q12 refers to

chromosome 22, the long arm (q), region 1 (closest to the centromere), band 2. If a deletion removes a portion of chromosome 22, the exact location of that deletion can be identified based on this numbering system.

Review Questions

Fill in the blanks:

1. A(n) _____ is an enzyme that digests the ends of linear nucleic acid molecules.
2. A(n) _____ is an enzyme that cuts both linear and circular nucleic acid molecules.
3. Bacterial chromosomes are found in a region of the cytoplasm called the _____.
4. One distinction between prokaryotic and eukaryotic chromosomes is that bacterial chromosomes have _____ origin of replication while eukaryotic chromosomes have _____.
5. Eukaryotes contain highly condensed DNA that lacks genes, these regions called _____ are not generally transcribed and appear as _____ bands on a Giemsa-stained chromosome.
6. _____ are highly repetitive DNA sequences that compose up to 10% of the human genome.
7. The ends of a linear chromosome are called _____ and the portion where spindle proteins attach is called the _____.
8. Two types of genes that are moderately repetitive include _____ and _____ genes.
9. The shorter piece of a chromosome is called the _____ arm while the longer piece is called the _____ arm.
10. _____ genes are found on the *p* arms of chromosomes 13, 14, 15, 21, and 22.

2 - Chromosome Compaction

Part 2 is divided into two sections. The first section discusses how large DNA molecules are compacted to fit inside virus particles, prokaryotic cells, and eukaryotic cells. This process of packaging DNA is called **chromosome compaction**. Chromosome compaction overcomes a significant problem for all organisms. For example, a virus called bacteriophage lambda can package a 17 micrometer (μm) long nucleic acid molecule into a virus particle less than 0.1 μm in diameter (~200 times compaction); the intestinal bacterium *E. coli* can package a chromosome 1.2 millimeters (mm) in length in a cell that is only 0.002 mm long (~1000 times compaction). Human cells package a genome that is 200,000 times longer than the diameter of the nucleus!

The second section within Part 2 explores the process of **X chromosome inactivation (XCI)**. X chromosome inactivation involves compacting one of the two X chromosomes found in the cells of female mammals to produce a condensed **Barr body** structure. This compaction of the X chromosome effectively silences one copy of every X-linked gene in the cell.

A. Chromosome Compaction Strategies

Bacterial Chromosome Compaction

The bacterial chromosome must be compacted approximately 1000 times to fit into the nucleoid region within a bacterial cell. To compact the DNA tenfold, the bacterial cell forms **microdomains** within the chromosome (see **Figure 2.1**). Each microdomain is a loop connected to a centralized core structure composed of DNA binding proteins. The *E. coli* chromosome forms 400 to 500 microdomains, each of which contains approximately 10,000 base pairs (bp) of DNA. Adjacent microdomains are further bundled together to create **macrodomain** regions (not shown in **Figure 2.1**). Each macrodomain contains 80–100 bundled microdomains. Microdomains and macrodomains are formed when the repetitive DNA sequences (see Part 1) within the bacterial chromosome bind to **nucleoid-associated proteins (NAPs)**.

To compact the bacterial chromosome even further, the microdomains are **supercoiled** (i.e., twists are introduced into the microdomains; see **Figure 2.1)**. **Topoisomerases** (see below) are the enzymes that direct the supercoiling of *E. coli* DNA.

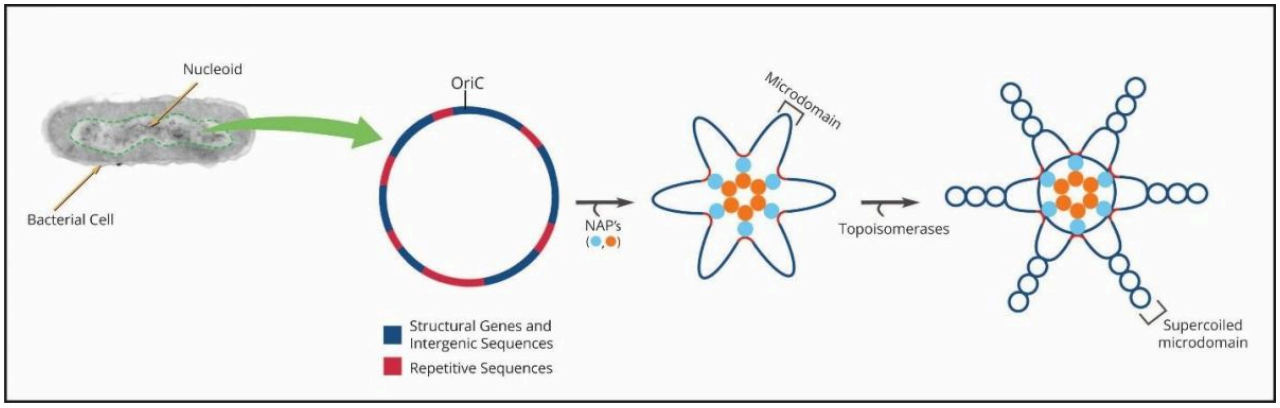


Figure 2.1 ***Bacterial Chromosome Compaction ---*** *Bacterial chromosome compaction involves the formation of microdomains (middle), followed by supercoiling (right). For the sake of simplicity, macrodomains are not included in this diagram. Prokaryote Cell pictured left adapted from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction) --- Image created by SL*

Key Questions

- What are the three levels of chromosome compaction in prokaryotic cells?
- How are microdomains formed?
- What are macrodomains?
- What is meant by supercoiling?

Supercoiling

Suppose a piece of linear double-stranded DNA is connected to two supports, one on each end of the molecule. Also suppose that the bottom support is held firmly in place, while the top support is twisted in the left-handed (counterclockwise) direction. DNA is naturally a right-handed double helix, meaning that the two DNA strands interact by hydrogen bonding to produce a double helix that rotates clockwise. Introducing counterclockwise twists into a right-handed double helix produces **underwinding**.

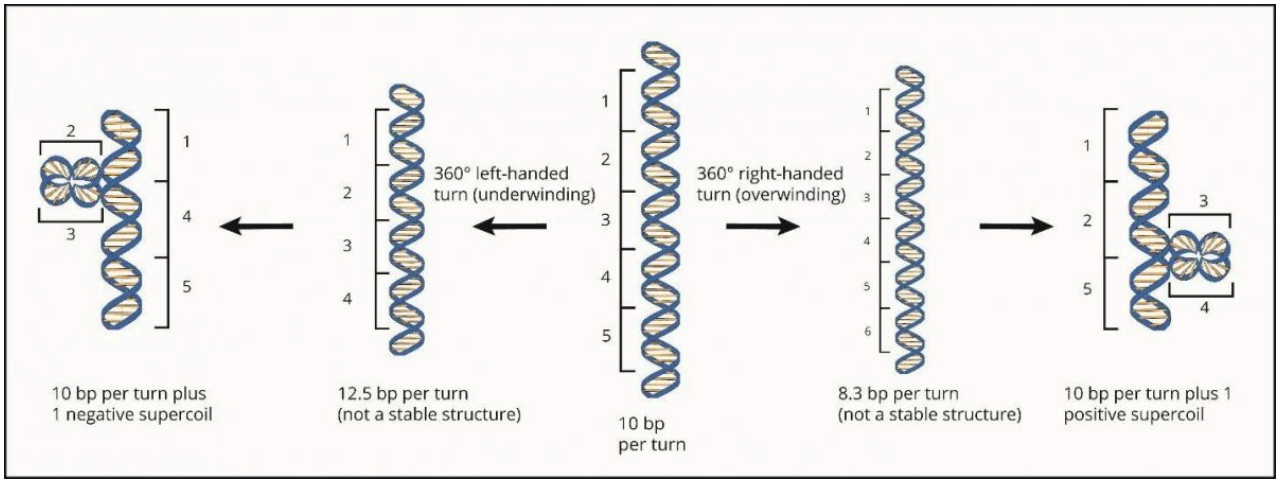


Figure 2.2 ***Negative and Positive Supercoiling –*** *The molecule in the center of the image contains five turns, with each turn containing 10 base pairs (bp). Underwinding by one turn produces an unstable structure with four turns, while overwinding by one turn produces an unstable structure with six turns. In both cases, the DNA double helix is stabilized by supercoiling --- Image created by SL*

Underwinding of the DNA produces fewer turns in the double helix. For example, a linear 50 base pair (bp) DNA molecule with five turns in the double helix (10 bp per turn) would have 12.5 bp per turn if it is underwound by one turn (50 bp/4 turns = 12.5 bp/turn) (see the left-hand side of **Figure 2.2**). This linear form of DNA with 12.5 bp per turn is an unstable structure that does not naturally occur in cells. Instead, the underwound DNA molecule produces a negative supercoil, an attempt by the DNA double helix to stabilize itself. This negative supercoil causes the distance between the ends of the DNA molecule to decrease (i.e. compaction). Negative supercoils are observed within bacterial chromosomes and function to compact the bacterial DNA into the nucleoid.

Overwinding, twisting the DNA double helix in the right-handed direction, increases the number of turns in the double helix. For example, a 50 base pair (bp) DNA molecule with five turns in the double helix (10 bp per turn) would have 8.3 bp per turn if it is overwound by one turn (50 bp/6 turns = 8.3 bp/turn) (see the right-hand side of **Figure 2.2**). This linear form of DNA with 8.3 bp per turn is an unstable structure that does not naturally occur in cells. The DNA molecule attempts to stabilize itself by producing a supercoil. This time, the supercoil is a **positive supercoil** that causes the distance between the ends of the DNA molecule to decrease (i.e. compaction).

A DNA molecule that lacks supercoils, has a single negative supercoil, or has a single positive supercoil can be converted into each other using **topoisomerase** enzymes (see below). These DNA molecules have the same base pair sequence and only differ in the degree of supercoiling. As a result, these three DNA molecules are considered to be **topoisomers** of each other.

Key Questions

- What is meant by positive and negative supercoiling?

Negative Supercoiling

Bacteria prefer negative supercoiling to positive supercoiling. In fact, typical bacterial chromosomes contain approximately one negative supercoil per 400 base pairs (bp) of DNA. Negative supercoiling is preferred because negative supercoiling:

- **Compacts the chromosomal DNA into the nucleoid of the cell**.
- **Promotes DNA strand separation**. Separation of the DNA strands is required for DNA replication prior to cell division and transcription to activate a gene.

Note that positive supercoiling compacts the chromosomal DNA to the same extent as negative supercoiling; however, positive supercoiling is inhibitory to DNA replication and transcription.

Key Questions

- What impact does supercoiling have on chromosome structure and function?
- Why is negative supercoiling preferred to positive supercoiling?

Topoisomerases

Topoisomerases are enzymes that supercoil DNA. There are two general classes of topoisomerases: **topoisomerase I** and **topoisomerase II**. Topoisomerase I enzymes are thought to mainly generate positive supercoils but can generate negative supercoils under certain conditions. Unlike topoisomerase II enzymes (see below), topoisomerase I enzymes are composed of a single protein subunit, do not cleave ATP during supercoiling, and only cut one of the two DNA strands during supercoil formation.

DNA gyrase from the bacterium *E. coli* is the best characterized example of a **topoisomerase II** enzyme**.** DNA gyrase cleaves ATP and uses the released energy to introduce negative supercoils into chromosomes. Further, DNA gyrase is composed of four protein subunits, two **A subunits** and two **B subunits**. Negative supercoils are generated by DNA gyrase using the following mechanism:

1. The A subunits of DNA gyrase bind to the DNA.
2. The A subunits function as an endonuclease to cut both strands of the DNA.
3. The B subunits pass another portion of the DNA molecule through the break using the energy released by cleaving ATP.
4. The DNA break is repaired producing an intact DNA double helix.

DNA gyrase generates two negative supercoils in the bacterial DNA per catalytic cycle; the production of each negative supercoil requires the cleavage of a single ATP molecule. For example, if the original bacterial DNA molecule had no supercoils, the same molecule would have two negative supercoils after DNA gyrase action (two ATP molecules cleaved). In bacteria, topoisomerase I and topoisomerase II compete to determine the overall level of supercoiling within the chromosome. Topoisomerase I and topoisomerase II enzymes also function in eukaryotic cells, as well.

The ability to produce negative supercoils in the DNA is required for bacterial survival. For example, a group of antibiotics called **quinolones** inhibit DNA gyrase. Since DNA gyrase is involved in both negative supercoiling and DNA replication (see Part 6), the bacterial cells are killed. One example quinolone antibiotic, **ciprofloxacin**, is used to treat patients infected with serious bacterial infections, including anthrax and typhoid fever.

Key Questions

- Describe how DNA gyrase introduces negative supercoils.
- What are some differences between the structure and function of topoisomerase I and DNA gyrase?

Eukaryotic Chromosome Compaction

Let us now consider how the human genome, which is over six feet in length, can be packaged into the nucleus within a cell. An important aspect of chromosome compaction in eukaryotes involves the association of the DNA double helix with proteins (both **histone** and **nonhistone proteins**) to form **chromatin**.

The basic structure of chromatin consists of chains of **nucleosomes** that resemble beads on a string (see Figure 2.3). A nucleosome consists of 146 or 147 base pairs (bp) of DNA negatively supercoiled around eight **histone** proteins. There are four types of histone proteins within a nucleosome; each histone protein is present in two copies. These histone proteins are called **H2A, H2B, H3**, and **H4**. Each histone protein consists of a **globular domain** and an extended, flexible region called a **histone tail**. The globular domain allows the individual histone proteins to bind to each other to form the nucleosome core, while the histone tails are enriched in positively charged amino acids, such as arginine and lysine. Recall that the backbone portion of the DNA double helix is negatively charged; thus, the histone tail and DNA backbone bind through electrostatic interactions.

Nucleosomes are connected by a **linker** DNA sequence that is approximately 50 bp long. Histone **H1**, also called the **linker histone**, as well as other nonhistone proteins bind to the linker region DNA. H1 and these nonhistone proteins play a role in further compaction of eukaryotic DNA into 30-nm fibers (see below).

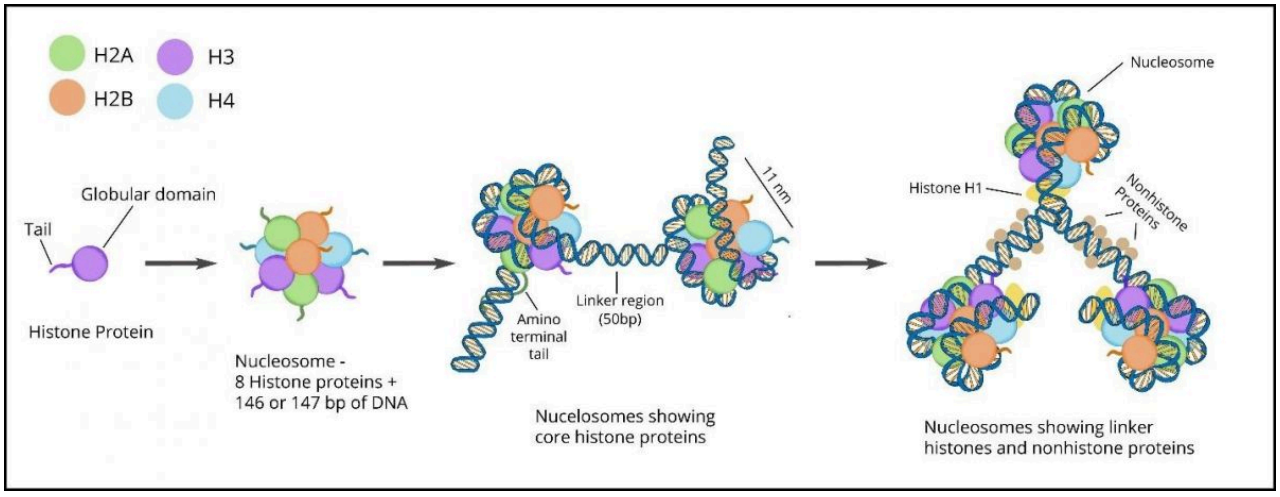


Figure 2.3 Nucleosomes are the first level of chromosome compaction. *Histone proteins contain a globular region and a positively charged histone tail. Eight histone proteins (two copies each of the H2A, H2B, H3, and H4 proteins) form the nucleosome core. The eight histones bind to 146 or 147 base pairs of DNA to form nucleosomes, which resemble beads on a string. The linker histone H1 and nonhistone proteins bind to the 50 base pair linker sequence between nucleosomes.* ***---*** *Image created by SL*

Key Questions

- What is a nucleosome?
- What are the names of the protein components within a nucleosome?
- Explain how the DNA and histone proteins assemble to make a nucleosome.
- What is a linker region?

30-nm Fiber

The second level of chromosome compaction in eukaryotes involves the association of a string of nucleosomes into a fiber that is 30 nanometers (nm) wide (**30-nm fiber**). The formation of the 30-nm fiber depends on the linker histone protein H1 and nonhistone proteins. Two 30-nm fiber structures have been proposed (see **Figure 2.4**):

- **Solenoid**. The solenoid involves tight interactions between nucleosomes to form a compact, symmetrical structure that resembles a cylinder. The solenoid form of the 30-nm fiber is advantageous because it allows a high degree of chromosome compaction; however, structural genes within the solenoid are difficult to activate by transcription.
- **Zigzag**. In the zigzag, nucleosome interaction is minimal and the linker regions are free to bend and twist. As a result, the overall structure of the zigzag 30-nm fiber is irregular. The zigzag form of the 30-nm fiber is advantageous because it promotes the transcription of structural genes, but the degree of chromosome compaction is not as great as with the solenoid form.

Recent evidence suggests that the zigzag form of the 30-nm fiber is found predominantly in cells; however, some scientists believe that the two forms of the 30-nm fiber can convert into each other depending on whether a region of DNA needs a high degree of chromosome compaction or gene activation.

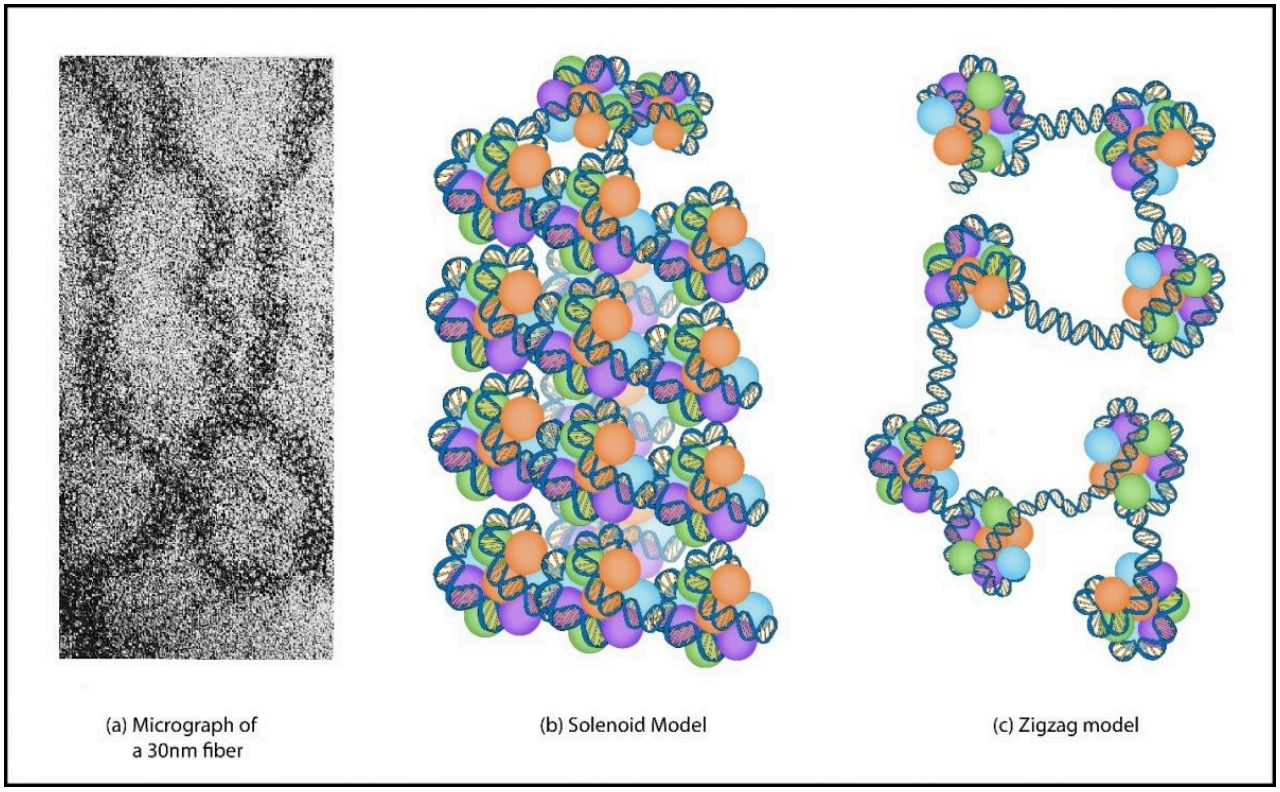


Figure 2.4 ***30-nm Fiber*** *–(a)* Photo courtesy of Dr. Barbara Hamkalo (b) Solenoid model (c) Zigzag model --- Image created by SL

Key Questions

- What are the advantages and disadvantages of the solenoid and zigzag 30-nm fiber models?

Radial Loop Domains

The third level of DNA compaction in eukaryotes involves the interaction of the 30-nm fibers with **nuclear matrix proteins** to form **radial loop domains** (see **Figure 2.5**). The eukaryotic radial loop domain is somewhat similar to the prokaryotic microdomain, with each radial loop consisting of 25,000–200,000 base pairs (bp) of DNA.

The nuclear matrix consists of two parts, the **nuclear lamina** and the **internal nuclear matrix**. The nuclear lamina portion of the nuclear matrix is composed of cytoskeletal proteins and lies adjacent to the inner surface of the nuclear membrane. The internal nuclear matrix, which likely includes hundreds of different protein types, forms a fine meshwork of filaments throughout the interior of the nucleus. DNA sequences called **matrix-attachment regions (MARs)** link the 30-nm fiber to the internal nuclear matrix. Anchoring the 30-nm fiber to the internal nuclear matrix results in the formation of radial loop domains.

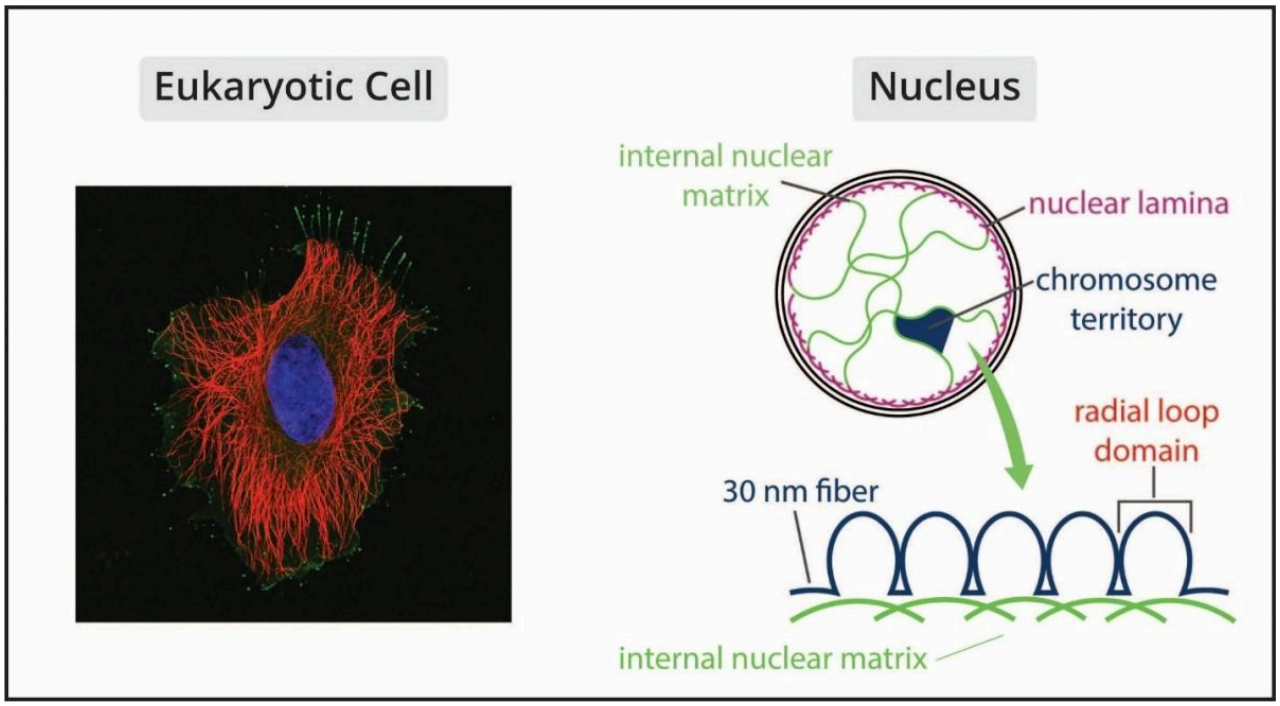


Figure 2.5 ***Radial Loop Domains.*** *Fluorescence micrograph of a eukaryotic cell (left). The DNA is shown in blue, the microtubule cytoskeleton is shown in red, while the actin cytoskeleton is shown in green. The nucleus of a eukaryotic cell (right) contains nuclear lamina and internal nuclear matrix protein fibers. The chromatin 30-nm fibers bind to internal nuclear matrix proteins to form radial loop domains.--- Image created by SL*

Key Questions

- What is a radial loop domain?
- How do the MARs and the internal nuclear matrix contribute to the formation of radial loop domains?

Chromosome Territories

The internal nuclear matrix bound to radial loop domains also functions to localize each chromosome within a unique region of the nucleus called a **chromosome territory**. These chromosome territories within the nucleus can be visualized when the individual chromosomes are stained with uniquely colored fluorescent dyes (see **Figures 2.5** and **2.6**).

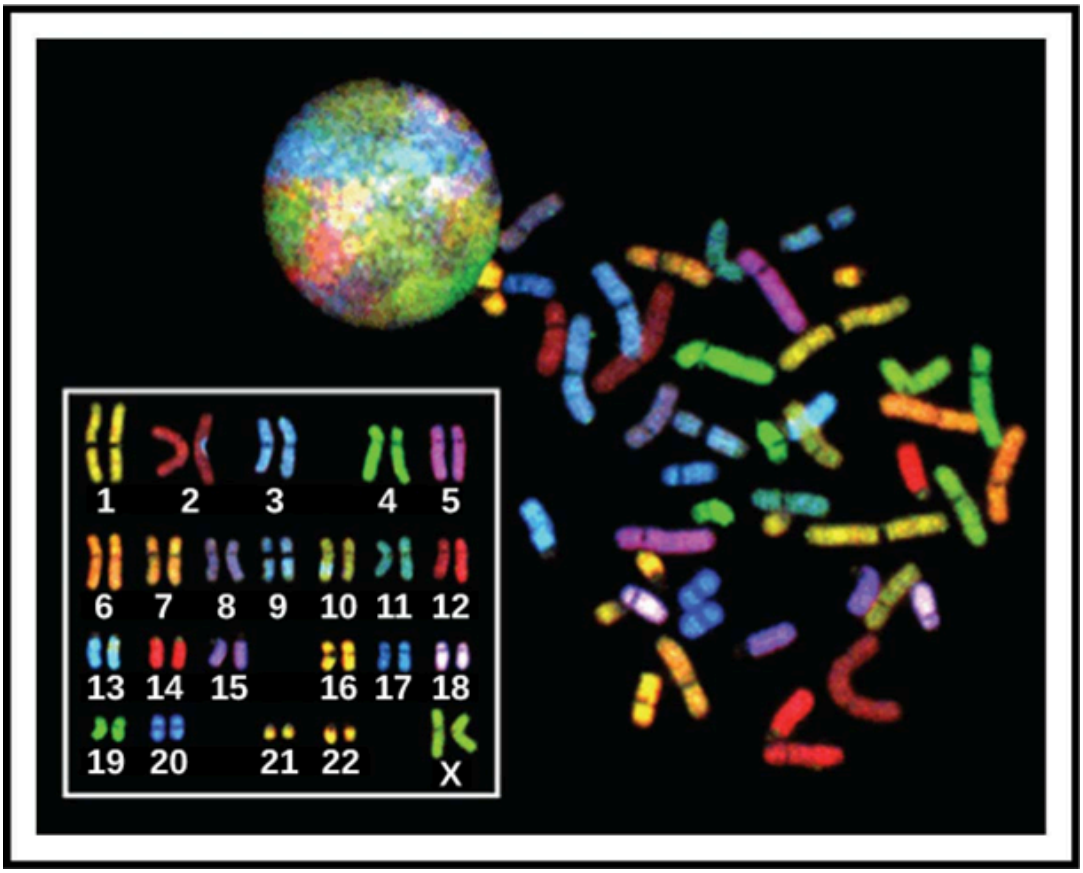


Figure 2.6 ***Chromosome Territories*** *--- Chromosome Territories was used from OpenStax (access for free at* https://openstax.org/books/biology-2e/pages/1-introduction

Key Questions

- What is a chromosome territory?

Heterochromatin and Euchromatin

The radial loop domains can assume two different structural conformations (see **Figure 2.7**):

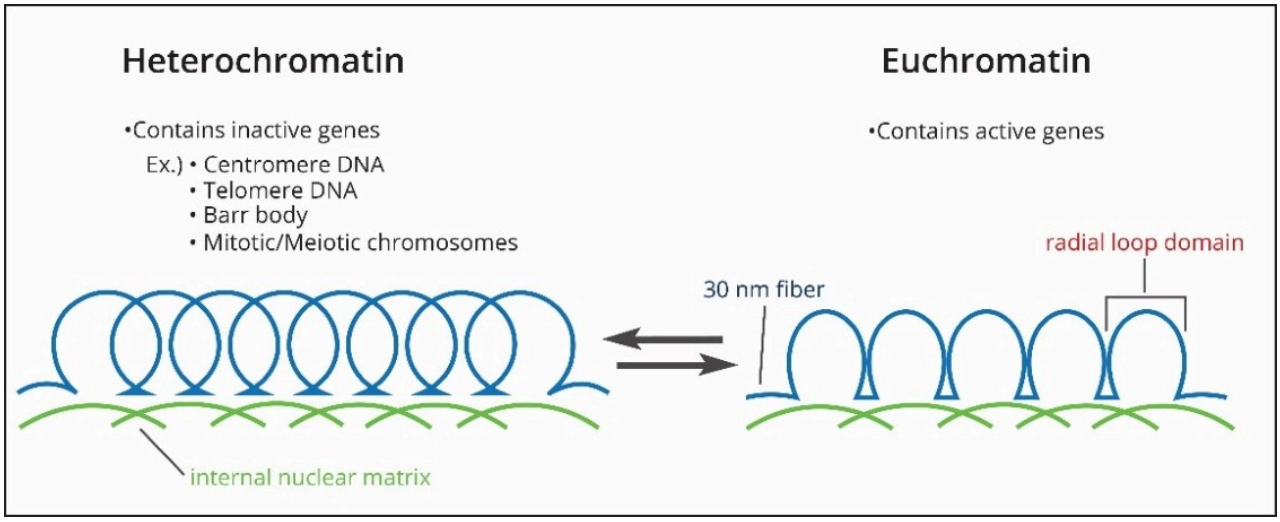


Figure 2.7 ***Heterochromatin and Euchromatin*** *--- Image created by SL*

- **Euchromatin.** Euchromatin is a region of chromatin in which the radial loops are not tightly compacted. Transcibed structural genes are typically found in euchromatin.
- **Heterochromatin.** Heterochromatin is defined as regions of the chromosome where the radial loops are tightly compacted. Most areas of heterochromatin either do not contain structural genes or the structural genes within these regions are not transcribed. There are two forms of heterochromatin:
    - **Constitutive heterochromatin.** The centromere and the telomeres regions of chromosomes always assume the heterochromatin conformation and, are therefore, considered to be constitutive heterochromatin. Constitutive heterochromatin often contains either moderately or highly repetitive DNA sequences (see Part 1).
    - **Facultative heterochromatin.** Facultative heterochromatin is a dynamic structure that can convert between the heterochromatin and euchromatin states. For example, when one of the two X chromosomes in female mammals is chosen to be inactivated during early embryogenesis, the X chromosome is converted from euchromatin to heterochromatin (see below). The heterochromatin form of the X chromosome is the **Barr body**. Further, each cell type in the body contains a unique pattern of facultative heterochromatin, meaning the distribution of facultative heterochromatin in a white blood cell is different from an epithelial cell. These differences in facultative heterochromatin ensure that each cell type transcribes a unique subset of structural genes. We will examine how changes in facultative heterochromatin influences gene activation in Parts 14 and 15.

Key Questions

- What is the difference between euchromatin and heterochromatin?
- How do the two types of heterochromatin differ from one another?

Scaffold

Nucleosomes, 30nm fibers, and radial loop domains are found in the chromatin of interphase cells. In contrast, during mitosis and meiosis, the chromosomes become even more compacted—10,000 times more compact than what is observed during interphase. In fact, the characteristic X-shaped chromosomes observed in mitosis and meiosis are so compact that scientists think that they are composed mainly of facultative heterochromatin. As a result, few structural genes are active during mitosis and meiosis.

To form highly compacted mitotic and meiotic chromosomes, the radial loops interact with a **scaffold** (see **figure 2.8**). The scaffold is a protein structure that ensures the radial loops throughout the chromosome are in the heterochromatin state. The structure of the scaffold is poorly understood; however, scientists believe the scaffold is composed of nonhistone proteins, including condensin (see below), and nuclear matrix proteins. The X-shaped metaphase chromosomes are produced by radial loop domains binding to the scaffold.

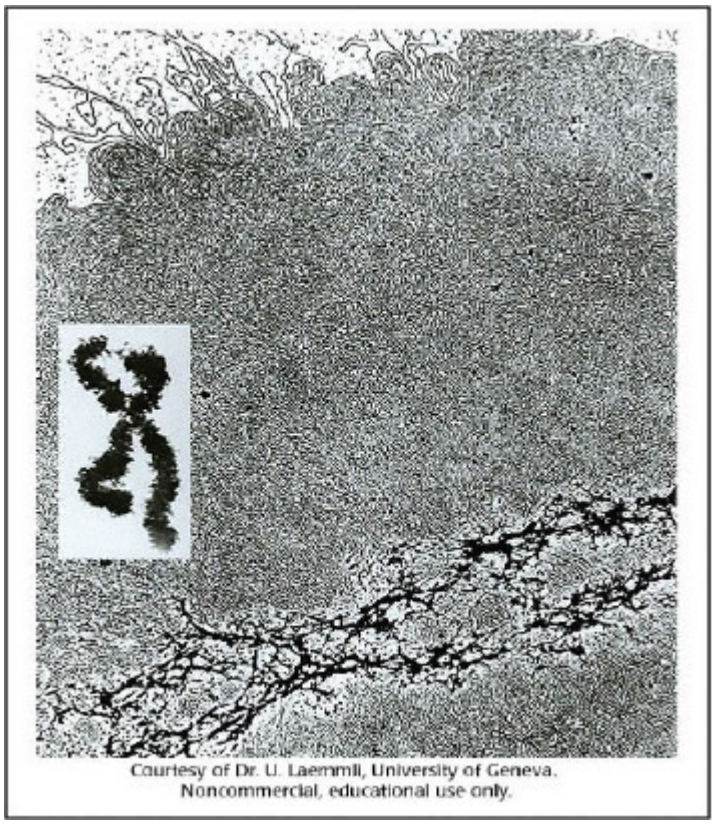


Figure 2.8 ***The Scaffold*** *– A mitotic chromosome was experimentally treated to release the DNA double helix (seen most clearly at the top of the image), while preserving the scaffold structure (darker area near the bottom of the image). --- Photo courtesy of Dr. U. Laemmli*

Key Questions

- What are the four levels of chromosome compaction in eukaryotes?
- Which levels are compaction are observed during interphase?
- Which levels of compaction are observed during mitosis and meiosis?

Condensin

The **condensin** protein plays an important role in the formation of mitotic and meiotic chromosomes. When a eukaryotic cell is in interphase, condensin is found in the cytoplasm of the cell. However, during mitosis and meiosis, the nuclear envelope breaks down and condensin can then bind to the chromosomes. Condensin is thought to link radial loop domains together and hold them in place, forming the dense heterochromatin observed in mitosis and meiosis. Condensin is a member of a group of proteins called **structural maintenance of chromosome (SMC)** proteins. Other members of the SMC family include the NAPs that play a role in bacterial chromosome compaction (see above). All SMC proteins cleave ATP and use the released energy to promote changes in chromatin structure.

Key Questions

- How does condensin contribute to chromosome compaction?

B. X Chromosome Inactivation (XCI)

Dosage Compensation

Because female animals have two X chromosomes and males have a single X chromosome, females can potentially produce twice as much of the protein products from X-linked structural genes as their male counterparts. However, we know that the level of X-linked protein production is similar males and females. This **dosage compensation** between males and females can be accomplished in several different ways. In mammals, one of the two X chromosomes is inactivated in females. For example, placental mammals randomly inactivate either the paternally-inherited or the maternally-inherited X chromosome in somatic cells. On the other hand, female marsupials inactivate the X chromosome they inherited from their father. Fruit flies conduct dosage compensation by increasing X-linked gene expression in males twofold. Finally, female nematode worms reduce gene expression on each X chromosome by 50% to accomplish dosage compensation.

Key Questions

- Why is dosage compensation important?
- How do humans accomplish dosage compensation?

Evidence for X Chromosome Inactivation (XCI)

In mammals, one of the X chromosomes experiences **X chromosome inactivation (XCI)**. XCI was first suggested by two lines of experimental evidence:

- **The studies of Murray Barr and Ewart Bertram**. Barr and Bertram found that somatic cells (i.e., non-gamete cells) from female cats contained a nuclear structure (**Barr body**), not found in the males (see **Figure 2.9**). The Barr body is a highly condensed, inactive X chromosome; few structural genes present within a Barr body are expressed.
- **The studies of Mary Lyon**. Female tortoiseshell (black and orange) and calico (black, orange, and white) cats have a distinctive coat pattern, containing patches of orange and black fur (**mosaics**). The mosaic phenotype occurs in female cats that are heterozygous for X-linked coat color alleles (X^bX^o). The X^b allele is expressed in black patches, while the X^o allele is silenced. In contrast, the X^o allele is expressed in orange patches, while the X^b allele is silenced. Mary Lyon suggested that the mosaic tortoiseshell and calico patterns are due to XCI.

Key Questions

- How did the experiments by Barr, Bertram, and Lyon demonstrate that XCI occurs?

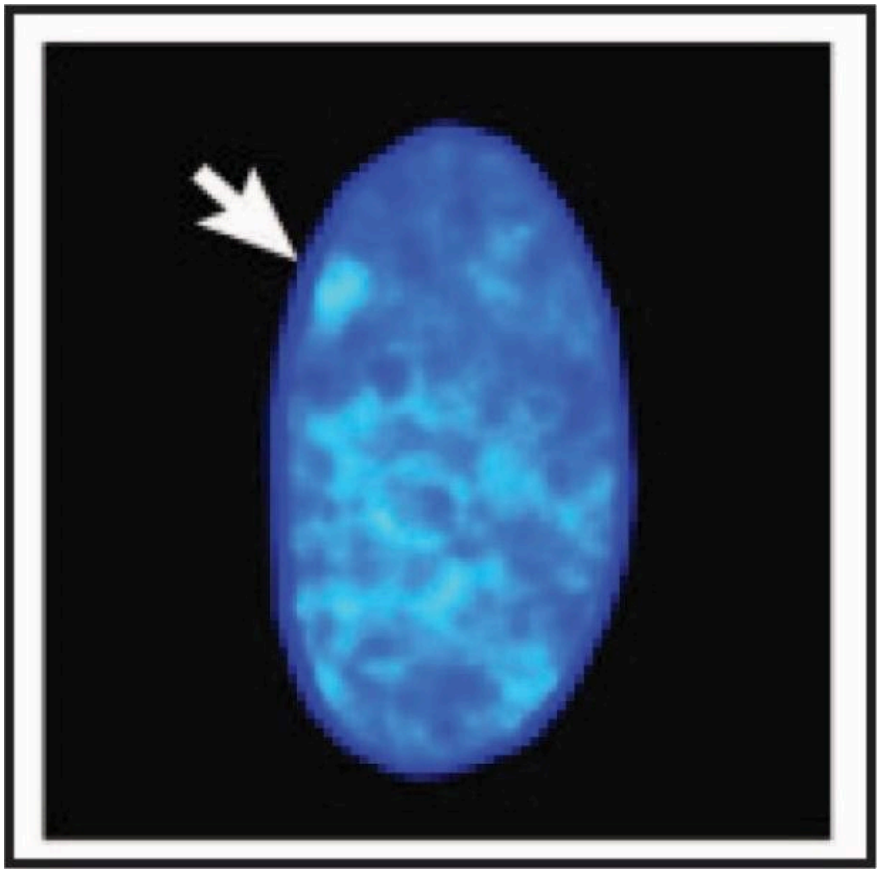


Figure 2.9 ***Evidence for X Chromosome Inactivation –*** *A) The arrow in the image indicates a Barr Body within the nucleus of a cell. --- Barr Body BMC by Stanley Gartler is used under CC BY 2.0) B) A calico cat displaying the mosaic orange and black phenotype.--- Tortoiseshell Cat by James Petts is used under CC BY-SA 2.0*

The Lyon Hypothesis

The **Lyon Hypothesis**, first proposed by Mary Lyon, provided a deeper understanding of X chromosome inactivation (XCI). In mice, fur color is controlled by two X-linked alleles: The X^B allele produces black fur color, while the X^b produces white fur. Consider a heterozygous female mouse (X^BX^b), with a mosaic phenotype, similar to the tortoiseshell and calico cat coat patterns discussed above. The Lyon hypothesis states that during embryonic development in mice, both of the X chromosomes are active in each embryonic cell. However, one of the X chromosomes in each embryonic cell is soon inactivated and becomes a Barr body. This inactivation process is random in each embryonic cell, one cell may inactivate X^B; a neighboring cell may inactivate X^b. The embryonic cell containing an active X^b (silenced X^B) divides to produce a white fur patch (see **Figure 2.10**). The embryoinc cell with an active X^B (silenced X^b) divides to produce a black patch. Collectively, these events produce the mosaic phenotype of the heterozygous mouse.

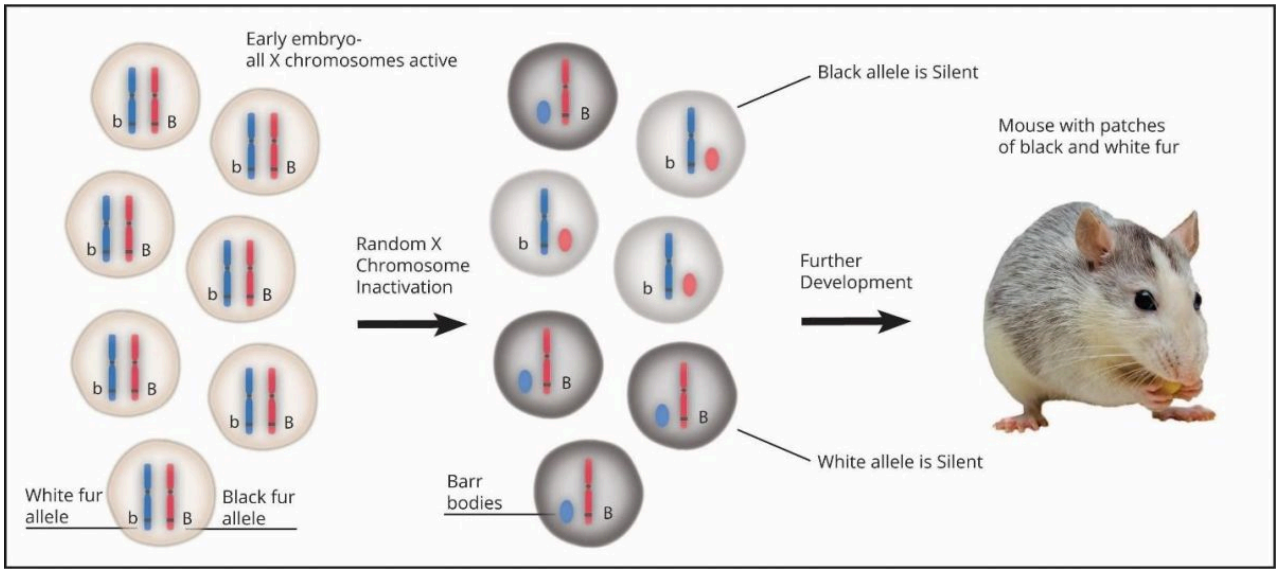


Figure 2.10 ***The Lyon Hypothesis*** *--- Image created by SL (mouse image by Pexels and is under CC0)*

One consequence of the Lyon hypothesis is that all female mammals (including humans) are thought to be mosaics. That means that in some areas of the body one X-linked allele is expressed; other areas of the body express the other X-linked allele. **Anhidrotic ectodermal dysplasia** provides evidence for human mosaicism. Anhidrotic ectodermal dysplasia is a human genetic disease caused by an X-linked recessive mutation. If a male possesses the recessive disease allele, he displays a variety of defects, including the absence of sweat glands. Heterozygous females are mosaics in which some areas of the body have sweat glands; other areas lack sweat glands.

Key Questions

- What is the Lyon hypothesis and how does it explain the mosaic coat phenotypes seen in mice and cats?

X-inactivation Center (*Xic*)

In females with two X chromosomes, one X chromosome is inactivated to produce a Barr body. In **Turner syndrome** females with one X chromosome, Barr bodies are not observed. In **Triplo-X syndrome** females with three X chromosomes, two Barr bodies are formed; and in **Klinefelter syndrome** males with two X chromosomes and a Y chromosome, a single Barr body is formed. Thus, it appears that mammalian cells can "count" the number of X chromosomes in a particular cell and ensure that only one X chromosome remains active.

X chromosome inactivation is controlled by a region within the X chromosome, near the centromere, called the **X-inactivation center (*Xic*)**. If the *Xic* is missing from one X chromosome and is present on the other X chromosome, then both X chromosomes remain active; two *Xics* must be present for one X chromosome to be inactivated. This result suggests that it is not the X chromosomes that are counted by the cell, *per se,* but actually the number of *Xics*. If two or more *Xics* are present in a cell, only one remains active. The additional *Xics* (and the X chromosomes) are inactivated.

Key Questions

- Why does the number of Barr bodies differ in genetic disorders of the X chromosome?

Xist and *Tsix*

The *Xic* region of the X chromosome contains two genes: ***Xist*** and ***Tsix*** (see **figure 2.11**).

- **The *Xist* gene**. The *Xist* (X-inactive specific transcript) gene is transcribed preferentially from the X chromosome that will be inactivated. The *Xist* gene produces an RNA molecule that is a **non-coding RNA (ncRNA).** Non-coding RNA molecules function directly in the cell and are not translated to make a protein product. The ***Xist* RNA** functions to recruit proteins that modify the structure of the X chromosome, converting an active X chromosome into a Barr body composed of heterochromatin.
- **The *Tsix* gene**. When the *Tsix* gene is transcribed, a ***Tsix*** ncRNA is made. The *Tsix* RNA is transcribed from both X chromosomes prior to XCI and likely allows the two X chromosomes to pair briefly during early embryonic development (see Part 15). The *Tsix* RNA is later transcribed preferentially from the active X chromosome, which in turn inhibits the production of the *Xist* RNA on the active X chromosome. Thus, the *Xist* RNA does not inactivate one X chromosome because the *Tsix* gene is transcribed. Interestingly, the *Tsix* gene overlaps the *Xist* gene but is transcribed in the opposite direction.

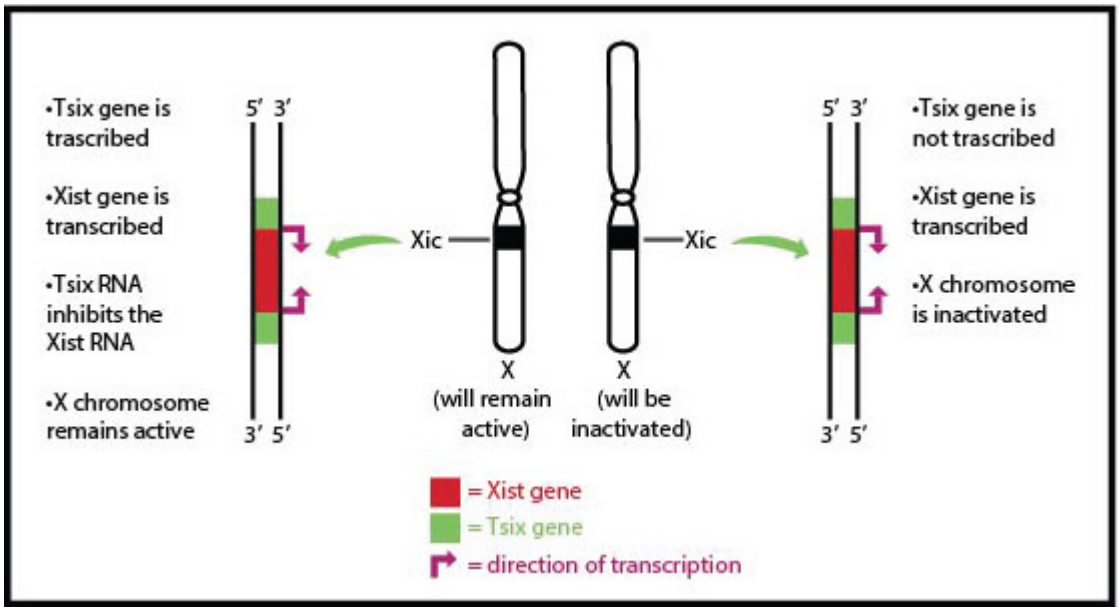


Figure 2.11 ***Mechanism of X Chromosome Inactivation*** *--- Image created by SL*

Key Questions

- How does the *Xist* gene promote the formation of a Barr body?
- How does the *Tsix* gene ensure that one X chromosome remains active?

The Three Stages of XCI

The process of XCI has three stages (see **Figure 2.12**):

1. The **initiation** stage involves a choice as to which X chromosome is inactivated. *Tsix* gene expression from the X chromosome that will remain active inhibits the production of the *Xist* RNA. The X chromosome that is inactivated does not express the *Tsix* RNA. As a result, the *Xist* RNA is produced, leading to the formation of a Barr body.
2. The **spreading** stage involves the actual inactivation of the chosen X chromosome. The *Xist* RNA participates in spreading by starting at the *Xic* and coating the X chromosome to be inactivated in both directions. This coating with *Xist* RNA allows proteins to recognize the X chromosome and compact it into heterochromatin, forming a Barr body.
3. **Maintenance** of the Barr body after cell division. Suppose a cell divides by mitosis. Prior to mitosis, this cell converts the Barr body back into euchromatin and replicates the DNA. The replicated X chromosome is then separated into the daughter cells and is silenced again by converting it back into a Barr body. Interestingly, the daughter cells have the ability to remember which X chromosome was inactivated in the parent cell prior to cell division. Thus, the two progeny cells both contain inactive copies of the same X chromosome as the parent cell.

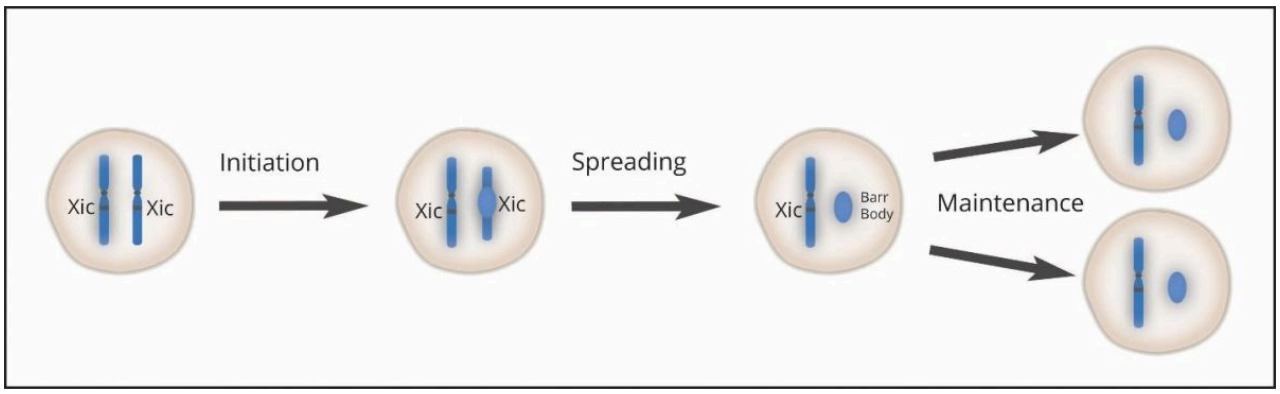


Figure 2.12 ***Stages of XCI*** *--- Image created by SL*

Key Questions

- Explain the three stages of XCI.

Review Questions

Fill in the blanks:

Chromosome Compaction Strategies in Prokaryotes and Eukaryotes

1. _____ are proteins that help compact the *E. coli* chromosome into microdomains.
2. The enzyme _____ produces both positive and negative supercoiling.
3. Topoisomerase _____ uses ATP to generate negative supercoils in bacterial chromosomes.
4. Histone _____ is also called the linker histone because it links two nucleosomes together.
5. A nucleosome contains _____ bp of DNA wrapped around _____ histone proteins.
6. Amino acids _____ and _____ are positively charged, and make up many of the amino acids in the histone tail.
7. The _____ form of the 30-nm fiber resembles a cylinder.
8. DNA sequences called _____connect the 30-nm fiber to the internal nuclear matrix.
9. _____ heterochromatin can be transcribed under certain cellular conditions.
10. A region in the nucleus where a particular chromosome is located is called a chromosome _____.
11. During interphase, condensin proteins are located in the _____ of the cell.

X Chromosome Inactivation

1. A normal female has _____ Barr body(bodies) whereas a female with Turner syndrome has _____ Barr body (bodies).
2. When the *Xist* gene is active, it transcribes an RNA molecule that attaches to the chromosome that becomes (active or inactive). Circle the correct answer.
3. When the *Tsix* gene is active, it transcribes an RNA molecule that is associated with the (active or inactive) X chromosome. Circle the correct answer.
4. The inactive X chromosome condenses into a tight nuclear structure called a _____.

3 - Chromosome Variation

In the first half of Part 3, we will consider deficiencies in chromosome structure. Deficiencies in chromosome structure refer to altering the total amount of genetic information on a chromosome (**deletions, duplications**), rearranging the order of genes on a chromosome (**inversions**), or moving genes from one chromosome to a nonhomologous chromosome (**translocations**).

In the second half of Part 3, we will consider situations in which the number of chromosomes within an individual varies (**variations in euploidy** and **aneuploidy**).

A. Changes in Chromosome Structure

Overview

We learned in Part 1 that most structural genes are unique DNA sequences, found as a single copy on a particular chromosome. However, since we have two copies of each chromosome (one copy of the homologous chromosome pair is inherited from dad; the other member of the pair is inherited from mom), each structural gene is actually present in two copies per genome. Changes to this general rule include the following (see **figure 3.1**):

- **Deletions**. A deletion occurs when a portion of a chromosome is missing. A deletion can be as small as a single base pair or can include the loss of several genes. The portion of the chromosome that is missing is called a **deficiency**. A person who suffers a deletion would have a single copy of one or several structural genes.
- **Duplications**. A duplication occurs when a portion of the chromosome is repeated. In a duplication, a single chromosome can have more than one copy of the same structural gene.

Mutations can also move structural genes from their normal location to a new location in the genome. The mutations that alter the location of a gene include:

- **Inversions**. Inversions involve changing the direction of an internal segment within a single chromosome. Inversions change the location of a structural gene within an individual chromosome.
- **Translocations**. A translocation occurs when a portion of a chromosome becomes attached to a nonhomologous chromosome. For example, a portion of chromosome 1 can be translocated to chromosome 5. There are three types of translocations:
    - **Simple (nonreciprocal) translocation**. A simple translocation occurs when a segment of one chromosome becomes attached to a nonhomologous chromosome. The chromosome receiving the DNA segment remains intact.
    - **Reciprocal translocation**. Reciprocal translocations involve nonhomologous chromosomes exchanging pieces. For example, one copy of chromosome 1 and one copy of chromosome 5 could exchange telomere regions.
    - **Robertsonian translocation**. Robertsonian translocations involve the fusion of the long (*q*) arms of two acrocentric chromosomes. For example, the *q* arm of one copy of chromosome 14 can fuse with the *q* arm of one copy of chromosome 21. The *p* arms of chromosomes 14 and 21 are lost (see below).

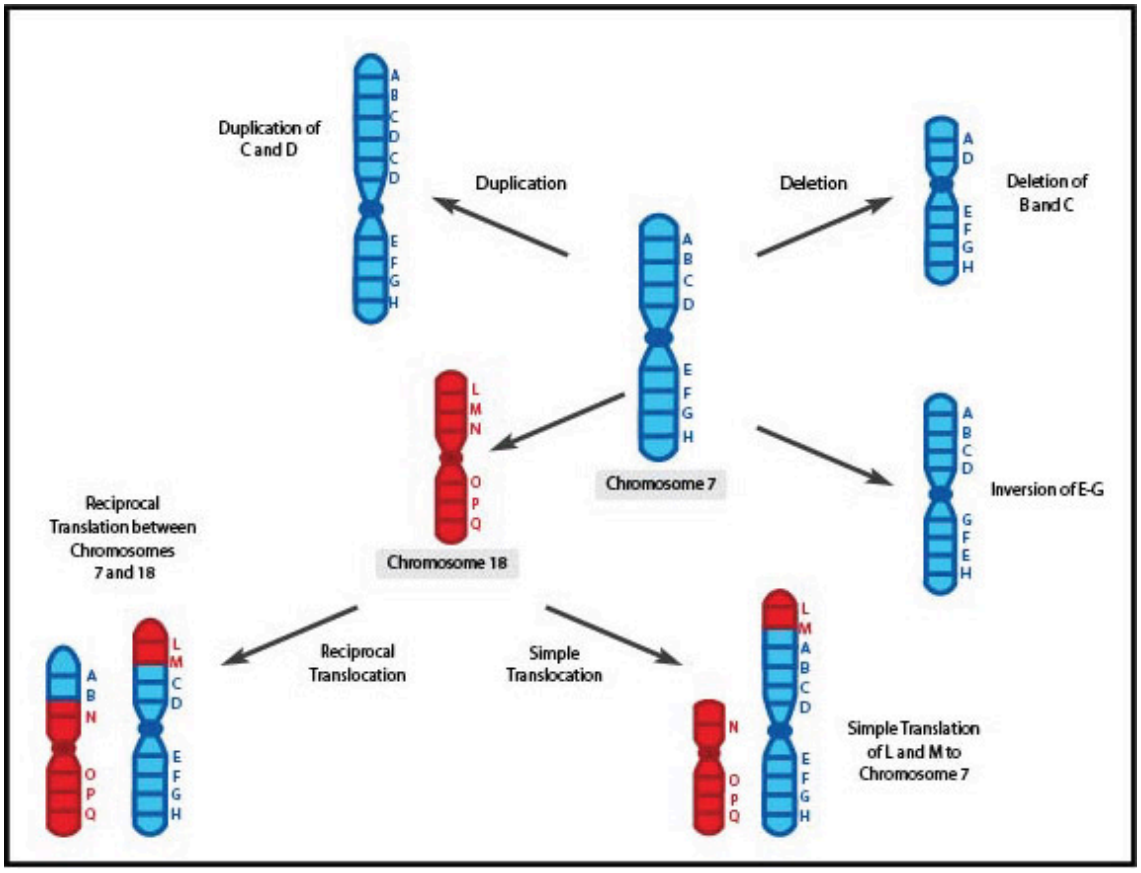


Figure 3.1 Changes in Chromosome Structure (Overview). *The Robertsonian translocation in not shown in the image. Image created by SL.*

Key Questions

- What are the four major types of chromosome structural defects?
- What are the differences between the three types of translocations?

Deletions

One or more DNA breaks can lead to the loss of a portion of the chromosome. This type of chromosomal aberration is called a **deletion** (see **figure 3.2**). A single break in a chromosome can result in a DNA fragment that contains a centromere and a DNA fragment that lacks a centromere. The centromere fragment is retained by the cell, while the DNA fragment that lacks the centromere is lost during cell division. This type of event is a **terminal deletion**. Terminal deletions are usually generated by endonuclease damage or by environmental factors, such as ionizing radiation, that break the DNA backbone. An **interstitial deletion** is a deletion within the interior of the chromosome and does not involve the telomere. Interstitial deletions are generated by defects in synapsis and crossing over during meiosis I (see below).

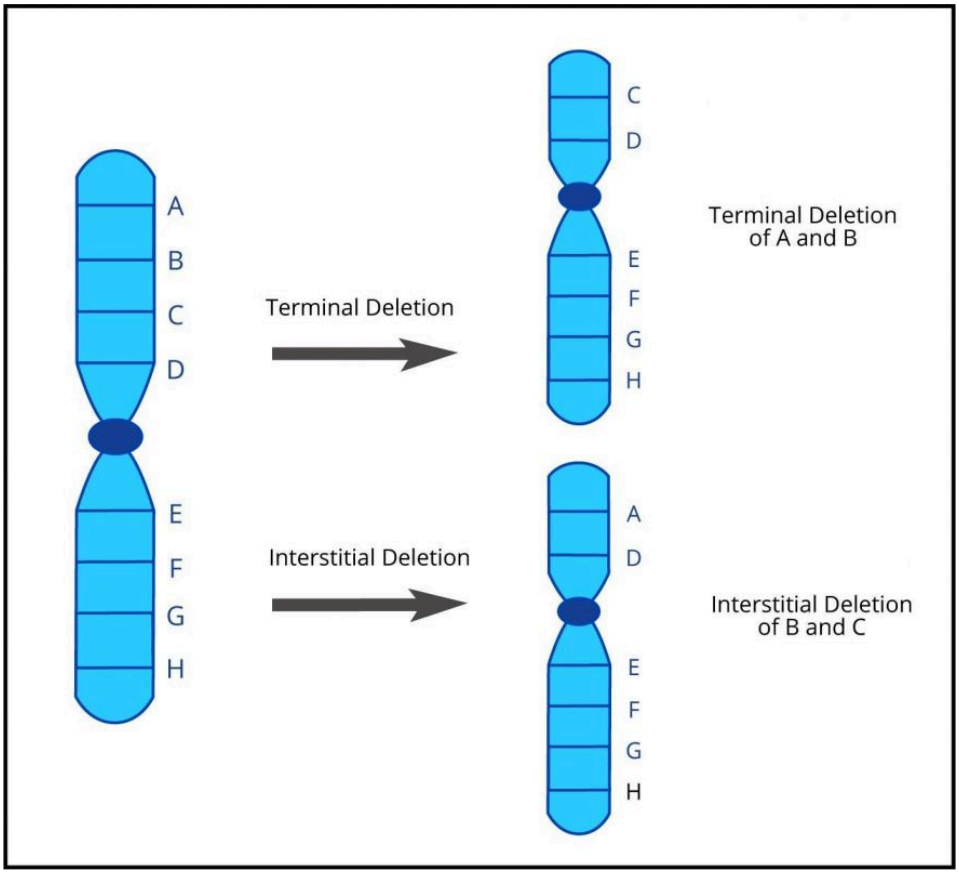


Figure 3.2 Deletions. *A terminal deletion involves the loss of a telomere region. An interstitial deletion involves the loss of genes within the chromosome. --- Image created by SL*

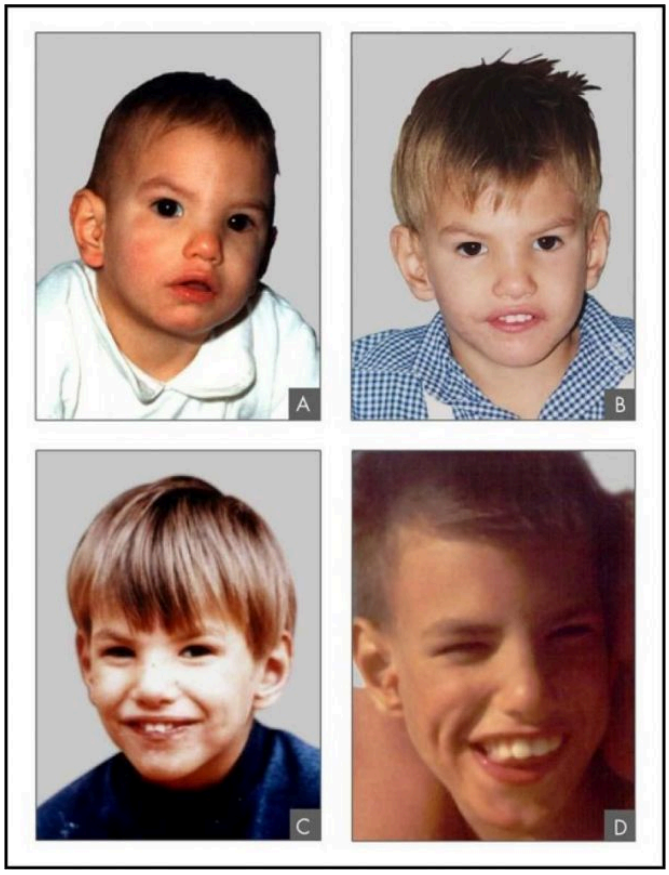
Key Questions

- What is the difference between an interstitial and a terminal deletion?

Cri-du-chat

In general, the larger the deletion (i.e., the more structural genes involved), the more severe the phenotypic consequences. Moreover, a detrimental phenotype can occur even though the individual may have a normal copy of the homologous chromosome, indicating that most deletions behave as dominant mutations.

Cri-du-chat (*46, 5p-*) is an example genetic disease caused by a deletion in the *p* arm of chromosome 5. Cri-du-chat occurs in 1 in 25,000–50,000 live births (see **figure 3.3**). Cri-du-chat is usually not inherited; instead, the disease is caused by the loss of the *p* arm of chromosome 5 during meiosis. A cri-du-chat individual usually has one normal copy of chromosome 5 and a **terminal deletion** copy of the same chromosome. The deletion in chromosome 5 can be quite small or can encompass much of the *p* arm; however, it is thought that the absence of a specific gene causes cri-du-chat. This missing gene encodes telomerase reverse transcriptase (*TERT*). We will learn about the function of *TERT* in Part 6. The cri-du-chat individual displays mental deficiencies, facial abnormalities, gastrointestinal, and cardiac complications. Those afflicted also tend to vocalize using a catlike cry, due to defects in the formation of the glottis and larynx.


A
B
C
D

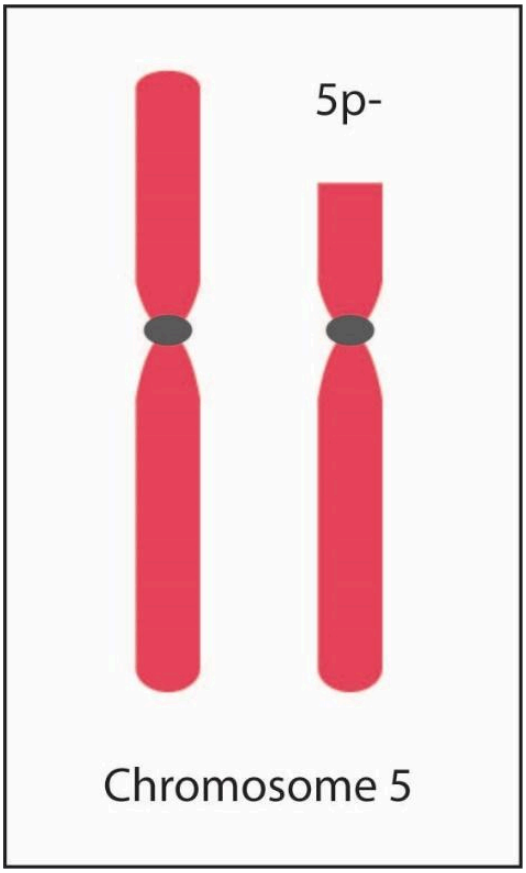


***Figure 3.3 Cri-du-cha**t A) A Cri-du-chat patient--- CriDuChat by Paola Mainardi is licensed under CC BY 2.0 B) Terminal deletion in chromosome 5 --- Image created by SL*

Key Questions

- Which change in chromosome structure produces cri-du-chat?

Duplications

A **duplication** produces two copies of a structural gene on a single chromosome. Since the homologous chromosome contributes another copy of the same structural gene, a person with a duplication has three copies of the structural gene, instead of two copies. As the region of the chromosome that is duplicated gets larger, the phenotypic effect on the individual becomes more severe. One example disease caused by a duplication is the neuropathic disease **Charcot-Marie-Tooth disease type 1A (CMT type 1A)**, produced by a duplication on chromosome 17.

Duplications and interstitial deletions can be produced simultaneously by the misalignment of synapsed homologous chromosomes during meiosis, followed by **unequal crossing over** (see **figure 3.4**). Unequal crossing over produces four gametes. One gamete contains an interstitial deletion chromosome, and a second gamete contains a chromosome with a duplication. The final two gametes contain the normal allele arrangement.

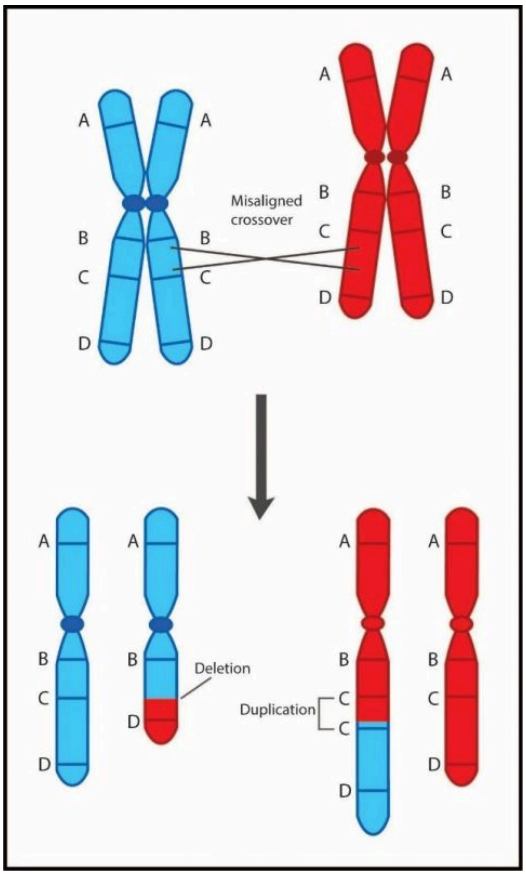


Figure 3.4 Unequal Crossing Over Produces an Interstitial Deletion and a Duplication *--- Image created by SL*

Key Questions

- How can a meiosis defect produce a gamete with a deletion and a gamete with a duplication?
- What human disease is caused by a duplication?

Duplications Can Produce Gene Families

Small duplications can sometimes be beneficial and are important in the formation of **gene families**, closely related genes that have similar but not identical functions. For example, the **globin gene family** in humans is thought to have been formed by multiple duplications from a single ancestral globin gene (see **figure 3.5**). To form the globin gene family, the ancestral globin gene was duplicated to produce two identical genes on the same chromosome. These two

genes then accumulated mutations independently over the course of thousands of generations to become specialized; one gene became a hemoglobin gene, the other became a myoglobin gene. Later, the hemoglobin gene duplicated additional times followed by divergence through the continued accumulation of mutations. The current globin gene family, consisting of fourteen member genes, includes genes that encode the protein subunits of hemoglobin, which is specialized to carry oxygen in the bloodstream, and the protein subunits of myoglobin, which carries oxygen within muscles. The globin gene family is a good example of how gene duplication can produce the genetic variability necessary to drive evolution.

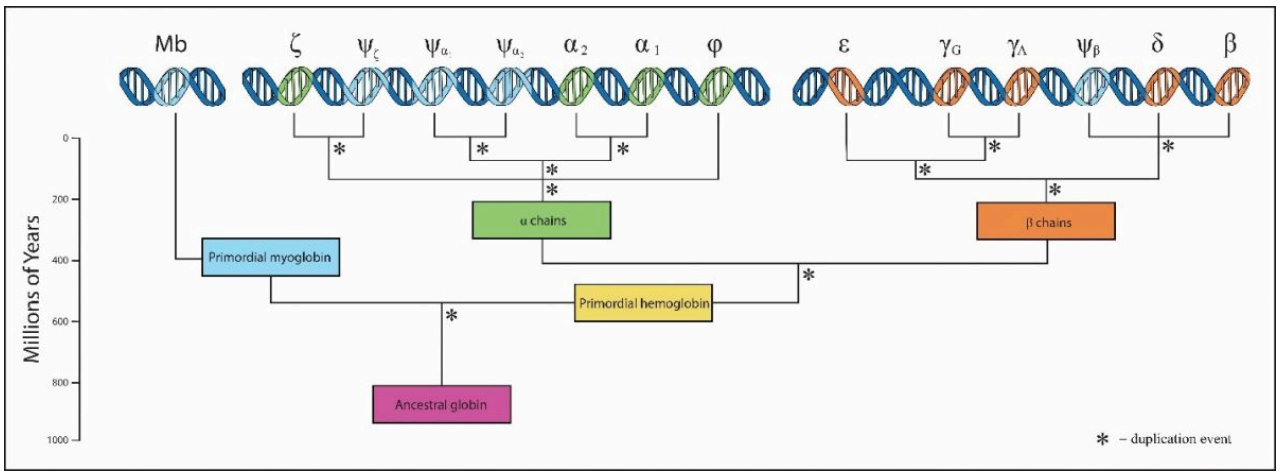


Figure 3.5 The Globin Family. *An ancestral globin gene duplicated to produce a primordial myoglobin and primordial hemoglobin gene. The primordial hemoglobin gene duplicated additional times to produce specialized globin genes.---Image created by SL*

Key Questions

- Describe how gene duplications and mutations can produce a gene family.

Types of Inversions

Inversions involve the rearrangement of genes along a single chromosome. An inversion can be thought of as breaking the chromosome in two places, flipping the DNA between the breaks, and sealing the DNA breaks. The total amount of genetic material (number of structural genes) in the chromosome does not change. Interestingly, inversions are quite common; about 2% of the human population carry a detectable inversion.

There are two types of inversions (see **figure 3.6**):

- **Pericentric inversion**. In a pericentric inversion, one chromosome break occurs in the *p* arm, while a second break occurs in *q* arm of the same chromosome. The central region of the chromosome, including the centromere, is located within the inverted region. Note that a pericentric inversion has the potential to change the position of the centromere within the chromosome.
- **Paracentric inversion**. In a paracentric inversion, two chromosome breaks occur within the same arm of the chromosome. The chromosome region between the two breaks is inverted, with the centromere of the chromosome lying outside of the inverted region. As a result, a paracentric inversion does not change the position of the centromere within the chromosome.

Most inversions have no phenotypic consequences; however, if one of the chromosome breaks that lead to an inversion occurs within a gene, then a change in phenotype can occur. For example, in **type A hemophilia**, the breakpoint of an inversion on the X chromosome occurs within the **factor VIII** gene. The encoded Factor VIII protein is required for proper blood clotting; this inversion produces a nonfunctional protein, leading to a deficiency in blood clotting (hemophilia). Further, the change in the position of a structural gene on a chromosome can alter the transcription of nearby genes. This alteration of transcription by an inversion is called a **position effect**. In some cases, the position effect can result in the overexpression of genes that regulate the cell cycle, resulting in cancer.

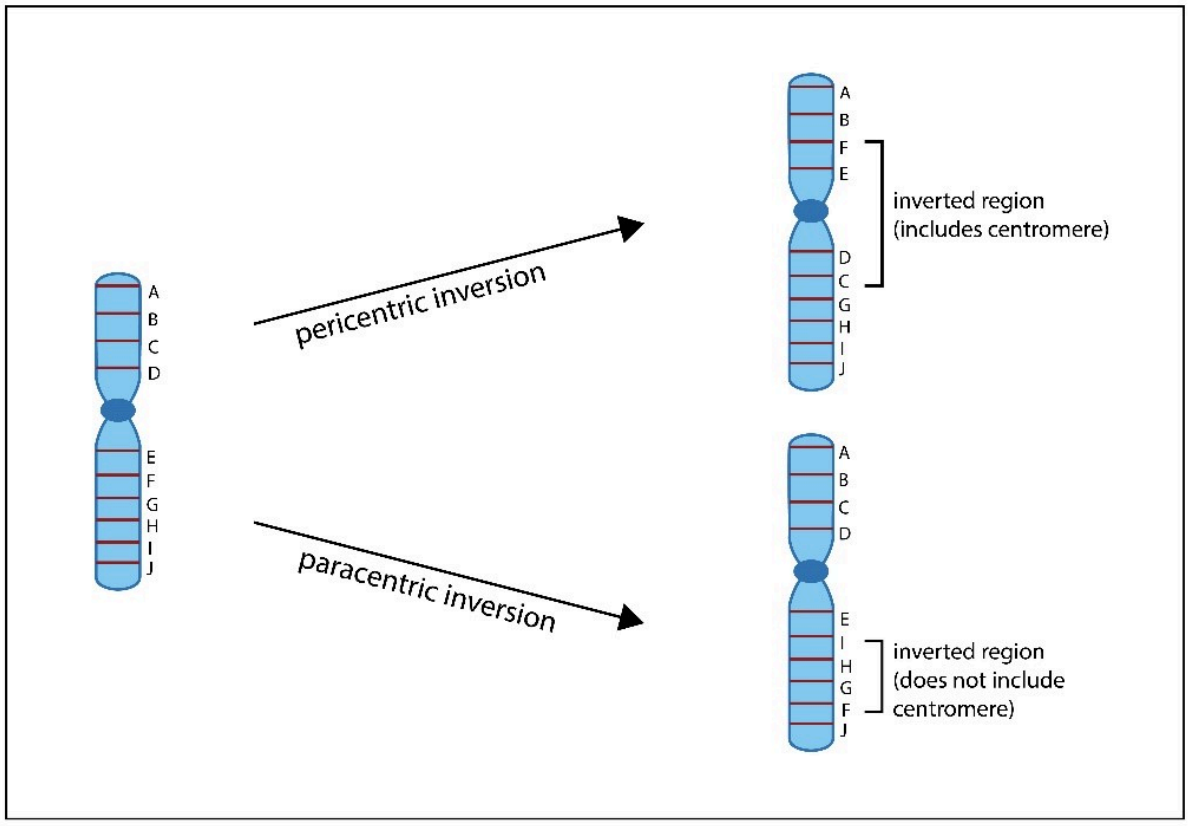


Figure 3.6 Pericentric and Paracentric Inversions. *Pericentric inversions include the centromere in the inverted region, while paracentric inversions do not involve the centromere. --- Image created by SL*

Key Questions

- What is the difference between a pericentric and a paracentric inversion?
- What human disease is caused by an inversion?

Crossing Over in Inversion Heterozygotes

An **inversion heterozygote** is an individual who has a chromosome with a normal gene arrangement, while the homologous chromosome contains an inversion. Even though an inversion heterozygote individual has a normal phenotype, they produce unusual gametes during meiosis. Recall that prior to meiosis, the two chromosomes within a homologous chromosome pair are copied by DNA replication, producing four sister chromatids (see **figure 3.7**). DNA replication is followed by synapsis (alignment) of the homologous chromosome pair during meiosis I. For the normal

chromosome and the inversion chromosome to synapse properly, one of the two chromosomes twists to form an **inversion loop**. After the inversion loop is formed, crossing over occurs between the two chromosomes within the homologous chromosome pair. Once crossing over is concluded, abnormal chromosomes are distributed to gamete cells.

- **Pericentric inversion.** In a pericentric inversion, the centromere lies within the inverted region of the chromosome. When crossing over occurs between the homologous chromosomes and after meiosis is completed, the following gametes are produced:
  - A gamete that contains a chromosome with the normal gene arrangement. This gamete will produce offspring with the normal gene arrangement and phenotype.
  - A gamete that contains an inversion chromosome. Thus, inversion chromosomes are passed from parents to offspring. The resulting offspring will suffer no negative phenotypic effects.
  - Two gametes that contain abnormal chromosomes. Each gamete contains a chromosome with a duplication of some genes and a deletion of other genes. The fertilized egg produced from either of these two gametes is not generally viable.

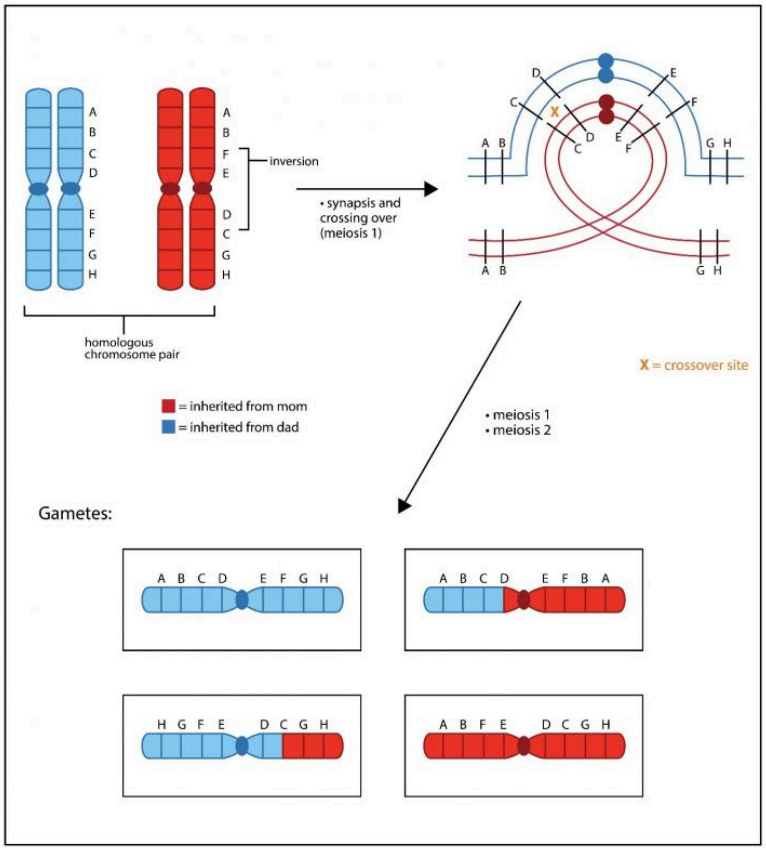


Figure 3.7 Meiosis in an Individual Heterozygous for a Pericentric Inversion. *One of the gametes has the normal allele arrangement, while a second gamete contains the inversion. Two of the gametes (bottom) contain simultaneous duplications and deletions. --- Image created by SL*

- **Paracentric inversion**. In a paracentric inversion, the centromere lies outside of the inverted region of the chromosome. After crossing over between the homologous chromosomes occurs and meiosis is completed, the following gametes are produced (see **figure 3.8**):
    - A gamete that contains a chromosome with the normal gene arrangement. This gamete will produce offspring with the normal gene arrangement and phenotype.
    - A gamete that contains an inversion chromosome. The inversion chromosome is passed from parents to offspring. The resulting offspring will suffer no negative phenotypic effects.
    - Two gametes that contain highly unusual chromosomes. One gamete contains an abnormal chromosome containing duplications and deletions, but more importantly, the chromosome lacks a centromere. This **acentric fragment** will be lost during cell division. The other gamete will contain an abnormal chromosome that has two centromeres (**dicentric chromosome**) along with duplications and deletions. Between the two centromeres of this dicentric chromosome is a region called a **dicentric bridge**. When a dicentric chromosome, attached to opposite spindle poles, tries to separate during anaphase of meiosis, it is torn apart. Breakage produces chromosome fragments that are missing genes.

