Chromosomes, Genes, and Traits: An Introduction to Genetics [Revised Edition]

CHROMOSOMES, GENES, AND TRAITS: AN INTRODUCTION TO GENETICS [REVISED EDITION]

AMANDA SIMONS

ROTEL (Remixing Open Textbooks with an Equity Lens) Project
Framingham, MA

ISBN 978-1-964276-08-3 (Print)
ISBN 978-1-964276-09-0 (Ebook)

CONTENTS

Part IV. Overview of Central Dogma and Replication

Part V. Transcription

Part VI. Translation and genetic code

Part VII. Mutation

Amanda Simons

Part VIII. Overview of Gene Regulation

Part IX. Eukaryotic Gene Regulation in Action: Examples from Development

Part X. Mendel and Basic Heredity

Part XI. Allele Interactions

Part XII. Multigenic Inheritance

Part XIII. Genetics of Sex

Part XIV. Family Trees and Pedigrees

Part XV. Linkage and Mapping

Part XVI. Epigenetics

Part XVII. Evolution and Ancestry

Part XVIII. DNA repair and cancer

INTRODUCTION

Amanda Simons

This is the second edition of an OER intended to serve as a starting point for an introductory Genetics course, funded by the ROTEL project, a multi-institutional collaboration among six Massachusetts colleges and universities.

This text includes foundational information about both transmission and molecular genetics. Compared to the first edition, there are additional chapters on gene regulation in eukaryotes, cancer biology, and evolution and ancestry. Note that the text intentionally does not include much detail about current techniques in molecular genetics or genomics. Those things change so rapidly, chapters would be out of date within a few semesters! This OER is meant to be supplemented with current resources for such topics.

You can access each chapter via the contents menu at the top left. Clicking the “+” after each chapter heading in the contents menu will let you navigate directly to subsections of each chapter. As you are reading through each chapter, you can navigate to the next section by clicking the “Next” button in the dark blue ribbon at the bottom right of each screen. There is also a button to navigate back to the previous section.Each chapter begins with a list of objectives. This is what you are meant to take away from the reading. As you read, think about how the text relates to those objectives.

Within most chapters are interactive “Test Your Understanding” questions. These are meant to help you reflect on important parts of the text and practice doing the things listed in the objectives. If you click though, you will get an immediate answer.At the end of each chapter are wrap-up questions that are intended to help you engage more deeply with the text and objectives. Many of these questions are open-ended, and answers are not provided. But they can be useful study tools, and they might spark interesting conversations with your instructor!

This is a work in progress. If you find broken links or typos, please let your instructor know so they can be corrected for the next edition.

LAND ACKNOWLEDGEMENT STATEMENT FOR THE ROTEL GRANT

As part of ROTEL Grant's mission to support the creation, management, and dissemination of culturally-relevant textbooks, we must acknowledge Indigenous Peoples as the traditional stewards of the land, and the enduring relationship that exists between them and their traditional territories. We acknowledge that the boundaries that created Massachusetts were arbitrary and a product of the settlers. We honor the land on which the Higher Education Institutions of the Commonwealth of Massachusetts are sited as the traditional territory of tribal nations. We acknowledge the painful history of genocide and forced removal from their territory, and other atrocities connected with colonization. We honor and respect the many diverse indigenous people connected to this land on which we gather, and our acknowledgement is one action we can take to correct the stories and practices that erase Indigenous People's history and culture.

Identified Tribes and/or Nations of Massachusetts

Historical Nations

- Mahican
- Mashpee
- Massachuset
- Nauset
- Nipmuc
- Pennacook
- Pocomtuc
- Stockbridge
- Wampanoag

Present-Day Nations and Tribes

- Mashpee Wampanoag Tribe
- Wampanoag Tribe of Gay Head Aquinnah
- Herring Pond Wampanoag Tribe
- Assawompsett-Nemasket Band of Wampanoags
- Pocasset Wampanoag of the Pokanoket Nation
- Pacasset Wampanoag Tribe
- Seaconke Wampanoag Tribe

- Chappaquiddick Tribe of the Wampanoag Indian Nation
- Nipmuc Nation (Bands include the Hassanamisco, Natick)
- Nipmuck Tribal Council of Chaubunagungamaug
- Massachusett Tribe at Ponkapoag

At the time of publication, the links above were all active.

Suggested Readings

Massachusetts Center for Native American Awareness

A guide to Indigenous land acknowledgment

'We are all on Native Land: A conversation about Land Acknowledgements' (YouTube video)

Native-Land.ca | Our home on native land (mapping of native lands)

Beyond territorial acknowledgments – âpihtawikosisân

Your Territorial Acknowledgment Is Not Enough

This land acknowledgement was based on the land acknowledgement of the Digital Commonwealth.

PART I
DNA STRUCTURE

Objectives

1. Describe the structure of DNA. Identify the following: Base, nucleotide, 5'end, 3'end, phosphate, deoxyribose, ribose, major groove, minor groove, double helix.
2. Compare and contrast the chemical structures of DNA and RNA.
3. Predict the sequence of one strand of DNA given the sequence of the complementary strand.
4. Recognize the difference between DNA, chromatin, chromatid, chromosome, and genome.
5. Describe how proteins are used to compact eukaryotic chromosomes.

Source material

Nucleic Acids and DNA Double Helical Structure modified from Openstax Biology, 2e, by Mary Ann Clark, Matthew Douglas, and Jung Choi. Selections from Chapter 3.5 (Nucleic Acids) and Chapter 14.2 (DNA Structure and Sequencing). Access for free at OpenStax.

Chromatin compaction modified from Online Open Genetics (Nickle and Barrette-Ng), available through Biology LibreTexts.

Introduction

Genetics is the study of the inheritance of traits from parent to offspring, from generation to generation. Humans have long been aware that the characteristics of an individual plant or animal in a population could be passed down through the generations: Offspring tend to look more like their biological parents than unrelated adults. Humans also knew that some characteristics (such as the size or color of fruit) varied between

individuals and that agricultural crops and domestic animals could be bred for the most favorable traits. Knowledge of these hereditary properties has been of significant value in the history of human development. However, human understanding of *how* these traits are inherited is a relatively new science.

It all comes down to DNA: the genetic material.

A common saying in biology is "structure dictates function", or, sometimes, "form follows function". In everyday terms, this means that the structure of an object – be it a part of an organism, part of a cell, or a molecule – determines what that object can be used for. To use an analogy: a wheel must be round to roll.

Until the mid-1950s, it was very difficult for scientists to imagine how a chemically simple molecule like DNA could possibly store enough information to build a biologically complex organism like a human. And it was even more difficult to understand how that information could be shared from generation to generation. The structure of DNA, however, once understood, immediately suggested how genetic information could be transmitted. The structure of the molecule made it possible.

The story of DNA structure also illustrates how modern science works, with each generation of researchers building on the work of previous generations and their own contemporaries: Each new insight is dependent on what has come before. Although James Watson and Francis Crick are the most well-known names in the history of DNA structure, their insights were dependent on the work of many other researchers as well.

In this module, we will look at the chemical structure of DNA as well as a few key experiments in the decades of work that led up to the determination of DNA structure.

THE CHEMICAL NATURE OF NUCLEIC ACIDS

Nucleic acids serve as a biological information storage and retrieval system. The two main types of nucleic acids are **deoxyribonucleic acid (DNA)** and **ribonucleic acid (RNA)**. DNA and RNA are called nucleic acids because they were identified in the nucleus of cells, and they function as acids: part of the molecule gives up a proton (H^+) and is negatively charged at physiological pH.

DNA is the information storage system: it is the genetic material in all living organisms, ranging from single-celled bacteria to multicellular mammals. RNA primarily serves in the retrieval of information from DNA. RNA has several different functions in cells, most of which are involved in protein synthesis. For example, messenger RNA (mRNA) is used an intermediate between DNA and protein synthesis. Other types of RNA—like rRNA, tRNA, and microRNA—are involved in the process of protein synthesis and its regulation. RNA is also used by some viruses as genetic material: although viruses are not living, viruses do share many biochemical characteristics with living things.

Both DNA and RNA are **polymers**. The prefix "poly" means "many", and the word "polymer" is used to describe any molecule that is composed of a long chain of smaller building blocks. The individual subunits of any polymer are called "monomers". This is illustrated in Figure 1.

The monomers of DNA and RNA are called **nucleotides**. The generic structure of a nucleotide is shown in Figure 2. Three components comprise each nucleotide: a nitrogenous **base**, a pentose (five-carbon) **sugar**, and one or more **phosphate** group (one phosphate group is shown in the nucleotide in Figure 2). The nucleotides are linked together one after another to form a **polynucleotide**, and it is the polynucleotide that is referred to as DNA or RNA.

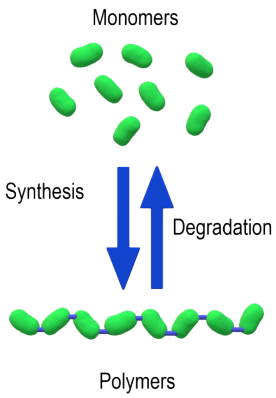


Figure 1. Polymers are long molecules synthesized from the chain-like linkage of smaller subunits called monomers.

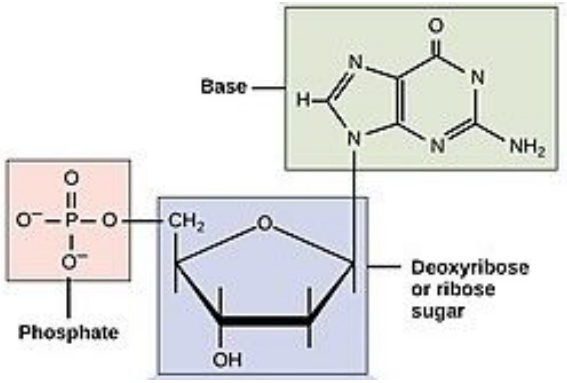


Figure 2. Structure of a nucleotide. Each nucleotide is comprised of three parts: one or more phosphates (one shown in pink here) linked to a sugar (shown in blue), which is attached to a nitrogenous base (shown in green).

Nucleic acids are, in fact, acids. Remember from chemistry that an acid is a molecule that can give up a proton (H+). Under acidic conditions, the oxygens of the phosphate are protonated (meaning they are bound to hydrogen to form -OH). Under physiological pH (around pH 7), the oxygens of the phosphate give up their protons to become negatively charged. You can see the negative charges on the oxygens of the pink-highlighted phosphate group in Figure 2. The polymers DNA and RNA are likewise negatively charged in the cell.

Please note: the structures shown in these images use an organic chemistry shorthand, where the "C's" representing carbon are not shown. Instead, if you see a "corner" in a structure with no atom indicated, you can assume there is a carbon at that position. Hydrogen atoms attached to carbons are also typically not shown unless they are specifically under discussion.

The nitrogenous base part of the nucleotide is highlighted in green. The nitrogenous bases contain nitrogen (thus giving them their name). They are bases because the nitrogen-containing groups (eg -NH_2) have a lone pair of electrons and can act as electron-pair donors. If the pH were decreased, those groups would accept a proton and become positively charged (-NH_3+). At physiological pH, though, the nitrogenous bases are uncharged.

Each nucleotide in DNA contains one of four possible nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T). RNA does not typically contain thymine. Instead, RNA contains the very similar base uracil, along with adenine, guanine, and cytosine. Uracil and thymine differ only by an extra methyl (-CH_3) group. The structures of these five bases are shown in Figure 3.

Adenine and guanine are classified as **purines**. The purine's general structure is two carbon-nitrogen rings. Scientists classify cytosine, thymine, and uracil as **pyrimidines**, which have a single carbon-nitrogen ring as their general structure (Figure 2). Each of these basic carbon-nitrogen rings has different functional groups attached to it, which differentiate the bases from one another. In molecular biology shorthand, we know the

nitrogenous bases by their symbols A, T, G, C, and U. DNA contains A, T, G, and C; whereas RNA contains A, U, G, and C.

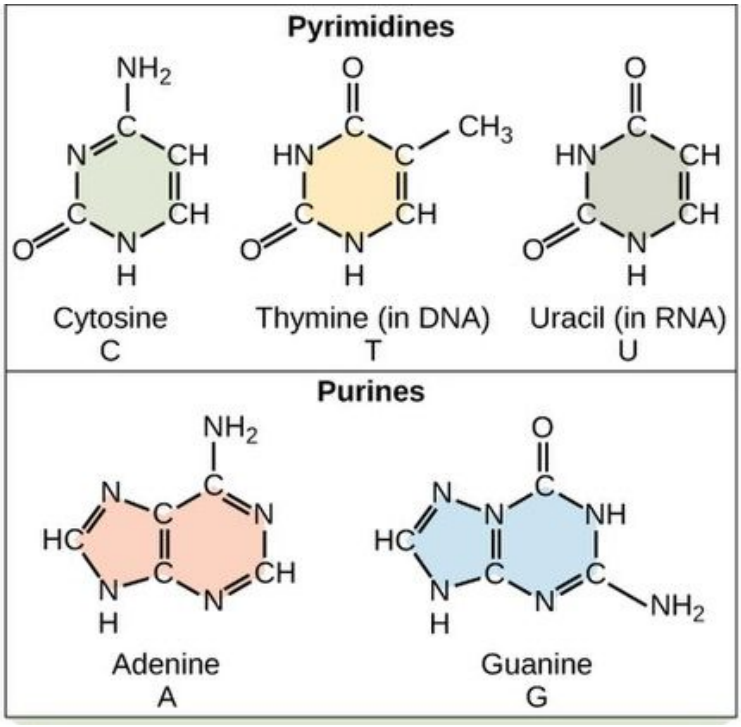


Figure 3. Nitrogenous bases include adenine, guanine, cytosine, thymine, and uracil. Adenine and guanine are composed of two carbon-nitrogen rings and are called purines. Cytosine, thymine, and uracil are single-ring structures called pyrimidines.

DNA and RNA nucleotides contain different sugars. The sugar in DNA is deoxyribose, and in RNA, the sugar is ribose. The difference between the sugars is the presence of the hydroxyl (-OH) group on the ribose's second carbon and hydrogen (-H) on the deoxyribose's second carbon. (Figure 4, bottom left).

To give scientists a way to talk about specific parts of the molecule, the carbon atoms of the sugar molecule are numbered as 1′, 2′, 3′, 4′, and 5′ (1′ is read as "one prime"). The prime distinguishes these atoms from those in the base, which are numbered without using a prime notation. As shown in Figure 4, the carbons are numbered, starting with the carbon attached to the base, which is 1'. Numbering continues around the ring in order, ending with 5' as the carbon linked to the phosphate. Note that only the 2', 3', and 5' carbons are highlighted in this image.

Given this numbering scheme, deoxyribose has an -H at the 2' position, while ribose has an -OH.

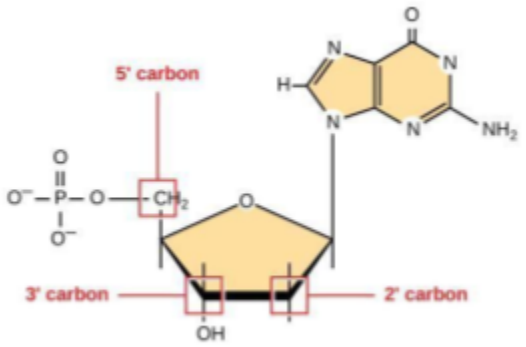


Figure 4. Sugar structure in DNA and RNA. The carbons of the sugar are numbered 1', 2', 3', 4', and 5', beginning with 1' nearest the base and 5' nearest the phosphate. DNA uses deoxyribose (with -H on the 2' carbon), and RNA uses ribose (with -OH on the 2' carbon).

Nucleotides are named in a way that indicates which base (adenine, guanine, thymine, cytosine, or uracil), which sugar, and how many phosphates are part of the molecule. Together, a base and a sugar make up a smaller unit that is called a **nucleoside**. *Nucleosides* containing adenine, guanine, cytosine, and thymine are called adenosine, guanosine, cytidine, and thymidine, respectively.

An RNA *nucleotide* containing adenine is **adenosine 5'monophosphate, adenosine 5'diphosphate,** or **adenosine 5'triphosphase**, depending on how many phosphate groups are part of the structure. These are often abbreviated AMP, ADP, and ATP for short. A DNA nucleotide containing adenine and three phosphate groups would be called 2'deoxyadenosine 5'triphosphate (or dATP for short). These structures are illustrated in Figure 5. The phosphate groups are named alpha (α), beta (β), and gamma (γ) for their proximity to the sugar, with the alpha phosphate closest to the sugar and the gamma farthest.

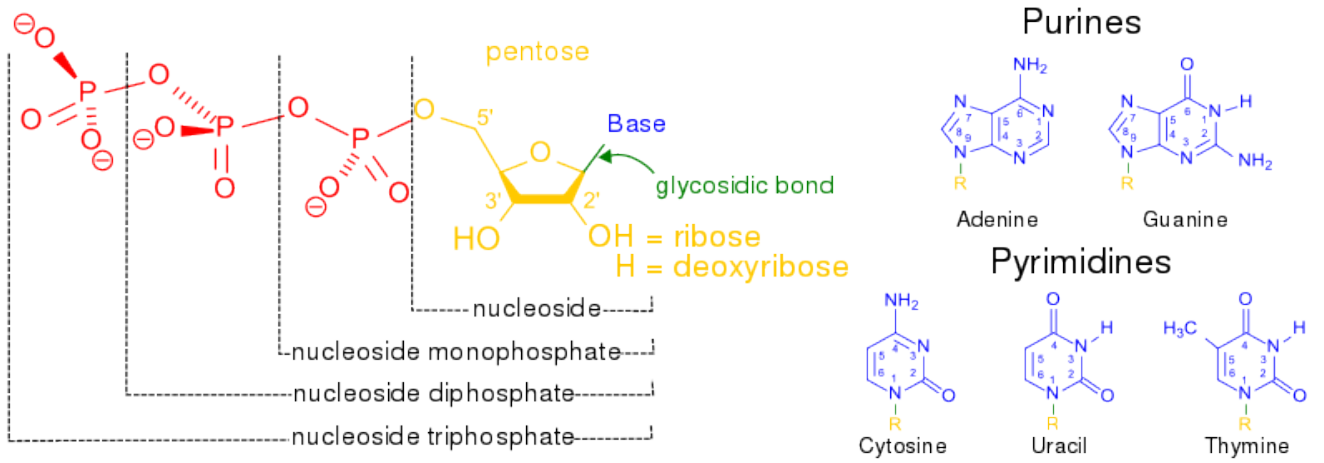


Figure 5. Nucleotide vs nucleoside structure.

The nucleotides combine with each other via phosphodiester bonds. The phosphate residue attached to the 5′ carbon of the sugar of one nucleotide forms a second ester linkage with the hydroxyl group of the 3′ carbon of the sugar of the next nucleotide in a dehydration reaction that forms a 5′-3′ **phosphodiester bond**. Once the nucleotides are linked, the dinucleotide will have a 5’ phosphate at one end, and a 3’-OH group at the other. A polynucleotide may have thousands of such phosphodiester linkages. Thus, in a polynucleotide, one end of the chain has a free 5′ phosphate, and the other end has a free 3′-OH. These are called the 5′ and 3′ ends of the chain.

In Figure 6, a dinucleotide of thymine and guanine are shown, with the 5’ end of the short dinucleotide chain at the top of the image, the 3’ end of the dinucleotide at the bottom, and the phosphodiester bond circled in red. On the left side of the structure, you can see phosphate-sugar-phosphate-sugar alternating from top to bottom, with the bases extending to the right from the sugars. This is called the phosphate-sugar backbone.

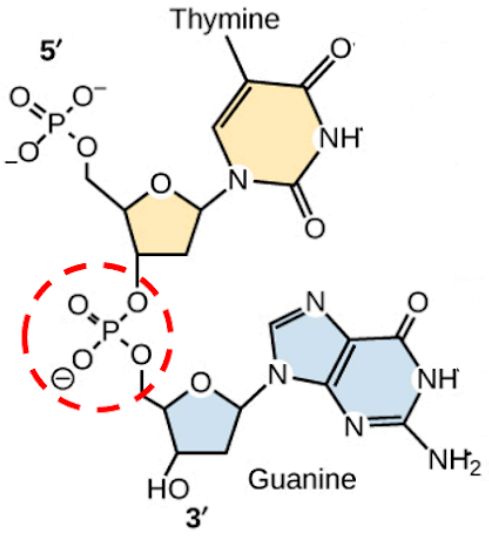


Figure 6. Dinucleotide of thymine and guanine, with phosphodiester bond circled in red

Media Attributions

- Synthesis and Degradation © Christinelmiller is licensed under a CC BY-SA (Attribution ShareAlike) license
- Structure of a nucleotide © CNX OpenStax, CC BY 4.0 , via Wikimedia Commons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Nitrogenous © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Sugar structure © Adapted from OpenStax Biology is licensed under a CC BY-SA (Attribution ShareAlike) license
- Nucleotide vs nucleoside structure © Wikimedia Commons is licensed under a Public Domain license
- Dinucleotide of thymine and guanine © OpenStax Biology is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license

DNA DOUBLE-HELIX STRUCTURE

Nucleotides thus come together to form long polymeric chains. In DNA, two chains pair **antiparallel** to one another via hydrogen bonds that link the bases in pairs. Antiparallel means that the two strands run in opposite directions, with the 5′ carbon end of one strand facing the 3′ carbon end of its matching strand, as shown in Figure 7.