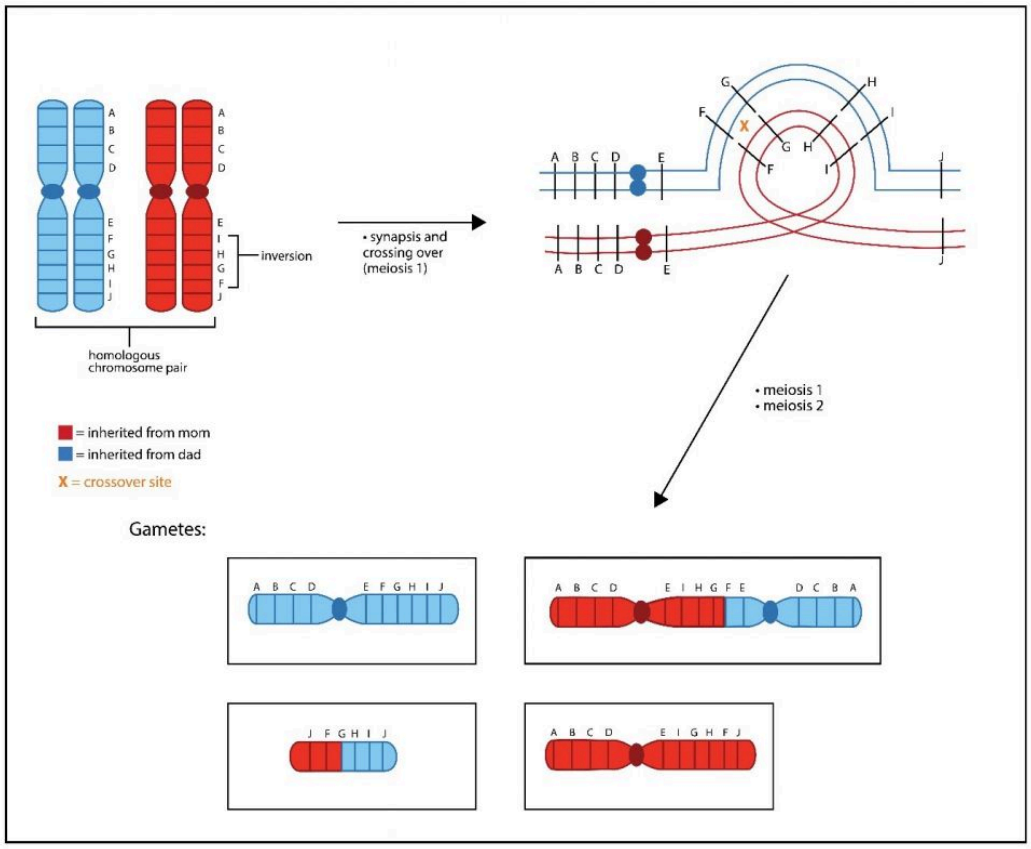


Figure 3.8 Meiosis in an Individual Heterozygous for a Paracentric Inversion. *One of the gametes has the normal allele arrangement, while a second gamete contains the inversion. Two of the gametes (bottom) contain simultaneous duplications and deletions; however, one gamete lacks a centromere (acentric), while the other gamete has two centromeres (dicentric).--- Image created by SL*

Importantly, 50% of the gametes produced by either pericentric or paracentric inversion heterozygotes fail to produce viable offspring. Thus, even though the inversion may not affect the individual's phenotype directly, the inversion causes a 50% reduction in fertility.

Key Questions

- Describe the four gametes produced by an individual who carries a pericentric inversion.
- Describe the four gametes produced by an individual who carries a paracentric inversion.
- What is a dicentric bridge, and why does it produce lethal products?

Reciprocal Translocations

A **translocation** occurs when a piece of a chromosome becomes attached to a nonhomologous chromosome. As mentioned earlier, there are three types of translocations: **simple translocations**, **reciprocal translocations**, and **Robertsonian translocations.** We will focus primarily on reciprocal and Robertsonian translocations.

Reciprocal translocations involve nonhomologous chromosomes exchanging pieces. Reciprocal translocations are formed by two general mechanisms (see **figure 3.9**):

- **Chromosome breakage and defective DNA repair**. Some chemicals or environmental agents can cause chromosomes to break at internal sites, forming reactive ends not protected by **telomeres**. Recall that telomeres are the structures found on the ends of linear eukaryotic chromosomes that are designed to prevent chromosome ends from sticking together (see Part 1). Cells contain repair enzymes to handle these situations, and in most cases, quickly repair these breaks. When nonhomologous chromosomes are broken simultaneously, the repair enzymes can sometimes inadvertently join nonhomologous chromosomes together, resulting in a reciprocal translocation.
- **Nonhomologous chromosomes crossing over**. If two nonhomologous chromosomes accidently synapse and undergo crossing over during meiosis I, a reciprocal translocation occurs.

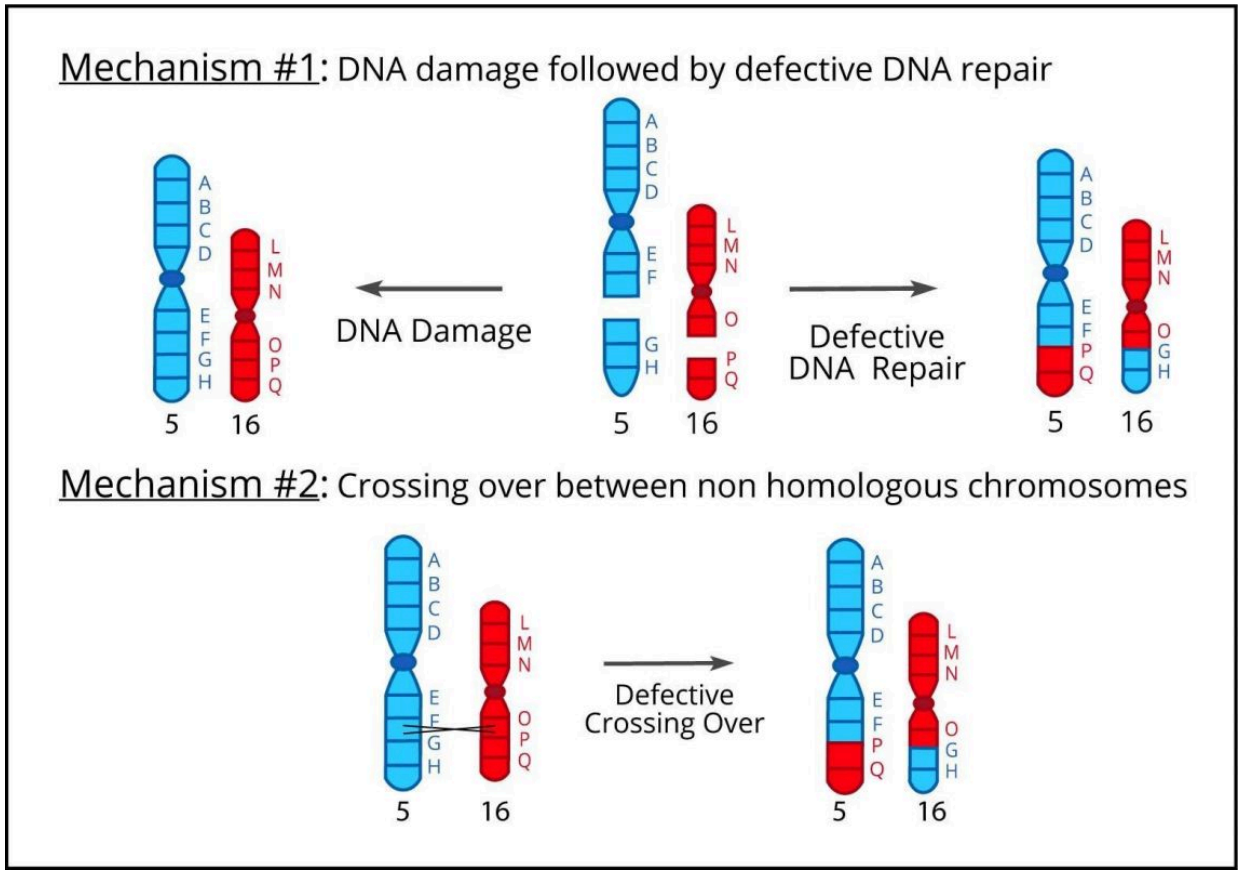


Figure 3.9 Mechanisms for Producing Reciprocal Translocations. *The chromosome number is shown beneath each chromosome. --- Image created by SL*

Key Questions

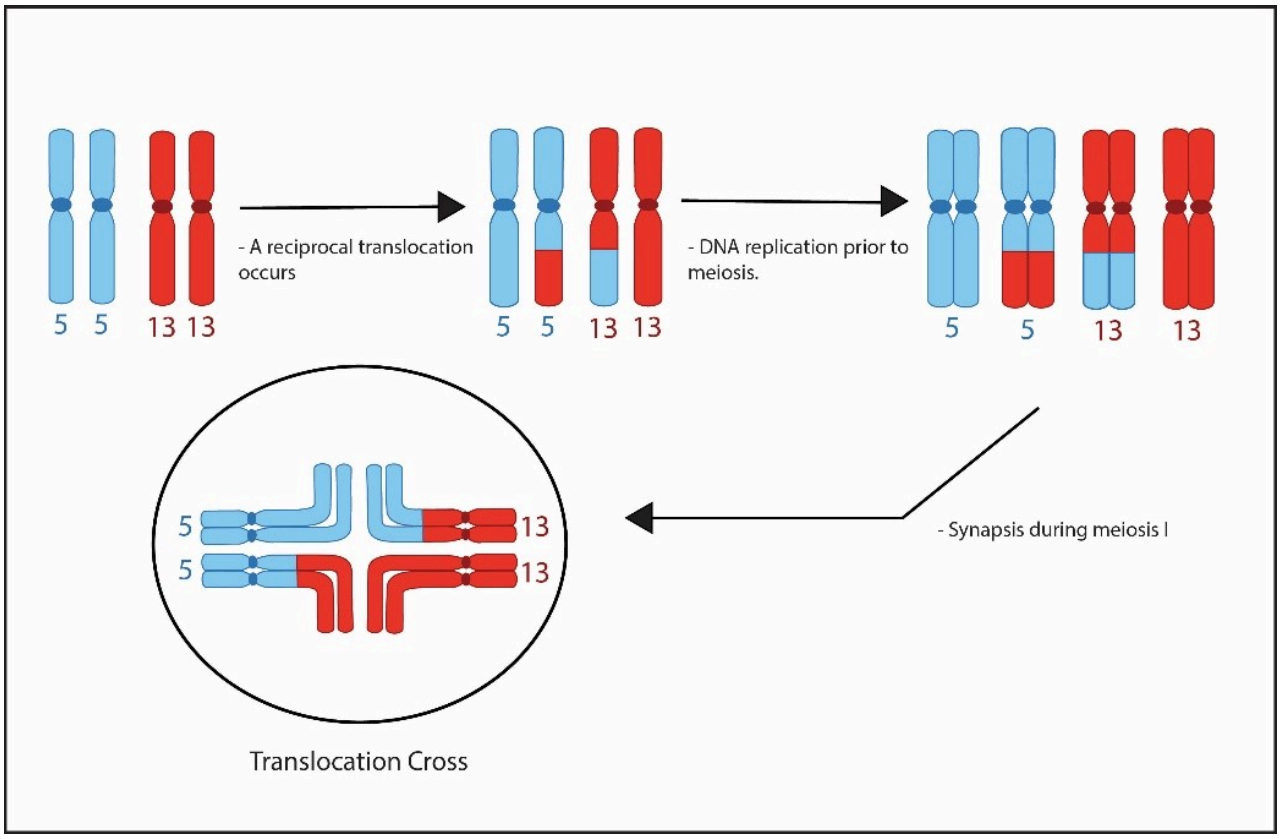
- Describe the two processes that generate reciprocal translocations.

Meiosis in Cells with Reciprocal Translocations

How do nonhomologous chromosomes that have experienced a reciprocal translocation synapse and then segregate into gametes during meiosis? During synapsis, the two pairs of homologous chromosomes (four chromosomes total) that include two individual chromosomes that have suffered a reciprocal translocation attempt to synapse. Because of the reciprocal translocation, the four chromosomes synapse to form a **translocation cross** (see **figure 3.10**).

For example, suppose a translocation cross is produced from a normal copy of chromosome 5, a normal copy of chromosome 13, and a situation in which the other copies of chromosomes 5 and 13 have undergone a reciprocal translocation. Prior to meiosis, these four chromosomes are copied by DNA replication to produce eight sister chromatids. In order to synapsis properly, the normal copy of chromosomes 5 and 13 end up diagonal from each other in the translocation cross, while the two translocation chromosomes are diagonal from each other (see **figure 3.10**). The chromosomes in the translocation cross can then segregate during anaphase I in three possible ways:

- **Alternate segregation** (common). The chromosomes diagonal from each other segregate into the same cells at the conclusion of meiosis I. For example, the normal copies of chromosomes 5 and 13 segregate with each other into the same daughter cell while the two translocation chromosomes segregate with each other into the same daughter cell. After meiosis II, there are two normal gametes and two gametes that carry translocations. Because none of these gametes are missing genes, all four gametes produced by alternate segregation are capable of producing viable offspring.
- **Adjacent-1 segregation** (common). The two chromosomes on the bottom half of the translocation cross (normal chromosome 13 and translocation chromosome 5) segregate with each other into the same daughter cell. The two chromosomes on the top half of the cross (normal chromosome 5 and translocation chromosome 13) segregate with each other into the same daughter cell at the conclusion of meiosis I. After meiosis II, all four gametes carry duplications and deletions and therefore are not generally viable.
- **Adjacent-2 segregation** (rare). The two chromosomes on the right half of the cross (normal chromosome 5 and translocation chromosome 5) segregate with each other into the same daughter cell, while the two chromosomes on the left half of the cross (normal chromosome 13 and translocation chromosome 13) segregate into the same daughter cell at the conclusion of meiosis I. One daughter cell receives, in essence, both copies of chromosome 5; the other daughter cell receives both copies of chromosome 13. Since both copies of chromosome 5 end up in the same cell and both copies of chromosome 13 end up in the same cell, adjacent-2 segregation can be considered a **nondisjunction** event (see below). After meiosis II, all four gametes carry duplications and deletions and are not generally viable.



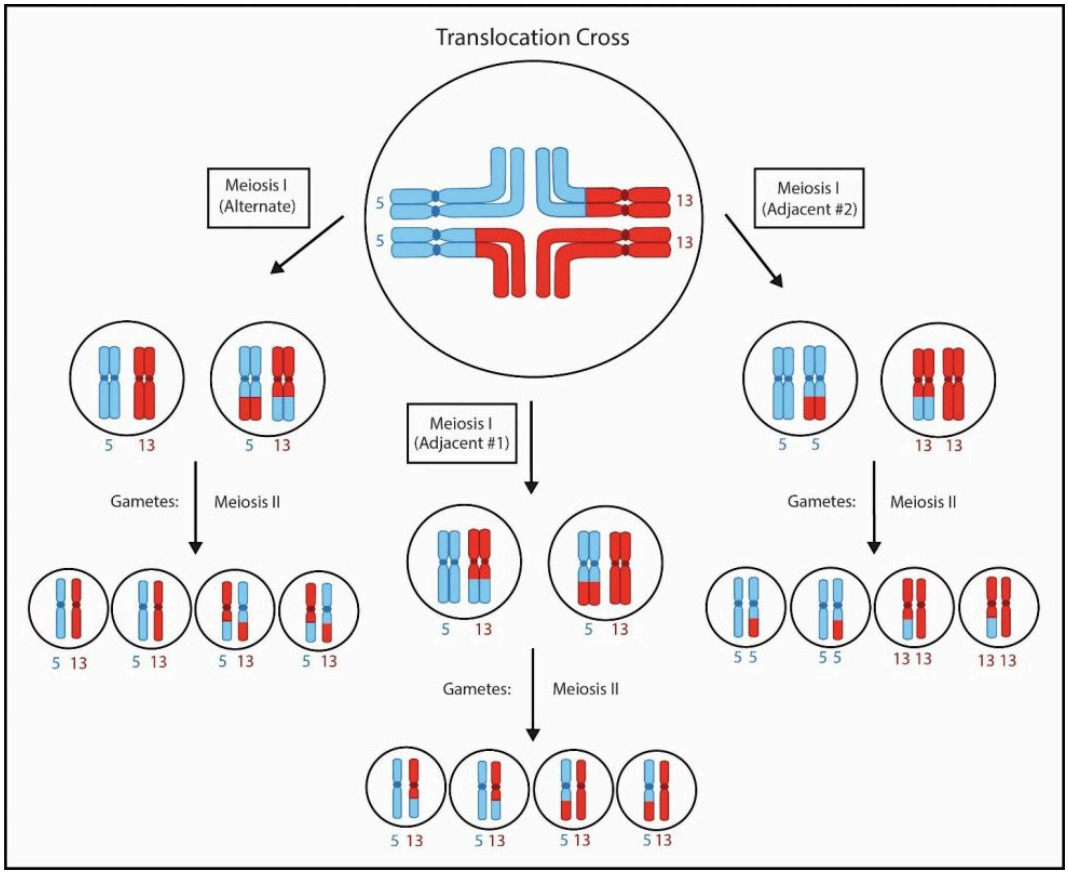


Figure 3.10 Formation of a Translocation Cross and Meiotic Chromosome Segregation – *Top) A reciprocal translocation produces a translocation cross during meiosis. Bottom) Chromosome segregation during meiosis I and II. Only alternate segregation can produce viable offspring.--- Image created by SL.*

Key Questions

- Which segregation pattern produces normal gametes?
- Why does adjacent-2 segregation occur so rarely?

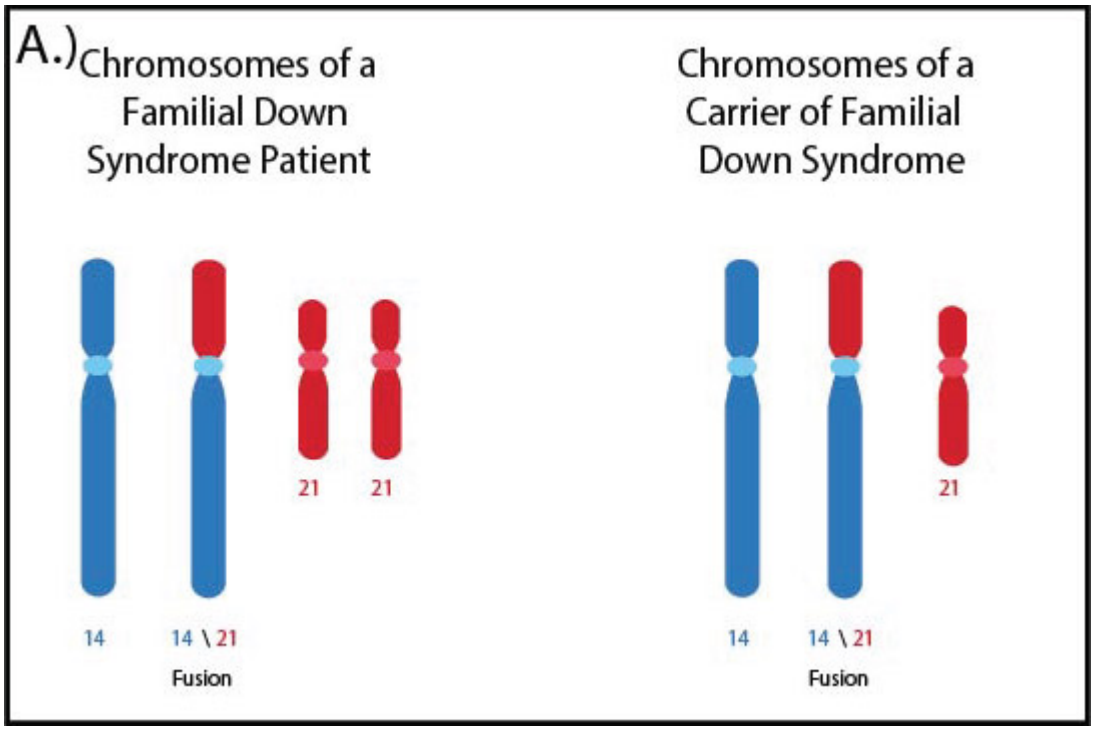
Robertsonian Translocations

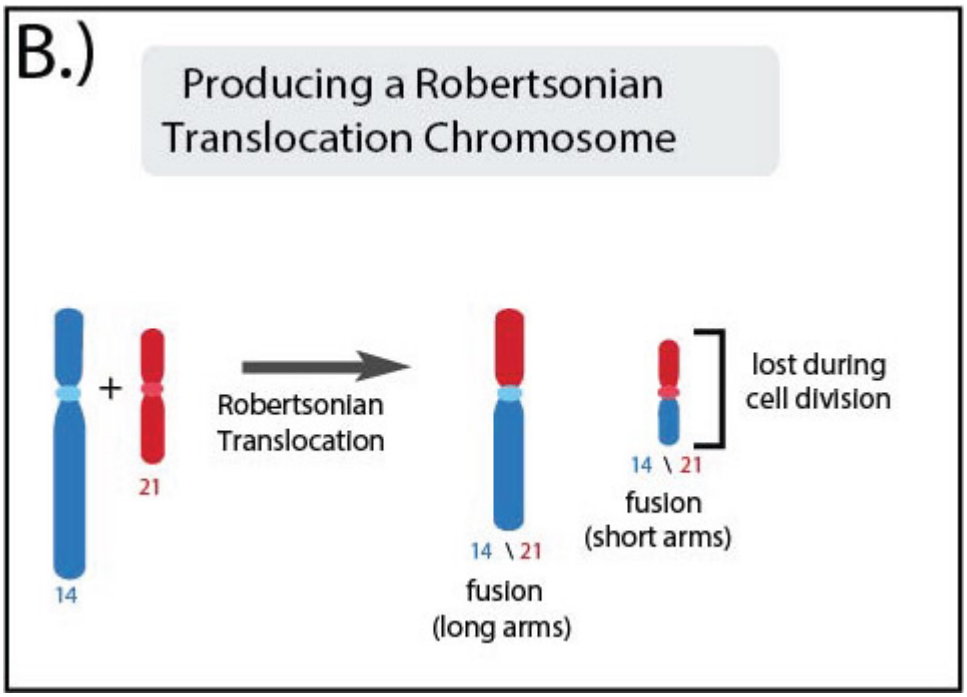
A rare form of Down syndrome called **familial Down syndrome** is inherited (see **figure 3.11**). In familial Down syndrome, a phenotypically normal parent can carry a translocation. This carrier individual has normal copies of chromosomes 14 and 21 and a chromosome that contains a fusion between the long (*q*) arms of chromosome 14 and 21. In this **balanced carrier** person, the short (*p*) arms of chromosome 14 and 21 have been lost, but since these regions carry repetitive DNA sequences that are found on other chromosomes in the genome (see Part 1), the individual can tolerate the loss of the two *p* arms. This type of translocation, involving the fusion of the long arms of two acrocentric chromosomes, is a **Robertsonian translocation**. The Robertsonian translocation, which involves only human chromosomes 13, 14, 15, 21, and 22 (i.e., acrocentric chromosomes), is the most common chromosome abnormality in humans.

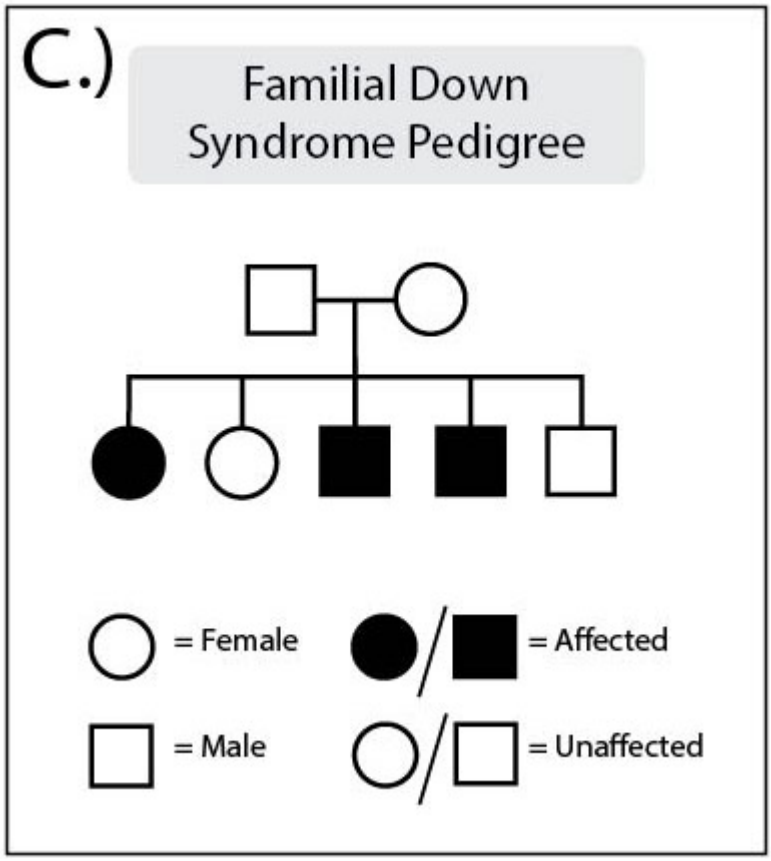
A problem occurs during meiosis in an individual that carries the Robertsonian translocation. In the case of a balanced carrier for familial Down syndrome, the normal chromosome 14, normal chromosome 21, and Robersonian

translocation chromosome replicate, synapse, and attempt to segregate into gametes during meiosis. There are six possible types of offspring that can be produced by the carrier individual:

- **Normal.** When a normal gamete containing chromosomes 14 and 21 produced by the carrier parent fuses with a normal gamete from the other parent during fertilization, an offspring is produced that contains two copies of chromosomes 14 and 21. This offspring is phenotypically normal.
- **Balanced carrier.** When a gamete containing the Robertsonian translocation between chromosomes 14 and 21 fuses with a normal gamete from the other parent during fertilization, a carrier offspring that contains one copy of chromosome 14, one copy of chromosome 21, and a Robertsonian translocation chromosome is produced. The total chromosome number in this person is 45 due to the loss of the short arms of chromosomes 14 and 21. This carrier is phenotypically normal but can produce familial Down syndrome offspring in the next generation.
- **Familial Down syndrome.** A gamete that contains the Robertsonian translocation between chromosomes 14 and 21 and a copy of chromosome 21 can fuse with a normal gamete from the other parent during fertilization. The offspring contains two copies of chromosome 21, one copy of chromosome 14, and the Robertsonian translocation chromosome. Since there are three copies of the long arm of chromosome 21 (trisomy-21), Down syndrome results. A familial Down syndrome patient has 46 total chromosomes. Conversely, conventional Down syndrome patients (see below) have 47 total chromosomes.
- **Unbalanced, lethal (three types of gametes).** Fifty percent of the gametes produced by a balanced carrier individual will not produce viable offspring. The resulting offspring produced from these gametes are either missing chromosome 14 (monosomy 14), missing chromosome 21 (monosomy 21), or have three copies of the long arm of chromosome 14 (trisomy 14).




B.)
Producing a Robertsonian Translocation Chromosome
+
21
14
Robertsonian Translocation
14 \ 21
fusion (long arms)
14 \ 21
fusion (short arms)
lost during cell division


C.)
Familial Down Syndrome Pedigree
= Female
= Affected
= Male
= Unaffected

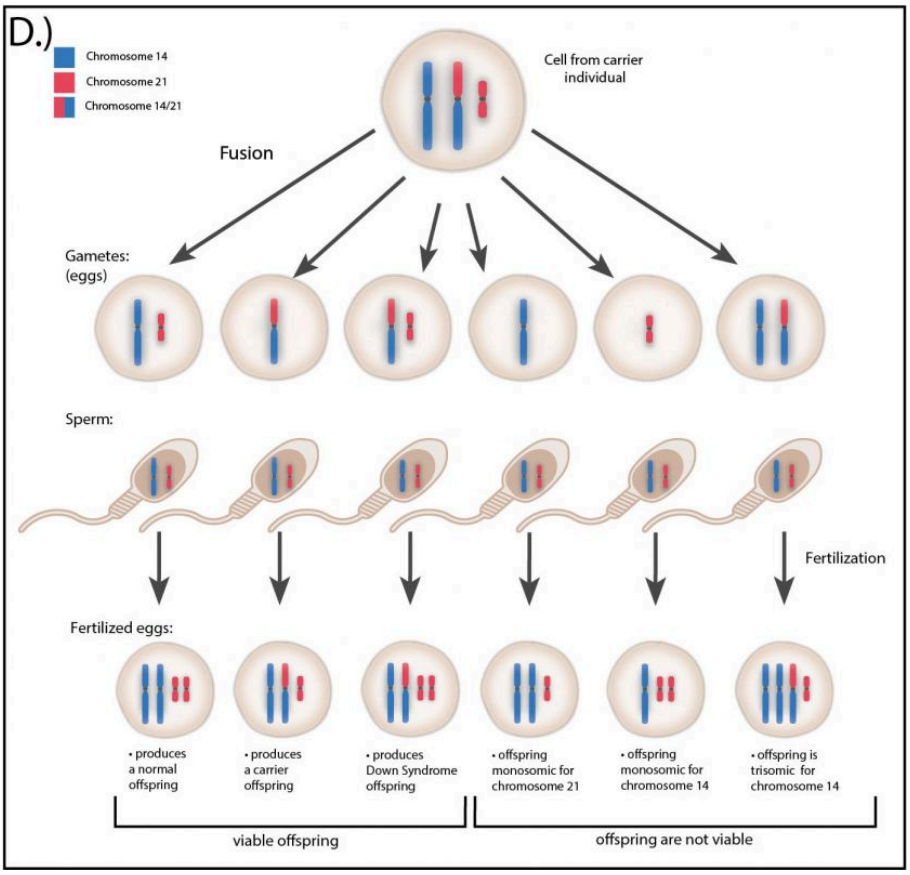


Figure 3.11 Robertsonian Translocation *- A) Chromosomes of a carrier and a familial Down Syndrome patient B) Mechanism of Robertsonian Translocation C) Familial Down Syndrome Pedigree D) Offspring Produced by a Familial Down Syndrome Carrier --- Images created by SL*

Key Questions

- What is a Robertsonian translocation?
- Which chromosomes are involved in familial Down syndrome?
- How many total chromosomes are found per cell in a balanced carrier individual?
- How many total chromosomes are found per cell in a familial Down syndrome patient?
- What repetitive DNA sequences are located on the *p* arms of the five acrocentric human chromosomes (see Part 1). Why would the loss of a few copies of these genes not be lethal to the cell?

B. Changes in Chromosome Number

Euploidy and Aneuploidy

Sometimes the total number of chromosomes within an individual can vary. These variations in chromosome number are placed into two categories (see **figure 3.12**):

- **Variations in euploidy.** Variations in euploidy involve changes in the total number of **chromosome sets** in a cell or individual. Recall that a chromosome set is all of the chromosomes inherited from one parent (i.e. humans contain two chromosome sets; each set contains 23 individual chromosomes). Variations in euploidy involves organisms or cells that are:
    - **Haploid** (n; n = the number of chromosomes within a set). Haploid organisms or cells have one chromosome set (i.e., one copy of every chromosome). Haploid is the normal state for gamete cells and some eukaryotic organisms.
    - **Diploid** (2n). Diploid organisms or cells have two chromosome sets. One chromosome set is inherited from the paternal parent; one chromosome set is inherited from the maternal parent. Diploid is the normal state for many eukaryotic organisms and for most of the cells in the human body.
    - **Triploid** (3n). Triploid organisms or cells have three chromosome sets. Triploid is an abnormal state for most organisms; however, there are examples of triploid plants. For example, seedless watermelon and banana plants are triploid.
    - **Polyploid**. Polyploid organisms or cells have more than two chromosome sets.
- **Aneuploidy.** Aneuploidy involves changes to the number of chromosomes within a chromosome set. For example, aneuploidy occurs when an individual is missing or has an additional chromosome within a set. Interestingly, aneuploidy of the sex chromosomes is usually better tolerated than aneuploidy of the autosomes (non-sex chromosomes) in many organisms. Aneuploid conditions include:
    - **Trisomy** (2n+1). A trisomic organism or cell has one more chromosome than normal. Trisomy is usually better tolerated than monosomy. The conventional form of Down syndrome is an example trisomic human condition (see below).
    - **Monosomy** (2n-1). A monosomic organism or cell is missing a single chromosome. Turner syndrome is an example monosomic human condition (see below).
    - **Disomy** (2n). Disomy is the normal state in which an organism or cell has two copies of a particular chromosome.
    - **Nullisomy** (2n-2). An organism or cell that is nullisomic is missing both copies of the chromosomes that constitute a homologous chromosome pair.

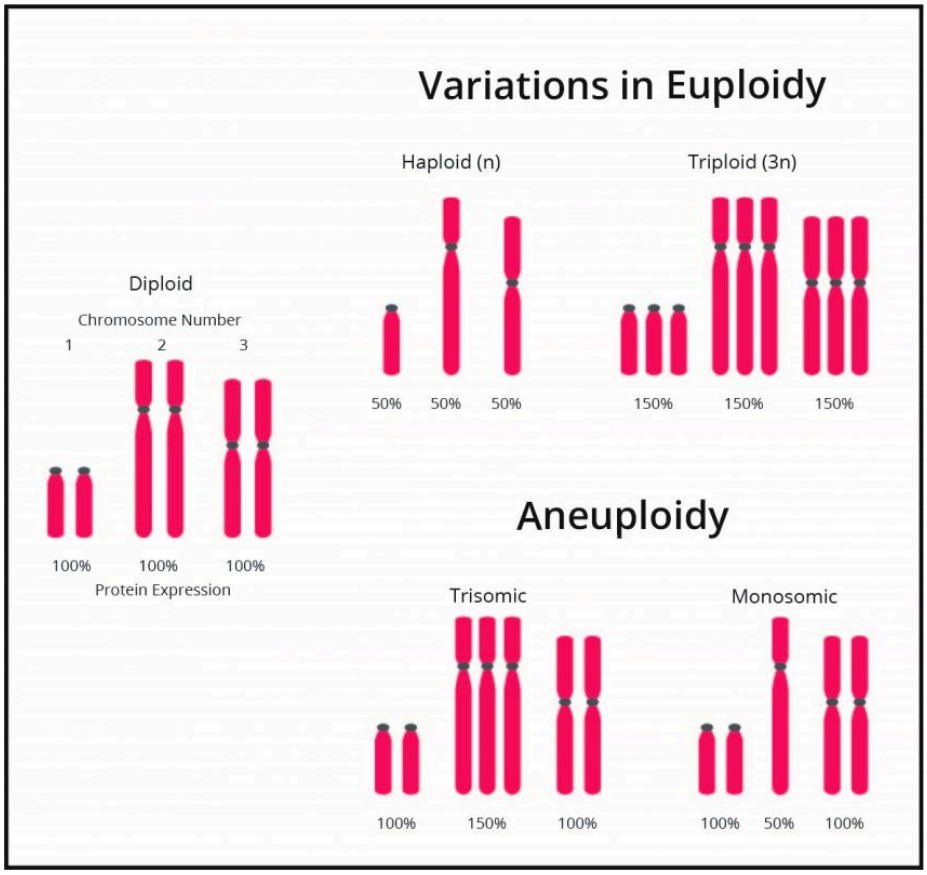


Figure 3.12 Changes in Chromosome Number (Overview). *Suppose an organism contains six total chromosomes organized into three homologous chromosome pairs (i.e., two sets of three chromosomes). Variations in euploidy changes the number of sets. Aneuploidy changes the number of chromosomes within a set. The percentages below each chromosome pair indicates the level of transcription for the structural genes found on particular chromosomes. Variations in euploidy and aneuploidy are often detrimental as each alters transcription levels of structural genes within the cell.--- Image created by SL*

Key Questions

- What is meant by variations in euploidy?
- What does aneuploidy mean?
- How many chromosomes are found in a trisomic human cell? (Note: human somatic cells contain 46 chromosomes.)
- How many chromosomes are found in a monosomic human cell?
- How many chromosomes are found in a triploid human cell?

Aneuploidy and Gene Expression

A phenotypically normal individual has two copies of most structural genes. When the number of structural genes is out of balance, the phenotype is often affected in a negative way. For example, in trisomic individuals with three copies of a particular chromosome, the amount of protein products produced from the three chromosomes are 150% of the normal expression level. These individuals produce too much protein product, so the phenotype is negatively affected (see **figure 3.12**). In the case of monosomy, the single copy of the chromosome can only produce protein products at 50%

of the normal level. Since these individuals produce lower amounts of protein product, the phenotype is negatively affected. Monosomic cells or individuals also have a second problem. In monosomic cells, recessive lethal alleles cannot be “masked” by the normal, dominant allele from the homologous chromosome.

In a trisomic or monosomic animal, the overproduction or underproduction of protein product decreases viability. However, it is worth noting that there are many natural varieties of plants that tolerate higher variations in euploidy. For example, wheat plants are hexaploid, some potato varieties are tetraploid, and wild strawberry plants can be octaploid.

Key Questions

- Why does trisomy have negative phenotypic consequences?
- What are the two reasons that monosomy produces negative phenotypic effects?

Aneuploidy in Humans

About 30% of all fertilization events in humans produces an embryo that is aneuploid. In most cases, the embryo does not survive to birth. That being said, there are some aneuploid human conditions that result in live births. These aneuploid conditions in humans include:

- **Trisomy 13 (*47,13+*)**. Trisomy 13 produces **Patau syndrome**, which occurs in 1 in 19,000 births. Patau syndrome causes mental and motor deficiencies, cleft palate, polydactyly (extra digits), microcephaly (a small head), defects in several organs, and an early death (usually by 3 months of age).
- **Trisomy 18 (*47, 18+*)**. Trisomy 18 produces **Edwards syndrome**, which occurs in 1 in 8,000 births. Edwards syndrome causes skeletal abnormalities such as elongated skulls, deformed hips, and facial deformities. Most infants with this syndrome are females, and death usually occurs within 4 months after birth.
- **Trisomy 21 (*47, 21+*)**. Trisomy 21 causes the conventional form of **Down syndrome**, which occurs in 1 in 800 births. Down syndrome results in mental deficiencies, almond-shaped eyes, flattened faces, round heads, and a short stature. As described earlier, there is another type of Down syndrome called **familial (inherited) Down syndrome**. Familial Down syndrome is caused by a Robertsonian translocation (described above).
- **XXY (*47, XXY*)**. An individual with XXY sex chromosomes has **Klinefelter syndrome**, which occurs in 1 in 1000 male offspring. Klinefelter syndrome results in infertility (no sperm production) and the formation of breast tissue. Klinefelter syndrome patients produce a single Barr body per cell.
- **XYY (*47, XYY*).** An individual with XYY sex chromosomes has **Jacobs syndrome**, which occurs in 1 in 1000 male offspring. Jacobs syndrome produces mild phenotypic effects.
- **XXX (*47, XXX*).** An individual with XXX sex chromosomes has **Triplo-X syndrome**, which occurs in 1 in 1500 female offspring. Triplo-X syndrome produces mild phenotypic effects. Triplo-X syndrome patients have two Barr bodies in each somatic cell.
- **XO (*45, X*)**. An individual with a single X chromosome has **Turner syndrome**, which occurs in 1 in 5000 female offspring. Turner syndrome females are short, have a webbed neck, and have reduced fertility. Turner syndrome patients do not produce Barr bodies.

The aneuploid conditions described above are the result of chromosome **nondisjunction**, a defect in chromosome segregation during meiosis (see below) in one of the two parents.

Key Questions

- The aneuploidies described above are essentially the only ones that result in live human births. Why do you think these aneuploidies are viable while aneuploidies of other chromosomes are not?

Endopolyploidy

Some tissues in an animal can contain cells that have more than two chromosome sets, whereas the somatic cells in the rest of the body are diploid. This situation is called **endopolyploidy**. For example, human liver cells can vary in euploidy (some cells are tetraploid or octaploid). Endopolyploidy allows liver cells to increase the production of protein products to meet the unique metabolic demands placed on liver cells. Moreover, the fruit fly *Drosophila* is a diploid organism containing four pairs of homologous chromosomes. However, the salivary glands of the fruit fly contain higher variations in euploidy. Endopolyploidy occurs when the homologous chromosomes pair with each other and then undergo several rounds of DNA replication without cell division. In fruit flies, DNA replication in this way produces **polytene chromosomes**, a thick bundle of identical DNA molecules, lying parallel to each other.

Key Questions

- What is endopolyploidy?
- Provide two examples of endopolyploidy.

Meiotic Nondisjunction

Nondisjunction occurs when chromosomes do not separate properly in either meiosis or mitosis. **Meiotic nondisjunction** produces aneuploid gamete cells that either have an extra chromosome or lack a chromosome (see **figure 3.13**). After fertilization, the resulting offspring will be either trisomic or monosomic. Meiotic nondisjunction can occur during anaphase of either meiosis I or meiosis II.

- **Meiosis I nondisjunction**. During meiosis I nondisjunction, the chromosomes within a homologous pair fail to separate from each other and instead segregate into the same daughter cell. All gamete cells produced from meiosis I nondisjunction are aneuploid. These aneuploid gametes will produce 50% trisomic offspring and 50% monosomic offspring.
- **Meiosis II nondisjunction**. In meiosis II nondisjunction, meiosis I proceeds normally; however, nondisjunction occurs in one of the two daughter cells. During meiosis II nondisjunction, the sister chromatids that constitute one duplicated chromosome fail to separate from each other. Two of the resulting haploid gametes are normal, while two of the gametes are aneuploid. Collectively, the four possible gametes will produce 50% normal offspring, 25% trisomic offspring, and 25% monosomic offspring.

The aneuploid human conditions described above (e.g., conventional Down syndrome, Klinefelter syndrome, Turner syndrome, etc.) are thought to be produced from either meiosis I or meiosis II nondisjunction.

On rare occasions, all chromosomes in a cell fail to separate properly during either meiosis I or II. This event is called **complete nondisjunction** and produces diploid gametes. If a diploid gamete fuses with a normal gamete, a triploid offspring is produced.

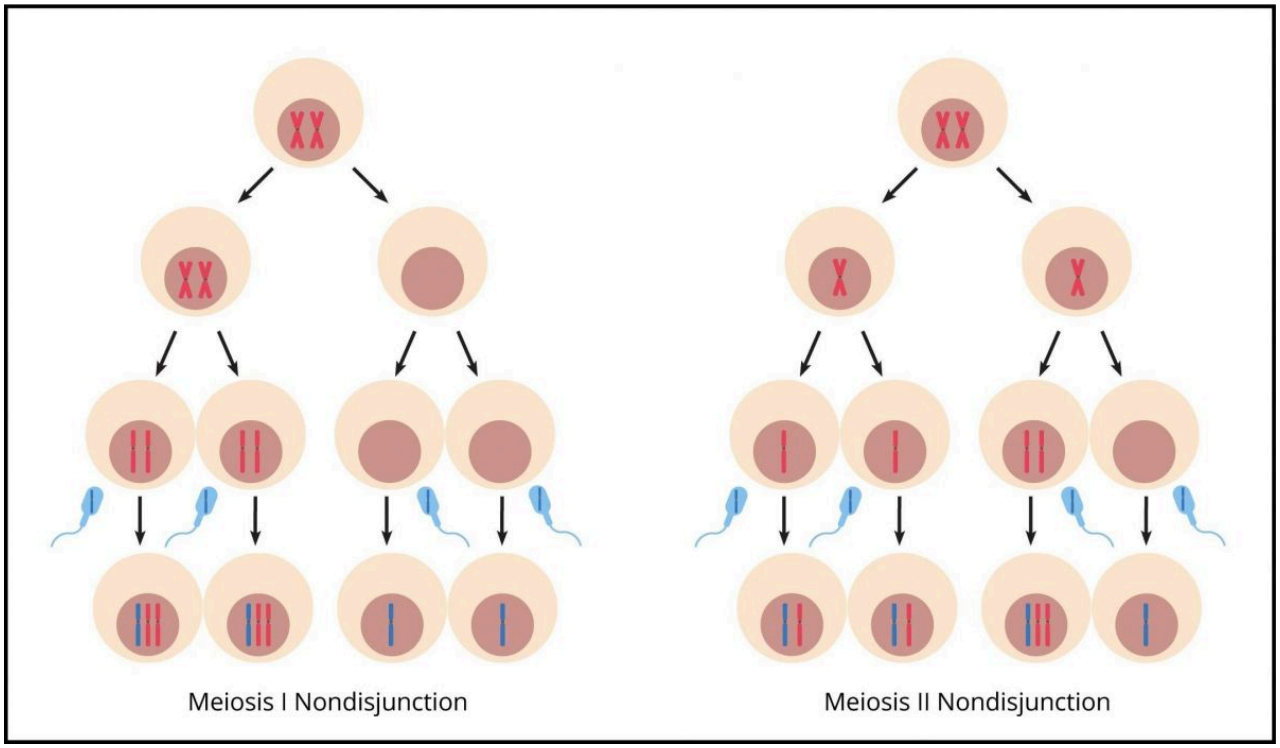


Figure 3.13 ***Meiotic Nondisjunction.*** *A homologous chromosome pair in the maternal parent experiences nondisjunction in meiosis I (left), while the sister chromatids fail to separate in the maternal parent during meiosis II nondisjunction (right). The paternal parent contributes a copy of the same chromosome (in blue). Meiosis I nondisjunction produces 100% aneuploid gametes and offspring. Meiosis II nondisjunction produces 50% aneuploid gametes and offspring.--- Image created by SL*

Key Questions

- What happens during meiosis I nondisjunction?
- Describe the four gametes produced by meiosis I nondisjunction.
- What happens during meiosis II nondisjunction?
- Describe the four gametes **produced** by meiosis II nondisjunction.

Mitotic Nondisjunction

Nondisjunction can also occur during mitosis (**mitotic nondisjunction**). During mitotic nondisjunction, the sister chromatids that constitute one duplicated chromosome fail to separate from each other during anaphase. Mitotic nondisjunction produces one daughter cell with three copies of a particular chromosome (trisomic), while the other daughter cell has one copy of the chromosome (monosomic) (see **figure 3.14; left panel**). All future daughter cells produced from the trisomic cell will also be trisomic, whereas all daughter cells produced from the monosomic cell will be monosomic. As a result, mitotic nondisjunction produces a **mosaic** phenotype with some trisomic tissues in the body, while other tissues are monosomic.

Sometimes chromosomes lose attachment to the mitotic spindle during anaphase (see **figure 3.14; right panel**). A detached chromosome is not retained in the nucleus, and is degraded by nucleases in the cytoplasm. This event produces one daughter cell with two copies of a particular chromosome (disomic), while the other daughter cell has

one copy of the chromosome (monosomic). All future daughter cells produced from the monosomic cell will be monosomic. This failure of a chromosome to attach to the mitotic spindle also results in a mosaic phenotype.

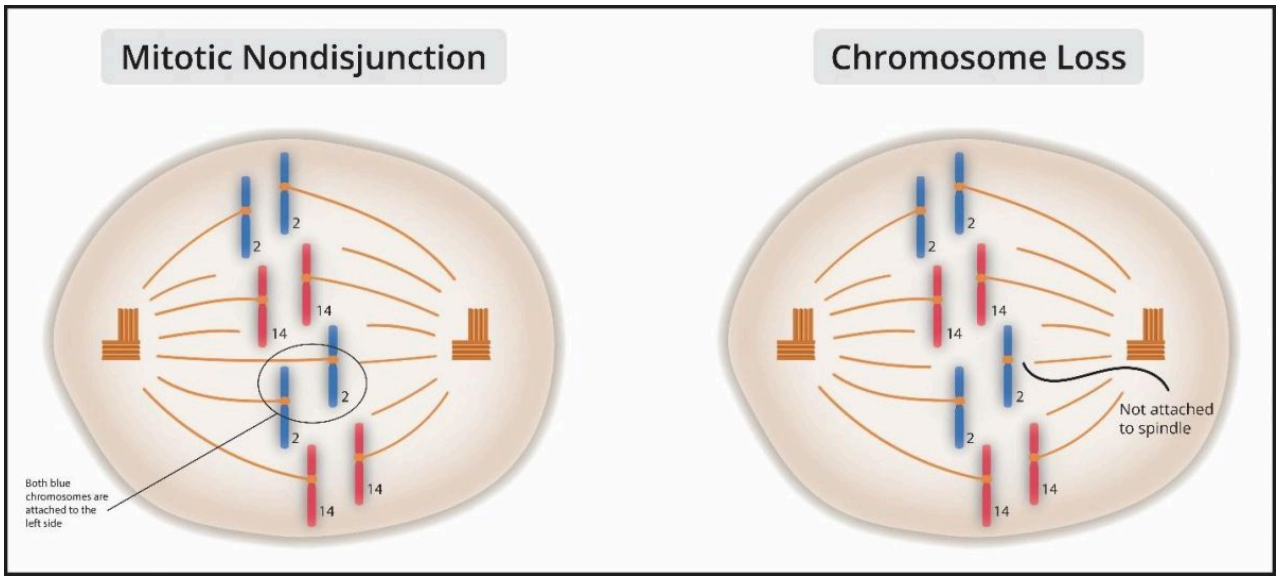


Figure 3.14 ***Mitotic Nondisjunction.*** *This figure highlights the fates of the sister chromatids derived from the two copies of chromosomes 2 and 14. --- Image created by SL*

Key Questions

- Describe the two processes that cause mitotic nondisjunction.

Review Questions

Fill in the Blanks:

1. A _____ is a change in chromosome structure that produces an acentric fragment and a dicentric chromosome during meiosis.
2. The disease _____ is caused by an inversion within the X chromosome.
3. A(n) _____ and a(n) _____ are two changes in chromosome structure that alter the amount of genetic information found on a chromosome.
4. _____ % of the gametes produced by a meiosis II nondisjunction event will result in trisomic offspring.
5. A terminal deletion in chromosome _____ causes _____, a disease that results in a malformation of the glottis and larynx.
6. Nondisjunction in _____ can produce trisomic and monosomic somatic cells.
7. The term _____ can describe a cell with *2n-2* chromosomes.
8. The globin gene family arose due to a _____ (name the chromosome mutation type).
9. A _____ translocation involves only the acrocentric chromosomes.
10. Reciprocal translocations of the _____ segregation type results in two normal gametes and two translocation gametes. All offspring produced from this event are normal.

4 - Gender Identity vs Biological Sex

Most of this reading assignment will be completed online. You may click the link below, access this online assignment in the Part 4 module in Canvas, or do a Google search using the keyword "HHMI Sex Verification Testing".

https://www.hhmi.org/biointeractive/testing-athletes?utm_source=BioInteractive+News&utm_campaign=63b97368c4-BioInteractive_News_Vol_109_2018_1_10_COPY_01&utm_medium=email&utm_term=0_98b2f5c6ba-63b97368c4-69814393

Please complete the four parts of the Click and Learn Assignment: Introduction, Human Development, Case Studies (Swimmer and Sprinter), Conclusion.

Sex Development in Mammals

Overview

In mammals, an individual's biological sex is partly determined by the sex chromosomes; females have two X chromosomes, while males have an X and a Y chromosome. Genes on these chromosomes are responsible for sex determination during embryonic development. Most noteworthy is the ***SRY*** gene, located on the *p* arm of the Y chromosome. In the early human embryo, the precursors of the sex organs are the same in both males and females. At approximately week six of embryonic development, the events that set-in motion male-specific gonad formation are initiated by the *SRY* gene. The protein product produced by the *SRY* gene is an activator for other male-specific genes, while acting as a repressor for female-specific genes. About two weeks later, embryos begin to take on either male or female specific phenotypes.

Because the X and Y chromosome form a homologous pair during meiosis I, synapsis and crossing over can move the *SRY* gene from the Y chromosome to the X chromosome. If the *SRY* gene moves to the X chromosome as the male parent produces gametes, upon fertilization an XX embryo can develop some male-specific traits. The same can also happen if there is a mutation in the *SRY* gene so that it no longer produces a functional protein. As a result, an XY embryo is unable to begin the key steps necessary for male sex differentiation. In both cases, the embryos grow without fully developing a male or female phenotype. The resulting condition is called a **disorder of sexual development (DSD)**. The overall occurrence of these DSDs is rare; about 1 in 5,500 children is born with ambiguous genitalia at birth. Imaging tests, such as an ultrasound, will usually show immature internal reproductive organs that are nonfunctional; therefore, leading to infertility.

Biological Sex Assignment

Despite a diagnosis of ambiguous genitalia, a **biological sex** (male or female) is assigned to the newborn infant, and the child is raised according to the determined biological sex. Diagnostic tests are performed to provide the physicians and the parents with as much information as possible to make a choice about the biological sex of the infant. Karyotype analysis can show which sex chromosomes are present, followed by physical examination of the external genitalia. Imaging tests show the internal structures of reproductive organs and tissues. Hormone tests are performed to examine sex hormone levels and DNA tests are done to identify sex-specific gene mutations. Once the physicians and the family make a choice as to the biological sex of the child, psychological support is offered to allow the child to grow

up as normal as possible. In some cases, reconstruction surgery can be performed to provide the child with external genitalia that fit the assigned sex. With support from family and friends, and the medical community, a child born with ambiguous genitalia can lead a happy life. However, there may arise new issues once the child reaches puberty. Secondary sex characteristics can fail to develop, or a child assigned to one sex at birth feels that they are the opposite sex.

Case Study in Sex Determination

E. Weil, "What If It's (Sort of) a Boy and (Sort of) a Girl?" The New York Times Magazine, September 24, 2006, pages 48-53.

Excerpts have been reprinted from this story.

When Brian Sullivan was born in New Jersey on August 14, 1956, doctors kept his mother, a Catholic housewife, sedated for three days until they could decide what to tell her. Brian was born with ambiguous genitalia. He spent the first 18 months of his life as a boy, until doctors performed exploratory surgery, found a uterus, and ovotestes (gonads containing both ovarian and testicular tissue) and told the parents that they'd made a mistake: Brian was actually a girl. Brian was renamed Bonnie, reconstructive surgery was performed to make her look more like a female on the outside and doctors counseled the family to throw away all pictures of Brian as a baby, move to a new town, and get on with their lives. The Sullivans did the best they could; they relocated, had three more children and did not speak of the circumstances around their oldest child's birth for many years. The doctors promised the parents that if they shielded Bonnie from her medical history, she would grow up normal, happy, heterosexual, and give them grandchildren.

Bonnie Sullivan spent most of her childhood and young-adult life extremely unhappy, feeling different from her peers, but unsure why. Around age 10, her parents told her that she had had an operation to remove her very large clitoris, but that everything turned out fine. At age 19, Bonnie started trying to access her medical records and succeeded when she was 22. She finally learned what happened to her as a baby. As a means of recovery from this startling news, Bonnie changed her name to Cheryl Chase, graduated from M.I.T. with a degree in math and then went on to study Japanese at Harvard. She threw herself into her work thinking that if she worked really hard, she would overcome her identity problems and finally be happy. After helping found a successful technology company in Japan, she realized that being happy was not going to happen until she found out the truth about who she was. She learned about the doctors' decision to give her reconstructive surgery to make her look more female, and why the medical community believes that surgery be done as early as possible based on their decision as to the sex of the child born with ambiguous genitalia.

Chase is now a leading activist about who has the right to decide what should be done with other people’s bodies. Is it the doctor, the parents, or the child who should decide biological sex? In 2004, she addressed the Human Rights Commission concerning the question of medical procedures of children born with ambiguous genitals. After the report was ratified, Chase commented, "What the Human Rights Commission has done is to recognize me as a human being. You've stated that just because I was born looking in a way that bothered other people doesn't mean that I should be excluded from human rights protections that are afforded to other people."

Disorders of sexual development are not the same as, say, a heart condition. Parents may feel entitled to make decisions based on the sense that they know what is right for their family members, and the reality is that in the case of these children, the right treatment for one child may not be the right treatment for all. These are not happy people, either. Some of them have isolated, difficult lives.

5 - Nucleic Acid Structure

In Part 5, we will first learn about the major experiments that established DNA as the genetic material of nearly all organisms (recall that we learned in Part 1 that some viruses use RNA as the genetic material). Then we will learn about the structure of both DNA and RNA.

A. DNA is the Genetic Material

Transformation

The physician Frederick Griffith was interested in developing a vaccine against the bacterium *Streptococcus pneumoniae*, one of the major causes of pneumonia, ear infections, and meningitis in children. Some strains (varieties) of *S. pneumoniae* produce a polysaccharide **capsule** that surrounds the bacterial cell wall. These encapsulated strains of *S. pneumoniae* are more virulent (disease causing) than strains that do not produce a capsule. Further, the encapsulated strains of *S. pneumoniae* form smooth colonies (called type S) on bacterial culture media, those strains without capsules form rough colonies (called type R) on culture media.

In 1928, Griffith showed the following (see **figure 5.1**):

- When smooth (type S) strains of *S. pneumoniae* were injected into mice, the mice died. The smooth strain of *S. pneumoniae* could be cultured from the blood of the dead mouse.
- When rough (type R) strains of *S. pneumoniae* were injected into mice, the mice survived. No bacteria were isolated from the blood of the living mouse.
- When smooth (type S) *S. pneumoniae* were heat-killed to lyse the bacterial cells, and the bacterial extract was then injected into mice, the mice survive. No bacteria were isolated from the blood of the living mouse.
- When live rough (type R) bacteria were mixed with heat-killed smooth (type S) bacteria and injected into mice, the mice died. Note that although neither the rough (type R) bacteria nor the heat-killed smooth (type S) bacteria killed mice on their own, the mixture of the two killed the mice. The living bacteria isolated from the blood of the dead mice were smooth (type S) bacteria.

In this final experiment, Griffith reasoned that some chemical released from the heat-killed smooth (type S) bacteria was internalized by the rough (type R) bacteria. This chemical could change phenotype, converting the rough (type R) bacteria into smooth (type S) bacteria. This change in phenotype is called **transformation**. The transformed bacteria then passed the type S trait to their progeny. As a result, the chemical responsible for transformation (the **transforming principle**) had properties of the genetic material. The transforming principle changed the phenotype of the bacterial cells and is inherited when the bacterial cell divides. Unfortunately, Griffith did not identify the chemical responsible for transformation.

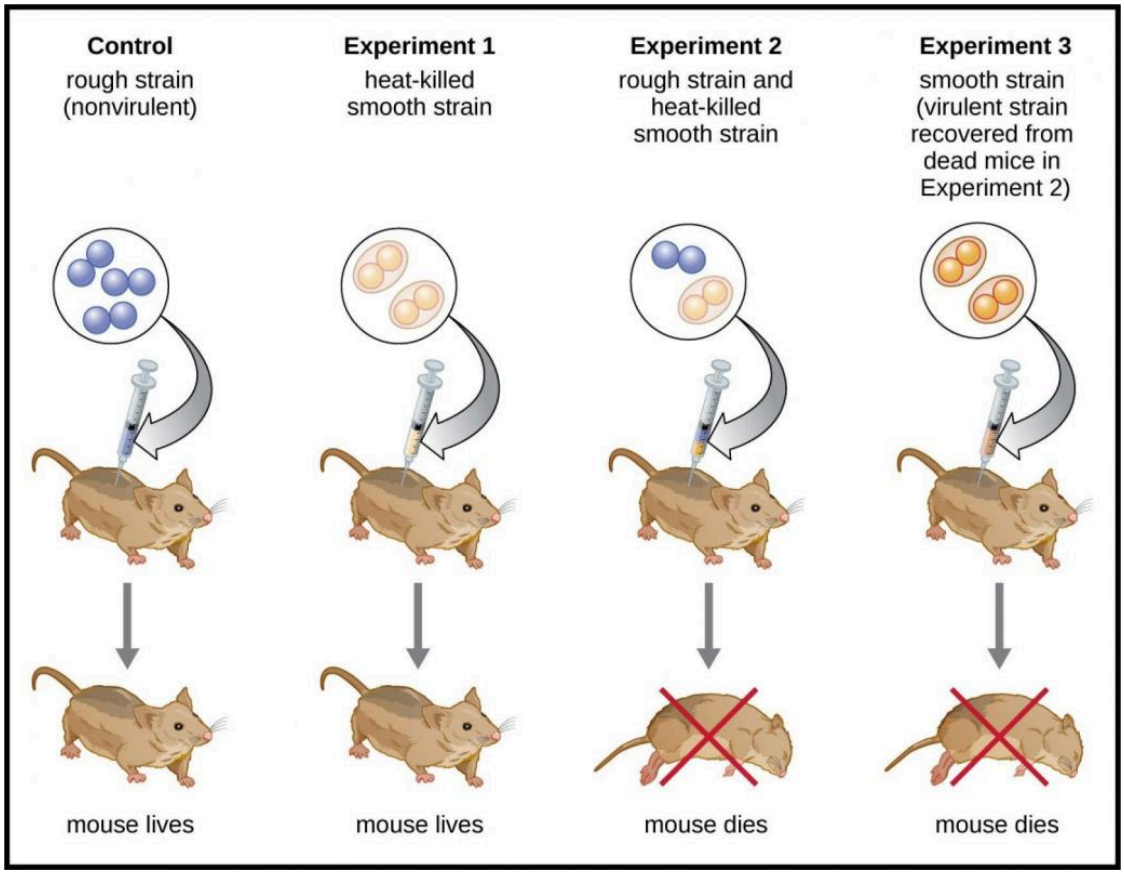


Figure 5.1 ***The Griffith Experiment ---*** *image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction)*

Key Questions

- Describe the Griffith experiment.
- What is meant by the “transforming principle”?
- When smooth (type S) bacteria were "heat-killed", what major class of molecules was denatured (destroyed)? What major class of molecules was not denatured?

DNA is the Transforming Principle

Oswald Avery, Maclyn McCarty, and Colin MacLeod wanted to identify the chemical responsible for transforming rough (type R) into smooth (type S) bacteria in Griffith's experiment. Avery and his colleagues focused on three candidate chemicals: DNA, RNA, and protein (see **figure 5.2**). Avery, McCarty, and MacLeod performed three experiments:

- In one experiment, rough (type R) bacterial cells were mixed with bacterial extracts from heat-killed smooth (type S) cells in the presence of **ribonuclease (RNase)** to digest RNA. When RNA was eliminated from the bacterial extract from type S cells, the rough bacteria were still transformed into smooth bacteria. Thus, digestion of RNA had no effect on transformation.
- In a second experiment, rough (type R) bacterial cells were mixed with bacterial extracts from heat-killed smooth (type S) cells in the presence of **protease** to digest proteins. When protein was eliminated from the bacterial extract from type S cells, the rough bacteria were still transformed into smooth bacteria. Thus, digestion of protein had no effect on transformation.
- In a final experiment, rough (type R) bacterial cells were mixed with bacterial extracts from heat-killed smooth (type S) cells in the presence of **deoxyribonuclease (DNase)** to digest DNA. Interestingly, the rough bacteria were not transformed into smooth bacteria when DNA was eliminated from the bacterial extract. Thus, the digestion of DNA prevented transformation.

Avery and colleagues concluded from these experiments that DNA (not RNA nor protein) was the chemical responsible for transforming rough bacteria into smooth bacteria in Griffith's experiment. Therefore, DNA is the genetic material of the bacterium *S. pneumoniae*.

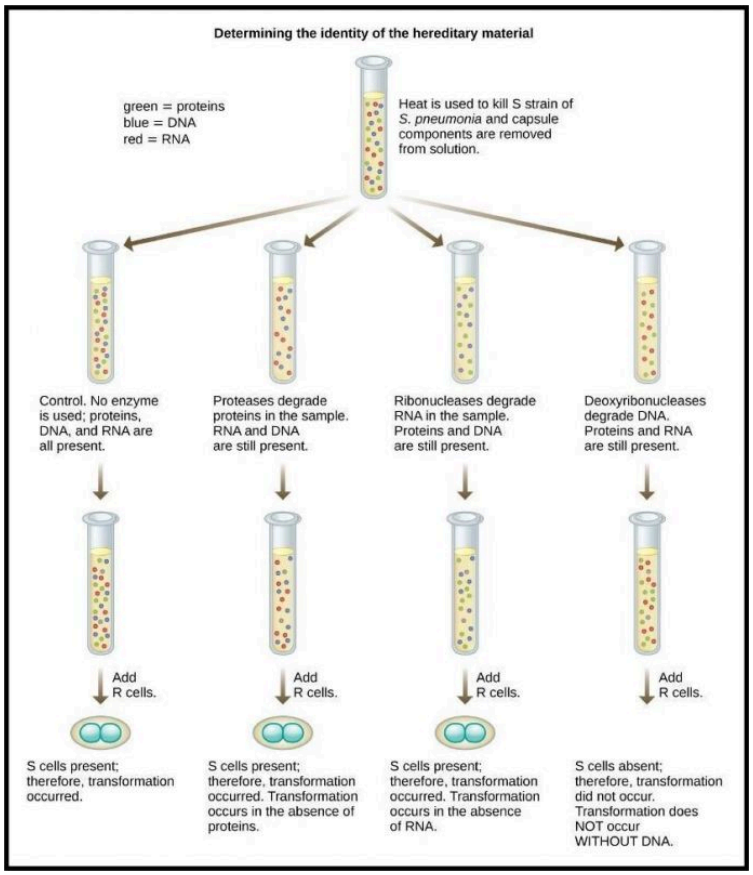


Figure 5.2 ***Avery, McCarty, MacLeod Experiment ---*** *image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction)*

Key Questions

- Describe the Avery, McCarty, and MacLeod experiment.
- What was the key finding of the experiment?

Bacteriophage T2

To confirm the results from Avery, McCarty, and MacLeod, Alfred Hershey and Martha Chase examined **T2 bacteriophage.** T2 bacteriophage is a type of virus that infects the bacterium *Escherichia coli* (see **figure 5.3**). T2 bacteriophages contain two molecular components: DNA and protein. The DNA of the bacteriophage is encased within a bacteriophage head structure made of protein. T2 bacteriophage also contains other protein structures, including a tube-like sheath, tail fibers, and a base plate. During the bacteriophage life cycle, T2 bacteriophage acts like a hypodermic needle, injecting the bacteriophage genetic material into a host *E. coli* cell. The genetic material of the bacteriophage then reprograms the host *E. coli* cell to shut off many host cell functions and instead produce progeny T2 bacteriophages. Hershey and Chase were interested in determining which of the two bacteriophage T2 components, DNA or protein, was responsible for producing progeny bacteriophage particles. In essence, Hershey and Chase were asking whether DNA or protein is the genetic material of bacteriophage T2.

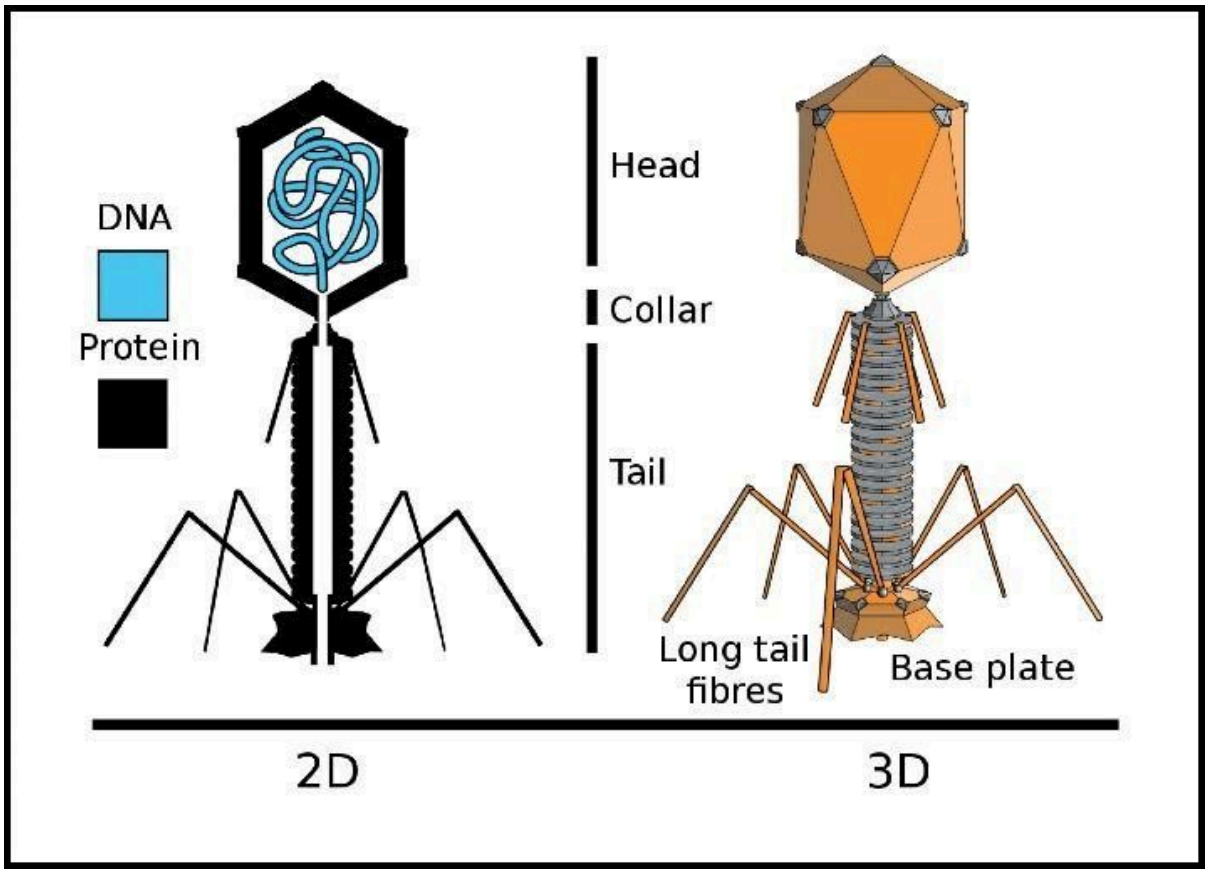


Figure 5.3 ***A) Typical Bacteriophage.*** *The bacteriophage was imaged in a scanning electron microscope.* ***---*** *licensed under CC BY 4.0* ***B) Bacteriophage Structure ---*** *Tevenphage by Adenosine licensed under CC BY-SA 2.5*

The Hershey-Chase Experiment

The Hershey and Chase experiment relied on two important experimental details:

- A kitchen blender can separate the bacteriophage components that remain attached to the surface of the host bacterial cell from the bacteriophage components that are injected into the cytoplasm of the *E. coli* cell.
- Bacteriophage proteins can be distinguished from bacteriophage DNA using radioactive labeling. Bacteriophage proteins were radiolabeled with ^{35}S (a radioactive isotope of sulfur) and the bacteriophage DNA was radiolabeled with ^{32}P (a radioactive isotope of phosphorus). Note that sulfur is found in proteins and not in DNA; phosphorus is a component of DNA and is not found in proteins. Thus, Hershey and Chase could determine whether DNA or protein is the bacteriophage T2 genetic material by determining whether ^{35}S or ^{32}P is injected into the host *E. coli* cell during a bacteriophage infection.

The Hershey and Chase experiment was done as follows (see **figure 5.4**):

1. In one experiment, bacteriophage T2 proteins were radiolabeled with ^{35}S. In another experiment, bacteriophage T2 DNA was radiolabeled with ^{32}P.
2. The radiolabeled bacteriophages were mixed in two separate reactions with *E. coli* cells to allow bacteriophage infections to occur.
3. After bacteriohage T2 injected its genetic material, the reactions were subjected to blending. During blending, the bacteriophage components that remained on the surfaces of the *E. coli* cells were released.
4. The infected *E. coli* were collected in a centrifuge. The empty bacteriophage components (phage head, tail, tail fibers) remain in the supernatant (liquid) after centrifugation. The *E. coli* cells and the bacteriophage genetic material are found in a pellet at the bottom of the centrifuge tube.
5. The amount of radioactivity in the supernatant and pellet was calculated.

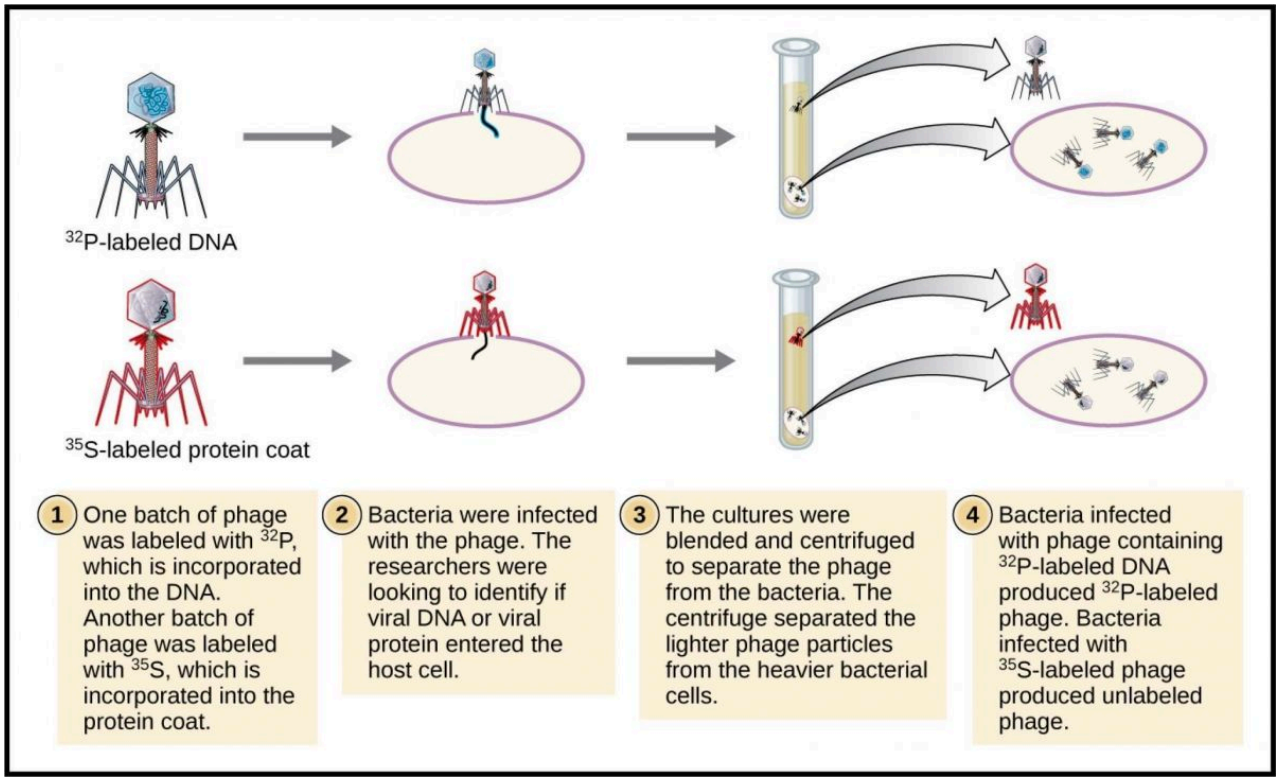


Figure 5.4 ***Hershey - Chase Experiment ---*** *image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction)*

Blending removed most of the ^{35}S from the *E. coli* cells in the pellet. Thus, proteins are not injected into *E. coli* to direct the formation of progeny T2 bacteriophages. In contrast, most of the ^{32}P was found in the host *E. coli* cells in the pellet after blending and centrifugation, indicating that DNA is injected into host *E. coli* cells. Note that after the completion of the bacteriophage life cycle, the progeny bacteriophages also contained ^{32}P; the progeny bacteriophages contained little ^{35}S. Thus, the ^{32}P labeled DNA is heritable. The Hershey and Chase experiment showed that DNA serves as the genetic material for bacteriophage T2. The work of Avery, McCarty, and MacLeod combined with the results of the Hershey-Chase experiment provided compelling evidence that DNA is the genetic material of viruses and bacteria.

Key Questions

- Describe the Hershey-Chase experiment.
- What was the key finding of the experiment?

DNA is the Genetic Material of Eukaryotes

Eukaryotic cells are not as easy to work with in the lab as bacteria and bacteriophages. As a result, it was easier to determine that DNA is the genetic material of bacteria and bacteriophages than eukaryotic cells. Evidence that DNA is the genetic material of eukaryotes relied on the combination of both indirect evidence and direct evidence.

Several lines of indirect evidence suggest that DNA is the genetic material of eukaryotic cells. Scientists reasoned that the genetic material of eukaryotes should be found within chromosomes because chromosomes are copied and distributed to daughter cells during mitosis and meiosis. Chromosomes contain both proteins and DNA; however, the DNA component is found exclusively in chromosomes (protein is found in the cell cytoplasm as well). In addition, a

diploid cell, which contains twice as many chromosomes as a haploid cell, also contains roughly twice as much DNA as a haploid cell. No such correlation was observed when the protein content of haploid and diploid cells was compared. Finally, **ultraviolet light (UV light)** causes mutations that affect the phenotype of a cell. The wavelength of UV light that produces the highest frequency of mutations corresponds to the wavelength of UV light that is absorbed most strongly by DNA. On the other hand, the wavelength of UV light absorbed most strongly by proteins does not alter phenotpye.

Recombinant DNA technology provided direct evidence that DNA is the genetic material of eukaryotic cells. In this technique, a DNA sequence from a eukaryotic cell is isolated and then introduced into a bacterial cell. This eukaryotic DNA sequence can then be transcribed by the bacterial cell to make a messenger RNA (mRNA); the mRNA is then translated by bacterial ribosomes to make a protein. The resulting protein often changes the phenotype of the bacterial cell. Moreover, the introduced eukaryotic DNA sequence is passed on to the progeny bacterial cells during bacterial cell division. As an example, recombinant DNA technology allowed scientists to insert the human insulin gene into bacteria. These bacterial cells then produce the human insulin protein (change in phenotype) and transfer the human insulin gene to daughter cells after cell division (inherited). The fact that introduced eukaryotic gene can result in protein production, can alter the phenotype of a bacterial cell, and can be passed to progeny bacterial cells provided strong evidence that DNA is the genetic material of eukaryotes.

Key Questions

- What are the three lines of indirect evidence that DNA is the genetic material of eukaryotes?
- What direct evidence shows that DNA is the genetic material of eukaryotes?

B. The Structure of DNA and RNA

Overview of Nucleic Acid Structure

Nucleic acid molecules (DNA and RNA) have four levels of structural complexity (see **figure 5.5**):

- **Nucleotide**. Nucleotides are the basic subunits (the building blocks) of nucleic acid molecules.
- **Nucleic acid strand**. A nucleic acid strand is a chain of nucleotides covalently linked together via **phosphodiester** bonds.
- **Double helix**. Two nucleic acid strands of either DNA or RNA can hydrogen bond together to form a double helix structure.
- **Chromosomes**. DNA molecules associate with NAP proteins or histone proteins to form prokaryotic and eukaryotic chromosomes. We discussed the structures of prokaryotic and eukaryotic chromosomes in Part 2.

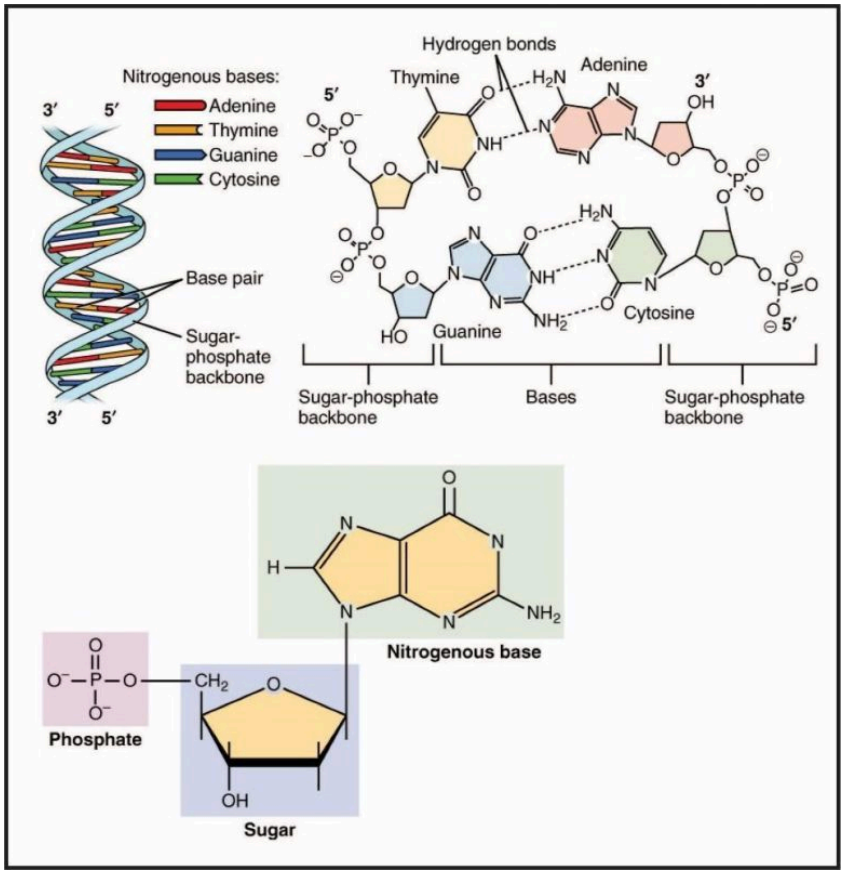


Figure 5.5 ***Overview of DNA Structure.*** *The basic subunits of DNA and RNA are nucleotides (bottom). Two nucleic acid strands can interact through hydrogen bonding (upper right). The general structure of a DNA double helix (upper left). This image is from OpenStax (access for free at* https://openstax.org/books/biology-2e/pages/1-introduction*)*

Nucleotides

A nucleotide is composed of the following molecular components (see **figure 5.6**):

- **One, two, or three phosphate groups**. When the cell synthesizes DNA and RNA strands, the nucleotides used as substrates contain three phosphate groups. However, When a nucleotide is incorporated into a DNA or RNA strand, two of the phosphate groups are released.
- **A pentose sugar**. In DNA, the pentose sugar is **deoxyribose**; in RNA, the pentose sugar is **ribose**. Deoxyribose and ribose are distinguished by the chemical groups attached to the carbon atoms within the pentose sugar. Specifically, the carbon atoms within both deoxyribose and ribose are numbered 1' (one-prime) to 5' (five prime) (**see figure 5.6**). The nitrogenous base (see below) in both deoxyribose and ribose is attached to 1' carbon. The 2' carbon in deoxyribose is attached to two hydrogen atoms, the 2' carbon in ribose is attached to a hydroxyl group and a hydrogen atom. As a result, one of the fundamental differences between a DNA nucleotide and an RNA nucleotide is the chemical group attached to the 2' carbon. The 3' carbon in both deoxyribose and ribose is attached to a hydroxyl group. It is important to note that the hydroxyl group attached to the 3' carbon plays a critical role in DNA and RNA synthesis (see Parts 6 and 9). The 4' carbon in both deoxyribose and ribose is involved in maintaining the integrity of the pentose ring structure. The 5' carbon in both deoxyribose and ribose is attached to a phosphate group. This phosphate group also plays an important role in DNA and RNA synthesis (see Parts 6 and 9).
- **A nitrogenous base**. There are two types of nitrogenous bases found in DNA and RNA nucleotides (see figure 5.6):
  - **Pyrimidines**. The pyrimidine nitrogenous bases consist of a single carbon/nitrogen ring structure. The pyrimidine bases are **cytosine (C)**, **thymine (T)**, and **uracil (U)**. T is found only in DNA; U is found only in RNA. C is found in both DNA and RNA.
  - **Purines**. The purine nitrogenous bases consist of a double carbon/nitrogen ring structure. The purines bases include **adenine (A)** and **guanine (G)**. A and G are found in both DNA and RNA.

The nucleotide substrates used to build DNA and RNA are called **deoxynucleoside triphosphates (dNTPs)** and **nucleoside triphosphates (NTPs)**, respectively. In the dNTP/NTP nomenclature "N" is the specific name of the nitrogenous base (A,U,C,G, and T); TP refers to triphosphate. For example, dTTP contains deoxyribose as the pentose sugar; deoxyribose is attached to thymine (1' carbon) with three phosphate groups attached to the 5' carbon. GTP contains ribose as the pentose sugar; ribose is attached to guanine (1' carbon) with three phosphate groups attached to the 5' carbon.

Key Questions

- What are the functions of the 1', 2', 3', 4' and 5' carbons in deoxyribose and ribose?
- How are purines and pyrimidines different?
- Which purines and pyrimidines are found in DNA?
- Which purines and pyrimidines are found in RNA?
- What is meant by "dNTP" and "NTP"?

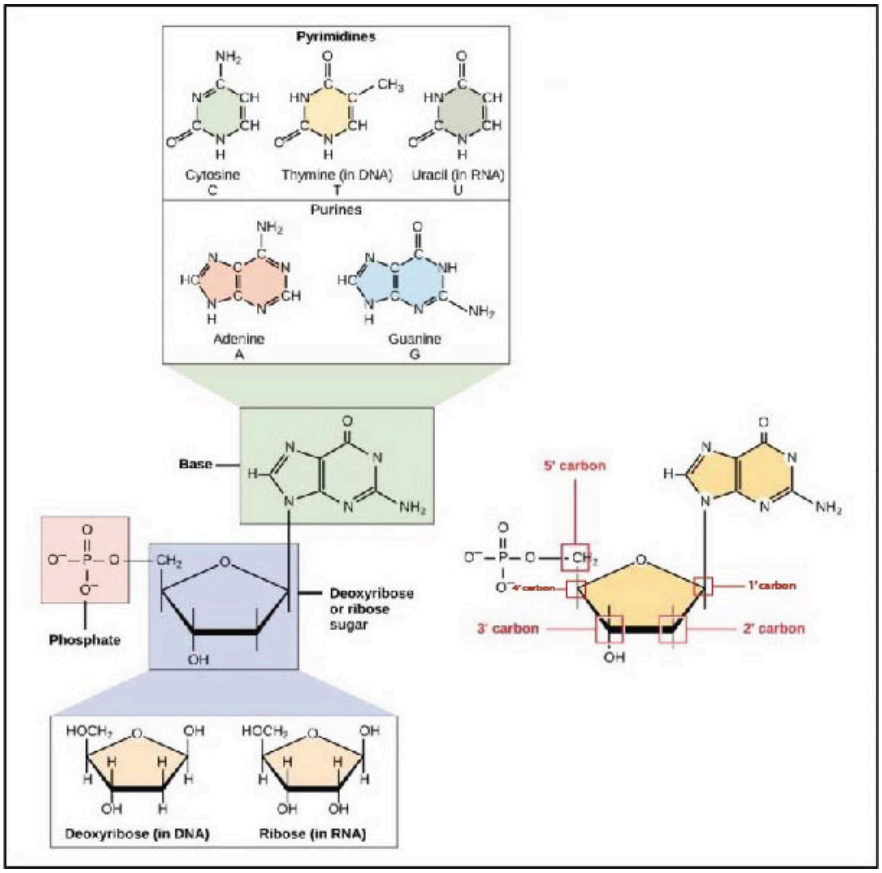


Figure 5.6 ***Nucleotide Structure*** *--- image used from OpenStax (access for free at* https://openstax.org/books/biology-2e/pages/1-introduction*)*

Nucleic Acid Strands

Nucleotides are covalently linked to form a nucleic acid **strand** (see **figure 5.7**). Specifically, the sugars of two adjacent nucleotides are linked together by **phosphodiester bonds** to form the **backbone** of the nucleic acid strand. The phosphate groups in the backbone give the nucleic acid strand a negative electrical charge.

A strand of DNA or RNA has **directionality** or **polarity**. The formation of phosphodiester bonds between nucleotides produces a nucleic acid strand in which the 5′ carbon of the nucleotide at one end of the strand contains a free phosphate group (the **5′ end** of the strand). The 3′ carbon of the nucleotide at the other end of the nucleic acid strand (the **3′ end** of the strand) is attached to a free hydroxyl group.

Each nucleic acid strand has a particular sequence of nitrogenous bases. The sequence of these nitrogenous bases on a single nucleic acid strand is written from the free 5′ end to the free 3′ end, for example 5′-TTGCAGG-3′. The sequence of the nitrogenous bases allows nucleic acid molecules to carry genetic information.

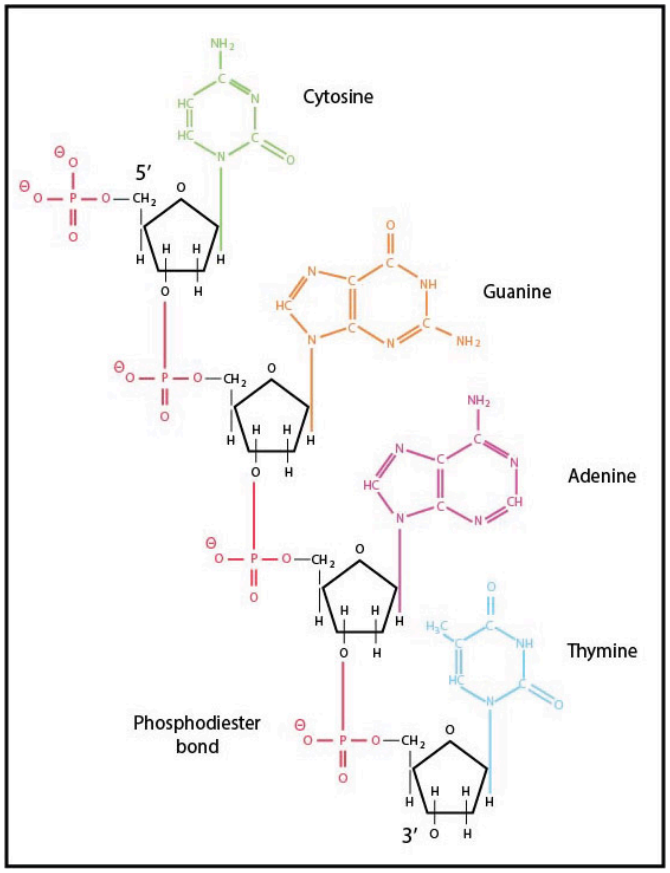


Figure 5.7 ***Structure of a DNA Strand.*** *The sequence of this nucleic acid strand is 5′-CGAT-3′. --- Image created by SL*

Key Questions

- How are adjacent nucleotides linked within a nucleic acid strand?
- How do you know which end of the nucleic acid strand is the 5′ end and which end is the 3′ end?

Determining the Structure of the Double Helix

James Watson and Francis Crick were awarded the Nobel Prize in 1962 for determining the structure of DNA; however, their work was built upon the contributions of other noteworthy scientists (see **figure 5.8**). For example, in the early 1950s, Linus Pauling, who won two Nobel Prizes himself, demonstrated that ball-and-stick models could describe the locations of individual atoms within biological molecules. Pauling is known for using ball-and-stick models to describe the secondary structures found within proteins. Watson and Crick mimicked Pauling's approach by building a ball-and-stick model of the DNA double helix.

The structure of the DNA double helix could not have been determined without the **X-ray diffraction** technique. X-ray diffraction involves subjecting a substance, such as DNA, to X-rays. When the X-rays pass through the DNA, the atoms within the DNA diffract the X-rays to produce a unique pattern on photographic film. This pattern can be interpreted using mathematics to determine the location of every atom within DNA. Rosalind Franklin and Maurice Wilkins used X-ray diffraction to determine that DNA has a helical structure and a diameter of 2 nanometers (2 nm), suggestive of two nucleic acid strands.

Moreover, Erwin Chargaff isolated DNA from many different organisms (bacteria, yeast, chickens, and humans) and then studied the nitrogenous base composition of this isolated DNA. Chargaff found that the total percentage of adenine (A) within any DNA molecule was nearly identical to the total percentage of thymine (T). Likewise, the percentage of cytosine (C) was nearly identical to the total percentage of guanine (G). These relationships are known as **Chargaff's rule**.

Watson and Crick used the observations/approaches of Pauling, Franklin, Wilkins, and Chargaff to build a ball-and-stick DNA model with the following features:

- The phosphate-sugar backbones of the two nucleic acid strands are on the outside of the DNA molecule.
- The nitrogenous bases are on the inside of the DNA molecule.
- The two nucleic acid strands are antiparallel (see below).
- Adenine forms two hydrogen bonds with thymine within the center of the DNA molecule; guanine forms three hydrogen bonds with cytosine. These hydrogen bonding interactions between nitrogenous bases are called **base pairing**.

Watson and Crick also proposed a mechanism by which the DNA double helix could be copied prior to cell division (**semi-conservative replication**). We will examine the process of semi-conservative replication in Part 6.

Key Questions

- What was the contribution of Pauling, Franklin and Wilkins, and Chargaff to the DNA story?
- Describe Watson and Crick's model of the DNA double helix.

The DNA Double Helix

Watson and Crick showed that the two phosphate-sugar backbones within DNA are found on the outside of the DNA molecule, directly exposed to water within the cell (see **figure 5.9**). Also, hydrogen bonds are formed between pairs of nitrogenous bases, called **base pairs (bp)**, located in the interior of the double-helix. Adenine always forms two hydrogen bonds with thymine, and guanine always forms three hydrogen bonds with cytosine. This relationship between nitrogenous bases is called the **AT/GC rule** or **Chargaff's rule**. Thus, the nitrogenous bases in one strand of DNA are **complementary** to the nitrogenous bases in the other DNA strand. Because of these hydrogen bonding interactions, DNA sequences with a higher proportion of GC base pairs are more stable than DNA molecules that are rich in AT base pairs. Additionally, there are 10 base pairs per complete turn of the DNA double helix. Every turn is 3.4 nm in length, meaning that each base pair within the DNA double helix is separated by 0.34 nm. The DNA double helix is 2 nm wide.

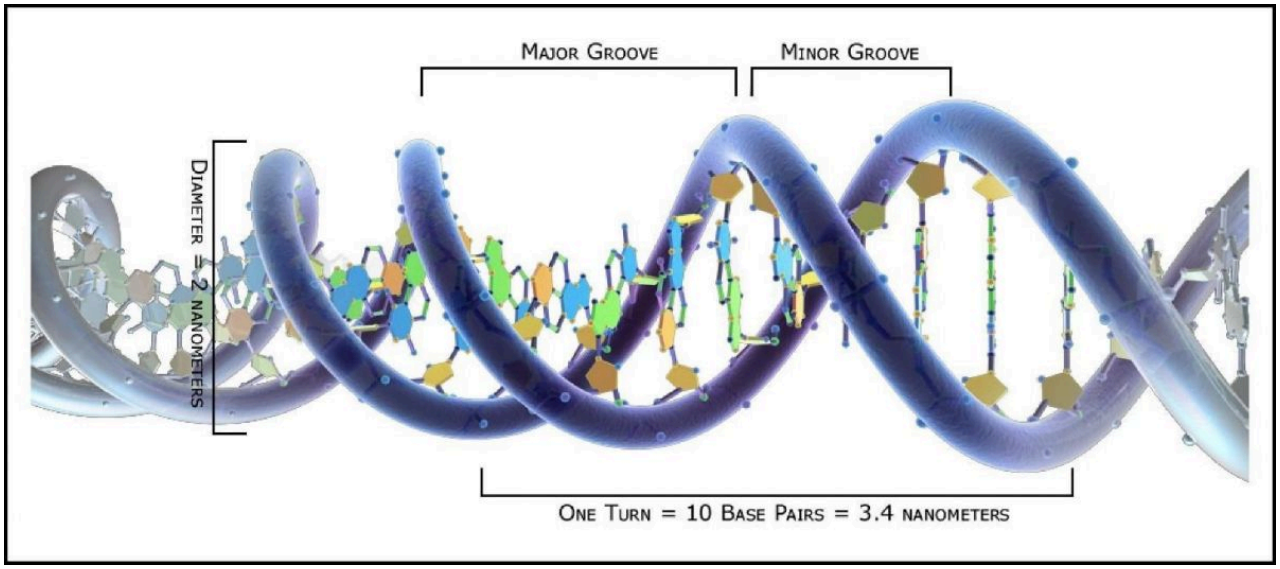


Figure 5.9 ***DNA Double Helix Structure.*** *The backbones of each DNA strand are represented by blue ribbons. [3D Science DNA Structure](#) by [3DScience.com](#) used under license [CC BY 2.5](#)*

The two nucleic acid strands of DNA are **antiparallel (**see **figure 5.10)**. One DNA strand starts with the free 5' phosphate group at the top of the DNA strand and ends with the free 3' hydroxyl group at the bottom. The other DNA strand runs in the opposite direction; the free 5' phosphate group is at the bottom end of the DNA strand, the free 3' hydroxyl group is at the top.

You can predict the sequence of one strand of DNA if you know the sequence of the other DNA strand. For example, if one DNA strand is 5'-GCCATG-3', then the opposite DNA strand is 3'-CGGTAC-5'. As a result, the two DNA strands are said to be **complementary** (see **figure 5.10**).

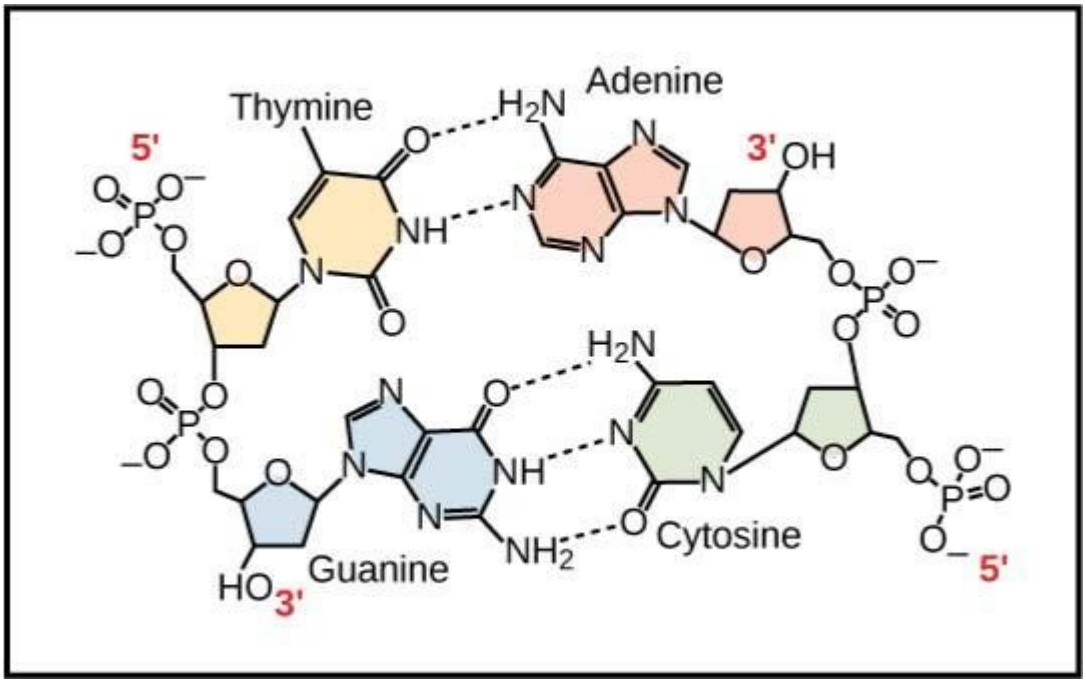


Figure 5.10 ***Complementary Base Pairs*** *--- Image used from OpenStax (access for free at [https://openstax.org/books/biology-2e/pages/1-introduction](#))*

Key Questions

- What is the distance between adjacent base pairs within DNA?
- How many base pairs are found per helical turn in DNA?
- How does DNA strand polarity relate to the antiparallel structure of DNA?

Other Features of DNA

The backbone of the DNA double helix is right-handed, meaning that the backbone turns in the clockwise direction as you look down the axis of the DNA molecule. Within the central part of the double helix, the nitrogenous bases form hydrogen bonds (A with T; G with C). The base pairs themselves are flat (planar) and stack on top of each other, much like the stairs of a spiral staircase. On the outside of the DNA double helix, there are **major** and **minor** grooves where the nitrogenous bases are directly exposed (see **figure 5.9** above). Proteins that bind to specific base pair sequences within the DNA double-helix interact mainly with the major groove and to a lesser extent with the minor groove. These proteins bind to the DNA double helix to control DNA replication and transcription.

Key Questions

- How do DNA-binding proteins recognize the nitrogenous bases in DNA?

Alternative Forms of DNA

The structure of the DNA double helix detailed above is the standard form of DNA **(B DNA)**. B DNA is the predominant form of DNA that is found in aqueous environments, including living cells. Interestingly, there are at least two alternative forms of DNA, called **A DNA** and **Z DNA** (see **figure 5.11**).

The A form of DNA is produced in the laboratory under high salt (low water) conditions. The A DNA structure is more compact than B DNA, has 11 base pairs (bp) per turn of the helix, and is 2.3 nm wide. Like B DNA, A DNA is a right-handed double helix; however, while the base pairs are perpendicular to the axis of the double helix in B DNA, the base pairs in A DNA are tilted relative to the axis of the molecule.

The Z form of DNA is a left-handed double helix, having a more extended structure. Z DNA has 12 base pairs per turn of the helix, and is only 1.8 nm wide. Z DNA is favored over B DNA when cytosine bases are modified by the addition of **methyl** ($-CH_3$) chemical groups. The functional significance of Z DNA is unknown; however, Z DNA may form in certain circumstances within living cells. Scientists speculate that Z DNA formation near a particular gene may influence whether the gene can be activated by transcription to produce an RNA molecule.

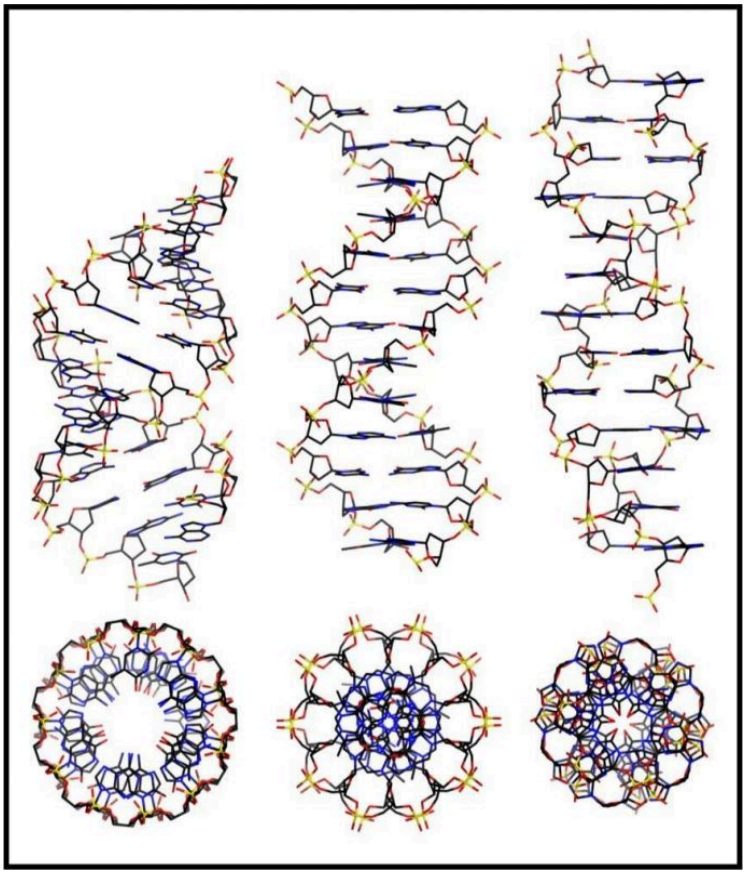


Figure 5.11 ***Left) A-DNA, Middle) B-DNA, Right) Z-DNA*** *---* *DnaConformations by Mauroesguerroto licensed under CC BY-SA 4.0*

Key Questions

- How do A-DNA, B-DNA, and Z-DNA differ from one another?

RNA Structure

Nucleic acid strands composed of RNA nucleotides have a similar structure as DNA strands. For example, RNA strands have a backbone composed of negatively charged phosphate groups, adjacent nucleotides within an RNA strand are linked by phosphodiester bonds, and the RNA strand has polarity (5' and 3' ends). However, RNA differs from DNA in that RNA molecules contain the pentose sugar ribose, RNA molecules are often single-stranded, and RNA strands are much shorter than DNA strands.

Since many RNA molecules are single-stranded, there is the possibility that nitrogenous bases in one part of an RNA molecule can form base pairs via hydrogen bonding with nitrogenous bases in another part of the same RNA molecule. These base pairing interactions form short regions of double-stranded RNA. One important RNA structure formed in this way is a **stem-loop (hairpin loop)**. For example, **transfer RNA (tRNA)** molecules, which play a critical role in the translation process (see Part 11), are noteworthy because each tRNA molecule contains three stem-loop structures (see **figure 5.12**).

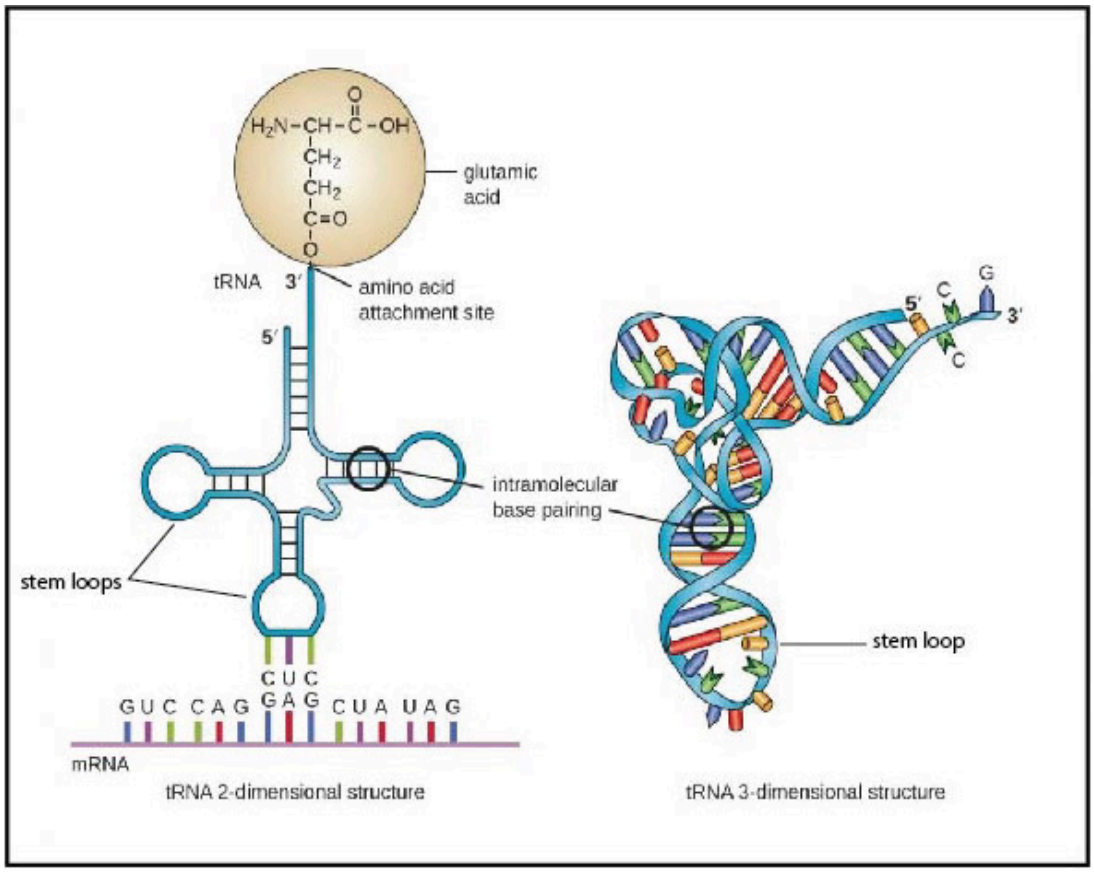


Figure 5.12 ***tRNA Structure*** *--- Image used from OpenStax (access for free at* https://openstax.org/books/biology-2e/pages/1-introduction*)*

Key Questions

- What are the structural differences between DNA and RNA?

Review Questions

Fill in the Blank:

1. In the Griffith experiment, the _____ strain of bacteria kills the mouse.
2. The bacterium _____ was used in the Hershey-Chase experiment, while the bacterium _____ was examined in the Frederick Griffith experiment.
3. Avery, McCarty, and MacLeod found that the enzyme _____ destroyed the transforming principle in bacteria, whereas the enzymes _____ and _____ did not.
4. The _____ experiment showed that DNA is the genetic material of bacteriophage T2.
5. In eukaryotes, a _____ cell has twice the amount of DNA as a _____ cell, but both cells have similar amounts of protein.
6. Each base pair within the DNA double helix is separated by _____ nanometers (nm).
7. The two purines found in RNA are _____ and _____.
8. The pentose sugar in dCTP is _____, while the pentose sugar in ATP is _____.
9. Proteins bind to DNA primarily at the _____ groove.
10. The _____ form of DNA exists in aqueous environments while the _____ form of DNA exists when cytosines are methylated.
11. The _____ experiment is sometimes called the “blender experiment”.

6 - DNA Replication

When James Watson and Francis Crick determined the structure of the DNA double helix, they noticed that the structure provided clues to how DNA is copied prior to cell division. This copying process is called **DNA replication** (see **figure 6.1**).

Overview of DNA Replication

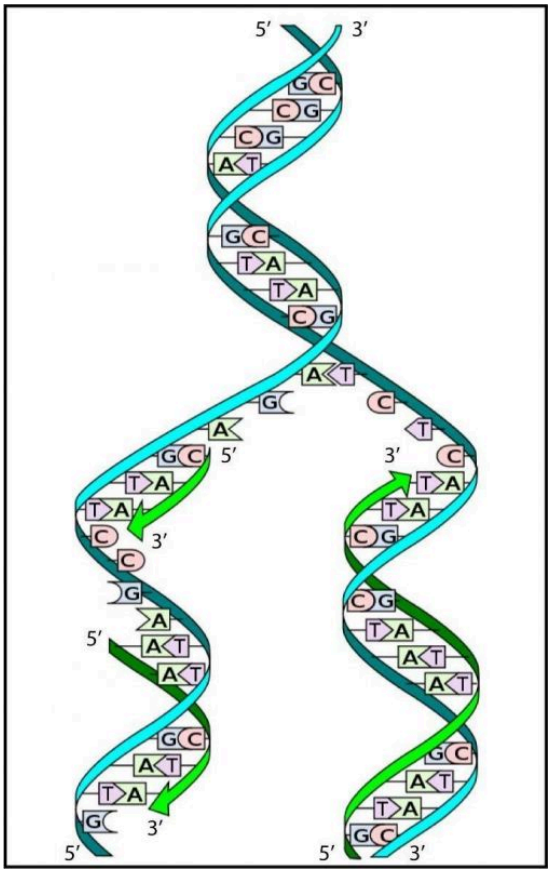


Figure 6.1 ***Overview of DNA Replication*** *--- Image created by SL*

Watson and Crick proposed that during DNA replication, the two original DNA strands within the double helix separate, and two new strands of DNA are synthesized. The two original DNA strands are called **template DNA strands** or **parental DNA strands**; each of the newly synthesized DNA strands is called a **daughter DNA strand**.

When DNA nucleotides (deoxyribonucleoside triphosphates or **dNTPs**) are used to generate the daughter DNA strands, the AT/GC rule is followed. Hydrogen bonds are formed between the nitrogenous bases within the incoming nucleotides and the template strand nitrogenous bases. Then a phosphodiester bond is formed between the free 5' phosphate on the incoming nucleotide and the free 3' hydroxyl group on the growing daughter DNA strand. The dNTPs used as the

substrates for DNA synthesis include **deoxyadenosine triphosphate (dATP)**, **deoxythymidine triphosphate (dTTP)**, **deoxycytidine triphosphate (dCTP)**, and **deoxyguanosine triphosphate (dGTP)**.

Key Questions

- What is a template (parental) DNA strand?
- What is a daughter DNA strand?
- What are the four dNTPs used in DNA replication?

A. DNA Replication in Bacteria

Origin of Replication in Bacteria

The site on the bacterial chromosome where DNA replication begins is the **origin of replication** (see **figure 6.2**). The bacterium *E. coli* has a single origin of replication called ***OriC***. *OriC* is a 275 base pair (bp)-long region that contains important DNA sequences, including:

- **AT-rich sequences**. These AT-rich sequences are significant as only two hydrogen bonds hold AT base pairs together in DNA. Less energy is required to separate AT-rich DNA sequences than GC-rich sequences, so the parental DNA strand separation that is required during DNA replication initiates at these AT-rich sequences.
- **DnaA box sequences**. The DNA replication protein **DnaA** binds to the DnaA box sequences to initiate template DNA strand separation. Template DNA strand separation occurs at the AT-rich sequences.
- **GATC methylation sequences**. Methylation of the adenine bases within each GATC methylation sequence serves as an activation signal for DNA replication.

DNA replication begins at *OriC* and proceeds in both directions (clockwise and counterclockwise) around the circular bacterial chromosome (**bidirectional replication**). Further, a **replicon** is defined as all of the DNA replicated from a single origin. Since the entire *E. coli* chromosome is replicated from a single origin, the chromosome is one replicon**.**

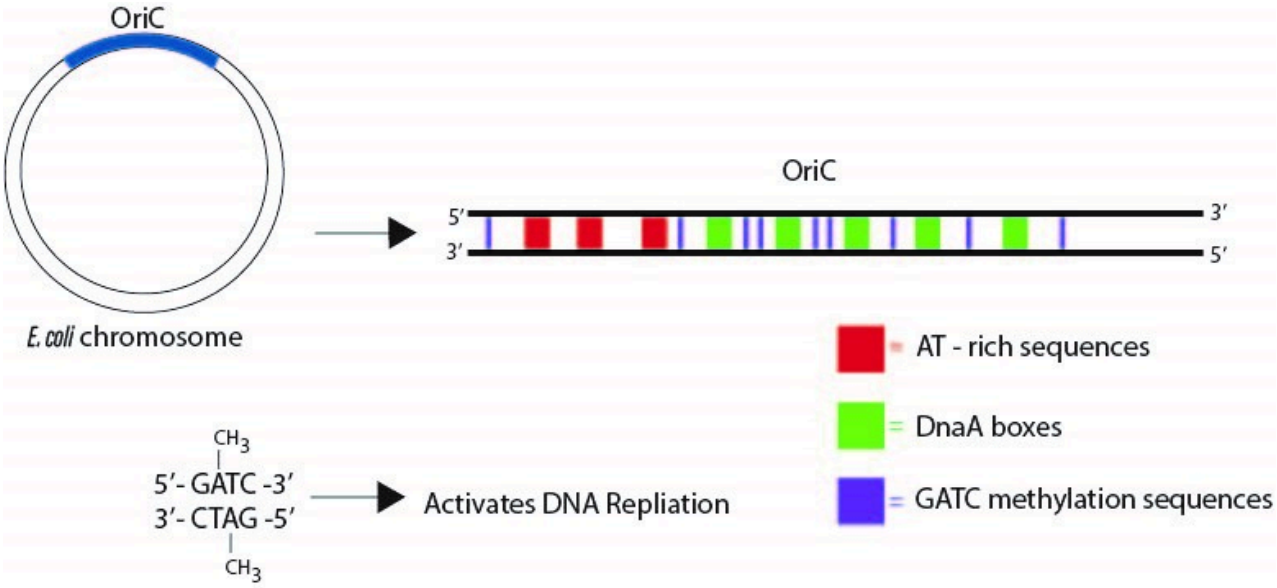


Figure 6.2 ***OriC in E.coli ---*** *Image created by KMD*

Key Questions

- What are the names and functions of the three DNA sequence types found in *OriC*?
- What is a replicon?

Replication Initiation

The steps involved in DNA replication in bacteria are (see **figure 6.3**):

1. **DnaA proteins bind to the DnaA box sequences**. When DnaA proteins bind to ATP, DnaA binds tightly to the DnaA box sequences within *OriC*.
2. **The origin forms a loop and the individual DNA strands separate.** Multiple copies of the DnaA bind to each other, forming a loop in the DNA. The DNA loop promotes DNA strand separation within the AT-rich sequences of *OriC*. This looping of the DNA and strand separation requires ATP cleavage by the DnaA protein. After ATP is cleaved, the DnaA proteins are released from *OriC*.
3. **A copy of DNA helicase binds to each of the two separated DNA strands.**
4. **The DNA helicases move along the template DNA strands, separating the DNA strands to form two replication forks**. Template DNA strand separation starts at *OriC* and moves in both directions around the circular bacterial chromosome. DNA helicase cleaves ATP and uses the released energy to catalyze DNA strand separation.
5. **Single-stranded DNA binding proteins (SSBPs) bind to the separated single-stranded template DNA strands.** SSBPs prevent the template DNA strands, separated by DNA helicase, from reforming hydrogen bonds, so that DNA replication can proceed.

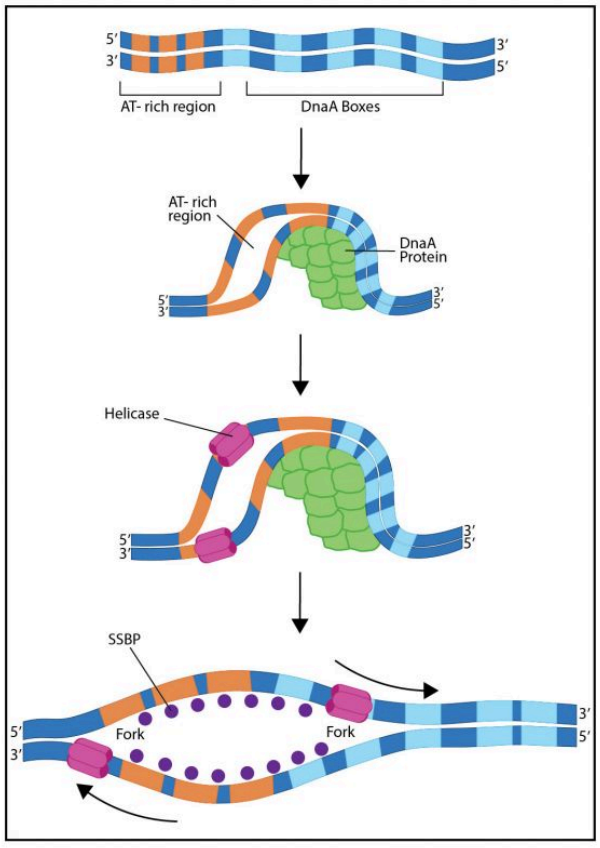


Figure 6.3 ***Replication Initiation in Bacteria*** *--- Image created by SL*

Coordinating DNA Replication with Cell Division

Most bacteria divide quickly; for example, the cell division time of *E. coli* is approximately 20 minutes. If DNA replication in *E. coli* does not keep up with the division of the cytoplasm, daughter cells will be formed that lack chromosomes. On the other hand, if DNA replication occurs too quickly, daughter *E. coli* cells would contain more than one copy of the chromosome.

How is DNA replication and division of the cytoplasm coordinated? *E. coli* coordinates these two processes by regulating how often DNA replication starts. There are two general ways to regulate the initiation of DNA replication:

- **Limiting the amount of active DnaA protein**. To initiate DNA replication, DnaA proteins must be bound to all DnaA box sequences within *OriC*. When a bacterial cell decides to replicate the DNA, there is only enough active DnaA proteins in the cell to bind to the DnaA box sequences within a single copy of *OriC*. After DNA replication occurs, there are now two copies of chromosome (and two copies of *OriC)* in the same cell. At this point, there is not enough active DnaA protein present in the cell to start a second round of DNA replication. By the time additional copies of the DnaA proteins are synthesized, the cytoplasm has divided producing two daughter cells.
- **Methylating GATC sequences**. The enzyme **DNA adenine methyltransferase (Dam)** recognizes the GATC methylation sequences in *OriC* and methylates the adenine nitrogenous bases in both DNA strands. Recall that there are numerous GATC methylation sequences in *OriC*. If every GATC sequence is methylated, DNA replication is initiated. After DNA replication, two DNA molecules are found in the same bacterial cell. Within each of these two molecules, the parental DNA strands contain methylated adenine, but the daughter DNA strands do not. A new round of DNA replication does not start until the Dam protein methylates the adenines within the daughter DNA strands (this can take several minutes). Thus, an *E. coli* cell has enough time to divide its cytoplasm prior to initiating a second round of DNA replication.

Key Questions

- What are the names and functions of the four proteins involved in DNA replication initiation in *E. coli*?

Replication Elongation

The elongation stage of DNA replication in bacteria consists of the following steps (see **figure 6.4**):

1. **RNA primers are synthesized**. After the template DNA strands have separated, small RNA strands (10-12 nucleotides long) are synthesized that form hydrogen bonds with the template DNA strands. These RNA **primers** provide the free 3’-OH groups required by DNA polymerases to initiate daughter DNA strand synthesis.

2. **DNA synthesis occurs by reading the template DNA strands.** Daughter DNA strands are synthesized in the 5’ to 3’ direction by adding dNTPs to free 3'-OH groups. However, because the template DNA strands are antiparallel to the daughter DNA strands, DNA polymerases read the template DNA strands in the 3’ to 5’ direction as the daughter DNA strands are synthesized. Note that as the DNA polymerase reads the template DNA strand 3' to 5' and synthesizes the daughter DNA strand 5' to 3', the DNA polymerase is moving in a single direction.

   Since DNA polymerases only synthesize the daughter DNA strands in the 5’ to 3’ direction, the two daughter DNA strands synthesized at each replication fork are made in opposite directions. One newly synthesized daughter DNA strand is called the **leading strand.** The leading strand is synthesized in the same direction that the replication fork is moving as the template DNA strands are separated. The leading DNA strand requires only one RNA primer and DNA synthesis is **continuous**. The other newly synthesized daughter DNA strand at each replication fork is the **lagging strand.** The lagging strand is synthesized as a series of **Okazaki fragments** (1000–2000 nucleotide-long DNA fragments) in the opposite direction the replication fork is separating the template DNA strands. Each Okazaki fragment is initiated by a single RNA primer; the lagging DNA strand is synthesized in a **discontinuous** (fragmented) manner.

3. **The RNA primers are removed.** Removing the RNA primers results in a gap between each Okazaki fragment.

4. **DNA synthesis fills the gaps left by the removed RNA primers.** DNA synthesis to fill the primer gaps occurs 5' to 3'.

5. **The adjacent Okazaki fragments are linked (ligated) together.** Ligation of the adjacent Okazaki fragments forms a continuous lagging DNA strand.

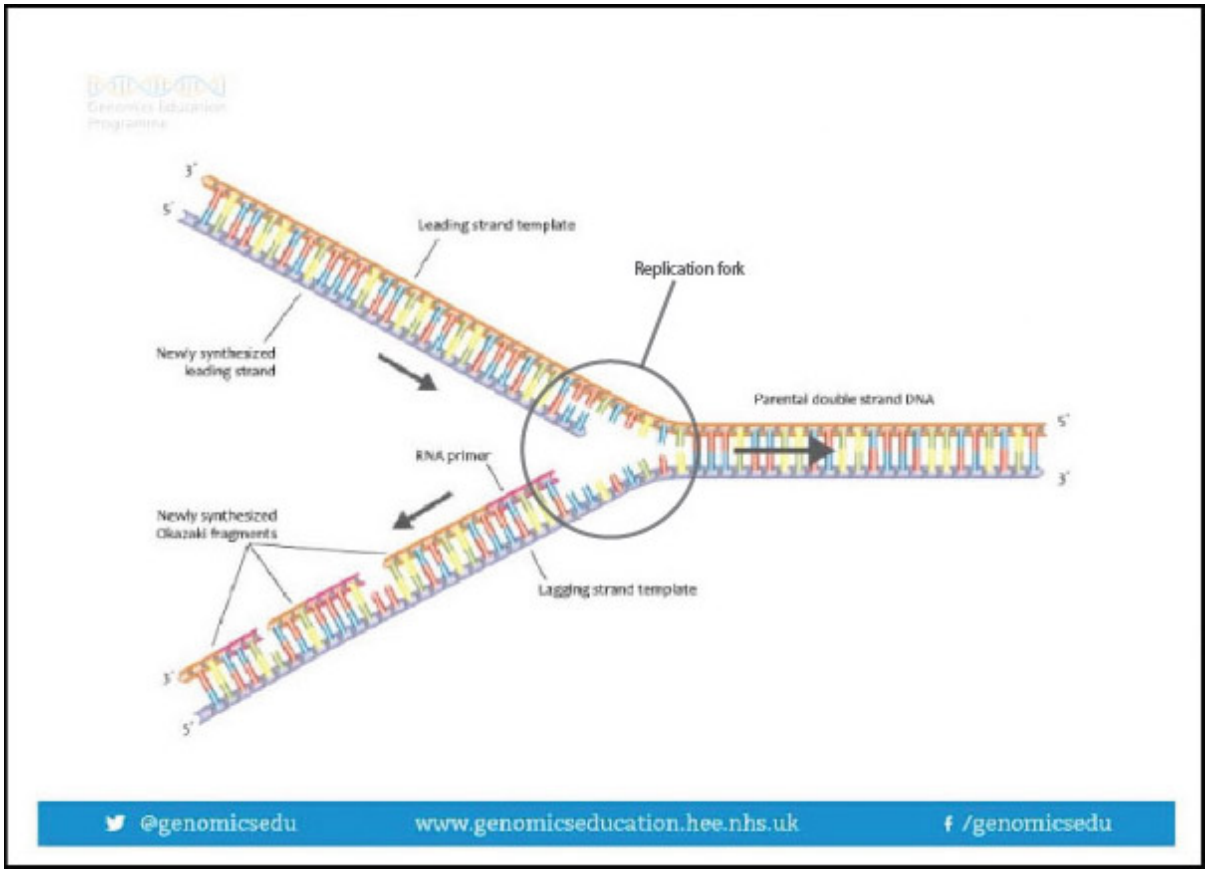


Figure 6.4 ***Replication Elongation ----*** *Image by Genomics Education Programme. Image licesensed under CC BY 2.0*

Key Questions

- What are the major events in the elongation stage of DNA replication in bacteria?
- In what direction do the DNA polymerases synthesize the daughter DNA strands?
- What is the difference between the leading and lagging DNA strands?

Proteins Involved in Elongation

The following proteins are involved in the elongation stage of DNA replication in bacteria (see **figure 6.5**):

- **DNA helicase**. DNA helicase separates the two parental DNA strands as the replication forks proceed from *OriC* clockwise and counterclockwise around the circular *E. coli* chromosome. DNA helicase uses the energy in ATP to break the hydrogens bonds between base pairs as the replication forks proceed.

- **Single-stranded DNA binding proteins (SSBPs).** SSBPs prevent the template DNA strands, separated by DNA helicase, from reforming hydrogen bonds.

- **DNA gyrase.** Since DNA is a right handed double helix, the separation of the parental DNA strands by DNA helicase produces **positive supercoiling** ahead of each replication fork. This positive supercoiling can be lethal to a bacterial cell if left unchecked. DNA gyrase functions to relieve this positive supercoiling by introducing **negative supercoils** ahead of each replication fork. DNA gyrase cleaves ATP and uses the released energy to form negative supercoils.

- **DNA primase.** To synthesize the daughter DNA strands, short RNA **primers** are synthesized by DNA primase. As mentioned earlier, the **leading strand** (DNA synthesis in the same direction as the movement of the replication fork) requires only a single RNA primer, while the **lagging strand** (DNA synthesis in the opposite direction as the movement of the replication fork) requires many RNA primers. Since DNA primase synthesizes an RNA nucleic acid strand (i.e., the primer), DNA primase cleaves RNA nucleotides (e.g., ATP, UTP, CTP, and GTP). As the RNA nucleotides are cleaved by DNA primase, two of the phosphate groups are released, while the remaining nucleoside monophosphates (e.g., AMP, UMP, CMP, and GMP) are incorporated into the synthesized primers.

- **The DNA polymerase III holoenzyme.** The DNA polymerase III holoenzyme synthesizes the daughter DNA strands in the 5′ to 3′ direction. A single DNA polymerase III holoenzyme synthesizes both the leading and lagging DNA strands at each replication fork simultaneously (see below). The DNA polymerase III holoenzyme synthesizes DNA using the nucleotides dATP, dTTP, dCTP, and dGTP as substrates. During daughter strand synthesis, these DNA nucleotides are cleaved, releasing two of the phosphate groups. The remaining nucleoside monophosphates (e.g., dAMP, dTMP, dCMP, and dGMP) are incorporated into the daughter DNA strands.

- **DNA polymerase I.** DNA polymerase I removes the RNA primers and synthesizes DNA to fill in the sequence gaps left by the removed primers. DNA synthesis by DNA polymerase I also occurs in the 5′ to 3′ direction. Like the DNA polymerase III holoenzyme, DNA polymerase I uses the nucleotides dATP, dTTP, dCTP, and dGTP as substrates as it synthesizes DNA. These DNA nucleotides are cleaved, releasing two of the phosphate groups. The remaining nucleoside monophosphates (e.g., dAMP, dTMP, dCMP, and dGMP) are incorporated into the synthesized DNA.

- **DNA ligase.** DNA ligase forms the final covalent bond that links adjacent Okazaki fragments into a continuous daughter DNA strand. DNA ligase uses the energy within ATP to synthesize the final covalent bond in the daughter DNA strand.

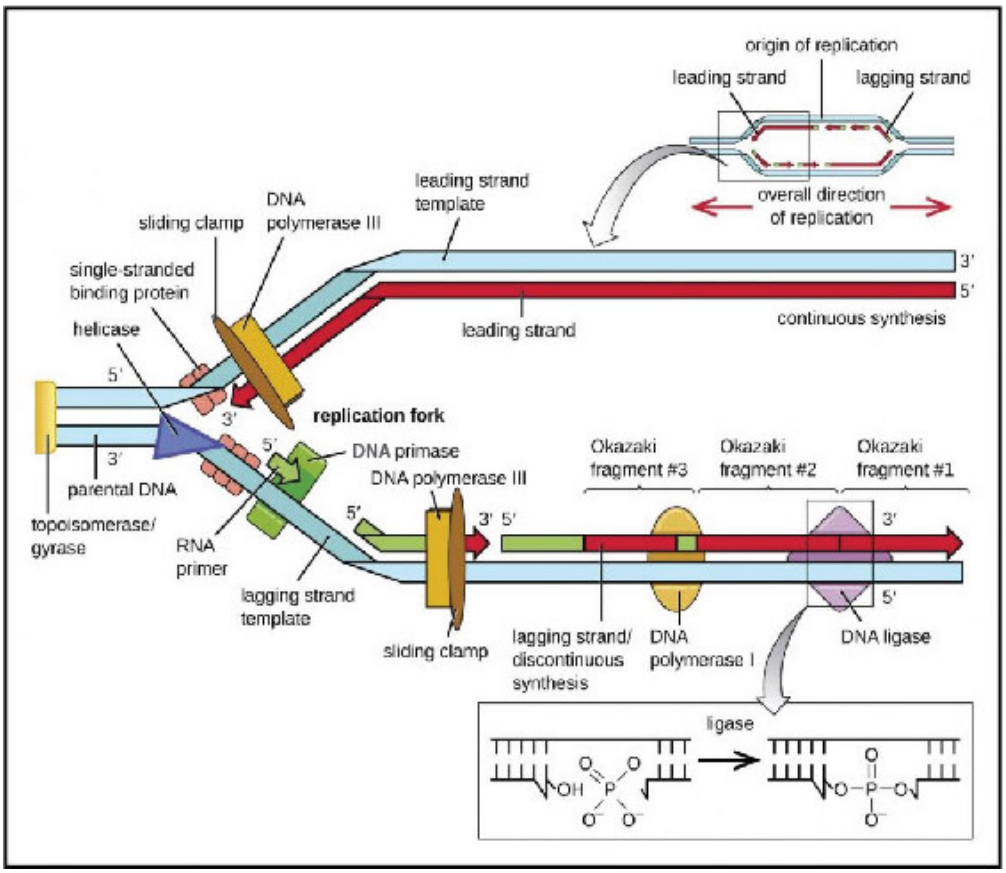


Figure 6.5 ***Bacterial Replication Proteins*** *--- This image is used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction)*

Key Questions

- What are the functions of the seven proteins involved in elongation in *E. coli*?
- List four replication elongation proteins that use ATP as energy.
- List two replication elongation proteins that use dNTPs as substrates for DNA synthesis.

DNA Polymerase III Holoenzyme

DNA polymerase III is a **holoenzyme** (multi-protein enzyme complex) composed of at least ten unique protein types (see **figure 6.6**). Moreover, each of these unique protein types within the DNA polymerase III holoenzyme is present in multiple copies, making the overall composition of the DNA polymerase III holoenzyme quite complex. The protein subunit composition of the DNA polymerase III holoenzyme is as follows:

- **Two alpha (α) protein subunits.** The α protein subunits of the DNA polymerase III holoenzyme carry out the **5' to 3' polymerase** activity to synthesize DNA. One α protein subunit synthesizes the leading DNA strand; the other α protein subunit synthesizes the lagging DNA strand.
- **Four beta (β) protein subunits.** The β protein subunits form sliding clamps that attach the two α subunits to the template DNA strands. These β subunits slide along with the template DNA strands during DNA replication, preventing the α subunits from falling off (increase DNA polymerase III holoenzyme processivity; see below).
- **Two epsilon (ε) protein subunits.** The ε protein subunits of DNA polymerase III possess **proofreading** activity (see below) that fixes mistakes made during DNA replication.
- **Accessory protein subunits.** The accessory protein subunits load the α and β subunits onto the RNA primers during lagging strand synthesis and maintain the overall stability of the DNA polymerase III holoenzyme.

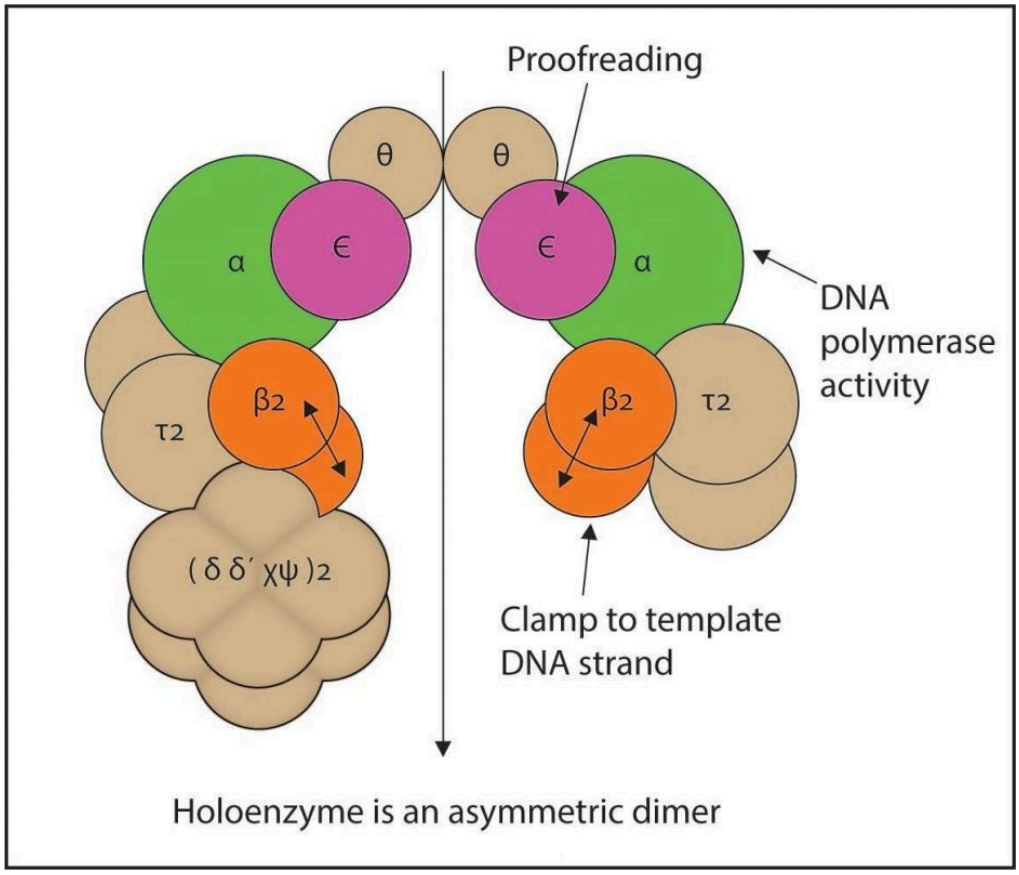


Figure 6.6 ***DNA Polymerase III Holoenzyme.*** *The direction the replication fork is moving is shown by the arrow in the center of the image (i.e. the replication fork is moving from top to bottom) The protein subunits on the left side of the image synthesize the lagging DNA strand; the protein subunits on the right synthesize the leading strand. The α subunits (green), β subunits (orange), ε subunits (pink), and accessory subunits (tan) are indicated. --- Image created by SL*

Key Questions

- What are the functions of the α, β, and ε subunits of the DNA polymerase III holoenzyme?

DNA Replication Proteins form Complexes

Many of the DNA replication enzymes described above are not physically separated. Each enzyme has a distinct function in DNA replication; however, many of these enzymes are physically linked to each other to form multiprotein "machines." For example, the **primosome** is a protein complex formed by DNA helicase and DNA primase. The primosome moves along the DNA separating the DNA strands and simultaneously synthesizing lagging strand RNA primers. Further, the primosome itself is part of a larger multi-subunit complex called the **replisome**. The replisome includes:

- The primosome components (DNA helicase, DNA primase).
- A DNA polymerase III holoenzyme (including the α, β, ε, and accessory protein subunits).

There is a single replisome per replication fork in the bacterium *E. coli*. Since a replicating bacterial chromosome has two replication forks, there are two replisomes per bacterial chromosome.

Key Questions

- What are the protein components of the primosome?
- What are the protein components of the replisome?

DNA Polymerases in Bacteria

In the bacterium *E. coli*, there are five DNA polymerase types. We will focus our attention on DNA polymerases I and III, as these two enzymes are involved in DNA replication. The other three DNA polymerases (DNA polymerase II, IV, and V) are involved in repairing bacterial DNA that has been damaged by environmental agents.

DNA polymerase III (also called the DNA polymerase III holoenzyme; see above) replicates the leading and lagging DNA strands (has **5' to 3' polymerase** activity). DNA polymerase III also contains a **proofreading activity** that removes DNA replication mistakes in the 3' to 5' direction (the so-called **3' to 5' exonuclease** activity; see below). **DNA polymerase I** is composed of a single protein subunit and functions to remove Okazaki fragment RNA primers in the 5' to 3' direction (i.e., the **5' to 3' exonuclease** activity). DNA polymerase I also fills in the gaps left by the removed RNA primers with DNA via its **5' to 3' polymerase** activity and has **3' to 5' exonuclease activity** (proofreading activity; see below).

All DNA polymerases have two unique features. First, DNA polymerases require a free 3'-OH group provided by the primer to begin DNA synthesis. The primer used within cells is RNA; however, DNA polymerases can use DNA primers to synthesize DNA as well. In fact, DNA primers are commonly used when synthesizing DNA in the lab (see Part 8). Second, DNA polymerases synthesize the growing daughter strand in the 5' to 3' direction only.

Key Questions

- What are the two enzymatic activities of DNA polymerase III holoenzyme?
- What are the three enzymatic activities of DNA polymerase I?
- What are two unique features of all DNA polymerases?

DNA Polymerase Mechanism

DNA polymerases use the chemical energy stored within the high energy phosphate bonds of deoxyribonucleoside triphosphate (**dNTP**) molecules to synthesize the daughter DNA strands. Specifically, the DNA polymerase mechanism involves (see **figure 6.7**):

1. The DNA polymerase reads a nitrogenous base in the template DNA strand and binds to the complementary dNTP according to the AT/GC rule. The incoming dNTP forms hydrogen bonds with the nitrogenous base in the template DNA strand.
2. The free 3'-OH group on the growing daughter DNA strand reacts with the phosphate groups on the incoming dNTP.
3. A high energy bond within the dNTP is broken releasing two of the phosphate groups in the form of **pyrophosphate (PP_i)**.
4. The released energy is used to synthesize a new phosphodiester bond between the 3' end of the growing DNA strand and the 5' end of the incoming nucleotide.

The DNA polymerase III holoenzyme is **processive**. Processivity means that the DNA polymerase III holoenzyme can add many nucleotides to a daughter DNA strand without falling off the template DNA strand. This processivity is due to the four β subunits (sliding clamps; see above) found within the DNA polymerase III holoenzyme.

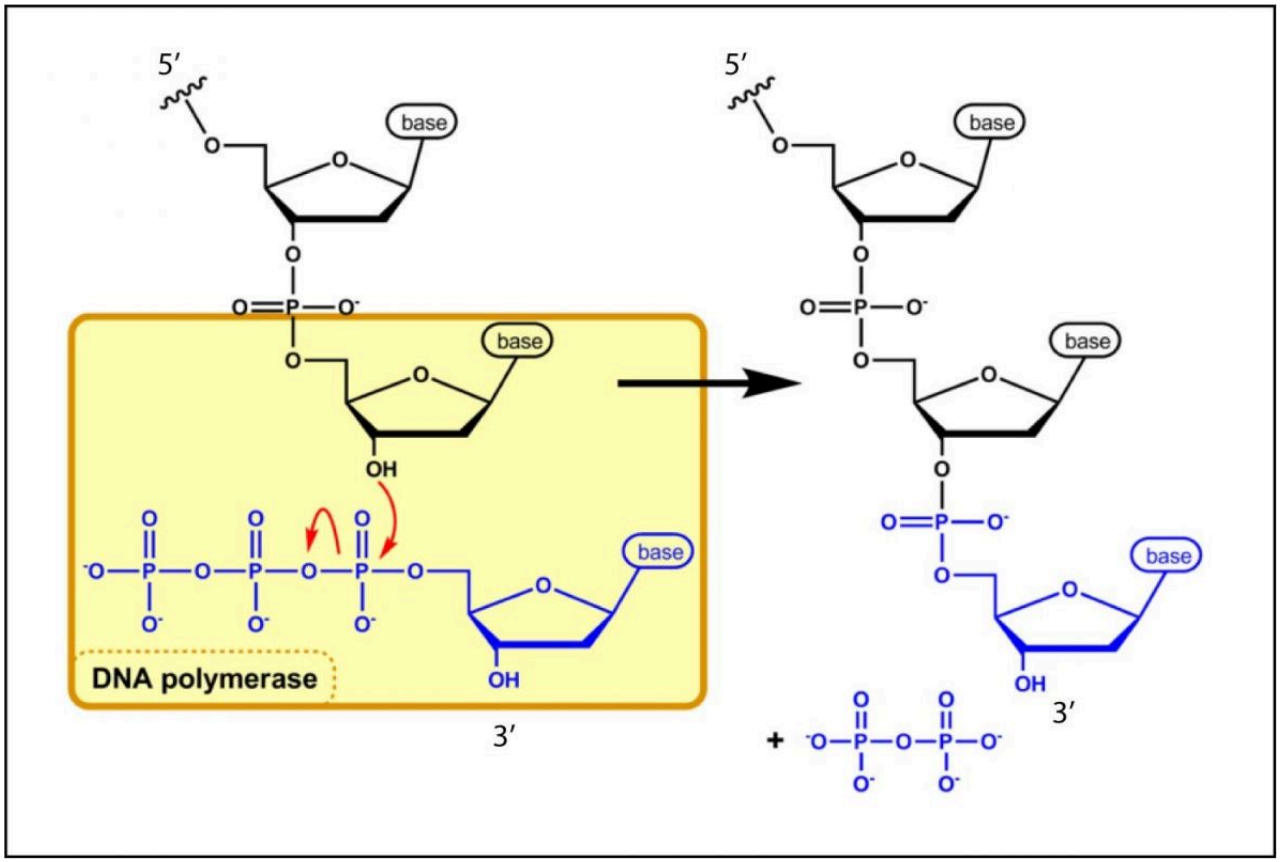


Figure 6.7 ***DNA Polymerase Mechanism*** *--- Image created by Michal Sobkowski and is licensed under CC BY 3.0.*

Key Questions

- Describe the DNA polymerase mechanism.
- What is meant by the phrase "DNA polymerases are processive?"

Proofreading by DNA Polymerases

DNA polymerases incorporate the wrong nucleotide (i.e., a nucleotide that forms base pairs that deviate from the AT/GC rule) into a daughter DNA strand rarely. For example, the DNA polymerase III holoenzyme is thought to incorporate the wrong nitrogenous base once in every 10–100 million nitrogenous bases in a daughter DNA strand. This accuracy during DNA synthesis is called **fidelity**; both DNA polymerase I and the DNA polymerase III holoenzyme are said to have high fidelity (low error rates). The fidelity of DNA polymerases is the combination of three factors:

- **The stability of the hydrogen bonds between AT and GC.** Mismatched nitrogenous base pairs fail to form hydrogen bonds altogether or result in less stable hydrogen bonds.
- **The active site of DNA polymerases is specific.** A covalent bond is not formed between the free 3'-OH group of the growing daughter DNA strand and the free 5' phosphate group of the incoming dNTP unless correct base pairing occurs.
- **Proofreading.** If an incorrect base pair is accidently formed, the DNA polymerase can pause, recognize the mismatch, and remove it (see **figure 6.8**). This **proofreading activity** occurs in the 3' to 5' direction on the daughter DNA strand and is sometimes called the **3' to 5' exonuclease activity** of the enzyme. Once proofreading is complete, the DNA polymerase can resume incorporating dNTPs into the growing daughter DNA strand in the 5' to 3' direction.

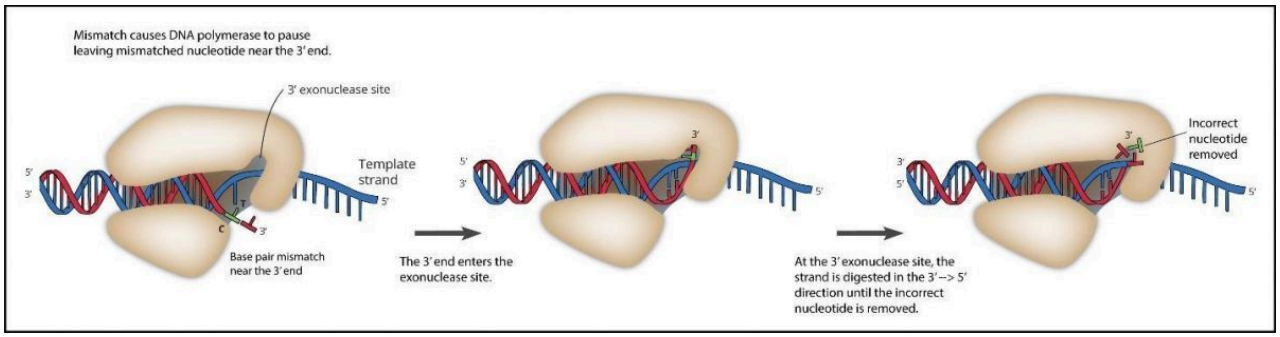


Figure 6.8 ***Proofreading*** *--- Image created by SL*

Key Questions

- What is meant by proofreading?
- Which enzymatic activity is responsible for proofreading?
- What is meant by the phrase, "DNA polymerases display high fidelity?"

Termination of Replication in Bacteria

DNA replication in *E. coli* terminates at specific locations within the circular chromosome called **termination (*ter*) sequences**. Since there are two replication forks moving in opposite directions around the circular chromosome, there are two *ter* DNA sequences. Each *ter* sequence (the **T1** and **T2** sequences) stops the advancement of one of the two replication forks (see **figure 6.9**). Proteins called **termination utilization substances (Tus)** bind to the T1 and T2 sequences. Tus proteins release the replisomes from the two replication forks, terminating DNA replication.

Once replication ceases, **DNA ligase** forms the final covalent bond between the 5' and 3' ends of each daughter DNA strand, resulting in two double-stranded circular *E. coli* chromosomes. These chromosomes can then be distributed to

daughter *E. coli* cells after cell division.

Occasionally, the two chromosomes produced by DNA replication are intertwined like the links in a chain. These intertwined DNA molecules are called **catenanes**. Catenanes must be separated prior to the division of the *E. coli* cytoplasm, so that each daughter cell receives a chromosome. **DNA gyrase** solves this catenane problem by cutting one chromosome (both DNA strands are cut), passing the other chromosome through the break, and sealing the break to generate two separate chromosomes that can be distributed properly to the daughter bacterial cells.

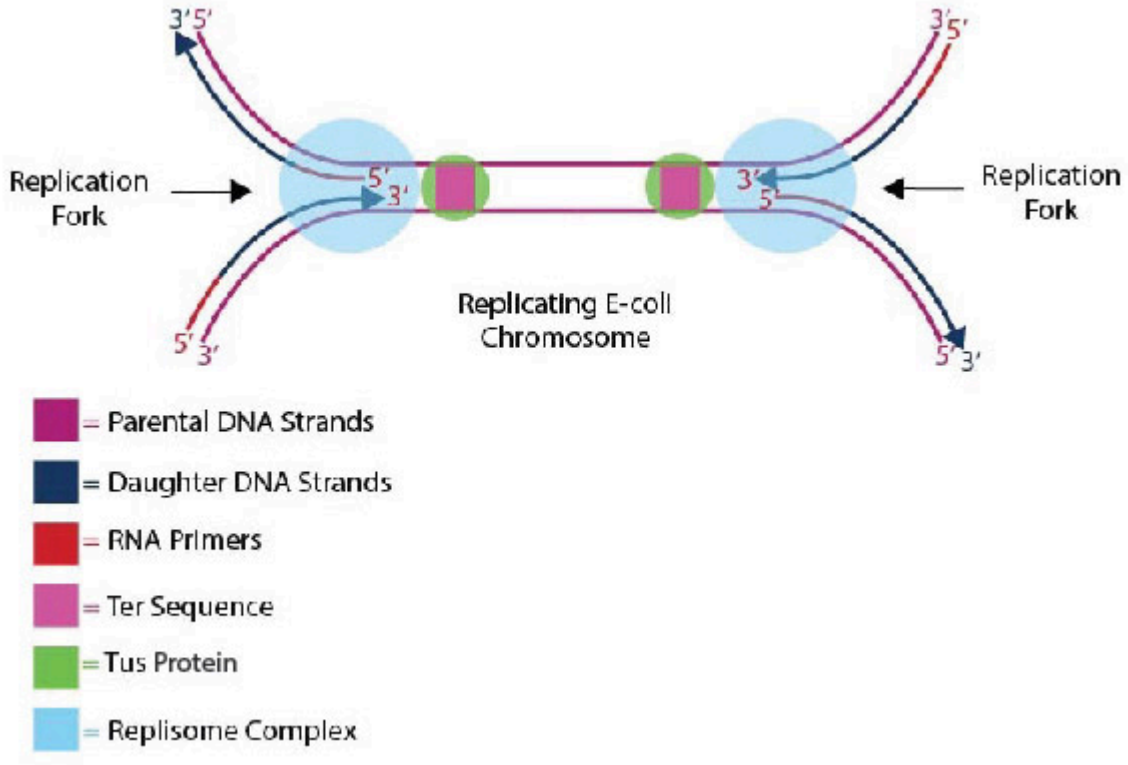


Figure 6.9 ***Termination of Replication in E. coli*** *--- Image created by KMD*

Key Questions

- What DNA sequences participate in replication termination in *E. coli*?
- What are the names and functions of the three proteins that participate in replication termination in *E. coli?*
- How are catenanes resolved?

B. DNA Replication in Eukaryotes

Eukaryotic Origins

Eukaryotic DNA replication is more complex than DNA replication in bacteria. This increase in complexity is because eukaryotic genomes are generally larger than prokaryotic genomes, and the genetic material in eukaryotes is organized into linear chromosomes. However, the good news is that the DNA replication process is similar in prokaryotes and eukaryotes and many of the DNA replication proteins (helicases, primases, and polymerases) identified in bacteria have

eukaryotic counterparts that function in the same way. In contrast, one major difference between prokaryotic and eukaryotic DNA replication is that eukaryotic chromosomes have multiple replication origins (see **figure 6.10**). Like bacteria, DNA replication proceeds bidirectionally from each origin, with the formation of two replication forks per origin. As DNA replication occurs, the replication forks from adjacent origins fuse, eventually producing two identical sister chromatids.

In a model eukaryotic organism, the bread yeast *Saccharomyces cerevisiae*, the 250–400 origins are called **ARS elements.** *S. cerevisiae* ARS elements have the following features:

- ARS elements are approximately 50 base pairs (bp) in length.
- ARS elements are AT-rich. The presence of numerous AT base pairs within the origin promotes DNA strand separation.
- ARS elements contain an **ARS consensus sequence (ACS)**. This ARS consensus sequence is the binding site for the ORC protein complex (see below).

The DNA replicated from a single ARS element is called a **replicon**. Since eukaryotic organisms have many origins, eukaryotes also have many replicons. For example, *S. cerevisiae* contains 250–400 replicons per genome, while the human genome is thought to contain approximately 25,000 replicons.

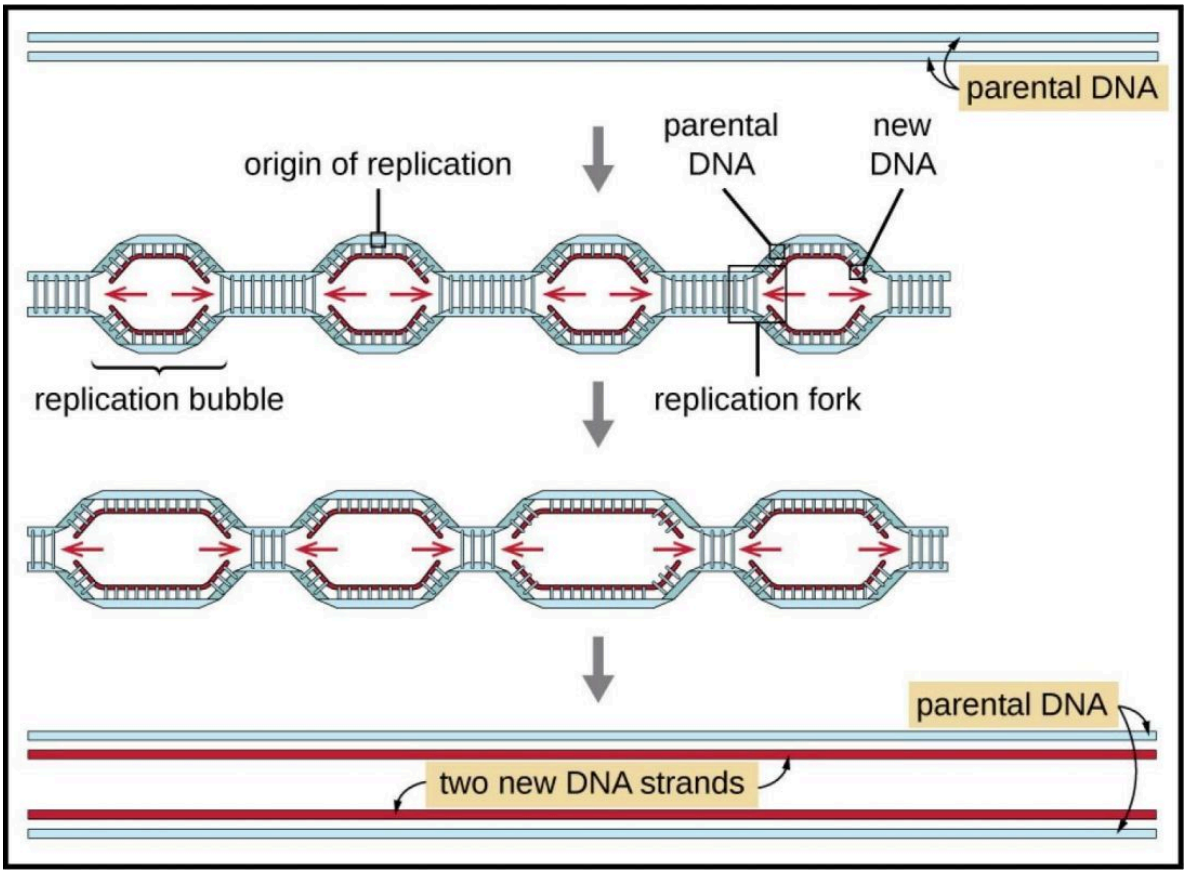


Figure 6.10 ***Eukaryotic Chromosomes Have Multiple Origins*** *--- This image is used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction)*

Key Questions

- How is DNA replication in prokaryotes and eukaryotes similar?
- What is one major differences between prokaryotic and eukaryotic replication?
- What are the features of an ARS element?

Replication Initiation in Eukaryotes

A multi-subunit **prereplication complex (preRC)** assembles on each ARS element and initiates DNA replication in eukaryotes (see **figure 6.11**). The preRC contains the following protein components:

- The **origin recognition complex (ORC)**. ORC binds directly to the ARS consensus sequence within each ARS element (origin).
- **Regulatory proteins**. Two regulatory proteins, **cdc6** and **cdt1**, bind to ORC and function to inhibit the initiation of DNA replication during the G_1, G_2, and M phases of the cell cycle. That way the initiation of DNA replication is tightly controlled; DNA replication can occur only during the synthesis (S) phase of the cell cycle. During S phase, cdc6 and cdt1 are phosphorylated by cellular kinases, causing cdc6, cdt1, and ORC to be released from the ARS element. DNA replication is then initiated.
- **MCM helicase.** Once the cdc6, cdt1, and ORC proteins are released, the MCM helicases catalyze the separation of the two parental DNA strands forming two replication forks. Like prokaryotic DNA helicases, the MCM helicases cleave ATP and use the released energy to form replication forks.

After the DNA strands have separated, **replication protein A (RPA)** prevents the separated DNA strands from reforming hydrogen bonds. The eukaryotic DNA polymerases can then begin the elongation stage of DNA replication.

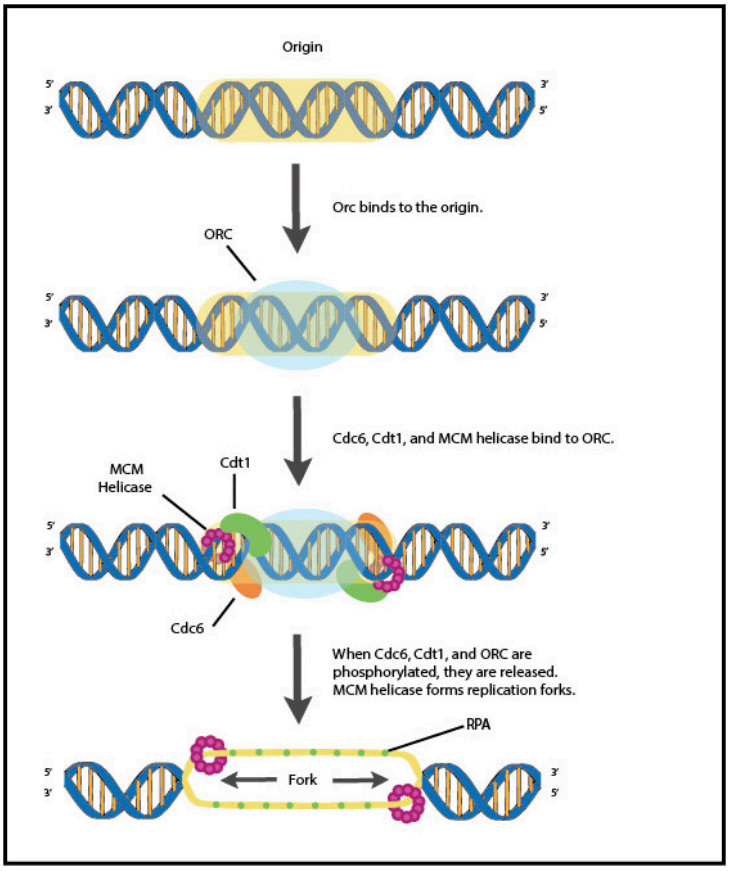


Figure 6.11 ***Replication Initiation in Eukaryotes*** *--- Image created by SL*

Key Questions

- What are the names and functions of the five proteins that participate in DNA replication initiation in eukaryotes?

Replication Elongation in Eukaryotes

MCM helicase continues DNA strand separation during the elongation phase of DNA replication, causing the replication forks to proceed in both directions away from each origin. **RPA** prevents the separated DNA strands from reforming hydrogen bonds. The separation of the DNA strands by MCM helicase generates positive supercoiling ahead of each replication fork. **Topoisomerase II** is located ahead of each replication fork and produces negative supercoiling to compensate for the positive supercoiling produced by MCM helicase. Topoisomerase II cleaves ATP to generate negative supercoils.

There are over a dozen different DNA polymerases in a typical eukaryotic cell. These eukaryotic DNA polymerases are named according to the Greek alphabet (α, β, γ, etc.). **DNA polymerases alpha (α), delta (δ)**, and **epsilon (ε)** are the DNA polymerases involved in replicating nuclear DNA in eukaryotes (see **figure 6.12**). DNA polymerase α binds to **DNA primase** to form a protein complex that synthesizes hybrid nucleic acid strands composed of 10 RNA nucleotides followed by 10–30 DNA nucleotides. These hybrid nucleic acid strands are used as primers by DNA polymerases δ and ε. DNA primase synthesizes the RNA component of the hybrid primer, while DNA polymerase α synthesizes the DNA component of the hybrid primer. Note that DNA polymerase α has both 5′ to 3′ polymerase and 3′ to 5′ exonuclease

(proofreading) activity. Once the primer is made, DNA polymerase α is released and is replaced by either DNA polymerase δ or DNA polymerase ε (i.e., the so-called **polymerase switch)**.

DNA polymerases δ and ε are the processive eukaryotic DNA polymerases. These two DNA polymerases bind to **proliferating cell nuclear antigen (PCNA)**, a protein that functions as a sliding clamp, increasing the processivity of DNA polymerases δ and ε. Once bound to PCNA, DNA polymerase ε synthesizes the leading strand, whereas the PCNA:DNA polymerase δ complex synthesizes the lagging DNA strand. Both DNA polymerases ε and δ contain 5′ to 3′ polymerase and 3′ to 5′ exonuclease (proofreading) activity. All three eukaryotic DNA polymerases (α, δ, and ε) cleave dNTPs during DNA synthesis. The released energy powers DNA replication, while the nucleoside monophosphates (dAMP, dTMP, dCMP, and dGMP) are incorporated into the growing daughter DNA strands.

Finally, **flap endonuclease** (**Fen1**) removes the RNA nucleotides of each primer, and **DNA ligase I** forms the final covalent bonds to link adjacent Okazaki fragments in the lagging DNA strands. DNA ligase I cleaves ATP during ligation.

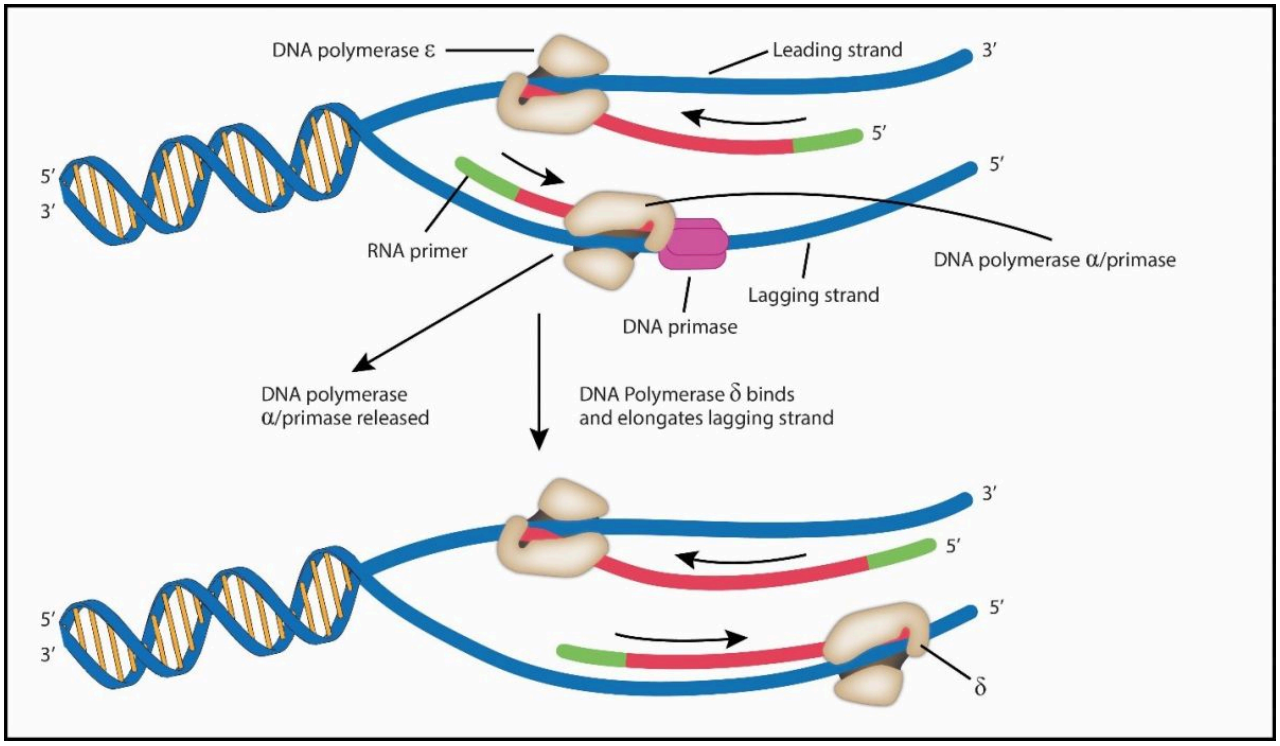


Figure 6.12 ***Replication Elongation in Eukaryotes*** *--- Image created by SL.*

Key Questions

- What are the eukaryotic equivalents of the *E. coli* enzymes DNA helicase, SSBPs, DNA gyrase, DNA primase, DNA polymerase III holoenzyme, DNA polymerase I, and DNA ligase?
- Which eukaryotic replication enzyme synthesizes the leading DNA strand?
- Which enzyme synthesizes the lagging DNA strand?
- Which eukaryotic replication elongation enzymes cleave ATP?
- Which enzymes cleave dNTPs?