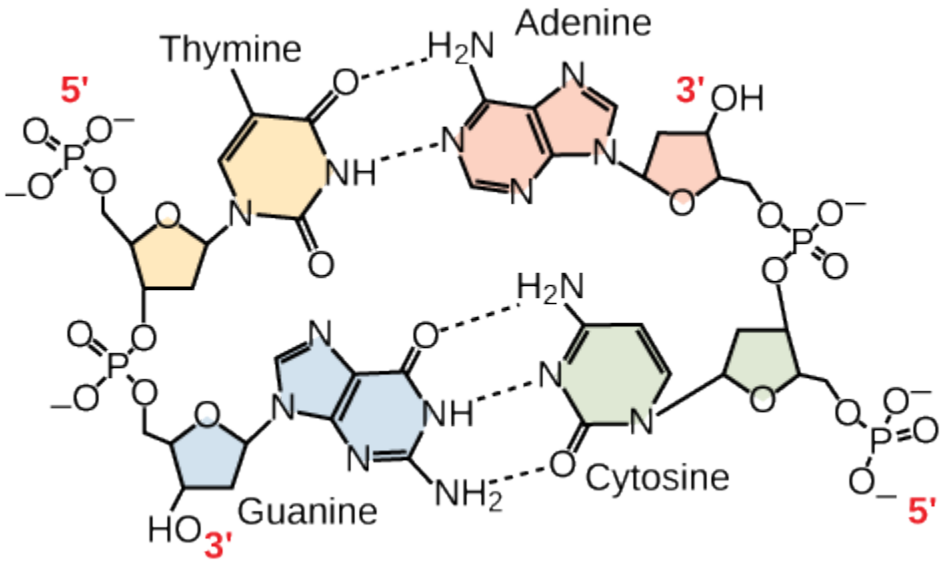


Figure 7. In a double-stranded DNA molecule, the two strands run antiparallel so that one strand runs 5' to 3' and the other 3' to 5'. The sugar-phosphate backbone is on the outside, and the bases are in the middle. A forms base pairs with T, and G base pairs with C.

Only certain base pairings are allowed. Adenine can pair with thymine, and guanine can pair with cytosine, as Figure 6 shows. Adenine and thymine are thus said to be **complementary**, as are cytosine and guanine. Each base pair consists of one purine and one pyrimidine, which makes each base pair approximately the same size. We therefore also say that the DNA strands are complementary to each other.

If we know the sequence of one strand, we can always determine the sequence of the other. For example, if the sequence of one strand has the sequence 5'-AATTGGCC-3', the complementary strand would have the sequence 3'-TTAACCGG-5'. We will see in the Replication and Transcription modules that this property of DNA makes it possible to both copy and use the genetic information stored in DNA.

A note on terminology: We are used to reading text left to right, from the top of a page to the bottom. But inside the cell, DNA is coiled around itself in three dimensions, and left/right and up/down do not have any meaning. Where would you start reading? So instead of thinking of DNA sequence from left-to-right, we think in terms of the 5' and 3' ends. The sequence 5'-AATTGGCC-3' is the same as 3'-CCGGTTAA-5', reading from the 5' end through the bases to the 3' end. For clarity, it is always best practice to label DNA sequence with the 5' and 3' ends. But by convention, if the 5' and 3' ends are not indicated, by convention it is assumed that the sequence is listed 5' to 3', left to right.

Under physiological conditions, the two paired chains coil around each other to form a double-helical molecule (Figure 8). The sugar and phosphate lie on the outside of the helix, forming the DNA's backbone. The nitrogenous bases are stacked in the interior, like a pair of staircase steps. Hydrogen bonds bind the pairs to each other.

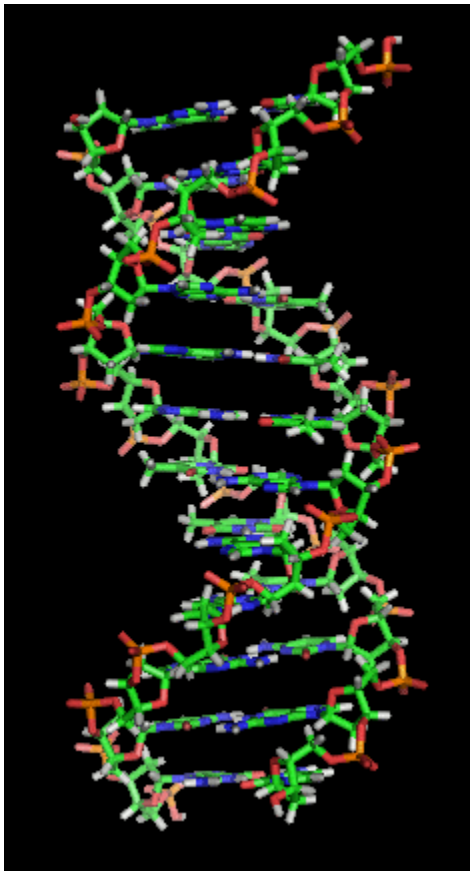


Figure 8. Native DNA is an antiparallel double helix. The phosphate backbone (the curvy lines) is on the outside, with bases on the inside. Each base from one strand interacts via hydrogen bonding with a base from the opposing strand.

Media Attributions

- Antiparallel DNA strands © LibreTexts is licensed under a CC BY-SA (Attribution ShareAlike) license
- DNA double helix © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

HOW DO WE KNOW? DETERMINING THE STRUCTURE OF DNA

DNA is the genetic material

Determining the structure of DNA involved many researchers and the work of several decades.

In the first half of the 20th century, there was considerable debate about the nature of the genetic material. Chromosomes were visible inside cells and were known to be hereditary units, but chromosomes contain both DNA and protein elements. It was known that both DNA and proteins were long polymers built from smaller building blocks. But how those building blocks were assembled was not known. Proteins, built from twenty different amino acids, were chemically far more complex than DNA, which contained only four different nucleotides, so most scientists at the time believed that protein would eventually be shown to be the genetic material.

A series of landmark experiments helped to convincingly determine that DNA was the genetic material. The works of Oswald Avery, Colin Macleod and Maclyn McCarty, and Alfred Hershey and Martha Chase were among these important contributions, but their work all depended on data from earlier studies.

Avery, Macleod, and McCarty, for example, built on earlier work by Frederick Griffith. Avery, Macleod, and McCarty showed that DNA from some bacteria – but not protein or RNA – could transform the characteristics of other bacteria. Thus, DNA passed genetic information from one organism to another. Their work was published in 1944; however, their conclusions were not universally accepted. Critics maintained that their experimental design might allow protein contamination in their DNA samples, thus not completely ruling out protein as the genetic material.

The work of Alfred Hershey and Martha Chase, published in 1951, provided additional evidence that DNA is the genetic material. Hershey and Chase were working with bacteriophages, which are viruses that infect bacteria. Hershey and Chase used radioactively labeled protein and DNA to show that only the DNA of the virus – and not its protein – is transferred to a host bacterium upon infection. Because the transferred DNA is sufficient for the virus to replicate in the host cell, producing more virus, it could be concluded that the genetic material of the virus was DNA.

As more evidence accumulated for DNA as the genetic material, other researchers were attracted to the study of DNA. By the late 1040s and early 1950s, many scientists were actively working to discover the structure of DNA. James Watson and Francis Crick were probably the most well-known of the researchers. However, like most scientific discoveries, the full history included work from many other scientists, all of whom were also, in turn, influenced by the work of other scientists.

Erwin Chargaff, for example, showed that for any DNA sample, the ratios of adenine and thymine were always equal, as were the ratios of thymine and cytosine. These equivalences became known as "Chargaff's rules", and we know now that Chargaff's rules are true because A and T are paired within the double helix, as are C and G.

These landmark experiments greatly contributed to the scientific environment of the 1950s. Because DNA was so chemically simple with only four nucleotide building blocks, it was difficult to understand **how** it could function as the genetic material. So the next big question to answer was how nucleotides were assembled. Understanding the structure would give a glimpse into how DNA functions.

Although many research groups were racing to solve this problem, two different methods were primarily used. These methods were X-ray diffraction and model-building. Both grew directly out of work from protein physicists – remember that protein was the favored molecule in the early 1900's, so a lot of work had already been done in that field.

Protein Structure: Linus Pauling, Robert Corey, and Herman Branson

In many ways, the story could begin with Linus Pauling. Pauling was a chemist who used quantum mechanics to study chemical bonding. His research interests were quite varied: over the course of his career, Pauling published research on chemical bonding and structure, the structure of biomolecules, explosives and weaponry, sickle-cell anemia, and the health benefits of vitamin C. for which he ultimately was awarded the Nobel Prize in Chemistry. Linus Pauling and his wife Ava, a human rights activist, are shown in the photograph in **Figure 9**.

Figure 9. Linus and Ava Pauling.

Over a long career, Linus Pauling studied chemical bonding and structure, biomolecules, explosives and weaponry, sickle-cell anemia, and vitamin C. He has the unique honor of having won two unshared Nobel Prizes: one for Chemistry in 1954, and the Nobel Peace Prize in 1962. Linus Pauling's life was not without controversy: Early years of his career coincided with the rise of Naziism and the Second World War, and he worked extensively on explosives and weaponry at the beginning of the Second World War. He was invited to head the Manhattan Project, the classified government program that ultimately led to the development of the atomic bomb, although he turned down the offer. In later years, Pauling rethought his stance on war and became an ardent pacifist, advocating against nuclear weapons. It was this work that lead to his 1962 Nobel Peace prize. His wife, Ava Pauling, was also a prominent human rights activist. Pauling's activism led to his condemnation as a communist during the Cold War, and for several years his passport was revoked and he was prohibited from international travel. Pauling's work on pacifism eventually led him to leave in 1964 his academic "home" of over four decades, the California Institute of Technology, in response to pressure from university leadership. Despite this, Pauling continued his work in chemistry (at University of California,San Diego and Stanford University) and his humanitarian work (advocating against military action in Vietnam, Cuba and Nicaragua, and the Persian Gulf).

By the late 1940s, Pauling had shifted his research focus from chemical bonding and structure to larger, more complex biomolecules, particularly proteins. Remember, protein had not yet been ruled out as the genetic material.

Like DNA, proteins are complex polymeric molecules. They are built as long chains of amino acid monomers. The chains fold into complex shapes. Pauling was interested in the chemical bonds that drive the folding of the chains.

To study these complex structures, Pauling employed the technique of X-ray diffraction. Briefly, this technique works by shining X-rays through an ordered crystal of protein (under proper conditions, proteins can form crystals much salt can). The X-rays do not pass straight through the crystal. Instead, they are bounced

aside or diffracted if they encounter an atom within the crystal. The diffracted beams are captured through photographic film, much like taking a picture by capturing light via a digital camera sensor. The diffraction pattern can be used to mathematically back-calculate the arrangement of the atoms in the crystal, determining the protein structure.

The technique of X-ray diffraction is difficult and involves complex mathematics, and in the mid-1930s, Pauling did not have the equipment or experience for this work. So, Pauling eventually recruited two additional experts in X-ray diffraction: Robert Corey and Herman Branson.

Robert Corey had done preliminary work using X-ray diffraction to determine the structure of biological molecules. He joined Pauling's lab at Caltech in 1937. While working with Pauling, Corey initially worked to determine the amino acid structure of several amino acids. During World War II, Corey worked to administer Pauling's research collaborations with the War Department, returning to the study of biological molecules in the late 1940s.

In 1948-1949, Pauling recruited Herman Branson (**Figure 10)**, a visiting researcher from Howard University with experience in X-ray diffraction and mathematical chemistry. By that point, Pauling and Corey had data regarding many features of protein structure, including the structure of several individual amino acids and, notably, the planar (flat) shape of the peptide bond linking adjacent amino acids. With these preliminary data as constraints, Branson was tasked with calculating all possible protein structures that could be consistent with the existing information and the X-ray diffraction patterns. He came up with two, including the alpha-helical structure which is one of two main secondary structures found in proteins, and Pauling, Corey, and Branson published the work in *Proceedings of the National Academy of the Sciences (PNAS)* in 1951. Later X-ray diffraction data from other labs confirmed Branson's calculations.

Figure 10. Herman Branson.

*Who contributed? Who gets the credit? In 1954, Linus Pauling was awarded an individual Nobel Prize for his work "**for his research into the nature of the chemical bond and its application to the elucidation of the structure of complex substances'.** Although up to three researchers may share a Nobel Prize, it is notable that this was an unshared award (Pauling's first of two unshared Nobel Prizes). This is, however, uncertainty regarding the role Pauling, Branson, and Corey played in this project, with evidence of disagreement among the collaborators. In later years Branson asserted that his contribution was much larger than his authorship on the PNAS paper suggested. While Branson's work was limited to the one-year project that led to proposing the alpha helix structure, certain inconsistencies between the X-ray diffraction data, Pauling's assumption regarding the water content of the helix, and Branson's measurements worried Pauling. [ref http://scarc.library.oregonstate.edu/coll/pauling/proteins/narrative/page33.html]. It was not until his scientific competitors appeared to be closing in that Pauling consented to publish the structures resulting from Branson's calculations – which, as history has shown, have proven largely correct. [https://www.pnas.org/doi/10.1073/pnas.2034522100#body-ref-ref7] Branson went on to hold prestigious positions in academia, acting as President of two different institutions: Central State University in Ohio, and Lincoln University in Pennsylvania.*

In addition to X-ray diffraction, Pauling's lab relied heavily on what came to be known as "model-building". Essentially, X-ray diffraction allowed calculation of certain bond angles, bond lengths, and geometries within larger molecules, much like observing the shape of a puzzle piece without knowing what the full puzzle looks like. Then, through a bit of trial and error, Pauling and his colleagues would figure out how the pieces could physically fit together, making sure that their final proposed structure was consistent with the existing data. Model building was remarkably successful for determining features of protein structure: after the alpha helix collaboration with Branson, Pauling and Corey went on to publish papers describing the beta sheet and other features of protein structure.

When the work of researchers like Avery, Macleod, McCarthy, Hershey, and Chase added up to suggest

that it was DNA, not protein, that was the genetic material, many researchers, including Pauling, turned their attention to determining the structure of DNA.

Model-building and X-ray diffraction: DNA structure

Using what he knew about the chemical composition of nucleotides, and with helices on his mind, Pauling proposed a helical structure for DNA as well. His structure was spectacularly wrong: he proposed a helix with three intertwined strands, with phosphates pointing inward to the center axis of the helix and bases extending outward. In this, he failed to address both Chargaff's ratios of A:T and C:G equivalency, as well as the chemistry of acids: he used the uncharged, protonated structure of the nucleotides even though, as an acid, the phosphates of DNA are negatively charged at pH 7. Negatively charged phosphates clustered in the center of a helix would repel one another, making Pauling's proposed model chemically impossible.

In Maurice Wilkins' lab at King's College, London, researcher Rosalind Franklin was using X-ray diffraction methods to understand the structure of DNA.

(b)

Figure 11. Scientist Rosalind Franklin used X-ray diffraction to produce this image of DNA, helping to elucidate its double-helix structure. A characteristic "X" pattern is produced when X-ray beams are passed through a helical structure.

Rosalind Franklin's X-ray diffraction methods involved shining a beam of X-rays through a sample of DNA, much like the X-ray diffraction patterns of protein crystals. As in protein diffraction, the X-ray beams bounce off of, or are diffracted by, atoms in the DNA structure. The pattern in which the beams are diffracted is determined by the arrangement of atoms in the molecule, so the structure can be determined from the diffraction pattern. One of Franklin's diffraction images is shown in Figure 11. The X-shape that is clearly visible in this image was immediately suggestive of a helix, as was well-known from X-ray images of protein helices.

While Franklin labored at the tedious calculations required to determine structure from X-ray data, along came the team of James Watson and Francis Crick. Watson and Crick were model builders using the very methods that Pauling had embraced. Bringing together information from many researchers, including Chargaff's rules for the ratios of the bases and, crucially, Franklin's X-ray diffraction data and calculations, which they had access to unbeknownst to Franklin, Crick and Watson used model-building techniques to come up with a model in which all the data fit together.

Watson and Crick proposed the structure that we now accept to be correct: DNA is made up of two strands that are twisted around each other to form a right-handed helix. Base pairing takes place between a purine and pyrimidine on opposite strands so that A pairs with T and G pairs with C as suggested by Chargaff's Rules. Thus, adenine and thymine are complementary base pairs, and cytosine and guanine are also complementary base pairs.

The final X-ray diffraction calculations from both Wilkins and Franklin provided the data to support Watson and Crick's model. All three lab groups – Watson and Crick, Wilkins, and Franklin – published papers describing the structure of DNA, back-to-back-to-back in 1953 in the same issue of Nature.

Watson and Crick's paper, the first of the three in the issue, provided a clear, modeled illustration of the B-form structure of DNA, which is the most common form in cells. The second paper was from Wilkins' research group and described the structure of A-form DNA, which occurs under conditions of low humidity. The final paper was Franklin's, which detailed the structure of B-form DNA as calculated from diffraction data. This paper was the only one of the three to provide experimental evidence for the B-form structure. In 1962, James Watson, Francis Crick, and Maurice Wilkins were awarded the Nobel Prize in Medicine. Franklin, who had died of ovarian cancer in 1958, was not eligible for recognition, as Nobel Prizes are not awarded posthumously.

Media Attributions

GEOMETRY OF THE DOUBLE HELIX

The X-ray diffraction images generated by Franklin, Wilkins, and others, allowed calculation of precise geometric arrangement of the atoms of the DNA molecule. In the double-helix, bases are paired together, stabilized by hydrogen bonds. The strands twist around each other counterclockwise, and they form a right-handed helix, as shown in figure 12 (right image).

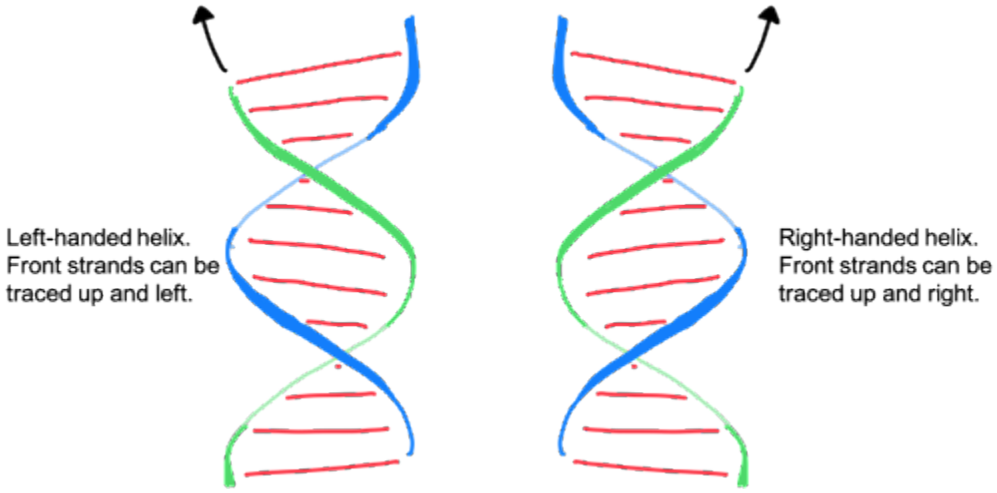


Figure 12. DNA is a right-handed helix. Left- and right-handed helices can be distinguished by looking at the front face of the helix. If the front strands trace up and to the right, it is a right-handed helix. The front face of a left-handed helix trace up and left.

Remember that the two strands are anti-parallel in nature, with the 3′ end of one strand facing the 5′ end of the other strand. The sugar and phosphate of the nucleotides form the backbone of the structure, whereas the nitrogenous bases are stacked inside, like the rungs of a ladder. This is seen in Figure 13a, where the light blue ribbons of the backbone coil around the outside of the helix. One important feature to note is that adenine and thymine form two hydrogen bonds, while cytosine and guanine form three hydrogen bonds, as shown in Figure 13b. Each base pair is a flat structure, roughly trapezoidal in shape. The base pairs form the center of the helix, much like the steps of a spiral staircase, with the sugar-phosphate backbone forming the vertical railing around the outside.

Because each base pair is flat but shaped more like a trapezoid than a rectangle, the base pairs have different dimensions on each side relative to the phosphate groups. The twisting of the two strands around each other thus results in the formation of uniformly spaced major and minor grooves (**Figure 13c**). On the right face

of the helix shown in Figure 13c, the minor groove is labeled. You can see that the red and yellow phosphate groups of the two strands are relatively close together, while on the left side, they are spaced farther apart (labeled as the major groove). This is also seen in Figure 13a, where the cartoon drawing shows the strands alternatingly farther apart and closer together.

The most common form of DNA found in a cell is called B-DNA. In B DNA, Each base pair is separated from the next base pair by a distance of 0.34 nm, and each turn of the helix measures 3.4 nm. Therefore, 10 base pairs are present per turn of the helix. The diameter of the DNA double-helix is 2 nm, and it is uniform throughout due to the consistent pairing of a larger purine with a smaller pyrimidine. These dimensions are also labeled in **Figure 13**.

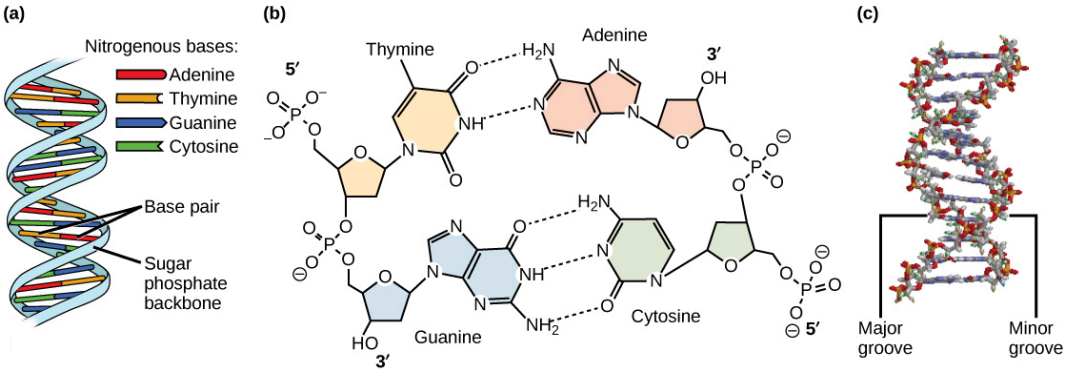


Figure 13. DNA has a double helical structure with sugar phosphate backbone on the outside and bases on the inside (a). AT base pairs have two hydrogen bonds, and GC base pairs have three, represented by dotted lines in (b). DNA has major and minor groove alternating along the length of the helix (c).

The structure we have been discussing, determined by Franklin, Watson, and Crick, is called B-DNA. The structure closely matches the structure of DNA found in the cell, which is an aqueous environment. Under other chemical conditions, DNA can assume slightly different conformations. A-DNA is a short, squat helix, with the bases slanted relative to the axis of the helix. It has a hollow core, as viewed down the center axis of the helix. It is shown on the left in **Figure 14**. A-DNA forms under dehydrating conditions, among others. DNA can even (very rarely) form a left-handed helix! The left-handed helix is called Z-DNA. Certain base sequences are more prone to forming Z-DNA. A-, B-, and Z-DNA are shown in Figure 14.

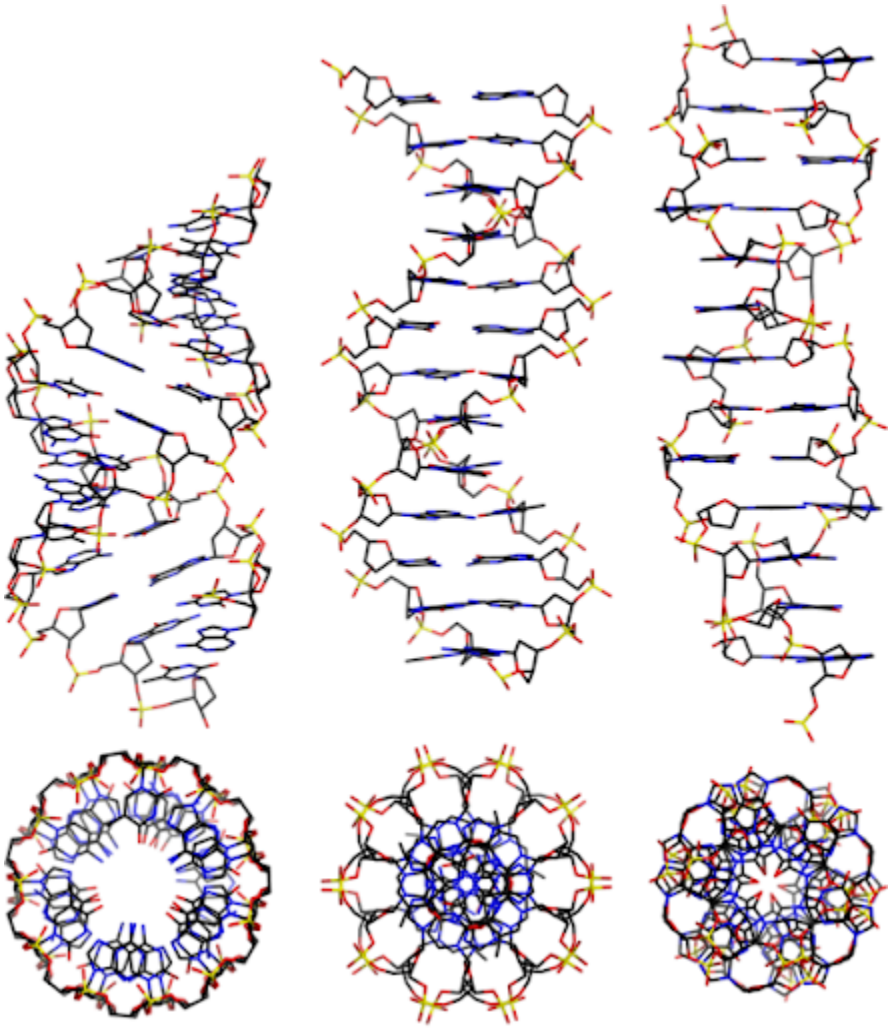


Figure 14 A-Form (left), B-Form (center), and Z-Form (right) DNA. B-DNA is the primary form in the cell. A- and B-DNA are right-handed, while Z-form is left-handed.

Media Attributions

- Left and right handed helix © Amanda Simons is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- A-, B-, and Z-DNA © Adapted from Biology LibreTexts, Working with Molecular Genetics (Hardison) is licensed under a CC BY-SA (Attribution ShareAlike) license

CHROMOSOME STRUCTURE

In eukaryotes, DNA is found in the nucleus as well as in the chloroplasts and mitochondria. In prokaryotes, the DNA is not enclosed in a membranous envelope. The cell's entire genetic content (all its DNA) is its **genome**, and the study of genomes is **genomics**.

An organism's genome is typically divided among a number of long DNA molecules called **chromosomes**. While prokaryotic genomes are typically made up of only one chromosome, eukaryotic genomes may have many. Human cells, for example, have 23 pairs of nuclear chromosomes (46 total) plus one mitochondrial chromosome.

Each chromosome can be thousands or even millions of base pairs long. The largest human chromosome (Chromosome 1) is 249 million base pairs long. As shorthand, biologists sometimes refer to lengths of DNA in kilobases or megabases, using the prefixes of the metric system. A thousand base pairs would be one kilobase, and human Chromosome 1 is 249 megabases. The human genome in total is 6 gigabases.

Prokaryotic chromosomes are typically circular, while eukaryotic chromosomes are typically linear in structure. In eukaryotic cells but not in prokaryotes, DNA forms a complex with histone proteins to form **chromatin**. The protein and DNA together make up the substance of eukaryotic chromosomes.

Before a cell can divide, the entire genome must be copied, so that there is a complete set of chromosomes available to be passed to each daughter cell. In the first half of the cell cycle, each eukaryotic chromosome is made up of one molecule. However, after each chromosome has been copied, the two daughter molecules remain linked together by specialized proteins. The two DNA molecules are called sister chromatids, and as long as they remain linked they are still collectively called one chromosome. This is illustrated in **Figure 15.** Note that although the replicated chromosome is drawn here (and in many other images!) as an X-shaped structure, in actuality the two chromatids are bound together all along their length. Individual chromatids cannot visually be distinguished using light microscopy, because they are so tightly associated with one another (see Figure 19 for an example).

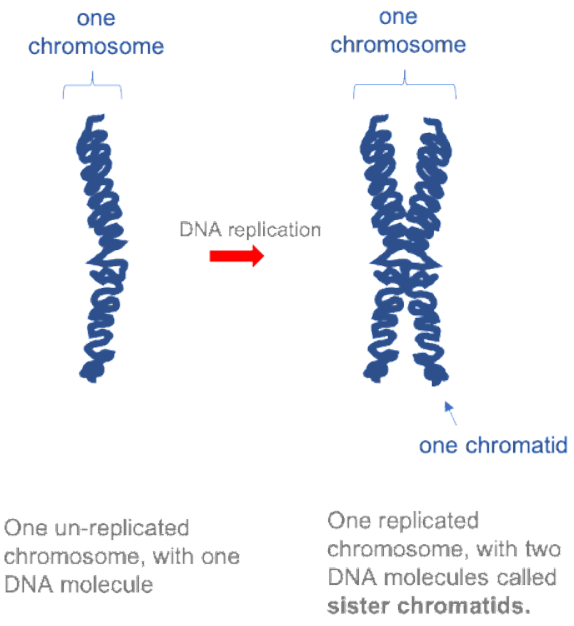


Figure 15. Before DNA replication, each chromosome is comprised of one DNA molecule. After replication, each chromosome is comprised of two DNA molecules. The two halves of a replicated chromosome are called chromatids.

A chromosome may contain tens of thousands of

genes. Many genes contain the information to make protein products. Other genes code for RNA products. All cellular activities are ultimately controlled by turning the genes "on" or "off" – using the genetic information to make RNA and protein products, or not. Whether or not a gene is active can have profound effects on the function of a cell, allowing differences in cell behavior, appearance, and function.

DNA can be highly compacted

If stretched to its full length, the DNA molecule of the largest human chromosome would be 85mm. Yet during mitosis and meiosis, this DNA molecule is compacted into a chromosome approximately 5μm long. Although this compaction makes it easier to transport DNA within a dividing cell, it also makes DNA less accessible for other cellular functions such as DNA synthesis and transcription. Thus, chromosomes vary in how tightly DNA is packaged, depending on the stage of the cell cycle and also depending on the level of gene activity required in any particular region of the chromosome.

Prokaryotic compaction

Prokaryotic chromosomes are circular DNA, with one continuous circular sugar-phosphate backbone. They are typically smaller than eukaryotic chromosomes. For example, the smallest human chromosome, chromosome 21, is 48 million base pairs (48,000,000 base pairs, or 48 megabases). The largest human chromosome, chromosome 1, *E coli,* a bacterium commonly used as a model organism in the lab, has a genome comprised of one circular piece of DNA, about 4,600,000 base pairs (or 4.6 megabases).

4.6 megabases is still 1.5 millimeters long – 10x longer than the length of the *E. coli* cell, which is only 1-2 micrometers. To fit a long piece of DNA into a relatively small cell, prokaryotes typically compact their DNA through supercoiling. This results from over- or under-winding the DNA, which causes it to curl up or writhe around itself. This is like what happens if you try to untwist or overtwist a piece of string comprised of smaller individual twisted strands. As shown in the video in Figure 16, over- or under-twisting a coil introduces writhes in the circular structure, packing what was a loose, floppy structure into a tightly coiled package. Most prokaryotes compact their DNA through underwinding, although some extremophiles use overwinding.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=57#h5p-1

Levels of compaction in eukaryotes

Eukaryotic chromosomes are linear, so writhes introduced by under-winding are not retained in the structure. Instead, eukaryotic cells use proteins to organize long, linear DNA molecules into compact structures. Collectively, the DNA and proteins together are called **chromatin**.

There are several different levels of structural organization in eukaryotic chromosomes, with each successive level contributing to the further compaction of DNA. Each level involves a specific set of proteins that associate with the DNA to compact it.

First, proteins called the **core histones** act as spool around which DNA is coiled twice to form a structure called the **nucleosome**. The core histones are called H2A, H2B, H3, and H4. Two of each core histone come together to form a histone octamer. The histones each have long tails extending from the core of the histone. These tails contain positively charged amino acids, which bind to the negatively charged DNA to hold it in place. About 150 bp of DNA wraps around each histone octamer, as shown in **Figure 17**.

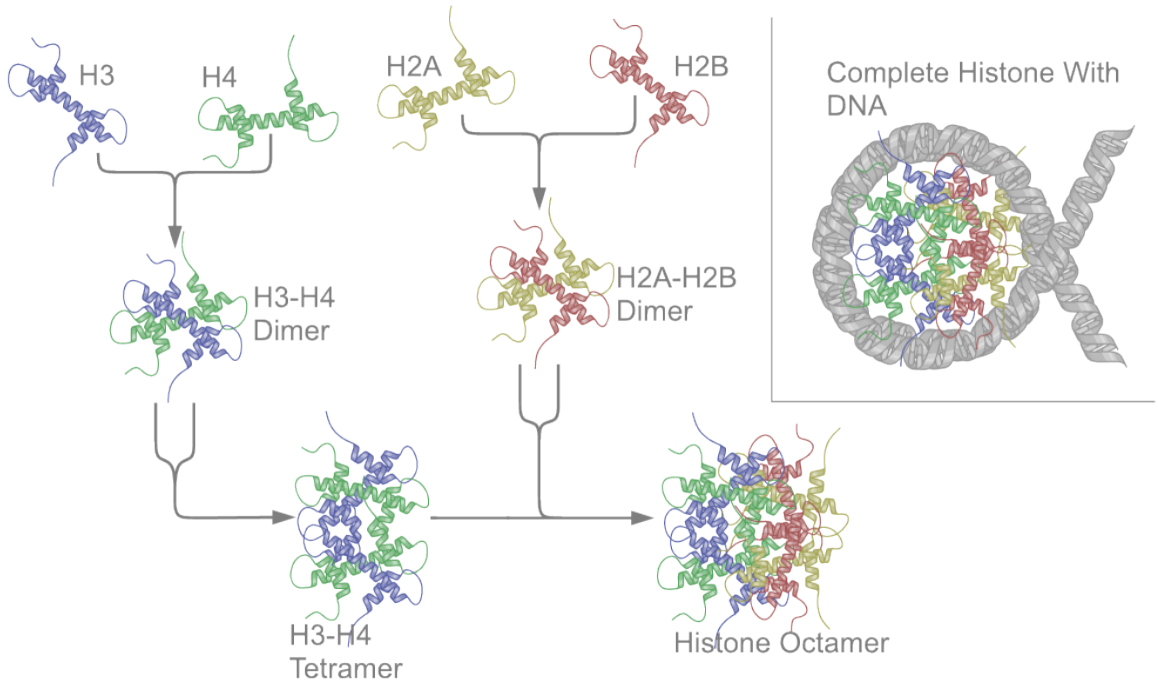


Figure 17 Histone proteins H2A, H2B, H3, and H4 are assembled into the histone octamer in a multistep process. Histone "tails" extending from the body of the octamer body. DNA wraps around the octamer, with histone tails holding the DNA in place through noncovalent bonds.

Nucleosomes are formed at regular intervals along the DNA strand, giving the molecule the appearance of "beads on a string" under an election microscope. At the next level of organization, **histone H1** helps to compact the DNA strand and its nucleosomes into a **30nm fiber**, so named because it is 30nm in diameter.

Subsequent levels of organization involve the addition of **scaffold proteins** that wind the 30nm fiber into coils, which are in turn wound around other scaffold proteins (**Figure 18**).

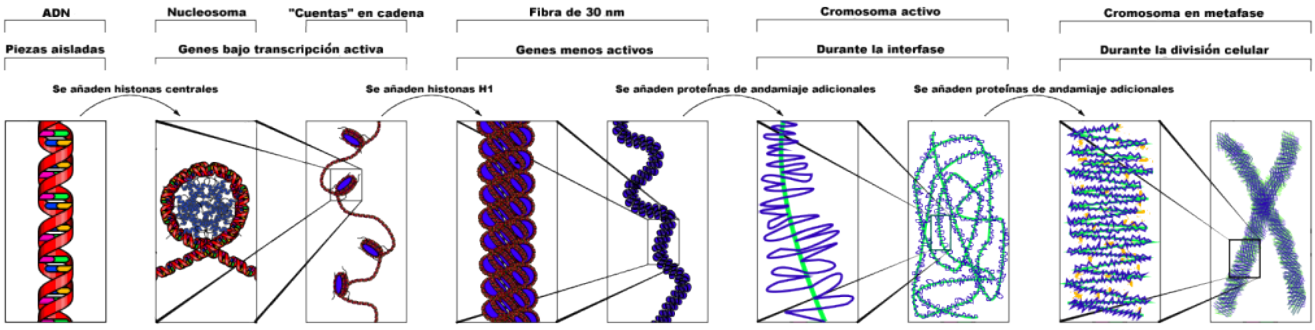


Figure 18 Successive stages of chromosome compaction depend on the introduction of additional proteins. In cells not actively dividing, DNA remains partially compact, as shown in the images labeled "interphase". The final, most compact structures are formed beginning only in metaphase of mitosis or meiosis.

For more loosely compacted DNA, only the first few levels of organization may apply. DNA is fully compacted, or **condensed**, only during cell division. During interphase of the cell cycle, the less compact DNA is said to be **decondensed**. Note again: Although replicated chromosomes are often depicted as "X" shaped structures, as in this image, scaffold proteins bind the length of sister chromatids. Under a microscope, replicated chromosomes can appear as a single linear shape since the two chromatids may not be resolved.

Chromatin Packaging Varies inside the Nucleus: Euchromatin & Heterochromatin

Chromosomes stain with some types of dyes, which is how they got their name: Chromosome means "colored body". Certain dyes stain some regions along a chromosome more intensely than others, giving some chromosomes a banded appearance. Researchers skilled in cytogenetics can use these characteristic banding patterns to identify specific chromosomes. This is discussed further in later modules.

There are two major types of chromatin, but these are more the ends of a continuous and varied spectrum. **Euchromatin** is more loosely packed, and tends to contain genes that are being transcribed. **Heterochromatin** is more densely compacted. It tends to contain many repetitive sequences and genes within heterochromatin tend not to be transcribed.

Morphological features of Chromosomes

Chromosomes also contain distinctive features, such as centromeres and telomeres. Both of these are usually heterochromatin. In most cases, each chromosome contains one **centromere**. Centromeres are regions of

the chromosome that are bound by proteins that link the centromere to microtubules that transport chromosomes during cell division. Under the microscope, centromeres of metaphase chromosomes can sometimes appear as constrictions in the body of the chromosome (**Figure 19**).

If a centromere is located near the middle of a chromosome, it is said to be **metacentric**, while an **acrocentric** centromere is closer to one end of a chromosome, and a **telocentric** chromsome is at, or near, the very end. In contrast, some species have a **holocentric** centromere, where no single centromere can be defined and the entire chromosome acts as the centromere.

Telomeres are repetitive sequences near the ends of linear chromosomes and are important in maintaining the length of the chromosomes during replication, and protecting the ends of the chromosomes from alterations.

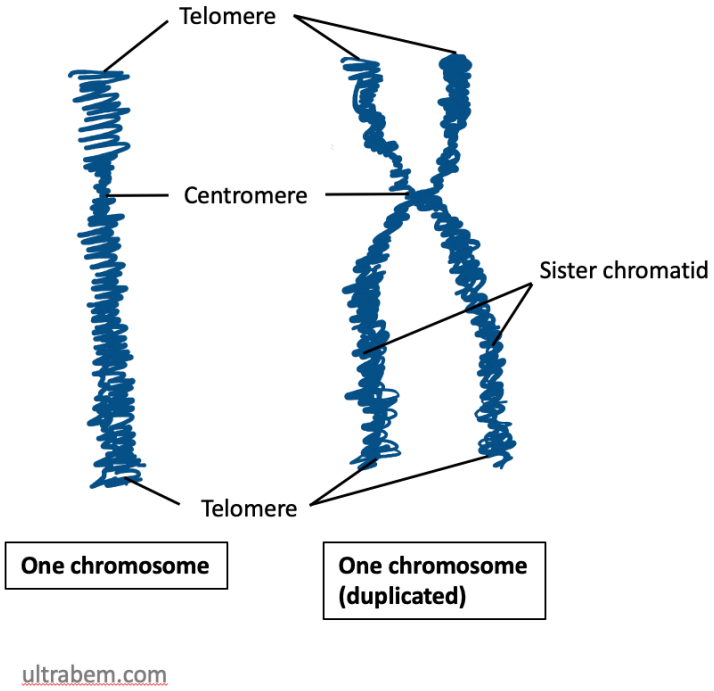


Figure 19. A (top) pair of metacentric chromosomes, with arrow indicating the centromere. The dark and light banding is due to areas of the chromosome staining differently from one another. Note that although these are replicated chromosomes, there is no visible separation between the chromatids, since they are bound together along their full length. B (bottom) Centromeric regions of a chromosome serve as attachment points for proteins during cell division. Telomeres are sequences found at the ends of linear chromosomes.

Some chromosome terminology: **Homologous** is a word that means similar but not identical. **Homologous chromosomes** are pairs of similar but non-identical chromosomes in which one member of the pair typically comes from the male parent, and the other comes from the female parent. Homologous chromosomes contain the same genes but not necessarily the same version (allele) of each gene. In Figure 20, three chromosomes are shown: two homologous large chromosomes and one smaller chromosome that is not homologous to the other two.

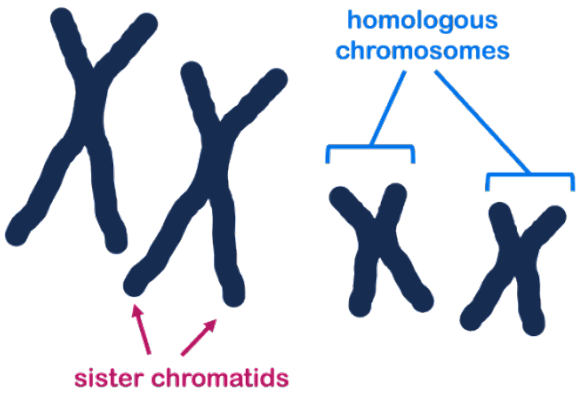


Figure 20. Relationships between chromosomes and chromatids.

Sister chromatids are the two DNA molecules that form halves of a single, replicated chromosome. In **Figure 20**, each chromosome has two sister chromatids.

Because a pair of sister chromatids is produced by the replication of a single DNA molecule, their sequences are essentially identical (same versions of each gene), differing only because of any DNA replication errors. Homologous chromosomes will be highly similar, but not identical. They may have different versions (alleles) of each gene.

Nuclear organization

Even when they are decondensed, the chromosomes are not randomly arranged in within the interphase nucleus. They often have specific locations within the nucleus and relative to one another. In **Figure 21**, each chromosome in the nucleus of a single cell has been stained a different color, using a technique called Fluorescent In Situ Hybridization (FISH). FISH is sometimes called chromosome painting, since it is so colorful! Rather than decondensed chromosomes tangling all over the nucleus like a big plate of spaghetti, each chromosome remains collected in a small area of the nucleus. As we will see in other modules, this likely helps in gene regulation.

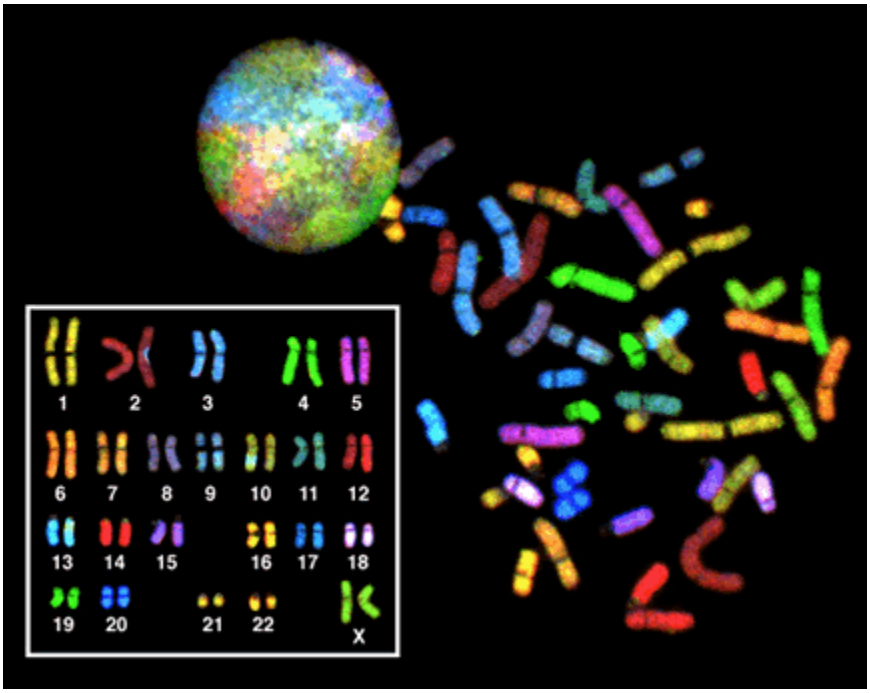


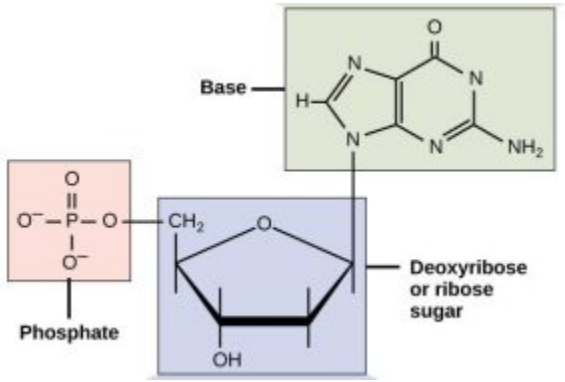
Figure 21. Top: FISH (Fluorescence in situ hybridization) labeling of all 24 human chromosomes (1-22, XX) in a human cell nucleus, each chromosome labeled with a different combination of fluorescent dyes to give it a unique color. Right: FISH-labeled condensed chromosomes collected from a metaphase cell. Bottom: Karyogram constructed by digitally manipulating the metaphase image to pair sister chromosomes and visually depict the cell's karyotype.