Replication at Chromosome Ends

The 3’ ends of the parental DNA strands within linear eukaryotic chromosomes present a potential problem during DNA replication. Suppose a primer is made for the daughter DNA strand directly opposite the 3’ end of the parental DNA strand. Once this primer is used for DNA synthesis, the primer is removed with the hope that DNA replication will fill in the primer gap. However, DNA polymerases cannot fill in the primer gap at the end of the chromosome because DNA polymerases require a 3’-OH group to begin DNA synthesis. As a result, this primer gap is not filled in, and the newly synthesized daughter DNA strand is slightly shorter than its template DNA strand. This end replication problem would result in the progressive shortening of daughter DNA strands with each round of DNA replication. Eventually, this shortening would delete genes and have a negative effect on the phenotype of the cell.

Eukaryotes solve this potential DNA replication problem by using **telomerase** to add moderately repetitive DNA sequences to the 3’ ends of the parental DNA strands prior to DNA replication (see **figure 6.13**). Telomerase is an unusual enzyme that contains a built-in RNA component (**TERC**) and a protein component (**TERT**). Thus, telomerase is an example of a **ribonucleoprotein**. The TERC component forms hydrogen bonds with the 3’ overhang DNA sequence at the ends of the two parental DNA strands. Once bound to the 3’ end of the parental DNA strands, TERT catalyzes the synthesis of additional telomere repeat sequences using the built-in TERC component of telomerase as a template. The synthesis of additional telomere repeats by telomerase occurs in the 5’ to 3’ direction. Because telomerase synthesizes DNA in the 5' to 3' direction and requires a 3'-OH group for DNA synthesis, telomerase is considered a DNA polymerase.

Once the 3’ end of the parental DNA strand is lengthened by telomerase, DNA replication of the daughter DNA strand can occur by the synthesis of a primer opposite the repeats added by telomerase. DNA synthesis from this newly added primer occurs using the DNA polymerase δ. Finally, the primer is removed by Fen1. Since the primer for the daughter DNA strand is made opposite the telomere repeat sequences added by telomerase, the loss of the primer does not affect structural genes or the phenotype of the daughter cell.

To sum this all up, telomerase lengthens the parental DNA strands prior to DNA replication, so that the replication enzymes can make the daughter DNA strands shorter. The net result is that the overall chromosome length does not change significantly because of DNA replication.

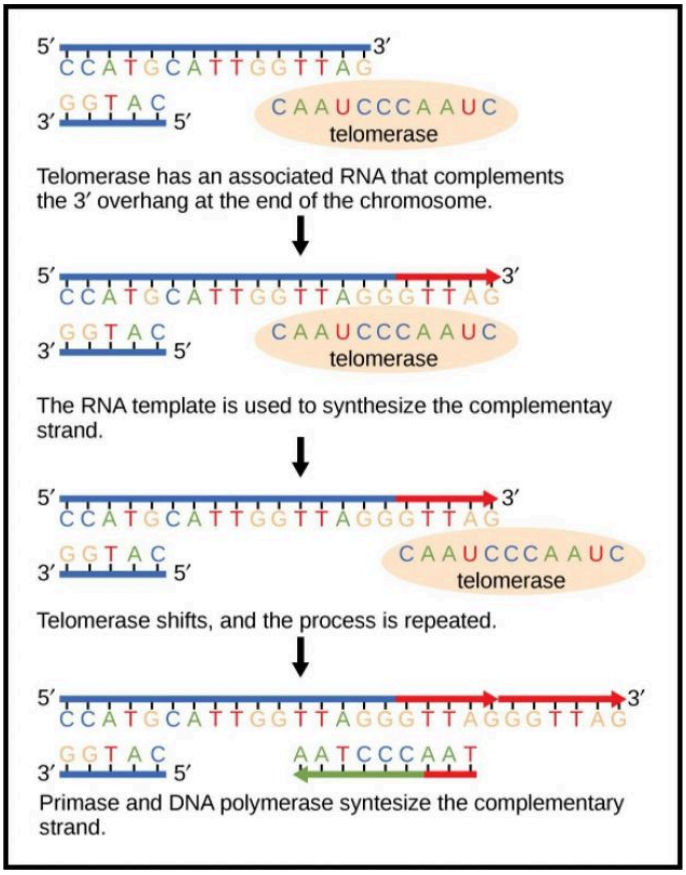


Figure 6.13 ***Telomerase Mechanism*** *--- This image is used from OpenStax (access for free at* https://openstax.org/books/biology-2e/pages/1-introduction*)*

Key Questions

- Describe the so-called "end replication problem" experienced by organisms with linear chromosomes.
- How is this end replication problem solved?
- What are the functions of the two components of telomerase?

Review Questions

Fill in the Blank:

1. The enzyme _____ methylates adenine to activate DNA replication in bacteria.
2. The enzyme _____ connects adjacent Okazaki fragments together during DNA replication in *E. coli.*
3. The _____ protein is the eukaryotic equivalent of SSBPs.
4. The enzyme _____ is composed of two subunits, called TERC and TERT.
5. During DNA replication, the template DNA strands are read by DNA polymerases in the _____ direction, while the daughter DNA strands are synthesized in the _____ direction.
6. Phosphorylation of _____ and _____ initiates DNA replication in eukaryotic organisms.
7. _____ is a eukaryotic enzyme that produces replication forks, while _____ is an *E. coli* enzyme that alleviates positive supercoiling ahead of each replication fork.
8. The _____ subunit of the DNA polymerase III holoenzyme is responsible for proofreading, while the _____ subunit is responsible for DNA synthesis.
9. _____ is an unusual DNA polymerase that contains a built-in RNA template molecule.
10. The enzyme _____ has both 5'- 3' polymerase and 5' - 3' exonuclease activity.
11. _____ binds directly to the ARS element, while _____ synthesizes the leading DNA strand in eukaryotes.

7 - Mutations and DNA Repair

The function of DNA is to store genetic information. To store genetic information effectively, it is important that the sequence of DNA remains unchanged from generation to generation. However, rare mistakes occur during DNA replication; recall one mistake is made every 10 to 100 million nucleotides incorporated into daughter DNA strands in bacteria. Further, environmental agents, such as ionizing radiation, ultraviolet light, and a myriad of chemicals can damage DNA. DNA damage is often detrimental; DNA damage worsens protein activity and negatively affects phenotype. As a result, it may seem that it would be beneficial to fix all replication errors or DNA damage: however, changes in the DNA sequence can be advantageous in some cases. Alterations in the DNA sequence can create new gene variants (alleles) in a population. Evolution requires this diversity of alleles; natural selection chooses which allele combinations survive and reproduce in a particular environment. For this reason, a low level of mutation is required for evolutionary change.

In this section, we will define mutation, discuss various ways mutations can be classified, learn the mechanisms used by a bacterial cell to repair damaged DNA, and discuss situations in which defective DNA repair leads to disease in humans.

A. Mutations

What is a mutation?

A **mutation** is a change in the genetic material that is inherited by daughter cells at the conclusion of cell division. In most cases, an incorrect nucleotide is replaced with the correct one before the cell divides; therefore, based on the definition above, this type of change would not be considered a mutation. Only nucleotide changes that remain after cell division are mutations.

When environmental agents (ionizing radiation, ultraviolet light) damage the DNA double-helix, **induced mutations** occur. In other cases, DNA polymerases inadvertently incorporate an incorrect nucleotide into the daughter DNA strand during DNA replication. This DNA replication error is called a **spontaneous mutation**.

Key Questions

- What is the definition of mutation?
- What is the difference between an induced and a spontaneous mutation?

Germline and somatic mutations

What determines if a mutation is inherited by offspring? **Gametes cells** arise from specialized cells called **germline cells**. In contrast, the non-gamete cells within the body (muscle cells, neurons, epithelial cells, etc.) are **somatic cells**. Therefore, when a mutation arises in a germline cell (**germline mutation**), the mutation is transmitted via the gametes to

the individual's offspring. In contrast, a mutation generated in a somatic cell (**somatic mutation**) is not transmitted to the individual's offspring.

The timing of the mutation can also influence its transmission to offspring. If a mutation arises in a cell during early embryonic development, then chances are good that the mutation will be found in both the germline and somatic cells. Since this mutation is found in germline cells, the mutation will then be transmitted to the next generation. However, when a mutation arises late in development or in the adult organism, the mutation may only be found in one type of somatic cell and will be absent from the germline cells. As a result, this somatic mutation will not be transmitted to the offspring.

The timing of the mutation can also affect the severity of the phenotype. If a somatic mutation arises early in the development of a particular organ, it is likely that the majority of cells that make up the organ will contain the mutation. When an organ is made up of mostly mutant cells, the phenotype of the organ is impacted, often in a negative way. Alternatively, if only a subset of cells in the tissue contains the mutation, the phenotypic effect can be comparatively mild.

Key Questions

- What is the difference between a germline and a somatic mutation?
- How likely is a germline mutation transmitted from parents to their offspring? How likely is a somatic mutation transmitted from parents to their offspring?
- If a mutation arises early in development, is it more or less likely to be transmitted to the germline?

Point Mutations

Sometimes a mutation alters a single nucleotide in the DNA molecule. When only one nucleotide has been altered, the mutation is classified as a **point mutation**. When a purine nitrogenous base is changed to another purine (e.g., adenine to guanine) or a pyrimidine nitrogenous base is exchanged for another pyrimidine (e.g., cytosine to thymine), the point mutation is a **transition** mutation. Alternatively, when a purine is exchanged for a pyrimidine (or vice versa), the point mutation is a **transversion** mutation.

Silent, Missense, and Nonsense Mutations

When the mutation occurs in the intergenic or the repetitive DNA sequences within chromosomes, the mutation often has no or mild phenotypic consequences. Even though these mutations outside of the structural genes do not alter the cell's phenotype, these mutations are useful as they define the **DNA fingerprint** of an individual. Each person's DNA fingerprint contains a unique collection of mutations within the intergenic and repetitive DNA sequences (see Part 1).

We will focus our attention on mutations that occur within structural genes. Remember that structural genes are transcribed into messenger RNA (mRNA) molecules. The mRNA is then translated when the ribosome reads three nucleotide-long **codons** to generate the amino acid sequence of the encoded protein (see Part 11). When one nucleotide has been exchanged for another nucleotide, but there is no change in the amino acid sequence of the encoded protein, we refer to the mutation as a **silent** mutation (see **Figure 7.1**). Silent mutations occur because the genetic code is **degenerate**, that is, more than one codon specifies the same amino acid.

WILD-TYPE 5’ AUG . UUC . GUG . CAC . UUA . AUC . UAG 3’

MET . PHE . VAL . HIS . LEU . ILE . STOP

SILENT 5’ AUG . UU**U** . GUG . CAC . UUA . AUC . UAG 3’

MET . **PHE** . VAL . HIS . LEU . ILE . STOP

MISSENSE 5’ AUG . UUC . GUG . **A**AC . UUA . AUC . UAG 3’

MET . PHE . VAL . **ASN** . LEU . ILE . STOP

NONSENSE 5’ AUG . UUC . GUG . CAC . U**A**A . AUC . UAG 3’

MET . PHE . VAL . HIS . **STOP**

FIGURE 7.1 Base Substitutions Can Affect Gene Structure and Function. The sequence of the wild-type RNA is indicated, with the amino acid sequence of the translated protein shown below it. In each mutation, the effect on the amino acid sequence of the encoded protein is indicated in bold. Silent mutations do not change the encoded amino acid. Missense mutations change a single amino acid in the encoded protein. Nonsense mutations change a codon into a stop codon.

When a point mutation substitutes a single amino acid, the mutation is a **missense** mutation. Missense mutations can sometimes cause disease. For example, **sickle cell anemia** is an autosomal recessive disease caused by a missense mutation in the structural gene that makes the beta globin protein subunits within hemoglobin. In sickle cell anemia, the missense mutation in the beta globin gene replaces the amino acid glutamic acid in codon six with the amino acid valine.

Sometimes a point mutation results in the formation of a premature stop codon (**nonsense mutation**). Nonsense mutations are usually more severe in phenotype than missense mutations because nonsense mutations cause the encoded protein to be shorter in length than normal. These shorter proteins are often nonfunctional.

Frameshift Mutations

In some cases, nucleotides are inserted into the sequence of a structural gene (i.e., an **insertion** mutation) or nucleotides are **deleted** from the sequence of the structural gene. Because the ribosome reads the encoded mRNA one codon at a time during translation (see Part 11), some insertions or deletions will change the **reading frame** of the mRNA molecule. When the reading frame changes, the entire amino acid sequence of the encoded protein changes from the point of the insertion/deletion (**indel**) site onward. A mutation that changes the reading frame is called a **frameshift mutation** (see **Figure 7.2**).

WILD-TYPE 5’ AUG . UUC . GUG . CAC . UUA . AUC . UAG 3’

MET . PHE . VAL . HIS . LEU . ILE . STOP

FRAMESHIFT 5’ AUG . UUC . G**C**U . GCA . CUU . AAU . CUA . G 3’

MET . PHE . **ALA** . **ALA** . **LEU** . **ASN** . **LEU** ….

FIGURE 7.2 Frameshift Mutations Can Affect Gene Structure and Function. By inserting a single cytosine base in the third codon (underlined), a reading frame shift occurs that changes the amino acid sequence of the protein and eliminates the stop codon.

Key Questions

- What is a transition mutation?
- What is a transversion mutation?
- What is the difference between a silent, missense, and a nonsense mutation?
- How do frameshift mutations affect the amino acid sequence of a protein?

B. DNA Repair Systems

Most alterations to the DNA sequence are recognized by cellular **DNA repair systems** immediately. Some of these DNA repair systems recognize changes in the width of the DNA double-helix. For example, the DNA double-helix always has a purine (adenine or guanine) paired with a pyrimidine (cytosine or thymine). Purine-pyrimidine base pairing results in a DNA molecule that is 2 nanometers (nm) wide. Repair enzymes run along the length of double stranded DNA as it is replicating, ensuring that the replicated DNA is 2 nm wide. If two purines are paired, a bulge in the DNA double-helix occurs, and if there are two pyrimidines paired, the double-helix narrows. Variations in the width of the DNA double-helix signal that damage has occurred, and DNA repair is required.

There are multiple DNA repair systems because there are different types of mutations to correct. Even though each DNA repair system has different components and recognizes a different aberration in the DNA, there are some features in common among the DNA repair systems. First, each pathway contains an enzyme that recognizes the DNA damage. Next, the DNA damage and often a few extra nucleotides are removed from one of the two DNA strands. Finally, the excised nucleotides are replaced by a DNA polymerase and the final phosphodiester bond in the damaged DNA strand is formed by DNA ligase.

There are six DNA repair systems:

1. **Proofreading**
2. **Mismatch repair**
3. **Base excision repair (BER)**
4. **Nucleotide excision repair (NER)**
5. **Homology directed repair (HDR)**
6. **Non-homologous end joining (NHEJ)**

You will only need to know about the first four DNA repair systems in our current BIO375 class. The homology directed repair (HDR) and non-homologous end joining (NHEJ) content will be added to BIO375 in the future.

Proofreading

The first-line DNA repair system involves the DNA polymerases we discussed in Part 6. When a DNA polymerase synthesizes the daughter DNA strand, the DNA polymerase occasionally inserts an incorrect nucleotide. Recall that DNA polymerases possess 3’ to 5’ exonuclease activity; the DNA polymerase can reverse directions, remove the incorrect nucleotide in the 3' to 5' direction, and then move 5' to 3' again, replacing the incorrect nucleotide with the correct one.

Mismatch Repair

If an incorrect base pair is formed during DNA replication and the mistake is not removed by proofreading, the **mismatch repair** system is activated to fix the mistake. Mismatch repair recognizes which of the two nucleotides in the base pair is correct and then removes the incorrect nucleotide. For example, suppose that cytosine in the daughter DNA strand has been paired accidently with adenine in the template DNA strand during DNA replication. When mismatch repair recognizes this mistake, it now faces a dilemma: which of the two bases is incorrect? If mismatch repair

randomly chooses the nucleotide to correct, then 50% of the time, it will remove the adenine in the template DNA strand, instead of the cytosine in the daughter DNA strand.

In mismatch repair, the cell must first distinguish the template and daughter DNA strands. In the bacterium *E. coli*, the template DNA strand is methylated by **DNA adenine methyltransferase (Dam)**, while newly synthesized daughter DNA strands are unmethylated. The proteins in the mismatch repair system recognize the methylated parental strand, and then mismatch repair removes the mismatched nucleotide within the unmethylated daughter DNA strand.

In addition to Dam, there are seven other proteins that make up the mismatch repair system in the bacterium *E. coli*: MutS, MutL, MutH, MutU, ExoI, DNA polymerase, and DNA ligase. These seven proteins function as follows:

1. The **MutS** protein slides along the DNA double helix and finds the mismatch.
2. The **MutH** protein binds to a nearby DNA sequence containing a methylated adenine in the template DNA strand. In essence, MutH is the protein that distinguishes the template from the daughter DNA strand.
3. The **MutL** protein binds to both MutS and MutH forming a loop in the DNA.
4. MutH is an endonuclease that makes a single-stranded DNA break in the backbone of the unmethylated (daughter) DNA strand. This break occurs between the G and A bases in the 5'-GATC-3' sequence in the daughter DNA strand.
5. The **MutU** protein binds to the MutH cut site in the daughter DNA strand. MutU has DNA helicase activity and functions to separate the daughter DNA strand from the parental DNA strand at the MutH cut site.
6. The **ExoI** protein is an exonuclease that degrades the damaged daughter DNA strand in the 5' to 3' direction starting at the cut site produced by MutH. Degradation of the damaged daughter DNA strand continues until ExoI removes the mismatched nucleotide.
7. The gap in the daughter DNA strand is filled in by a **DNA polymerase** (i.e., either the DNA polymerase holoenzyme or DNA polymerase I).
8. The final covalent bond in the newly synthesized daughter DNA strand is formed by **DNA ligase**.

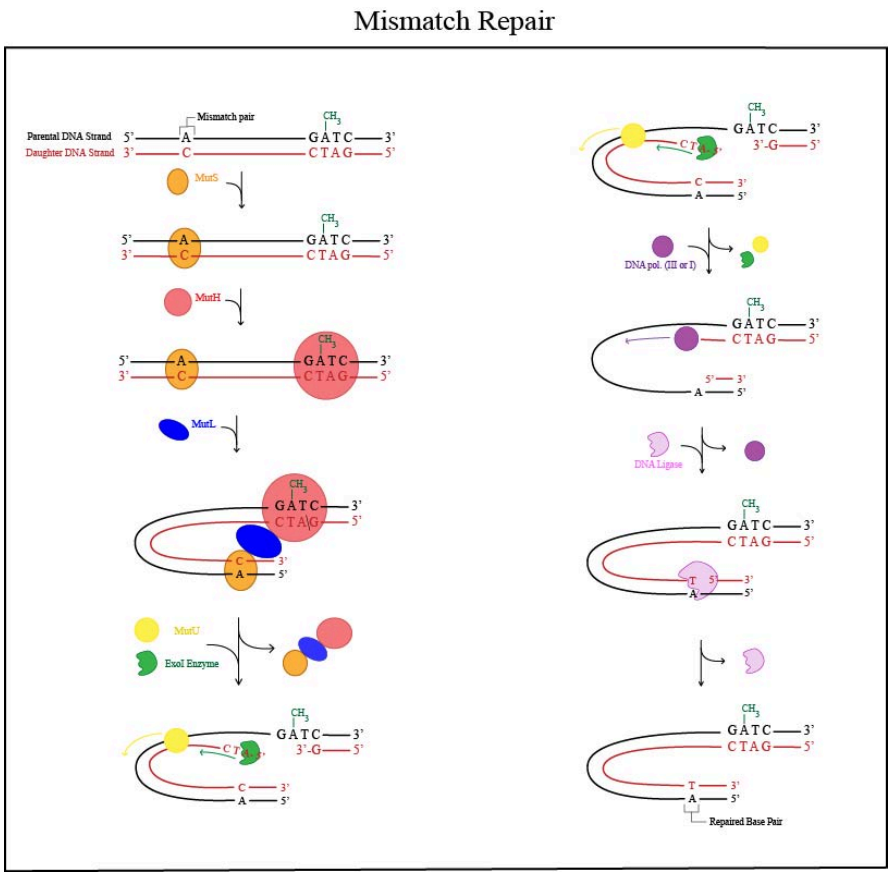


Figure 7.3 - Mismatch repair.

Key Questions

- What is the purpose of mismatch repair?
- Describe the functions of MutS, MutL, MutH, MutU, ExoI, DNA polymerase, and DNA ligase proteins in mismatch repair.
- Which one of the four Mut proteins is responsible for identifying the mismatched base?

Base Excision Repair (BER)

Base excision repair (BER) corrects abnormal nitrogenous bases that are sometimes formed in DNA. For example, occasionally cytosine in DNA can be spontaneously converted into uracil. Note that the conversion of cytosine to uracil is a transition mutation that does not distort the width of the DNA double helix. BER in the bacterium *E. coli* repairs this transition mutation as follows:

1. The enzyme **uracil DNA glycosylase (UNG)** recognizes uracil in the DNA. UNG cleaves the covalent bond that links uracil to deoxyribose (i.e., breaks the covalent bond attached to the 1’ carbon of deoxyribose). The uracil base is released from the rest of the nucleotide, creating an **abasic** site in the DNA. Note that the nucleotide at the abasic site still contains deoxyribose and the phosphate group; the nucleotide is just missing the uracil nitrogenous base.
2. The abasic site is detected by , **apurinic/apyrimidinic endonuclease 1 (APE1)**, which cleaves the phosphodiester bond at the 5’ end of the abasic site. This nick in the DNA backbone generates the free 3'-OH group required by DNA polymerase I.
3. **DNA polymerase I** uses its 5’ to 3’ exonuclease activity to remove the nucleotide (i.e., the deoxyribose sugar and phosphate group) at the abasic site. DNA polymerase I then uses its DNA synthesis activity (5' to 3' polymerase activity) to replace the removed nucleotide with the correct nucleotide.
4. **DNA ligase** forms the final phosphodiester bond in the DNA strand.

Note that there are similar pathways to repair other unconventional nitrogenous bases that are sometimes found in DNA (e.g. to remove the nitrogenous base hypoxanthine (H) that forms spontaneously from adenine). Instead of UNG, another gycosylase releases hypoxanthine; the other enzymes in BER work the same as described above.

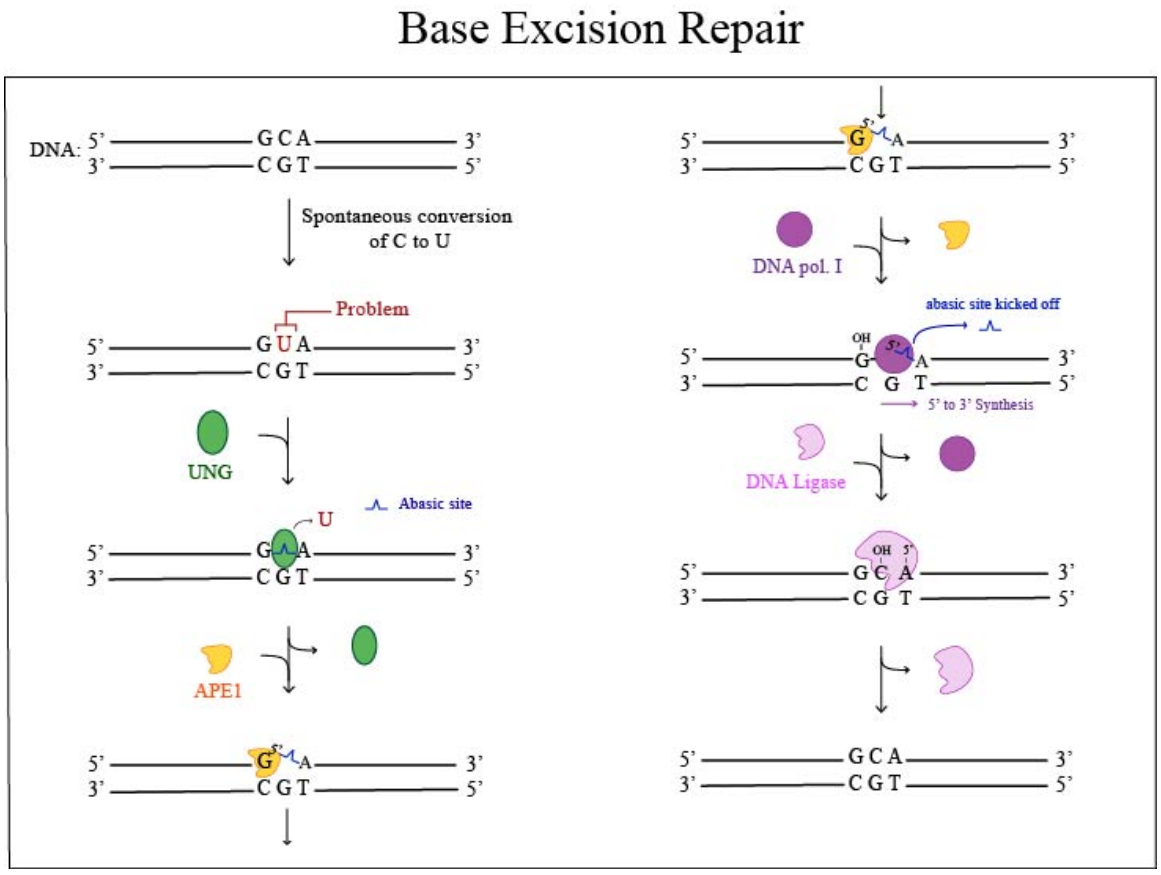


Figure 7.4 - Base excisior repair (BER).

Key Questions

- What is an abasic site?
- Explain the function of UNG, APE1, DNA polymerase I, and DNA ligase in base excision repair.

Nucleotide Excision Repair (NER)

Ultraviolet (UV) light exposure can lead to the formation of a **pyrimidine dimer** mutation in DNA that distorts the structure of the DNA double-helix. A pyrimidine dimer occurs when UV light causes additional covalent bonds to form between two adjacent pyrimidines in the same DNA strand. When this type of mutation occurs, the hydrogen bonds between the two adjacent pyrimidines and the bases in the other DNA strand are broken. During DNA replication or transcription, when the DNA or RNA polymerase encounters a pyrimidine dimer in the template DNA strand, the polymerases either stop replication or transcription altogether, or incorrect nitrogenous bases are placed in the synthesized strand opposite the pyrimidine dimer.

In *E. coli*, pyrimidine dimers can be repaired with the **nucleotide excision repair (NER)** system. NER involves six proteins: UvrA, UvrB, UvrC, UvrD, DNA polymerase, and DNA ligase. These six proteins remove a segment of DNA including the pyrimidine dimer and then replace the removed nucleotides via DNA replication. NER occurs as follows:

1. A complex consisting of two **UvrA** proteins and one **UvrB** protein scans the double stranded DNA in search of a pyrimidine dimer.
2. Once a pyrimidine dimer is identified, the UvrA/UvrB complex pauses over the dimer. The UvrA proteins are released, while **UvrC** attaches to UvrB at the dimer site.
3. UvrC is an endonuclease that cuts the damaged DNA strand on each side of the pyrimidine dimer.
4. **UvrD**, which is a DNA helicase, separates the two DNA strands, releasing the short segment of damaged DNA, including the pyrimidine dimer itself. UvrB, UvrC, and UvrD are released.
5. A **DNA polymerase** (i.e., either the DNA polymerase III holoenzyme or DNA polymerase I) fills in the gap, using the other DNA strand as a template.
6. **DNA ligase** forms the final covalent bond in the newly synthesized DNA.

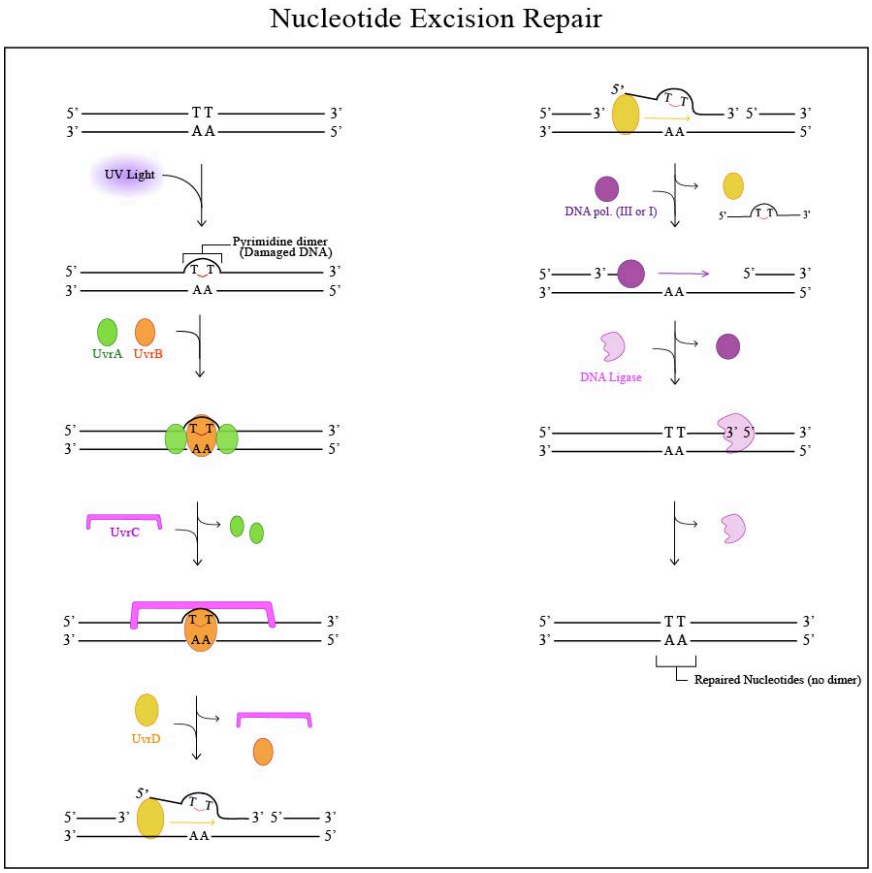


Figure 7.5 - Nucleotide excision repair (NER).

Key Questions

- Which nitrogenous bases can potentially form pyrimidine dimers?
- What are the functions of the UvrA, UvrB, UvrC, UvrD, DNA polymerase, and DNA ligase proteins in NER?
- Which Uvr protein is the helicase that unwinds and releases the damaged DNA?

Homology Directed Repair (HDR) (future content)

Nonhomologous End Joining (NHEJ) (future content)

Human Genetic Diseases Associated with Faulty DNA Repair

The DNA repair mechanisms in the bacterium *E. coli* described above have counterparts in human cells. Although the proteins are not identical, much of the repair process is similar. Mutations in the genes that give rise to mismatch repair enzymes have been noted in human cells. When both copies of the repair gene are inactivated, cancer can occur. Specifically, a type of inherited cancer called **hereditary nonpolyposis colorectal cancer** occurs because of mutations in the genes that produce human mismatch repair proteins. Mutations in the base excision repair system in humans have also been implicated in **colon cancer**.

Epithelial cells in the skin are constantly exposed to UV light from the sun. As a result, the nucleotide excision repair system plays an important role in repairing pyrimidine dimers that form in the DNA of epithelial cells. **Xeroderma pigmentosum (XP)**, an autosomal recessive disease, occurs when one of the seven human genes involved in nucleotide excision repair are inactivated by mutation. Xeroderma pigmentosum patients are sensitive to sunlight (due to the inability to repair DNA damage caused by UV light) and are predisposed to forming skin cancer.

Insert Xeroderma pigmentosum image here.

Key Questions

- What inheritance pattern is most often associated with diseases caused by defects in the DNA repair systems? Why do you think this is so?

C. Trinucleotide Repeat Expansions and Disease

In humans, there are DNA sequences in which a three nucleotides is repeated consecutively along a DNA strand (**trinucleotide repeat**). In most cases, a parent transmits these repeats to their offspring without any change in the repeat number. However, in a few human genetic diseases, this repeat can expand over the course of generations due to slippage of the DNA polymerases during replication. Once the expansion exceeds a threshold size, the function of the encoded protein is altered, causing disease.

Huntington's disease (HD) is an example of a neurodegenerative disease in humans caused by the expansion of a 5'-CAG-3' repeat within the structural gene ***HTT***. Although the exact function of the encoded HTT protein is unknown, it is believed that *HTT* plays an important role in the function of neurons. During protein synthesis, 5'-CAG-3' encodes the amino acid glutamine. In unaffected people, there can be between 10 and 35 repeats of the 5'-CAG-3' codon in the *HTT* gene. These repeats lead to a string of 10-35 glutamine amino acids within the encoded protein. As long as the number of 5'-CAG-3' repeats remains below 35, the encoded HTT protein functions normally. However, in some families, the number of repeats can extend beyond 35, ranging from 36 to 120 repeats. The longer string of encoded glutamine amino acids causes the HTT protein to degrade into small, toxic fragments. As these toxic protein fragments accumulate, neurons die prematurely. Over time, brain activity is altered, leading to symptoms of Huntington's disease, such as uncontrolled body movements, emotional problems, and decreased ability to learn and to make decisions. The number of 5'-CAG-3' repeats in the *HTT* gene correlates with the severity of the disease; a patient with more 5'-CAG-3' codon repeats typically has more severe symptoms than a person with fewer repeats.

Insert a Huntington's disease figure here.

Key Questions

- Why do trinucleotide expansions above a certain threshold cause human disease?

Review Questions

Fill in the blank:

1. A bulge in the DNA double strand occurs when two _____ form a base pair.
2. Exposure to radiation and chemicals is responsible for causing _____ mutations.
3. A single nucleotide substitution in a structural gene that exchanges an amino acid for a stop codon is called a _____ mutation.
4. A deletion of four nucleotides from the coding region of a gene would result in a _____ mutation.
5. In each DNA repair system, _____ forms the final covalent bond in the damaged DNA strand after the correct nucleotide has been added.
6. If a mutation occurs in the DNA such that guanine is paired with uracil, then the most likely DNA repair system to recognize and correct this mutation would be _____ .
7. In mismatch repair, _____ recognizes and binds to the methylated adenine on the parental DNA strand.
8. _____ is the DNA helicase used in nucleotide excision repair.
9. _____ is an example enzyme that has 3' – 5' exonuclease activity.
10. Xeroderma pigmentosum is caused because of a defect in the _____ repair system.
11. CAG encodes the amino acid _____ and when the number of CAG repeats exceeds _____, neurological symptoms appear consistent with Huntington's disease.

8 - Polymerase Chain Reaction (PCR)

PCR Reaction Components

The **polymerase chain reaction (PCR)** is essentially DNA replication in a test tube. In the laboratory, PCR can be used to copy or **amplify** any DNA sequence of interest. PCR has a myriad of applications. For example, PCR can be used in forensics to make copies of the DNA molecules left by a suspect at a crime scene, scientists can use PCR to make many copies of a gene to study gene structure and function, and finally, PCR can be used to determine if an individual is infected with a microbe, such as the human immunodeficiency virus (HIV) or SARS-CoV-2 (the virus that causes COVID-19).

Suppose our goal is to study the human insulin gene. We could use PCR to make millions of copies of the insulin gene as part of our research. To accomplish this goal, our PCR reaction would contain the following components:

- **Template (target) DNA.** This double-stranded template DNA molecule includes the gene or segment of DNA that is to be copied (amplified). In our scenario, the template DNA would be a fragment of chromosome 11 that contains the human insulin gene.
- **DNA primers.** These two single-stranded DNA primers (each approximately 20 nucleotides in length) are designed to bind to the template DNA molecule on each side of the gene that will be amplified by PCR. One DNA primer would bind to the left of the insulin gene on one of the two template DNA strands, the other DNA primer would bind to the right of the insulin gene on the other template DNA strand.
- **dNTPs** (dATP, dCTP, dTTP, and dGTP). The dNTPs provide the energy for DNA replication in PCR. Moreover, the nitrogenous base, the deoxyribose sugar, and one of the phosphate groups in each dNTP is incorporated into the growing daughter DNA strand.
- **Thermostable DNA polymerase.** PCR uses a DNA polymerase called ***Taq* polymerase**, isolated from the thermophilic bacterium *Thermus aquaticus. Taq* polymerase is a thermostable DNA polymerase that functions similarly to DNA polymerase I from *E. coli*. A thermostable DNA polymerase is used because PCR involves a series of heating steps (see below) that would denature other DNA polymerases, including all of the DNA polymerases introduced in Part 6.
- **Buffer.** The buffer maintains the proper pH for the PCR reaction and contains enzyme cofactors, such as magnesium ions, necessary for *Taq* polymerase activity.

Key Questions

- What are the functions of the five components of a PCR reaction?

PCR Cycles

Once the PCR reaction is prepared containing each of the five components above, the PCR reaction is subjected to multiple **PCR cycles** (see **figure 8.1**). Each PCR cycle has the following three steps:

1. **The template (target) DNA (i.e., the human insulin gene) is denatured by heat treatment (95°C for 1 minute).** The denaturing step breaks the hydrogen bonds within the target DNA molecule, causing the individual template DNA strands to separate from each other.
2. **The DNA primers anneal (form hydrogen bonds) with complementary sites on the template DNA strands (55°C for 1 minute).**
3. ***Taq* polymerase synthesizes DNA (72°C for 2 - 3 minutes).** *Taq* polymerase synthesizes the daughter DNA strands by adding dNTPs to the free 3'-OH groups provided by the two DNA primers This DNA synthesis step is sometimes called the primer **extension** step.

Each PCR cycle described above is repeated typically 30-35 times. The number of DNA copies (i.e., the number of copies of the human insulin gene) is doubled at the conclusion of each cycle. The total number of template (target) DNA copies made during PCR can be estimated using the following equation:

$$n = a \times 2^b$$

- n = the number of double-stranded template (target) DNA copies made during PCR (i.e., the number of copies of the human insulin gene).
- a = the number of template (target) DNA copies at the beginning of the PCR experiment. Often this is assumed to be a single template DNA molecule (i.e., a single human insulin gene).
- b = the number of PCR cycles that have occurred.

In the lab, the PCR cycles are accomplished by mixing the above five PCR reaction components in a test tube, followed by placing the PCR reaction in a **thermocycler** device. The thermocycler automates the number of cycles, the temperature of each step within a cycle, and the length of each step within a cycle.

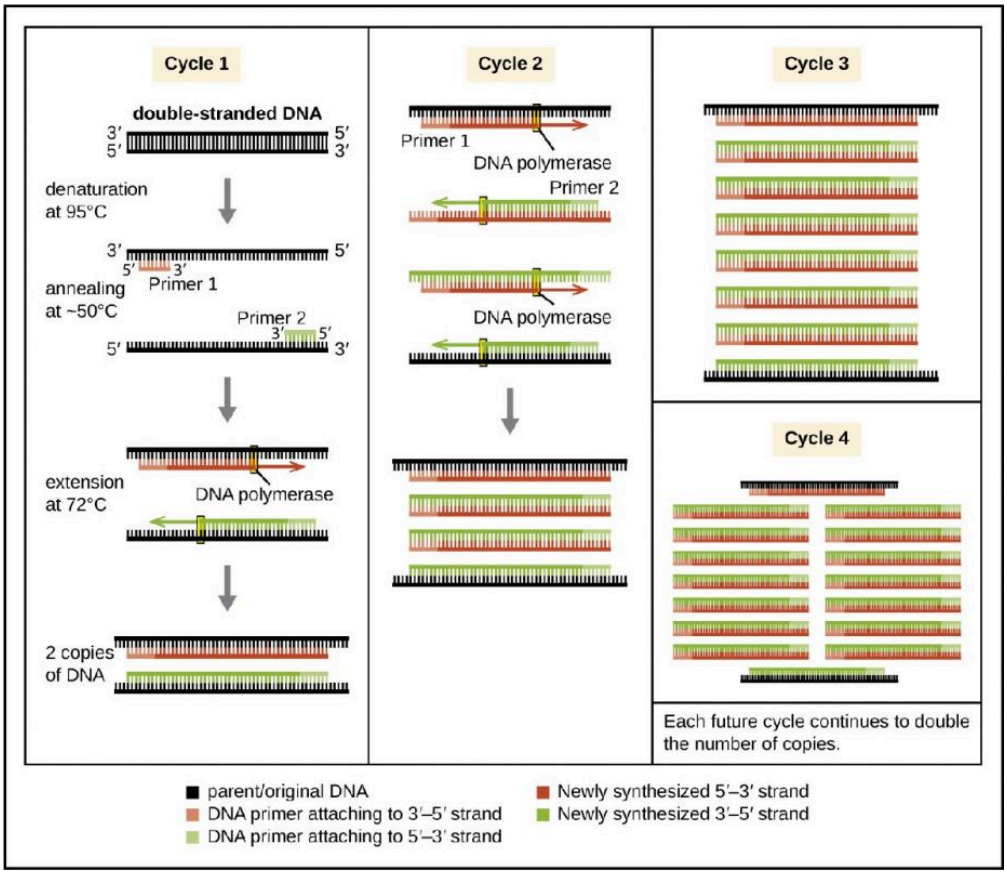


Figure 8.1 PCR Cycles. *Each PCR cycle includes a denaturing step in which the template DNA strands separate, a primer annealing step, and a DNA synthesis (extension) step in which the daughter DNA strands are synthesized by Taq polymerase. The number of double-stranded DNA copies is doubled at the end of each cycle. --- Image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction*

Key Questions

- What is happening in each of the three steps in a typical PCR cycle?
- Describe how PCR can be used to amplify a gene of interest.

Agarose Gel Electrophoresis

Agarose gel electrophoresis is used in the laboratory to visualize DNA. For our purposes, agarose gel electrophoresis can be used to determine whether the PCR experiment successfully copied (amplified) the human insulin gene. Since agarose gel electorphoresis separates DNA molecules from each other based upon size, agarose gel electrophoresis can also be used to analyze the size (in base pairs) of the insulin PCR products (see **Figure 8.2**).

To perform an agarose gel electrophoresis experiment:

1. **The agarose gel is prepared.** To prepare an agarose gel, agarose (a polysaccharide powder purified from marine algae) is dissolved in a buffer solution by heating in a microwave. This melted agarose solution is then poured into a plastic mold and allowed to cool to form a gel. The resulting agarose gel contains a meshwork (matrix) of agarose polymers. The consistency of the matrix depends on the concentration of agarose used. High agarose concentrations produce a matrix with small pores, useful for separating small DNA molecules. Low agarose concentrations produce a matrix with larger pores, useful for separating large DNA molecules. Moreover, a fluorescent dye called **ethidium bromide** is added to the agarose gel before it solidifies. Ethidium bromide binds to the DNA within the agarose gel and fluoresces orange in the presence of ultraviolet light. Thus, ethidium bromide provides a convenient way to visualize the DNA molecules as they separate within the agarose gel.
2. **The PCR products are loaded into depressions (wells) created at one end of the gel.** A sample called a **molecular marker (DNA ladder)** is also loaded into one well of the gel. This molecular marker sample contains DNA fragments of known size for comparison to the PCR product.
3. **An electric field is applied across the agarose gel.** This electric field causes charged molecules to migrate from one end of the agarose gel to another. Typically, the end of the gel that contains the wells is placed near the negative electrode; the opposite end of the gel is placed near the positive electrode. Since DNA is negatively charged, DNA will migrate from the wells into the agarose gel. DNA migration continues through the agarose gel matrix towards the positive electrode.
4. **The DNA samples separate according to size.** Smaller DNA molecules migrate farther through the agarose gel matrix than larger DNA molecules.
5. **The DNA is visualized using ultraviolet light**. The ethidium bromide in the agarose gel binds to the DNA as the DNA migrates through the gel. The ethidium bromide bound to the DNA fluoresces orange in the presence of ultraviolet light.

If PCR amplification of the insulin gene was successful, a band will be seen in the agarose gel when exposed to ultraviolet light.

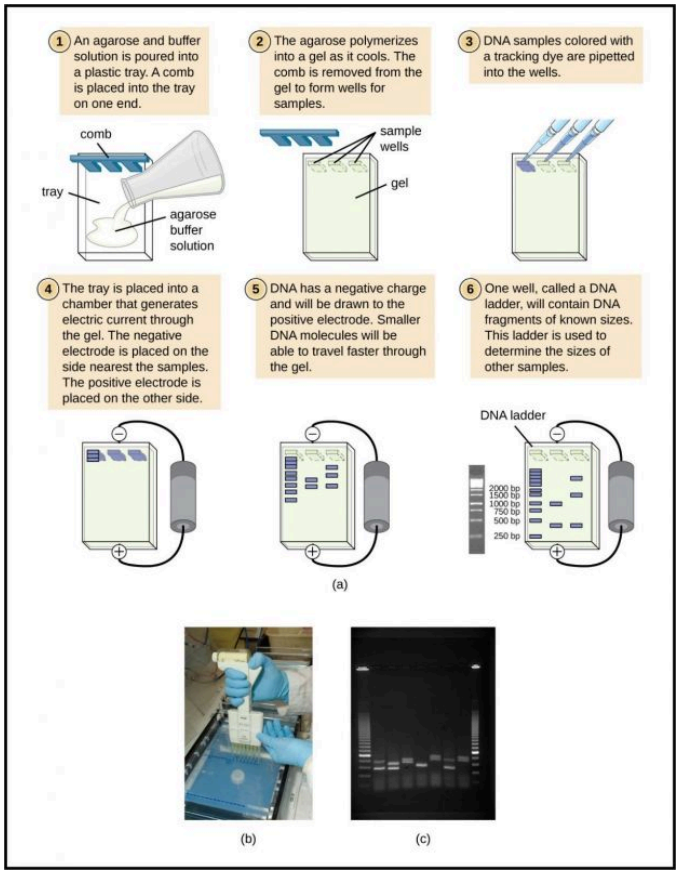


Figure 8.2 ***Agarose Gel Electrophoresis*** *--- Image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction*

Key Questions

- What is an agarose gel?
- Describe how DNA molecules separate in an agarose gel.
- Describe how agarose gel electrophoresis can be used to determine if the insulin gene was successfully amplified by PCR.

Reverse Transcription PCR (Future Content)

PCR can also be used to amplify a particular mRNA molecule; however, mRNAs are not used directly as templates for the PCR reaction. To amplify a mRNA molecule, **reverse transcription** is done prior to PCR. Reverse transcription converts the mRNA molecules within a cell into a collection of DNA molecules called **complementary DNAs (cDNAs)**. These cDNAs can then be used as templates in PCR.

Reverse transcription to produce cDNAs involves the following steps:

1. Purify the mRNA molecules from a cell.
2. Mix the mRNA molecules with a **poly-dT primer**. The mRNA molecules produced from eukaryotic genes contain approximately 250 adenine bases at the 3' end of the mRNA (see the Part 10). This sequence of 250 A bases is called a **poly-A tail.** The poly-dT primer is composed of a series of T nucleotides that form hydrogen bonds with the poly-A tails on these mRNA molecules. The poly-dT primer also provides the free 3'-OH group required for DNA synthesis.
3. Add **reverse transcriptase** and **dNTPs**. Reverse transcriptase is a DNA polymerase isolated from certain types of viruses, including the human immunodeficiency virus. Reverse transcriptase is a unique DNA polymerase because it uses an RNA template to synthesize DNA (the other DNA polymerases we have studied use a DNA template to synthesize DNA). The newly synthesized strand of DNA generated by reverse transcriptase is complementary (forms hydrogen bonds) to the mRNA. Thus, after reverse transcription, an RNA:DNA double-helix is present in the reaction.
4. Add **ribonuclease H (RNase H)** to digest portions of the mRNA. RNase H is a viral endonuclease that cuts phosphodiester bonds within the RNA component of an RNA:DNA double-helix . One consequence of RNase H treatment is that the mRNA is cleaved into fragments; each mRNA fragment is essentially a primer, containing the free 3'-OH groups required to synthesize a second DNA strand.
5. Add **DNA polymerase I** and **DNA ligase** from *E. coli* to synthesize the second DNA strand to form a double-stranded cDNA molecule. Recall that DNA polymerase I has both 5' to 3' exonuclease and 5' to 3' polymerase activities that replace the RNA bases with DNA bases, while DNA ligase forms the final phosphodiester bonds to convert Okazaki fragments into a continuous DNA strand.

Reverse transcription produces a collection of double-stranded cDNA molecule that correspond to the entire mRNA collection within a cell. This collection of various cDNA molecules can then serve as the templates for PCR. Since the primers used in PCR are typically designed to be specific for a particular gene sequence, PCR amplification copies of a single cDNA type, corresponding to a single type of mRNA. In essence, this **reverse transcription PCR** process has amplified a single type of mRNA from the cell.

Insert a figure here that illustrates reverse transcription PCR.

Key Questions

- What is the starting material for reverse transcription?
- What are cDNAs?
- Describe the function of the poly-dT primer, reverse transcriptase, RNase H, DNA polymerase I, and DNA ligase during reverse transcription.

Real-time PCR (Future Content)

Often, the goal of PCR is to make many copies of a particular DNA sequence. The success of the PCR experiment is determined by analyzing the PCR product on an agarose gel, as described above. In other cases, the goal of PCR is to determine how many copies of the template DNA molecule are present at the beginning of the PCR experiment, before PCR amplification occurs. To determine the number of template DNA molecules present in a sample, a modification of PCR, called **real-time PCR** or **quantitative PCR (qPCR)** is used.

Real-time PCR allows a scientist to monitor the production of PCR products in real-time (i.e., as the reaction is occurring in the thermocycler). If the concentration of template DNA molecules in the reaction is low prior to the start of PCR, then it takes more PCR cycles to produce the number of PCR products required for detection. Alternatively, if the concentration of template DNA molecules in the reaction is high prior to the start of PCR, then detectable products are

formed in earlier PCR cycles. Thus, the real-time PCR technique is especially powerful, as it allows researchers to quantitatively measure the concentration of template DNA sequences. Moreover, if the template molecules are cDNA molecules produced from mRNA by reverse transcription, then real-time PCR can provide a quantitative measure of how actively a gene is transcribed (the more active the gene, the more mRNA molecules are produced by the gene).

Key Questions

- What is an advantage of using real-time PCR?

TaqMan (Future Content)

Real-time PCR involves a modification to the conventional PCR approach. In addition to the five PCR components described earlier, a real-time PCR reaction contains a **probe** molecule. A common probe that is used in many real-time PCR reactions is **TaqMan**. The TaqMan probe is a short DNA molecule that is designed to form hydrogen bonds to one of the two template DNA strands downstream (in the 3' direction) from one of the two DNA primers. The 5’ and 3’ ends of the TaqMan molecule have been modified; the 5’ end of TaqMan is attached to a fluorescent **reporter** molecule, while the 3’ end of TaqMan is attached to a **quencher** molecule. When the reporter and the quencher are in close proximity (i.e., within the same probe molecule), the quencher molecule inhibits the fluorescence produced by the reporter molecule. When the reporter molecule is separated from the quencher molecule (i.e., when the probe is degraded by the 5' to 3' exonuclease activity of *Taq* DNA polymerase), fluorescence occurs.

During the primer annealing step in a real-time PCR reaction, both a primer and the TaqMan probe binds to one of the two template DNA strands. During the DNA synthesis step, *Taq* polymerase synthesizes the daughter DNA strand toward the TaqMan probe bound to the template DNA strand. When *Taq* polymerase encounters the TaqMan probe, *Taq* polymerase uses its 5’ to 3’ exonuclease activity to begin digesting TaqMan. As the TaqMan probe is digested, the reporter molecule is separated from the quencher molecule and fluorescence occurs. If the number of template DNA molecules in the reaction is low, fluorescence is low in early PCR cycles and begins to be detectable in later PCR cycles. If the number template DNA molecules in the reaction is high, fluorescence is detectable in the earlier PCR cycles.

The thermocycler used in real-time PCR is modified to detect the fluorescence emitted by the reporter molecule as it is released from the quencher molecule during DNA replication.

Insert a figure here that illustrates a real-time PCR cycle.

Key Questions

- What are the six major components of a real-time PCR reaction?
- What are the functions of the two parts of a TaqMan probe?
- Describe how real-time PCR using a TaqMan probe can be used to determine the concentration of template DNA molecules in a reaction.

9 - Transcription

When a gene is activated, the gene is **transcribed**, producing an RNA intermediate. **Structural genes** are genes that are transcribed to produce **messenger RNA (mRNA)** molecules. The mRNA molecule is then **translated** to make a protein product. **Nonstructural genes** are also transcribed to produce RNA molecules; however, the RNA molecule is not translated and instead functions directly in the cell. These functional RNA molecules, called **noncoding RNAs (ncRNAs),** include transfer RNA molecules (tRNAs), ribosomal RNA molecules (rRNAs), and the *Xist* and *Tsix* RNA molecules discussed in Part 2.

Key Questions

- What is the difference between a structural and a nonstructural gene?

A. Transcription in Bacteria

Expression of Structural Genes

What factors determine whether a bacterial structural gene is **expressed**; in other words, the gene is activated to make a mRNA molecule? Gene expression requires the interaction between **transcription factor** proteins and specific DNA sequences near the gene.

The DNA sequences that regulate the expression of a particular structural gene include (see **figure 9.1**):

- **Regulatory DNA sequences.** Regulatory DNA sequences influence how often transcription starts. These DNA sequences are located, in most cases, adjacent to a structural gene.
- **The promoter sequence.** The promoter consists of DNA sequences that determine where transcription starts. The promoter is typically adjacent to the controlled structural gene.
- **The terminator sequence.** The terminator consists of DNA sequences that signal termination of transcription by causing the RNA polymerase to dissociate from the DNA.

Transcription produces an RNA molecule that is complementary to the **template** or **antisense** strand of DNA. The other DNA strand, the one that forms hydrogen bonds with the template DNA strand, is called the **coding** or **sense** DNA strand. The coding DNA strand is identical in sequence to the RNA transcript, except that the RNA molecule contains uracil (U) instead of thymine (T).

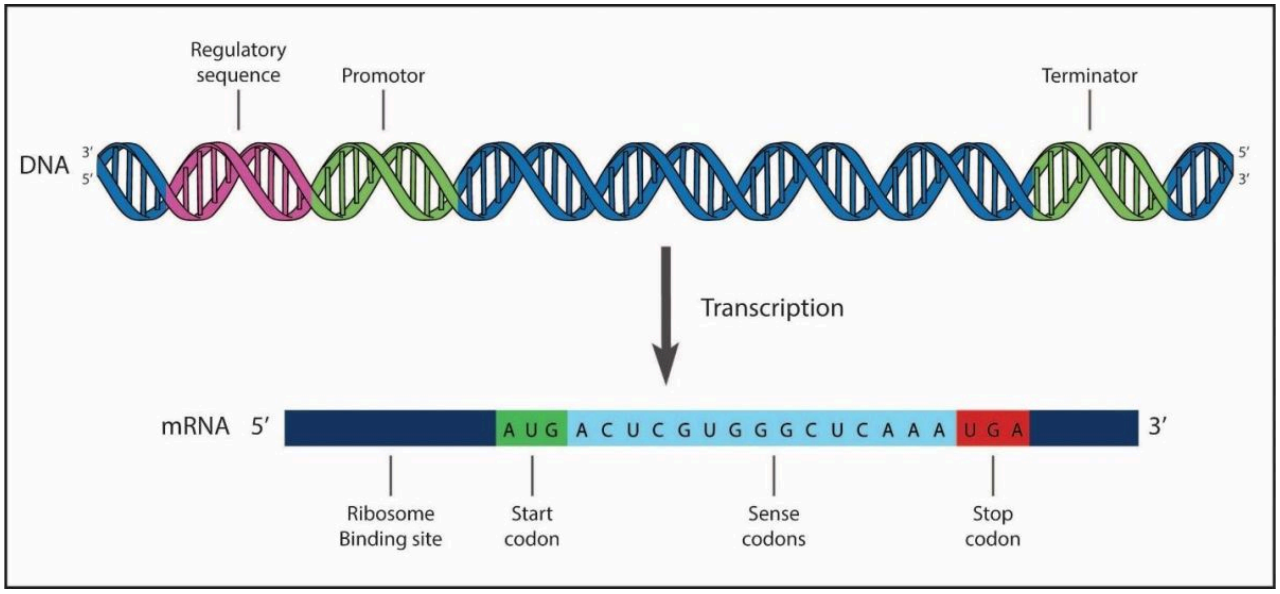


Figure 9.1 ***DNA sequences that control transcription.*** *--- Image created by SL*

Key Questions

- What are the functions of the three DNA sequences that regulate transcription?
- What is the difference between the template and the coding DNA strands?

Transcription Stages

Transcription of structural genes in the bacterium *E. coli* has the following three stages (see **figure 9.2**):

1. **Initiation.** During the initiation stage of transcription in bacteria, a transcription factor protein called **sigma (σ) factor** guides the **RNA polymerase** to the promoter.
2. **Elongation.** During elongation, the RNA polymerase bound to the promoter acts as a DNA helicase, separating the two DNA strands, forming an **open complex.** RNA polymerase then reads the template DNA strand while synthesizing a complementary mRNA transcript.
3. **Termination.** The termination stage of transcription involves the release of the RNA polymerase and the mRNA molecule from the DNA.

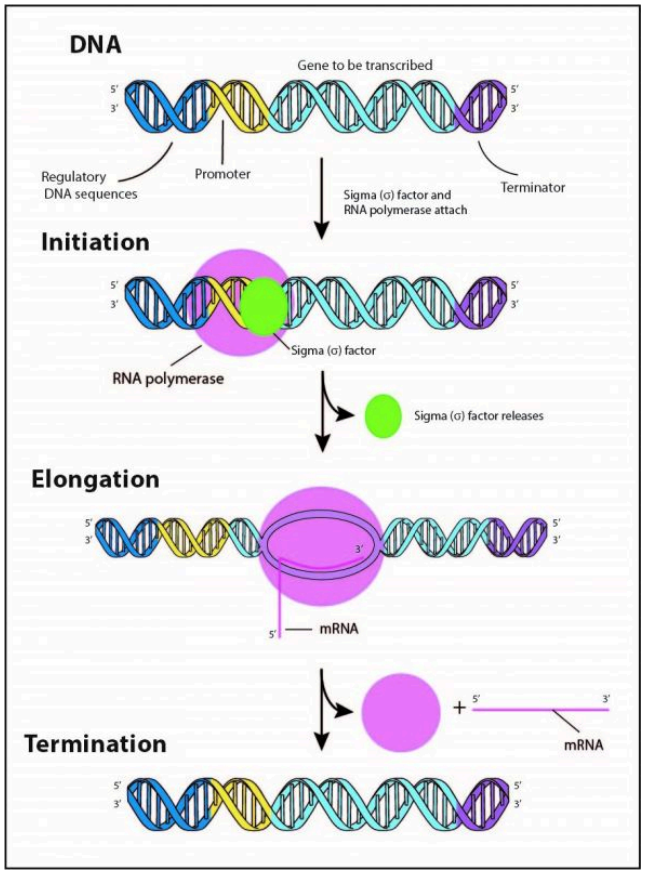


Figure 9.2 ***Transcription Overview*** *--- Image created by SL.*

Key Questions

- What is happening during the three stages of transcription?
- What is the function of sigma factor?

Promoter Structure in Bacteria

The bacterial **promoter** is located **upstream** (typically drawn to the left) of the structural gene to be transcribed and serves as a docking site for the sigma (σ) factor protein and later RNA polymerase. DNA sequence elements within the promoter are numbered relative to the **+1 site**, the first nucleotide in the template DNA strand that is transcribed (see **figure 9.3**). Important DNA sequences within the bacterial promoter include the following:

- **-35 sequence.** The -35 sequence (5'-TTGACA-3' in the coding DNA strand) allows high transcription rates because it serves as part of the binding site for the **sigma (σ) factor** protein. The -35 sequence is located approximately 35 base pairs (bp) upstream of the transcription start site (+1 site).
- **-10 sequence.** The -10 sequence (5'-TATAAT-3' in the coding DNA strand) is essential for transcription in prokaryotes because it serves as the second part of the sigma factor binding site. Moreover, the -10 sequence is AT-rich, promoting the separation of the two DNA strands, a requirement for transcription. The -10 sequence is located approximately 10 bp upstream of the transcription start site (+1 site).
- **+1 site.** The +1 site is the **transcription start site**. The nitrogenous base at the +1 site in the coding DNA strand is usually adenine (A). Since the mRNA and the coding DNA strand have the same sequence, the first nitrogenous base in the mRNA is also adenine (A).

Both the -35 and -10 sequences described above (5'-TTGACA-3' and 5'-TATAAT-3') are **consensus sequences**, meaning that they are the "average" sequences found when the DNA sequences of many *E. coli* promoters are compared. Some bacterial promoters are **strong promoters**, whereas others are **weak promoters**. The difference between strong and weak promoters largely depends on how closely the promoter DNA sequence in question matches the -35 and -10 consensus sequences. Strong promoters initiate transcription frequently, while weak promoters initiate transcription less frequently.

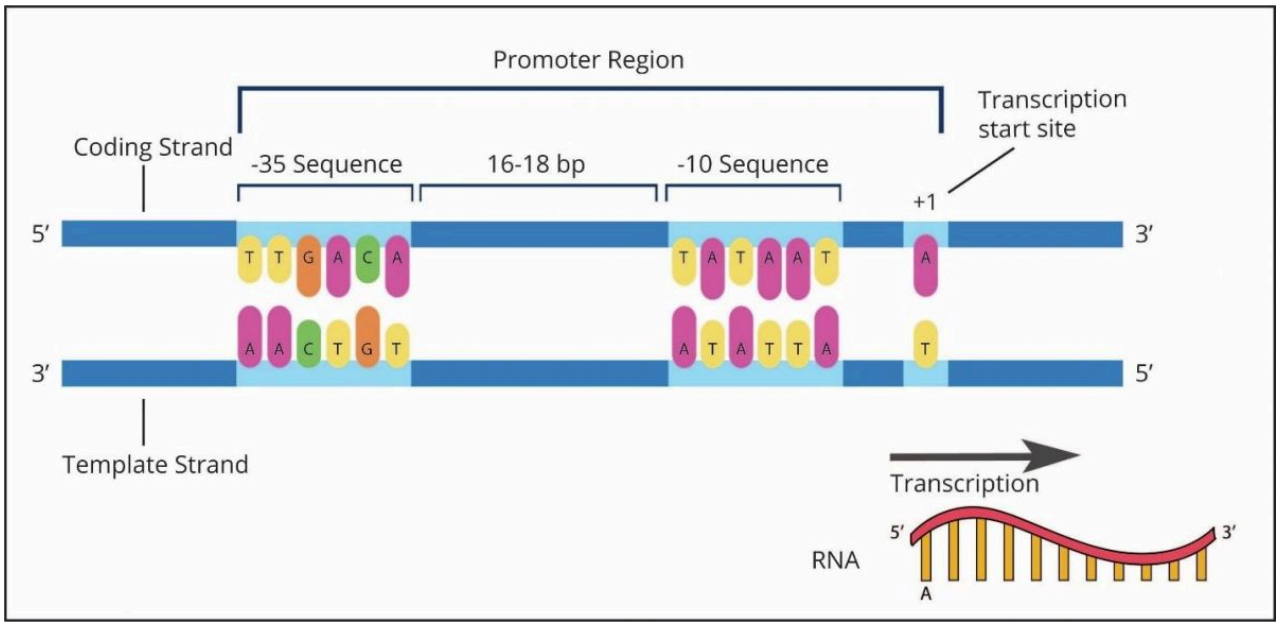


Figure 9.3 ***Bacterial Promoter*** *--- Image created by SL*

Key Questions

- What is the function of the -35 sequence?
- What are the two functions of the -10 sequence?
- What is the function of the +1 site?
- What is the difference between a strong and a weak promoter?

Bacterial RNA Polymerase

In the bacterium *E. coli*, the **RNA polymerase core enzyme** is composed of five protein subunits (α_1, α_2, β, β', and ω) (see **figure 9.4**). The two α subunits and the ω subunit function to assemble the enzyme and bind to the DNA sequence to be transcribed. The RNA molecule is synthesized between the β and β' subunits.

The RNA polymerase core enzyme (α_1, α_2, β, β', and ω subunits) associates with the **sigma (σ) factor** protein to form the **RNA polymerase holoenzyme**. *E. coli* makes at least eight different types of sigma factor proteins, depending on the environmental conditions encountered by the cell. For example, the main sigma factor in *E. coli* is called the **housekeeping sigma factor** or **σ^{70}** protein. The σ^{70} protein functions to guide the RNA polymerase core enzyme to the promoters of structural genes required for the viability of the *E. coli* cell in a typcial environment (e.g., body temperature with plenty of carbon and nitrogen sources). In addition to the σ^{70} protein, there are specialized sigma factor proteins that guide the RNA polymerase core enzyme to survival genes when an *E. coli* cell encounters stressful environments. These specialized sigma factors include a nitrogen starvation sigma factor (σ^{54}), a carbon starvation sigma factor (σ^{38}), and a heat shock sigma factor (σ^{32}). Because the sigma (σ) factor proteins regulate transcription, the sigma (σ) factor proteins are example **transcription factor proteins**

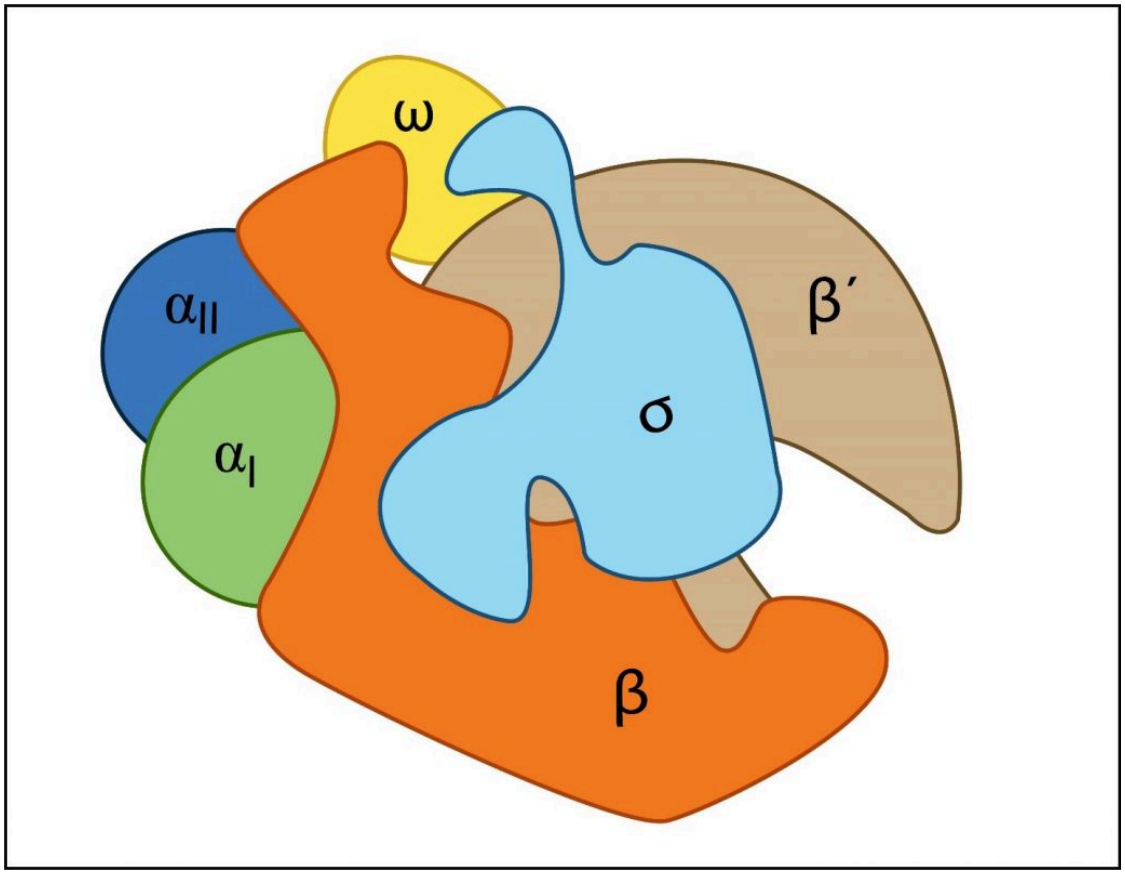


Figure 9.4 ***RNA Polymerase Holoenzyme Subunits*** *--- Image created by SL*

Key Questions

- Why does *E. coli* make several different types of sigma factor proteins?
- What is the difference between the RNA polymerase core enzyme and the RNA polymerase holoenzyme?

Transcription Initiation in Bacteria

Transcription initiation in bacteria (*E. coli*) occurs as follows:

1. The RNA polymerase holoenzyme recognizes the promoter via sigma (σ) factor binding to the -35 and -10 DNA sequences. At this stage, the RNA polymerase holoenzyme:DNA complex is called a **closed complex** because the two DNA strands are still hydrogen bonded together.
2. The AT hydrogen bonds within the -10 sequence are broken forming an **open complex**. The RNA polymerase core enzyme is the DNA helicase that separates the two DNA strands at the -10 sequence.
3. A short RNA molecule is synthesized beginning at the +1 sequence; however, the RNA polymerase core enzyme is still attached to σ factor. Sigma (σ) factor is still bound to the -10 and -35 DNA sequences.
4. The sigma (σ) factor protein is released, freeing the RNA polymerase core enzyme.
5. Once the sigma (σ) factor protein is released, transcription transitions to the elongation phase as the RNA polymerase core enzyme incorporates additional nucleotides at the 3’ end of the RNA transcript.

Key Questions

- Describe the initiation phase of transcription in bacteria.

Elongation in Bacteria

The elongation phase of transcription in bacteria involves RNA synthesis by the RNA polymerase core enzyme (see **figure 9.5**). The *E. coli* RNA polymerase core enzyme has the following features:

- The RNA polymerase core enzyme does not require a primer for RNA synthesis (in other words, no 3'-OH group is required to initiate transcription). Because the RNA polymerase core enzyme has no primer requirement, the first nucleotide incorporated into the mRNA has three phosphate groups attached to the 5' carbon.
- The RNA polymerase core enzyme has DNA helicase activity, separating the two DNA strands during transcription elongation.
- The RNA polymerase core enzyme reads the template DNA strand in the 3' to 5' direction.
- The RNA polymerase core enzyme synthesizes the mRNA in the 5' to 3' direction.
- The RNA polymerase core enzyme catalyzes the formation of a covalent bond between the 3'-OH of the growing RNA strand and the 5' phosphate group on the incoming **nucleoside triphosphate (NTP)**. The NTPs used by the RNA polymerase core enzyme are ATP, UTP, CTP, and GTP. The NTP molecules are cleaved during transcription, releasing pyrophosphate (PP_i) during the RNA synthesis reaction.
- RNA synthesis follows the AT/GC rule except that uracil is found in RNA (in other words, transcription follows the AU/GC rule).
- The RNA polymerase core enzyme does not have proofreading activity (no 3' to 5' exonuclease activity). As a result, the mRNA molecule made during transcription sometimes contains mistakes.
- The RNA polymerase core enzyme reforms hydrogen bonds within the two DNA strands after the open complex has passed by. As the RNA polymerase core enzyme rewinds the DNA double helix, the RNA transcript trails behind the core enzyme as a single-stranded RNA molecule.

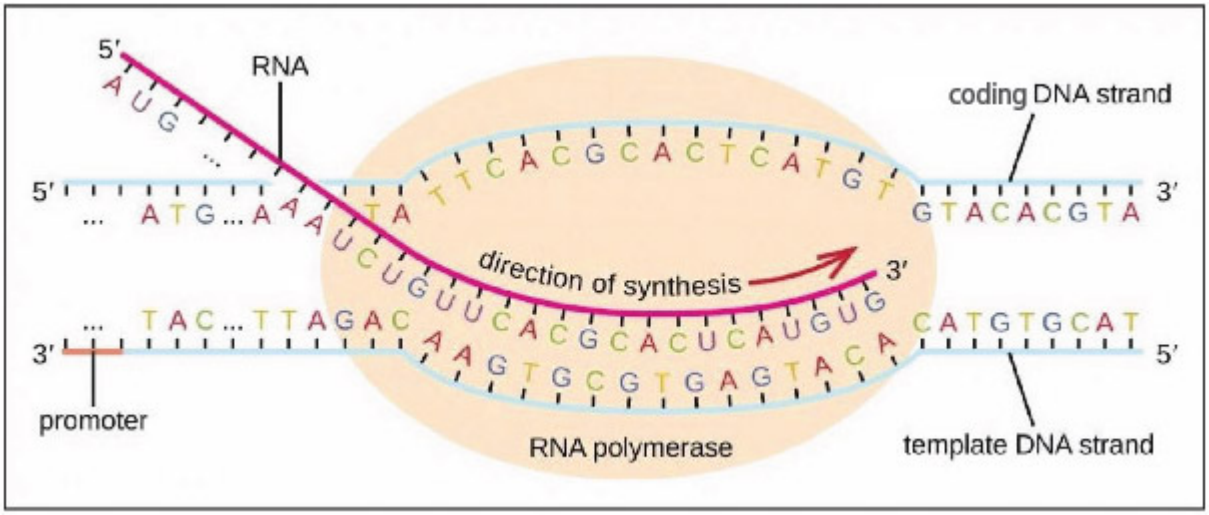


Figure 9.5 ***Transcription Elongation*** *--- Image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction)*

Key Questions

- What are the similarities and differences between the RNA polymerase core enzyme and the DNA polymerases discussed in Part 6?
- Which protein functions as the DNA helicase for transcription?
- What molecules provide the energy for transcription?

Transcription of Multiple Genes

Not all genes use the same DNA strand as the template strand. In **figure 9.6**, genes A and B use the bottom DNA strand as the template strand for RNA synthesis, because the promoter is located to the left of the gene. Alternatively, gene C

uses the top DNA strand as the template strand, as the promoter is located to the right of the gene. Genes A and B are transcribed left to right, while gene C is transcribed right to left.

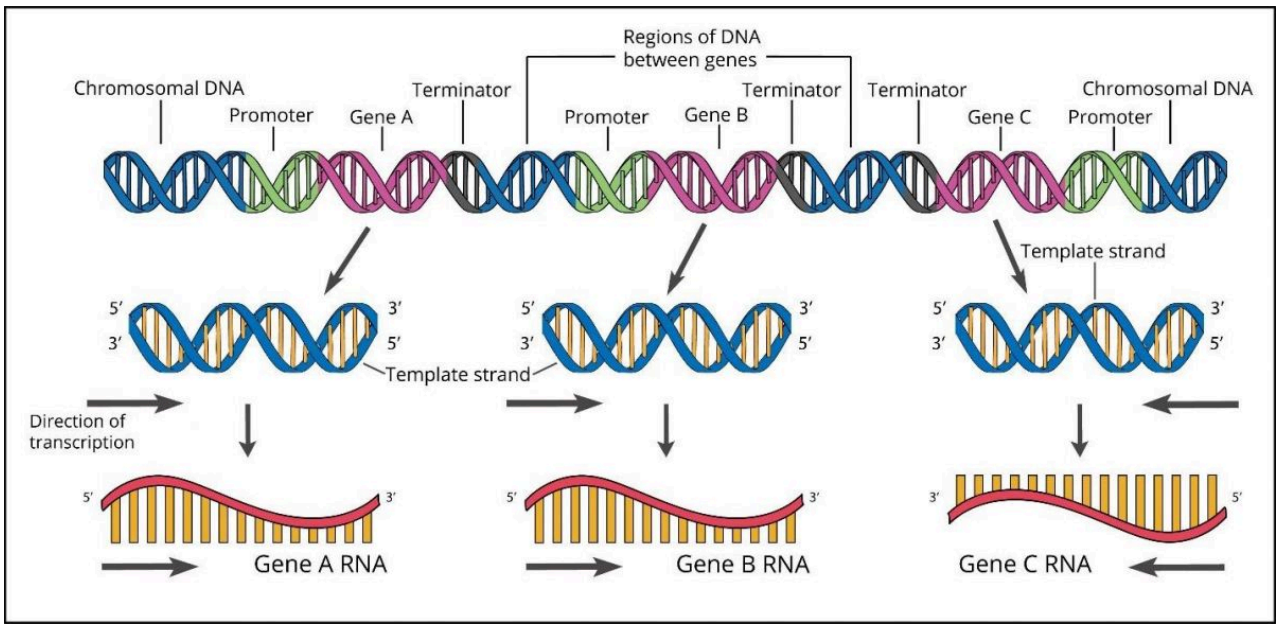


Figure 9.6 ***Transcription of Multiple Genes*** *--- Image created by SL*

Rho (ρ)-Dependent Termination

While the RNA polymerase core enzyme is synthesizing a mRNA molecule, an RNA-DNA double helix molecule is formed within the enzyme. Transcriptional termination involves weakening the hydrogen bonds within this RNA-DNA double helix, resulting in dissociation of the RNA (and the RNA polymerase core enzyme) from the DNA.

Transcriptional termination can occur in two different ways in the bacterium *E. coli*:

- **Rho (ρ)-dependent termination**
- **Rho (ρ)-independent termination**

The rho (ρ)–dependent mechanism of termination requires binding between the **rho (ρ) protein**, a helicase that breaks the hydrogen bonds within an RNA-DNA double helix, and an RNA sequence near the 3’ end of the mRNA transcript called the **rho utilization site (*rut*)** (see **figure 9.7**). The ρ-dependent mechanism of transcription termination also requires the formation of a secondary structure within the RNA transcript called a **stem-loop** or **hairpin loop**. The stem-loop is formed when guanine (G) and cytosine (C) bases are produced in the mRNA as the RNA polymerase core enzyme reads the terminator DNA sequence. The stem-loop, composed of hydrogen bonds between these G and C nucleotides within the same mRNA molecule, slows the RNA polymerase core enzyme during transcription. The rho (ρ) protein then catches up with the RNA polymerase, separates the RNA from the template DNA strand, and releases the RNA transcript and the RNA polymerase core enzyme from the DNA. Transcription is terminated.

Key Questions

- What three components are involved in rho (ρ)-dependent termination?
- What are the functions of each of these components in rho (ρ)-dependent termination?

Rho (ρ)-Independent Termination

The rho (ρ)-independent termination mechanism does not require rho (ρ) protein or the *rut* RNA sequence (see **figure 9.7**). In rho (ρ)-independent termination of transcription, a stem-loop structure is formed in the newly synthesized RNA that slows the RNA polymerase core enzyme. This pausing of the RNA polymerase is aided by the **NusA** protein. While the RNA polymerase slows down, a uracil-rich region is synthesized in the RNA because the RNA polymerase core enzyme is copying an adenine-rich region in the template DNA strand. Recall that each uracil base in the mRNA forms two hydrogen bonds with each adenine base in the template DNA strand. This weak base pairing between U and A bases tends to break spontaneously, releasing the mRNA and RNA polymerase, terminating transcription.

The mechanism that is used for transcription termination depends on the gene. About 50% of *E. coli* genes use the rho (ρ)-dependent mechanism, the other 50% of genes use the rho (ρ)-independent mechanism. An individual gene does not use both termination mechanisms.

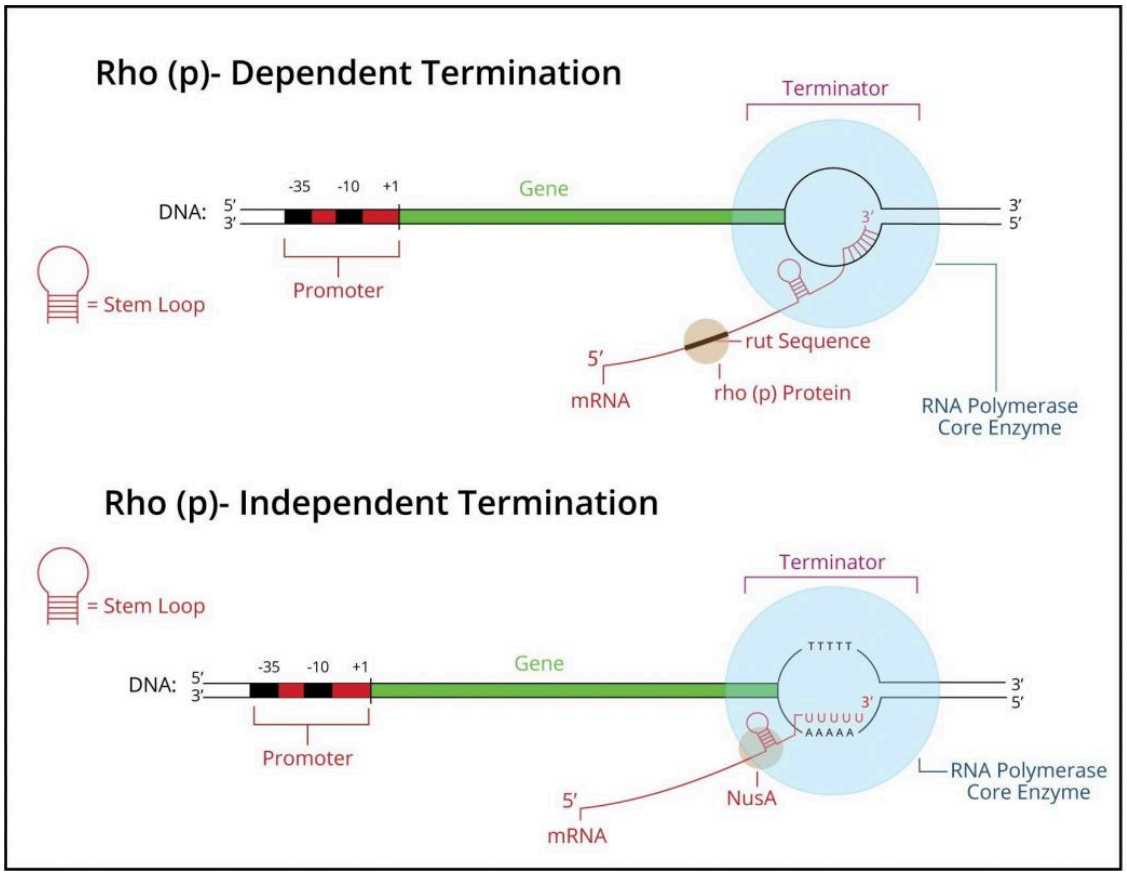


Figure 9.7 ***Transcription Termination in Bacteria*** *--- Image created by SL*

Key Questions

- What three components are involved in rho (ρ)-independent termination?
- What are the functions of each of these components in rho (ρ)-independent termination?