Media Attributions

- NA replication © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Nucleosome structure © Wikimedia is licensed under a CC0 (Creative Commons Zero) license
- Estructuras cromatina © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- metacentric chromosomes © LibreText is licensed under a CC0 (Creative Commons Zero) license
- Chromosome structure © wikipedia is licensed under a CC0 (Creative Commons Zero) license
- chromosomes and chromatids © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- FISH (Fluorescence in situ hybridization) © Wikipedia is licensed under a Public Domain license

WRAP-UP QUESTIONS

1. What are the three main functional groups in a nucleotide?
2. What functional group makes the DNA molecule an acid?
3. What charge do DNA and RNA carry in the cell?
4. Name three chemical differences between the structures of DNA and RNA.
5. In the nucleotide structure below, circle the 5' carbon. Draw a square around the 3' carbon.



6. 10% of the nucleotides in a double-stranded DNA are thymine. Use Chargaff's rules to determine the percentage of C, G, and A.

7. Sort these nitrogenous bases into purines and pyrimidines: Cytosine, Guanine, Uracil, Adenine, Thymine.

8. If one strand of a DNA molecule has the sequence 5'CGGAGT3', what is the sequence of the second strand? Be sure to label the 5' and 3' ends.

9. Fill in the blanks. The most common form of DNA in a cell is __DNA. This form of DNA is a _____-handed helix.

10. In the image of DNA molecular structure in **Figure 18**, find and label the 5' and 3' ends of each strand. Find and label and the bases adenine, guanine, cytosine, and thymine.

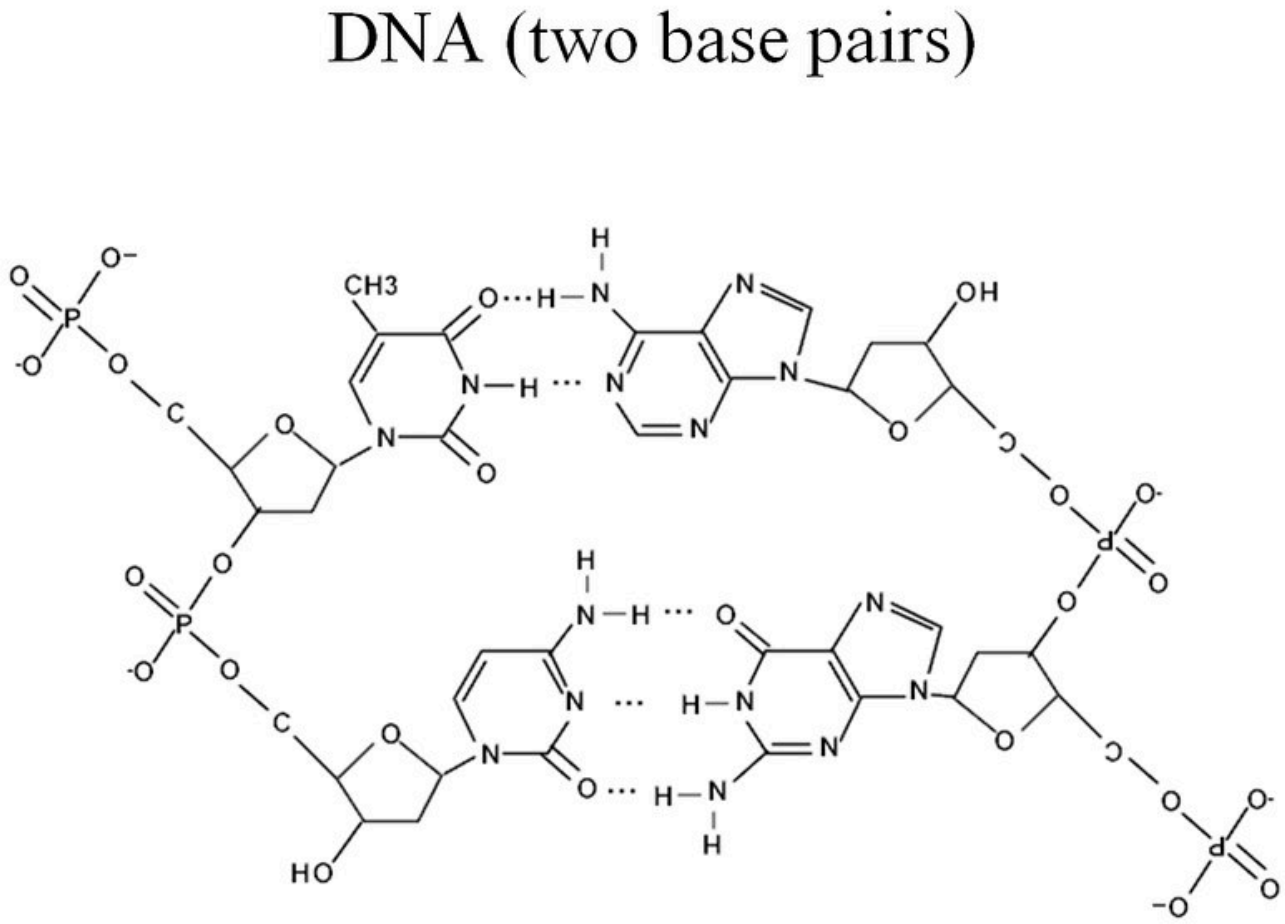


Figure 18. DNA molecular structure.

11. Name the four histone proteins that are found in a nucleosome core.

Science and Society

12. Nobel Prizes are not awarded posthumously, but they are also not awarded to more than three individual researchers at a time. If Rosalind Franklin had lived, who should have won the Nobel Prize for DNA structure? Whose contributions – Watson, Crick, Franklin, or Wilkins — "count most"? In your response, use outside sources as references to justify your opinion.

Media Attributions

- Question 5 Figure © Adapted from OpenStax Biology is licensed under a CC0 (Creative Commons

PART II

GENOME STRUCTURE

Objectives

1. Recognize the difference in genome structure between eukaryotes and prokaryotes.
2. Define ploidy, diploid, haploid, aneuploid.
3. Define karyogram, karyotype, ideogram.
4. Describe the genome of a cell or organism as a multiple of n, where n=haploid number of chromosomes.
5. Recognize that most of a eukaryotic genome is not protein-coding sequence.
6. Define LINE, SINE, and transposable element.

Introduction

The genome of prokaryotic organisms is typically composed of a single circular DNA molecule, or chromosome. In eukaryotic species, the nuclear genome is typically divided among multiple individual DNA molecules, or chromosomes. Eukaryotic chromosomes are linear. In both bacterial and eukaryotic chromosomes, protein-coding genes are interspersed with other DNA sequence.

This chapter looks more at how prokaryotic and eukaryotic chromosomes are organized, described, and visualized.

PROKARYOTIC GENOME STRUCTURE

Figure 1 shows a diagram of the circular chromosome from one strain of the bacteria Escherichia coli, modified from a recent research paper. The E. coli genome is 5 million base pairs long, or 5 megabase pairs (5Mbp). There are about 5,000 genes within that chromosome. The diagram shows the location of the genes arranged along the chromosome, with the coding sequence of each gene, or cds for short, highlighted in blue. The coding sequence is the information that will be used to produce a protein (discussed more in later chapters). Remember that DNA is double-stranded: gene coding sequences can be on either strand, and the two concentric circles of blue marks indicate on which strand the coding sequence is found[1].

Bacterial genomes typically do not have very much DNA sequence separating genes. Depending on species, 90-95% of the genome may be coding sequence[2]. Some of the interspersed DNA includes regulatory sequence that is important for determining under which conditions a gene might be used to produce a protein.

1. Sellera, F. P. *et al.* Genomic Analysis of a Highly Virulent NDM-1-Producing Escherichia coli ST162 Infecting a Pygmy Sperm Whale (Kogia breviceps) in South America. *Front. Microbiol.* **13**, (2022).

2. Rogozin, I. B. *et al.* Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res* **30**, 4264–4271 (2002).

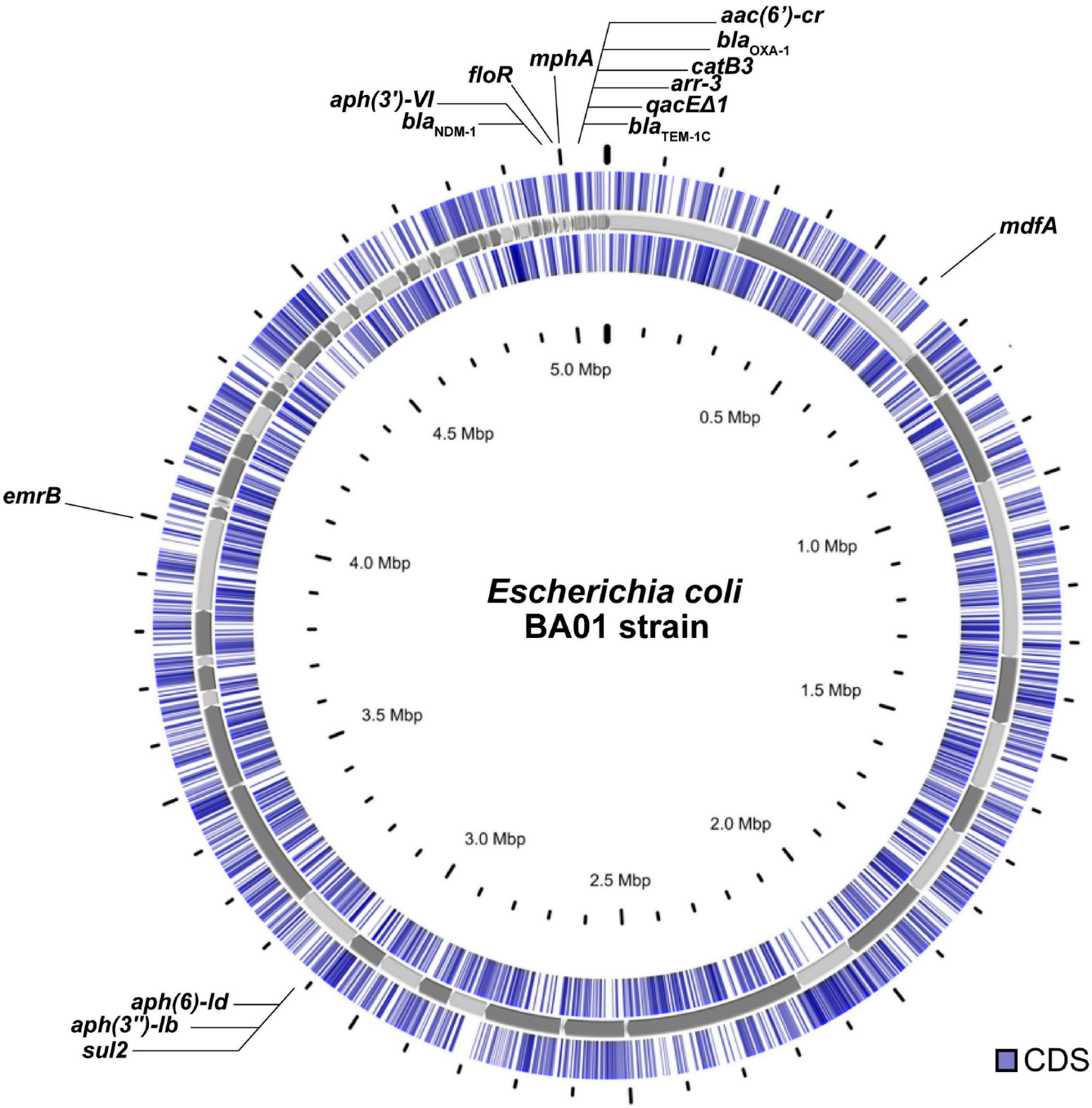


Figure 1. Example bacterial genome.Bacterial chromosomes are circular. In this example genome, the location of genes are shown in blue. Gene coding sequences (CDS) may be found on either strand of the double helix. The two concentric circles represent the coding sequence that is found on each strand of the helix.

In addition to the circular genome, bacteria may also harbor additional extra-chromosomal DNA molecules called **plasmids**, illustrated in **Figure 2**. Each bacterial cell typically only houses one copy of the genome, but plasmids are usually present in multiple copies. Plasmid DNA, like the bacterial chromosome, is circular, but

usually much smaller (measured in a few kilobases rather than megabases). Like the bacterial genome, plasmid DNA is copied by the cell and passed to daughter cells during cell division. Plasmid DNA can carry genes.

Antibiotic resistance genes are often encoded on plasmids, and the exchange of plasmids among bacterial cells can play a role in the spread of bacteria resistant to antibiotics. You can read more about the role plasmids play in the spread of antibiotic resistance in this article from the American Society for Microbiology[3].

Although not discussed extensively in this text, bacteria can, under some conditions, take up plasmid DNA from their environment or from another bacterial cell through a process called conjugation.

Plasmid DNA can also readily be transferred to both prokaryotic and eukaryotic cells in the lab, so plasmids are commonly used to study mechanisms of genetics.

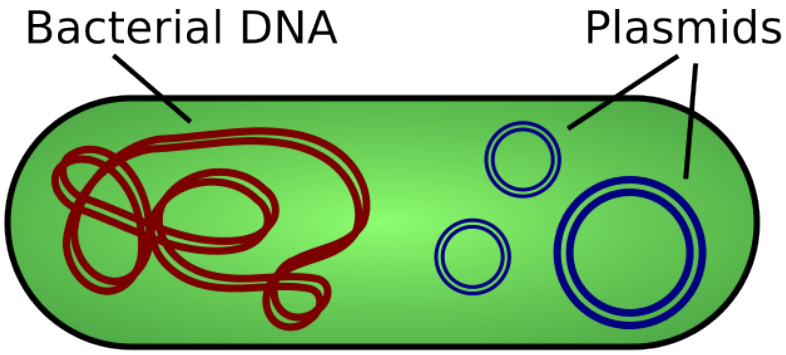


Figure 2. In addition to the bacterial chromosome, bacterial cells may harbor extrachromosomal plasmid DNA. Plasmids are variable in size but typically much smaller than the bacterial genome. A plasmid may be present in multiple copies in the cell. Image source:

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=65#h5p-92

Media Attributions

- Escherichia coli BA01 strain © frontiers is licensed under a CC BY-SA (Attribution ShareAlike) license

3. Plasmids and the Spread of Antibiotic Resistance Genes. *ASM.org* https://asm.org:443/Articles/2023/January/Plasmids-and-the-Spread-of-Antibiotic-Resistance-G.

- Plasmid © User:Spaully via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

EUKARYOTIC GENOME STRUCTURE

The genome of each eukaryotic species has a characteristic number of linear chromosomes found in the nucleus of each somatic cell; for example, humans have 46 nuclear chromosomes, mice have 40, and apples have 34.

In addition, mitochondria and chloroplasts also house their own DNA! So, in total, the human genome has 46 nuclear chromosomes, plus a mitochondrial chromosome. The apple genome has 34 nuclear chromosomes, plus a mitochondrial chromosome and a chloroplast chromosome.

The **ploidy** of an organism's genome describes the number of copies of each chromosome the organism typically has. A **haploid** genome consists of one copy of each chromosome, and a haploid organism likewise only has one copy of each chromosome (prokaryotes are haploid). A **diploid** organism is one in which each chromosome is represented in two copies, or **homologs**. Humans are diploid organisms with 23 pairs of nuclear chromosomes, for 46 total.

The homologs of a pair of chromosomes are very, very similar to one another (99.9% identical for humans![1]), but they will have small differences in DNA sequence.

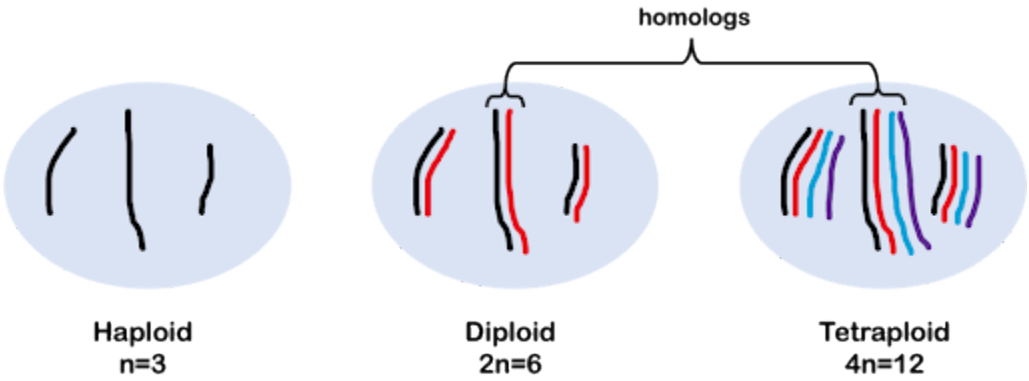


Figure 3. The ploidy of a cell describes the number of copies of the genome. A haploid cell has only a single copy of each chromosome. A diploid cell has two copies of each chromosome. A tetraploid cell has four copies of each chromosome. The copies of a chromosome are called homologs. They are very similar in structure and sequence, but not identical.

1. Zhang, X.-O., Pratt, H. & Weng, Z. Investigating the Potential Roles of SINEs in the Human Genome. Annu. Rev. Genom. Hum. Genet. 22, 199–218 (2021).

As discussed more extensively in the module on Meiosis and Mitosis, we use the variable "n" to indicate the number of chromosomes in a single, haploid copy of the genome, with a multiplier to indicate ploidy. For example, 2n indicates a diploid with two copies of each chromosome, 3n indicates a triploid with three copies of each chromosome, and 4n indicates a tetraploid with four copies of each chromosome (**Figure 3**).

In this text, we primarily focus on genetics using diploid organisms (2n), but it's important to note that other organisms have different genome copy numbers. For example, some bananas are triploid (3n) and some wheat is hexaploid (6n). Strawberries are an interesting example: different cultivars of strawberries are diploid, tetraploid, pentaploid, hexaploid, octoploid, or pentaploid!

You will also see the term **aneuploidy**. **Aneuploid** describes a cell or individual organism with an extra (or missing) portion of the genome compared to what is expected for that species.

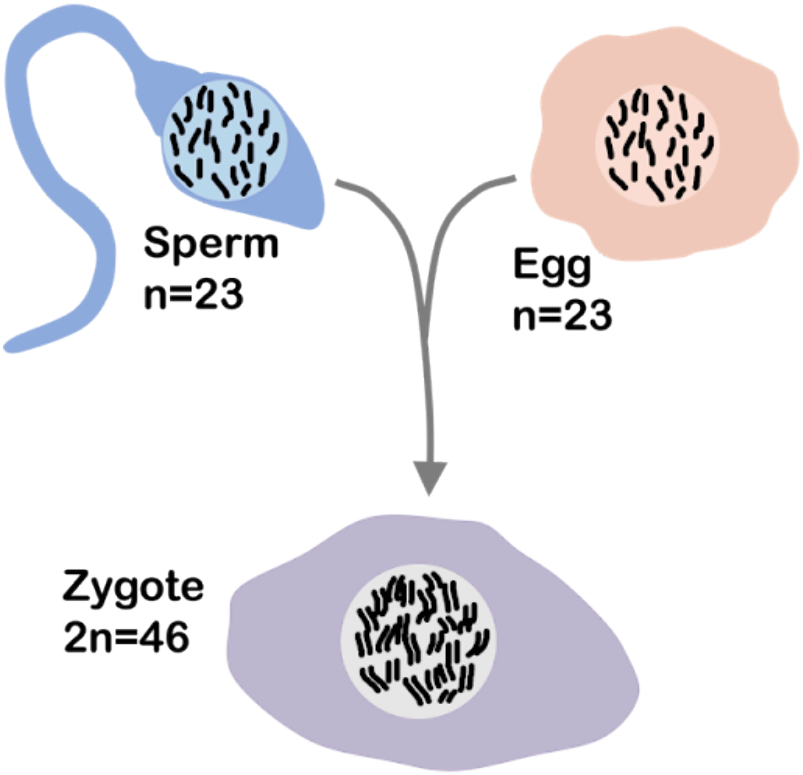


Figure 4. Human cells are diploid with 46 pairs of chromosomes (2n=46). Haploid sperm (n=23) and haploid egg (n=23) come together to form a diploid zygote. The zygote divides via mitosis to give rise to the rest of the cells of the mature body.

During sexual reproduction, only one copy from each pair is passed to offspring. Reproductive cells like egg or sperm, produced through the process of meiosis, have half the number of copies of the chromosome.

Human gametes (egg or sperm) are haploid, with 23 chromosomes (n=23). An egg and sperm come together to form a diploid zygote that develops into a mature organism. Each human thus typically 23 paternally-inherited nuclear chromosomes and 23 maternally-inherited nuclear chromosomes, with 46 total (2n=46), as shown in **Figure 4**.

Diploid offspring thus inherit one copy of each chromosome from each parent, for two copies total.

For organisms with more than two copies of the genome, gametes will still contain half the total chromosomes. So an octoploid strawberry, which has eight copies of each chromosome, will produce gametes that have four copies of each chromosome. Sometimes the genome of such plants would be written as 2n = 8x = 56 rather than just 8n = 56.

Organisms with an odd number of chromosome sets (like seedless watermelon and many cultivars of bananas) are often sterile: they cannot reproduce sexually. This is because the process of gametogenesis fails during meiosis I when chromosomes would normally pair. With an odd number of chromosome sets, pairing isn't possible. Seedless watermelon are produced from a cross between a diploid and a tetraploid cultivar. New triploid banana plants are produced via vegetative propagation: rooting a new plant from outshoots of an older one.