B. Transcription in Eukaryotes

Transcription is important to a eukaryotic cell, as the activation of a structural gene allows eukaryotic cells to adapt to environmental changes (e.g., the presence of a hormone in the blood can activate transcription; see Part 14). Moreover, many eukaryotic organisms are multicellular, so genes need to be transcribed at the right time during development and in the correct cell type. For example, genes involved in building the central nervous system should be transcribed during embryonic development. Genes that encode proteins involved in muscle contraction should be transcribed in muscle cells and not transcribed in other cell types, such as white blood cells. These phenotypic differences are due to transcription, as all cell types (neurons, muscle cells, white blood cells) in the body contain an identical collection of genes.

DNA Sequences Control Eukaryotic Transcription

The transcription of eukaryotic genes is controlled by several types of DNA sequence elements, including the following (see **figure 9.8**):

- **Core promoter.** The core promoter determines where the RNA polymerase will bind to the DNA and begin transcription**.** The core promoter contains two important DNA sequence elements:
  - **TATA box (-25 sequence).** The TATA box (5'-TATAAAA-3' sequence in the coding DNA strand) is located approximately 25 base pairs upstream of the transcriptional start site. The TATA box serves as the binding site for the general transcription factor protein TFIID (see below). The TATA box is also rich in AT base pairs, promoting DNA strand separation.
  - **Transcription start site (+1 site).** The +1 site is the first nitrogenous base in the template DNA strand that is transcribed into an RNA nucleotide.

For a eukaryotic gene to be transcribed, the TATA box and the +1 site must be present. However, if these two sequences are the only DNA sequences present upstream of a gene, the gene is transcribed at a low, yet constant rate, the so-called **basal** level of transcription.

- **Regulatory DNA sequences.** Regulatory DNA sequences function to either transcribe the gene above the basal level or transcribe a gene below the basal level. Regulatory DNA sequences serve as the binding sites for **regulatory transcription factor** proteins that influence the ability of RNA polymerase to recognize the core promoter efficiently. Regulatory DNA sequences include:
  - **Enhancers.** Enhancer DNA sequences stimulate the transcription of the controlled gene above the basal level. Enhancer DNA sequences are the binding sites for **activator** proteins.
  - **Silencers.** Silencer DNA sequences down-regulate transcription of the controlled gene below the basal level. Silencer DNA sequences are the binding sites for **repressor** proteins.

The DNA sequences that influence transcription of an adjacent gene are called ***cis*-acting DNA elements.** *Cis*-acting DNA elements include the core promoter, enhancer, and silencer sequences. The transcription factor proteins that bind to these *cis*-acting DNA elements are called ***trans*-acting factor proteins**. *Trans*-acting factors proteins, also called **transcription factor proteins**, include activator proteins, repressor proteins, and the general transcription factor (GTF) proteins (see below).

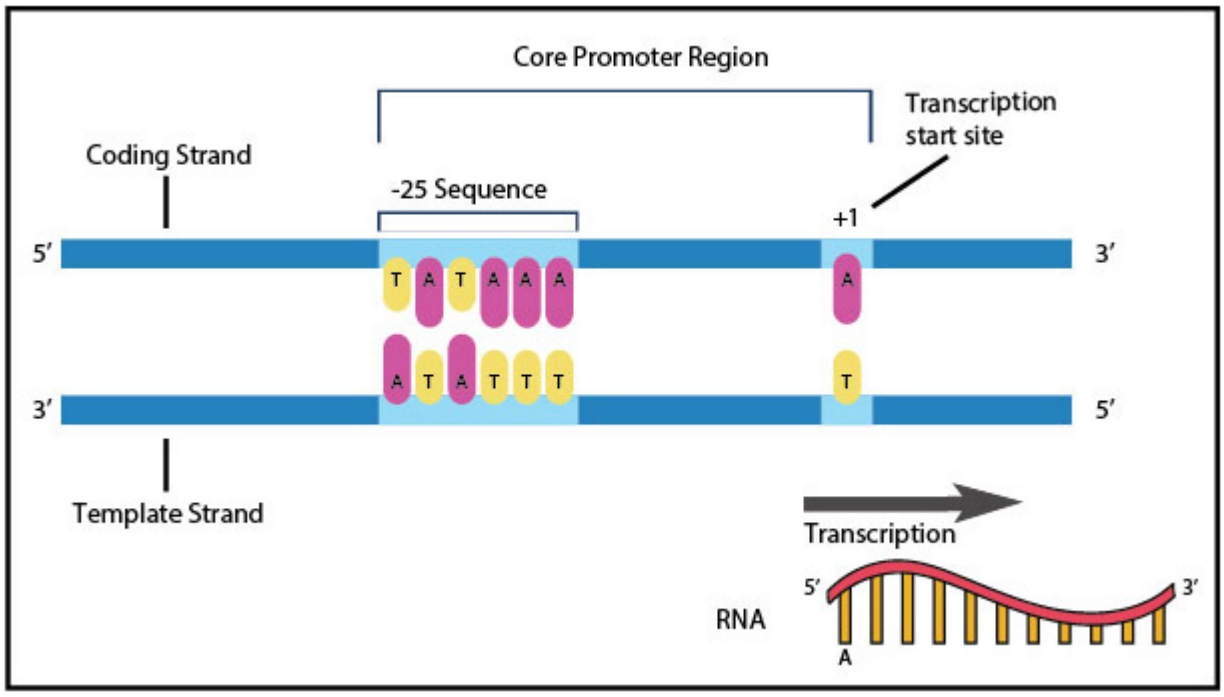


Figure 9.8 ***Eukaryotic Core Promoter*** *--- Image created by SL*

Key Questions

- What are the names of the two sequence features within the core promoter?
- What are the two functions of the TATA box?
- What are names and functions of the two regulatory DNA sequences that influence the transcription of eukaryotic genes?
- What are the names of the proteins that bind to these two regulatory DNA sequences?

RNA Polymerases in Eukaryotes

In eukaryotes, there are three types of RNA polymerases that handle transcription:

- **RNA polymerase I.** RNA polymerase I transcribes most of the eukaryotic ribosomal RNA (rRNA) genes to make rRNA molecules. We will learn in Part 11 that rRNA molecules are noncoding RNA molecules that play a critical role in the translation process.
- **RNA polymerase II.** RNA polymerase II transcribes eukaryotic structural genes. Recall that structural genes produce mRNA molecules upon transcription. In this section, we will focus our attention on RNA polymerase II.
- **RNA polymerase III.** RNA polymerase III transcribes all eukaryotic transfer RNA (tRNA) genes. We will learn in Part 11 that tRNA molecules are noncoding RNA molecules that play a critical role in translation, functioning to deliver amino acids to the ribosome.

Key Questions

- What types of genes do the three eukaryotic RNA polymerases transcribe?

Initiation in Eukaryotes

Both **basal** (constant, low level) transcription and **regulated** (above or below the basal level) transcription of structural genes in eukaryotes require the following proteins (see **figure 9.9**):

- **RNA polymerase II**.
- **General transcription factor (GTF) proteins.** The GTF proteins function like the bacterial sigma (σ) factor protein; the GTFs deliver RNA polymerase II to the core promoter and regulate RNA polymerase II function. There are six major GTF proteins in eukaryotes:
  - **TFIID.** The TFIID protein binds to the core promoter by recognizing the TATA box (-25 sequence). TFIID is actually a multi-subunit protein "machine" composed of multiple protein subunits. One of these protein subunits is the **TATA-binding protein (TBP)** that binds directly to the TATA box (-25 sequence).
  - **TFIIA.** The TFIIA protein helps TFIID bind to the TATA box (-25 DNA sequence).
  - **TFIIB.** The TFIIB protein binds to TFIID and recruits the RNA polymerase II/TFIIF protein complex to the core promoter.
  - **TFIIF.** The TFIIF protein is associated with RNA polymerase II. When the TFIIF protein binds to TFIIB, RNA polymerase II is located at the +1 sequence.
  - **TFIIH.** TFIIH is another multi-subunit protein complex. One protein subunit within the TFIIH complex is a DNA helicase that breaks the hydrogen bonds at the TATA box (-25 sequence). Another protein subunit within the TFIIH complex is a kinase, phosphorylating RNA polymerase II to activate transcription. TFIIH uses the chemical energy within ATP to activate RNA polymerase II.
  - **TFIIE.** TFIIE assists TFIIH to separate the two DNA strands, activating transcription.

The association of RNA polymerase II with the six GTF proteins listed above forms a **preinitiation complex.** The preinitiation complex is also called the **basal transcription apparatus.**

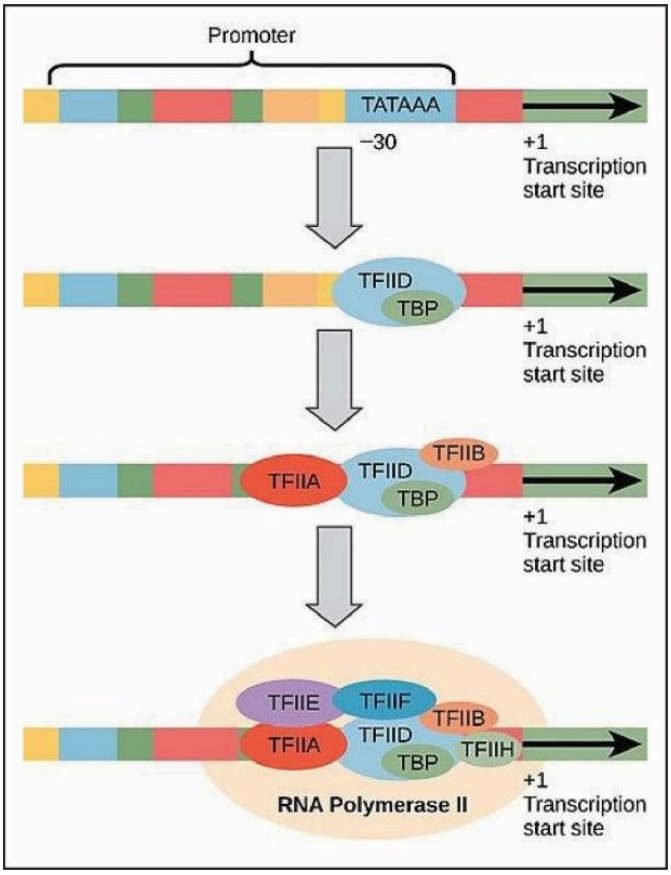


Figure 9.9 ***Transcription Initiation in Eukaryotes*** *--- Image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction)*

Key Questions

- Which GTF binds to the core promoter?
- Which GTF acts as a bridge to connect the GTF bound to the core promoter to the GTF bound to RNA polymerase II?
- Which GTF is the DNA helicase that separates the two DNA strands?
- Which GTF activates RNA polymerase II?

General and Regulatory Transcription Factors

Transcription factor proteins influence the ability of RNA polymerase II to bind to a eukaryotic core promoter. A huge number of eukaryotic genes encode transcription factor proteins; it is estimated that as many as 1000 human genes encode proteins that regulate transcription! There are two categories of transcription factor proteins:

- **General transcription factor (GTFs) proteins.** The GTFs include the **TFIID, TFIIA, TFIIB, TFIIF, TFIIE**, and **TFIIH** proteins described above. The GTFs function to recruit RNA polymerase II to the core promoter and activate RNA polymerase II to begin transcription. The GTFs are required for all transcription events. If these transcription factors are the only ones involved, the gene is transcribed at a low, yet constant level, the so-called **basal level**. GTFs are also required for transcription rates above and below the basal level.
- **Regulatory transcription factor proteins**. Regulatory transcription factor proteins function to either increase the rate of transcription above the basal level or decrease the rate of transcription below the basal level (see **figure 9.10**). **Activator** proteins are regulatory transcription factor proteins that bind to enhancer DNA sequences and increase the level of transcription above the basal level. Conversely, **repressor** proteins bind to silencer DNA sequences and decrease transcription below the basal level. Many regulatory transcription factors are only expressed in certain cell types or at certain times during embryonic development, thus playing a critical role in cell-specific or time-specific transcription.

The DNA binding sites (core promoter, enhancer, and silencer sequences) for these transcription factor proteins tend to be near the genes they control. As a result, the DNA sequences are called ***cis*-acting DNA elements**. However, these *cis*-acting DNA elements do not need to be immediately adjacent to the core promoter. Some enhancers and silencers can be within the gene they control or can be thousands of base pairs away. The transcription factor proteins (GTFs, activators, and repressors) that bind to the *cis*-acting DNA elements are ***trans*-acting factor proteins**.

Since transcriptional control requires both input from a myriad of DNA sequences and proteins, some component in the cell needs to interpret the various activation and repression signals to provide an overall signal to RNA polymerase II. A large multi-subunit **mediator** protein complex regulates the interaction between RNA polymerase II and the activator and repressor proteins. Mediator thus serves as a link between transcription factors that bind to enhancer and silencer DNA sequences and RNA polymerase II, thereby determining the overall rate of transcription.

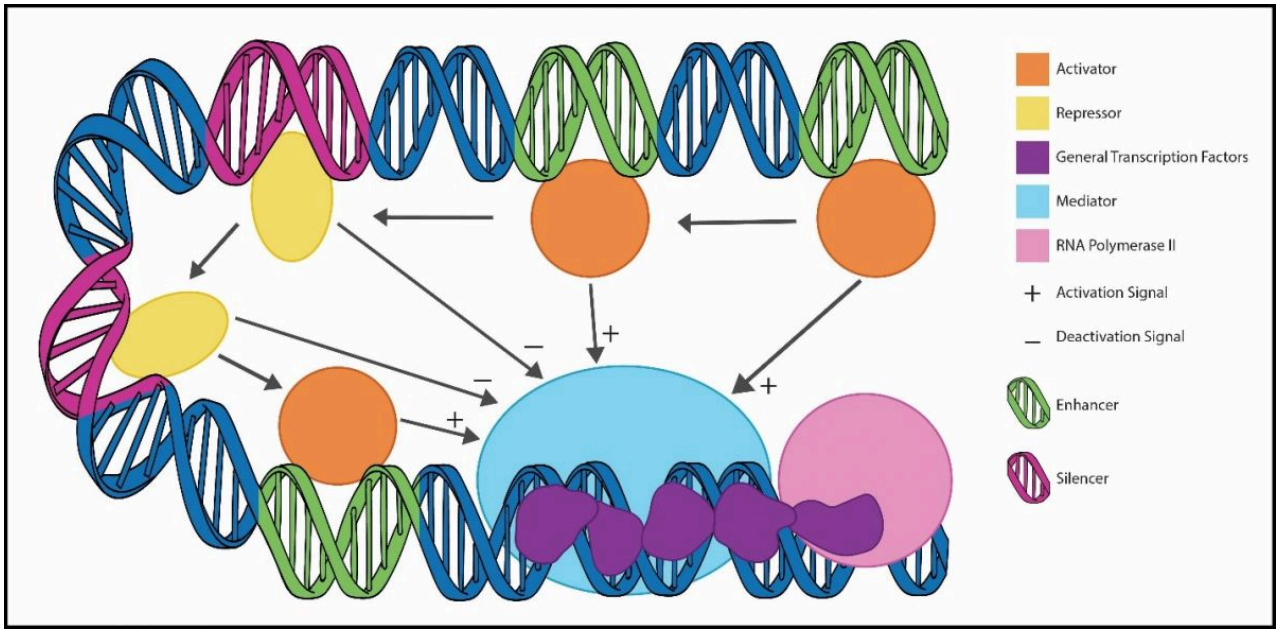


Figure 9.10 ***Regulatory Transcription Factors and Mediator.*** *Mediator (light blue) interprets the activation signals from activator proteins (orange) bound to enhancer DNA sequences (green) and the repression signals from repressor proteins (yellow) bound to silencer DNA sequences (magenta). Mediator then communicates an overall transcription signal (an activation signal in this case) to the general transcription factor proteins (purple) and RNA polymerase II (pink). RNA polymerase II is positioned on the +1 site (not shown) and transcribes the gene towards the right. --- Image created by SL*

Key Questions

- What are three examples of *cis*-acting DNA elements?
- What are three examples of *trans*-acting factor proteins?
- What is the function of the mediator protein complex?

Transcription Elongation in Eukaryotes

The elongation step in eukaryotic transcription is virtually identical to the transcription elongation step in prokaryotes. RNA polymerase II in eukaryotes has the same functional capabilities as the RNA polymerase core enzyme from *E. coli*.

Key Questions

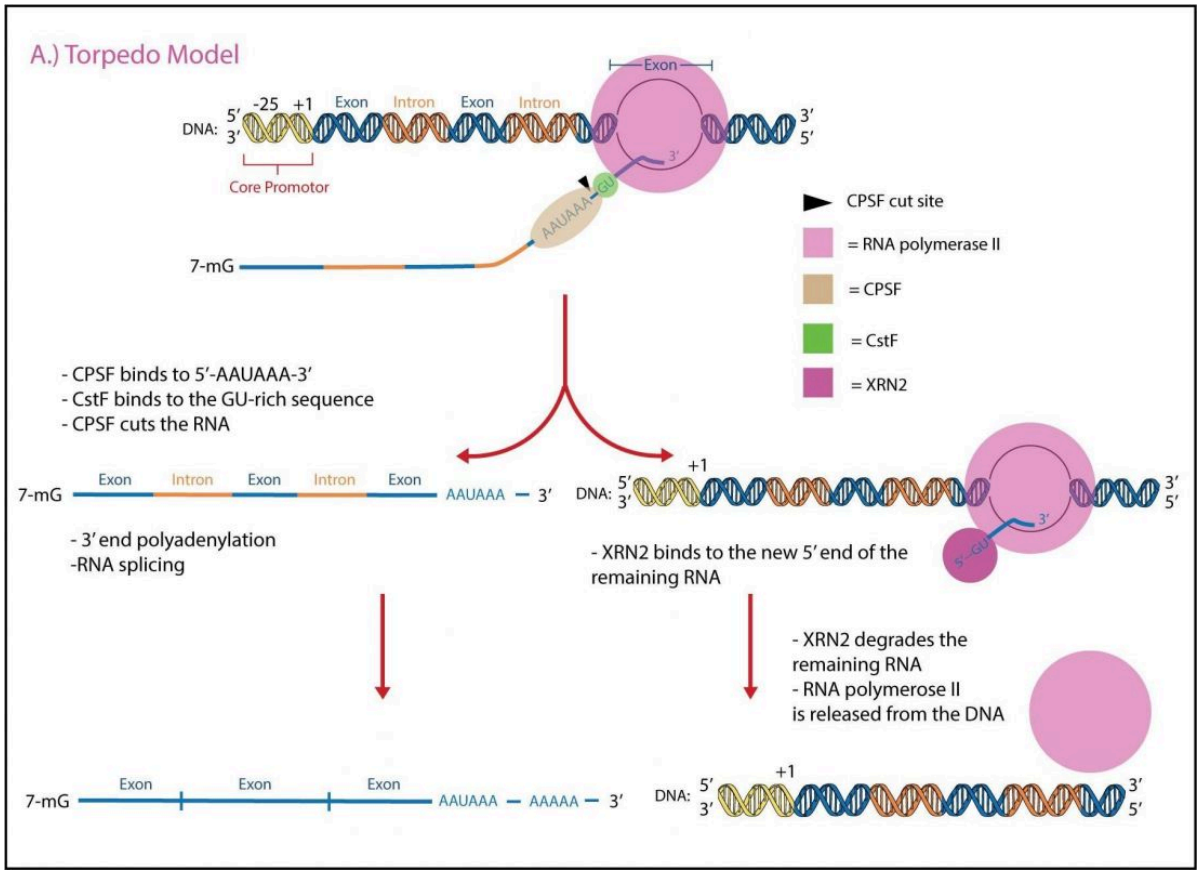
- What are the names of the two proteins that act as DNA helicases in eukaryotic transcription?

Transcription Termination in Eukaryotes

Transcriptional termination in eukaryotes occurs during the process of **3' end polyadenylation**, a modification to the 3' ends of eukaryotic mRNAs. We will cover 3' end polyadenylation in more detail in Part 10. In short, an endonuclease called **cleavage and polyadenylation specificity factor (CPSF)** binds to a **polyadenylation signal sequence** (5'-AAUAAA-3') near the 3' end of the mRNA. CPSF then cuts the mRNA approximately 20 nucleotides downstream (towards the 3' end of the mRNA) from the polyadenylation signal sequence. Cleavage of the mRNA by CPSF releases the mRNA from RNA polymerase II.

After CPSF releases the mRNA from RNA polymerase II, there are two potential ways that RNA polymerase II can be released from the DNA, thereby terminating transcription:

- **Torpedo model**. The torpedo model involves a 5' to 3' exonuclease called **XRN2** degrading the remaining RNA linked to RNA polymerase II and dissociating RNA polymerase II from the DNA (see **figure 9.11a**). Note that the torpedo model shares some similarities to the rho (ρ)-dependent termination mechanism in *E. coli*.
- **Allosteric model**. When RNA polymerase II transcribes the portion of the gene that produces the polyadenylation signal sequence, the RNA polymerase is destabilized and is released from the DNA (see **figure 9.11b**). Note that the allosteric model shares some similarities to the rho (ρ)-independent termination mechanism in *E. coli*.


A.) Torpedo Model
DNA:
-25
+1
Exon
Intron
Exon
Intron
Exon
5′
3′
3′
5′
Core Promotor
AAUAAA
GU
3′
7-mG
CPSF cut site
= RNA polymerase II
= CPSF
= CstF
= XRN2
- CPSF binds to 5′-AAUAAA-3′
- CstF binds to the GU-rich sequence
- CPSF cuts the RNA
7-mG
Exon
Intron
Exon
Intron
Exon
AAUAAA
3′
- 3′ end polyadenylation
-RNA splicing
DNA:
+1
5′
3′
3′
5′
3′
5′-GU
- XRN2 binds to the new 5′ end of the remaining RNA
- XRN2 degrades the remaining RNA
- RNA polymerose II is released from the DNA
7-mG
Exon
Exon
Exon
AAUAAA
AAAAA
3′
DNA:
+1
5′
3′
3′
5′

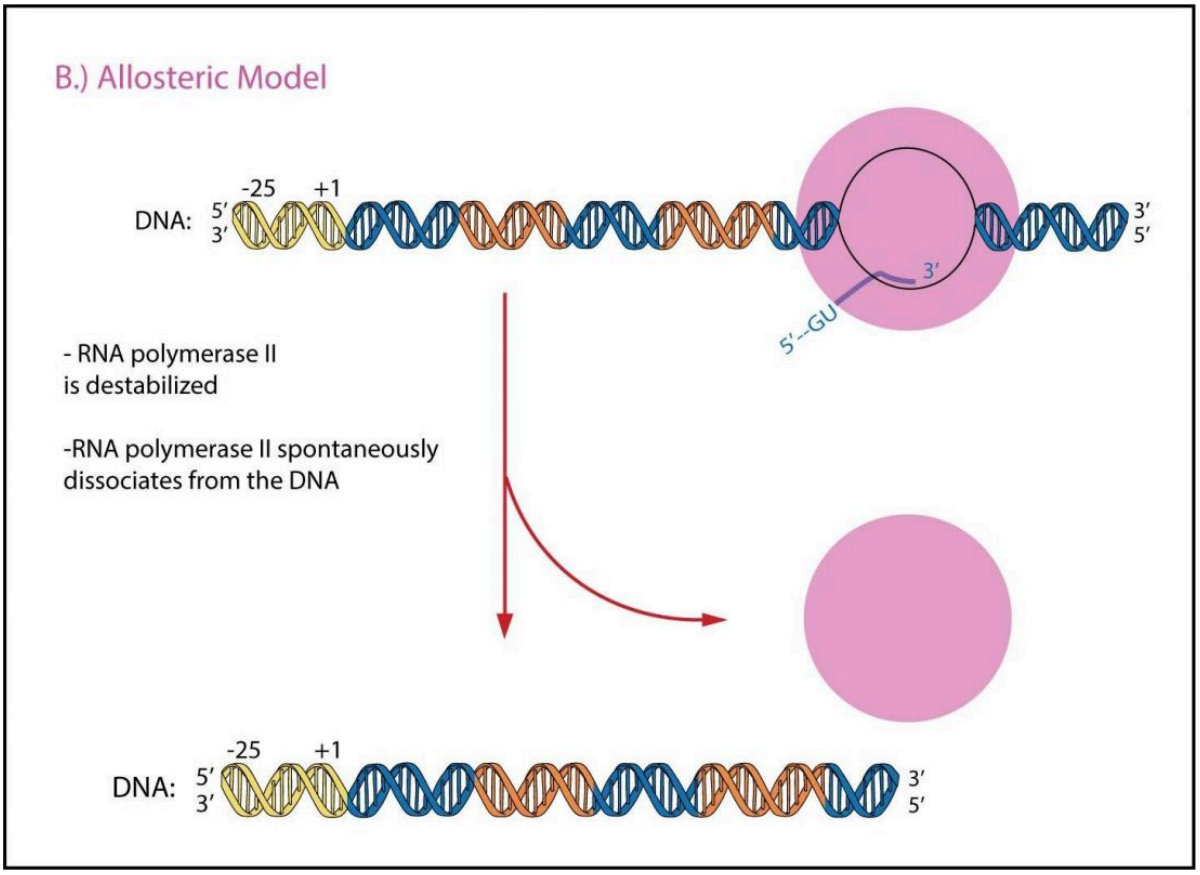


Figure 8.11 ***Transcription Termination in Eukaryotes*** *A) Torpedo Model B) Allosteric Model --- Images created by SL*

Key Questions

- What is the difference between the torpedo and the allosteric models of transcription termination?

Review Questions

Fill in the blank:

1. When structural genes are expressed, they produce _____RNA molecules; when nonstructural genes are expressed, they produce _____RNA molecules.
2. _____ is a GTF protein that has both DNA helicase and kinase activity.
3. The _____ protein binds to the -10 and -35 sequences.
4. The RNA polymerase holoenzyme consists of the _____ protein subunits and the _____ factor protein.
5. The TATA box (-25 sequence) is the binding site for the _____ protein.
6. The _____ protein binds to the *rut* sequence found in 50% of bacterial mRNA molecules.
7. RNA polymerase _____ is responsible for transcribing eukaryotic structural genes.
8. Phosphorylation of _____ helps to activate transcription in eukaryotes.
9. A(n) _____ protein binds to an enhancer sequence in the DNA to activate transcription above the basal level, while a(n) _____ protein binds to a silencer sequence in the DNA to decrease transcription below the basal level.
10. The _____ protein causes the RNA polymerase core enzyme to pause at the stem loop in the rho (ρ)-independent mechanism.
11. DNA replication requires the use of DNA helicase to unwind double-stranded DNA, while transcription in bacteria uses the _____ to unwind double-stranded DNA.

10 - RNA Modifications

After the RNA molecule is produced by transcription (Part 9), the structure of the RNA is often modified prior to being translated into a protein product. These **RNA modifications** apply mainly to eukaryotic RNA transcripts.

Key Questions

- Which group of organisms modify their RNA transcripts?

Overview of RNA Modifications

The modifications to eukaryotic RNA transcripts include the following:

- **5’ end capping.** 5’ end capping involves the attachment of a modified nucleotide called **7-methylguanosine (7-mG)** to the 5’ end of RNA molecules. The added 7-mG is sometimes called the **5’ cap**.
- **3’ end polyadenylation.** 3’ end polyadenylation involves the addition of a string of adenine (A) nucleotides to the 3’ end of the RNA molecule. The added sequence of A nucleotides is called the **polyA tail**.
- **RNA splicing.** Most eukaryotic genes are **split genes,** being composed of both intron DNA sequences and exon sequences. For split genes, initial transcription in the nucleus produces a **precursor mRNA (pre-mRNA)** molecule. This pre-mRNA then goes through 5' end capping, 3' end polyadenylation, and finally RNA splicing. During RNA splicing, the **intron** sequences are removed from the pre-mRNA and discarded (see **Figure 10.1**). The remaining exon RNA segments are spliced together to produce a **mature mRNA** molecule that is transported to the cytoplasm of the cell for translation.
- **RNA processing.** RNA processing involves cutting larger RNA transcripts into smaller ones. RNA processing involves both **exonucleases** (removing nucleotides from the ends of the RNA transcript) or **endonucleases** (cleaving the RNA transcript at an internal site). The ribosomal RNA (rRNA) molecules that are essential components within ribosomes (see Part 11) experience RNA processing after transcription.
- **RNA editing.** RNA editing involves changing the nucleotide sequence of the mRNA molecule prior to translation.
- **Base modification.** During base modification, nitrogenous bases within the RNA transcript are covalently modified by the addition of chemical groups, such as methyl groups.

The remainder of Part 10 will focus on the first three RNA modifications: 5' end capping, 3' end polyadenylation, and RNA splicing.

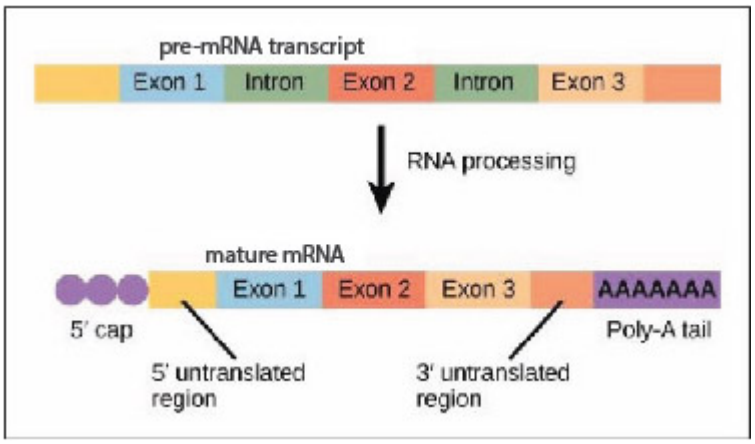


Figure 10.1 ***RNA Modifications Overview*** *--- Image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction*

Key Questions

- What is the difference between a pre-mRNA and a mature mRNA molecule?
- What is the difference between an intron and an exon DNA sequence?
- What is meant by 5′ end capping?
- What is meant by 3′ end polyadenylation?

5′ End Capping

The 5′ end of the pre-mRNA molecule is modified by the addition of a **7-methylguanosine (7-mG)** nucleotide. The process of adding the 7-mG to the pre-mRNA is **5′ end capping**. 5′ end capping is the first RNA modification, occurring as soon as the 5' end of the pre-mRNA emerges from RNA polymerase II during transcription. 5′ end capping (see **Figure 10.2**) involves the following enzymes:

1. **RNA 5′-triphosphatase.** Recall that RNA polymerases do not require a primer to initiate transcription (see Part 9). As a result, the first nucleotide incorporated into the RNA has three phosphate groups attached to the 5' carbon. RNA 5'-triphosphatase removes one of the three phosphates from this nucleotide.
2. **Guanylyltransferase.** Guanylyltransferase cleaves GTP to produce GMP and pyrophosphate (PP_i). Guanylyltransferase then attaches the phosphate group of the GMP molecule to the two phosphate groups on the nucleotide at the 5′ end of the pre-mRNA transcript. It is important to note that an unusual 5′ to 5′ linkage is formed, placing three phosphate groups between the 5' carbons on two adjacent nucleotides.
3. **Methyltransferase.** Methyltransferase attaches a methyl group to the added guanine nitrogenous base, resulting in the 7-mG cap.

The 7-mG cap on eukaryotic mRNAs has at least three functions. The 7-mG cap:

- Serves as a binding site for proteins that transport the mRNA from the nucleus to the cytoplasm of the cell.
- Serves as a recognition site for translation factor proteins that help the ribosome bind to the mRNA. Once the ribosome binds to the mRNA, translation begins (see Part 11).
- Protects the 5′ end of the mRNA transcript from exonuclease digestion.

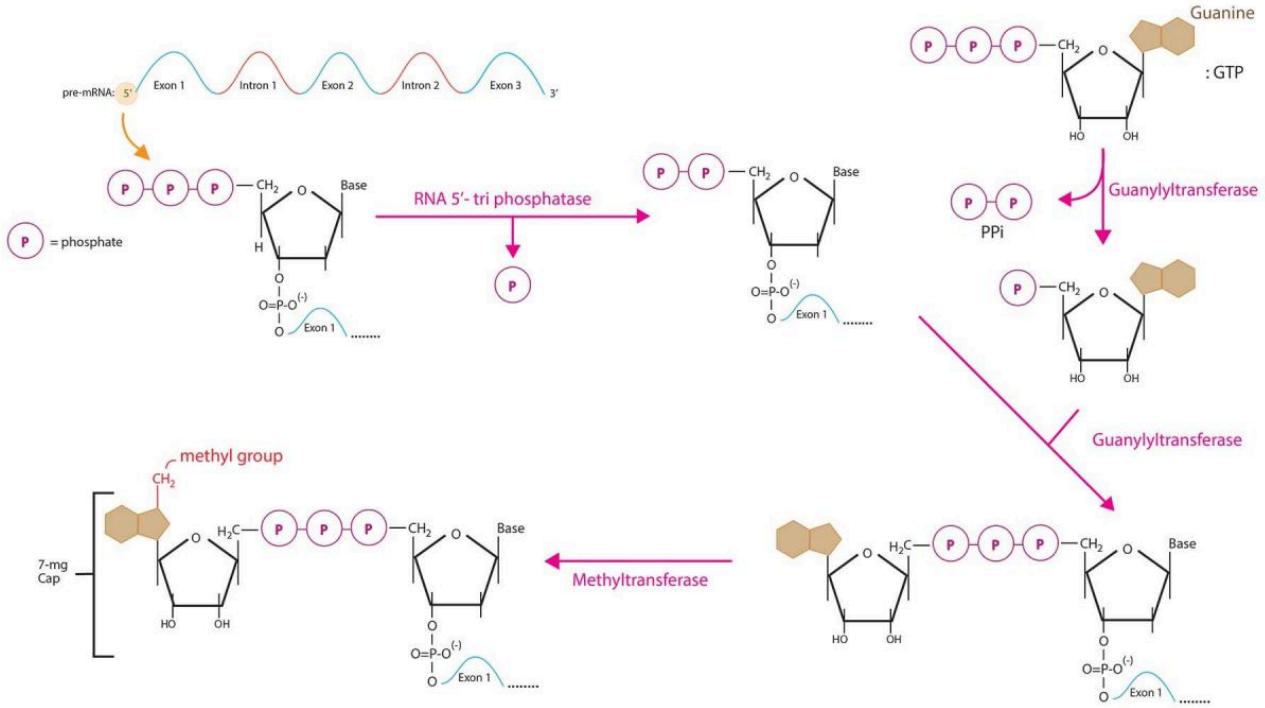


Figure 10.2 ***5' end capping mechanism*** *--- Image created by JET*

Key Questions

- How does the 7-mG structure contribute to translation?
- Which nucleotide triphosphate provides the energy for 5' end capping?
- What is unusual about the covalent bonds between 7-mG and the rest of the pre-mRNA molecule?

3' End Polyadenylation

The 3' end of the pre-mRNA is modified by the addition of a **polyA tail**, a string of approximately 250 adenine (A) nucleotides. The process of adding a polyA tail to the mRNA transcript (see **Figure 10.3**), called **3' end polyadenylation**, involves:

1. The detection of two recognition sequences (**polyadenylation signal sequences**) near the 3' end of the pre-mRNA molecule. The first polyadenylation signal sequence, 5'-AAUAAA-3', is recognized by the endonuclease **cleavage and polyadenylation specificity factor (CPSF)** protein. The second polyadenylation signal sequence, enriched in guanine and uracil bases, is called the GU-rich sequence. This GU-rich sequence is the binding site for the **cleavage stimulatory factor (CstF)** protein. When CstF and CPSF bind to their respective polyadenylation signal sequences, CstF activates CPSF.
2. The CPSF protein cleaves the pre-mRNA between the two polyadenylation signal sequences. When CPSF cleaes the pre-mRNA molecule, the pre-mRNA is released from RNA polymerase II. The new 3' end of the pre-mRNA is then available for the addition of a polyA tail.
3. **Poly(A)-polymerase (PAP)** attaches 250 adenine nucleotides to the newly generated 3' end of the pre-mRNA transcript. PAP is an unusual RNA polymerase that does not require a template and only forms phosphodiester bonds between adenine nucleotides.

The polyA tail on the mRNA has at least three functions. The polyA tail functions to:

- Protect the 3’ end of the pre-mRNA transcript from exonuclease degradation.
- Promote the transport of the mRNA from the nucleus to the cytoplasm of the cell.
- Help the ribosome bind to the mRNA to initiate translation.

The 3’ end polyadenylation process occurs after 5’ end capping, but prior to RNA splicing. In fact, 3’ end polyadenylation assists in terminating transcription in eukaryotes by the torpedo model (see Part 9).

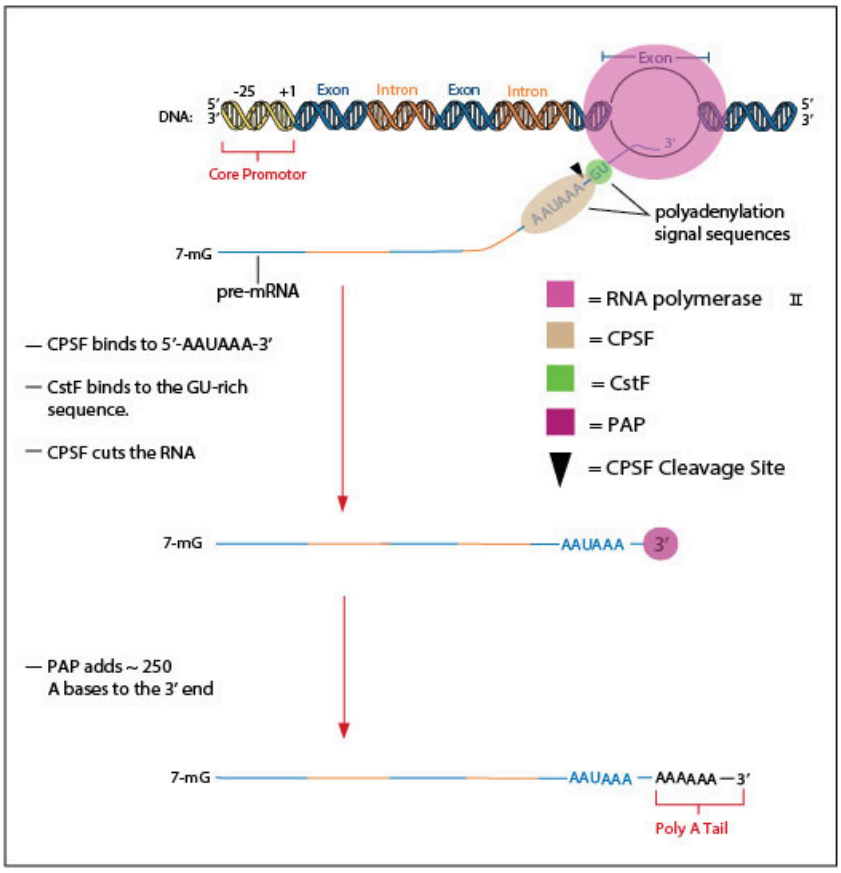


Figure 10.3 ***3’ end polyadenylation mechanism*** *--- Image created by SL*

Key Questions

- How does 3’ end polyadenylation contribute to transcription termination in eukaryotes?
- Describe the functions of CPSF, CstF, and Poly(A)-polymerase during 3' end polyadenylation.

Splicing of Group I and Group II Introns

There are three general mechanisms used by eukaryotes to remove introns from RNA molecules. The **group I** and **group II** mechanisms are limited to certain types of eukaryotes or certain organelles within a eukaryotic cell. For example, the group I mechanism removes the introns found in ribosomal RNA (rRNA) molecules in certain protozoa. The group II mechanism removes the introns found in the mRNA and transfer RNA (tRNA) molecules produced by mitochondrial and chloroplast genes. The **spliceosome mechanism** is the major mechanism that is used to remove introns from pre-mRNA transcripts in the nucleus of eukaryotic cells.

- **Removing group I introns.** RNA splicing of group I introns occurs by **self-splicing**, meaning that the precursor RNA molecule catalyzes the removal of its own intron (see **Figure 10.4**). These catalytic precursor RNA molecules are examples of a unique group of molecules classified as RNA enzymes **(ribozymes)**. In fact, the discovery of the Group I ribozyme was the first demonstration that a molecule other than a protein could serve as an enzyme. The ribozyme mechanism to remove group I introns occurs as follows:

    1. A free **guanosine nucleoside** (guanine nitrogenous base covalently linked to a ribose sugar; no phosphate groups) binds to a pocket within the intron. The guanosine nucleoside bound to the intron serves as an enzyme cofactor (i.e., assists the ribozyme in catalysis) for the remaining steps in the reaction.
    2. A break forms at the junction between the 3' end of the first exon and the 5' end of the intron.
    3. The released 3' end of the first exon then cleaves the phosphodiester bond between the 3' end of the intron and the 5' end of the second exon.
    4. A phosphodiester bond is formed that links the first and second exons together, generating a mature RNA molecule. The intron is released and degraded.

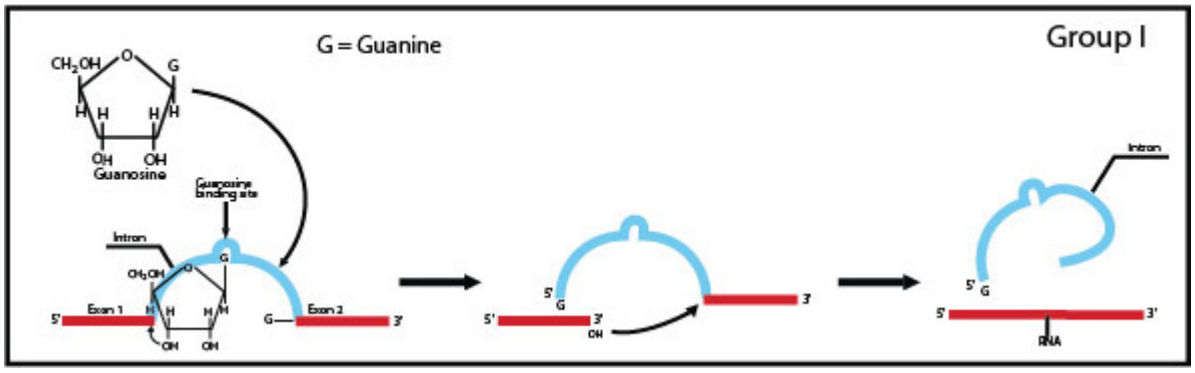


Figure 10.4 ***Removing group I introns*** *--- Image created by SL*

- **Removing group II introns.** RNA splicing of group II introns also occurs by self-splicing, meaning that the precursor RNA is an RNA enzyme (ribozyme) that removes its own intron (see **Figure 10.5**). The self-splicing of group II introns involves:

    1. The 2'-OH group of an **adenine nucleotide** within the intron cleaves the phosphodiester bond between the 3' end of the first exon and the 5' end of the intron. In this reaction, the adenine nucleotide serves as an enzyme cofactor for the reaction.
    2. The released 3' end of the first exon then cleaves the phosphodiester bond between the 3' end of the intron and the 5' end of the second exon.
    3. A phosphodiester bond is formed that links the first and second exons, generating a mature RNA transcript. The intron is released and degraded.

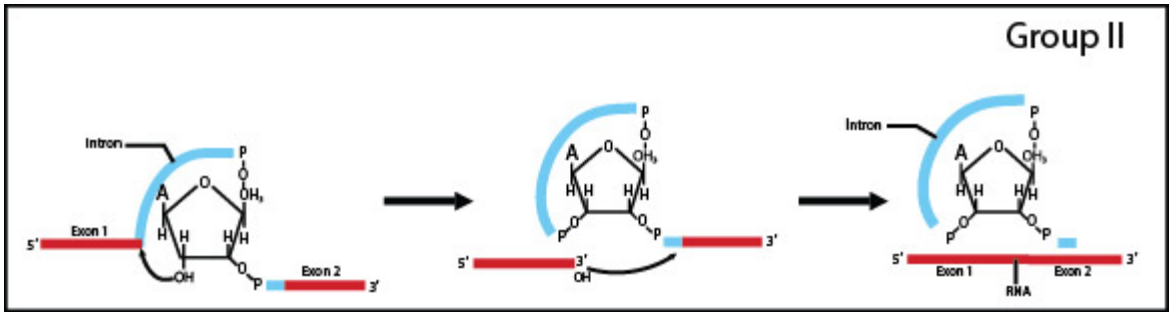


Figure 10.5 ***Removing group II introns*** *--- Image created by SL*

Key Questions

- What is a ribozyme?
- Describe the major events that occur in the Group I and Group II splicing mechanisms.
- What molecules serve as enzyme cofactors in the Group I and Group II splicing mechanisms?

Removal of Introns by Spliceosomes

Transcription of most structural genes in the nucleus of eukaryotic cells produces pre-mRNA molecules; the removal of the introns within these pre-mRNA molecules involves a large multi-subunit **spliceosome** complex. To remove introns from the pre-mRNA, the spliceosome binds to recognition sequences within the intron (see **Figure 10.6**). These **intron recognition sequences** include:

- **The 5′ splice site.** The 5′ splice site is a 5′-GU-3′ RNA sequence at the 5′ end of the intron.
- **The branch site.** The branch site is an adenine nucleotide (A) near the middle of the intron RNA sequence.
- **The 3′ splice site.** The 3′ splice site is an 5′-AG-3′ RNA sequence at the 3′ end of the intron.

The spliceosome complex contains multiple subunits; these subunits are called **small nuclear ribonucleoproteins** or **snRNPs** (“snurps”). Each snRNP within the spliceosome complex is composed of a **small nuclear RNA (snRNA)** molecule that acts as an RNA enzyme (ribozyme) to remove the introns from the pre-mRNA molecule. snRNPs are also composed of proteins that function to stabilize snRNP structure.

The spliceosome splicing mechanism occurs as follows:

1. The **U1** snRNP binds to the 5′ splice site within the intron RNA sequence, while the **U2** snRNP binds to the branch site adenine within the intron.
2. Additional snRNPs called **U4, U5**, and **U6** bind to the intron. These five snRNPs (U1, U2, U4, U5, and U6) form the spliceosome complex.
3. The intron forms a loop bringing the two exon sequences to be linked close together.
4. The 5′ splice site within the intron is cut by U1, and the 5′ end of the intron is covalently linked to the 2′-OH group of the branch site adenine, forming an RNA loop structure called a **lariat**.
5. The U1 and U4 snRNPs are released.
6. The 3′ splice site within the intron is cut by the U5 snRNP.
7. A phosphodiester bond is formed that links the two exons together to form the mature mRNA molecule.
8. The intron is released along with the U2, U5, and U6 snRNPs.

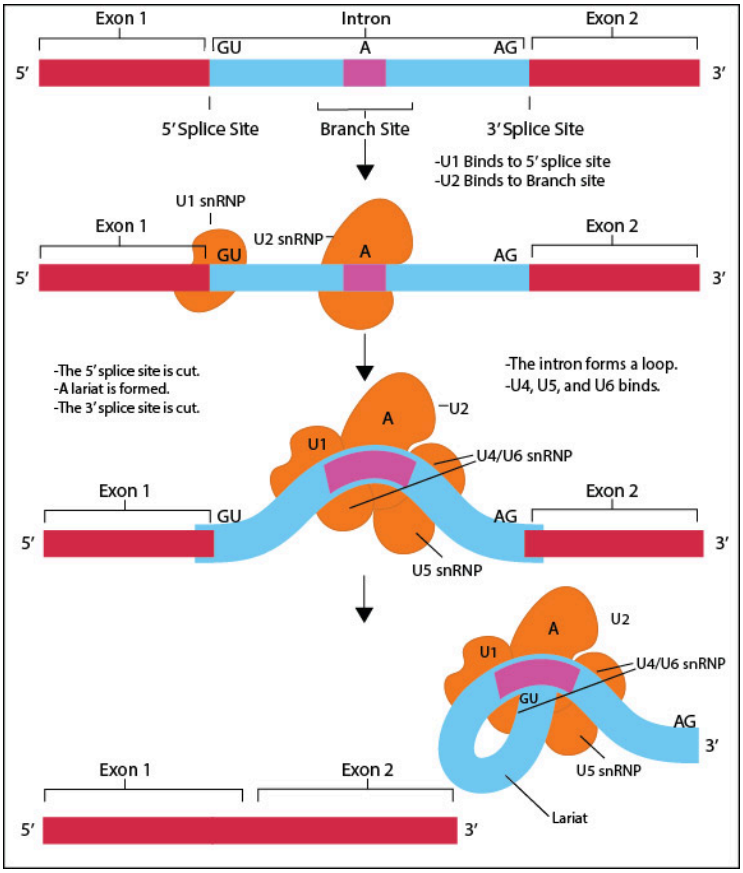


Figure 10.6 ***Spliceosome splicing*** *--- Image created by SL*

Key Questions

- Which two splicing mechanisms are found in human cells?
- What are the names of the three RNA sequences found within introns removed by the spliceosome?
- Which snRNP component is the ribozyme?
- What are the functions of the U1 and U5 snRNPs?

Identifying Introns Using R-Loop Experiments

Introns were initially identified within eukaryotic genes by performing **R-loop (hybridization) experiments**. These R-loop experiments relied on separating the two DNA strands within a gene, allowing a mRNA molecule to form hydrogen bonds (hybridize) with the template DNA strand, and adding the coding strand DNA, which attempts to form hydrogen bonds with the template DNA strand. Finally, the resulting nucleic acid structure was examined in an electron microscope. Below are the results expected from two R-loop experiments, one experiment involving the pre-mRNA (before RNA modifications), the other experiment involving the mature mRNA (after RNA modifications).

- **Gene hybridized to the pre-mRNA.** The pre-mRNA forms hydrogen bonds with the template DNA strand preventing the coding DNA strand from binding. Because the coding DNA strand fails to bind to the template DNA strand, the coding DNA strand extends outward from the RNA-DNA hybrid region. This loop where the coding DNA strand cannot bind to the template DNA strand is called an RNA displacement loop or **R loop** (see **Figure 10.7A**).
- **Gene hybridized to the mature mRNA.** Hybridization between the template DNA strand and the mature mRNA forces the intron DNA sequences in the template DNA strand to loop out, because the mature mRNA lacks intron sequences. Adding the coding DNA strand produces R-loops with intervening regions of double-stranded DNA (i.e., the intron sequences within the template and the coding DNA strands form hydrogen bonds) called **intron loops** (see **Figure 10.7B**).

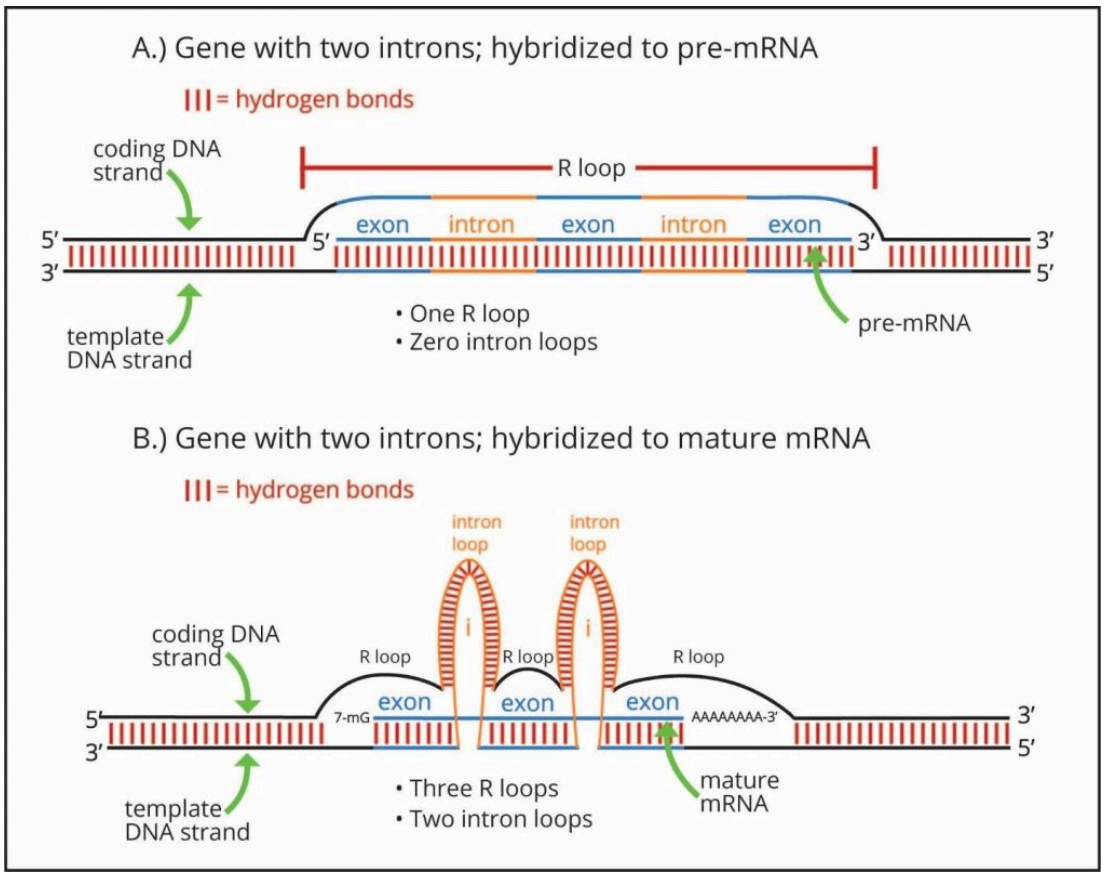


Figure 10.7 ***R-Loop Results*** *--- Image created by SL*

Key Questions

- Suppose a gene contains four introns and is hybridized with its mature mRNA. How many R loops would be observed in the electron microscope at the end of an R-loop experiment? How many intron loops would be observed?

Identifying Introns by Comparing gDNA with cDNA

Introns within genes can also be identified by comparing the length of a **genomic DNA (gDNA)** version of a gene to the **complementary DNA (cDNA)** version of the same gene. gDNA is the version of a gene found in the genome; the gDNA version of a gene contains both introns and exons. cDNA is produced in the laboratory by **reverse transcription** (see Part 8). Reverse transcription converts mature mRNA into a cDNA molecule using the viral enzyme **reverse**

transcriptase. Since the cDNA molecule is produced from the mature mRNA, cDNA molecules contain exons but lack introns. The gDNA version of the gene, which contains introns, will be longer than the cDNA version of the same gene, which lacks introns.

The **polymerase chain reaction (PCR)** technique (see Part 8) can be used to make billions of copies of the gDNA and the cDNA versions of any gene of interest. The gDNA and cDNA PCR products are then separated by size using **agarose gel electrophoresis** (see Part 8). The size difference between the gDNA and the cDNA copy of the gene can be easily observed on an agarose gel (see **Figure 3.2**).

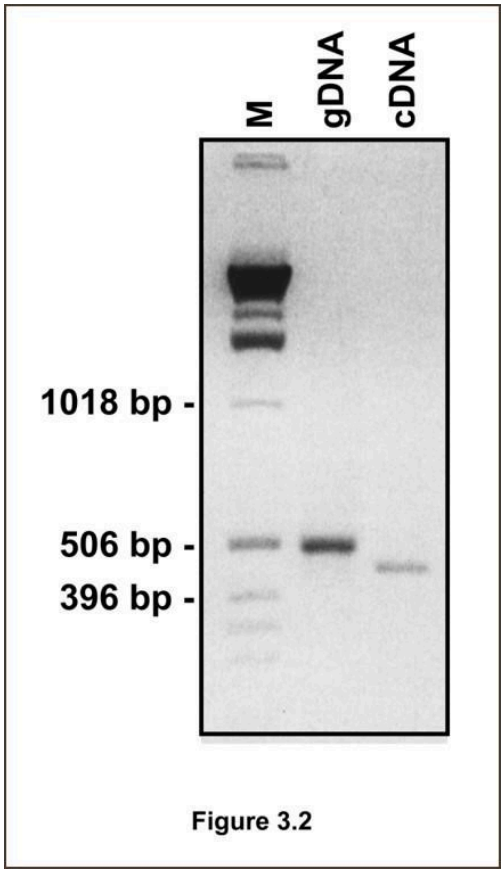


Figure 10.8 ***Comparing gDNA to cDNA to identify* introns.** *The gene in question contains a 75 base pair (bp) intron. --- Image provided by K. Mark DeWall*

Key Questions

- What is the difference between gDNA and cDNA?
- How can comparing gDNA to cDNA on an agarose gel help you indentify an intron?

Alternative Splicing

Alternative splicing involves splicing a single type of pre-mRNA molecule in different ways to produce multiple mature mRNA molecules (see **Figure 10.9**). Each of these mature mRNAs can then produce slightly different proteins upon translation. These distinct, yet related protein **isoforms**, all derived from a single gene, can have specialized functions. Alternative splicing is beneficial in that it allows eukaryotes to carry fewer genes in the genome, permitting a relatively small number of genes the flexibility to encode a vast array of proteins. In humans, it is estimated that 30–60% of the

genes in the genome are alternatively spliced. As a result, the human genome, which contains approximately 23,000 structural genes, can produce at least ten times that number of unique protein products.

One example of alternative splicing involves the human α-tropomyosin gene, a gene involved in muscle contraction. The α-tropomyosin gene contains 14 exons and 13 introns. The α-tropomyosin gene contains two types of exons:

- **Constitutive exons.** Constitutive exons are included in all of the α-tropomyosin mature mRNAs products of alternative splicing. These exons likely encode amino acid sequences that maintain the general three-dimensional structure of the encoded α-tropomyosin protein.
- **Alternative exons.** Alternative exons vary between α-tropomyosin mature mRNAs. In one cell type, one combinations of alternative exons are spliced together with the constitutive exons to make a mature mRNA molecule. In another cell type, a different combination of alternative exons are spliced together with the constitutive exons to make a mature mRNA molecule. The result is two related proteins that have slightly different functions to meet the unique needs of these two different cell types.

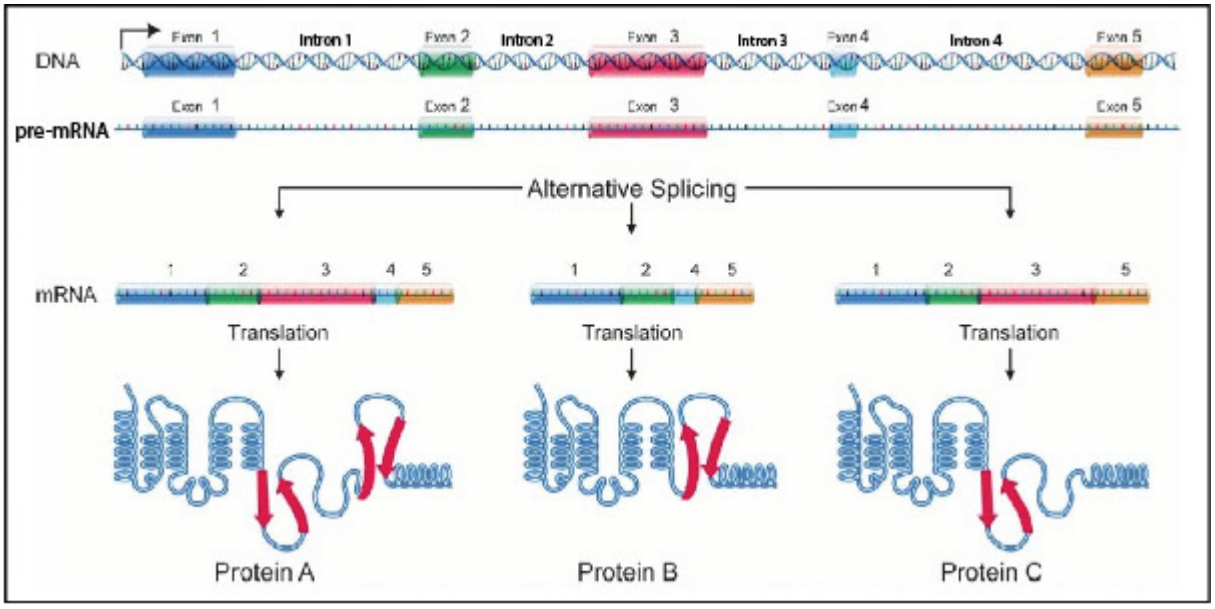


Figure 10.9 ***Alternative splicing allows one gene to produce three proteins.*** *In this example, exons 1, 2, and 5 are constitutive exons, while exons 3 and 4 are alternative exons. --- DNA Alternative Splicing by National Human Genome Research Institue and is used under CC0*

Key Questions

- Why is alternative splicing advantageous?
- What are protein isoforms?
- What is the difference between a constitutive exon and an alternative exon?

Patterns of Alternative Splicing

Alternative splicing is regulated by **splicing factor** proteins. These splicing factor proteins help the spliceosome choose which intron splice sites to cut during RNA splicing. Each cell type has a different collection of splicing factor proteins, allowing different RNA splicing patterns to occur in each cell type.

Here are some common alternative splicing patterns observed in eukaryotic cells:

- **Exon Skipping.** Some splicing factor proteins act as **splice repressors**. Splice repressor proteins prevent the spliceosome from recognizing a particular 3' splice site within an intron (see Figure 10.10). When a splice repressor protein blocks a 3′ splice site within an intron, the 3′ splice site in the next intron is chosen for splicing instead, and the intervening exon is removed from the pre-mRNA molecule (**exon skipping**).
- **Alternative 5' and 3' Splice Sites.** In addition to the 5′ splice site, the branch site, and the 3′ splice sites discussed earlier, there are other pre-mRNA sequences involved in RNA splicing. These additional sequence elements, often located within a nearby exon, can promote the use of a particular 5′ or 3′ splice site. For example, some potential 5′ or 3′ splice sites in the pre-mRNA are poorly recognized by the spliceosome. In certain cell types, the binding of a **splice activator protein** to a **splice enhancer sequence** within a nearby exon promotes the use of these otherwise poorly recognized 5′ or 3′ splice sites (see **Figure 10.10**). When a splice activator protein binds to a splice enhancer sequence, an exon is included in the mature mRNA (i.e., the exon is not skipped).
- **Mutually Exclusive Exons.** In some cases, splicing events are coordinated between different cell types to ensure that unique protein isoforms are produced by each cell type. For example, suppose there are four exons (three introns) in a pre-mRNA molecule. During splicing in one cell type, exon two is consistently retained in the mature mRNA, while exon three is spliced out. In a different cell type, exon two is always spliced out, while exon three is.01233 retained in the mature mRNA. Exons one and four are found in the mature mRNAs in both cell types and are thus constitutive exons.

Scientists are still learning the true complexity of alternative splicing. It appears that alternative splicing patterns are cell-type and developmental stage specific. Moreover, mutations often lead to aberrant splicing patterns. This aberrant splicing produces abnormal protein isoforms and in some cases, disease phenotypes.

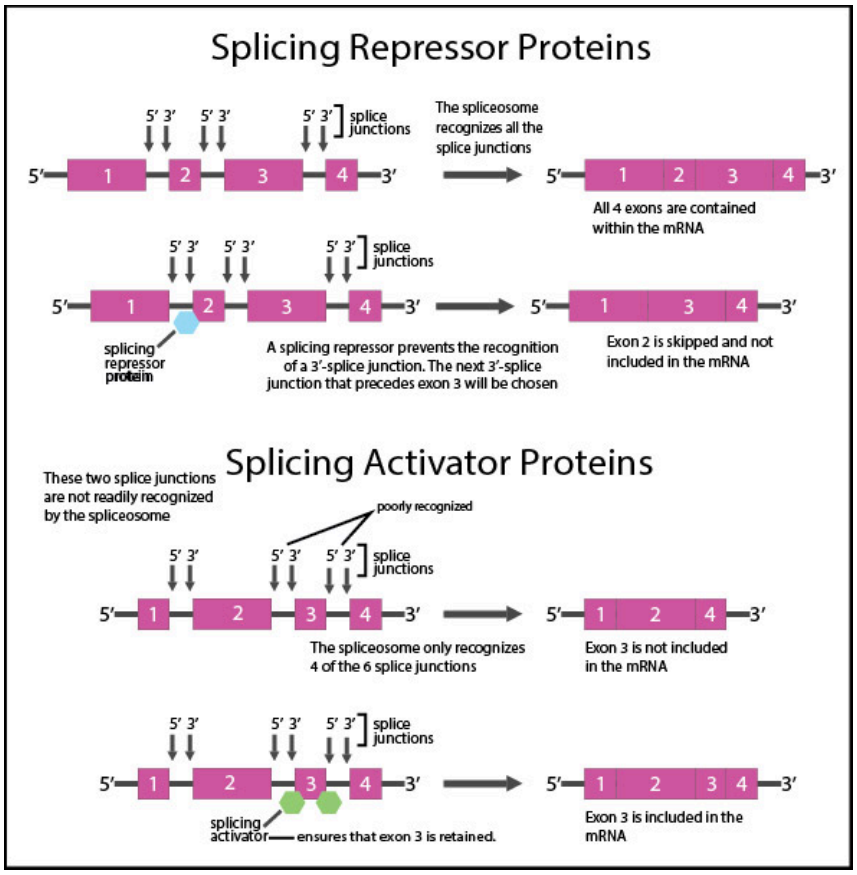


Figure 10.10 ***Splicing repressor and activator proteins*** *--- Image created by SL*

Key Questions

- What happens when a splice repressor protein binds to the 3’ splice site within an intron?
- What effect would a splice activator protein binding to a splice enhancer sequence have on alternative splicing?
- What is meant by the term mutually exclusive exons?

Review Questions

Fill in the blank:

1. _____ is an endonuclease that releases the pre-mRNA from RNA polymerase II to terminate transcription.
2. _____ is an enzyme that attaches two nucleotides together via a 5’ to 5’ linkage.
3. One function of the 7-mG cap is to _____.
4. A _____ protein prevents the spliceosome from binding to a 3’ splice site.
5. _____ is an enzyme that adds adenine nucleotides to the 3' end of a pre-mRNA. These adenine nucleotides are added in the 5' to 3' direction.
6. The Group I intron splicing mechanism uses the nucleoside _____ as a cofactor during catalysis, while the _____ intron splicing mechanism uses an adenine nucleotide as a cofactor during catalysis.
7. The U2 snRNP binds to the _____ site of the pre-mRNA.
8. Spliceosome subunits are composed of two components: proteins and _____.
9. _____ is a pattern of alternative splicing in which an exon is always retained in one cell while that same exon is always skipped in another cell.

11 - Translation

Translation is the cellular process that converts the language of nucleic acids contained within the mRNA molecule (A, U, G, C) into the language of amino acids in a synthesized protein (methionine, alanine, histidine, etc.). This translation process relies on a **genetic code** that converts nucleic acid sequence into amino acid sequence. Note that the genetic code that is translated into protein functions within the mRNA molecule, not in the DNA.

The mRNA molecule is read by the **ribosome** during translation as a group of three consecutive nucleotides, a so called **triplet code**. Each combination of three nucleotides within the mRNA sequence is a **codon**. In addition to the codons, the mRNA contains several RNA sequence features that play roles in translation (**figure 11.1**). These sequence features are listed from 5′ to 3′ along the mRNA molecule below:

- **5′ untranslated region (5′-UTR)**. The 5′-UTR of the mRNA is not translated by the ribosome. Instead, the 5′-UTR allows the ribosome to bind to the mRNA and positions the ribosome to interact properly with the start codon (see below). In bacteria, the 5′-UTR of the mRNA contains a **Shine-Dalgarno sequence** that serves as the binding site for the ribosome. In eukaryotes, the 5′-UTR contains the **7-methylguanosine (7-mG)** cap added during RNA modification (see Part 10), which serves as the ribosome binding site. The 5'-UTR in eukaryotes also contains a **Kozak sequence**, which positions the ribosome to recognize the start codon.
- **Start codon**. The **start codon** in the mRNA codes for the first amino acid in the synthesized protein. The start codon is usually **5'-AUG-3'** and encodes the amino acid ***N*-formylmethionine (fmet)** in bacteria and **methionine (met)** in eukaryotes.
- **Sense codons.** After the start codon, the mRNA is read as consecutive three nucleotide-long **sense codons**. Each sense codon specifies an amino acid in the protein; the specific amino acid sequence of the protein depends on the sequence of sense codons. Note that since the start codon and the sense codons encode amino acids, the start and sense codons constitute the **coding region** of a gene.
- **Stop codon.** The stop codon in the mRNA serves as a signal to end translation. The possible stop codons within prokaryotic and eukaryotic mRNAs include **5′-UAG-3′**, **5′-UAA-3′**, or **5′-UGA-3′**; each mRNA uses one of these three stop codons. The stop codon does not encode an amino acid within the synthesized polypeptide.
- **3′ untranslated region (3′-UTR).** The 3′-UTR of the mRNA, located downstream of the stop codon, plays a role in transcription termination. In prokaryotes, the 3′-UTR contains the *rut*, stem-loop, and the uracil-rich sequences that function in rho (ρ)-dependent and rho (ρ)-independent transcriptional termination (see Part 9). In eukaryotes, the 3′-UTR contains the polyadenylation signal sequences recognized by the CPSF and CstF proteins during transcription termination (see Parts 9 and 10).

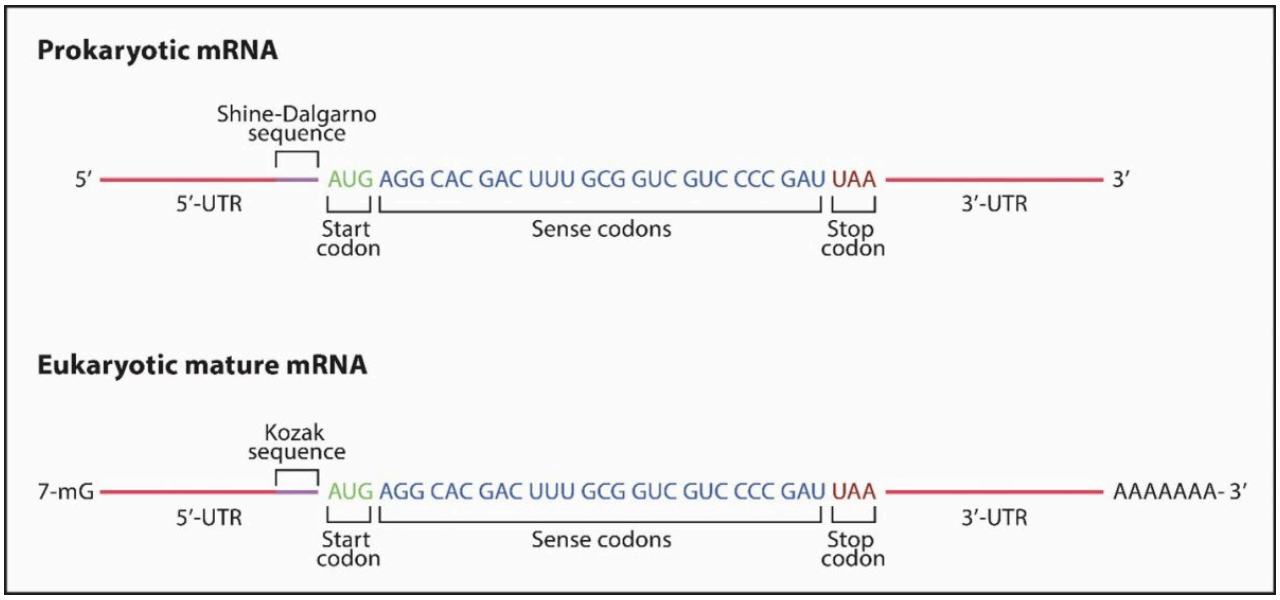


Figure 11.1 ***- Prokaryotic and eukaryotic mRNA features.*** *--- Image created by SL*

Key Questions

- Identify one important sequence feature found in the 5′-UTR of bacterial mRNAs.
- Identify two important sequence features found within the 5′-UTR of eukaryotic mRNAs.
- What amino acid is encoded by 5′-AUG-3′ in bacteria?
- What amino acid is encoded by 5′-AUG-3′ in eukaryotes?
- What is the function of the coding region?
- What is the function of the stop codon?
- Is the stop codon part of the coding region?
- What is the function of the 3′-UTR?

The Genetic Code

There are 20 different types of amino acids typically incorporated in proteins; however, there are 64 possible combinations of four nucleotides (A, U, G, and C) in a triplet codon RNA sequence. As a result, there are more codons than amino acid possibilities (see **figure 11.2**). Thus, the genetic code is **degenerate**, meaning that more than one codon encodes a particular amino acid type. For example, the amino acid valine (val) is encoded by four codons (5′-GUU-3′, 5′-GUC-3′, 5′-GUA-3′, and 5′-GUG-3′) that differ only in the 3′-most base (**wobble base**). These four codons are called **synonymous codons** because they all encode valine.

The genetic code is also **unambiguous**, meaning that a particular codon sequence only encodes one type of amino acid. For example, the codon 5′-AAA-3′ always encodes the amino acid lysine (lys). Further, the genetic code is **commaless**, meaning that codons are read by the ribosome consecutively one triplet sequence after another, with no spacer nucleotides between codons. The codons are also **nonoverlapping**, meaning each nucleotide in the coding region of the mRNA is a member of only one codon. Finally, the genetic code is **universal**, meaning that the same genetic code is used in all organisms.

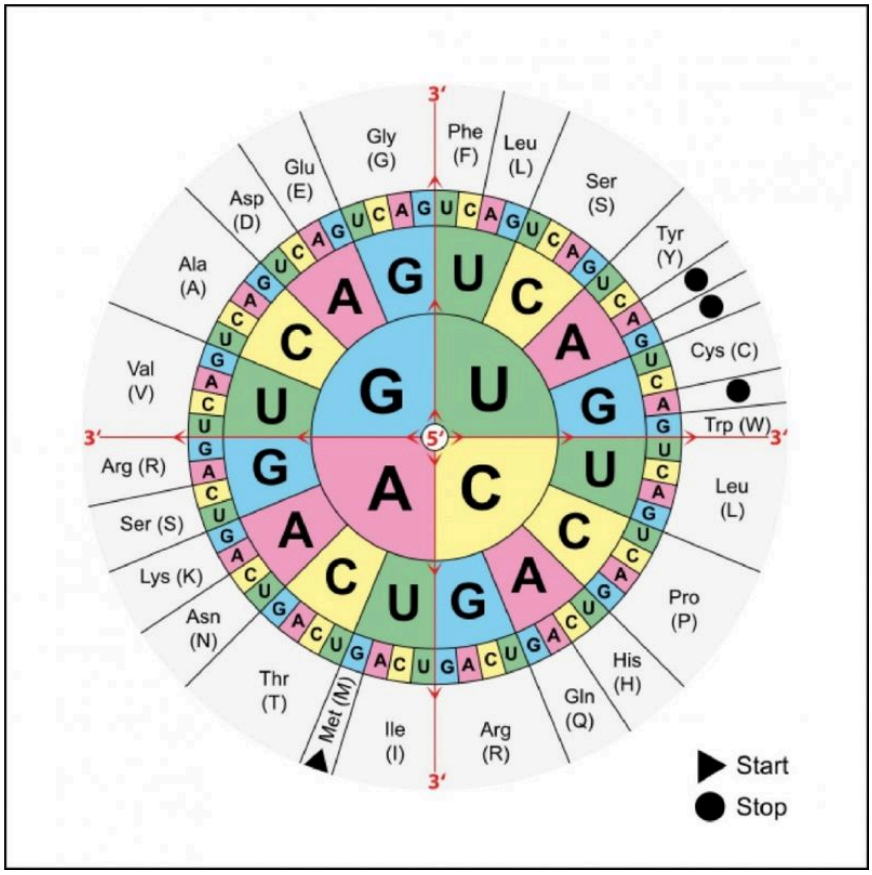


Figure 11.2 ***The genetic code.*** *This representation of the genetic code is read from the center of the circle (the 5'-most base in the codon) to the periphery of the circle (the 3'- most base in the codon). The encoded amino acids are indicated on the outside of the circle. Aminoacids Table was created by Mouagip and is used under CC0*

Key Questions

- How do the words degenerate, unambiguous, commaless, nonoverlapping, and universal apply to the codons in a mRNA sequence?

Directionality of Polypeptide Chains

The polypeptide chains synthesized during translation have directionality (polarity); however, since the language of proteins is different than the language of nucleic acids, the labels 5' and 3' do not apply to the polarity of polypeptide chains. Instead, the amino acid encoded by the start codon (closer to the 5' end of the mRNA) contains a free amino (NH_3^+) chemical group and thus is said to be the **amino** or **N-terminus** of the polypeptide chain**.** As the polypeptide chain is synthesized, **peptide bonds** are formed between the carboxyl group (COO^-) of the growing amino acid chain and the amino groups of incoming amino acids (see **figure 11.3**). Peptide bond formation involves a condensation reaction that releases water as a product. The final amino acid added to a polypeptide contains a free carboxyl chemical group and is called the **carboxyl terminal** or C-terminal end of the polypeptide chain. The C-terminal end of the polypeptide corresponds to the codon immediately before the stop codon (closer to the 3' end of the mRNA).

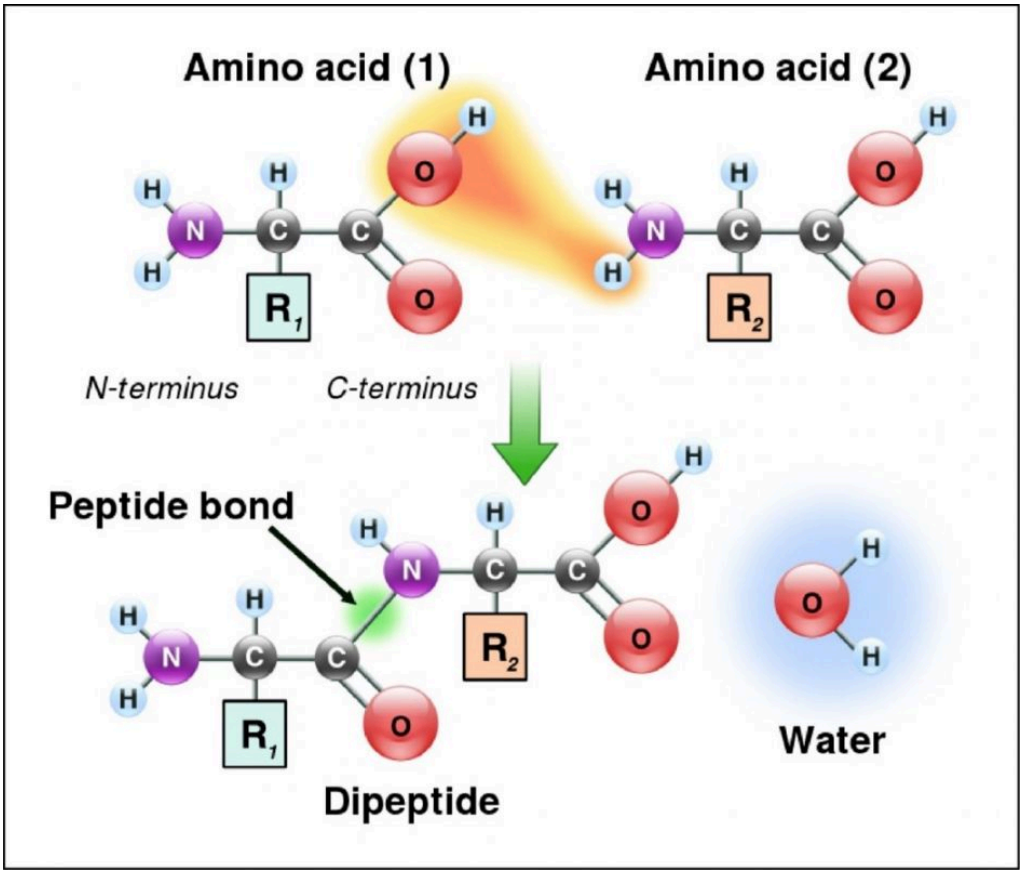


Figure 11.3 ***Peptide bond formation*** *--- Image Peptide Bond Formation was created by Yassine Mrabet and used under CC0*

Key Questions

- How do the N- and C-terminal ends of a protein compare to the 5' and 3' ends of an mRNA molecule?

tRNAs are Adaptor Molecules

Transfer RNA molecules **(tRNAs)** act as **adaptor molecules** in translation. tRNAs function to:

- **Recognize the codon nucleotide sequence within the mRNA.** This recognition process involves hydrogen bonding between the mRNA codon and an **anticodon** nucleotide sequence within the tRNA.
- **Carry a specific amino acid.** tRNA molecules carry amino acids to the ribosome. The carried amino acid is then incorporated into the growing polypeptide chain. The particular amino acid transported to the ribosome is specified by the anticodon sequence in the tRNA.

There are many different types of tRNA molecules in a cell. Each tRNA type is encoded by a different gene in the genome and has a unique anticodon sequence. tRNA types are also distinguished by the amino acid carried. For example, $tRNA^{val}$ carries the amino acid valine, while $tRNA^{phe}$ carries the amino acid phenylalanine. tRNA molecules are example **noncoding RNAs (ncRNAs)**, meaning that tRNA molecules themselves are untranslated; however, tRNAs function directly in the cell.

Key Questions

- What are the two functions of a tRNA molecule?

Features of tRNAs

Even though different tRNA types ($tRNA^{val}$ RNA^{phe}) have unique anticodons and carry unique amino acids, all tRNA molecules share similar structural features (**figure 11.4**):

- **tRNA molecules are small**. Each tRNA type is 75–90 nucleotides in length.
- **tRNA molecules are shaped like a cloverleaf**. Three stem-loop structures are formed by hydrogen bonding within the tRNA molecule. The anticodon RNA sequence is located within one of these stem-loop structures (i.e., the **anticodon loop**).
- **tRNA molecules contain a 3′ single-stranded region called an acceptor stem**. The acceptor stem region is covalently linked to the amino acid.
- **tRNA molecules contain variable regions**. The tRNA variable regions differ between tRNA types (e.g., $tRNA^{val}$ compared to $tRNA^{his}$) in the number of nucleotides and the specific nucleotide sequence that they contain. Variability in these regions can make the stem loops of different tRNA types larger or smaller.
- **tRNA molecules contain unusual nitrogenous bases**. Many tRNA types have unusual nitrogenous bases, such as inosine (I), methylinosine (mI), and dimethylguanine (m_2G). These unusual nitrogenous bases are created from the four conventional RNA nitrogenous bases (A, U, G, C) after transcription of the tRNA molecule has occurred.