Chromosomes vary dramatically in size, even within a species. For example, the smallest human chromosome is 48 megabases long (48 million bases), while the largest is 249 megabases. Note that although single-celled organisms and simpler eukaryotes do tend to have smaller genomes than higher eukaryotes, the complexity of an organism, the number of base pairs in the genome, the number of genes, and the number of chromosomes are not well correlated to one another. **Table 1** compares the nuclear genomes for several diploid model organisms used regularly in life sciences research.

Table 1. Comparison of nuclear genomes

Eukaryotic Species	Approximate size of nuclear haploid genome	Number of nuclear chromosomes in haploid genome	Approximate number of genes
Human *Homo sapiens*	3.1 billion bases	n = 23	20,500
Chimpanzee *Pan troglodytes*	3.2 billion bases	n = 23	23,500
Mouse *Mus musculus*	2.7 billion bases	n = 20	22,500
Fruit fly *Drosophila melanogaster*	144 million bases	n = 4	14,000
Roundworm *Caenorhabditis elegans*	100 million bases	n = 6	20,000
Corn *Zea mays*	2.4 billion bases	n = 10	40,000
Mustard weed *Aradbidopsis thaliana*	120 million bases	n = 5	27,500

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=67#h5p-93

A **karyotype** is a written description of an individual's complete set of chromosomes. The karyotype of a typical human chromosomal male would be written, "46, XY". A karyotype can also include any aneuploidies: abnormalities or deviations from expected chromosome number. For example, the karyotype of a male with Down syndrome, who has an extra copy of chromosome 21, would be written, "47, XY +21". The genome of an individual who is aneuploid for sex chromosomes might be described as "47, XXX".

A **karyogram** is a visual depiction of a karyotype. It is an image of the chromosomes from a single cell,

produced by manipulating microscopic images of a cell in metaphase. A karyogram from a human cell is shown below in **Figure 5**, reprinted from the chapter on DNA structure.

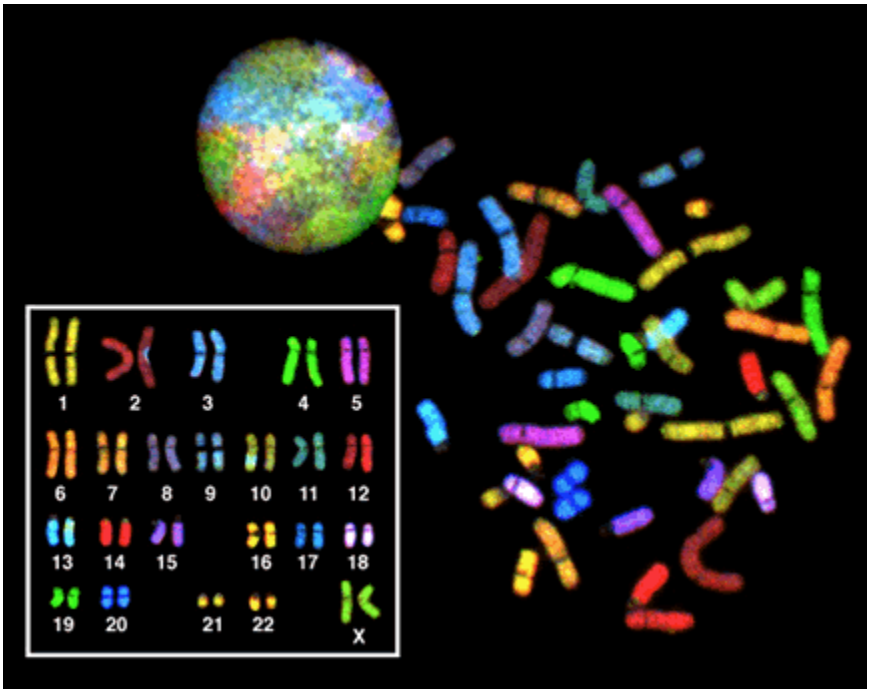


Figure 5. Human Spectral Karyogram.

To produce a karyogram, cells undergoing division are used to produce a chromosome smear, where the chromosomes from a single cell are spread out on a microscope slide and separated visually. An image of the chromosome smear is then digitally cut apart and reassembled to pair the maternal and paternal copies of each chromosome next to one another. This makes it easy to see any chromosomal abnormalities.

Figure 5 was produced by staining the chromosomes using a technique called "SKY", or spectral karyotyping. Each chromosome appears a different color because it is tagged with a different combination of fluorescent molecules. At the top left of **Figure 5**, the round colorful circle is an interphase nucleus. The different colored patches are areas of the nucleus occupied by different uncondensed chromosomes. Note that without the colored staining, the chromosomes would just be part of a single, uniform mass: there are no distinct boundaries between the chromosomes.

On the right in the image is a spread of chromosomes from a single mitotic nucleus. Individual condensed chromosomes have been released from the cell and are spread out on a microscope slide. The chromosomes are stained many different colors, but if you look carefully, you will see that there are two chromosomes of each color.

In the inset image on the bottom left of **Figure 5**, this metaphase spread has been digitally manipulated to pair up the homologous pairs. You'll see that the pairs of the chromosomes are the same ones from the spread: For example, the red C-shaped copy of chromosome 2 can be found toward the bottom right of the

spread, and its homolog appears to have a blue spot in the karyogram because it is slightly overlapped by a blue chromosome 3 toward the top left of the spread.

Karyograms are produced by staining chromosomes by other methods as well. For example, one of the most common dyes used is Giemsa stain, which has greater affinity for A-T rich regions of DNA. When DNA is stained with Giemsa, the chromosomes appear banded: light in color in areas where the DNA has a high percentage of G-C base pairs, and dark where the DNA has a high percentage of A-T base pairs. Because each chromosome has a different DNA sequence, each chromosome has a characteristic striped (or "band") pattern: those bands can be used to identify individual chromosomes under the microscope. Other methods of staining exist as well: one method stains centromeres, another stains G-C rich regions, and another stains telomeres more darkly than the rest of the chromosome.

Figure 6 shows a Giemsa stained karyogram. Although Figure 3 and Figure 4 were produced using different staining techniques, both karyograms arrange the chromosomes in the same pattern: By convention, human karyograms are arranged with chromosomes 1 to 5 in the top row, 6 to 12 in the second, 13 to 18 in the third, and 19 to 22 and XY in the bottom.

Chromosomes numbered 1-22 are called the **autosomes**: both males and females have two copies of each. The X and Y chromosomes are a little different. These are the **sex chromosomes**, so-called because they play a role in determining sex phenotype. Mammalian females, including humans, typically have two X chromosomes, with males having one X and one Y. The karyogram in **Figure 3** was obtained from a chromosomal female: you can see two similar-length green X chromosomes. In the typical male karyogram in **Figure 4**, you can see the X and Y chromosomes paired in the bottom right of the karyogram. The Y chromosome is much shorter than the X chromosome.

The chromosomes in the karyograms are arranged from largest to smallest: the chromosomes were numbered based on their size. The only exceptions to this are the sex chromosomes (X and Y), and chromosome 21, which is smaller than chromosome 22. Although chromosome 22 was originally thought to be the smallest, later analysis showed that chromosome 21 is shorter.

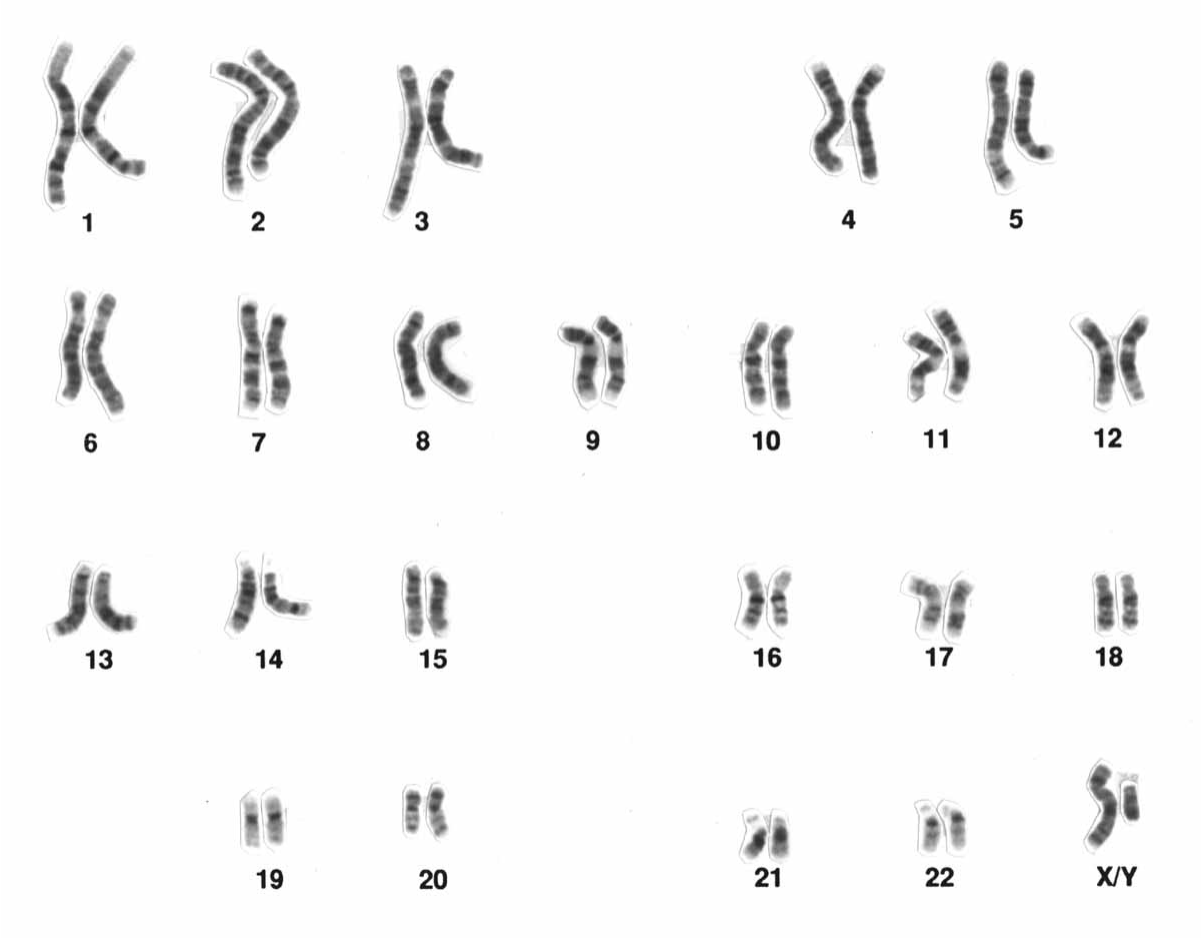


Figure 6. Giemsa banded karyogram of a typical human male karyotype.

The patterns produced by Giemsa staining, as in **Figure 6**, are called G-banding. Because each chromosome has a characteristic banding pattern, the bands have been used to identify features – and sometimes abnormalities – of individual chromosomes. Loci (or locations) on a chromosome are often described by their position relative to G-bands. The G-bands are numbered based on their position relative to the centromere. Chromosome maps called **ideograms** diagram the bands of a chromosome. A human Chromosome 12 ideogram is shown in **Figure 7**.

Briefly, the short arm of the chromosome is called p, the long arm is called q, and the bands are numbered outward from the centromere. So band 12q13.11 is on the long arm of chromosome 12, and it is closer to the centromere than band 12q24.32. The bands on a chromosome do not correspond to individual genes. Each band may include millions of base pairs and hold hundreds of genes.

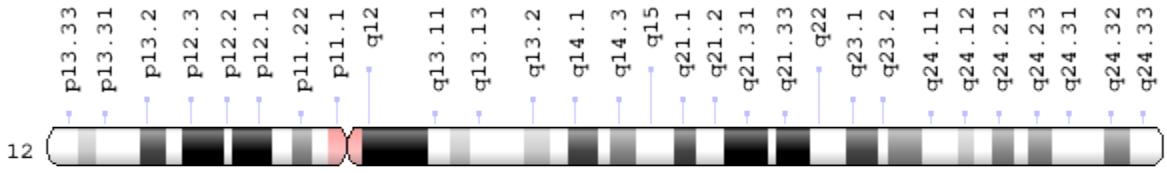


Figure 7. Chromosome 12 ideogram.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=67#h5p-94

Media Attributions

- Ploidy of a cell © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Human gametes © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- FISH (Fluorescence in situ hybridization) © Wikipedia is licensed under a Public Domain license
- Human male karyotpe © Wikipedia is licensed under a Public Domain license
- Human chromosome 12 ideogram © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

EUKARYOTIC GENOMES ARE MOSTLY NON-CODING SEQUENCE

In contrast to prokaryotic genes, which have a relatively high gene density with little interspersed DNA, eukaryotic chromosomes have a much lower gene density. For example, depending on strain, an *E coli* genome may have about 5000 genes spread over 5,000,000 base pairs. This is about one gene per 1,000 base pairs. But for the human genome, with about 20,000 genes in about 3 billion base pairs, there is one gene per 150,000 base pairs!

Less than 2% of the human genome is protein-coding gene sequence. So what's in that remaining 98%?

Some of the 98% is regulatory sequence, used in the control of cellular processes like replication, transcription, and translation. But about 50% of the human genome is repetitive DNA, much of for which we still do not know the function. Other organisms can have even more: for example, about 80% of the maize (corn) genome is repetitive DNA. Repetitive elements can be clustered together in the genome in tandem repeats, or they can be interspersed throughout the genome.

About 13% of the human genome is made up of **short interspersed nuclear elements**, or **SINEs**. SINEs are about 100-400 base pairs long. There are about 1.8 million SINEs in the human genome! This includes over a million copies the most common SINE, the 300 base pair *Alu* sequence4. An additional 20% of the human genome is made up of **LINEs,** or **long interspersed nuclear elements.** The most common human LINE is LINE1, which is about 6,000 base pairs long and repeated about 500,000 times in the genome.

Both SINEs and LINEs are ancient remnants of transposons. Transposons are elements that can move, or "jump", locations in the genome. Although this sounds like science fiction, these "jumping genes" likely played a big role in evolution. Most SINEs and LINEs remaining in the genome of humans and other organisms are no longer are mobile, having accumulated enough mutations over evolutionary time that they do not retain the sequence information necessary for movement. But some still do!

SINEs and LINEs, while not encoding protein themselves, may nevertheless affect the regulation of nearby genes. Although the repetitive DNA in eukaryotic genomes was once thought of as "junk" DNA, there is increasing evidence that these sequences can, in fact, affect gene expression, cell function, and organismal phenotypes.

An interactive H5P element has been excluded from this version of the text. You can view it

online here:
https://rotel.pressbooks.pub/genetics/?p=69#h5p-95

SUMMARY

Bacteria are haploid. A bacterial genome consists of a single, circular chromosome. A bacterial genome has relatively high gene density, consisting mostly of protein coding sequence. Eukaryotic genomes are linear and usually made up of multiple chromosomes. The vast majority of a eukaryotic genome is not protein coding sequence. Much of the noncoding sequence is repetitive DNA, including LINEs and SINEs. A karyotype is a written description of an individual's chromosomes, while a karyogram is a visual representation of the karyotype.

WRAP-UP QUESTIONS

1. Describe three differences between prokaryotic and eukaryotic genomes.

2. Define “ploidy”, and compare the ploidy of a bacterium, a human sperm, and a human skin cell.

3. What percentage of the human genome is protein-coding? Is it true to say that the remaining percentage does not affect an organism’s phenotype? Explain your reasoning.

4. Initially grouped together as “junk” DNA, in recent years more and more data suggests that repetitive DNA can have a profound effect on an organism’s traits. Do a quick websearch to find one example of a SINE or LINE that affects human health.

PART III
MITOSIS AND MEIOSIS

Objectives

1. Describe the difference between a DNA molecule, a chromatid, and a chromosome.
2. List the stages of the cell cycle and describe what happens during each stage.
3. Compare and contrast meiosis and mitosis and recognize under which conditions each occurs.
4. List the stages of meiosis and mitosis and describe what happens during each.
5. Predict the number of chromosomes and DNA molecules and DNA content for a cell at a given stage of meiosis or mitosis.
6. Describe the consequences to the ploidy of a genome of aberrant separation of chromosomes.

Source Material

- Selected images and text modified from Open Genetics Lectures[1]. Chapter 14: Mitosis and the Cell Cycle
- Selected images and text remixed from Online Open Genetics (Nickle and Barrette-Ng), available through Biology LibreTexts[2]. Chapter 2: Chromosomes, Mitosis, and Meiosis

1. Locke, J. 'Open Genetics Lectures' textbook for an Introduction to Molecular Genetics and Heredity (BIOL207). (2017) doi:10.7939/DVN/XMUPO6.

2. Nickle and Barrette-Ng. Open Online Genetics. in Open Online Genetics (2016).

Introduction

Living things all have a genome, the collection of DNA used to store genetic information. But how is that genome transmitted to offspring? This module looks at the process of cell division, including maintenance of the genome from generation to generation.

As a multicellular organism grows from a single cell, every cell contains (nearly) the same genome. That means that the genetic information must be shared as new cells are made. This requires two things: the genome must be copied, and one copy of the genome must be distributed equally to each of the daughter cells as parent cells divide. The process of dividing the chromosomes among daughter cells is called **mitosis**. This is described as **equatorial cell division** since the chromosomes must be *equally* distributed to daughter cells.

Some organisms can reproduce asexually, solely through mitotic or equatorial division. For example, plants can be propagated from roots or cuttings. In these cases, the offspring are genetically identical to the parent.

For organisms that sexually reproduce, genetic content from two parents is combined in their offspring. But each generation has the same number of chromosomes. So sexually-reproducing organisms have an additional need for **reductive cell division** when they produce reproductive cells like egg or sperm. During the process of **meiosis**, cells are produced with only half of the genetic content – it's called reductive cell division because the number of chromosomes has been *reduced*. Then, two cells with half content come together to form a fused cell with a full genome that has genetic information from both parents.

In this module, we will look first at an overview of the cell cycle, during which the cell copies its genome and prepares for cell division. We then look at the process of mitosis, which ensures faithful distribution of chromosomes to daughter cells. We compare that to the process of meiosis, or reductive cell division, which produces cells with half of a diploid genome in preparation for sexual reproduction. And, finally, we explore the consequences of errors in meiosis and mitosis.

THE CELL CYCLE

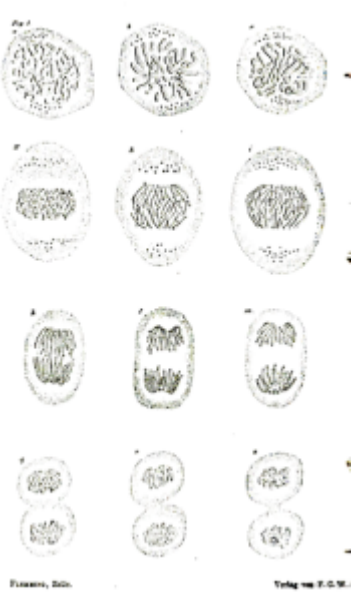


Figure 1. Hand drawn images of dividing cells. From Zellsubstanz, Kern und Zelltheilung, by Walther Fleming, 1882.

If living cells are observed under the microscope, they visibly change in **morphology** (size and shape) as they prepare for and then undergo cell division. One of the first biologists to observe this was Walther Fleming, who developed methods for staining cells that allowed him to see chromosomes. A cell actively undergoing division was easy to identify. It changed shape, pulling away from other nearby cells. The nucleus disappeared, but individual chromosomes became distinguishable from one another and moved through the cell in a distinctive pattern that was consistent from dividing cell to dividing cell. An example of hand-drawn images from his 1882 book, *Zellsubstanz, Kern und Zelltheilung* (*Cell Substance, Nucleus, and Cell Division*) is shown in **Figure 1.**

Fleming observed that the sequence of events drawn in Figure 1 all happened within a one-hour period, ending with the separation of two daughter cells. This cycle would repeat every 24 hours, but there was not much visible change in the intervening 23 hours. These two distinct phases of the cell became known as **mitosis** and **interphase**[1]**.**

The most rapidly dividing eukaryotic cells take around 24 hours to divide, but in other cell types division might only occur every few days (or even weeks or months or not at all!). In those cells, interphase is lengthened, but the process of mitosis is consistently around 1-2 hours.

1. Uzbekov, R. & Prigent, C. A Journey through Time on the Discovery of Cell Cycle Regulation. Cells 11, 704 (2022).

Although there is not much change visible under the light microscopy during interphase, in fact the cell performs a carefully orchestrated series of tasks during this time. Interphase is divided further into the stages of G_1, S, and G_2. This is illustrated in **Figure 2**.

Immediately following mitosis, the cells enter the G_1 (Gap 1) phase of the cell cycle. During this phase, the cell performs normal metabolic functions and grows in size.

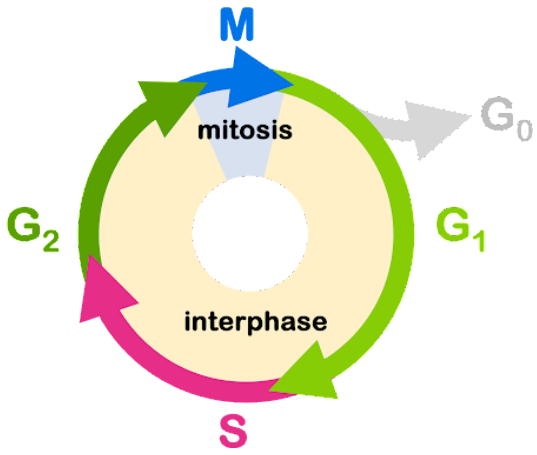


Figure 2. The cell cycle can be divided into 4 stages: G_1, S, G_2, and M, in order. G_1, S, and G_2 collectively make up interphase of the cell cycle. Cells can also exit the cell cycle into the non-dividing phase of G_0.

S phase is next. During S phase of the cell cycle, DNA Synthesis occurs. During this phase of the cell cycle, the cell must copy, or **replicate**, its entire genome in preparation for cell division. One complete copy must be available for each of the two daughter cells produced during cell division. The module on Replication describes this process more in detail. During S phase of the cell cycle, the DNA content of the cell doubles, with each chromosome now consisting of two chromatids. The chromatids remain connected via proteins called **cohesins**, and the total number of chromosomes does not change (**Figure 3**).

In G_2 phase (Gap 2), the cell continues to grow and prepares for cell division. And in M phase (Mitosis), the cell carefully separates the two copies of the genome, partitioning them so the cell can divide into two daughter cells.

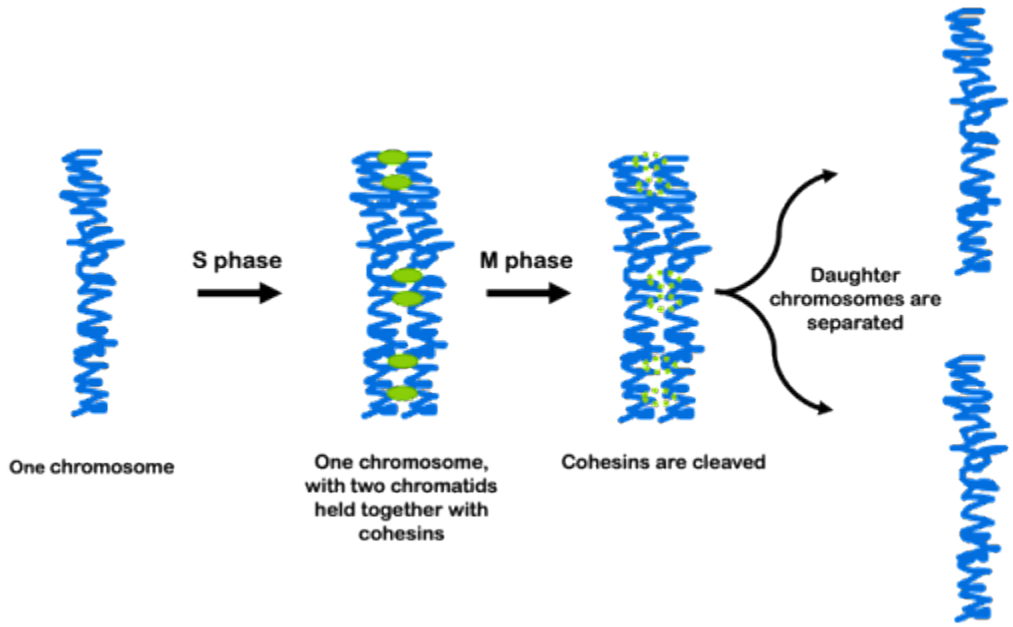


Figure 3. Before DNA replication, each chromosome is composed of one DNA molecule. After replication in S phase of the cell cycle, each chromosome is composed of two DNA molecules. The two halves of a replicated chromosome, called chromatids, are held together by proteins called cohesins until they are separated during M phase.

The progression from one stage to the next is driven by a series of cell cycle proteins called **cyclins**, and it is regulated by several **checkpoints**. At each checkpoint, the cell has mechanisms in place to ensure that conditions are right to move on to the next stage. For example, at the checkpoint between before the transition to S phase, the cell cycle will pause if there are insufficient cellular resources or DNA damage is detected. Replication of damaged DNA could have profound consequences for the cell and its daughters. If the cell proceeds through this checkpoint to S phase, the cell is now committed to cell division. At the G_2/M checkpoint, the cell cycle will pause if DNA is incompletely replicated or if DNA damage is detected. If there are not two complete copies of the genome, this will impact the genomic integrity of daughter cells. There are additional checkpoints during M phase, which are discussed later in this module.

A cell can also exit the cell cycle and enter G_0 phase from G_1. Cells in G_0 are described as **quiescent**. They still perform normal metabolic function, but they do not divide. Some cells can re-enter G_1 phase in response to environmental triggers, but others can remain in G_0 indefinitely: some eukaryotic cells like mature neurons or cardiac muscle cells may remain in G_0 for the lifetime of an individual. This contributes to the difficulty in healing such tissues after injury: damaged cells cannot be regenerated or replenished if they are not dividing.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=80#h5p-2

Media Attributions

- The cell cycle © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- DNA replication © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

DNA CONTENT THROUGH THE CELL CYCLE

The stages of the cell cycle can be assessed experimentally by measuring the DNA content in the cell. As discussed in the module on Genome Structure, a genome of a cell can be described in terms of the ploidy (copies of the genome) as well as the number of chromosomes in the genome. For example, human cells are described as 2n=46, where 2 indicates a diploid genome, and 46 indicates the total number of chromosomes. If 2n = 46, then simple math tells us n=23: the human genome has 23 pairs of chromosomes. While this textbook most often focuses on diploid (2n) cells and organisms, it is important to remember that some organisms have other ploidy: haploid (1n), triploid (3n), tetraploid (4n) and other ploidies exist as well, and are common in some plants.

At the start of the cell cycle in G1 phase, a human cell has DNA chromosomes described as 2n (diploid), each chromosome consists of one DNA molecule, and DNA content described as 2C. But by the end of S phase, DNA has been replicated, and each chromosome consists of two DNA molecules/chromatids (Figure 3). By the end of S phase, content is described as 4C: the number of chromosomes (2n) hasn't changed, but the total DNA content has doubled because each chromosome now consists of two chromatids. The DNA content for cells in each phase of the cell cycle is shown in **Figure 4**.

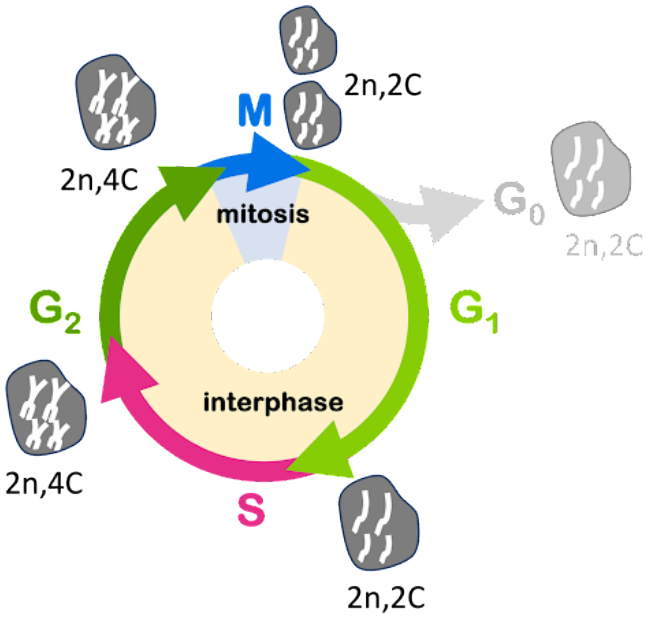


Figure 4. The DNA content of a cell changes with the stage of the cell cycle. After replication, a cell has twice as much DNA content (C) as in G1, although the total number of chromosomes (n) has not changed.

DNA content can be measured experimentally by using techniques like flow cytometry. Flow cytometry can measure the DNA content of single cells by labeling them with a fluorescent dye that binds to DNA: the cells are passed through a sensor that detects fluorescence, and the signal in each cell is measured. "Cyto" refers to cell, so "flow cytometry" refers to the measurement of the cells as they flow through the sensor.

Typically, an experiment measuring DNA content will present data from a population of cells, as shown in **Figure 5**, plotting the number of cells counted vs the fluorescent signal. The x-axis is often presented as DNA content rather than fluorescence, since the fluorescent signal is proportional to the amount of DNA in a cell.

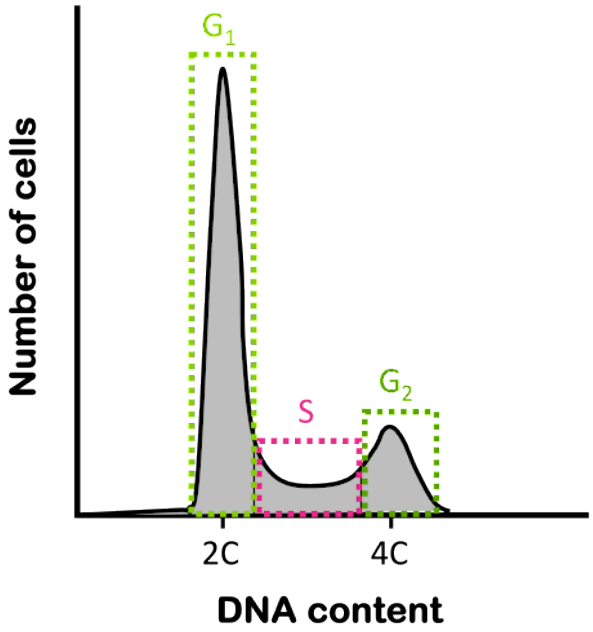


Figure 5. Flow cytometry to measure DNA content. Cells are stained with a dye that fluoresces when in contact with DNA. A flow cytometer measures the fluorescence in one cell at a time. The data are plotted as the number of cells observed vs their DNA content. An asynchronous population of cells typically appears like this, with a large peak representing cells in G1, a smaller peak or cells in G2, and cells in the middle representing S phase (with an intermediate DNA content). Not drawn to scale.

When an asynchronous population of diploid cells is measured for DNA content, a two-peak pattern like that seen in **Figure 5** is observed: the greatest number of cells typically have a DNA content of 2C and are presumed to be in G1 phase of the cell cycle. A smaller number of cells have a DNA content of 4C and are in G2 phase. And some cells have an intermediate amount of DNA content: these are in S phase of the cell cycle and are in the process of making additional DNA. The abundance of cells with each DNA content is proportional to the length of each stage of the cell cycle: if about 40% of the cell cycle is spent in G1 phase,

about 40% of cells in an asynchronous population will be in G1 phase. This technique is often used to measure the effects of experimental manipulation on the cell cycle. It also can be used to measure the ploidy of a cell.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=84#h5p-3

Media Attributions

- DNA content of a cell © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- DNA content graph © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

M PHASE: MITOSIS

During mitosis, or M phase, the cell separates the duplicated copies of the genome and then separates the cytoplasm via cytokinesis. The mitotic cell cycle is how new somatic (nonreproductive) cells are generated in multicellular organisms. It is important during development and as cells are replenished due to age and injury in a mature organism.

The process of mitosis requires that each daughter cell have the same genome as the parent: exactly one copy of each chromosome. To make this happen, the cell keeps sister chromatids together after replication, connected via **cohesion** proteins**.** The chromatids then only separate during mitosis. Mitosis has four distinct stages: prophase, metaphase, anaphase, and telophase, with the physical division of the daughter cells, cytokinesis, beginning simultaneously with events of telophase. Some textbooks may count differently, including an intermediate prometaphase as a fifth stage, and some may count cytokinesis as a sixth. The stages of mitosis are shown in **Figure 6**.

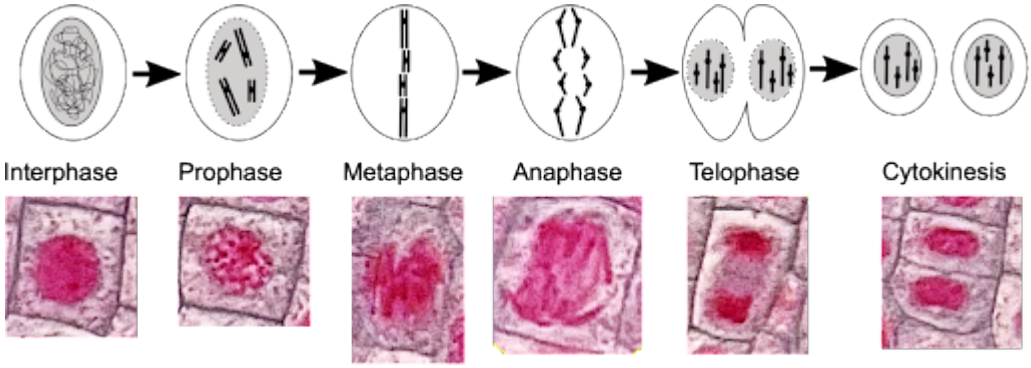


Figure 6. Stages of interphase and mitosis. Top: drawing of each stage of mitosis, with a 2n = 4 cell. Modified from Original-M. Deyholos/L. Canham-CC:AN. Bottom: Microscope images of onion root cells in each stage of mitosis.

During interphase of the cell cycle, the nucleus is clearly visible within the cell. Individual chromosomes are not visible, and the nuclear membrane forms a barrier between DNA inside the nucleus and the cytosol. During the **prophase** of mitosis, the chromosomes condense, becoming more compact. At this stage, they begin to be visible under a microscope as individual chromosomes. The nuclear membrane dissolves, and there is no distinct separation from the cytosol.

In late prophase (sometimes distinguished as **prometaphase**), the chromosomes begin to migrate toward the center of the cell. In **metaphase**, the chromosomes align at the center (or **equatorial plate**) with one chromatid oriented toward each pole of the cell.

The movement and alignment of the chromosomes requires the action of the **mitotic spindle**. The mitotic spindle is made up of microtubules, long polymeric proteins which dynamically rearrange in mitosis to form the spindle. A cartoon of the spindle is shown in **Figure 7**. The microtubules (green) extend outward from microtubule organizing centers at the poles of each cell. Microtubules of the spindle attach to the chromosomes via **kinetochore proteins** (pink dots) that bind to chromosomes (blue) near the centromere. Spindle formation also requires the assembly of polar microtubules that extend outward from the poles but do not directly contact the chromosomes.

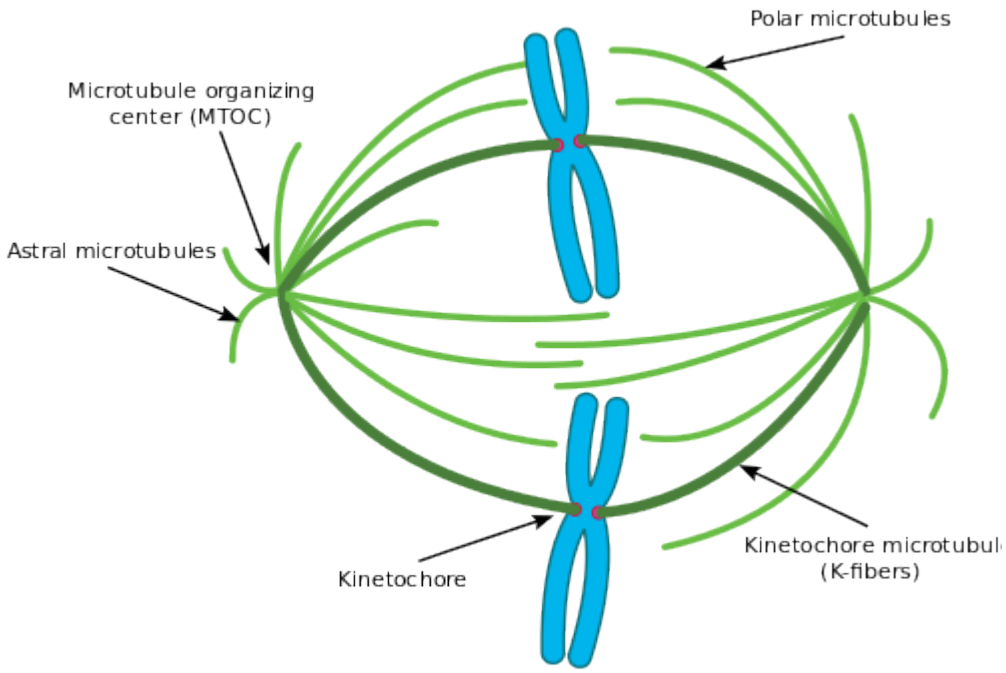


Figure 7. Mitotic spindle apparatus.

You can see the spindle via fluorescent microscopy in **Figure 8**. In Figure 8, microtubules are stained green and DNA is stained blue in sequential images that show metaphase, late anaphase, and cytokinesis. Note: The spindle is not seen in the cells in Figure 6, because the spindle proteins are not stained with the techniques used.

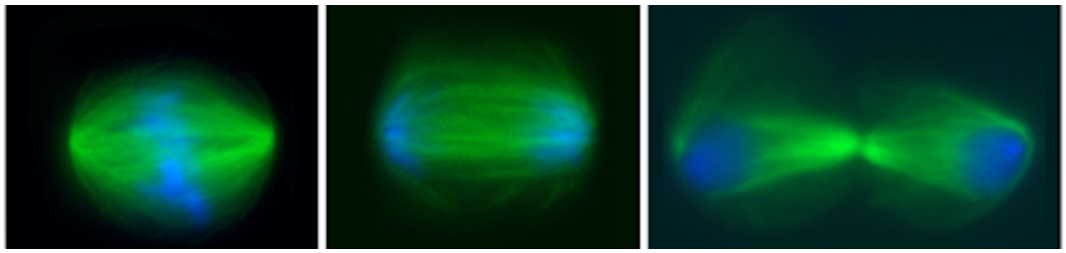


Figure 8. Mitosis in Drosophila melanogaster showing fluorescently labeled chromosomes (blue) and microtubules (green) at metaphase, anaphase and telophase (from left to right).

Once the chromosomes have assembled at the equatorial plate, the cohesins connecting the chromatids are dissolved. This allows the separation of chromatids, which are pulled to opposite poles of the cell by the mitotic spindle. This stage is called **anaphase**. In **telophase**, the separated chromatids (now unreplicated chromosomes) arrive at the poles of the cells, and nuclei reform.

Division of the cytoplasm is called **cytokinesis**, and it begins even as the separated chromatids are arriving at the poles. Because cytokinesis refers specifically to the splitting of the cytoplasm, it is considered distinct from telophase even though the two events happen simultaneously. Cytokinesis happens via different mechanisms dependent on cell type. Animal cells undergo cytokinesis as the cytosol is pinched off by a contractile ring of cytoskeletal proteins, as seen in the right panel of Figure 8. Plant cells separate by forming a new cell wall from vesicles that accumulate at the center of the cell. (This is seen in the telophase and cytokinesis cells shown in **Figure 6**.)

Accurate distribution of chromosomes to daughter cells requires the alignment of chromosomes at the equatorial plate, attachment to the mitotic spindle, and full separation of the chromatids. The M-phase checkpoint ensures that the chromosomes are properly attached to the spindle. If they are not properly attached, mitosis will pause.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=89#h5p-4

MEIOSIS

In organisms that replicate sexually, in each new generation genetic information is remixed and recombined from two parental cells. This happens when two parental **gametes** – each of which contain half of a genome – fuse together.

A cell about to undergo meiosis is called a **meiocyte.** Meiosis is mechanistically like mitosis but involves two rounds of cell division. In meiosis I, homologous chromosome pairs separate and are distributed among daughter cells. Note: this process is often described as **segregation** of the homologs. This word has a very negative meaning in a modern social context, referring primarily to segregation of ethnic or racial groups of people. In a biological context, it means a physical separation and compartmentalization. It is a holdover from work like Gregor Mendel's in the 1860's. He used the word to describe of the equal segregation of dominant and recessive characters among offspring.

Meiosis I is the reductional stage of mitosis since it reduces the total number of chromosomes. In meiosis II, sister chromatids separate, just as in mitosis. Meiosis II is an equatorial division just like mitosis, since it does not change chromosome number. Meiosis I and II both follow the same general sequence of events as mitosis: prophase, metaphase, anaphase, and telophase. In nomenclature, these stages are distinguished from one another by the addition of the Roman numerals I or II. For example, "prophase", with no Roman numeral, indicates the stage of mitosis. "Prophase I" indicates prophase of meiosis I, and "prophase II" indicates prophase of meiosis II. Meiosis I and II include two rounds of cytokinesis as well: for each parent meiocyte, there are thus four daughter cells produced. An overview of the processes of Meiosis I and Meiosis II is in **Figure 9**.

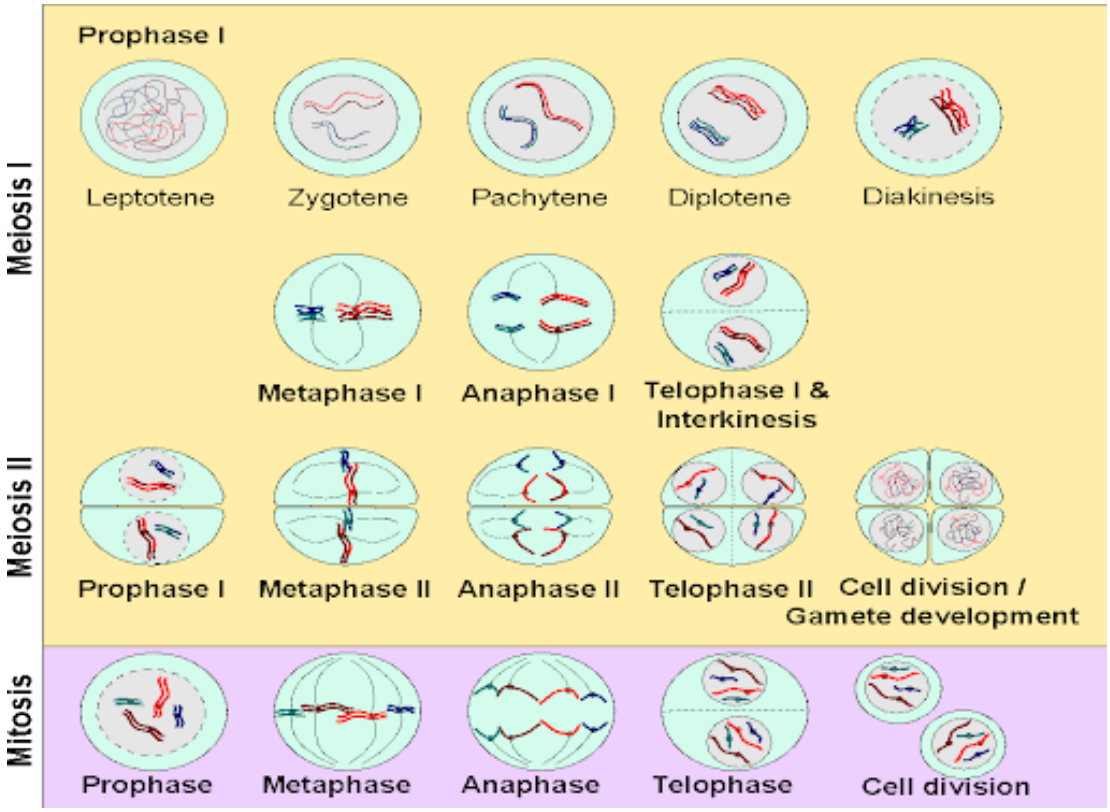


Figure 9. The stages of meiosis I and meiosis II. This example uses a diploid animal with 2 chromosome sets, so 4 chromosomes in total: Red, Maroon, Blue and Teal. Crossover events are shown between the two closest non-sister chromatids, but in reality can happen between all four chromatids. Prophase I is divided into stages. Leptotene is defined by the beginning of chromosome condensation, though chromosomes are still long. Zygotene chromosomes are still long, but you can readily identify chromosomes as they are starting to pair. Pachytene chromosomes are thickening and fully synapsed. During Diplotene one can begin to see the individual chromatids and chiasmata. In Diakinesis, chromosomes are fully condensed and the nuclear membrane dissolves. In Metaphase I, the synapsed chromosomes align along the metaphase plate and then the synapse breaks in Anaphase I. Meiosis I is completed with Telophase I and potentially interkinesis, completing the reductional division. Meiosis II is an equational division where the chromosomes align in Metaphase II similarly to Mitosis and complete Anaphase II and Telophase II, leaving 4 haploid gametes formed.