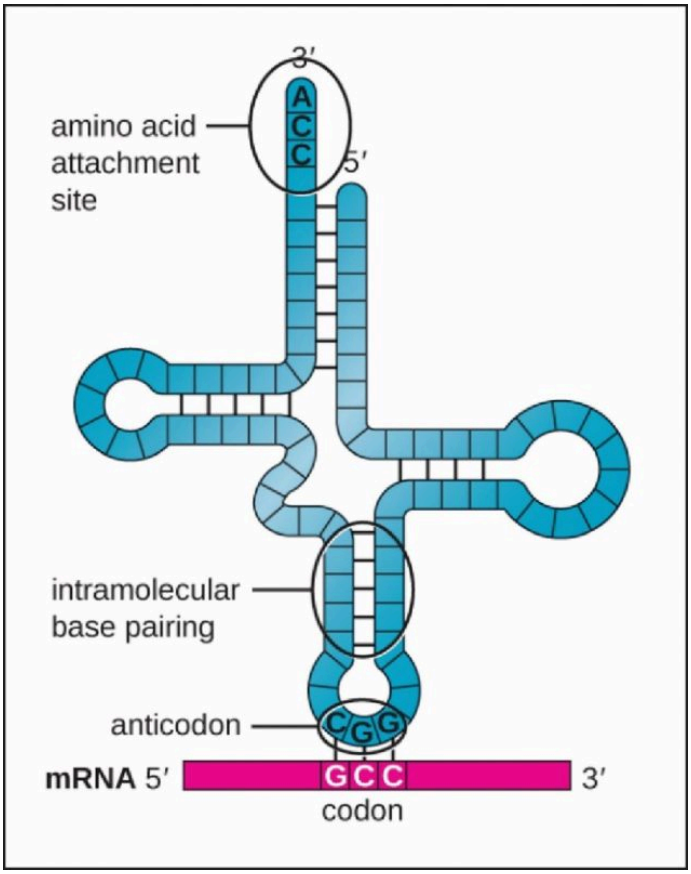


Figure 11.4 ***tRNA Structure.*** *tRNA molecules have three stem loop structures and an acceptor stem. --- Image used from OpenStax (access for free at* https://openstax.org/books/biology-2e/pages/1-introduction*)*

Key Questions

- Where is the anticodon in a tRNA molecule?
- Where is the amino acid attached to a tRNA molecule?

Charging tRNAs

The process of attaching an amino acid to the 3' end of a tRNA molecule is called **charging**. Charging tRNAs involves enzymes called **aminoacyl-tRNA synthetases**. Aminoacyl-tRNA synthetases have the following features:

- There are twenty different pools of aminoacyl-tRNA synthetases per cell, one pool for each of the twenty amino acid types.
- Aminoacyl-tRNA synthetases are named for the amino acid that is attached to the tRNA. For example, alanyl-tRNA synthetases attach the amino acid alanine (ala) to a tRNA molecule, while prolyl-tRNA synthetases attach the amino acid proline (pro) to a tRNA molecule.
- Aminoacyl-tRNA synthetases attach the incorrect amino acid to a tRNA rarely. In fact, it is estimated that aminoacyl-tRNA synthetases attach the wrong amino acid only one time per 100,000 tRNA molecules charged.

The mechanism used by an aminoacyl-tRNA synthetase to charge a tRNA molecule has the following steps (see **figure 11.5**):

1. The aminoacyl-tRNA synthetase binds to its preferred amino acid type and ATP. For example, the alanyl-tRNA synthetases bind to the amino acid alanine and ATP.
2. The ATP molecule is cleaved, PP_i (pyrophosphate) is released, and the remaining AMP is covalently linked to the amino acid. When AMP is linked to the amino acid, the aminoacyl-tRNA synthetase changes its conformation (shape) to produce the active form of the enzyme. This active enzyme conformation allows the aminoacyl-tRNA synthetase to bind to a tRNA molecule.
3. The aminoacyl-tRNA synthetase binds to a tRNA molecule that contains an anticodon sequence specific for the amino acid carried by the aminoacyl-tRNA synthetase. For example, an alanly-tRNA synthetase would only bind to tRNA molecules that have anticodons that would form hydrogen bonds with alanine codons in the mRNA. If the anticodon on the incoming tRNA is incorrect, the tRNA is released from the aminoacyl-tRNA synthetase and a new tRNA binds instead.
4. The aminoacyl-tRNA synthetase attaches the amino acid to the acceptor stem (3' single-stranded region) of the tRNA.
5. AMP is released.
6. The **charged tRNA** (i.e., tRNA covalently linked to the correct amino acid) is released from the aminoacyl-tRNA synthetase. This charged tRNA can then be delivered to the ribosome.

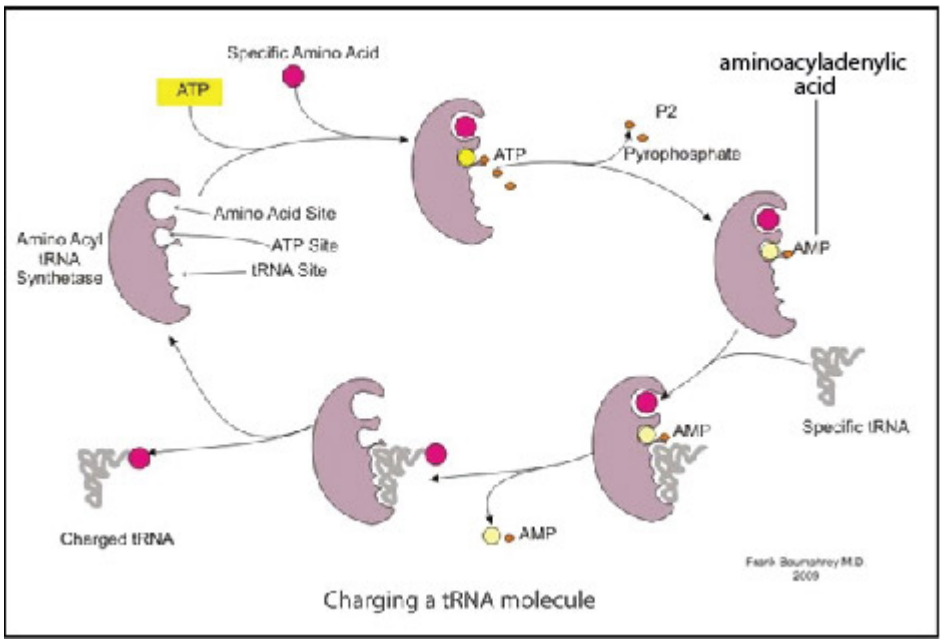


Figure 11.5 ***Charging a tRNA Molecule ---*** *Image by Dr. Frank Boumphrey used under license CC BY-SA 3.0*

Key Questions

- Which molecule provides the energy for the tRNA charging reaction?
- Explain the major events that occur during tRNA charging.

Wobble

Most synonymous codons (codons that encode the same amino acid) have identical bases at the first two nucleotide positions (the 5'-most base and the middle base of the codon). These first two bases obey the AU/GC base pairing rules when forming hydrogen bonds with the tRNA anticodon. Degeneracy in the genetic code typically occurs at the 3'-most base within the codon; the 3'-most base in the codon does not have to form conventional base pairing interactions with the 5'-most base of the anticodon, leading to "wobble." The rules that govern codon-anticodon at this wobble position

interactions are called the **wobble rules** (see **Table 11.1**). For example, a U base in the wobble position of the codon (3'-most base) can form hydrogen bonds with A, G, or I in the wobble position of the anticodon (5'-most base). Similarly, a U in the wobble position of the anticodon (5'-most base) can form hydrogen bonds with either A or G in the wobble position of the codon (3'-most base).

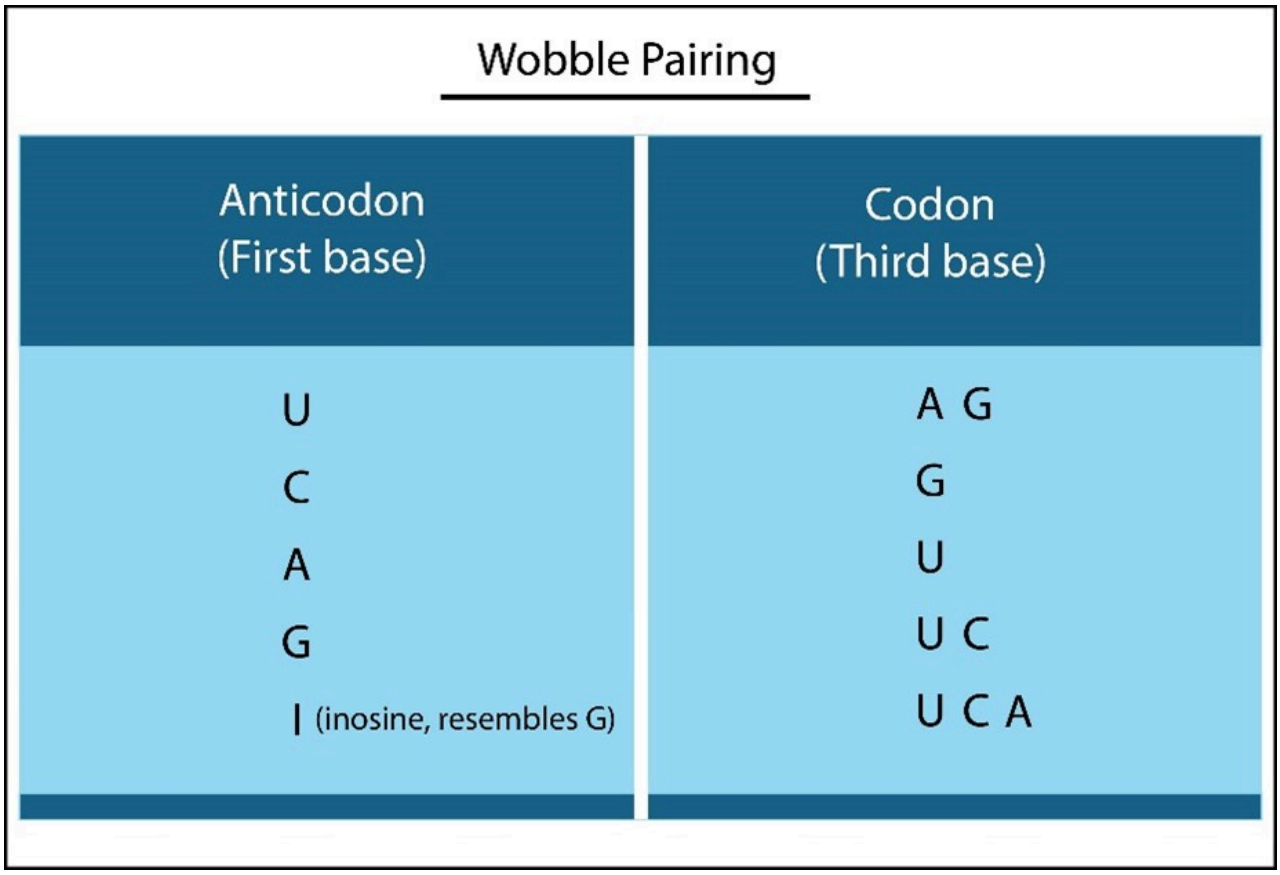


Table 11.1 ***Wobble Rules Table.***

A group of tRNA types, with slightly different anticodon sequences, which recognize the same mRNA codon sequence are called **isoacceptor tRNAs**. For example, suppose we have the codon 5'-UUU-3' in the mRNA. tRNA molecules containing either 3'-AAA-5', 3'-AAG-5', or 3'-AAI-5' anticodons can recognize this mRNA codon, according to the wobble rules (see Table 11.1). Note that these isoacceptor tRNAs are charged by the same aminoacyl-tRNA synthetase and therefore, bear the same amino acid (phenylalanine). Alternatively, a tRNA with the 3'-GGU-5' anticodon can recognize two codons: 5'-CCA-3'and 5'-CCG-3' according to the wobble rules. Both codons encode proline.

Why is this wobble phenomenon advantageous to cells? Wobble gives the cell tremendous flexibility in codon:anticodon interactions. A single mRNA codon can be recognized by more than one tRNA anticodon, and a single tRNA anticodon can bind to more than one mRNA codon. This flexibility allows translation to occur at a reasonable rate and allows a cell to save energy in the synthesis of tRNA molecules. Instead of assembling 61 different types of tRNA molecules (one type for each possible sense codon, minus the three stop codons), a cell can get away with synthesizing only 30-40 types of tRNA molecules, with some tRNAs recognizing multiple mRNA codon sequences.

Key Questions

- What is meant by wobble?
- What are isoacceptor tRNAs?
- What are the two advantages of wobble?

Prokaryotic Ribosomes

The multisubunit **ribosome** is the central figure in the translation process. The ribosome binds to the mRNA, serves as a site for mRNA codon:tRNA anticodon recognition, breaks the covalent bond between the 3' end of the tRNA and the amino acid, and synthesizes peptide bonds between amino acids.

A single bacterial cell is thought to contain approximately 10,000 ribosomes; these bacterial ribosomes are called **70S ribosomes**. Note that the "S" designation of ribosomes correlates roughly with the overall three-dimensional shape of the ribosome component; the larger the molecule, the larger the "S" value. The 70S ribosome consists of the following subunits (see **figure 11.6**):

- **Small subunit (30S subunit).** The small subunit of the ribosome is composed of several different ribosomal proteins and a **16S ribosomal RNA (rRNA)** molecule. The 16S rRNA molecule helps the ribosome bind to the bacterial mRNA molecule (see below).
- **Large subunit (50S subunit)**. The large subunit of the ribosome is composed of many ribosomal proteins, a 5S rRNA, and a **23S rRNA** molecule. The 23S rRNA molecule is the catalytic core of the ribosome; the 23S rRNA is the RNA enzyme (ribozyme) that forms peptide bonds between amino acids during prokaryotic translation (see below).

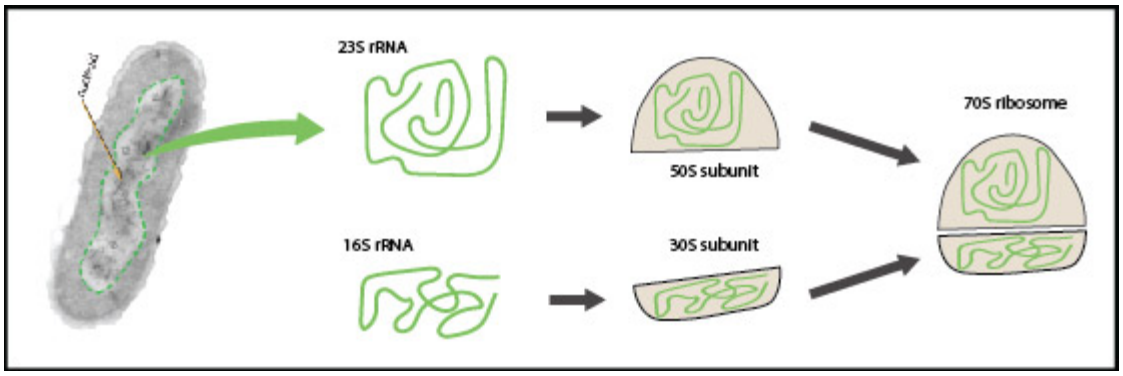


Figure 11.6 ***Prokaryotic Ribosomes*** *--- Prokaryotic cell (Left) used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction).Image created by SL*

Eukaryotic Ribosomes

Eukaryotic cells have cytosolic ribosomes called **80S ribosomes**. Mitochondrial and chloroplast ribosomes resemble prokaryotic 70S ribosomes. The 80S ribosomes in the cytosol synthesize most cellular proteins and consist of the following subunits (see **figure 11.7**):

- **Small subunit (40S).** The 40S subunit is composed of many proteins and an 18S rRNA molecule.
- **Large subunit (60S).** The 60S subunit is composed of a collection of proteins, 5S, 5.8S, and **28S rRNA** molecules. The 28S rRNA molecule is the RNA enzyme (ribozyme) that forms peptide bonds between amino acids during eukaryotic translation (see below).

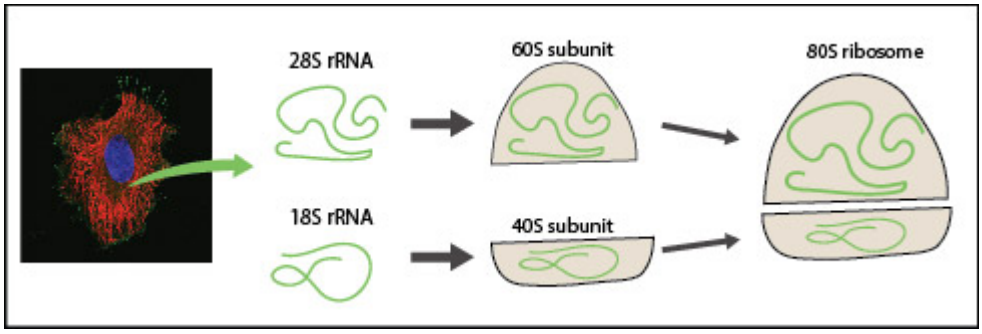


Figure 11.7 ***Eukaryotic ribosomes*** *--- Image created by SL*

Key Questions

- Compare and contrast the structural features of prokaryotic vs. eukaryotic ribosomes.
- What is the name of the enzyme that forms peptide bonds within the 70S ribosome?
- What is the name of the enzyme that forms peptide bonds within the 80S ribosome?

Overview of Translation

Translation in prokaryotes and eukaryotes occurs in three stages (see **figure 11.8**):

1. During **initiation**, the large and small ribosomal subunits assemble near the mRNA start codon with an **initiator tRNA** molecule that has an anticodon sequence specific for the start codon mRNA sequence. Once all translation components have assembled correctly, the initiator tRNA is located at the P site (see below) within the ribosome.
2. During **elongation**, the ribosome **translocates** (moves) along the mRNA in the 5' to 3' direction, reading triplet codons within the mRNA, converting the codon sequences into a chain of amino acids. During elongation, three tRNA binding sites within the ribosome are used. These tRNA binding sites are the:
   - **Aminoacyl site (A site).** Charged tRNAs enter the ribosome at the A site.
   - **Peptidyl site (P site).** During a translation elongation cycle (see below), the polypeptide chain attached to the tRNA molecule in the P site is transferred to the tRNA in the A site as a peptide bonds is formed.
   - **Exit site (E site)**. The E site allows uncharged tRNA molecules to exit the ribosome.
3. When the ribosome reaches a stop codon, translation **terminates**. The mRNA and the polypeptide chain are released from the ribosome, and the large and small subunits of the ribosome dissociate from each other. The ribosome subunits and the mRNA can be recycled to synthesize another copy of the protein.

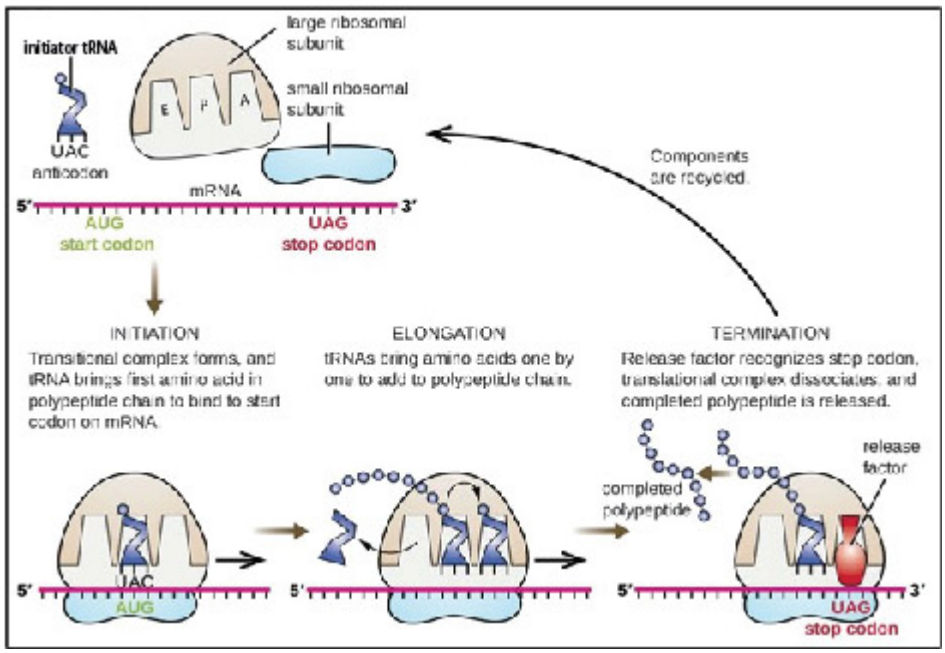


Figure 11.8 ***Overview of Translation ---*** *image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction), modified by SL*

Key Questions

- What major events are occurring during the initiation, elongation, and termination stages of translation?
- What are the functions of the A, P, and E sites?

Shine-Dalgarno Sequence

The **5'-untranslated region (5'-UTR)** of prokaryotic mRNAs plays a critical role in translation initiation, as the 5'-UTR contains the ribosome-binding site. Specifically, the 5'-UTR contains a recognition sequence (5'-AGGAGGU-3') called the **Shine-Dalgarno sequence** that forms hydrogen bonds with a complementary sequence within the **16S rRNA** of the small ribosomal subunit (30S) (see **figure 11.9**). Hydrogen bond formation between the Shine-Dalgarno sequence and the 16S rRNA promotes ribosome assembly. Additionally, the Shine-Dalgarno sequence positions the P site of the ribosome properly at the start codon to initiate polypeptide synthesis.

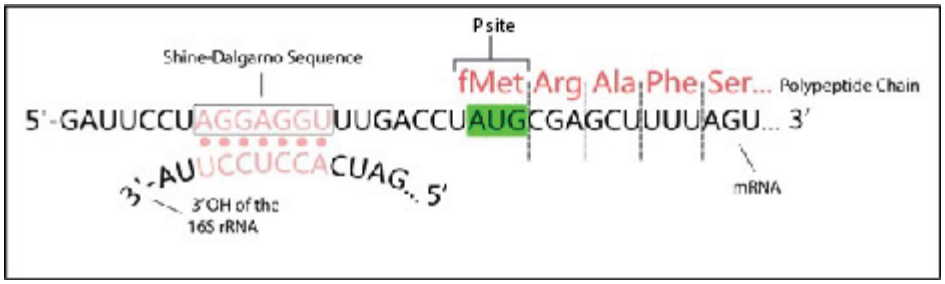


Figure 11.9 ***The Shine-Dalgarno Sequence ---*** *image created by Alejandro Porto and modified by SL. Used under license CC BY-SA 3.0*

Key Questions

- What is the function of the Shine-Dalgarno sequence?

Translation Initiation in Bacteria

Translation initiation in the bacterium *E. coli* includes the following steps (see **figure 11.10**):

1. **The mRNA binds to the small ribosomal subunit (30S) via the Shine-Dalgarno sequence.** An initiation factor protein called **IF3** promotes the formation of hydrogen bonds between the 16S rRNA (ribosome) and the Shine-Dalgarno sequence (mRNA).
2. **The initiator tRNA (tRNA^{fmet}) recognizes the start codon within the P site of the ribosome.** The initiator tRNA anticodon forms hydrogen bonds with the mRNA start codon in the P site. In bacteria, this initiator tRNA is covalently attached to a modified form of methionine called ***N*-formylmethionine (fmet)**. Thus, the N-terminal amino acid in all newly synthesized prokaryotic proteins is fmet. Binding of the tRNA^{fmet} to the start codon requires the initiation factor protein **IF2.** IF2 cleaves GTP to load the tRNA^{fmet} into the ribosome P site. The complex of the 30S subunit, the mRNA, IF3, and IF2 is called the **initiation complex**.
3. **The 50S subunit of the ribosome is added to the initiation complex.** The addition of the 50S ribosome subunit requires the release of IF2 and IF3.

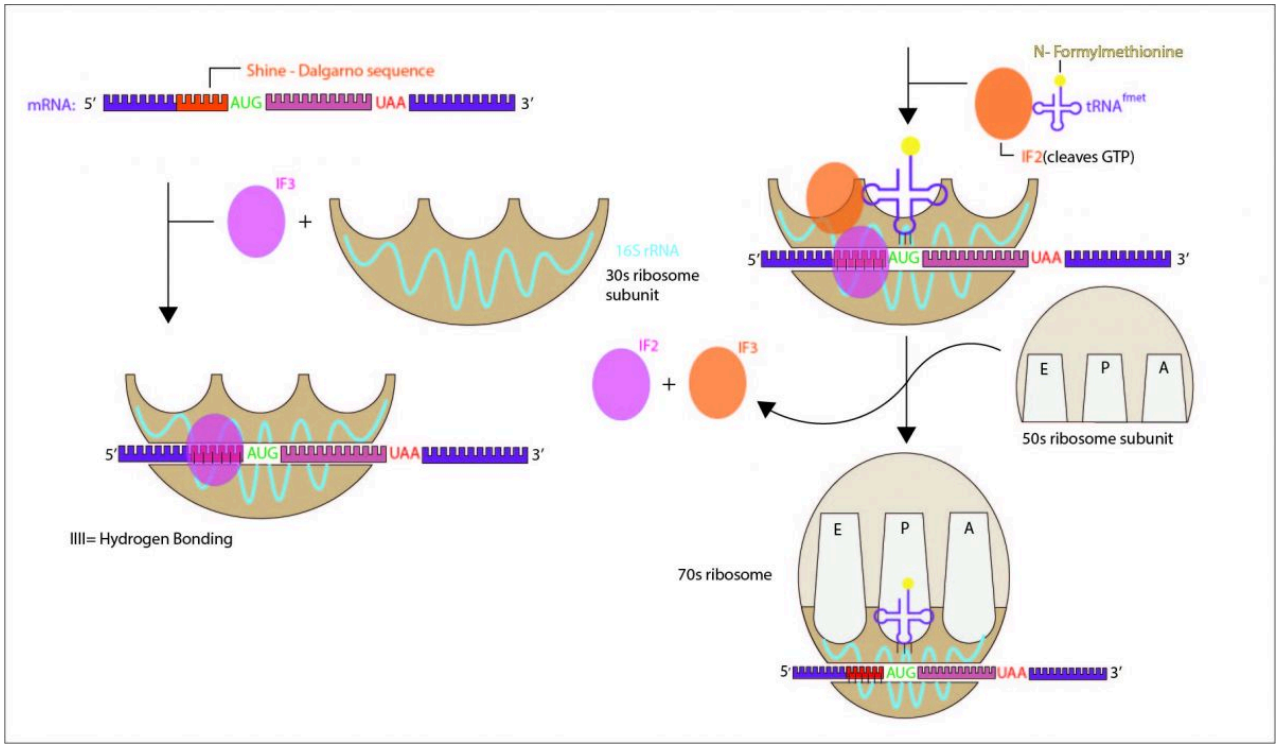


Figure 11.10 ***Translation initiation in bacteria*** *--- Image created by JET*

Key Questions

- How do IF2 and IF3 function in translation initiation in bacteria?
- What molecule provides the energy for translation initiation in bacteria?

Translation Initiation in Eukaryotes

Translation initiation is more complex in eukaryotes than in prokaryotes. Key differences between eukaryotic and prokaryotic translation initiation include (see **figure 11.11**):

- More initiation factor proteins are involved in eukaryotes. These initiation factor proteins are called **eukaryotic initiation factors (eIFs)**.
- The initiator tRNA in eukaryotes is **$tRNA^{met}$**. This initiator tRNA is charged with the amino acid **methionine (met)**.
- There is no Shine-Dalgarno sequence in the eukaryotic mature mRNA. Instead, the 5' UTRs of eukaryotic mature mRNAs contain a **7-methylguanosine (7-mG)** cap and a **Kozak sequence**.

The process of translation initiation in eukaryotes is as follows:

1. **The eIF2 protein brings $tRNA^{met}$ to the 40S ribosomal subunit**. eIF2 cleaves GTP to deliver $tRNA^{met}$ to the 40S ribosomal subunit.
2. **Assembly of a multiprotein initiation complex on the 7-methylguanosine (7-mG) cap on the mature mRNA.** This initiation complex includes:
   - the 40S ribosome subunit.
   - an **eIF4** protein.
   - a complex composed of eIF2 and $tRNA^{met}$
   - **cap-binding protein 1 (CBP1)**. CBP1 is the protein responsible for recognizing the 7-mG structure on eukaryotic mature mRNAs.
3. **Identification of the start codon**. The 40S ribosomal subunit and associated proteins move from the 7-mG cap 5’ to 3’ along the mRNA looking for a start codon. This scanning process requires the cleavage of ATP. Not all possible start codons (5’-AUG-3’) are chosen to initiate translation. The 5’-AUG-3’ that is chosen is within the **Kozak sequence:** 5’-GCC(A or G)CC**AUG**G-3’. When the $tRNA^{met}$ forms hydrogen bonds with the start codon, $tRNA^{met}$ is in the P site of the ribosome.
4. **Addition of the 60S ribosomal subunit to form the active 80S ribosome**. When the 60S ribosome subunit is added, eIF2, eIF4, and CBP1 are released.

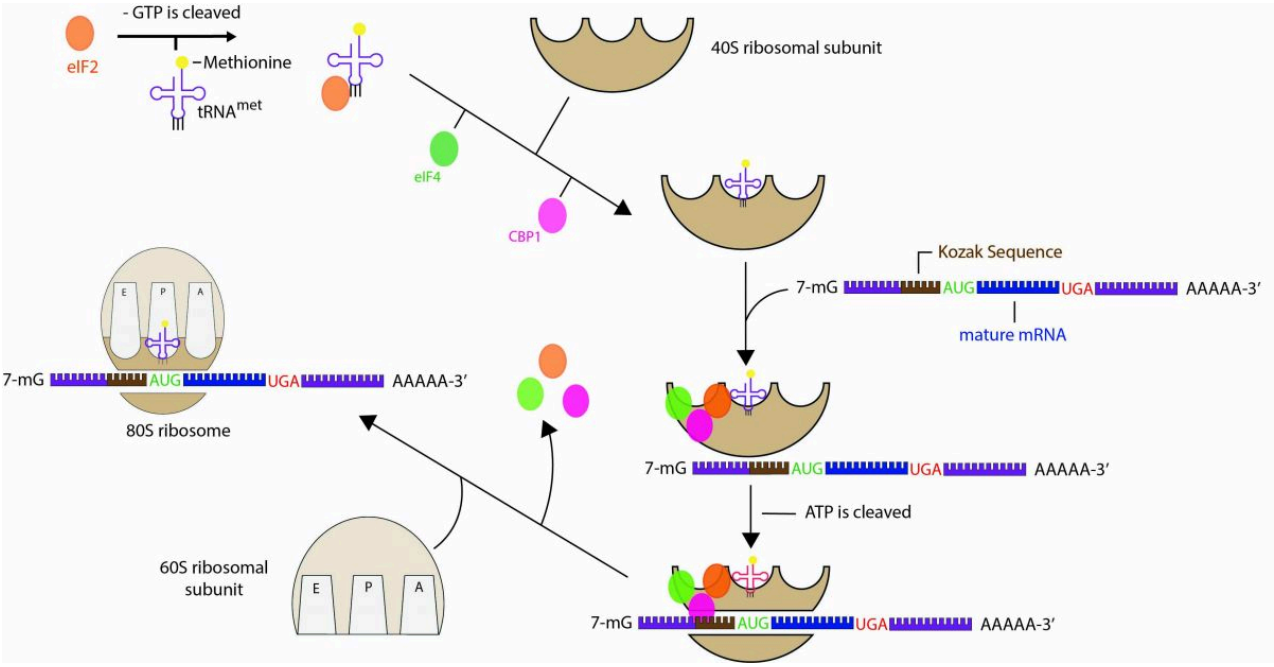


Figure 9.11 - Translation initiation in eukaryotes.

Key Questions

- How do eIF2 and CBP1 function in translation initiation in eukaryotes?
- What is the function of the Kozak sequence?
- Which two nucleotides provide the energy for translation initiation in eukaryotes?

Translation Elongation

Translation elongation in bacteria involves a series of cycles, one cycle per sense codon in the mRNA. At the beginning of each cycle, the tRNA molecule located within the P site of the ribosome is either attached to single amino acid (beginning of elongation cycle one) or a chain of amino acids (later elongation cycles; see **figure 11.12**). The A and E sites of the ribosome are empty. Each elongation cycle then proceeds as follows:

1. **A charged tRNA is delivered to the ribosome A site**. The prokaryotic **elongation factor** protein **EF-Tu** is responsible for loading a charged tRNA into the A site of the ribosome. This loading of a charged tRNA into the A site requires the cleavage of GTP.
2. **The codon:anticodon hydrogen bonds are checked**. Of course, the anticodon within the newly delivered tRNA must form complementary hydrogen bonds with the mRNA codon according to the wobble rules. An incorrect tRNA bound to the A site is recognized by the **16S rRNA**. When an incorrect anticodon is encountered, polypeptide synthesis is halted until the mismatched tRNA is released from the ribosome. This editing function of the 16S rRNA molecule allows for high fidelity in protein synthesis; it is estimated that only one mistake is made per 10,000 incorporated amino acids, or about one time in every 20 proteins synthesized by the ribosome.
3. **The formation of a peptide bond**. The amino acid or amino acid chain bound to the tRNA in the P site is transferred to the tRNA in the A site via the formation of a new peptide bond. This enzyme reaction is catalyzed by the **23S rRNA** component of the 50S ribosomal subunit; the 23S rRNA is an example of a ribozyme (RNA enzyme). The 23S rRNA ribozyme is also called **peptidyl transferase**.
4. **Ribosome translocation**. The ribosome **translocates** (moves) in the 5' to 3' direction one codon along the mRNA. The tRNA (with the attached polypeptide chain) that was in the A site moves to the P site. The uncharged tRNA that was in the P site moves to the E site and exits the ribosome. A bacterial elongation factor protein called **EF-G** cleaves GTP to translocate the ribosome to the next sense codon in the mRNA.
5. **Steps 1–4 are repeated for each sense codon in the mRNA**. The rate of protein synthesis is 15–18 cycles (i.e., 15-18 amino acids) per second in prokaryotes.

Translation elongation works similarly in eukaryotes; however, there are some minor differences:

- The 28S rRNA component of the 60S ribosomal subunit functions as the peptidyl transferase to synthesize peptide bonds between amino acids. The 28S rRNA is another example of a ribozyme (RNA enzyme).
- Eukaryotes have **eukaryotic elongation factor** proteins **(eEFs)** that function similarly to the EF-Tu and EF-G proteins from bacteria. The eukaryotic **eEF1** protein functions similarly to EF-Tu, delivering charged tRNA molecules to the A site of the ribosome. The eukaryotic **eEF2** protein functions similarly to EF-G in bacteria, translocating the ribosome to the next mRNA codon after peptide bond formation has taken place. Both eEF1 and eEF2 cleave GTP molecules.

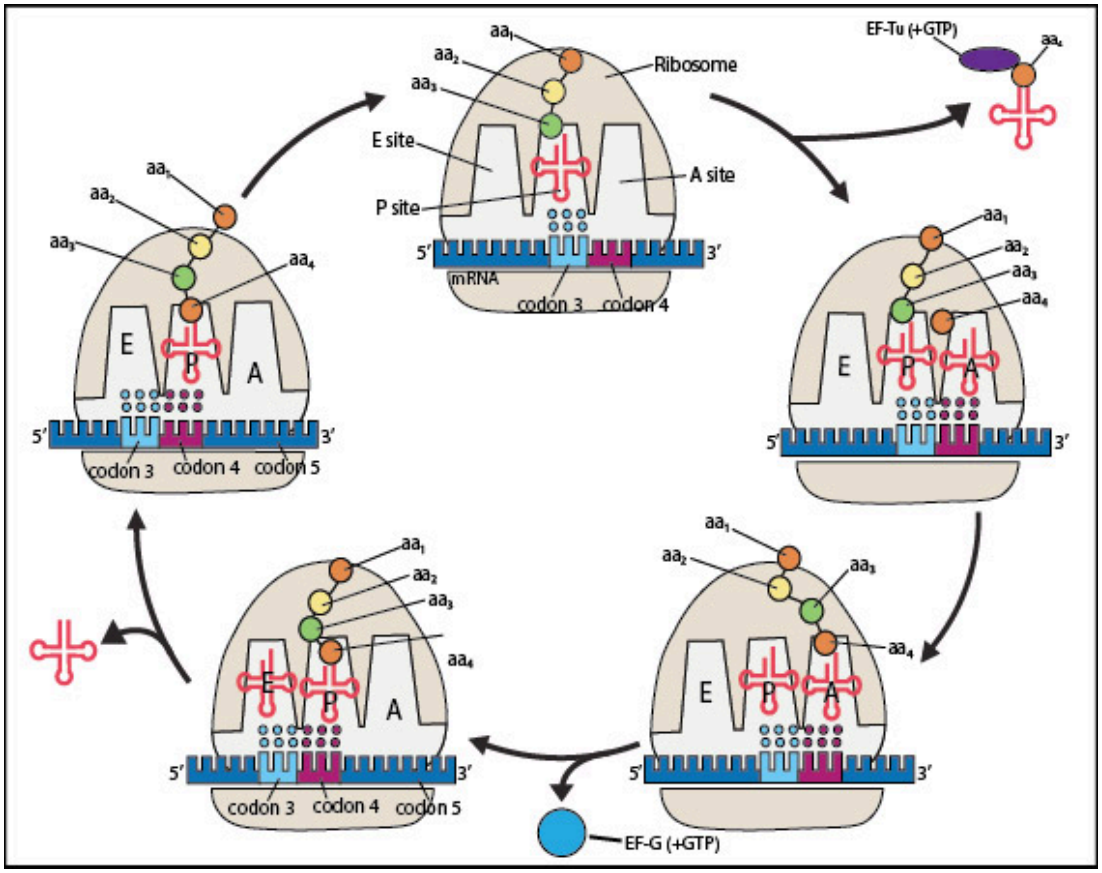


Figure 11.12 ***A translation elongation cycle in bacteria*** *-- The beginning of the translation cycle is shown at the top of the image. Image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction), modified by SL*

Key Questions

- Describe the events involved in one translation elongation cycle.
- Explain when GTP is used during the elongation cycle.
- What is the eukaryotic equivalent of 23S rRNA, EF-Tu, and EF-G?

Translation Termination

Translation termination in bacteria occurs as follows (see **figure 11.13**):

1. As the ribosome approaches the 3' end of the mRNA, a **stop codon (5'-UGA-3', 5'-UAG-3',** or **5'-UAA-3')** moves into the A site of the ribosome.
2. The stop codon is recognized by a **release factor** protein. The release factor protein mimics the three-dimensional structure of a tRNA molecule. In bacteria, the **release factor 1 (RF1)** protein recognizes the stop codons 5'-UAA-3' and 5'-UAG-3'. The **release factor 2 (RF2)** protein recognizes the 5'-UAA-3' and 5'-UGA-3' stop codons.
3. The release factor protein cleaves the covalent bond between the tRNA and the polypeptide chain in the ribosome P site. As a result, the tRNA and the polypeptide chain are released from the ribosome. The polypeptide chain folds to become the mature protein. This release step requires the release factor protein to cleave GTP.
4. The ribosomal subunits, mRNA, and release factor protein dissociate, terminating translation. The ribosome subunits and the mRNA can be recycled to begin translation again.

Translation termination works essentially the same way in eukaryotes. One minor diference is the use of a single **eukaryotic release factor (eRF)** protein that recognizes all three stop codons. eRF cleaves GTP to terminate translation in eukaryotes.

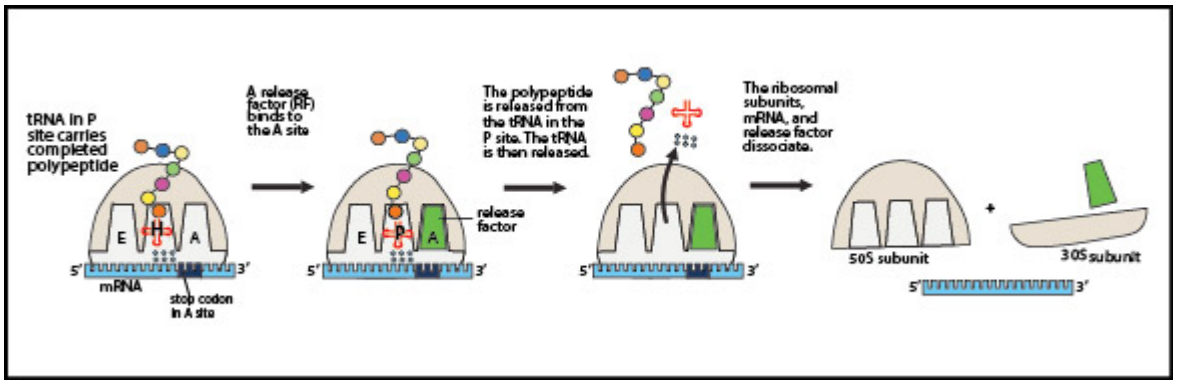


Figure 11.13 ***Translation termination*** *--- Image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction), modified by SL*

Key Questions

- What is the difference between RF1 and RF2?
- What molecule provides the energy to terminate translation?
- Are some of the molecules involved in translation recycled? If so, which ones?

Polysomes and Coupled Transcription and Translation

In the electron microscope, multiple ribosomes (**polyribosomes** or **polysomes**) are often observed attached to a single mRNA transcript (see **figure 11.14**). Polysomes have been observed attached to both prokaryotic mRNA and eukaryotic mature mRNA transcripts.

Since prokaryotic cells lack nuclei, transcription and translation are **coupled**, meaning that translation can begin before transcription is completed. In eukaryotes, transcription and translation are uncoupled. Transcription and RNA modifications occur within the nucleus. Translation occurs later, after the mature mRNA has been transported from the nucleus to the cytoplasm.

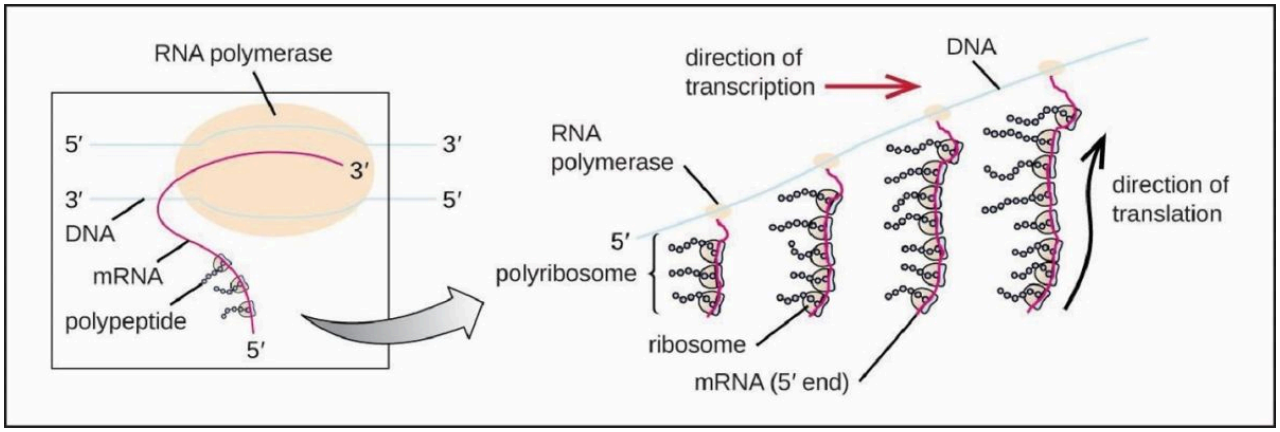


Figure 11.14 ***Polyribosomes and coupled transcription/translation in bacteria --- Image used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction)***

Key Questions

- Why is it advantageous to have multiple ribosomes translating a mRNA molecule simultaneously?
- Why is prokaryotic transcription and translation coupled?
- Why is eukaryotic transcription and translation uncoupled?

Review Questions

Fill in the blanks:

1. The _____ portion of a eukaryotic mature mRNA contains the polyadenylation signal sequences.
2. The genetic code is said to be _____, which means that each codon only specifies one type of amino acid. (For example, 5'-UUU-3' always encodes phenylalanine)
3. The cloverleaf structure of the tRNA is such that the _____ is a stem-loop that forms hydrogen bonds with the mRNA codon, and the _____ is the single-stranded region where an amino acid is attached.
4. _____ is an enzyme that links the amino acid alanine to a tRNA molecule.
5. The 30S and 50S ribosomal subunits in a prokaryotic cell combine to form a final size of _____ while in eukaryotes the _____ and _____ ribosomal subunits combine to form a final size of _____.
6. The initiator tRNA binds to the _____ site of the ribosome, whereas the remaining charged tRNAs bind to the _____ site of the ribosome.
7. An *E. coli* protein called _____ delivers charged tRNAs to the ribosome during translation elongation.
8. The ribosome moves along the mRNA in the _____to _____ direction.
9. _____ is a eukaryotic translation factor protein that delivers charged tRNAs to the ribosome during elongation.
10. _____ is a eukaryotic translation factor that helps the ribosome move one codon in the 3' direction along the mRNA. The equivalent bacterial protein is called _____.
11. The _____ protein in bacteria mimics the structure of a tRNA and recognizes 5'-UGA-3' in the A site of the ribosome.
12. An *E. coli* protein called _____ delivers $tRNA^{fmet}$ to the ribosome during translation.

12 - Gene Cloning

Gene cloning involves removing a gene from the genome of an organism and then placing that isolated gene into the genome of a bacterial cell. The bacterial cell is then responsible for maintaining this foreign gene. For examples, the bacterial cell copies the foreign gene by DNA replication, the bacterial transcribes the foreign gene to make a messenger RNA (mRNA) molecule, and the bacterial cell translates the mRNA to make a protein product. Many important human proteins, including insulin (for diabetes patients) and factor VIII (for type A hemophilia patients), have been produced in large quantities by bacterial cells via this gene cloning technique.

Overview

In the early 1970s, scientists isolated the DNA from two different organisms and covalently linked them together to form a hybrid DNA molecule in a test tube. This new hybrid DNA molecule, which contained DNA from two sources, is a **recombinant DNA molecule**. This experiment demonstrated the first use of **recombinant DNA techniques,** methods to manipulate DNA molecules outside of a living organism. Since then, many advances have made recombinant DNA techniques common practice among biologists.

Our discussion will focus on a process called **gene cloning** (**figure 12.1**). Gene cloning involves isolating a particular gene of interest and inserting that gene into a **vector** DNA molecule. Commonly used vectors include the circular **plasmid** DNA molecules found in many bacteria. The resulting recombinant DNA molecule, composed partly of the gene of interest and partly the plasmid vector, is then introduced into a host bacterial cell by **transformation**. The host bacterial cell maintains the recombinant DNA molecule, so the gene of interest (i.e., the **cloned gene**) can be studied in more detail. Once gene cloning is complete, the cloned gene can be used in:

- **DNA sequencing experiments.** DNA sequencing provides the base pair sequence of the cloned gene.
- **Mutagenesis experiments.** Mutations can be introduced at desired locations within the cloned gene. The phenotypic consequences of these mutations can then be studied.
- **Protein expression studies.** Cloned genes can be expressed via transcription and translation to make a protein product. Protein expression allows scientists to examine the function of the protein product encoded by a cloned gene or the cloned gene can be expressed at high levels for medical purposes (i.e., to express the human insulin or factor VIII proteins).
- **Gene therapy.** Cloned genes can be introduced into human cells as a treatment for genetic diseases.

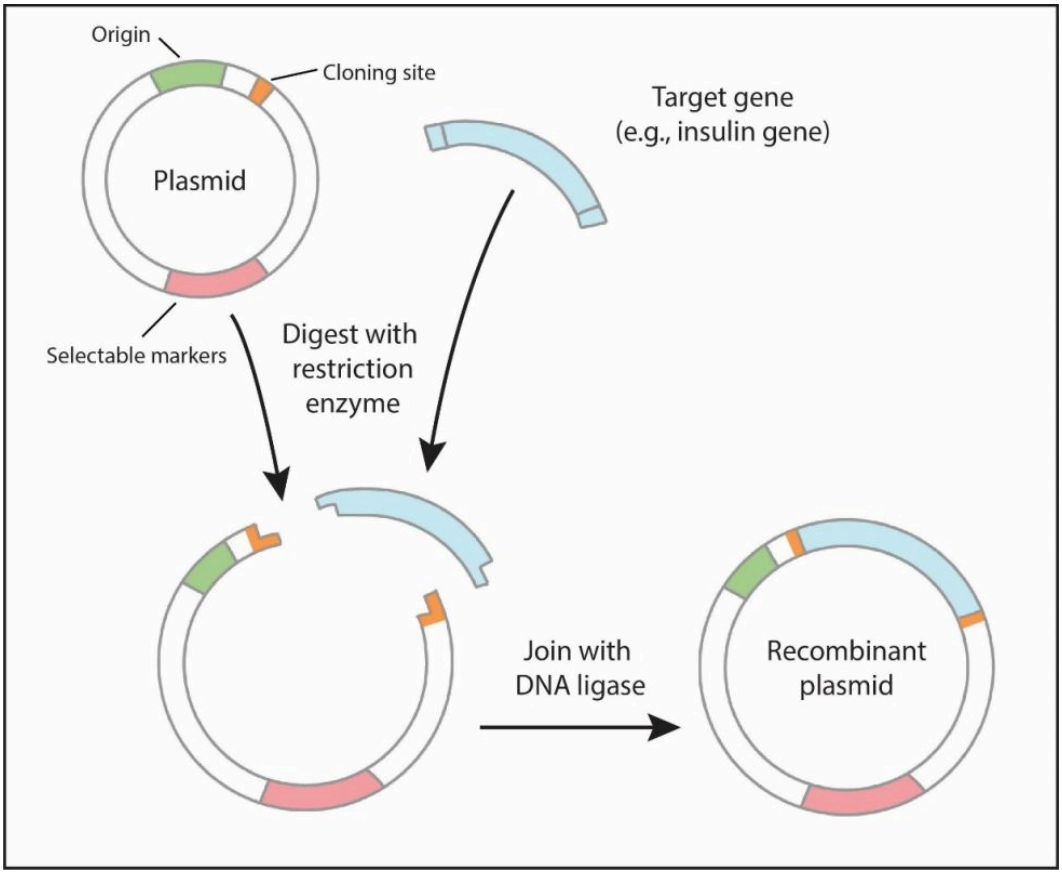


Figure 12.1 ***Gene cloning overview*** *- Image by Khan Academy and modified by SL*

Key Questions

- What is a recombinant DNA molecule?
- What is meant by the term gene cloning?
- Why is it useful to clone a gene?

Vectors

As mentioned earlier, gene cloning involves removing a gene of interest from an organism and inserting that gene into a **vector** DNA molecule (**figure 12.2**). The resulting recombinant DNA molecule is then introduced into a **host** organism for maintenance. The host organism is often the bacterium *Escherichia coli*.

From this point on, let us assume that we are interested in studying the insulin gene isolated from the human genome. The overall goal of our gene cloning experiment is to insert the human insulin gene into a vector DNA molecule and introduce the recombinant DNA molecule into the bacterium *E. coli*. In our scenario, the vector DNA molecule will function to:

- **Carry the insulin gene** (also called an **insert** or **cloned gene**). The insulin gene inserted into a vector can then be recognized and maintained by the host *E. coli* cell.
- **Copy the insulin gene**. Vectors are capable of efficient DNA replication within the host *E. coli* cell, and host *E. coli* cells often contain many copies of the vector. Thus, if the insulin gene is inserted in a vector, DNA replication of the vector will produce many copies of the insulin gene in each host *E. coli* cell.

Plasmids, small circular DNA molecules found in many bacteria, often serve as vector molecules (**figures 12.1** and **12.2**). These plasmids are not part of the bacterial chromosome. Plasmids contain:

- An **origin of replication**. The origin allows the plasmid to be replicated efficiently within the host cell. Some origins, such as *OriC*, allow replication of the plasmid within a particular host cell species (i.e., specifically the bacterium *E. coli*). Other origins allow efficient replication in different host species. The origin also determines the plasmid **copy number**. High copy number plasmids contain "strong" origins that allow frequent DNA replication to produce hundreds of plasmid DNA molecules per cell. Low copy number plasmids have less efficient origins, resulting in relatively few plasmid molecules per cell.
- **Unique gene cloning sites**. Gene cloning sites are plasmid DNA sequences that function as potential insertion sites for the cloned DNA sequence (i.e., the human insulin gene).
- **Selectable markers**. Plasmids contain selectable marker genes that confer an antibiotic resistant phenotype to the host bacterial cell. Thus, by growing host bacterial cells in the presence of an antibiotic, the researcher can ensure that the bacterial cells contain a plasmid. Common selectable marker genes include the ***amp^R*** gene, which confers resistance to the antibiotic ampicillin, and the ***tet^R*** gene, which confers resistance to the antibiotic tetracycline. In the case of the *amp^R* gene, a bacterial cell that is resistant to ampicillin contains the plasmid; a bacterial cell that is sensitive to ampicillin does not contain the plasmid. Ampicillin resistant bacteria survive when grown on agar plates that contain ampicillin; ampicillin sensitive bacteria die when grown on agar plates containing ampicillin.

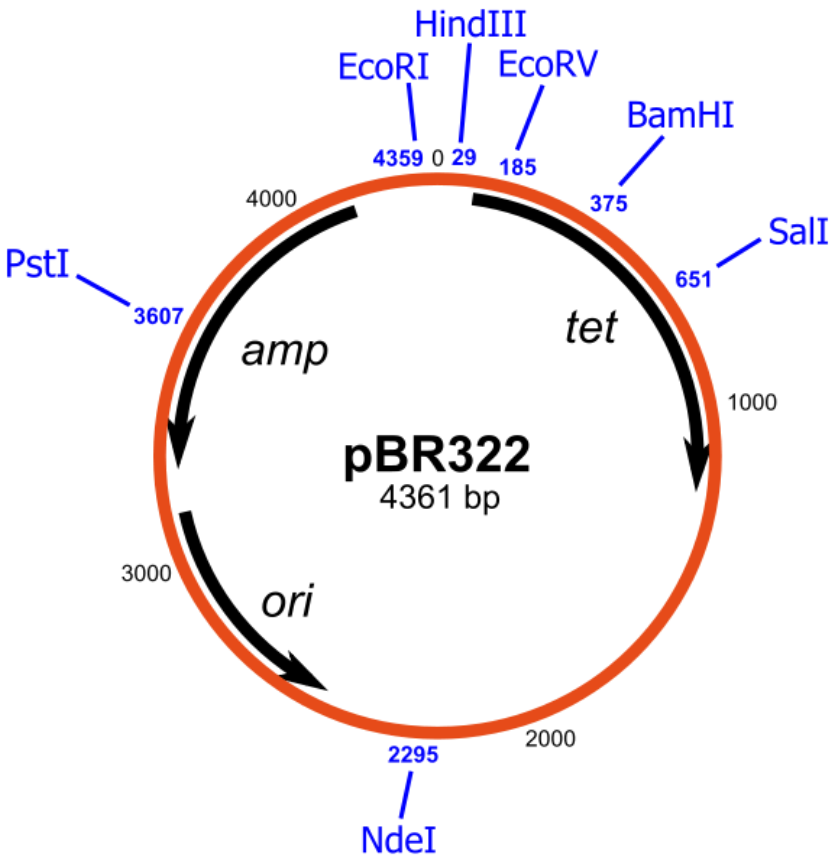


Figure 12.2 ***The bacterial plasmid vector pBR322.*** *Important DNA sequences within the plasmid include the origin of replication (ori), the selectable marker genes (amp, tet) and gene cloning sites (EcoRI, HindIII, etc.).*

Key Questions

- What is a vector?
- Identify two types of molecules that can serve as vectors.
- Describe the functions of three types of DNA sequences that allow plasmids to serve as vectors.
- What is a host cell?

Restriction Enzymes

How do we take the insulin gene from a human chromosome and insert it into a plasmid vector DNA molecule?

Restriction enzymes are important molecule tools used in the gene cloning process. Restriction enzymes are endonucleases that recognize specific DNA sequences (**restriction enzyme sites**) and cleave the phosphodiester bonds within both DNA strands. The restriction enzyme site is typically a **palindrome** DNA sequence. For example, the restriction enzyme ***Eco*RI** isolated from the bacterium *E. coli* cuts the DNA sequence 5'-GAATTC-3'. The complementary strand is 3'-CTTAAG-5', which is identical to the original restriction enzyme sequence but in the reverse orientation (i.e., a palindrome). *Eco*RI cuts both DNA strands within this palindromic DNA sequence between the G and A nitrogenous bases.

The natural function of a restriction enzyme is to protect the bacterial cell from foreign DNA, particularly bacteriophage DNA injected into the bacterial cell during a phage infection. Several hundred restriction enzymes have been isolated from bacteria and are available commercially for purchase.

Key Questions

- What is a restriction enzyme and a restriction enzyme site?
- In terms of DNA sequences, what is meant by the term sequence palindrome?

Producing Recombinant DNA Molecules

How can we use the restriction enzyme ***EcoRI*** to clone the human insulin gene into a plasmid vector DNA molecule?

Suppose that the restriction enzyme *Eco*RI recognizes the restriction enzyme sites shown below in both the insulin gene and in the plasmid DNA molecule (see **figure 12.3**) Note that the plasmid is cut by *Eco*RI at a single site, while *Eco*RI cuts on both ends of the human insulin gene (the target gene in the figure 12.3).

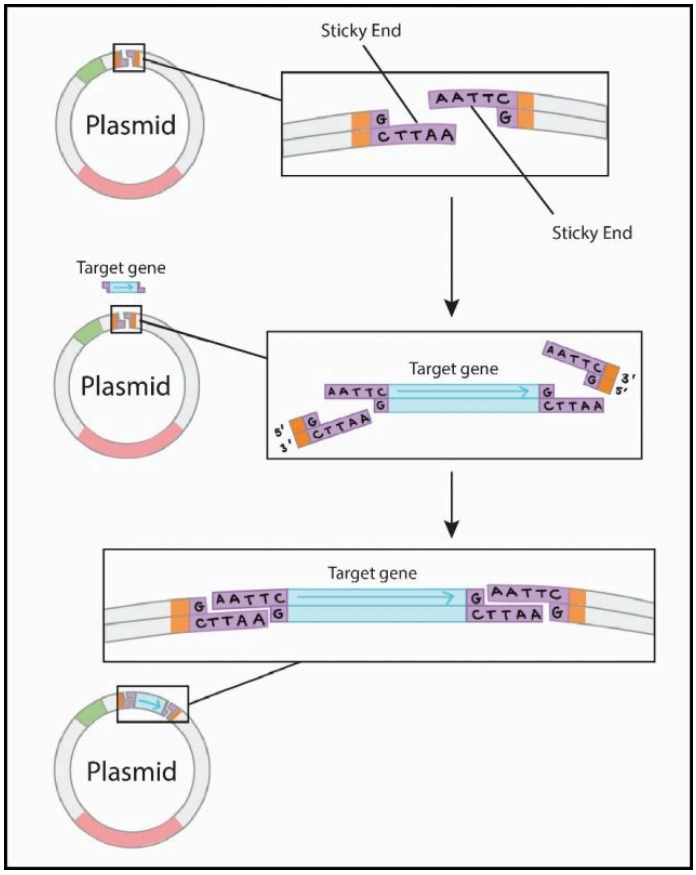


Figure 12.3 ***Cut Site and Sticky Ends*** *- Image from Khan Academy and modified by SL*

*Eco*RI cleaves the cloning site within the plasmid and the insulin gene to produce complementary single-stranded regions called **sticky ends**. When mixed, the sticky ends of the insulin DNA form hydrogen bonds with the sticky ends from the vector DNA. **DNA ligase** then forms the final covalent bonds in each DNA strand, covalently linking the insulin gene into the plasmid vector.

Key Questions

- What are sticky ends?
- What is the function of DNA ligase in a gene cloning experiment?

A Typical Gene Cloning Experiment

The entire procedure used to clone the insulin gene into a plasmid vector is outlined below (**figure 12.4**). For the purposes of this hypothetical experiment, let us assume that the plasmid DNA molecule contains a restriction enzyme site (cloning site), the *amp*^{R} gene as a selectable marker, *OriC,* and the *lacZ* gene. The function of the *lacZ* gene will be described below.

1. **The plasmid vector and the chromosome containing the human insulin gene are cut with the same restriction enzyme.** A restriction enzyme cuts the chromosomal DNA into many small pieces; one of these chromosome pieces contains the insulin gene. The same restriction enzyme is used to cut the plasmid DNA within the cloning site. The DNA fragment containing the insulin gene and the plasmid DNA molecule have complementary sticky ends.

2. **The cleaved plasmid and the insulin gene are mixed.** Three different events can occur:

    - The plasmid sticky ends hydrogen bond with each other. This situation produces an intact plasmid molecule that does not include the insulin insert.
    - Chromosomal DNA fragments that do not include the insulin gene form hydrogen bonds with the plasmid sticky ends. The recombinant DNA molecule produced will contain the wrong insert.
    - The insulin gene fragment forms hydrogen bonds with the plasmid sticky ends.

3. **DNA ligase is added.** DNA ligase catalyzes the covalent linkage of the insert DNA fragment to the plasmid DNA molecule. Some of these recombinant DNA molecules contain the desired insulin gene insert.

4. **Transformation of host *E. coli* cells**. The recombinant DNA molecules are now introduced into *E. coli* host cells for maintenance. When *E. coli* cells are treated with calcium, the bacteria become **competent** to take up DNA from the environment. When the recombinant DNA molecule is added to these competent bacterial cells, and the bacteria are shocked by a brief heat treatment, the recombinant DNA molecule is taken into the cytoplasm of the *E. coli* (**transformation**).

5. **Host bacteria are grown on agar plates that contains ampicillin**. Two scenarios are possible after transformation:

    - **Some *E. coli* cells were not transformed** (i.e., do not contain a recombinant DNA molecule). These bacterial cells cannot grow on agar plates containing ampicillin because the bacteria lack the amp^R gene.
    - **Some *E. coli* cells were transformed**. Since the plasmid vector contains the amp^R gene, transformed cells are now resistant to ampicillin. As a result, these transformed cells grow on agar plates that contain ampicillin. It is important to note that this population of growing bacteria contains three types of plasmids:
        - Some plasmids lack inserts altogether.
        - Some plasmids contain other chromosomal DNA fragments as the insert (i.e., the wrong insert).
        - Some plasmids contain the insulin gene as the insert (i.e., the correct insert).

6. **Identify colonies that contain an insert by the blue-white screening method**. Many cloning experiments are designed so that the restriction ezyme cutting site within the plasmid is located the ***lacZ*** DNA sequence. The *lacZ* gene produces the enzyme β-galactosidase. Cloning the insert disrupts the *lacZ* gene, preventing β-galactosidase production. Disruption of the *lacZ* gene allows researchers to distinguish bacteria that contain an insert versus bacteria that do not contain an insert.

    - **Recombinant plasmid vector without an insert**. In this case, the plasmid contains an intact *lacZ* gene. The plasmid *lacZ* gene produces β-galactosidase.
    - **Recombinant plasmid vector that contains an insert**. Since the presence of an insert disrupts the *lacZ* gene, no β-galactosidase is produced.

How do we determine if β-galactosidase is produced? This is done by plating the bacterial cells on an agar plate that not only includes ampicillin but also **IPTG**, a chemical that activates the *lacZ* gene to produce the β-galactosidase protein. The agar plate also contains **X-Gal**, a chemical substrate for the β-galactosidase enzyme.

- **β-galactosidase is expressed (no insert)**. The bacterial colony is blue as β-galactosidase converts X-Gal, which is colorless, into a blue product.
- **β-galactosidase is not expressed (insert present)**. The colony is white because β-galactosidase is nonfunctional and thus cannot convert X-Gal into a blue product.

7. **Identify colonies that contain the insulin insert**. All the white colonies contain an insert; however, at this point, we cannot distinguish the white colonies that contain the insulin insert from the white colonies that have other DNA fragments as inserts. Several techniques can be used to identify which bacterial cells contain a recombinant vector with an insulin gene. For example, vectors isolated from white colonies can be cut with the same restriction enzyme used at the beginning of the cloning experiment to release the insert. The digested DNA can then be analyzed by agarose gel electrophoresis. The presence of a DNA fragment of the appropriate size for the insulin gene indicates that the chosen colony likely contains the insulin insert. Alternatively, the polymerase chain reaction (PCR), using primers specific for the insulin gene can be used to amplify the insulin insert. If a PCR product is produced using these insulin primers, then the insulin gene was cloned successfully. Finally, determining the nucleotide sequence of the cloned insert by DNA sequencing will determine if the recombinant molecule contains the insulin gene.

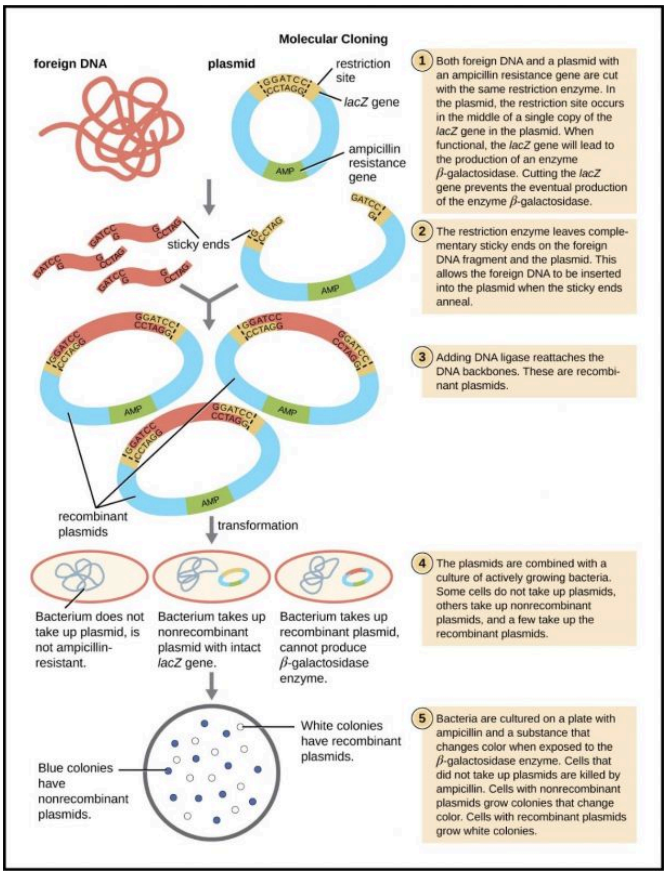


Figure 12.4 ***Gene cloning overview ---*** *This image is used from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction*

Key Questions

- Describe the seven steps involved in cloning the insulin gene into a bacterial plasmid.
- How are bacterial cells that were not transformed eliminated in the cloning experiment?
- How can blue-white screening be used to identify recombinant DNA molecules that contain an insert?
- How can you identify the recombinant DNA molecule that contains the insulin insert?

13 - The lac Operon

In Part 13, we will begin learning about the mechanisms that regulate **gene expression** in bacteria. Gene expression refers to processes that activate structural genes, producing a mRNA molecule by transcription and a functional protein product by translation. Specifically, we will study the expression of the ***lac* operon** system in the bacterium *E. coli*. The *lac* operon contains the structural genes that produce protein products that function to metabolize lactose for energy production.

Some bacterial genes are always transcribed. These genes that are always expressed are called **constitutive** or **housekeeping genes**. Note that constitutive genes produce constitutive or housekeeping proteins. Housekeeping proteins are required for the normal functioning of the bacterial cell, the so-called housekeeping functions. The genes that produce the proteins involved in glycolysis are example housekeeping genes.

Regulated genes change expression under different environmental conditions. In one environment, the regulated gene is transcribed, while in another environment the regulated gene is silenced. The mRNAs produced from regulated genes are translated to make **inducible proteins**. Inducible proteins are tightly controlled so that thousands of copies of the protein may be present in certain environments, while only a few or no copies of the protein are produced in other environments. Regulated genes and their protein products are advantageous because they allow bacteria to adapt to changing environments, competing for available resources, such as carbon or nitrogen.

Key Questions

- What is the difference between a constitutive and a regulated gene?
- What is one metabolic process that is always occurring in a bacterial cell (i.e., involves housekeeping genes)?

Inducible Genes

Gene regulation in bacteria often involves controlling the initiation of transcription. Transcriptional regulation requires the binding of **regulatory transcription factor proteins** to **regulatory DNA sequences** near the promoter region of a gene. These regulatory transcription factor proteins function to either enhance or inhibit sigma (σ) factor protein and RNA polymerase core enzyme binding to the promoter. Regulatory transcription factor proteins include:

- **Repressor proteins**. Repressor proteins decrease how often transcription starts (**negative control**).
- **Activator proteins.** Activator proteins increase how often transcription starts (**positive control**).

Repressor and activator proteins contain DNA binding domains and also have binding sites for small organic molecules (sugars, amino acids, or nucleotides) called **effectors**. When an effector molecule binds, the three-dimensional structure of the repressor or activator protein changes. This change in protein shape influences the ability of the activator protein or repressor protein to bind to the DNA.

How do bacteria turn a regulated gene from an off state to an on state? For example, how does a bacterium produce the enzymes necessary to metabolize the sugar lactose when lactose becomes available in the environment? An **inducer** effector molecule causes transcription to increase (**figure 13.1**). Inducers can function in two different ways:

- **The inducer binds to a repressor protein**. When the inducer molecule binds to the repressor protein, the repressor protein is released from a binding site on the DNA, and transcription of the nearby structural gene increases.
- **The inducer binds to an activator protein**. In this case, the activator protein cannot bind to the DNA unless the inducer is present. When the inducer molecule binds to the activator protein, the activator can bind to the DNA and transcription of the nearby structural gene increases.

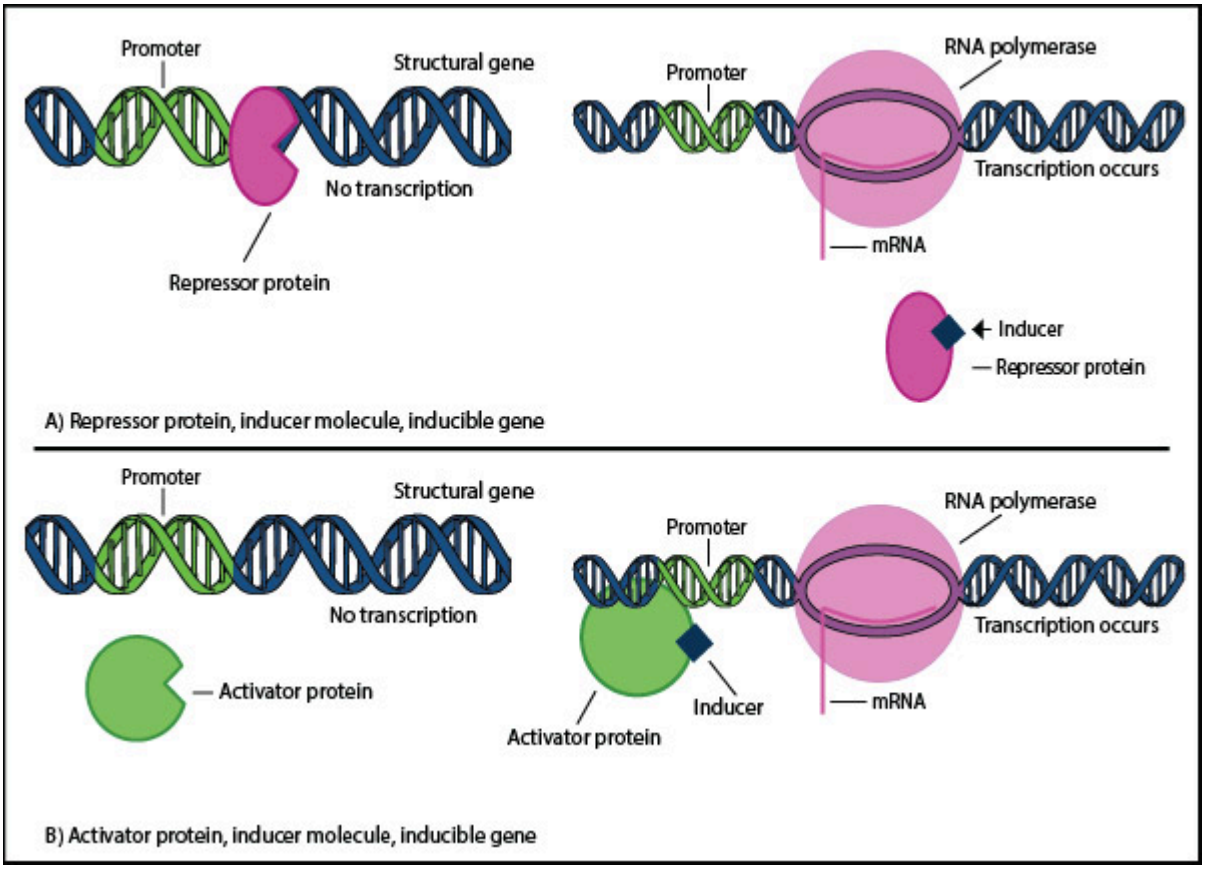


Figure 13.1 ***Inducing a Gene*** *--- Image created by SL*

Key Questions

- How do repressor and activator proteins affect transcription?
- What is an effector molecule?
- Describe two ways that an inducer can activate the transcription of a gene.

Repressible Genes

How do bacteria turn a regulated gene from an on state to an off state? For example, how does a bacterial cell stop producing the enzymes required to make the amino acid tryptophan, when there is plenty of tryptophan in the environment? The presence of effector molecules inhibits transcription in two ways (**figure 13.2**):

- **A corepressor effector molecule binds to a repressor protein.** Without the corepressor, the repressor protein does not bind to the DNA. When the corepressor molecule binds the repressor protein, a conformational change occurs in the repressor. The repressor protein can then bind to the DNA and inhibit transcription of a nearby structural gene.
- **An inhibitor effector molecule binds to an activator protein.** In this case, the activator protein is normally bound to the DNA and activates transcription. When the inhibitor molecule binds to the activator protein, a conformational change causes the activator protein to be released from the DNA, and transcription of a nearby structural gene ceases.

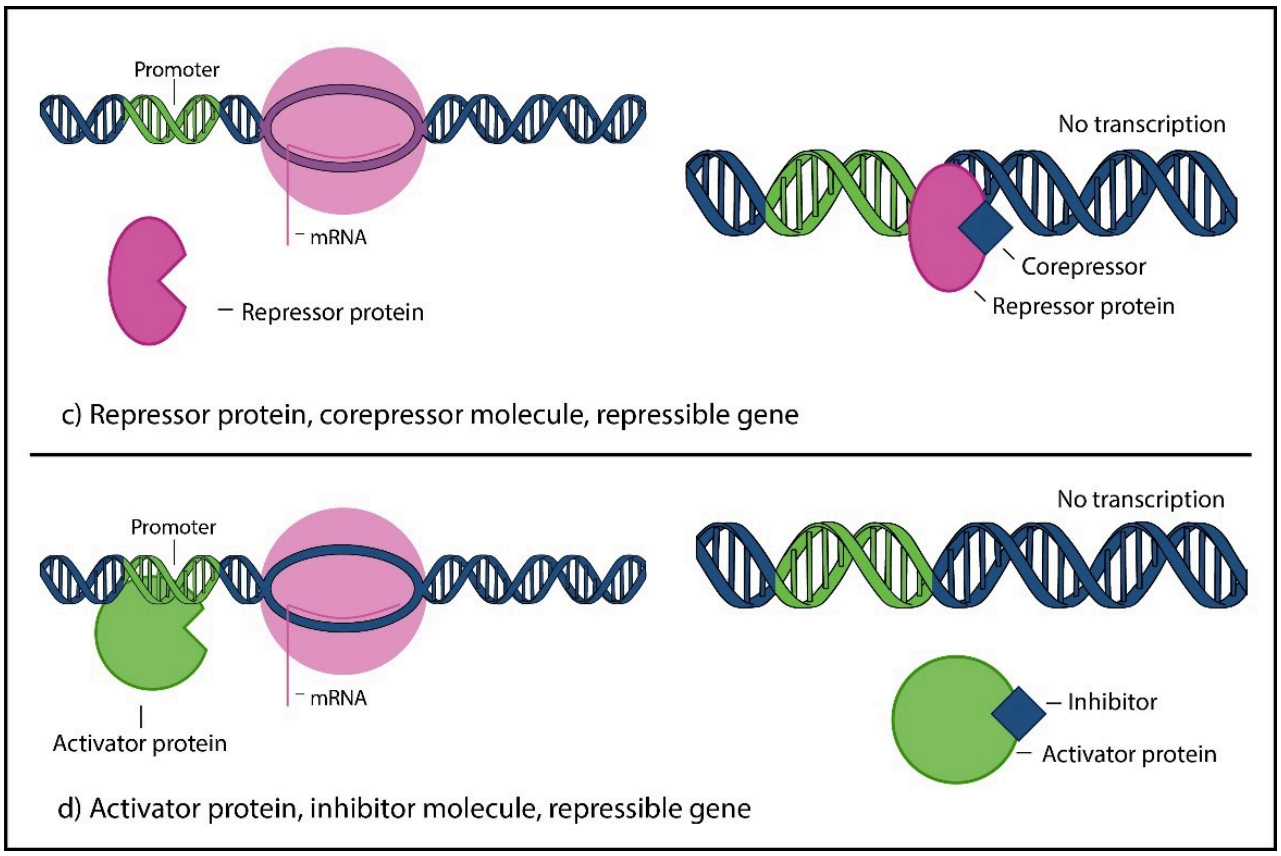


Figure 13.2 ***Repressing a Gene*** *--- Image created by SL*

Key Questions

- Describe how a corepressor and an inhibitor can turn a transcribed gene to an off state.

Enzymes Involved in Lactose Metabolism in *E. coli*

Now we will turn our attention to a specific example of gene regulation in the bacterium *E. coli,* involving the regulation of the structural genes involved in lactose metabolism. Lactose is a sugar that can be used as a carbon and energy source for the bacterium *E. coli* when the preferred carbon and energy source, glucose, is limited. Lactose breakdown by an *E. coli* cell involves three enzymes (**figure 13.3**):

- **Lactose permease.** Lactose permease is a cytoplasmic membrane protein involved in the transport of lactose from the environment into the cytoplasm of the *E. coli* cell.
- **Beta (β)-galactosidase.** β-galactosidase cleaves the lactose imported by lactose permease, producing the monosaccharides galactose and glucose. Galactose and glucose can then be metabolized by the *E. coli* cell to produce energy. β-galactosidase also catalyzes a side reaction that converts lactose into the effector molecule **allolactose**. Importantly, allolactose is one of the two inducers for the *lac* operon; allolactose binds to the *lac* repressor protein and releases the repressor protein from the DNA (see below).
- **Galactoside transacetylase.** Galactoside transacetylase converts atypical forms of lactose into forms that can be metabolized readily by β-galactosidase.

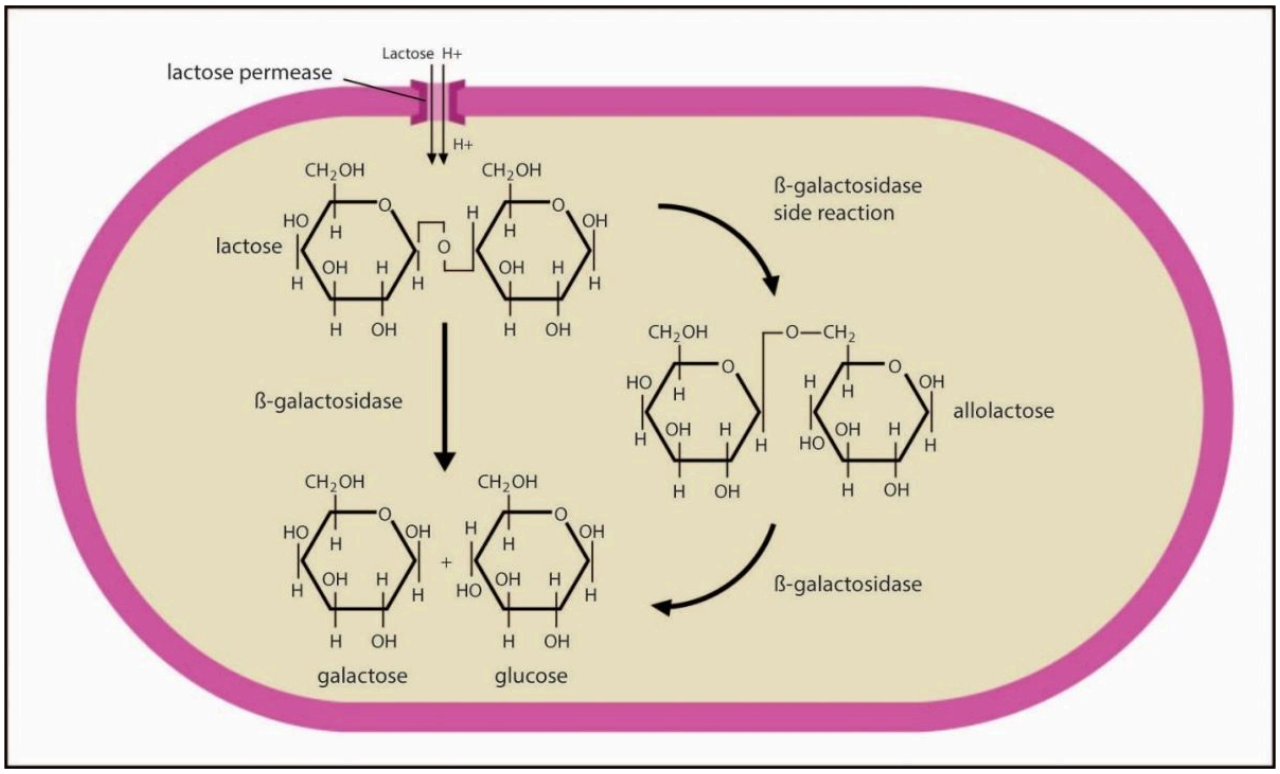


Figure 13.3 ***The E. coli enzymes involved in lactose metabolism.*** *Lactose permease moves lactose across the cytoplasmic membrane into the cytoplasm of the E. coli cell. β-galactosidase converts lactose into glucose and galactose. β-galactosidase also forms the inducer allolactose. The enzyme galactoside transacetylase is not shown in the figure. --- Image created by SL*

Key Questions

- What are the functions of the three bacterial enzymes involved in lactose breakdown?

Operons

In bacteria, a group of structural genes can be under the control of a single group of regulatory DNA sequences, a single promoter sequence, and a single terminator sequence. This grouping of structural genes is an **operon** (**figure 13.4**). The organization of structural genes into operons allows all of proteins involved in a single biochemical pathway (e.g., lactose metabolism) to be regulated in a coordinated way. When an operon is transcribed, a **polycistronic mRNA** is produced that contains the coding regions for multiple individual proteins.

Typical operons contain a **promoter**. Recall that the promoter serves as the binding site for the sigma (σ) factor protein and contains the transcription start site (+1 site) for the operon. Operons also contain an **operator** DNA sequence that serves as a repressor protein binding site, an **activator binding site** where an activator protein binds, **structural genes** that encode proteins, and a **terminator** sequence that signals the end of transcription. Recall that transcriptional terminators in bacteria work either using the rho (ρ)-dependent or rho (ρ)-independent mechanism.

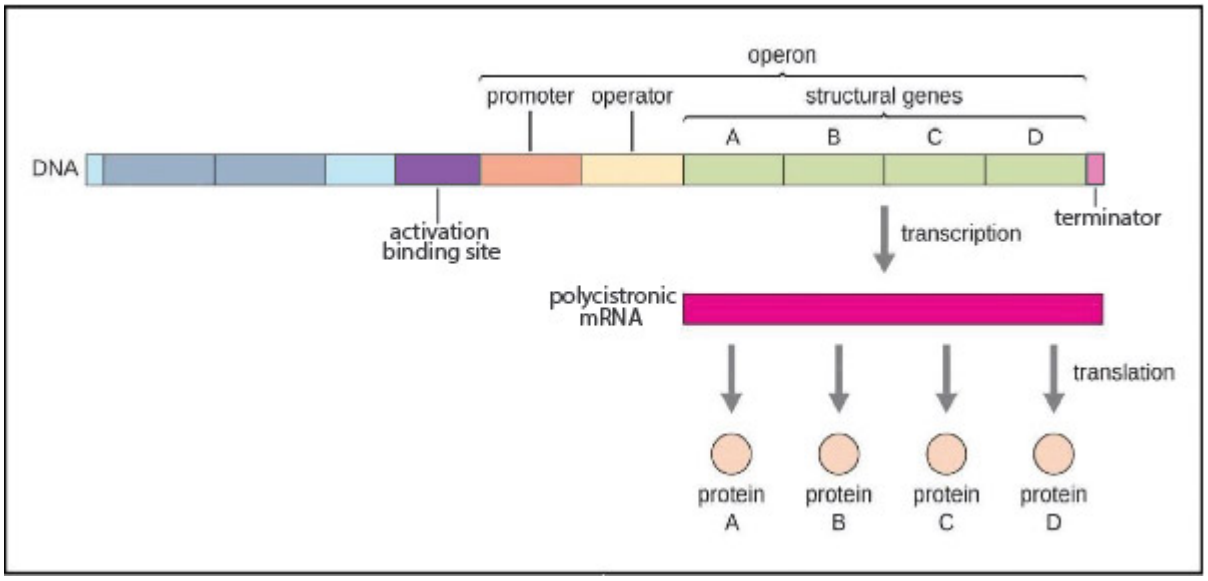


Figure 13.4 ***Operon Structure*** *--- This image is used from OpenStax (access for free at [https://openstax.org/books/biology-2e/pages/1-introduction)](https://openstax.org/books/biology-2e/pages/1-introduction)*

Key Questions

- What is an operon?
- Why is it advantageous to organize structural genes into operons?
- What is a polycistronic mRNA?
- What is an operator and an activator binding site?

The Lactose (*lac*) Operon

François Jacob and Jacques Monod first described transcriptional regulation by studying lactose metabolism in *E. coli*. Jacob and Monod won the Nobel Prize in 1965 for their work. Lactose metabolism in the bacterium *E. coli* requires regulating genes within the **lactose (*lac*) operon**. The *lac* operon contains the following DNA sequences and structural genes (**figure 13.5**):

- **CAP site.** The CAP site is a DNA sequence that serves as the binding site for the **catabolite activator protein (CAP)**.
- ***lac* promoter (*lacP*).** The *lacP* DNA sequence contains the -35 sequence, the -10 sequence, and the +1 site. *lacP* determines where transcription of the *lac* operon will begin, serving as the binding site for the sigma (σ) factor protein. Recall that sigma (σ) factor directs the RNA polymerase core enzyme to the +1 site.
- **Operator site (*lacO*).** *lacO* is the binding site for the *lac* repressor protein.
- ***lacZ.*** *lacZ* is the structural gene that encodes the enzyme **β-galactosidase**.
- ***lacY.*** *lacY* is the structural gene that encodes the enzyme **lactose permease**.
- ***lacA.*** *lacA* is the structural gene that encodes the enzyme **galactoside transacetylase**.
- ***lac* terminator.** The *lac* terminator is the DNA sequence involved in terminating transcription of the *lac* operon. The *lac* operon is terminated by the rho (ρ)-dependent mechanism.

Upstream of the *lac* operon is another structural gene, called ***lacI***, that contains its own promoter and terminator. The *lacI* gene encodes the ***lac* repressor** protein. The *lac* repressor protein binds to the *lacO* sequence and turns off the expression of the *lac* operon (in other words, the *lac* operon displays **negative control** via the *lac* repressor). The *lacI* gene is a constitutive (housekeeping) gene and is therefore always transcribed.

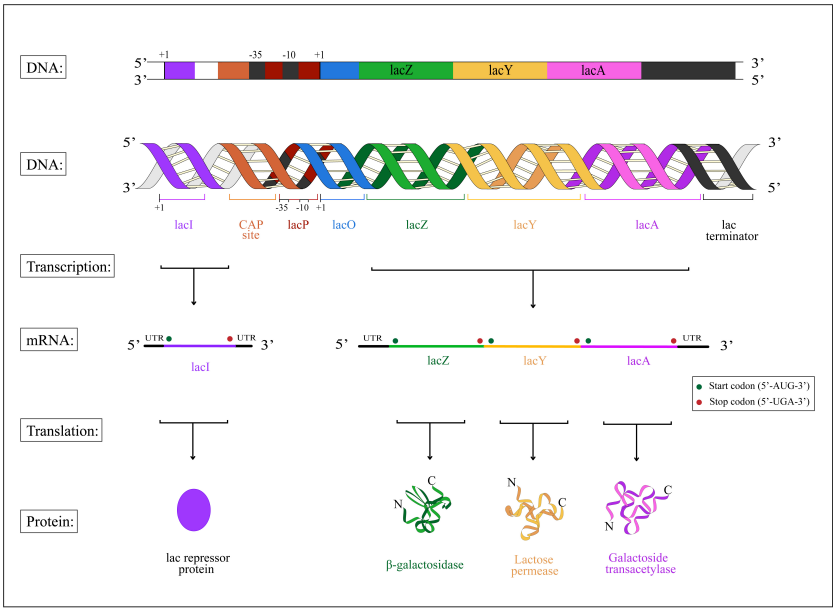


Figure 13.5 ***Lac operon structure*** *--- Image created by Alex Baff.*

Key Questions

- What are names of the three structural genes of the *lac* operon?
- What are the names and functions of the four regulatory DNA sequences within the *lac* operon?
- What is the function of the *lacI* gene?

lac Operon Expression

In the absence of lactose, **repression** of the *lac* operon occurs as follows (**figure 13.6** and **13.7**):

1. *lacI* is a constitutive gene, meaning that it is always transcribed. Transcription of the *lacI* gene produces a *lacI* mRNA that is then translated to produce the *lac* repressor protein.
2. The *lac* repressor protein binds to the operator (*lacO*) DNA sequence.
3. Sigma (σ) factor and the RNA polymerase core enzyme do not bind efficiently to *lacP* when the *lac* repressor is bound to *lacO*. As a result, the three structural genes (*lacZ*, *lacY*, and *lacA*) of the *lac* operon are transcribed at a low level, producing only a few copies of the β-galactosidase, lactose permease, and galactoside transacetylase proteins per cell.

When lactose becomes available in the environment, the *lac* operon is **induced** as follows:

1. The few copies of lactose permease expressed by *E. coli* move lactose across the cytoplasmic membrane into the cytoplasm of the cell.
2. The few copies of β-galactosidase present convert lactose into **allolactose**.
3. Allolactose binds to the *lac* repressor protein.
4. The conformation of the *lac* repressor protein changes.
5. The *lac* repressor protein is released from the *lacO* DNA sequence.
6. Sigma (σ) factor and the RNA polymerase core enzyme bind to *lacP* efficiently when the *lac* repressor protein is released from the DNA.
7. The structural genes of the *lac* operon (*lacZ*, *lacY*, and *lacA*) are actively transcribed to produce polycistronic mRNAs. The polycistronic mRNAs are then translated to produce thousands of copies each of the β-galactosidase, lactose permease, and galactoside transacetylase proteins.
8. Lactose is metabolized by the *E. coli* cell.

Note that when lactose is no longer present in the environment, the *lac* operon resets. Allolactose is released from the *lac* repressor protein, causing the *lac* repressor to bind once again to *lacO*. The excess β-galactosidase, lactose permease, and galactoside transacetylase proteins in the cytoplasm are eventually degraded.

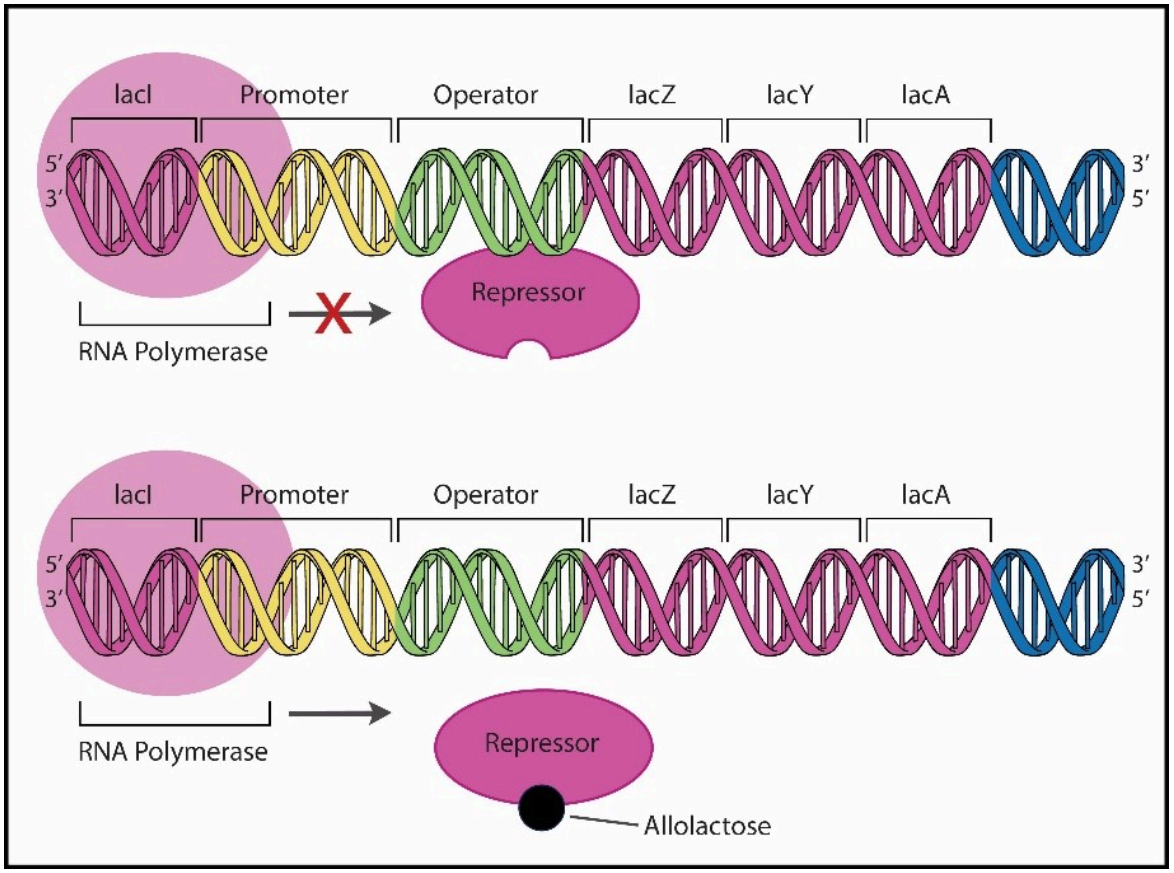


Figure 13.6 ***Lac Operon in the Absence/Presence of Lactose*** *--- Image is borrowed from OpenStax (access for free at https://openstax.org/books/biology-2e/pages/1-introduction) modified by SL*

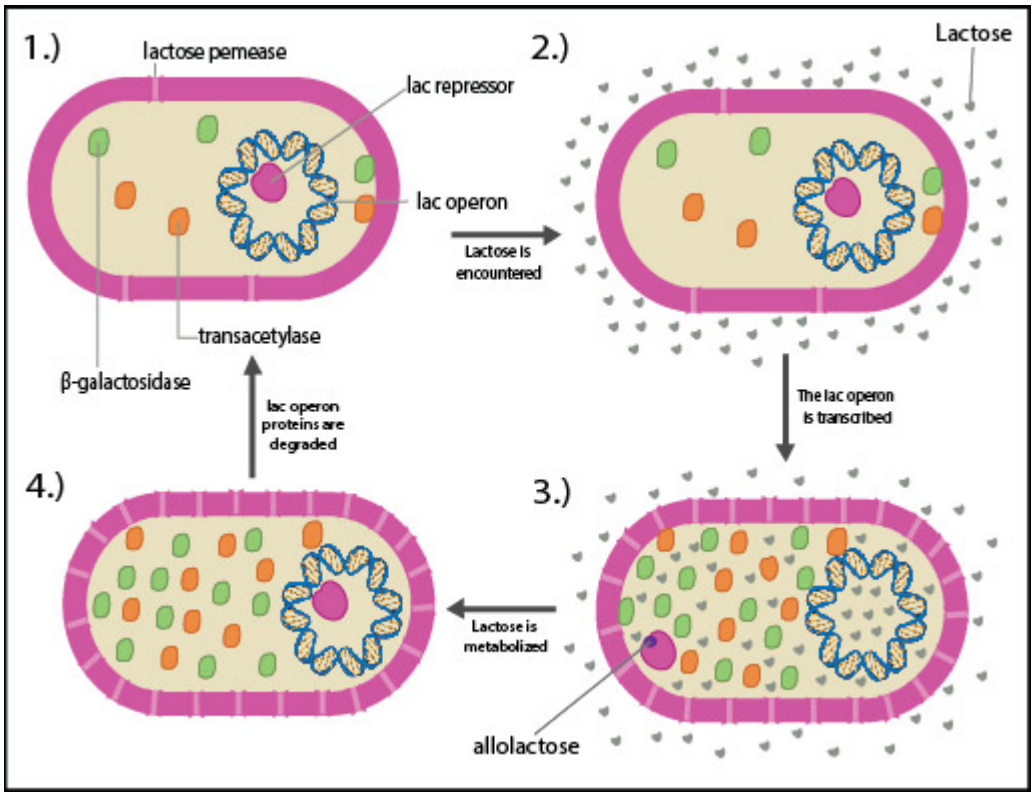


Figure 13.7 ***Summary of lac Operon Induction*** *--- Image created by SL*

Key Questions

- Describe the steps involved in inducing the *lac operon.*

The *lacI*$^-$ Strain Displays Constitutive Expression of the *lac* Operon

To gain an appreciation of the genetics involved in *lac* operon regulation, let us turn our attention to the experiment that determined the function of the *lacI* gene product, which we now know makes the *lac* repressor protein. When François Jacob and Jacques Monod first started studying lactose metabolism, they identified a mutant strain of *E. coli* that they named *lacI*$^-$. In this *lacI*$^-$ mutant strain, the enzymes involved in lactose metabolism were always produced, even in the absence of lactose. Thus, the *lacI*$^-$ mutation is a **constitutive mutation** producing constitutive expression of the *lac* operon. How could this phenotype be explained?

Jacob and Monod reasoned that the *lacI*$^-$ constitutive phenotype could be explained in two ways:

- The *lacI*$^-$ mutation produces a defective activator protein that activates transcription in the presence and absence of lactose (constitutive activator protein hypothesis).
- The *lacI*$^-$ mutation produces a defective repressor protein that fails to inhibit transcription when lactose is absent (defective repressor protein hypothesis).

Mutant and Merozygote Strains of *E. coli*

To distinguish between the two hypotheses indicated above, Jacob and Monod examined two strains of *E. coli*. Jacob and Monod studied the *lacI*$^-$ strain described earlier, and they studied an unusual strain of *E. coli* called a **merozygote**, or partial diploid.

Recall that bacteria typically have a single circular chromosome; however, bacteria also contain small circular **plasmid** DNA molecules in addition to the chromosome. These plasmids are commonly adapted for use in gene cloning experiments (see Part 12). A common type of plasmid is the **F plasmid** that functions in bacterial fertility (i.e., DNA transfer between bacteria). The merozygote strain that Jacob and Monod examined in their experiments contained a modified F plasmid (**F' plasmid**), which contained a *lacI* gene and the *lac* operon. Thus, *E. coli* cells that contain an F' plasmid are merozygotes (partial diploids), containing a copy of *lacI* and the *lac* operon genes on both the chromosome and on the F' plasmid.

The merozygote strain used by Jacob and Monod contained *lacI*$^-$ on the chromosome and a wild-type copy of the gene (*lacI*$^+$) on the F' plasmid; this *E. coli* strain was in essence a *lacI*$^+$/*lacI*$^-$ heterozygote. The other DNA sequences within the *lac* operon (*lacP*, *lacO*, *lacZ*, *lacY*, and *lacA*) were wild-type and were found on both the chromosome and the F' plasmid. Thus, the *E. coli* merozygote strain was homozygous for *lacP*, *lacO*, *lacZ*, *lacY*, and *lacA*.

Key Questions

- What is a merozygote?
- Why might it be advantageous for an *E. coli* cell to have two copies of each *lac* operon gene?

The Jacob and Monod Experiment

The Jacob and Monod experiment compared the *lacI* $^{-}$ *E. coli* strain (*lac* operon is always expressed) to the *lacI* $^{+}$*/lacI* $^{-}$ merozygote *E. coli* strain. The experiment was done as follows:

1. The mutant (*lacI* $^{-}$) and the merozygote (*lacI* $^{+}$*/lacI* $^{-}$) strains were grown in separate flasks.

2. Each bacterial culture was then split into two smaller flasks, a control flask and an experimental flask. For example:

- Mutant strain (*lacI* $^{-}$)
  - Control (flask 1)
  - Experimental (flask 2)
- Merozygote strain (*lacI* $^{+}$*/lacI* $^{-}$)
  - Control (flask 3)
  - Experimental (flask 4)

3. Lactose was added to the experimental reactions (flasks 2 and 4).

4. The bacterial cultures were incubated to allow transcription of the *lac* operon and translation of the β-galactosidase, lactose permease, and galactoside transacetylase proteins.

5. The bacterial cells in each flask were then lysed to release the β- galactosidase, lactose permease, and galactoside transacetylase proteins.

6. The β-galactosidase levels in each of the four bacterial cell lysates were measured. β-galactosidase can convert the chemical β-O-nitrophenylgalactoside (β-ONPG), which is colorless, into galactose and O-nitrophenol, which is yellow. Note that if a yellow product is formed, β-galactosidase is present (i.e. the *lac* operon was transcribed).

7. The O-nitrophenol (yellow product) levels in each lysate were measured using a spectrophotometer.

Key Questions

- Explain the purpose of flasks 1–4 in the Jacob and Monod experiment.
- Explain what is occurring when a reaction turns yellow.

lacI $^{+}$ Produces a Diffusible Repressor Protein

When Jacob and Monod did their experiments, yellow color was observed in flasks 1 and 2. Thus, in the *lacI* $^{-}$ strain, β-galactosidase is produced in the absence and in the presence of lactose (i.e., expressed constitutively).

In the merozygote strain, no yellow color was produced in the absence of lactose (flask 3); however, two times the yellow product was produced in the presence of lactose (flask 4). This means that β-galactosidase is not produced when lactose is absent because the *lacI* $^{+}$ gene on the F' plasmid produces a protein (i.e., the *lac* repressor protein) that binds to both the chromosomal and F' plasmid copies of *lacO* in the cell, and thus inhibits expression of both genes that make β-galactosidase. Remember that bacteria do not have a nuclear membrane, so both the host chromosome and the F' plasmid are found in the cytoplasm. This *lac* repressor protein diffuses throughout the

cytoplasm of the cell and can bind to **any** *lacO* sequence. Because the *lac* repressor can bind to any operator in the cell, the *lac* repressor is said to be an example of a ***trans*-acting factor**.

When lactose is present, lactose is converted to allolactose, and allolactose releases the *lac* repressor proteins from both copies of *lacO*. The *lac* operons on both the chromosome and on the F' plasmid are now expressed (flask 4). The expression of two copies of the *lacZ* gene produces two times as much β-galactosidase protein, leading to the production of two times the yellow color in flask 4.

This experimental result provided supporting evidence for the defective repressor hypothesis. It is worth noting that in the case of the *lacI*$^-$ constitutive activator hypothesis, the *lac* operon would have been expressed by the merozygote strain in both the absence (flask 3) and in the presence of lactose (flask 4). Thus, both flasks 3 and 4 should have produced two times the yellow color at the conclusion of the experiment.

Key Questions

- Explain how the flask 3 result showed that the *lac* repressor protein is an example of a *trans*-acting factor.

Glucose and the *lac* Operon

The *lac* operon can also be regulated by glucose. Glucose is the preferred carbon and energy source used by *E. coli,* so the genes involved in glucose breakdown (**catabolism**) are expressed constitutively (always transcribed). Thus, in the presence of glucose, the *lac* operon is not needed, so transcription of the *lac* operon is turned off (so-called **catabolite repression**). When glucose levels decrease and lactose is present, this catabolite repression is alleviated, and the *lac* operon is transcribed. Lactose is then used by the *E. coli* cell as the carbon and energy source. The sequential use of sugars—first glucose, followed by lactose—is called **diauxic growth** (**figure 13.8**).

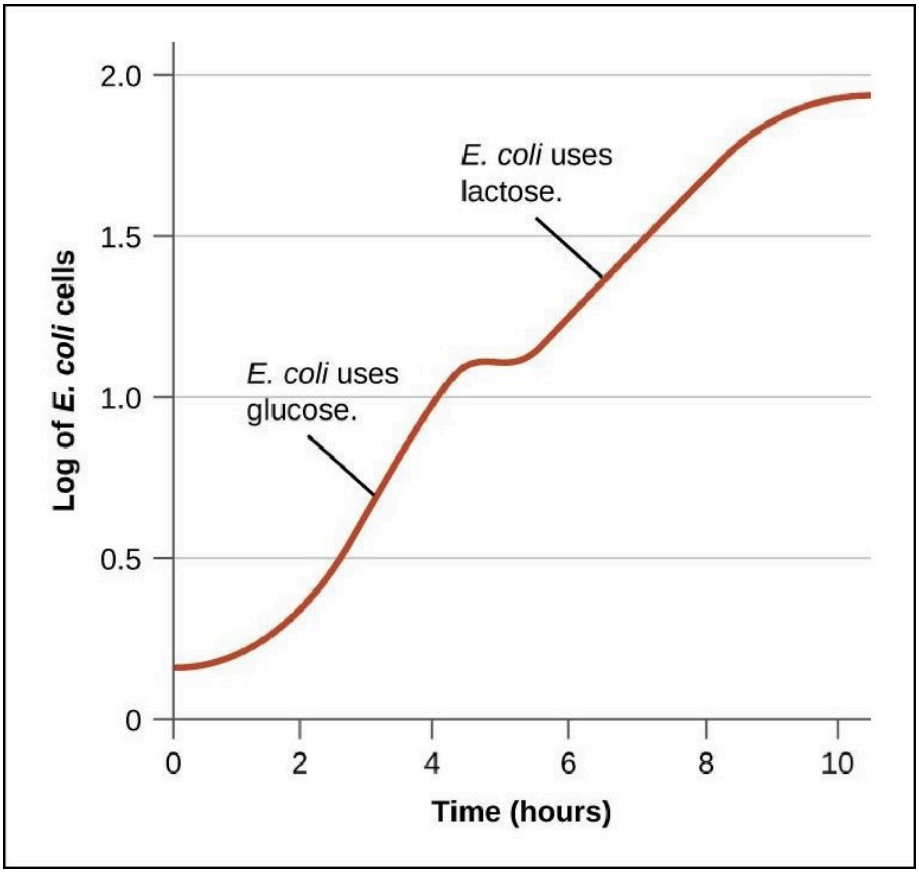


Figure 13.8 Diauxic growth of E. coli. *E. coli uses glucose first, then switches to using lactose. --- Image created by SL*

How is the *lac* operon repressed (turned off) by glucose? Glucose repression of the *lac* operon involves a(n):

- **Regulatory transcription factor protein**. The regulatory transcription factor protein that is responsive to glucose levels is the **catabolite activator protein (CAP).** The CAP protein binds to the CAP binding site in the *lac* operon, which is located immediately upstream of the *lac* operon promoter (*lacP*).
- **Effector molecule**. It would be reasonable to assume that glucose is the effector molecule involved in catabolite repression. However, the effector involved in glucose regulation is **cyclic AMP (cAMP)** (**figure 13.9**).

cAMP is produced from ATP by the enzyme **adenylyl cyclase**. When glucose is present in the environment, adenylyl cyclase activity is inhibited, and cellular cAMP levels are low. When glucose levels in the environment are low, adenylyl cyclase activity increases, resulting in higher levels of cAMP in the cell.

When the CAP protein binds to cAMP, the CAP protein changes conformation (shape) and can then bind to the CAP site in the DNA. As a result, *lac* operon transcription is activated. In fact, for sigma (σ) factor and the RNA polymerase core enzyme to bind efficiently to the *lac* promoter and transcribe the *lac* operon, CAP must be bound to the CAP site.

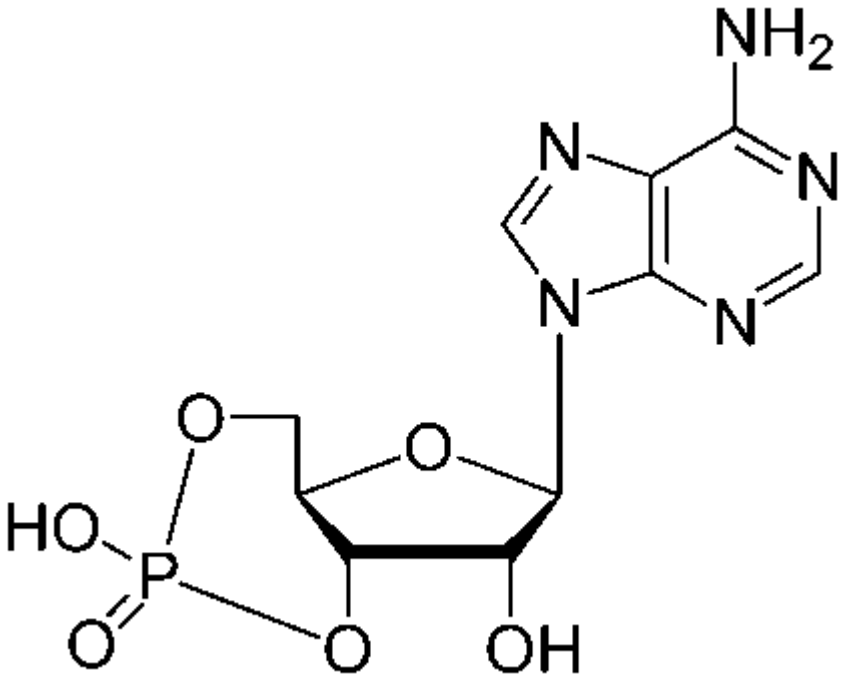


Figure 13.9 ***Cyclic AMP (cAMP) structure***

Key Questions

- Describe how the *lac* operon responds when glucose levels are high or low (assuming that lactose is present).
- What is the function of adenylyl cyclase?
- Explain the relationship between cAMP levels and CAP function.
- Describe the relationship between CAP binding to the CAP site and the transcription of the *lac operon.*

Regulation of the *lac* Operon by the *lac* Repressor and CAP

The interaction between a positive regulatory signal (CAP) and a negative regulatory signal (the *lac* repressor protein) makes transcriptional regulation of the *lac* operon more complicated. What happens to *lac* operon expression when an *E. coli* cell encounters the following environmental conditions (**figure 13.10**)?

- **No glucose or lactose in the environment.** In this environment, cAMP levels are high, and the cAMP is bound to the CAP protein. The cAMP:CAP complex binds to the CAP site on the DNA and tries to promote transcription. However, in the absence of lactose, there is no allolactose, so the *lac* repressor is bound to the *lacO* DNA sequence, preventing transcription. **In this environment, the lac operon is not transcribed efficiently.**
- **Glucose is present in the environment; however, there is no lactose.** In this case, cAMP levels are low. At low cAMP levels, the CAP protein does not bind to the CAP site on the DNA. The absence of bound CAP protein is a negative signal that inhibits transcription. Since there is no lactose, there is no allolactose, so the *lac* repressor is bound to the *lacO* DNA sequence. **In this environment, the *lac* operon is not transcribed efficiently.**
- **Glucose and lactose are present in the environment.** In the presence of glucose, cAMP levels are low. Thus, cAMP is not bound to the CAP protein, and CAP does not bind to the CAP site on the DNA. The absence of bound CAP protein is a negative signal that inhibits transcription. In the presence of lactose, allolactose binds to the *lac* repressor, releasing it from the operator. The release of the *lac* repressor tends to promote transcription. **In this environment, the inability of the CAP protein to bind to the CAP site results in inefficient *lac* operon transcription**.
- **Glucose is absent but lactose is present in the environment.** In this case, cAMP levels are high. cAMP binding to CAP protein changes the conformation of CAP, allowing the CAP protein to bind to the CAP site on the DNA. This serves as a positive signal that tends to promote transcription. When lactose is present, allolactose is present. The *lac* repressor protein binds to allolactose and therefore is released from the operator DNA sequence. **In this environment, the *lac* operon is transcribed.**

In summary, there is only one way the lac operon is transcribed efficiently: glucose must be absent from the environment, and lactose must be present.

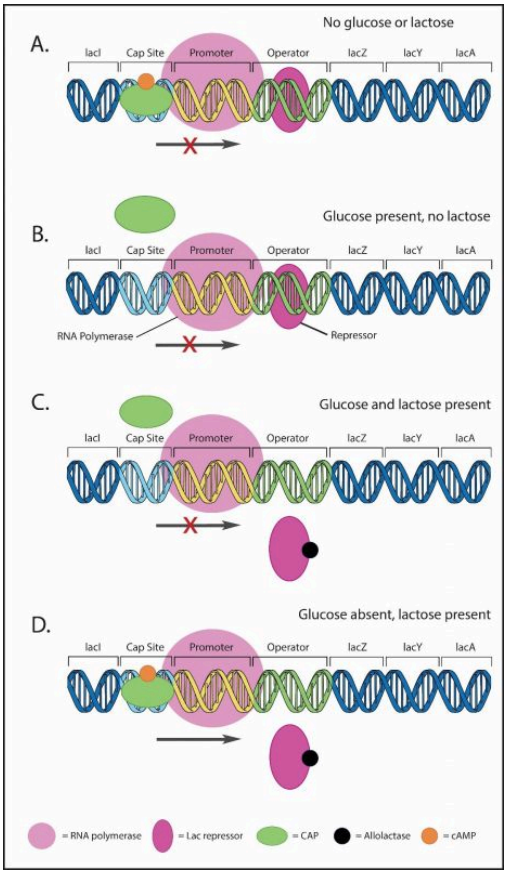


Figure 13.10 ***Lac Operon Expression Under All Cellular Conditions*** *--- Image created by SL*

Key Questions

- Why is the *lac* operon **off** if neither glucose nor lactose are present in the cell?
- Why is the *lac* operon **off** if glucose is present and lactose is absent in the cell?
- Why is the *lac* operon **off** if both glucose and lactose are present in the cell?
- Why is the *lac* operon **on** if glucose is absent and lactose is present in the cell?

Other Gene Regulation Mechanisms in Bacteria

The *lac* operon is an example of how regulation of transcription initiation can control gene expression. There are other ways to control the expression of a gene in bacteria, including attenuating transcription, regulating translation, and regulating the protein product produced by a structural gene following translation (**posttranslational regulation**).

- **Regulation of transcription**.
    - **Regulating transcription initiation**. Controlling how often transcription starts involves regulating sigma (σ) factor (and the RNA polymerase core enzyme) binding to the promoter. Regulatory transcription factors (activator and repressor proteins) promote or inhibit sigma (σ) factor binding. The *lac* operon is an example of this type of gene regulation.
    - **Attenuation of transcription**. Attenuation involves activating transcription to begin producing a mRNA molecule; however, transcription is terminated prematurely before the entire mRNA is made. Many bacterial operons are regulated by attenuation.
- **Regulation of translation**.
    - **Translation repressor proteins.** Translation repressor proteins prevent the initiation of translation. These translation repressor proteins bind to the Shine-Dalgarno sequence on the mRNA, preventing the 16S rRNA component of the ribosome from binding to the mRNA.
    - **Antisense RNA.** Antisense RNA molecules are produced by *E. coli* cells to form hydrogen bonds with the Shine-Dalgarno sequence of a particular mRNA, preventing the 16S rRNA component of the ribosome from binding to the mRNA. Antisense RNA molecules are examples of **noncoding RNAs (ncRNAs)**.
- **Posttranslational regulation**.
    - **Feedback inhibition**. Feedback inhibition is a situation in which the chemical product of a metabolic pathway binds to and inhibits the enzymes that made the chemical product in the first place.
    - **Covalent modification**. Covalent modification involves altering the structure and function of a protein by attaching phosphate groups, methyl groups, sugars, or lipids.

Key Questions

- What is an advantage of transcriptional regulation?
- What is an advantage of posttranslational regulation?

Review Questions

Fill in the blank:

1. An _____ protein is a type of regulatory transcription factor that increases transcription.
2. _____ molecules are organic compounds that regulate the functions of repressor and activator proteins.
3. _____ is the enzyme that catalyzes the cleavage of the disaccharide lactose into two simple sugars.
4. Identify each of the following genes of the *lac* operon:

- The _____ gene encodes the *lac* repressor.
- The _____ gene encodes lactose permease.
- The _____ gene produces an enzyme that converts atypical isomers of lactose into forms that can be used in the bacteria cell.

5. The DNA binding site for the *lac* repressor protein is called _____.
6. The two inducers of the *lac* operon are _____ and _____.
7. A merozygote contains two copies of a gene; one gene is located on the _____ while the other gene is located on the _____.
8. _____ is an enzyme that is inhibited by glucose.
9. Lactose and glucose regulate expression of the lactose operon. The highest expression of the *lac* operon occurs when _____ is absent and _____ is present.

14 - Gene Regulation in Eukaryotes

Comparing Gene Regulation in Prokaryotes and Eukaryotes

The *lac* operon provides an excellent example of how bacteria perform gene regulation in response to an environment that lacks glucose yet contains lactose. In the case of the *lac* operon, we learned that gene regulation involves an activator protein (CAP) and a *lac* repressor protein. The effector molecules cAMP and allolactose regulate CAP and the *lac* repressor binding to regulatory DNA sequences (CAP site, operator) near the promoter for the *lac* operon. Ultimately the binding of the CAP and the *lac* repressor proteins determine if the sigma (σ) factor protein and the RNA polymerase core enzyme activate transcription.

Even though gene regulation in prokaryotes and eukaryotes is similar (e.g., both involve activator proteins, repressor proteins, effector molecules, and regulatory DNA sequences), eukaryotic gene regulation is more complex. This complexity is needed to produce multicellular eukaryotic organisms with cells in each tissue having unique phenotypes. For example, a white blood cell (leukocyte) and a muscle cell have the same collection of structural genes; however, gene regulation ensures that a leukocyte expresses leukocyte-specific proteins, while a muscle cell expresses muscle-specific proteins. Further, many eukaryotic organisms progress from a fertilized egg through complex developmental stages to produce the mature adult. Gene regulation ensures that embryonic genes are expressed only during embryonic development, while other genes are expressed solely in the adult.

Regulation of a typical eukaryotic gene involves **combinatorial control**. For example, a single eukaryotic gene can be regulated by a combination of:

- **Activator** proteins binding to **enhancer** DNA sequences**.**
- **Repressor** proteins binding to **silencer** DNA sequences**.**
- **Regulation of activator and repressor protein function.** This regulation of activator and repressor proteins involves effector molecules, covalent modification, and protein-protein interactions.
- **Modifying the structure of chromatin to activate or repress transcription.** Modifying chromatin involves chemically modifying histone proteins or altering the arrangement of nucleosomes near the core promoter of a gene.
- **DNA methylation to silence transcription.** The methylation of cytosine bases near the core promoter region of a gene inhibits transcription.

Key Questions

- What is meant by combinatorial control?
- What combinations of processes influence the transcription of a eukaryotic gene?

Core Promoter vs. Regulatory Promoter

We learned in Part 9 that transcription in eukaryotes involves several types of DNA sequences. The **core promoter**, for example, determines where RNA polymerase II will bind to the DNA and begin transcription**.** The core promoter includes the **TATA box (-25 sequence),** which serves as the binding site for the general transcription factor protein **TFIID** and the **+1 site**, the first base in the template DNA strand that is transcribed by RNA polymerase II. For transcription to occur, the TATA box and the +1 site must be present. If these two sequences are the only sequences present upstream of a gene, the gene will be transcribed at a low, yet constant rate (the so-called **basal** level of transcription).

In addition to the core promoter, many eukaryotic genes include a **regulatory promoter** (see **figure 14.1**)**.** The components of the regulatory promoter are required for transcription levels higher than the basal level provided by the core promoter. A common regulatory promoter component that is present in many eukaryotic genes is the **CAAT box**. The CAAT box is located at -80 and has the sequence 5'-GGCCAATCT-3'. Another common regulatory promoter component is a **GC box** (5'- GGGCGG – 3') located at -100. The CAAT and GC boxes are the binding sites for certain **activator** proteins. Thus, the CAAT and GC boxes can be considered enhancers adjacent to many eukaryotic structural genes.

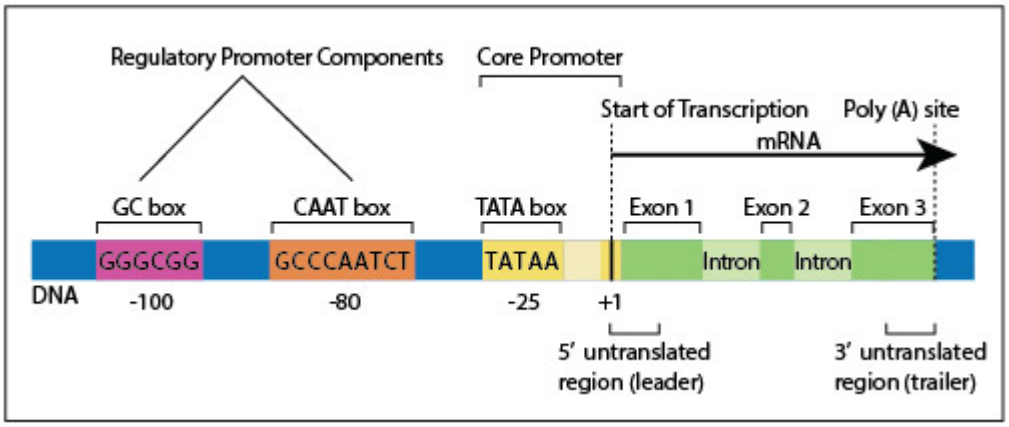


Figure 14.1 ***Core and Regulatory Promoter*** *--- Image created by SL*

Key Questions

- What is meant by basal transcription?
- What is the function of the regulatory promoter?
- What are the names of two common DNA sequences found in the regulatory promoters of eukaryotic genes?

General and Regulatory Transcription Factors

Eukaryotic **transcription factors** are proteins that influence the ability of RNA polymerase II to bind to a eukaryotic core promoter. There are two categories of transcription factor proteins:

- **General transcription factor proteins (GTFs).** The general transcription factor proteins include the **TFIID, TFIIA, TFIIB, TFIIF, TFIIE**, and **TFIIH** proteins described in Part 9. These proteins function to recruit RNA polymerase II to the core promoter to begin transcription. The general transcription factors are required for all transcription events. If these general transcription factors are the only proteins involved, the gene is transcribed at the basal level. The general transcription factors are also required for transcription rates above this basal level.
- **Regulatory transcription factor proteins**. Regulatory transcription factor proteins function to regulate transcription by either increasing transcription above the basal level or decreasing transcription below the basal level. An **activator** protein increases the level of transcription above the basal level; a **repressor** protein decreases the level of transcription below the basal level. Many activator and repressor proteins are only expressed in certain tissues or at certain times during development, thus playing a critical role in tissue-specific or time-specific gene expression.

Transcription factors proteins are ***trans*-acting factors** (i.e., can regulate genes found throughout the genome) and bind to DNA sequences called ***cis*-acting elements** (i.e., the DNA binding sites near the controlled gene) (see **figure 14.2**). However, these *cis*-acting elements are not always adjacent to the core promoter. Some *cis*-acting elements are within the gene that they control or can be thousands of base pairs away.

Recall that the **mediator** protein complex communicates the signals from activator and repressor proteins to RNA polymerase II. Mediator thus serves as a link between regulatory transcription factors, the GTF proteins, and RNA polymerase II, thereby determining the overall rate of transcription.

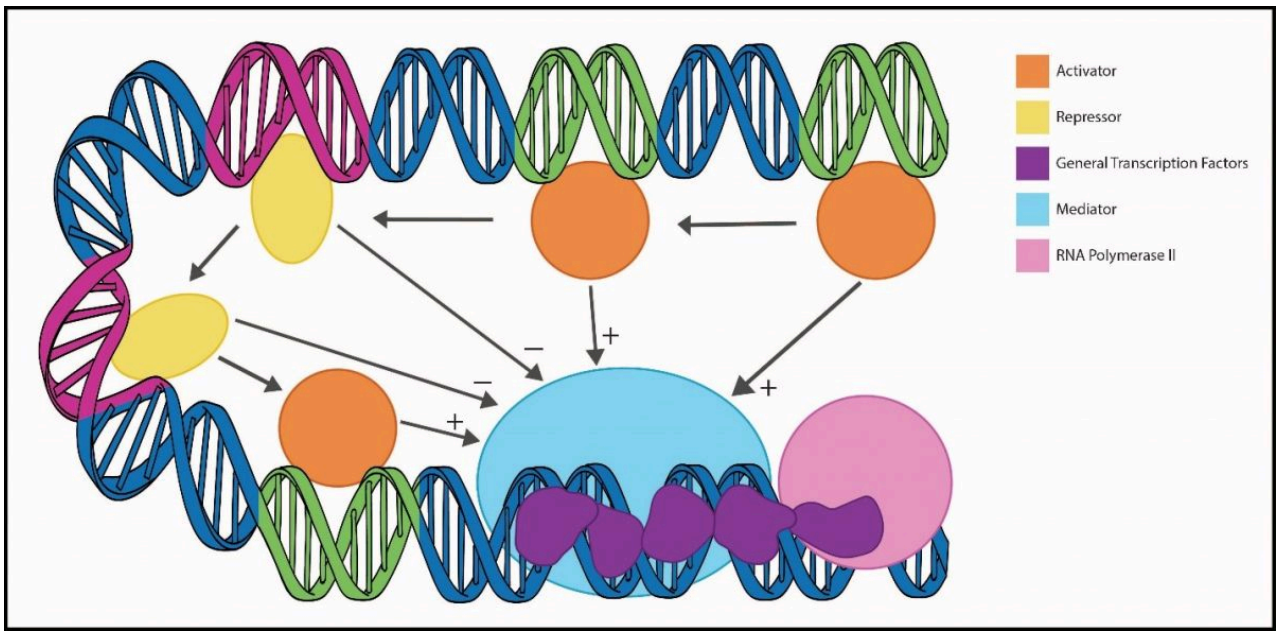


Figure 14.2 ***Trans-acting factors binding to cis-acting elements****. In this case, mediator interprets three activation signals and two silencing signals. Overall, transcription is increased above the basal level. --- Image created by SL.*

Key Questions

- Review the functions of TFIID, TFIIH, and mediator.
- Which transcription components are considered *trans*-acting factors?
- Which transcription components are considered *cis*-acting elements?

Enhancers and Silencers

Other regulatory DNA sequences assist the core promoter and regulatory promoter to regulate transcription by serving as the binding sites for transcription factor proteins. The binding of regulatory transcription factors to these DNA sequences may:

- **Increase the rate of transcription.** Transcription can increase 1000-fold when **activator** proteins bind to **enhancer** DNA sequences (**up-regulation**). Activator proteins and enhancer DNA sequences are generally responsible for tissue-specific expression of a gene.
- **Decrease the rate of transcription.** Transcription can decrease below the basal level when **repressor** proteins bind to **silencer** DNA sequences (**down-regulation**). Repressor proteins and silencer DNA sequences are generally responsible for tissue-specific repression of a gene.

A particular gene can be regulated by transcription factor proteins bound to different combinations of enhancer and silencer DNA sequences (see **figure 14.2**). The combination of the transcription factor proteins and regulatory DNA sequences involved determines the transcription pattern of the gene.

Key Questions

- Review the functions of activator proteins, repressor proteins, enhancer DNA sequences, and silencer DNA sequences (see Part 9).

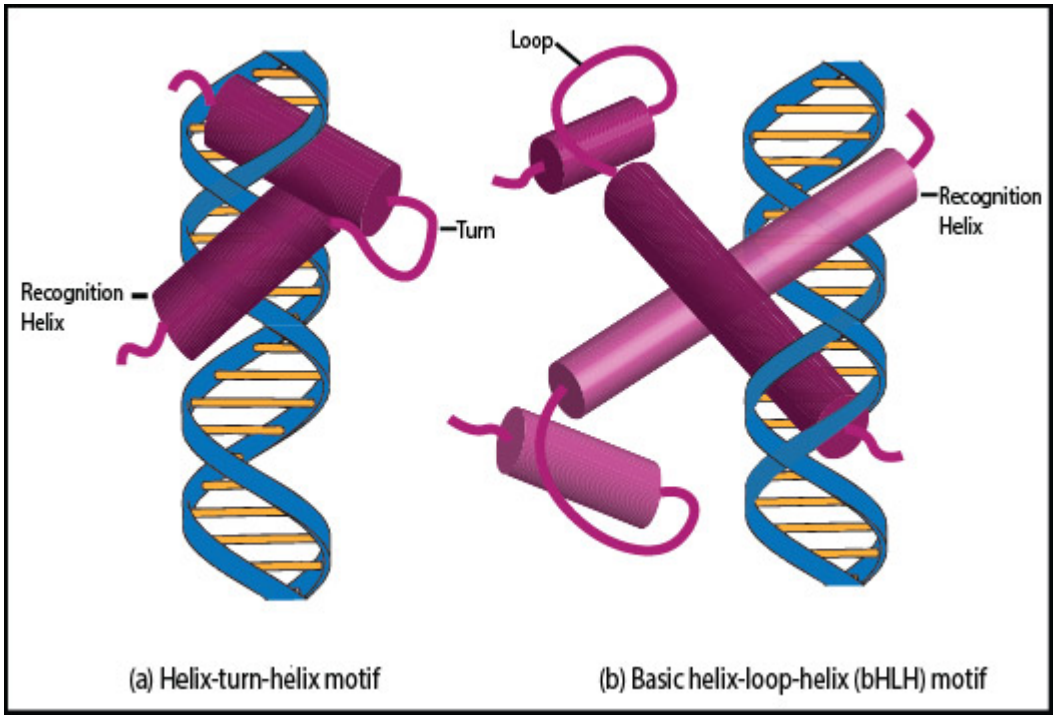
Structural Features of Transcription Factors

Transcription factor proteins have been identified in many organisms, including viruses, bacteria, fungi, plants, and animals. Nearly all transcription factor proteins contain conserved structural features that are important in either binding to regulatory DNA sequences, effector molecules, or other transcription factor proteins. For example, most transcription factor proteins contain **α-helices**, a type of protein secondary structure. An α-helix is produced when certain amino acids in the polypeptide sequence interact through hydrogen bonding to produce a helical structure. Importantly, the α-helix is the proper width to bind to the major groove in DNA. Thus, the α-helix is often used by transcription factors proteins to recognize specific base pair sequences located in the major groove of the DNA.

Four common **structural motifs** are found in transcription factor proteins. These structural motifs, based upon the α-helix structure described above, include (see **figure 14.3**):

- **Helix-turn-helix (HTH) motif**. The HTH motif includes two α-helices separated by a "turn" of 3-4 amino acids. One α-helix is called the **recognition helix**, and functions to bind to a specific base pair sequence in the DNA major groove. This recognition helix also includes basic (positively charged) amino acids that bind to the negatively charged DNA backbone. The helix-turn-helix motif is found in both prokaryotic and eukaryotic transcription factor proteins. For example, many of the transcription factor proteins that we have discussed contain the HTH motif including **sigma (σ) factor**, the ***lac* repressor** protein, and the **catabolite activator protein (CAP)**.
- **Basic helix-loop-helix (bHLH) motif**. The bHLH motif is similar to the helix-turn-helix motif, containing a recognition helix that binds to the DNA major groove. However, instead of a turn, bHLH transcription factors have an amino acid loop to connect two α-helices. bHLH transcription factors play an important role in cell division and differentiation. For example, the **MyoD** and **c-myc** proteins are transcription factor proteins that contain the bHLH motif. The MyoD protein activates muscle-specific genes, while the c-myc protein activates genes involved in cell division.
- **Zinc finger motif**. The zinc finger motif is composed of a finger-like structure composed of an α-helix (i.e., the recognition helix) and two β-strands (another type of protein secondary structure). Electrostatic interactions between zinc ions (Zn^{2+}) and negatively charged amino acid side chains within the transcription factor protein stabilize the zinc finger motif. Steroid hormone receptors, including the **glucocorticoid receptor** protein (see below), **testosterone receptor** protein, and the **estrogen receptor** protein contain zinc finger motifs.
- **Leucine zipper motif**. The leucine zipper motif not only contains a recognition helix, but also contains a second α-helix with many hydrophobic leucine amino acids. When the leucine-rich α-helices of two leucine zipper transcription factors interact, they form a **coiled-coil** to exclude water. The coiled-coil resembles a zipper with interlocking leucine amino acids. The DNA sequence is bound by recognition helices that extend from the coiled-coil region of these two transcription factor proteins. The **CREB** protein (see below) contains a leucine zipper motif.

It is important to note that all four transcription factor motif structures described above permit transcription factor proteins to bind to each other. Two identical transcription factor proteins interact to form a transcription factor **homodimer,** or two different transcription factor proteins interact to form a **heterodimer**. For example, both the CAP protein and the *lac* repressor proteins are homodimers, composed of two identical transcription factor protein with HTH motifs. Higher order interactions (trimers, tetramers) are also possible when transcription factor proteins bind to each other.



Caption

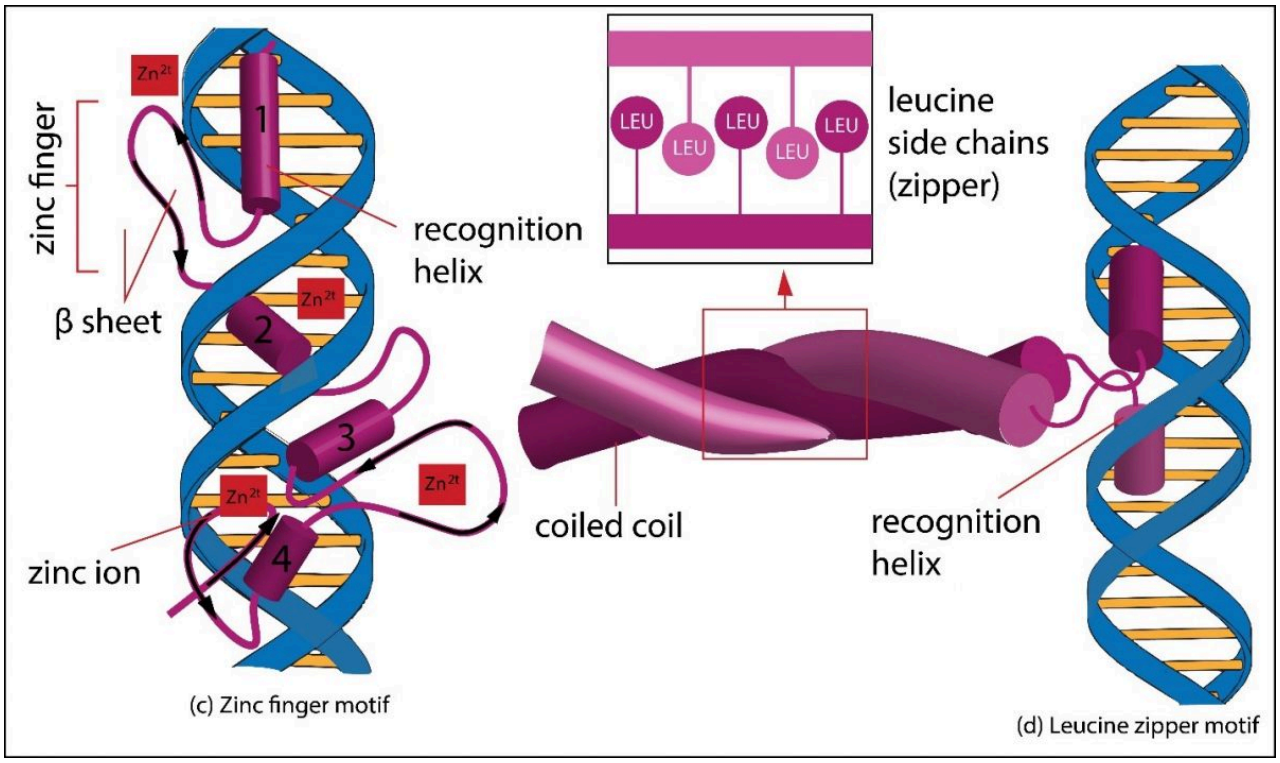


Figure 14.3 ***Transcription Factor Structural Motifs*** *a) Helix-turn-helix motif b) Basic helix-loop-helix motif c) Zinc finger motif d) Leucine zipper motif --- Images created by SL*

Key Questions

- What are three examples of transcription factor proteins that contain the helix-turn-helix (HTH) motif?
- What are two examples of transcription factor proteins that contains the basic helix-loop-helix (bHLH) motif?
- What are three examples of transcription factor proteins that contains the zinc finger motif?
- What is an example of a transcription factor protein that contains the leucine zipper motif?
- What protein secondary structure is found in all transcription factor structural motifs?
- What is meant by a transcription factor homodimer or heterodimer?

Mechanisms to Regulate Transcription Factor Proteins

If an activator protein is present in a cell, it does not always bind to an enhancer DNA sequence and up-regulate transcription. Similarly, a repressor protein does not always bind to a silencer DNA sequence and repress transcription. The DNA-binding activities of activator and repressor proteins are regulated in three ways:

- **Effector binding.** Small effector molecules bind to transcription factor proteins, change the conformation (shape) of the transcription factor, and influence the ability of the transcription factor protein to bind to enhancer or silencer DNA sequences. In animals, steroid hormones such as **glucocorticoid**, **testosterone**, and **estrogen** are effector molecules that regulate the functions of transcription factor proteins.
- **Transcription factor dimerization.** The formation of transcription factor homodimers or heterodimers influences binding to enhancer or silencer DNA sequences.
- **Covalent modification.** The addition of phosphate groups **(phosphorylation)** to activator or repressor proteins can stimulate binding to enhancer or silencer DNA sequences.

Note that for a particular gene, one or more of the above mechanisms may be involved in regulating gene expression. For example, the glucocorticoid receptor transcription factor protein (see below) is regulated by effector binding and dimerization, while the CREB transcription factor protein is regulated by dimerization and covalent modification.

Key Questions

- Describe the three ways that activator and repressor proteins can be regulated.
- What are three examples of eukaryotic effector molecules?

Regulating Transcription Through TFIID

Regulatory transcription factor proteins (activator and repressor proteins) influence the ability of RNA polymerase II to transcribe a gene. However, these regulatory transcription factor proteins do not typically bind to RNA polymerase II directly. Instead, transcription factor proteins communicate DNA binding indirectly to RNA polymerase II through other protein complexes. Eukaryotic regulatory transcription factors influence RNA polymerase II activity through **TFIID**, **mediator**, the enzymes involved in **chromatin remodeling,** and the enzymes involved in **DNA methylation**.

Consider first the regulation of RNA polymerase II through the TFIID protein. Recall that TFIID is the general transcription factor protein that binds to the TATA box (the -25 sequence) within the core promoter. TFIID recruits the other five general transcription factors (TFIIA, TFIIB, TFIIF, TFIIH, and TFIIE) that bring RNA polymerase II to the +1 site and activate RNA polymerase II to begin transcription. Suppose an activator protein binds to an enhancer DNA

sequence (see **figure 14.4**). This activator protein then encourages TFIID to bind to the TATA box, and TFIID then recruits the other general transcription factors and RNA polymerase II to the +1 site. As a result, transcription is up-regulated. Suppose instead that a repressor protein binds to a silencer DNA sequence adjacent to a gene. The repressor protein then prevents TFIID from binding to the TATA box. The absence of TFIID on the core promoter prevents the other general transcription factors and RNA polymerase II from binding to the core promoter. As a result, transcription is down-regulated.

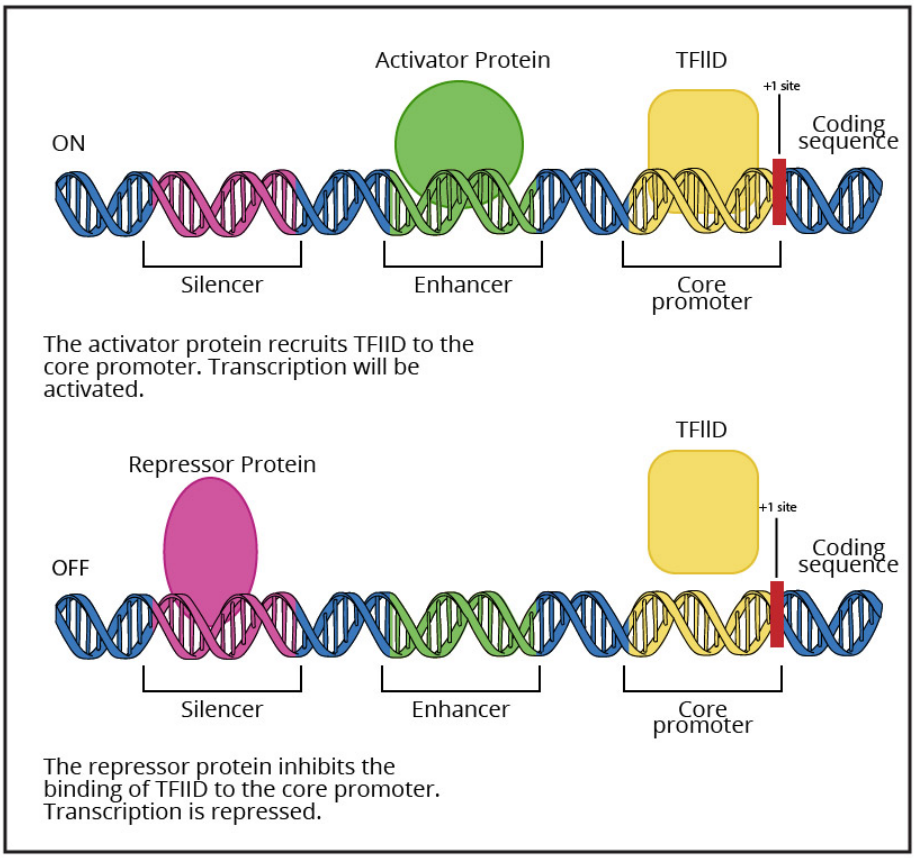


Figure 14.4 ***Regulating TFIID*** *- Image created by SL*

Key Questions

- How do activator and repressor proteins influence TFIID?

Regulating Transcription Through Mediator

Mediator is a protein complex that mediates the interaction between the regulatory transcription factors (i.e., activator and repressor proteins) and RNA polymerase II. If mediator activates RNA polymerase II, transcription begins. Suppose an activator protein binds to an enhancer DNA sequence (see **figure 14.5**). The activator protein in turn activates mediator, and mediator then activates the general transcription factor protein **TFIIH**. Next, TFIIH acts as a helicase to separate the template and coding DNA strands. TFIIH also acts as a kinase, phosphorylating RNA polymerase II to begin transcription.

Suppose a repressor protein binds to a silencer DNA sequence instead. The repressor protein then inhibits the activity of mediator. As a result, mediator fails to activate TFIIH, and TFIIH fails to separate the template and coding DNA

strands. TFIIH also fails to phosphorylate RNA polymerase II, preventing the initiation of transcription. Note that the DNA between the enhancer/silencer DNA sequences and the core promoter can form a loop to permit the proteins described above to bind to each other.

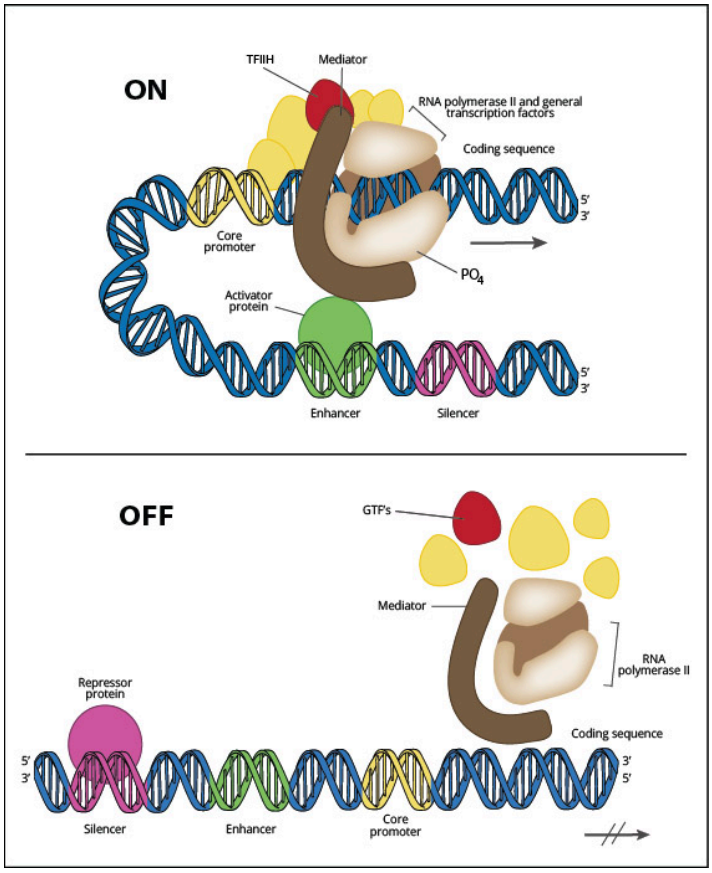


Figure 14.5 ***Regulating Mediator*** *--- Image created by SL*

Key Questions

- How do activator and repressor proteins influence the activity of mediator?

Transcription Activation Using the Glucocorticoid Receptor

Now let's apply what we have learned so far to two examples of gene regulation in the human body. The first example shows how steroid hormones produced by endocrine glands activate the transcription of genes. For example, **glucocorticoid hormones (GCs)** are released by the adrenal glands in response to fasting, as well as physical activity. The GCs lead to an increase in glucose synthesis, an increase in protein metabolism, an increase in fat metabolism, and a decrease in inflammation.

Glucocorticoid hormones increase the transcription of a gene above the basal level as follows (see **figure 14.6**):

1. The glucocorticoids are steroid hormones, which are nonpolar in structure. As a result, the glucocorticoids cross the cytoplasmic membrane and enter the cytoplasm of a target cell.
2. Glucocorticoids act as effector molecules by binding to an activator protein called **glucocorticoid receptor** that is found in many cell types. Prior to glucocorticoid binding, the glucocorticoid receptor is bound to **HSP90** proteins. HSP90 helps maintain the proper three-dimensional shape of the glucocorticoid receptor, so that glucocorticoid receptor can bind to glucocorticoid hormones produced by the adrenal glands. HSP90 is released when glucocorticoid hormone binds to glucocorticoid receptor.
3. Glucocorticoid binding changes the conformation (shape) of the glucocorticoid receptor, exposing a **nuclear localization signal (NLS).** The NLS is a polypeptide sequence that helps to target the glucocorticoid receptor (with bound glucocorticoid hormone) to the nucleus of the cell.
4. Two glucocorticoid receptors with bound glucocorticoid hormones form a homodimer in the cytoplasm of the cell.
5. The glucocorticoid receptor:glucocorticoid homodimer complex travels to the nucleus of the cell.
6. The glucocorticoid receptor:glucocorticoid homodimer complex binds to two adjacent enhancer DNA sequences called **glucocorticoid response elements (GREs)**. GREs are common enhancers found adjacent to many genes involved in metabolism.
7. The glucocorticoid receptor: glucocorticoid homodimer complex bound to the GRE sequences activates transcription.

Other steroid hormones, such as estrogen and testosterone, are effector molecules that activate transcription by binding to similar cytoplasmic transcription factor proteins. For example, estrogen binds to estrogen receptor proteins to activate transcription, while testosterone binds to testosterone receptor proteins to activate transcription. Both the estrogen receptor and testosterone receptor proteins are regulated by dimerization.

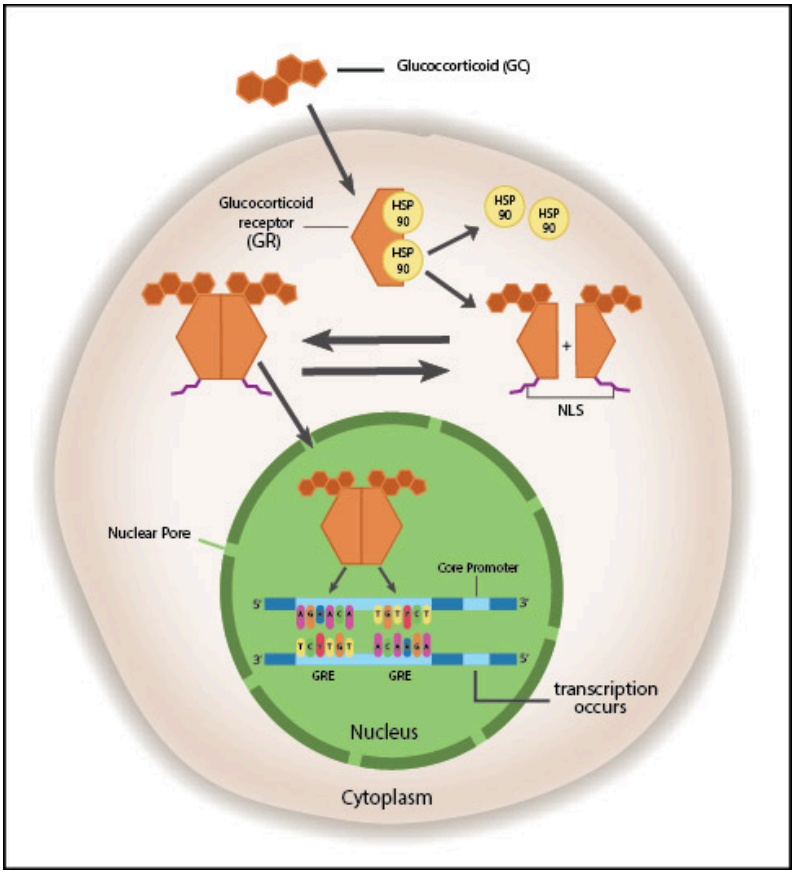


Figure 14.6 ***Transcription Regulation by Glucocorticoid*** *--- Image created by SL*

Key Questions

- How does the production of glucocorticoid by an adrenal gland lead to transcriptional activation of a target gene?

Transcription Activation via CREB

Unlike glucocorticoid, many signaling molecules in the body, such as peptide hormones, growth factor proteins, and cytokine proteins, are not able to diffuse through the cytoplasmic membrane into the cytoplasm of the target cell. Instead, these signaling proteins bind to cell receptors on the surface of a target cell, and the receptor binding signal is then transmitted to the nucleus to activate transcription. Our second example of gene regulation demonstrates how transcription is up-regulated when receptor binding activates the transcription factor protein **cAMP response element-binding protein (CREB)**. Transcription activation via CREB occurs when (see **figure 14.7**):

1. A receptor protein embedded in the cytoplasmic membrane binds to a peptide hormone, growth factor, or cytokine protein.
2. Receptor binding activates a **G protein**.
3. The G protein activates **adenylyl cyclase** inside the cell, which converts ATP into **cAMP**.
4. cAMP binds to and activates **protein kinase A (PKA).**
5. PKA moves into the nucleus and phosphorylates an inactive CREB protein homodimer.
6. The phosphorylated CREB protein homodimer binds to two adjacent enhancer sequences called **cAMP response elements (CREs).**
7. The phosphorylated CREB homodimer bound to the CRE sequences activates transcription.

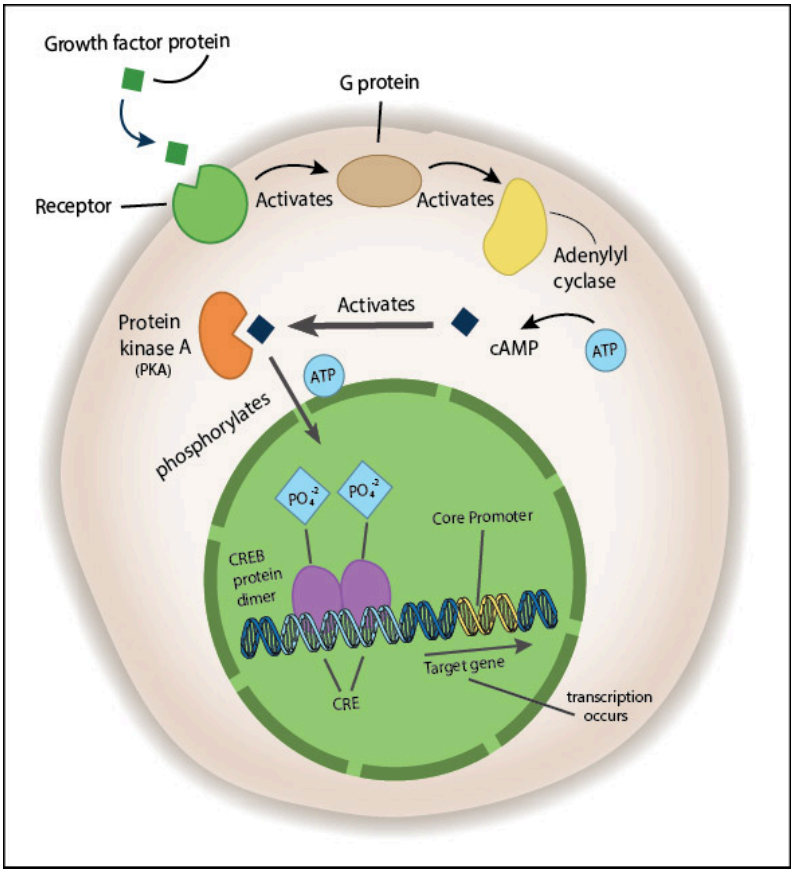


Figure 14.7 ***Transcriptional Regulation by CREB*** *--- Image created by SL*

Key Questions

- What is CREB and CRE?
- How does the binding of a signaling protein to a receptor lead to transcriptional activation of a target gene via the CREB pathway?

Chromosome Compaction and Transcription

The arrangement of nucleosomes on the DNA can also influence the transcription of a nearby gene (for a review of nucleosomes, refer to Part 2). For a gene to be transcribed, RNA polymerase II must be able to bind to the core promoter. If the core promoter region of a gene contains tightly packed nucleosomes (**heterochromatin**), RNA polymerase II struggles to find the core promoter. As a result, the heterochromatin form of DNA is said to be in a **closed conformation** and transcription is limited. Regions of the chromosome with loosely packed or absent nucleosomes are called **euchromatin** (**open conformation**). RNA polymerase II can better access a core promoter located in euchromatin, and as a result, transcription occurs more readily.

Recall that chromatin is a dynamic structure with a specific gene alternating between the closed (heterochromatin) and open (euchromatin) conformations depending on the needs of the cell. When an activator protein binds to an enhancer DNA sequence, chromatin is converted to the open conformation. When a repressor protein binds to a silencer DNA sequence, chromatin is converted to the closed conformation.

Key Questions

- Review the structure of a nucleosome and the terms heterochromatin and euchromatin (see Part 2).
- What is the difference between the open conformation and the closed conformation?

Arrangement of Chromatin at the β-globin Gene

As an example of how chromatin structure can influence the transcription of a gene, consider the human β-globin gene (see **figure 14.8**). The β-globin gene, which encodes the β-globin protein components of hemoglobin, is not normally expressed in many cell types, including fibroblast cells. When the DNA region that encompasses the β-globin gene from fibroblasts was analyzed with respect to nucleosomes, scientists discovered that nucleosomes were found at approximately 200 base pairs (bp) intervals from the -3000 to +1500 region of the gene. Note that this closed conformation region from -3000 to +1500 includes the regulatory promoter, core promoter, and the beginning portion of the β-globin gene. This heterochromatin arrangement of nucleosomes makes the β-globin promoter inaccessible to the general transcription factors (GTFs) and RNA polymerase II. As a result, the β-globin gene is not transcribed in fibroblasts.

The β-globin gene is expressed in erythroblasts (precursor red blood cells). When the nucleosome arrangement surrounding the β-globin gene was examined in erythroblasts, a different result was observed. Nucleosomes were displaced from the -500 to +200 region of the β-globin gene. This open conformation (euchromatin) area includes the regulatory promoter, core promoter, and the beginning portion of the β-globin gene. Thus, the GTFs and RNA polymerase II can access the regulatory and core promoter region in erythroblasts, leading to the transcription of the β-globin gene.

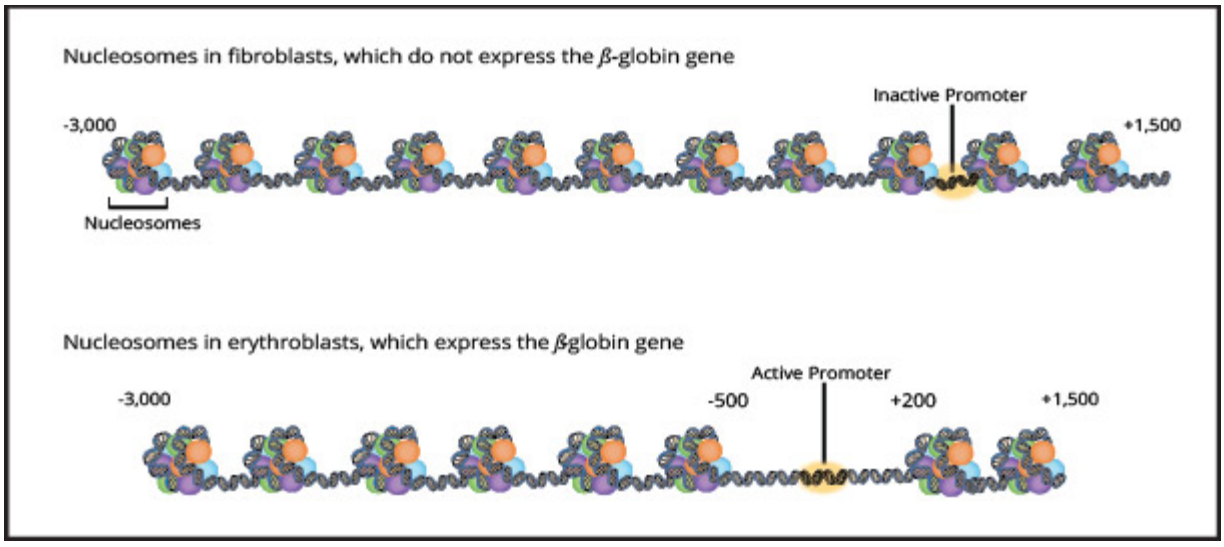


Figure 14.8 ***Nucleosome arrangement on the B-globin gene*** *--- Image created by SL*

Key Questions

- In terms of the regulatory and core promoter for the β-globin gene, describe the difference between chromatin structure in fibroblasts and erythroblasts.

Histone Acetylation

The results from fibroblasts and erythroblasts discussed above suggest that nucleosomes can be altered to influence transcription. Alterations in chromatin structure to promote transcription include the **covalent modification** of histone proteins and the rearrangement of nucleosomes within the promoter region by **ATP-dependent chromatin remodeling** (see **figure 14.9**).

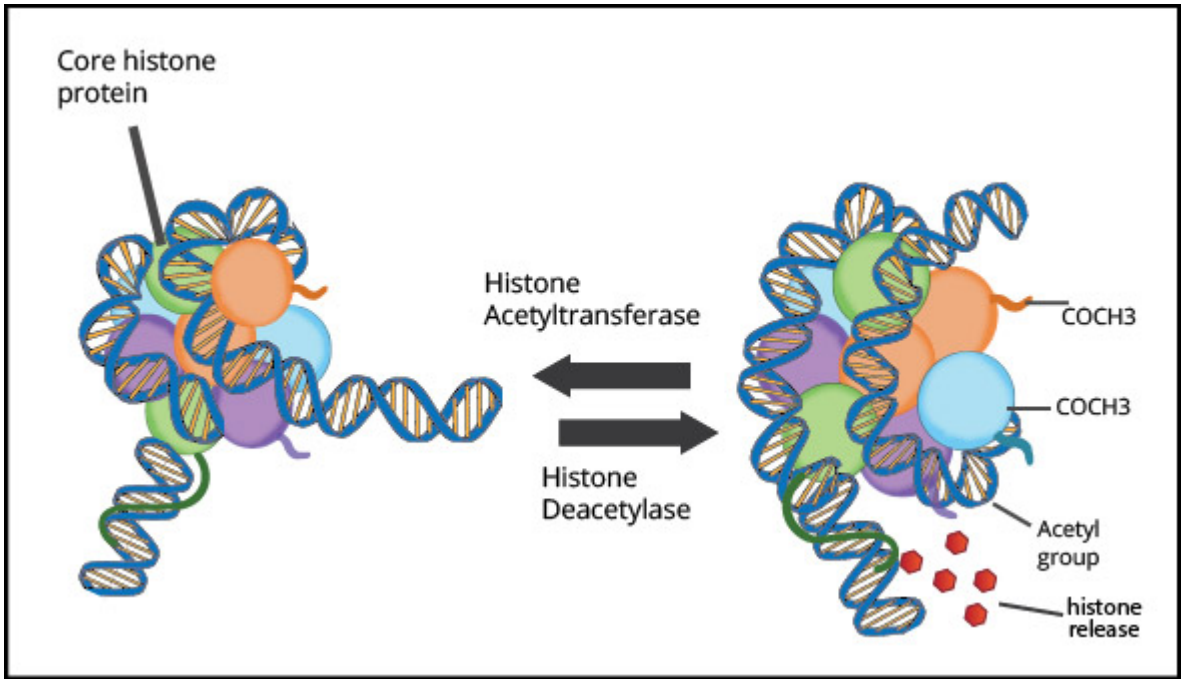
Covalent modification includes the **acetylation** of histone proteins within nucleosomes. Enzymes called **histone acetyltransferases (HATs)** add acetyl chemical groups to the tail regions within histone proteins (refer to the Part 2 reading for a description of histone structure). Acetylation neutralizes the positive charge on lysine amino acids within the histone tail, disrupting the interaction between the histone tail and the negatively charged DNA backbone. As a result, neutralization of the positive charges on the histone tails causes the histones to release from the DNA; the DNA is now more accessible for transcription. When transcription needs to be turned off, the histones are modified by **histone deacetylase (HDAC)** proteins that remove the acetyl groups from histones, restoring the positive charge on the histone tail. The histone tails once again bind to the negatively charged DNA backbone, and the chromatin is converted from the open to the closed conformation (heterochromatin), decreasing transcription of the gene.

Note that when an activator protein binds to an enhancer DNA sequence, the activator recruits HATs to the promoter, activating transcription. Alternatively, when repressor proteins bind to silencer DNA sequences, HDACs are recruited to the promoter, silencing transcription.

ATP-dependent Chromatin Remodeling

The ATP-dependent chromatin remodeling process uses the energy in ATP to alter the spacing of the nucleosomes in the promoter region near a gene (see figure 14.9). One example of an ATP-dependent chromatin remodeling enzyme is the multi-subunit **SWI/SNF** protein complex. The SWI/SNF protein complex performs at least two types of chromatin remodeling:

- SWI/SNF changes the distribution of nucleosomes along the DNA, creating large gaps between adjacent nucleosomes. When these larger gaps between nucleosomes includes the core promoter region of a gene, transcription in activated.
- SWI/SNF can replace the standard histone proteins (H2A, H2B, H3, and H4; see Part 2) within a nucleosome with **histone variant proteins.** The presence of these histone variant proteins within the modified nucleosome increases transcription.



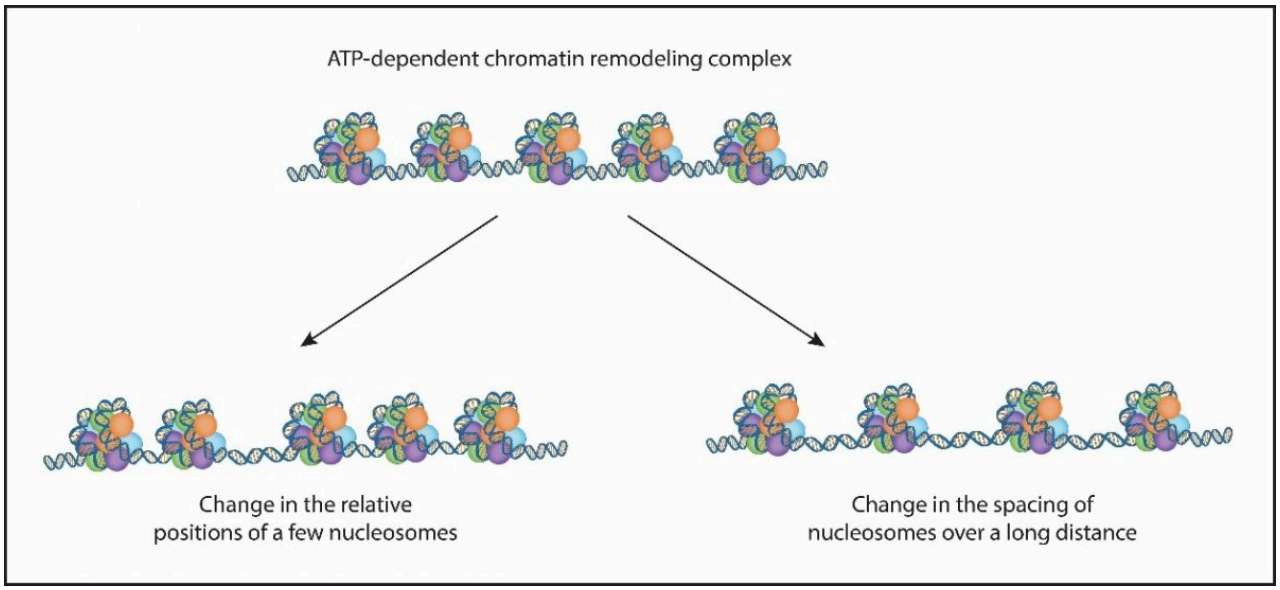


Figure 14.9 ***Histone Acetylation and ATP-Dependent Chromatin Remodeling*** *--- Images created by SL.*

Key Questions

- When a HAT is active, what effect does this have on transcription?
- When a HDAC is active, what effect does this have on transcription?
- What is the function of the SWI/SNF protein complex?