The biggest difference between Prophase I and mitotic Prophase is that, during Prophase I, homologous chromosomes **synapse**, or pair, forming a **synaptonemal complex**. This holds the homologs together and is important for the proper separation of chromosomes, ensuring that exactly one chromosome from each pair is destined for each daughter cell. It is also important for the exchange of genetic information between the homologous pairs, called crossing over or meiotic **recombination**.

The condensation of chromosomes and formation of the synaptonemal complex can be observed microscopically, and so Prophase I can be further divided based on those events. **Leptotene** is the first stage after interphase, when chromosomes begin to condense. Next is the **zygotene** stage, when homologous pairs

come together to form the synaptonemal complex. During the **pachytene** stage, the chromosomes are paired all along their length, and there is crossing over of genetic information between the homologous chromosomes. The pairs begin to partially separate during **diplotene**, making crossovers visible in the space between the pairs. The nuclear envelope dissolves during **diakinesis**.

Because genetic information is exchanged reciprocally between parental chromosomes, crossing over (**Figure 10**) means that chromosomes do not have to be inherited intact from generation to generation, but it does mean that no genetic information is lost. Crossing over thus plays an important role in increasing genetic diversity while maintaining genomic integrity. Although only one crossover is shown per chromosome in **Figure 10**, multiple crossovers can and do occur between homologs. The daughter chromosome could be a patchwork sequence from one or the other parental chromosome.

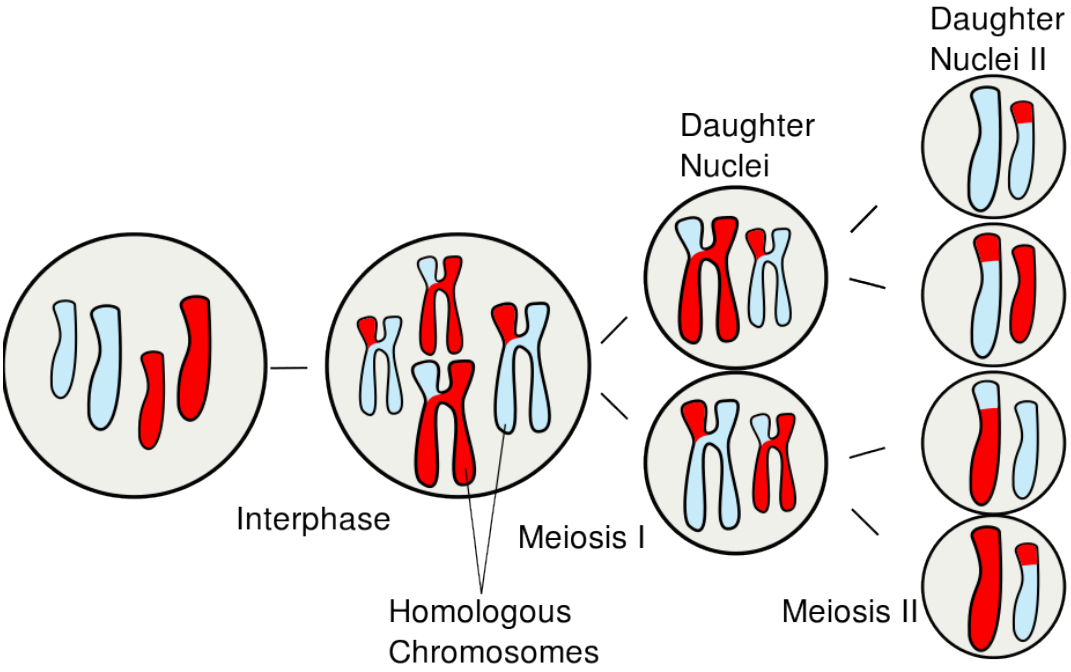


Figure 10. Crossing over during meiosis results in recombinant chromosomes with a combination of genetic information from both parents.

Following the diakinesis stage is **metaphase I**. As in the metaphase of mitosis, the chromosomes align on the equatorial plate of the cell. However, because the homologous chromosomes are paired, they orient in a way that chromosome pairs (rather than sister chromatids) are destined to be partitioned. The alignment of the pairs is what determines the combination of alleles in the daughter cells. With the transition to **anaphase I**, the homologous pairs separate and migrate toward their respective poles of the cell. During **telophase I**, nuclei reform and cytokinesis may separate the two cells into daughters. Because only one chromosome from each pair is distributed to each daughter cell, the daughter cells are haploid (1n, 2C) as they enter meiosis II.

The stages of meiosis II are much like mitosis. Chromosomes condense in **prophase II** and align on the equatorial plate during **metaphase II.** Chromatids separate and segregate during **anaphase II**, and nuclei reform and the cells separate during **telophase II** and cytokinesis. At the end of meiosis II, the DNA content of the daughter cells is 1n, 2C.

The stages of meiosis and mitosis are summarized in **Figure 11**, with DNA content for each stage indicated.

As in mitosis, checkpoints during meiosis ensure that each step is completed before the cell divides. Meiotic checkpoints monitor for synapse formation and resolution of crossovers, preventing cycle progression if these steps are not completed appropriately[1].

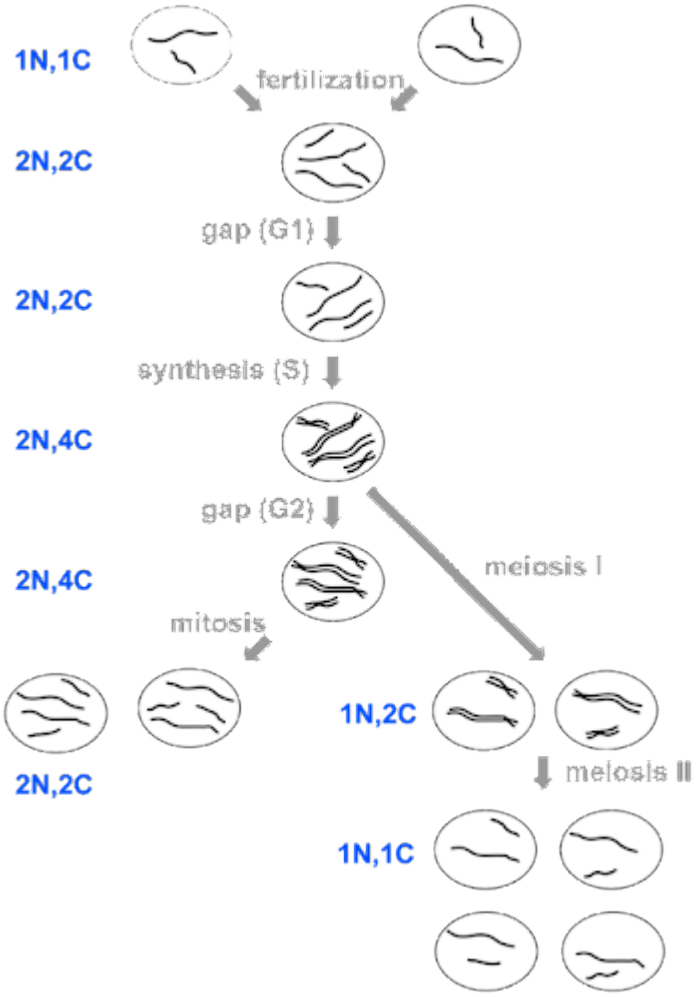


Figure 11. DNA content during meiosis and mitosis.

1. Subramanian, V. V. & Hochwagen, A. The Meiotic Checkpoint Network: Step-by-Step through Meiotic Prophase. Cold Spring Harb. Perspect. Biol. 6, a016675 (2014).

Figure 12. A mule is an infertile hybrid of a horse and a donkey.

Interestingly, these checkpoints may affect fertility in hybrid offspring from different species. For example, a mule is a hybrid of a horse (2n = 64) and a donkey (2n = 62), with 63 chromosomes. The odd number of chromosomes cannot pair during meiosis I, and the failure of synapse leads to a block to gametogenesis in mid-meiosis[2].

Media Attributions

- The stages of meiosis I and II © Open Genetics Lectures, Chapter 15 is licensed under a CC BY-NC (Attribution NonCommercial) license

2. Han, H. et al. Testicular Characteristics and the Block to Spermatogenesis in Mature Hinny. Asian-Australas. J. Anim. Sci. 29, 793–800 (2016).

MATURATION OF GAMETES AND ORGANISM LIFE CYCLE

The general steps of meiosis and mitosis are similar from cell to cell and organism to organism. But the specifics can vary: Although meiosis involves two cell divisions and ends with four haploid cells, the fate of those cells can be quite different from species to species, and even among sexes.

In gonadal male animals, the meiocytes are called primary spermatocytes. The products of meiosis I are called secondary spermatocytes, and the four products of meiosis II are called spermatids. The spermatids mature, growing tails and becoming functional sperm cells.

In female animals, the meiocytes are called primary oocytes. The primary oocytes begin meiosis I in the ovaries before birth, but they arrest in prophase I. After puberty, meiosis continues, but the products of meiosis I are not equal. To maximize the nutrients destined for the eventual egg cell, the primary oocyte divides asymmetrically into a secondary oocyte and a small, nonviable cell called a **polar body**. The polar body gets half of the chromosomes, but not much else: The much-larger secondary oocyte gets almost all the cell resources, including most organelles of the cell. The secondary oocyte arrests in metaphase II until after ovulation – and even until after fertilization! If the oocyte is fertilized, this triggers the completion of meiosis II. A second polar body is formed, with the fertilized egg again keeping most of the cellular resources. In this manner, even though there are two meiotic divisions, only one functional gamete is produced from meiosis in gonadal females. (**Figure 13**)

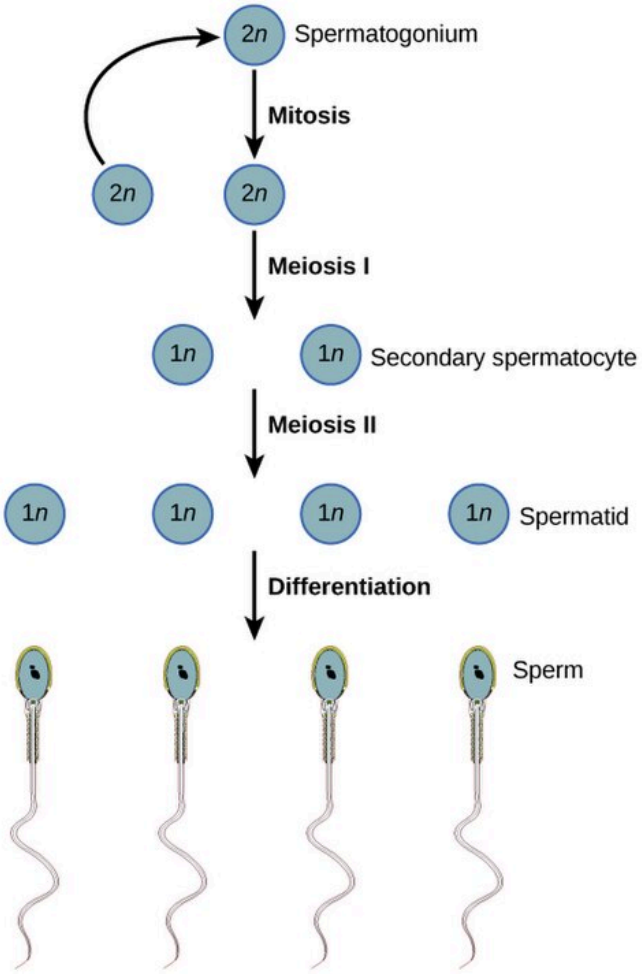


Figure 13. Comparison of meiosis of spermatocytes and oocytes.

In other species, there are further differences. For example, in plants haploid gametophytes undergo additional rounds of the mitotic cell cycle before they mature into pollen grains, which are the male reproductive cells. And in fungi, the organism spends much of its life cycle haploid, only forming a transient diploid cell for reproduction.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=98#h5p-5

CHANGES IN PLOIDY DUE TO MIS-SEGREGATION OF CHROMOSOMES

Improper meiosis or mitosis can result in cells with abnormal chromosome numbers. These are genetic mutations that affect entire chromosomes and therefore affect many genes. They can have a profound effect on the phenotype of a cell or of an organism.

If an error in chromosome number happens during meiosis, it will result in a **germline** mutation in a gamete. The mutation will be present in every cell of an offspring produced from that gamete. If an error in chromosome number happens during mitosis, it will result in a **somatic** mutation. Only the mitotic daughter cell – and its daughter cells – will be affected. This results in an organism with a **mosaic** genome: some cells will have a different genome than others. **Somatic** mutations cannot be passed to offspring through sexual reproduction, since meiocytes would not have the mutation.

Aneuploidy

Aneuploidy results when an extra chromosome or part of a chromosome is gained or lost from the cell. For example, **monosomy** describes a genome that only has one copy of one of the chromosomes. **Trisomy** means there are three copies of one of the chromosomes. A partial aneuploidy affects only part of a chromosome, with an extra (or missing) part of a chromosome.

Down syndrome, or Trisomy 21, is an example of an aneuploidy in humans, occurring in about 1/700 births. Sex chromosome aneuploidies also occur relatively often in the human population: as many as 1/500 people may have a sex chromosome aneuploidy. Other germline aneuploidies are more uncommon in humans: most result in embryonic death early in development. This is likely connected to the number of genes that are unbalanced by the extra or missing chromosome. Chromosome 21 is the smallest number of human chromosomes, and thus has the fewest number of genes impacted, making trisomy-21 less lethal than other trisomies that impact more genes.

Aneuploidies can result from **nondisjunction** during either mitosis or meiosis: the failure of either chromatids or chromosome pairs to separate before cytokinesis. Nondisjunction seems to happen more frequently with age in humans.

Figure 14 shows examples of nondisjunction in meiosis I (right) and meiosis II (left). Nondisjunction will always result in one daughter cell having an extra copy of a chromosome, and the other a missing one. It can happen in oogenesis or spermatogenesis, and if the gamete forms a zygote, the resulting offspring will have a change in chromosome number in every cell of their body.

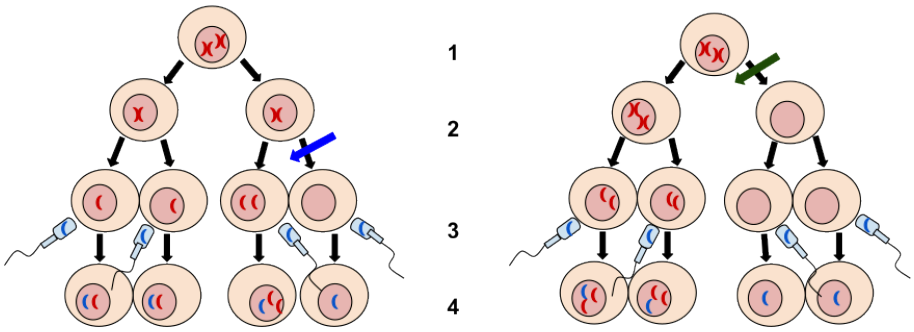


Figure 14. Nondisjunction can occur during meiosis I (right), meiosis II (left), or mitosis, resulting in daughter cells with too many or too few copies of a chromosome.

If nondisjunction happens during mitosis, it will result in a somatic mutation. A somatic mutation affects the cell and any mitotic daughter cells, making the individual **mosaic** for the mutation, but it will not be passed to offspring unless it specifically affects the reproductive cells. As an example in humans, some people have mosaic Down syndrome: some, but not all, of their cells have an extra copy of chromosome 21. This can happen through mis-segregation during mitosis early in development. All subsequent mitoses will produce daughter cells with an extra chromosome. Mosaic Down syndrome can also occur if nondisjunction in meiosis results in an embryo with trisomy 21, and during subsequent mitosis one of the cells of the embryo loses the extra copy of chromosome 21.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=102#h5p-6

Polyploidy

An extra copy of one chromosome (or part of a chromosome) is an aneuploidy, but **polyploidy** results from the addition of an entire copy of the genome. Remember that cells with two copies of the genome are diploid. Cells could lose a copy of the genome and become haploid, or they can gain additional copies of the chromosomes.

Polyploidy might result when a cell undergoes replication but fails to undergo cytokinesis in mitosis, or through a failure of meiosis that does not reduce chromosome number. This can happen through a number

of mechanisms, including a lack of cytokinesis, misalignment of spindle poles, or an atypical meiosis that skips either meiosis I or meiosis II[1].

Figure 15. A chocolate chip cookie.

Although aneuploidies typically cause reduced viability, in some species, especially plants, polyploidy is possible. This is likely because of gene balance: a biological pathway depends on having the correct ratios of interacting genes – too many or too few of interacting partners will affect function. This is kind of like having too many chocolate chips and not enough cookie dough to hold them together. But having a whole extra set can be better tolerated (you can double a recipe just fine).

We see this in a lot of cultivated plants. For example, most commercially available strawberries are octoploid, and some wheat strains are hexaploid.

Geneticists need a way to distinguish between the genomes of somatic cells and gametes, even in organisms with greater ploidy. Such organisms still make gametes with half as many copies of the genome. For example, octoploid strawberries produce tetraploid gametes, and hexaploid wheat strains produce triploid gametes.

So here's where notation can get a little confusing: Polyploids are described using the letter "X" to indicate ploidy. And "n" indicates the number of chromosomes in a gamete. A somatic cell is always described as 2n, and a gamete = n.

Humans (diploid) are 2n= 2x = 46, since somatic cells have two copies of each chromosome. Octoploid strawberries are 2n = 8x = 56, and n = 4x = 28, with 8 copies of the genome in somatic cells and 4 copies of the genome in gametes.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=102#h5p-7

Media Attributions

- The-cell-cycle-194×300-1

1. De Storme, N. & Geelen, D. Sexual polyploidization in plants – cytological mechanisms and molecular regulation. New Phytol. 198, 670–684 (2013).

DIFFERENCES IN PLOIDY AND THE IMPACT TO MEIOSIS

Polyploid organisms with even-numbered ploidies (2x, 4x, 6x, 8x) can produce gametes via meiosis with pairs of chromosomes assembling in prophase I as bivalents, or multiples assembling in a higher-order association (e.g. a tetravalent for a tetraploid organism). But cells with odd-numbered ploidies or aneuploidies cannot typically undergo meiosis, since an odd number of chromosomes cannot be evenly divided during meiosis I.

Cells with odd-ploidies can, however, undergo *mitosis*, and so we do see examples of viable monoploid and triploid organisms – they just aren't fertile for sexual reproduction.

An example of a triploid organism is the most common variety of bananas seen in U.S. grocery stores. Wild bananas are generally smaller than the Cavendish variety found in grocery stores. Wild bananas also have seeds, as shown in **Figure 16**.

Figure 16. Wild bananas are smaller and seeded compared to Cavendish bananas.

The Cavendish variety is larger and triploid. They most likely arose from the fusion of diploid gamete with a haploid gamete of a wild banana like the one in Figure 15. But because they are triploid, they are sterile: they do not undergo meiosis, and they are therefore seedless. Cavendish bananas are propagated from suckers that grow out of the base of a banana plant. These suckers are genetically identical to the parent because they grow from mitotic cell divisions rather than meiosis. About 47% of bananas grown worldwide – and about 99% of bananas produced for export – are Cavendish bananas. And they are all genetically identical[1].

Some trivia! Some seedless fruits have similar genetics to the banana: they've arisen from the fusion of gametes with differing ploidy, and they cannot undergo meiosis so do not produce seeds. Seedless watermelon is like this. But others, like seedless grapes, are different. Seedless grapes have a disruption in either meiosis or seed maturation, but it is due to different underlying genetic mechanisms unrelated to ploidy.

Odd-number genomes are rare in animals, but an exception is seen in many species of bees, wasps, and ants,

1. World's Bananas Are in 'Imminent Danger' of Disease. Newsweek https://www.newsweek.com/worlds-bananas-are-clones-and-they-are-imminent-danger-publish-monday-5am-1321787 (2019).

which use ploidy to determine sex. In these animals, males develop from unfertilized eggs and are haploid (or monoploid). Females develop from fertilized eggs and are diploid. Male cells cannot undergo meiosis since there are no chromosome pairs. Instead, sperm are produced via mitosis. Diploid female bees use meiosis to produce eggs.

SUMMARY OF THE CELL CYCLE

- Mitosis is equatorial cell division that results in two genetically identical daughter cells. It is used to produce new cells in a multicellular organism and in asexual reproduction. Meiosis is reductive cell division used to produce gametes for sexual reproduction. It produces four daughter cells with half the DNA content of the parent cell.
- The cell cycle consists of G_1, S, G_2, and M stages. G_1, S, and G_2 stages together make up an interphase. The DNA content of a cell doubles during S phase, when replication occurs, but chromosome number does not change because chromatids stay connected. Chromatids separate in M phase, or mitosis.
- Mitosis consists of prophase, metaphase, anaphase, and telophase, followed by cytokinese. Meiosis consists of two divisions: meiosis I and meiosis II. Meiosis I is divided into prophase I, metaphase I, anaphase I, and telophase I. Homologous pairs separate in meiosis I, resulting in haploid daughter cells. Meiosis II is divided into prophase II, metaphase II, anaphase II, and telophase II. Sister chromatids separate in meiosis II.
- The process of meiosis is different between sexes, with all four products of meiosis producing viable sperm, but only one product of meiosis producing a viable egg.
- Nondisjunction of chromosomes and other errors during mitosis and meiosis can lead to aneuploidies and polyploids. Polyploid genomes are common in plants. Aneuploid cells and cells with an odd ploidy cannot undergo meiosis because they arrest in meiosis I when chromosomes cannot form pairs.