Overview of DNA Methylation

Silencing of gene expression in many eukaryotes involves the **methylation** of DNA sequences near the core promoters of genes. The methyl groups that are added to the DNA double helix block the major groove of the DNA, preventing the

recognition helices (see above) within activator proteins to enhancer sequences from binding to the DNA. Cytosine bases within CG-rich sequences called **CpG islands** are typically targets for DNA methylation. Not surprisingly, many CpG islands are located near the core promoters of genes (see **figure 14.10**). Typical CpG islands are 1,000 – 2,000 base pair (bp) long sequences that contain multiple **CpG sites** (i.e., many 5'-CG-3' dinucleotide sequences in a row). Within CpG islands, adding methyl groups to the cytosine bases on both DNA strands is called **full methylation**. Full methylation inhibits transcription.

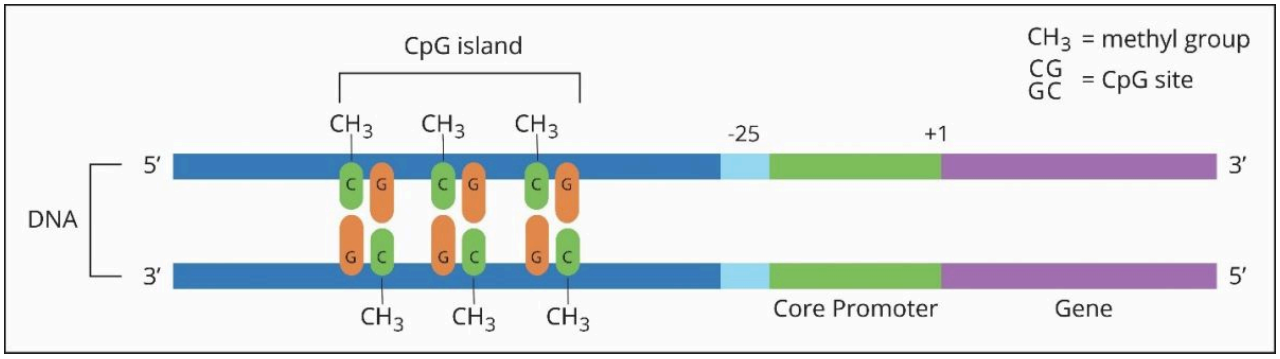


Figure 14.10 ***Overview of DNA Methylation.*** *CpG islands are the targets for DNA methylation to silence a gene. --- Image created by SL*

Housekeeping genes encode proteins that are required for the maintenance of a cell. For example, the structural genes that produce the enzymes involved in glycolysis are housekeeping genes The promoters of these housekeeping genes are typically unmethylated and as a result, housekeeping genes are always transcribed. **Tissue-specific genes** are only expressed in certain cell types. In cell types in which these genes are not expressed, the CpG island near the promoter is fully methylated. In cell types in which the gene is expressed, the CpG island near the promoter is unmethylated. As a final example, the inactive X chromosome (Barr body) in female mammals contains methylated CpG islands adjacent to most structural genes; this high degree of CpG island methylation renders the Barr body transcriptionally silent.

Key Questions

- How does methylation alter the structure of DNA?
- Where are many CpG islands located?
- In terms of DNA methylation, what is the difference between a housekeeping gene and a tissue-specific gene?

Methylation Blocks Activator Proteins and Recruits HDACs

DNA methylation is thought to silence the transcription of a nearby gene in two general ways. First, methylation at a CpG island near the promoter of a gene prevents an activator protein from binding to an enhancer DNA sequence (see **figure 14.11**). DNA methylation inhibits activator binding because the methyl group on cytosine prevents the recognition helix (see above) within activator proteins from binding to the DNA major groove. Second, methylated CpG islands near promoters serve as the binding sites for **methyl-CpG-binding proteins**. When a methyl-CpG-binding protein binds to a methylated CpG island, the methyl-CpG-binding proteins recruit histone deacetylases (HDACs). HDACs then remove the acetyl groups from histone tails, converting the core promoter region of the gene into the heterochromatin (closed) state. Transcription of the nearby gene is therefore inhibited.

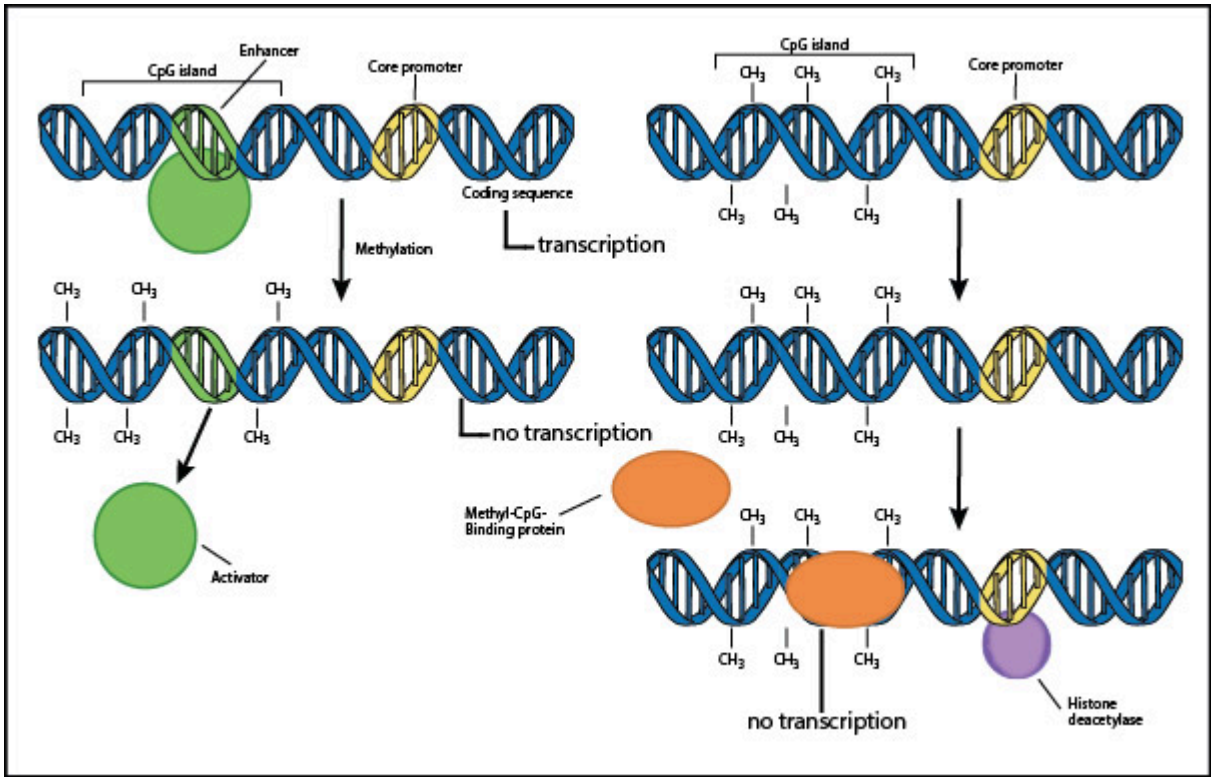


Figure 14.11 ***Methylation Inhibits Transcription*** *--- Image created by SL*

Key Questions

- Describe the two ways that DNA methylation can inhibit transcription.

DNA Methylation is Preserved During Cell Division

The DNA methylation pattern in the cell is established by a process called ***de novo* methylation** (see **figure 14.12**). *De novo* methylation converts unmethylated DNA to fully methylated DNA (i.e., both DNA strands are methylated). *De novo* methylation is thought to occur during embryonic development or when cells differentiate to form tissues. Unfortunately, the details of *de novo* methylation are poorly understood.

The DNA methylation pattern established during *de novo* methylation is preserved during cell division; if a CpG island is fully methylated in a cell prior to mitosis, the same CpG island is fully methylated in the two daughter cells at the conclusion of mitosis. **Maintenance methylation** ensures that the daughter cells produced by mitosis maintain the same methylation pattern as the parental cell. For instance, suppose that fully methylated DNA is replicated. Because the DNA replication machinery does not methylate nitrogenous bases, the daughter DNA strands produced do not contain methylated cytosine bases. Thus, the daughter double-stranded DNA molecules are initially **hemimethylated**, with a methylated parental strand and an unmethylated daughter DNA strand. This hemimethylated DNA is recognized by **DNA methyltransferase**, which subsequently methylates the cytosine bases on the daughter DNA strands, thus preserving the DNA methylation pattern established in the parental cell.

Methylation of DNA explains a genetic phenomenon called **genomic imprinting**. In oogenesis (egg cell formation) or spermatogenesis (sperm cell formation), a specific gene is methylated by *de novo* methylation. Following fertilization, the methylation pattern is maintained as the fertilized egg begins to divide. For example, if the paternal allele for a gene

is fully methylated by genomic imprinting, that paternal allele remains fully methylated in the cells of the offspring. We will discuss genomic imprinting more in Part 15.

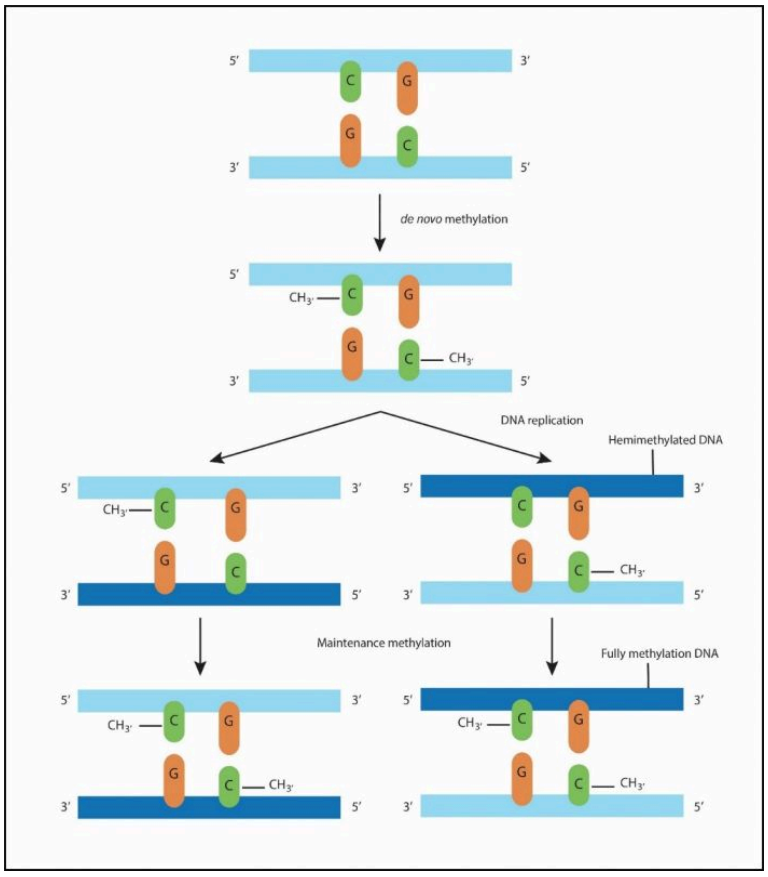


Figure 14.12 ***Preserving DNA Methylation During Cell Division*** *--- image created by SL*

Key Questions

- What is the difference between *de novo* and maintenance methylation?
- Which enzyme is responsible for maintenance methylation?
- What is meant by genomic imprinting?

Insulators

In eukaryotes, the processes that regulate the expression of one structural gene, such as activators/repressor proteins binding, histone acetylation, and DNA methylation do not necessarily influence the regulation of an adjacent gene. **Insulator** DNA sequences define the boundaries between genes (see **figure 14.13**); an insulator DNA sequence ensures that the gene regulation processes that affect one gene do not affect nearby genes. Insulator DNA sequences:

- **Serve as the binding sites for proteins that act as physical barriers for the HATs, HDACs and SWI/SNF protein complexes.** For example, suppose a gene is flanked by two insulator DNA sequences, and HATs modify histone tails and activate transcription of the gene. Because the proteins bound to insulators serve as physical barriers to the HATs, genes beyond the insulator sequences are not activated.
- **Serve as the binding sites for proteins that limit the effects of enhancer/silencer sequences.** Suppose that Gene A has an adjacent enhancer DNA sequence. Gene B is also near the enhancer DNA sequence. A protein bound to the insulator DNA sequence between Genes A and B ensures that the enhancer only activates Gene A; the transcription of Gene B is unaffected. Insulators can limit the effects of silencer DNA sequences in a similar way.

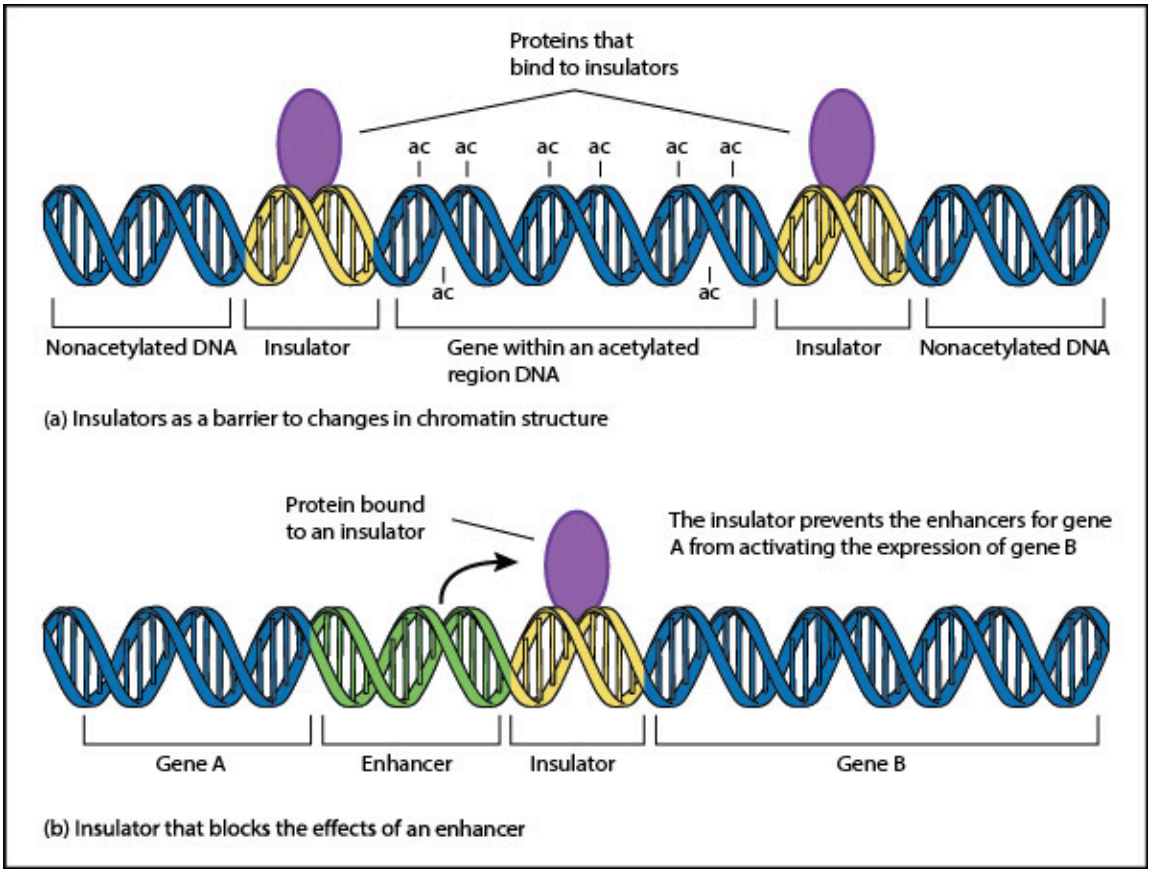


Figure 14.13 ***Insulators*** *--- Image created by SL*

Key Questions

- What is an insulator DNA sequence?
- How do insulators ensure that gene regulation is limited to a single gene?

Part 14 Review

Fill in the blank:

1. The core promoter consists of two consensus DNA sequences located at position _____ and _____.
2. The general transcription factor (GTF) proteins are _____.
3. Some examples of regulatory transcription factor proteins are _____, which increase transcription and _____, which decrease transcription below basal levels.
4. Transcription factor proteins contain structural motifs. Two transcription factors with the _____ motif interact and form a coiled coil. Two alpha-helices are part of a _____ motif seen in proteins involved in muscle cell differentiation.
5. The interaction of two identical transcription factor proteins to produce one molecule is called a _____.
6. One example of a steroid hormone is _____.
7. A glucocorticoid receptor is bound to _____ until a glucocorticoid hormone molecule binds to the receptor.
8. CREB is a (protein OR DNA sequence; circle the correct answer), whereas CRE is a (protein OR DNA sequence; circle the correct answer).
9. Upon its activation in the CREB system, protein kinase A (PKA) enters the nucleus and phosphorylates CREB which then leads to transcription being (turned ON or turned OFF; circle the correct answer).
10. Histone acetyltransferases add acetyl groups to _____ amino acids on the histone tail.
11. Acetyl groups are removed from histone tails by enzymes called _____.
12. Methyl groups added to cytosine bases usually project into the (major OR minor; circle the correct answer) groove of the DNA.
13. Housekeeping genes are usually (methylated OR unmethylated; circle the correct answer) while tissue-specific genes are (methylated or unmethylated; circle the correct answer) in cells that do not express the gene.
14. _____ methylation ensures that the methylation pattern continues in the daughter cells produced by mitosis.

15 - Epigenetics

Epigenetics involves cellular processes that alter the expression of genes and change the phenotype of an individual; however, these processes do not alter the nucleotide sequence within the DNA. As a result, epigenetic changes are not considered to be mutations. Instead, epigenetic mechanisms modify the structure of the DNA or the chromatin (i.e., the histones within nucleosomes) surrounding a gene or, in one case, alters the structure of an entire chromosome. These structural modifications either activate or silence transcription.

Key Questions

- What is meant by epigenetics?

Timing of Epigenetic Processes

The epigenetic factors that modify the DNA or alter chromatin structure are either established during the formation of gamete cells, embryonic development, or in the adult organism as a response to environmental agents. Processes that promote epigenetic changes during gamete formation include **genomic imprinting.** In genomic imprinting, the epigenetic changes established during gamete formation in one of the two parents are passed to their offspring. These epigenetic changes that are inherited are said to display **epigenetic inheritance**.

Epigenetic changes established during embryonic development include **X chromosome inactivation (XCI)**, and the processes that govern the **differentiation** of embryonic cells into adult cell types, including muscle cells, neurons, or epithelial cells.

Environmental factors that influence epigenetic changes in an adult organism include diet, stress, the unique environment of space, and the toxins found in cigarette smoke.

Epigenetic changes are permanent in the individual. For example, when an epigenetic change is established in a cell, this affected cell divides by mitosis, and the epigenetic changes are preserved in the daughter cells. This allows daughter cells to "remember" the epigenetic changes of the parental cell. Even though the epigenetic changes may be permanent in the individual, most epigenetic changes are erased during gamete formation and as a result, are not passed on to offspring. An exception to this general rule is genomic imprinting (see below).

Key Questions

- Which epigenetic process occurs during gamete formation?
- What is meant by epigenetic inheritance?
- What epigenetic processes occur during development?
- List some environmental factors that promote epigenetic changes.

Epigenetic Mechanisms

Three major epigenetic mechanisms influence the transcription of genes (see **figure 15.1**). These epigenetic mechanisms include:

- **DNA methylation.** We learned in Part 14 that **CpG islands** adjacent to structural genes are targets for DNA methylation. If the CpG island near a gene has a low level of methylation **(hypomethylation)**, the gene is actively transcribed. Conversely, a high level of CpG methylation **(hypermethylation)** corresponds to a silenced gene. We will investigate how DNA methylation influences the phenotype of an organism by considering the phenomenon of **genomic imprinting**.
- **Histone modifications.** Covalent modifications to histone tail domains represent a second type of important epigenetic modification. Two major histone modifications will be discussed in this section:
  - **Acetylation of histone tails.** The addition of acetyl groups by **histone acetyltransferases (HATs)** neutralizes the positive charges within the histone tail, activating transcription. On the other hand, **histone deacetylases (HDACs)** function to remove the acetyl groups from the histone tails, resulting in a tighter interaction between histone proteins and the DNA backbone. If histone deacetylation occurs near the promoter of a gene, transcription of the gene is inhibited.
  - **Methylation of histone tails.** The methylation of a lysine amino acid at position 4 within the histone H3 tail domain activates genes, while the methylation of a lysine at position 27 within histone H3 silences genes. The enzymes that add methyl groups to histone tails are called **histone methyltransferases**; the enzymes that remove methyl groups from histone tails are called **histone demethylases**. We will investigate how histone methylation is involved in activating and deactivating genes during embryonic development.
- **RNA-associated silencing.** RNA-associated silencing involves the use of specialized types of **noncoding RNA (ncRNAs)** molecules to silence the expression of genes. An example of RNA-associated silencing involves **X chromosome inactivation (XCI)**. Recall that during XCI, the ***Xist*** gene produces a noncoding RNA molecule (***Xist* RNA**) that inactivates an X chromosome. **MicroRNAs (miRNAs)** are another group of small ncRNAs that function in RNA-associated gene silencing. When a miRNA forms base pairs with a particular mRNA, the mRNA is degraded prior to translation. It is estimated that nearly 60% of all structural genes in the human genome are regulated by miRNAs.

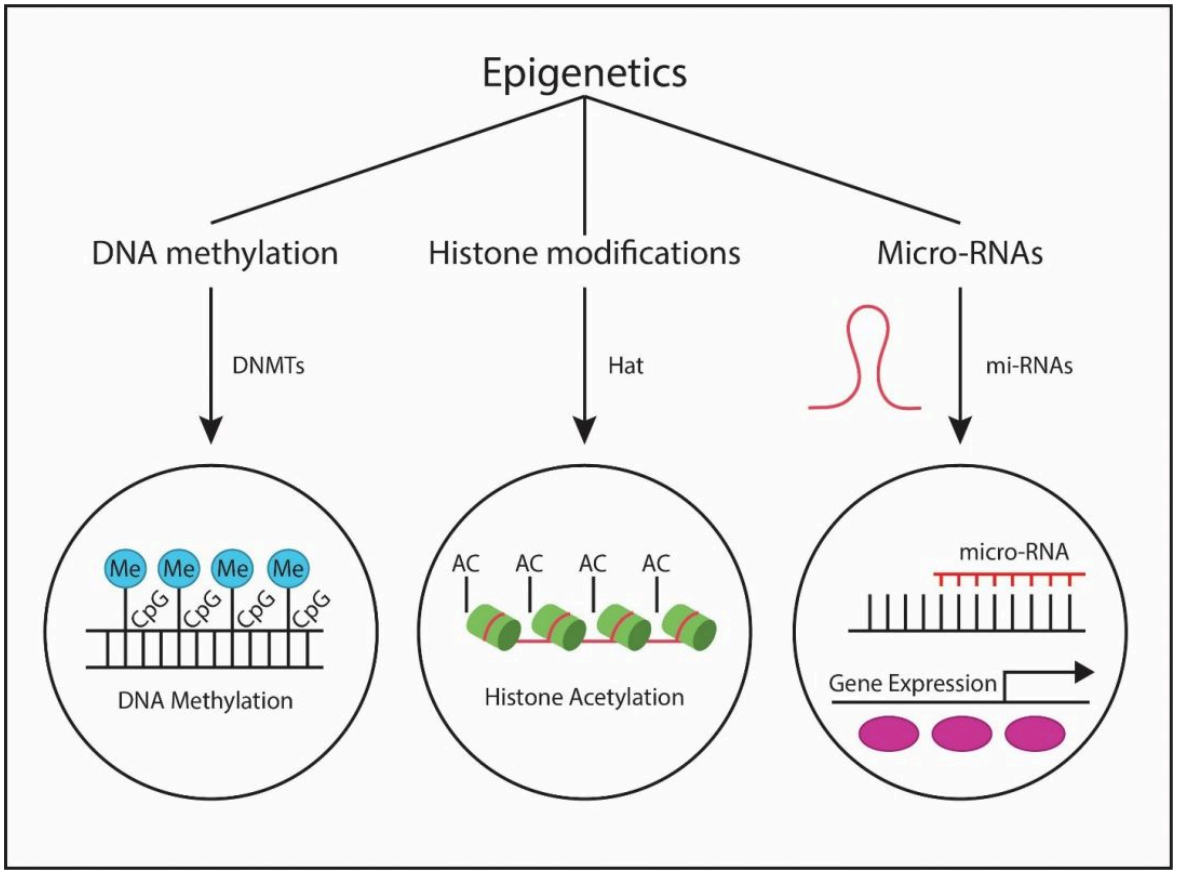


Figure 15.1 ***Epigenetic Mechanisms*** *--- Image created by SL*

Key Question

- Describe the three major mechanisms that promote epigenetic changes.
- Methylation of CpG islands in the DNA typically silences transcription. What effect does methylating histone H3 have on transcription?
- What are two examples of ncRNAs that participate in epigenetics?

Genomic Imprinting

Genomic imprinting involves inheriting a silenced gene from one parent. Since the active copy of the gene is inherited from the other parent, genomic imprinting causes the offspring to only express one of the two possible alleles that control a trait (**monoallelic expression**). The genomic imprint (in other words, the DNA methylation pattern) is established on the allele during gamete formation in one of the parents, is passed on to the offspring, and is retained throughout the lifetime of the offspring.

A well known example of genomic imprinting involves the regulation of the **insulin-like growth factor 2 (*Igf2*)** gene that contributes to body size in mice (see **figure 15.2**). There are two *Igf2* alleles in the population: the *Igf2* allele produces normal body size, while the *Igf2*$^{-}$ allele produces dwarf body size. In the case of the *Igf2* gene, the maternally-inherited allele is silenced, resulting in the offspring expressing the paternally-inherited allele only. For example, suppose a homozygous dwarf female (*Igf2*$^{-}$ *Igf2*$^{-}$) mouse is mated to a homozygous normal male (*Igf2 Igf2*) mouse. All the offspring are normal body size because the offspring inherited the active *Igf2* allele from the father. The *Igf2*$^{-}$ allele

inherited from the mother has been silenced and does not contribute to phenotype. Note that the offspring have the *Igf2 Igf2* $^{-}$ genotype. Alternatively, when a homozygous normal female (*Igf2 Igf2*) mouse is mated to a homozygous dwarf male (*Igf2* $^{-}$ *Igf2* $^{-}$) mouse, all the offspring are dwarf because the offspring inherited the active *Igf2* $^{-}$ allele from their father. The *Igf2* allele inherited from the mother has been silenced and does not contribute to phenotype. Note that these offspring are also heterozygous (*Igf2 Igf2* $^{-}$*)*. The results of these two crosses violate Mendel's laws of inheritance; the two crosses produce offspring with the same genotype (*Igf2 Igf2 -),* yet have different phenotypes.

Figure 15.2 ***An example of genomic imprinting****. A dwarf (left) and a normal (right) mouse.*

Key Questions

- What is meant by monoallelic expression?
- In the case of body size in mice, which *Igf2* allele is expressed? Which allele is silenced?

Genomic Imprinting Stages

The genomic imprinting mechanism has three stages (see **figure 15.3**):

1. **Establishment of the imprint**. In the case of the *Igf2* gene, imprinting occurs during egg formation, silencing the maternal allele (*Igf2* $^{-}$ in **figure 15.3**) for the gene. The maternal allele remains silent through fertilization. During sperm formation, the paternal allele (*Igf2*) remains active, so the offspring will be normal in size. The two heterozygote (*Igf2 Igf2* $^{-}$*)* offspring mice in **figure 15.3** express the paternal allele.
2. **Maintenance of the imprint**. After fertilization and subsequent cell divisions in the offspring mouse (*Igf2 Igf2* $^{-}$ genotype*),* the maternal allele is maintained in a silenced form. The mouse only expresses the paternally inherited allele.
3. **Erasure and reestablishment**. In both the male and female offspring, the imprint is erased when these offspring mice form their own gametes. After erasing the imprint, the imprint can then be reestablished depending on the sex of the offspring mouse:

    In the female offspring mice (*Igf2 Igf2* $^{-}$ genotype), both *Igf2* alleles are silenced during the formation of gametes. Note that 50% of the eggs have the silenced *Igf2* allele, while 50% of the eggs have the silenced *Igf2* $^{-}$ allele. As a result, in the female offspring, the imprint is reestablished during gamete formation. The imprinted (silenced) alleles are then passed by the female mouse to her offspring.

    - In the male offspring (*Igf2 Igf2* $^{-}$ genotype), both *Igf2* alleles remain active during the formation of gametes. 50% of the sperm cells have the active *Igf2* allele, while 50% of the sperm cells have the active *Igf2* $^{-}$ allele. Thus, in males, the imprint is not reestablished. The active *Igf2* alleles are then passed by the male mouse to his offspring.

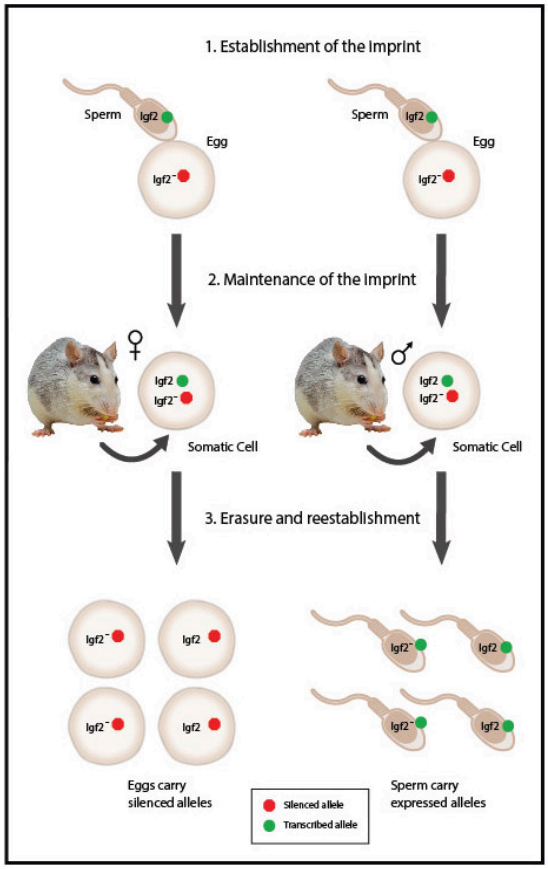


Figure 15.3 ***Stages of Genomic Imprinting*** *--- Image created by SL*

Key Questions

- Describe the events that are occurring during the three stages of genomic imprinting.

Genomic Imprinting Mechanism

Genomic imprinting involves DNA methylation patterns established during gamete formation. Genomic imprinting also involves several DNA sequences located near the *Igf2* gene (see **figure 15.4**). The *Igf2* gene in mice is located near another gene called ***H19***. The function of the *H19* gene is currently unknown; however, an enhancer sequence that functions to regulate the transcription of the *Igf2* gene is located next to the *H19* gene. DNA methylation occurs at two DNA sequences on each side of the *Igf2* gene. The first DNA sequence is the **imprinting control region (ICR)** and is located between the *H19* and *Igf2* genes. A second DNA sequence called the **differentially methylated region (DMR)** is located downstream of *Igf2*.

During the formation of egg cells, both the ICR and the DMR sequences are unmethylated. The absence of methylation allows **CTC-binding factor (CTCF)** proteins to bind to 5'-CTC-3' trinucleotide sequences within both ICR and DMR. The CTCF proteins bound to the ICR and DMR sequences also bind to each other, forcing a loop to form in the DNA. This loop containing the *Igf2* is considered a heterochromatin structure. When the *Igf2* gene is found within heterochromatin, an activator protein fails to bind to the enhancer DNA sequence adjacent to *H19*. As a result, the *Igf2* gene is silenced in egg cells.

During the formation of sperm cells, the ICR and DMR sequences are methylated by the ***de novo*** **methylation pathway.** CTCF proteins fail to bind to methylated ICR and DMR sequences, preventing the formation of a DNA loop containing the *Igf2* gene. Without the DNA loop, the *Igf2* gene is essentially located within euchromatin. In the absence of loop formation, an activator protein binds to the enhancer next to the *H19* gene, and the *Igf2* gene is transcribed. Note that even though DNA methylation usually silences genes by preventing activator proteins from binding to enhancer DNA sequences (see Part 14); in the case of the *Igf2* gene, DNA methylation prevents the binding of proteins that form heterochromatin. As a result, in mice, the methylation of DNA sequences near the *Igf2* gene activates transcription.

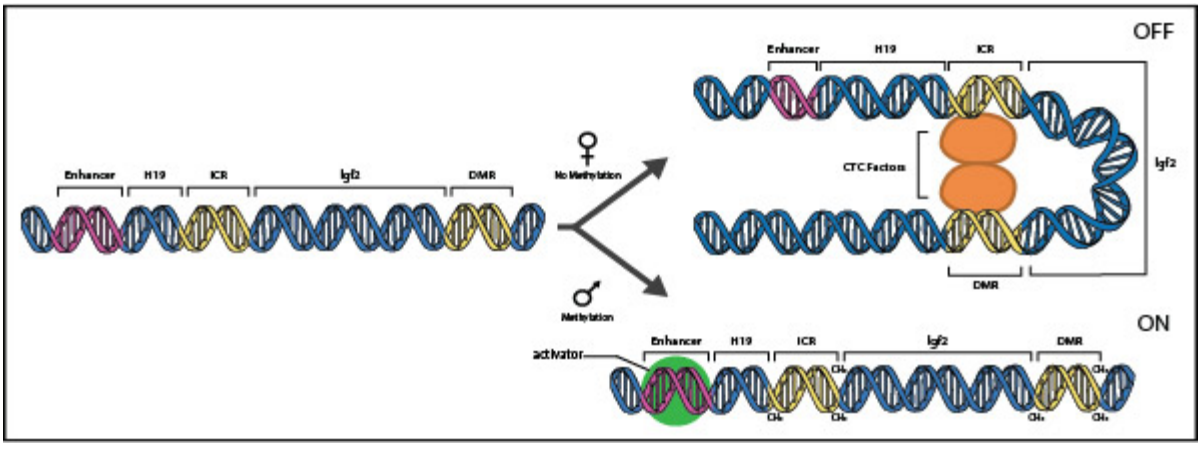


Figure 15.4 ***Genomic Imprinting Mechanism*** *--- Image created by SL*

Key Questions

- How do the activator protein, enhancer sequence, ICR sequence, DMR sequence, CTCF proteins, and a DNA loop contribute to the transcription of the *Igf2* gene?

Angelman and Prader-Willi Syndromes

Genomic imprinting plays an important role in two genetic diseases in humans: **Angelman syndrome (AS)** and **Prader-Willi syndrome (PWS)**. AS patients are thin, hyperactive, display mental deficiencies, have involuntary muscle contractions, and seizures. In contrast, PWS patients have an uncontrollable appetite, obesity, diabetes, small hands/feet, and like AS patients, have mental deficiencies.

In addition to genomic imprinting, both AS and PWS involves an identical deletion in the long arm (*q* arm) of chromosome 15 (see **figure 15.5**). This region of chromosome 15 contains a small group of genes that are either maternally or paternally imprinted. For example, in AS, a gene on chromosome 15 called ***UBE3A*** is imprinted (silenced) during sperm formation, meaning that a sperm cell contains a silenced *UBE3A* allele. If an offspring inherits this silenced *UBE3A* allele from the father and inherits a deletion copy of chromosome 15 (missing the *UBE3A gene*) from the mother, the offspring has no active *UBE3A* alleles. The absence of a active *UBE3A* allele produces the AS disease phenotype.

The genes involved in PWS have not been determined; however, candidate genes on chromosome 15 include ***SNRPN*** (encodes a splicing factor protein) and ***NDN***. In PWS, the *SNRPN* and *NDN* genes are imprinted (silenced) during egg formation. If the offspring inherits silenced *SNRPN* and *NDN* alleles from the mother and inherits a deletion copy of chromosome 15 (missing the *SNRPN* and *NDN* genes) from the father, the offspring lack active *SNRPN* and *NDN* alleles. The absence of active *SNRPN* and *NDN* alleles is thought to produce the PWS disease phenotype.

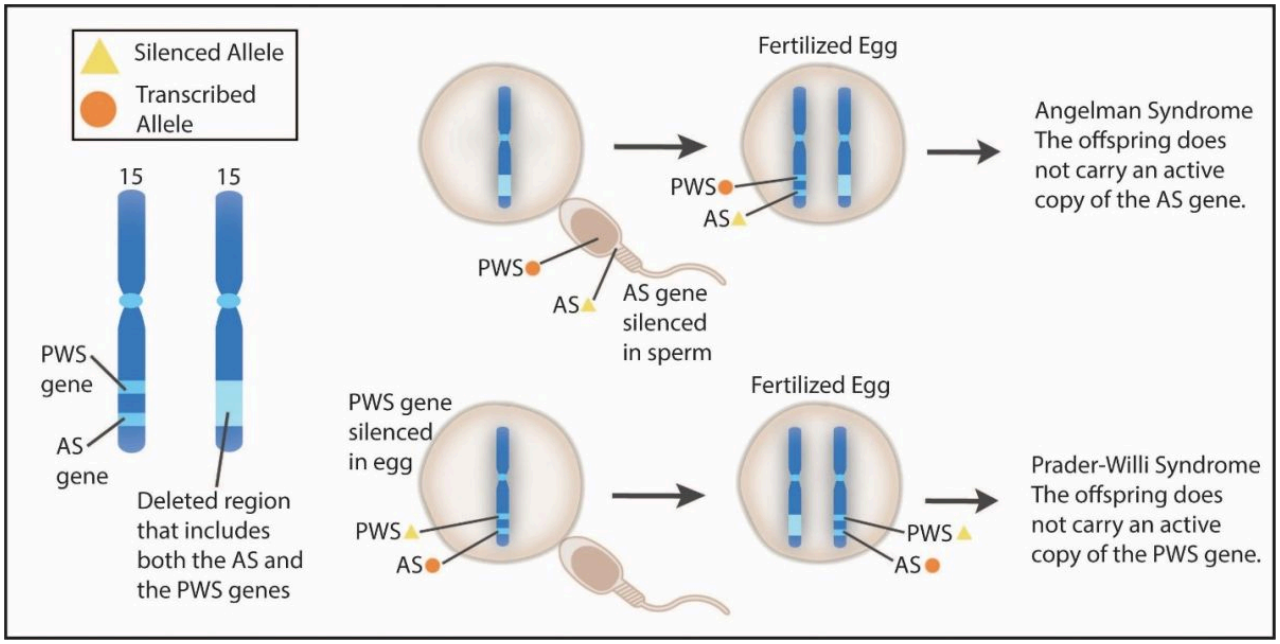


Figure 15.5 ***Mechanism of AS and PWS.*** *The AS gene in the figure represents the UBE3A gene described in the text, while the PWS gene in the figure represents the SNRPN and NDN genes described in the text. --- Image created by SL*

Key Questions

- What defect in chromosome structure contributes to both AS and PWS?
- Describe the two copies of chromosome 15 in a patient with AS.
- Describe the two copies of chromosome 15 in a patient with PWS.

X Chromosome Inactivation Mechanism

Now let's learn how RNA-associated silencing contributes to epigenetics. We learned previously (see Part 2) that during embryogenesis, one of the two X chromosomes in female mammals is randomly chosen for inactivation. This random inactivation process is called **X chromosome inactivation (XCI).** After XCI occurs, the inactive X chromosome is maintained in a transcriptionally silent state with each cell division.

In Part 2, we learned that each X chromosome contains a region near the centromere called the **X inactivation center (*Xic*)** that plays an important role in XCI. Within the *Xic* are two genes, the ***Xist*** and ***Tsix*** genes. The *Xist* gene is expressed preferentially from the X chromosome that will be inactivated, while the *Tsix* gene is expressed from the X chromosome that will remain active. The XCI process involving the *Xist* and *Tsix* genes occurs as follows (see **figure 15.6**):

1. Prior to XCI, a group of activator proteins called **pluripotency factors** bind to enhancer sequences on both X chromosomes and activate the transcription of both copies of the *Tsix* gene. *Tsix* expression produces *Tsix* RNA molecules, which inhibit the expression of the *Xist* genes on both X chromosomes. The two X chromosomes are active at this point.
2. The *Xic* regions on the two X chromosomes interact, causing the X chromosomes to pair. The pairing of the X chromosomes lasts less than an hour and involves the pluripotency factor proteins and CTCF proteins binding to both X chromosomes.
3. The pluripotency factors and CTCF proteins shift from both X chromosomes to just one of the two X chromosomes. The X chromosome that now contains the pluripotency factors and CTCF proteins continues to express the *Tsix* RNA and will remain active. The other X chromosome (without the pluripotency factors and CTCF proteins) silences *Tsix* expression and begins to express the *Xist* RNA. As a result, the X chromosome that expresses *Xist* will be inactivated.
4. The expressed *Xist* RNA molecules begin to bind to each other and to the X chromosome destined for inactivation. The *Xist* RNA binds initially to the *Xic* but later spreads in both directions along the X chromosome.
5. The *Xist* RNA produces the following epigenetic changes to the inactivated X chromosome:
    - *Xist* RNA recruits the ***de novo*** **methylation** proteins to the X chromosome that will be inactivated. *De novo* methylation occurs on CpG islands throughout the X chromosome, silencing approximately 80% of the X-linked genes.
    - *Xist* RNA recruits **histone methyltransferases** that function to add three methyl groups to a lysine amino acid located at position 27 within the histone H3 protein. As described below, the methylation of lysine 27 on histone H3 inhibits transcription.

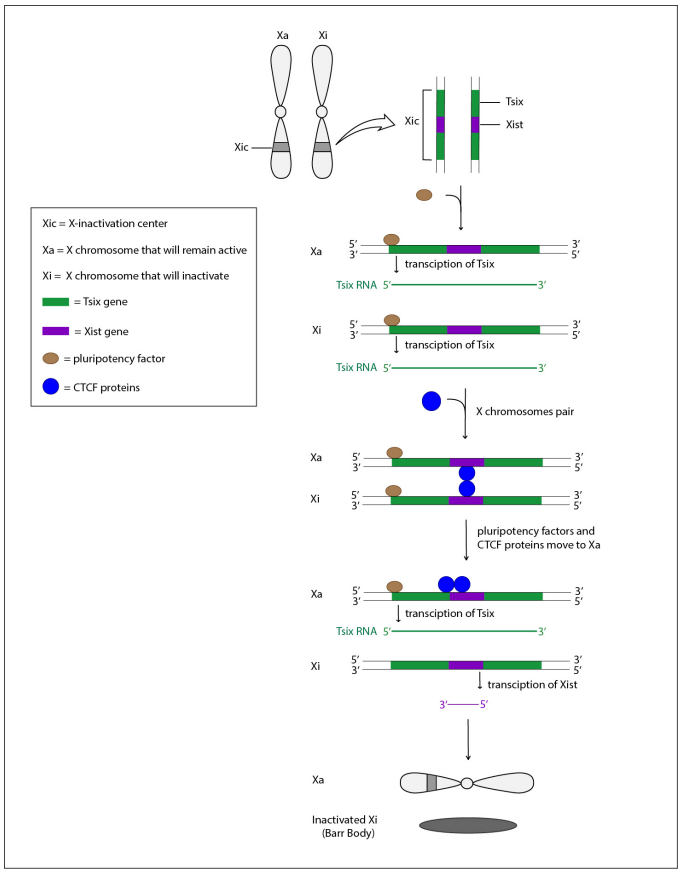


Figure 15.6 - X chromosome Inactivation (XCI).

Key Questions

- How do pluripotency factor proteins and CTCF proteins contribute to XCI?
- What is the function of the *Tsix* RNA?
- How is the *Xist* gene activated?
- How does the *Xist* RNA contribute to the formation of a Barr body?

Fragile X Syndrome

Methylation of CpG sites on the X chromosome contributes to **fragile X syndrome**. Fragile X syndrome is the most common form of inherited mental retardation, affecting 1 in 4000 males and 1 in 8000 females. Fragile X syndrome is named because of a site on the X chromosome that looks like a gap (see **figure 15.7**). This gap region tends to break and is therefore called a **fragile site**.

A **trinucleotide repeat expansion** (**TNRE**) mutation also contributes to fragile X syndrome. In the TNRE that causes fragile X syndrome, the number of copies of a 5'-CGG-3' sequence increases from generation to generation due to DNA polymerase slippage during DNA replication. When the number of 5'-CGG-3' trinucleotides exceeds 230 copies, disease symptoms are produced. Recall that a similar TNRE mutation is responsible for Huntington's disease (see Part 7). In fragile X syndrome, the 5'-CGG-3' trinucleotide repeats on the X chromosome are found in the first exon of the ***FMR1*** gene. This beginning region of *FMR1* is transcribed to produce the 5'-UTR in the mRNA. The expansion of the

trinucleotide repeat is thought to form multiple CpG sites within exon 1 of the *FMR1* gene. These numerous CpG sites near the beginning of the *FMR1* gene can become hypermethylated, silencing the transcription of the *FMR1* gene. Since the protein product of the *FMR1* gene is known to be expressed in the brain, the silencing of the *FMR1* gene is thought to prevent protein production, leading to disease symptoms.

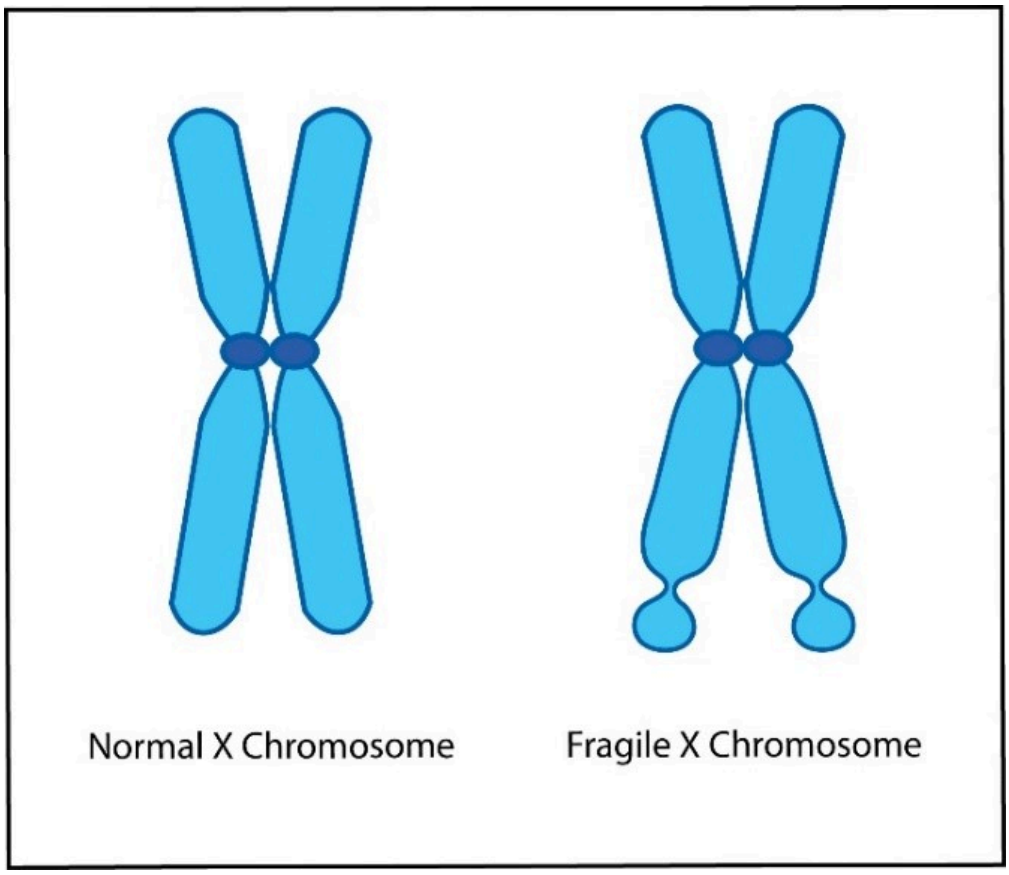


Figure 15.7 ***Top-Fragile X Chromosome Structure.*** *Image created by SL.* ***Bottom-Fragile X Syndrome Patient.*** *Image by Peter Saxon and used under license CC BY-SA 4.0*

Key Questions

- What is a TNRE?
- How does TNRE and DNA methylation contribute to fragile X syndrome?

Epigenetics in Embryonic Development

Now let's learn how histone modifications contribute to epigenetics. Epigenetic processes are critical in the embryonic development of multicellular organisms. Embryonic development, starting with a fertilized egg and eventually producing an entire adult organism, initiates with the activation of genes that produce the overall body plan. For example, a group of genes called ***Hox*** specify the structures that form on the anterior and posterior portions of the body. The *Hox* genes are actively transcribed during embryonic development when body parts are forming; *Hox* gene transcription is not needed in the adult organism. Epigenetic factors permanently silence *Hox* genes after the *Hox* protein products have been used to help form the body plan.

Further, epigenetic processes that occur in development ensure that the different cell types in the body have specific phenotypes. For instance, muscle-specific genes are actively transcribed in muscle cells, whereas genes that specify another fate (neuron, epithelial cell) are permanently silenced in muscle cells. Two protein complexes, called the **trithorax group (TrxG)** and the **polycomb group (PcG)**, are thought to regulate the epigenetic changes that occur during embryonic development and the differentiation of cell types. The TrxG protein complex is involved in gene activation processes, while the PcG protein complex is involved in gene silencing processes. The TrxG and PcG complexes are

both **histone methyltransferases**, which accomplish epigenetic changes by adding methyl groups to the tail domains of histone H3. The TrxG complex recognizes histone H3 and adds three methyl groups (**trimethylation**) to a lysine amino acid at position 4 within the histone tail. Trimethylation of lysine 4 within histone H3 is an activating epigenetic mark. Alternatively, PcG recognizes histone H3 and adds three methyl groups to a lysine at position 27; this modification to lysine 27 is a silencing epigenetic mark. Note that inactive X chromosomes (i.e., Barr bodies) have abundant trimethylation of histone H3 at lysine 27 (see above).

We will now consider how a PcG complex silences transcription, by describing how the *Hox* genes are inactivated after the *Hox* proteins have been used to help determine the body plan during embryogenesis (see **figure 15.8**). The silencing of the Hox genes occurs by:

1. A **PRE-binding protein** binds to a **polycomb response element (PRE)** near the *Hox* genes. The PRE-binding protein is a repressor, while the PRE is a silencer DNA sequence.
2. The PRE-binding protein (repressor) recruits a PcG protein complex called **PRC2** to the promoter region of the *Hox* gene.
3. PRC2 trimethylates lysine 27 of histone H3 within multiple nucleosomes near the *Hox* promoter.
4. Trimethylation of histone H3 at lysine 27 inhibits transcription of the *Hox* gene by preventing TFIID and RNA polymerase II from binding to the *Hox* gene core promoter.
5. Transcription of the *Hox* gene is silenced.

It is important to note that the epigenetic silencing of the *Hox* gene is maintained during subsequent cell divisions, ensuring that the *Hox* genes remain silent in the adult organism.

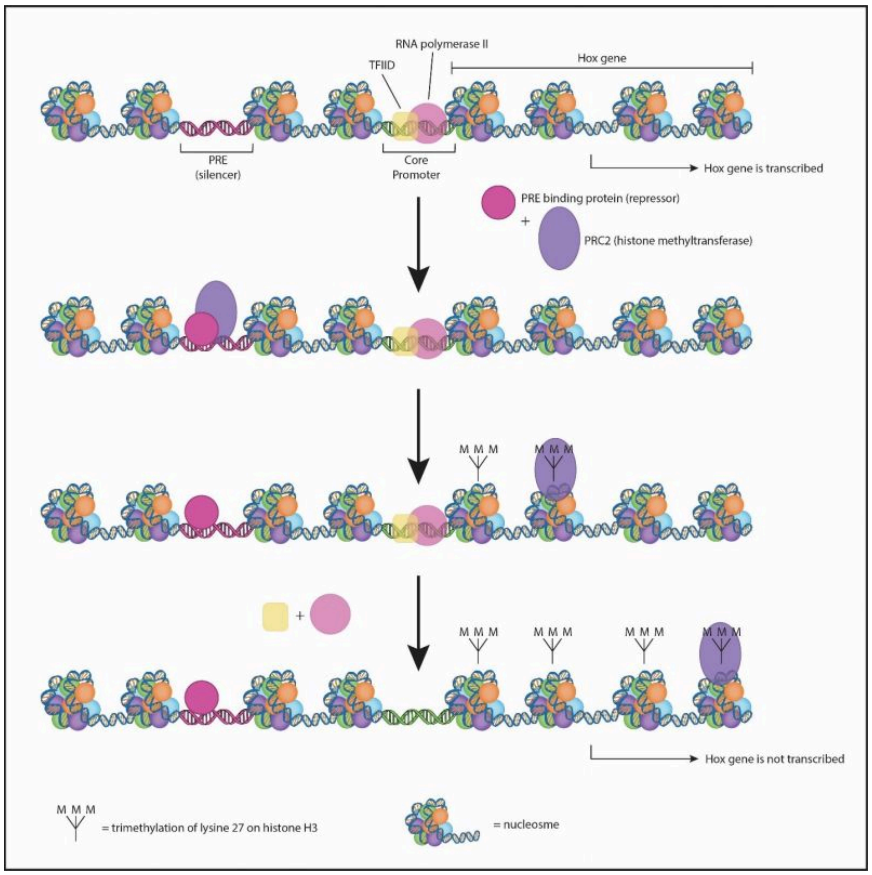


Figure 15.8 - ***PcG histone methyltransferase silences the Hox genes*** *--- image created by SL*

Key Questions

- What is the function of the *Hox* genes?
- How do the TrxG and PcG protein complexes contribute to embryonic development and tissue differentiation?
- Describe how trimethylation of histone H3 can lead to gene activation in some cases and to gene silencing in other cases.
- Describe how the PRE sequence, PRE-binding proteins, and the PRC2 protein complex contributes to the silencing of the *Hox* genes.

The *Agouti* Phenotype in Mice

One of the best examples of how environmental changes contributes to epigenetics involves the ***Agouti*** gene in mice. The protein product of the *Agouti* gene catalyzes yellow pigment formation in the hairs of developing mouse pups. The *Agouti* gene has three alleles in the population: *A*, *a*, and A^{vy}. If a mouse has the *AA* or *Aa* genotype, the mouse coat color is agouti (brown). If the hairs from an agouti mouse are examined closely, each hair contains a stripe of yellow pigment sandwiched between layers of black pigment. Thus *Agouti* mice have normal yellow pigment production. In mice that have the *aa* genotype, the mouse is black due to the inability to produce yellow pigment.

The A^{vy} allele results in the overexpression of the *Agouti* gene. Homozygous A^{vy} mice do not survive; however, if mice have AA^{vy} or $A^{vy}a$ genotypes, a variety of phenotypes are possible; some mice are yellow, some are mottled with black and yellow fur patches, and some are pseudoagouti with hairs that are mostly black with a little yellow pigment. The extent of the yellow fur color reflects the degree of A^{vy} allele expression. If the A^{vy} allele is highly overexpressed, then a yellow coat is produced. Intermediate levels of A^{vy} allele overexpression produces the mottled phenotype. If the A^{vy} allele displays low levels of overexpression, then the pseudo-agouti coat is produced. Interestingly, mice that have high levels of A^{vy} allele overexpression (yellow fur) are also prone to obesity, diabetes, and cancer (see **figure 15.9**). Mice with the *AA*, *Aa*, and *aa* genotypes are lean in appearance and are less susceptible to diabetes and cancer. Moreover, mice with lower A^{vy} allele overexpression (i.e., pseudoagouti) are also leaner and more resistant to diabetes and cancer.

Figure 15.9 Yellow and Wild Type Mice. *The yellow phenotype is the result of high A^{vy} allele overexpression. The wild-type mouse on the right has the AA genotype. Yellow mice have a higher incidence of obesity, diabetes, and cancer than wild-type mice. --- image provided by R. Jirtle and D. Dolinoy and used under license CC BY 3.0*

Key Questions

- What are the phenotypes of mice with high levels of A^{vy} allele overexpression?
- What are the phenotypes of mice with low levels of A^{vy} allele overexpression?

The *Agouti* Phenotype is Influenced by Diet and Bisphenol A

The variation in coat color phenotypes among A^{vy} heterozygotes can be partially explained by the diet of their mother during pregnancy. When pregnant female mice are fed a diet supplemented with the vitamins **folic acid** and **vitamin B_{12}**, the offspring that are heterozygous for the A^{vy} allele tend to have darker coats and are leaner compared to the heterozygous offspring of mice fed a diet that lacks these vitamins. Moreover, the offspring of pregnant mice fed the diet rich in folic acid and vitamin B_{12} had higher levels of CpG island methylation adjacent to the A^{vy} allele than the heterozygous offspring of mice fed a non-supplemented diet. These results suggest that supplementing the diets of pregnant mice with folic acid and vitamin B_{12} increases DNA methylation near the A^{vy} allele in the offspring, leading to decreased A^{vy} allele overexpression and decreased risk of obesity and cancer. Importantly, these results showed that the environment of a mother mouse (eating a diet supplemented with folic acid and vitamin B_{12}) influences the expression of a gene in her pups.

Another environmental agent that affects the *Agouti* phenotype is the chemical **bisphenol A (BPA)**, a chemical found in many plastics, including plastics that were at one time commonplace in water bottles. The exposure of pregnant female

mice to BPA produces more A^{vy} heterozygote offspring that have yellow coats, obesity, and cancer compared to the heterozygous offspring of mice not exposed to BPA. BPA is thought to inhibit the DNA methylation process, resulting in low levels of CpG island methylation near the A^{vy} allele. As a result, the A^{vy} allele is overexpressed in these mice, producing yellow coats, obesity, and cancer. Incidentally, the addition of folic acid and vitamin B_{12} to the diet of these pregnant mice counteracted the negative effect of BPA. Again, the environment of the mother mouse influenced the expression of a gene in her pups.

Key Questions

- How does the consumption of folic acid and vitamin B_{12} by a pregnant mouse influence methylation of CpG islands, the expression of the A^{vy} allele, and the phenotype of her heterozygous offspring?
- How does the exposure of a pregnant mouse to BPA influence methylation of CpG islands, the expression of the A^{vy} allele, and the phenotype of her heterozygous offspring?

Epigenetics and Cancer

Cancer is a condition characterized by uncontrolled cell division. Multiple mutations are typically required to convert a normal cell into a cancerous cell. If some of these mutations occur in the structural genes involved in DNA methylation, histones acetylation, or histone methylation, the epigenetic markings of many genes are altered. As a result, these mutations producing genes that are not regulated correctly; these genes are either overactive or not expressed sufficiently. For example, higher than normal expression of **oncogenes** can result in higher rates of cell division, promoting cancer. Alternatively, lower than normal expression of cancer-preventing **tumor-suppressor genes** can also promote the formation of cancer.

Mutations in the genes involved in DNA methylation have been associated with certain cancers. For example, mutations in the gene that produces DNA methyltransferase have been associated with acute myeloid leukemia. Note that the result of these mutations would be decreased methylation of the CpG islands adjacent to many genes, including oncogenes. As a result, these mutations lead to higher oncogene expression and higher rates of cell division.

Mutations in the genes involved in histone modifications have also been linked to certain cancers. For example, mutations in the genes that produce histone acetyltransferases (HATs) have been associated with colorectal, breast, and pancreatic cancer. In this case, the defective HAT would result in lower expression of many genes, including tumor-suppressor genes. Since tumor suppressor genes encode repressor proteins that silence cancer genes, the overall effect is a higher rate of cancer formation.

Finally, certain chemicals are known to produce the epigenetic changes associated with cancer. For example, the **polycyclic aromatic hydrocarbons (PAHs)** found in tobacco smoke are associated with lung, breast, stomach, and skin cancer. These PAHs are thought to contribute to cancer by altering the DNA methylation patterns adjacent to many genes, including oncogenes and tumor-suppressor genes.

Key Questions

- What is an oncogene and a tumor-suppressor gene?
- Explain how mutations in the DNA can have epigenetic consequences, potentially leading to cancer.
- What are PAHs?

Epigenetic Therapy

As we have seen, epigenetic processes are associated with several human diseases (AS, PWS, fragile X syndrome, acute myeloid leukemia, colorectal cancer). Scientists and physicians are interested in the possibility of treating diseases by converting the abnormal methylation or acetylation patterns in diseased cells back to the normal state. These types of restorative changes in epigenetic patterns are called **epigenetic therapy**. Inhibiting the DNA methylation process could reactivate silenced tumor suppressor genes in some cancers. One way to do this is to use DNA methyltransferase inhibitors, such as **5-azacytidine**. Similarly, histone deacetylases (HDACs) remove acetyl groups from histone tails, potentially silencing tumor suppressor genes. HDAC inhibitors, such as **phenylbutyric acid**, could reverse this effect, activating the silenced genes.

Key Questions

- What is meant by epigenetic therapy?
- How do 5-azacytidine and phenylbutyric acid contribute to epigenetic therapy?

The Epigenomes of Identical Twins are Not Identical

Identical twins have the same DNA sequences. They also have the same epigenetic markings in their genome (**epigenome**) when they are born. However, beginning at birth, the epigenetic processes in the twins behave independently of each other, so that later in life, the twins have very different epigenomes. Twin studies suggest that environmental factors influence the epigenetic patterns within the genome and may explain why individuals with the same DNA sequences do not necessarily have the same overall phenotype. For example, if one twin has been diagnosed with schizophrenia, the identical twin has only a 40-50% chance of having schizophrenia, despite the twins having the same DNA sequences. This difference in phenotype may be explained by the different environmental factors encountered by each twin during their lifetime, resulting in distinctive epigenetic markings in each genome.

Another example of how the environment can influence the epigenomes of identical twins involves Scott and Mark Kelly (see **figure 15.11**). Scott spent a year on the International Space Station, while his twin brother Mark stayed home. A recent NASA study showed that time in space altered the expression of many of Scott's genes, including those genes involved in response to hypoxia (oxygen depletion) and inflammation. Moreover, the expression of approximately 7% of Scott's genes has not returned to baseline levels even after spending several years on Earth since his time on the space station. This alteration in Scott's gene expression is thought to be the result of epigenetic changes resulting from the unique environment of outer space.

Figure 15.11 ***Astronauts Scott and Mark Kelly ---*** *photo taken by NASA*

Key Questions

- What is an epigenome?
- Are the epigenomes of identical twins the same? Why or why not?

16 - Genome Editing

Genome editing allows scientists to introduce targeted changes to the DNA of an isolated cell or an entire organism. For example, scientists can insert a DNA sequence, delete a DNA sequence, or modify the DNA sequence of any gene via genome editing. The goal of genome editing is to change the phenotype of a cell in a way controlled by the researcher.

The applications of genome editing are staggering. Genome editing can be used in research to better understand the role of a gene and its protein product in cellular structure and function. Genome editing can also be used as a treatment for genetic diseases by replacing a mutant gene that causes the disease with the normal (wild-type) version of the gene. Finally, genome editing can be used to enhance the yield of crops or give desirable traits to livestock.

Key Questions

- What is meant by genome editing?
- What are some of the applications of genome editing?

Genome editing (overview)

Genome editing works by recognizing a specific **target DNA sequence** in the genome. After recognition of the target DNA sequence, an **endonuclease** cuts both DNA strands. The cell then tries to fix the double-stranded DNA break by rejoining the two ends of the severed DNA molecule; however, the repair mechanisms involved are error-prone, introducing extra nucleotides or deleting nucleotides at the cut site. The insertion or deletion of a single base or two bases within the coding region of a gene changes every codon downstream of this insertion/deletion (**indel**) site. This type of mutation, referred to as a **frameshift mutation**, produces a defective protein product (see Part 7).

Key Questions

- How does genome editing work?
- What is meant by an indel mutation and a frameshift mutation?

Genome editing systems

There are three major genetic technologies that can be used to edit DNA sequences within isolated cells or entire organisms:

- **Zinc-finger nucleases (ZFNs).** ZFNs are enzymes engineered in the lab to contain two parts: a **zinc-finger motif** and an endonuclease. The zinc-finger motif allows the ZFN to bind to the target DNA sequence (see Part 14). One ZFN attaches to one DNA strand at the target site; a second ZFN binds to the other DNA strand about ten base pairs away. The endonucleases come together and cut both DNA strands between the ZFN binding sites. Because the ZFN binds and then cuts a specific target DNA sequence, ZFNs only create a single genome edit at a time.
- **Transcription activator-like effector nucleases (TALENs).** Like the ZFNs, TALENs are enzymes designed by researchers to include both a DNA-binding region and an endonuclease region. The TALEN DNA-binding protein domain can be engineered to bind to any target DNA sequence. Once bound to the DNA, the TALEN endonuclease domain cuts both strands of the target DNA sequence, allowing the creation of a single genome edit at a time.
- **CRISPR-Cas9.** The CRISPR-Cas9 system is the newest, most powerful, and versatile genome editing technique. CRISPR-Cas9 can be used to create a single genome edit or multiple genome edits simultaneously.

ZFNs and TALENs have many drawbacks, including the high cost and time involved in engineering the DNA binding domains within the nucleases and the inefficient cutting of the target DNA sequence. Although the ZFNs and TALENs have been used to successfully edit genes, the science world has embraced CRISPR-Cas9 due to its lower cost, higher efficiency, and potential to create multiple genome edits simultaneously. Because of its widespread current use and promising future, CRISPR-Cas9 will serve as the subject for the remainder of this chapter.

Key Questions

- Describe the three major genome editing technologies.
- Why do scientists prefer CRISPR-Cas9?

The CRISPR-Cas9 genome editing system

CRISPR is an acronym for the **clustered regularly interspaced short palindromic repeats (CRISPR)** system. The CRISPR-Cas9 system has two molecular components (see **figure 16.1**):

- A **single guide RNA (sgRNA)** The sgRNA consists of a single-stranded RNA molecule called **crRNA** that forms hydrogen bonds with a specific target DNA sequence. The crRNA is covalently linked to a stem-loop RNA sequence called **tracrRNA.** The tracrRNA binds to and activates the Cas9 endonuclease to cut the double-stranded DNA at the target site.
- A **CRISPR-associated endonuclease protein (Cas)**. The Cas protein is a non-specific endonuclease that cuts double-stranded DNA when activated by the tracrRNA. The genome editing system described below uses the **Cas9** enzyme isolated from the bacterium *Streptococcus pyogenes*.

Additionally, the DNA sequence targeted by the sgRNA needs to contain a **protospacer adjacent motif (PAM)** sequence, as Cas9 binds to the PAM sequence to position itself while it cuts both strands of the DNA. The PAM sequence is a DNA consensus sequence consisting of 5'-NGG-3', where N is any of the four DNA bases (A, T, C, or G). The PAM sequence is in the **nontarget DNA strand**; the nontarget DNA strand does not form hydrogen bonds with the crRNA component within the sgRNA. The PAM in the nontarget DNA strand is located 3-4 nucleotides in the 3’ direction (downstream) from the site that will be cut by Cas9.

The CRISPR-Cas9 system creates genome edits as follows:

1. Cas9 binds to the PAM sequence in the nontarget DNA strand.
2. The target and nontarget DNA strands are separated from each other. The Cas9 enzyme is the helicase that separates the two DNA strands.
3. The crRNA attempts to form hydrogen bonds with the target DNA strand. If the crRNA forms proper hydrogen bonds with the target DNA strand, then genome editing continues. If hydrogen bonds fail to form, the Cas9 enzyme is released and binds to another PAM sequence in the genome.
4. The binding of the crRNA to the target DNA strand activates tracrRNA, which in turn, activates Cas9.
5. Cas9 enzyme cuts both DNA strands 3-4 nucleotides in the 5' direction (along the nontarget strand) from the PAM site.
6. Once both strands of the DNA have been cut by Cas9, the cell's DNA repair systems attempt to fix the break in the dsDNA and, in doing so, add a few bases, delete a few bases, or insert a completely new piece of DNA.

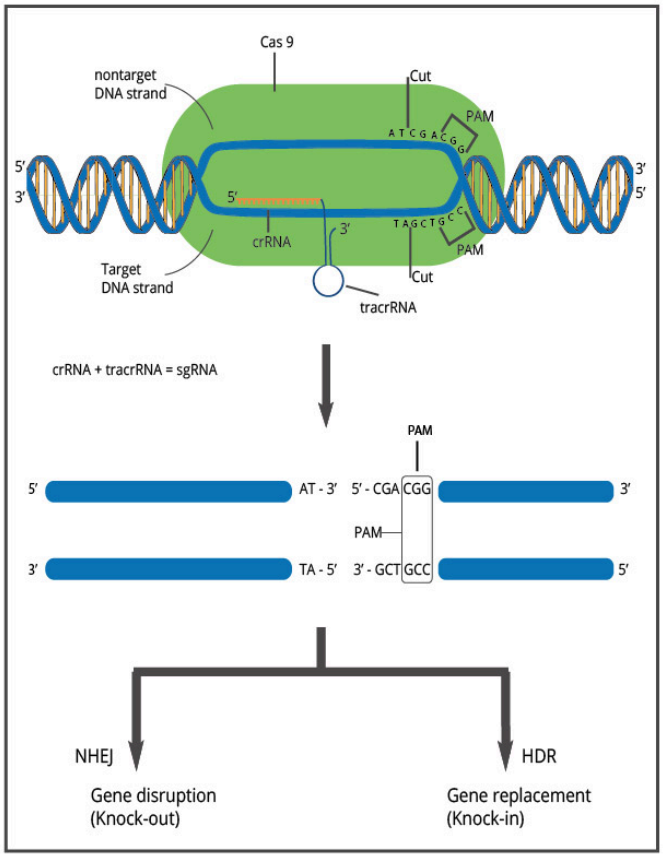


Figure 16.1 ***The CRISPR-Cas9 System*** *--- Image created by SL*

Key Questions

- What is meant by the target and nontarget DNA strands?
- What are the names of the two components within a sgRNA molecule?
- Describe how crRNA, tracrRNA, Cas9, and the PAM contribute to the CRISPR-Cas9 genome editing system.

What is the natural function of CRISPR-Cas9?

The CRISPR-Cas9 system is thought to be analogous to an immune system, protecting bacteria against invading bacteriophages (viruses that infect bacteria). During an infection, the bacteriophage genome is injected into the cytoplasm of the bacterial cell. The bacteriophage DNA is cut by nucleases, and a portion of the bacteriophage genome

is stored in the **CRISPR gene locus**. Overall, the CRISPR gene locus in bacteria consists of clusters of repetitive DNA sequences (short palindromic repeats that are 30-40 base pairs in length) separated by bacteriophage DNA sequences called **spacers**. In essence, the spacer sequences within the CRISPR locus are a library of previous bacteriophage infections (see **Figure 16.2**).

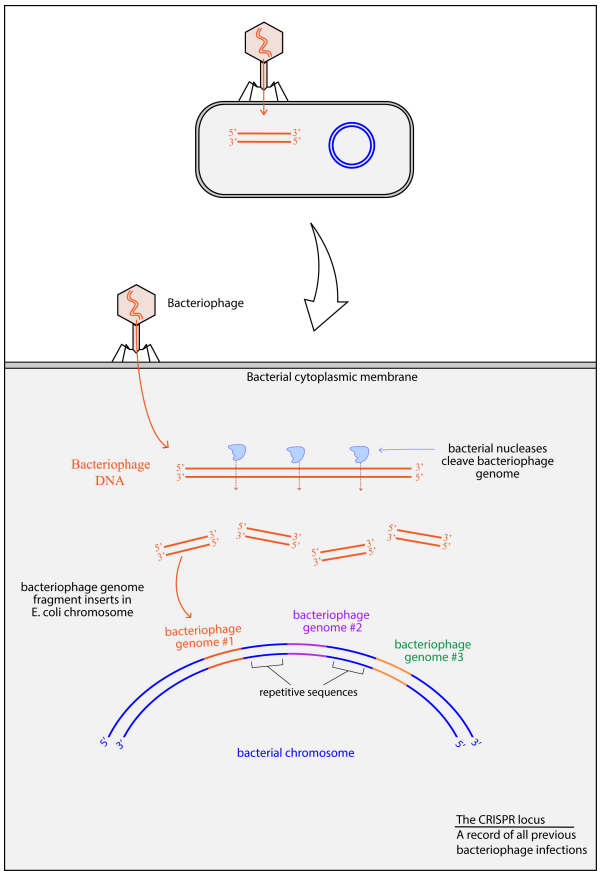


Figure 16.2 - The CRISPR locus is a library of previous bacteriophage infections. *Fragments of bacteriophage genomes are stored as spacers in the bacterial chromosome. Image created by Alex Baff.*

Upon reinfection with the same bacteriophage, the CRISPR gene locus is transcribed to produce two types of RNA molecules (see **Figure 16.3**). The spacer DNA sequence (i.e., the bacteriophage genome) is transcribed to produce the single-stranded CRISPR RNA **(crRNA)** to form hydrogen bonds with the DNA of the infecting bacteriophage. Another gene in the CRISPR locus is transcribed to make the **transactivating crRNA (tracrRNA)**. The crRNA and the tracrRNA from hydrogen bonds with each other and then bind to the Cas9 endonuclease. Note that the tracrRNA contains the stem-loop that activates Cas9. The crRNA:tracrRNA:Cas9 complex then binds to a PAM sequence in the DNA of the invading bacteriophage. The two DNA strands within the bacteriophage DNA are separated and the crRNA forms hydrogen bonds with the **target DNA strand**, while the **nontarget DNA strand** is moved out of the way. Finally, the Cas9 protein makes **double-stranded breaks** (DSB) in the DNA of the bacteriophage, thereby destroying the bacteriophage genome and inhibiting the bacteriophage infection.

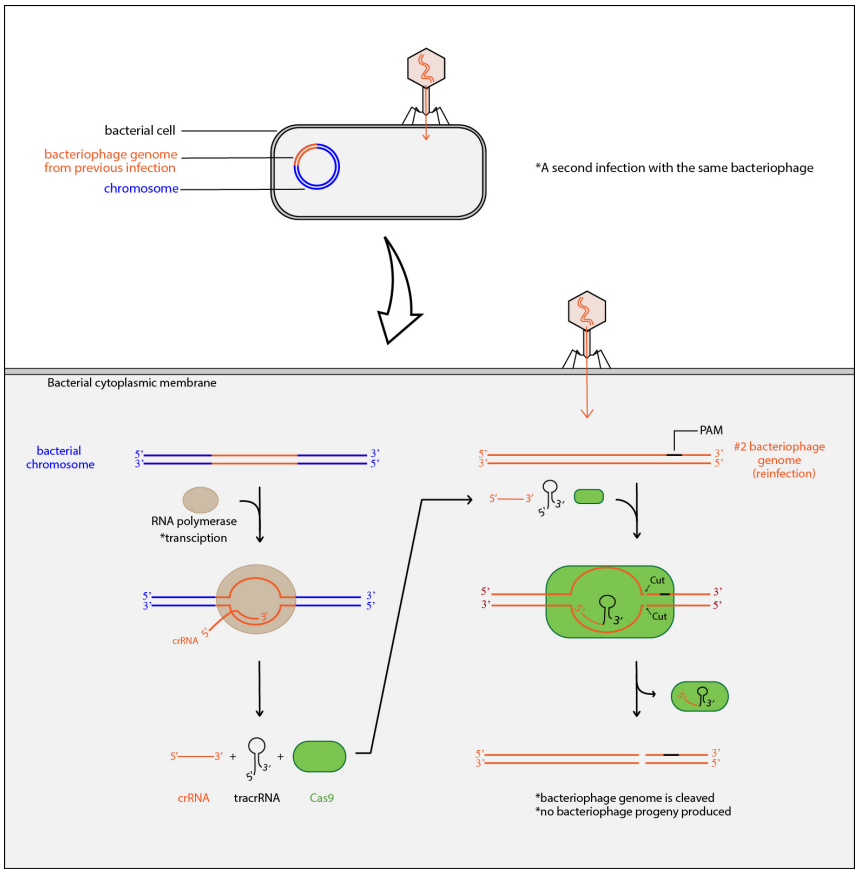


Figure 16.3 - The CRISPR-Cas9 destroys the bacteriophage genome upon reinfection. *Image created by Alex Baff.*

Key Question

- How is the CRISPR-Cas9 system beneficial to a bacterial cell?
- How are spacers generated?
- Describe how the CRISPR-Cas9 system destroys the DNA of an invading bacteriophage.

Applications of CRISPR-Cas9

Genome editing via CRISPR-Cas9 involves designing a 20 nucleotide-long crRNA sequence that forms hydrogen bonds with a target DNA sequence of interest. This crRNA is covalently linked to the tracrRNA that forms the RNA stem-loop to activate Cas9. The crRNA linked to the tracrRNA is the **sgRNA** component of the CRISPR-Cas9 system. Both the sgRNA and Cas9 DNA sequences are ligated into separate cloning sites within a plasmid vector, and the plasmid vector is introduced into a cell of interest, including a eukaryotic cell. Transcription of the cloned genes leads to the production of both the sgRNA and Cas9 molecules.

When the sgRNA binds to a target DNA sequence, Cas9 produces a **double-stranded break (DSB)** in the DNA. When the cell attempts to repair these DSBs, the cell can undergo the **non-homologous end joining (NHEJ)** DNA repair pathway. NHEJ is not perfect, and insertion or deletion of a few nucleotides occurs (these mutations are called **indels**) as the DSB is repaired. Recall that indels cause a frameshift during translation that ultimately prevents the eukaryotic gene

from making a functional protein product. Therefore, CRISPR-Cas9 genome editing followed by NHEJ allows the researcher to produce a gene **knock-out** cell line or organism. The knock-out fails to produce a functional protein product.

Double strand breaks in the DNA can also lead to another type of DNA repair known as **homology directed repair (HDR)**. In this case, DNA repair allows the insertion of a **donor sequence** at the location of the DSB, instead of repairing the break by inserting or deleting a few nucleotides. The donor DNA can be engineered to contain a mutant form of a gene. This approach allows the researcher to insert a mutant gene in the place of a wild-type gene to study the effects of the mutation on the cell. Alternatively, the donor DNA sequence can contain a wild-type version of a gene that replaces the mutant form of the gene within the cell. The replacement of a gene with a different allele of the same gene produces a **knock-in** cell.

CRISPR-Cas9 is a convenient genome editing system to use because if a scientist wishes to study a different gene, the scientist designs a new 20 nucleotide-long crRNA that forms hydrogen bonds with the new target gene, all of the other components (i.e., tracrRNA, Cas9) of the CRISPR-Cas9 system remain the same. Moreover, the use of multiple unique crRNA sequences allows the alteration of several genes in the genome simultaneously.

Key Questions

- How is gene cloning used in genome editing?
- Describe how NHEJ can be used to create a knock-out cell.
- Describe how HDR can be used to create a knock-in cell.

Challenges associated with CRISPR-Cas9

Before we explore the ethics of genome editing, let us investigate some of the challenges of using the CRISPR-Cas9 system. In a typical experiment, the researcher will introduce the CRISPR-Cas9 vector into a population of eukaryotic cells. Because the process of genome editing is inefficient, the experiment will result in three groups of cells in the population: those in which no editing occurred, those in which one of the two alleles of a gene is edited, and those with both alleles edited. If the knock-out approach is used, the researcher will want to study cells that have no functional copies of the gene; therefore, the researcher will need to identify those cells with both alleles edited. Determining the DNA sequence of the target gene in individual cells is one of the easiest ways to confirm that the desired changes have taken place.

The crRNA is designed to target a specific gene in the genome; however, sometimes a 20 nucleotide-long crRNA can bind to more than one DNA sequence in the genome simultaneously. This raises the possibility that the CRISPR-Cas9 system will cut the DNA at undesired locations within the genome, producing **off-target effects**. Because the locations of these off-target cut sites are difficult to predict, treating cells with CRISPR-Cas9 can have unintended consequences on the cell or organism.

Key Questions

- What are two challenges associated with CRISPR-Cas9 genome editing?

The ethics of genome editing

Many scientists are interested in using CRISPR-Cas9 to treat human genetic diseases, especially diseases for which there is currently no treatment. There are two main ways that human genome editing can be used to treat disease: inactivation of a mutant gene to remove its effects on the cell (using the NHEJ knock-out approach) or insertion of a functional allele to replace a mutant one (using the HDR knock-in approach).

With the promise that CRISPR-Cas9 brings, there is also uncertainty about the ethics of this technique, particularly when applied to humans. Most researchers agree that if we have the tools necessary to treat a genetic disease, we should use those tools to improve the lives of patients. However, considerable disagreement exists as to whether the CRISPR-Cas9 technique should be used to modify the germ-line cells that produce gametes or embryos. Current laws in the United States prohibit the use of human genome editing in gamete-producing cells. Research on human embryos is permitted if the treated human embryos are destroyed before day 14 of development and are not implanted into the womb.

An important issue to consider with genome editing is that of informed consent. An adult can give consent for genome editing that can potentially treat their genetic disease, but when that treatment extends to future generations, there is no way to obtain consent (i.e., the developing fetus cannot give consent). Public opinion remains divided as to who has the right to make the decision for the fetus; is it the person who develops from the embryo, parents, or the government?

In November 2018, the press announced that a researcher in China used CRISPR-Cas9 to successfully edit the *CCR5* gene in human twins (one received the edit while the other did not). Knocking out this gene is expected to prevent the treated child from contracting a human immunodeficiency virus (HIV) infection. To say the scientific world was upset about this announcement is an understatement. This was the first time that a human baby was born after genome editing was performed. The reason why this announcement was not received with congratulations was, in part, due to the lack of informed consent and the failure to make sure that no off-target effects took place before implanting the embryos into the birth mother. In fact, it is uncertain if the parents were informed as to the genome editing experiment, or if they were coerced into giving their consent.

There may never be an international agreement concerning genome editing that can be enforced by all nations. Even when there is agreement as to what is ethical and what is not, there will always be individuals or nations who will carry out research that is contrary to the moral beliefs of others. Important questions to consider include how do we establish laws concerning the ethical practice of scientific research, and how do we penalize those who knowingly disobey those laws?

In 1956, mathematician and biologist Jacob Bronowski wrote that, as scientists, “We ought to act in such a way that what is true can be verified to be so,” an expression of his belief that it is our right and our duty to explore the unknown and seek truth. The point is that maybe having large international committees decide what should be practiced and what should be prohibited is not the real question, but rather how can society ensure that research done is based on the priniciple of seeking truth to better the lives of humankind?

Key Questions

- Should genome editing be done on gamete-producing cells, embryonic cells, or somatic cells? Why or why not?