WRAP-UP QUESTIONS

1. What is the difference between a chromosome, a chromatid, and a DNA molecule?
2. The human genome in somatic cells is 2n=46.

a. What is the chromosome number in human egg and sperm cells?
b. For each of the following stages of the cell cycle, indicate how many chromosomes and how many DNA molecules would be in a human cell.

- G_1
- G_2
- Prophase I of mitosis
- Prophase II of mitosis

3. The mouse genome is 2n=40. For each of the following stages of meiosis or mitosis, indicate chromosome number (n) and DNA content (C).

a. Somatic cell in G_1 phase of the cell cycle
b. Somatic cell in G_2 phase of the cell cycle
c. Primary spermatocyte in Prophase I
d. Secondary spermatocyte in Prophase II
e. Spermatozoa (sperm cell) after meiosis II is complete

4. **Figure 17** shows flow cytometric analysis of DNA content for human colon cancer cells treated with several different drugs that affect the cell cycle8. Etoposide and Paclitaxel are chemotherapy drugs currently in use. VX-680 was in clinical trials for leukemia and colon cancer but was discontinued.

a. In the DMSO control image, label the areas that correspond to G_1, S, and G_2 phases of the cell cycle.
b. The x-axis indicates DNA content (C). What scale do each of the hash marks represent?
c. Etoposide triggers a checkpoint-induced arrest of the cell cycle[1]. Based on these data, at which

1. Massey, A. J. Multiparametric Cell Cycle Analysis Using the Operetta High-Content Imager and Harmony Software with PhenoLOGIC. PLOS ONE 10, e0134306 (2015).

checkpoint do you think the cell cycle was blocked? Explain your reasoning.

d. Both Paclitaxel and VX-680 induce the appearance of additional peaks, indicated by the arrows in Figure 17. What do you think those peaks represent, and what might have happened to the cell cycle to cause these results?

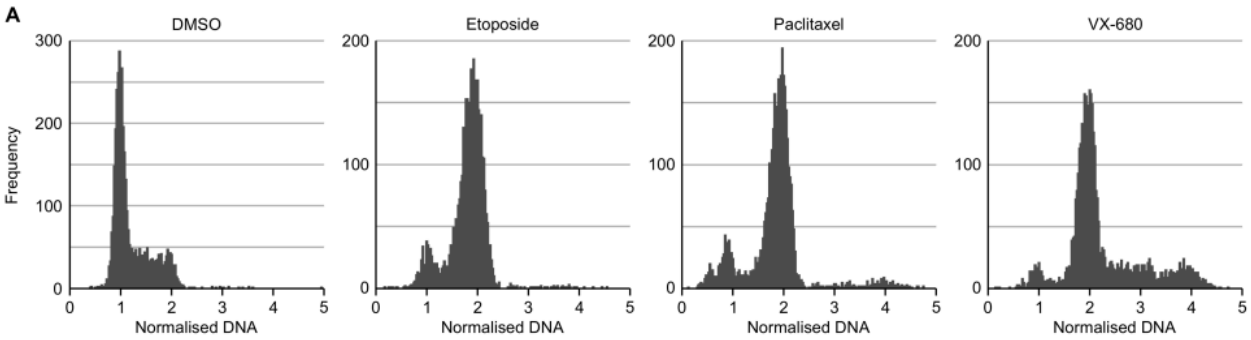


Figure 17. Flow cytometry analysis of colon cancer cells after drug treatment. DMSO is a control.

5. Double strand breaks in chromosomes can result in afragment of a chromosome without a centromere, as shown in **Figure 18**.

a. What role does the centromere play during mitosis and meiosis?
b. What might be the consequence to daughter cells if mitosis occurred before this damage could be repaired?

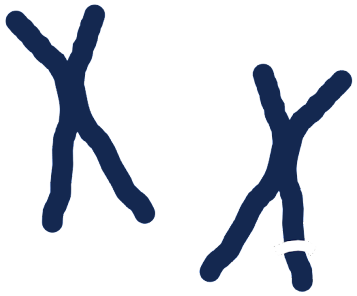


Figure 18. Double strand break separates part of a chromosome from the centromere.

6. Some people have mosaic Down Syndrome: they have some cells with 46 chromosomes (46, XX or 46, XY) and some with 47 (47, XX +21 or 47, XY+21). Nondisjunction in what cell types might cause mosaic Down Syndrome? Explain your reasoning.

7. Seedless watermelons are triploid, offspring of a cross between diploid and tetraploid parents. They are seedless because the odd-ploidy genome cannot complete meiosis. The tetraploid watermelons are produced by treating diploid

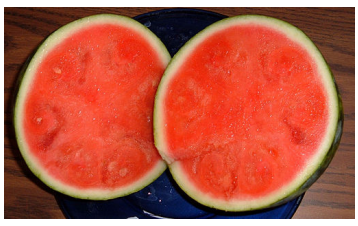
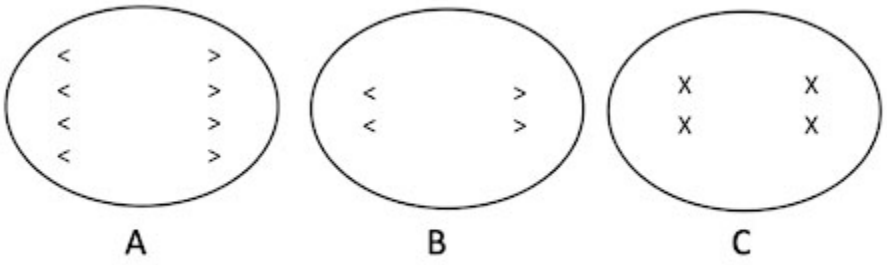


Figure 19. Figure 20 Seedless watermelon.

watermelon shoots with colchicine, a drug that disrupts microtubules[2]. Explain why disrupting microtubules could lead to tetraploidy in some treated plants.

8. The illustrations below show dividing cells from the same organism.

a. Indicate whether each illustration shows anaphase, anaphase I, or anaphase II.
b. How would you describe the ploidy of the parent organism: 2n=?



The following questions were adapted from *Online Open Genetics (Nickle and Barrett-Ng)*, Chapter 2.

9. For a diploid with 2n=16 chromosomes, how many chromosomes and chromatids are per cell present in the gamete and zygote and immediately following G1, S, G2, mitosis, and meiosis? Bread wheat (*Triticum aestivum*) is a hexaploid. Using the nomenclature presented in class, an ovum cell of wheat has n=21 chromosomes. How many chromosomes in a zygote of bread wheat?

10. Bread wheat (*Triticum aestivum*) is a hexaploid. Using the nomenclature presented in class, an ovum cell of wheat has n=21 chromosomes. How many chromosomes in a zygote of bread wheat?

11. For a given gene:

a. What is the maximum number of alleles that can exist in a 2n cell of a given diploid individual?
b. What is the maximum number of alleles that can exist in a 1n cell of a tetraploid individual?

2. Seedless Watermelon Breeding – Cucurbit Breeding. https://cucurbitbreeding.wordpress.ncsu.edu/watermelon-breeding/seedless-watermelon-breeding/.

c. What is the maximum number of alleles that can exist in a 2n cell of a tetraploid individual?

12. Why is aneuploidy more often lethal than polyploidy?

13. Which is more likely to disrupt gene balance: polyploidy or duplication?

14. For a diploid organism with 2n=4 chromosomes, draw a diagram of all of the possible configurations of chromosomes during normal anaphase I, with the maternally and paternally derived chromosomes labeled.

Science and Society

15. Virtually all of the bananas sold in grocery stores are genetically identical Cavendish bananas, but this was not always true. Up until the mid-1950's, the Gros Michel banana was the main variety of banana sold in the US. The Gros Michel bananas, like the Cavendish, are triploid and can only be propagated through cuttings.

a. Why can organisms with an odd number ploidy typically not reproduce via sexual reproduction?
b. What are the risks of a clonal population? To answer this, use outside sources to find out why the Gros Michel banana is no longer the top banana.
c. What is the evolutionary advantage to sexual reproduction?

PART IV

OVERVIEW OF CENTRAL DOGMA AND REPLICATION

Objectives

1. Describe the chemistry of DNA replication and the function of the players involved, including leading strand, lagging strand, Okazaki fragment, DNA pol I, DNA pol III, primase, primer, sliding clamp, clamp loader, single strand binding protein, helicase, ligase
2. Draw a replication bubble and/or replication fork, with leading and lagging strands and all 5' and 3' ends labeled.
3. Explain why eukaryotic chromosomes get shorter with every round of replication
4. Explain the process by which telomerase extends the ends of chromosomes after replication, and recognize that telomerase is only active in certain cell types

Source material

- Replication and Transcription introduction remixed from "Book: Cells – Molecules and Mechanisms (Wong)" by E. V. Wong, LibreTexts is licensed under CC BY-NC-SA. Available free here: https://bio.libretexts.org/@go/page/16085. Selections from Chapter 7 (DNA) and Chapter 8 (Transcription)[1].

1. Book: Cells – Molecules and Mechanisms (Wong). Biology LibreTexts https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Cells_-_Molecules_and_Mechanisms_(Wong) (2018).

- DNA is semi-conservative, remixed from Online Open Genetics (Nickle and Barrette-Ng), available through Biology LibreTexts[2].

Introduction

DNA is a genetic information storage system: it holds the information needed to build an organism. The genetic information storage system has two main requirements. First, the information must be heritable, able to be passed from one generation to the next. Second, the information must be accessible and able to be used by the organism without damaging or destroying the information itself. In the next two modules, Replication and Transcription, we will talk about how these two things are accomplished.

But DNA itself is not a molecule that "does things". Instead, a collection of other biologically active molecules act upon DNA to read the information and build protein molecules according to instructions in DNA. It is largely the protein molecules that "do" the jobs needed by cells and organisms. So how do we get from DNA to protein? And how is the information shared from parent to offspring?

Recall from the DNA structure module that DNA is composed of two intertwined, complementary strands. If those strands are separated, it's easy to determine from the sequence of one strand what the sequence of the missing strand should be. You can practice this below.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=117#h5p-8

2. Nickle and Barrette-Ng. Open Online Genetics. in Open Online Genetics (2016).

Cells do something similar when they **replicate**, or copy, their DNA. The strands of the double-helix are separated, and two new strands are constructed by creating a complement for the old ones. The old strands are thus used as **templates** for new strands (**Figure 1**). The new strands are called **daughter** strands. Because each daughter strand retains (or conserves) one intact strand from the parent, DNA replication is said to be **semi-conservative.**

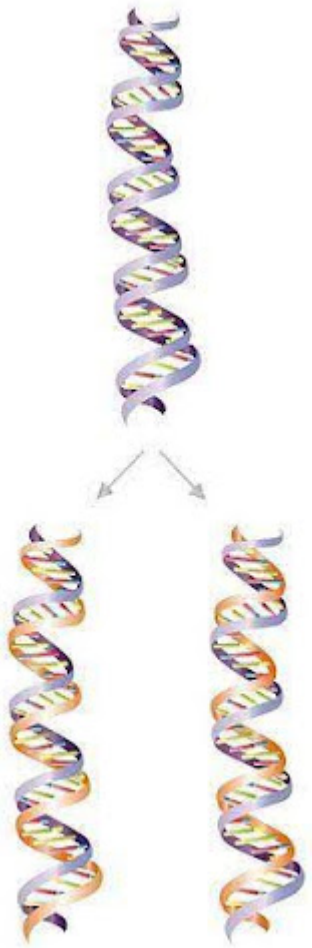


Figure 1. DNA replication is semi-conservative.

Thus, two daughter double helices result from one round of replication. When a cell divides, one daughter double helix is passed to each of the two daughter cells. In this way, genetic information is passed from parent to offspring.

So how is the information used? The DNA sequence is not random: it is organized into discrete units called **genes**. To use an analogy, the characters you are reading on this page are not a random collection of letters, spaces, numbers, and symbols: within the long string of characters are recognizable words and sentences.

Likewise, the genes within DNA contain recognizable sequences that flag parts of the DNA to be **transcribed** into RNA. During the process of **transcription**, segments of DNA are read by transcription enzymes, and a molecule of RNA is constructed to be complementary to the parent DNA. In this way, the genetic information stored in the DNA molecule is transferred to a new chemical form. It's almost like making a photocopy of a page of a library book: the book itself is not altered, but an extra copy of the information is created. The new copy is not exactly the same form as the original – it's not bound in a hardcover book, it's only one page and not the whole text, and the paper probably feels a little different – but in many ways it's also much more useable. You can fold up the photocopy and carry it with you out of the library, you can make notes on it, etc.

Some RNA molecules have specific roles in the cell. Some are enzymes, some regulate cell functions. We will discuss some of these functional RNAs in upcoming modules. But other RNA molecules are specifically used as templates to build proteins. The process of protein synthesis is called **translation.** To continue with the analogy: if the book you'd photocopied was a cookbook, the protein might be the food you cooked from the recipe.

These three processes together: replication, transcription, and translation, are called the **Central Dogma** of molecular genetics. The Central Dogma describes the flow of information from DNA: from DNA to daughter DNA, and from DNA to RNA to protein. It is called the central dogma because the ideas provide a framework for nearly all our understanding of genetics.

Since the Central Dogma was proposed and named by Francis Crick in 1957[3], there have been numerous examples of "exceptions to the rule": other ways information flows through a biological system. Some viruses, for example, including HIV and SARS-Cov2, have an enzyme called reverse transcriptase that creates a DNA molecule from an RNA template. Likewise, RNA molecules can, under certain circumstances, be created using an RNA template. And diseases called **prion** diseases result from an information transfer from protein to protein. But these exceptions mostly expand our understanding of molecular genetics; they do not counteract our understanding of the flow of information from DNA to protein.

The Central Dogma is illustrated in **Figure 2**, with the orange arrows indicating the flow of information through the Dogma's core processes of DNA replication, transcription, and translation. The green arrows indicate additional ways in which information can be transferred. In pink font are listed the molecular machines that perform each process in the cell.

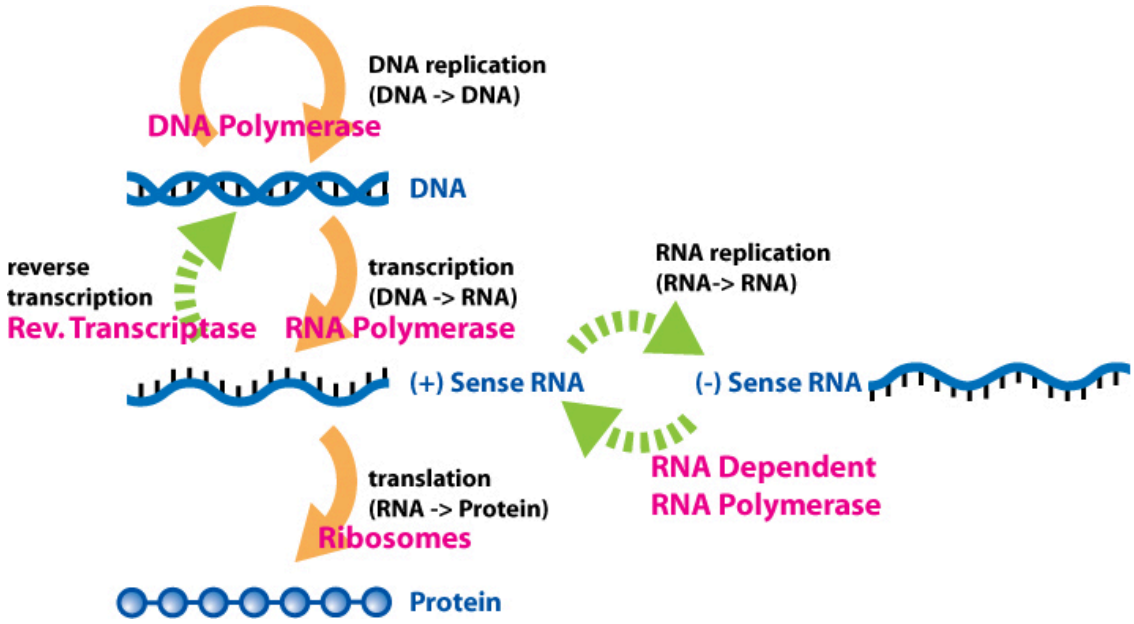


Figure 2. The Central Dogma of molecular genetics describes the flow of information from DNA to RNA to protein.

The remainder of this module focuses on the processes of replication and transcription.

3. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. PLoS Biol. 15, e2003243 (2017).

CHEMISTRY OF REPLICATION AND TRANSCRIPTION

DNA replication and transcription are similar in chemistry. Both processes use **nucleotide triphosphates,** or NTPs, as building blocks to assemble a new polymer (**Figure 3**). And both replication and transcription use an enzyme called a **polymerase** to add nucleotides to the 3' end of a growing polymer.

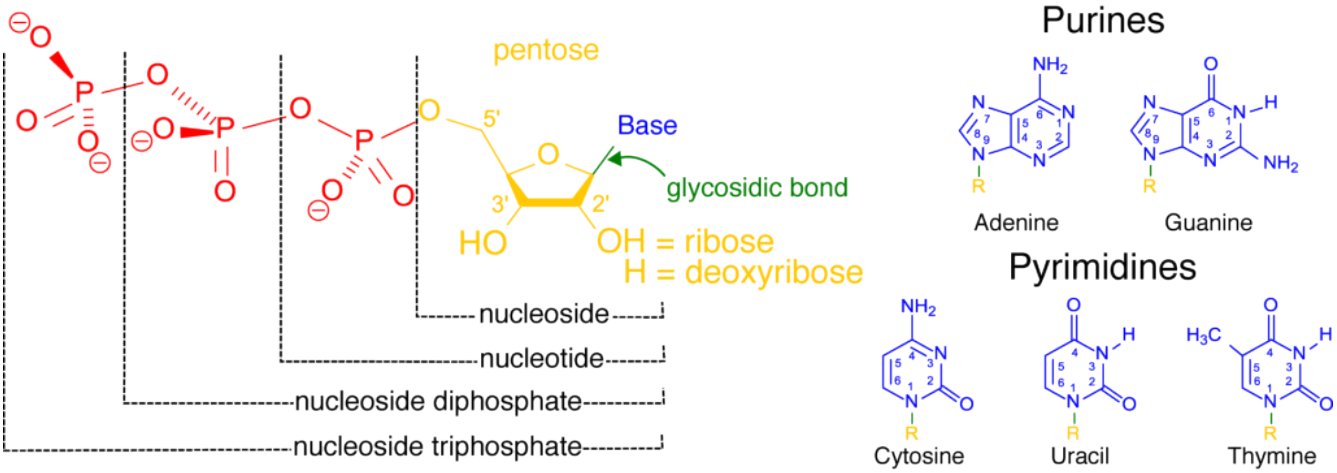


Figure 3. Structure of a nucleotide. A nucleotide may include one, two, or three phosphates (red) linked to the 5' carbon of a five-carbon sugar (yellow). DNA nucleotides use the sugar deoxyribose, with an OH group on the 2' carbon. RNA nucleotides use the sugar ribose, with a hydrogen (H) on the 2' sugar. A nitrogenous base (blue) is connected to the sugar via glycosidic bond. The structures of five bases are shown. DNA uses bases adenine, guanine, cytosine, and thymine, while RNA uses uracil instead of thymine.

A nucleotide may include one, two, or three phosphates (red in **Figure 3**) linked to the 5' carbon of a five-carbon sugar (yellow). The phosphate closest to the sugar is called the alpha phosphate, the middle one is the beta phosphate, and the one farthest from the sugar is the gamma phosphate.

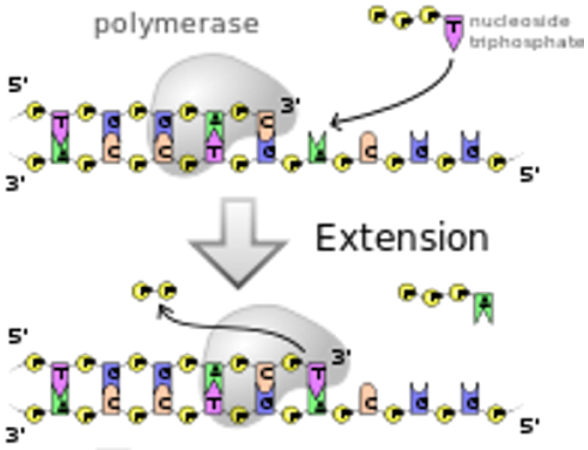


Figure 4. The chemistry of DNA and RNA synthesis. Both replication and transcription use a single-stranded DNA template. New strands are synthesized by from nucleotide triphosphates. The beta and gamma phosphates are lost as a phosphodiester bond is formed between the 3' end of the growing chain and the incoming nucleotide.

As a reminder, DNA nucleotides use the sugar **deoxyribose**, with an OH group on the 2' carbon. RNA nucleotides use the sugar **ribose**, with a hydrogen (H) on the 2' sugar. A nitrogenous base (blue) is connected to the sugar via glycosidic bond. The structures of five bases are shown. DNA uses bases adenine, guanine, cytosine, and thymine, while RNA uses uracil instead of thymine. The nucleotides are named for the base they include, so the four DNA nucleotides are adenosine, guanosine, cytidine, thymidine, and uridine. The RNA nucleotides are often abbreviated ATP, CTP, GTP, and UTP. The DNA nucleotides are often abbreviated dATP, dCTP, dGTP, and dTTP.

The general process of replication and transcription is shown in **Figure 4**. The bottom longer strand in the image is part of the original DNA molecule. The top strand is the newly synthesized strand, which forms complementary base pairs with the template. New nucleotides are added to the 3' end of the daughter strand in a reaction catalyzed by a **polymerase** enzyme. Because of this, we say that both replication and transcription proceed in a 5' to 3' manner.

Note that the template and daughter strands are antiparallel in orientation, with the 5' end of the daughter aligning with the 3' end of the template. Polymerases therefore move toward the 5' end of the template.

Adenosine triphosphate (used for RNA) is the same ATP molecule that powers cellular reactions. dATP is almost identical – it is just missing the 2' OH group. Just like the breakdown of ATP powers cellular reactions, the high-energy bonds connecting the phosphates of all the dNTPs and NTPs serve as an energy store to power DNA and RNA synthesis. When an incoming nucleotide is added to the growing strand, the high-energy bond connecting the alpha and beta phosphates is broken. The energy released powers the formation of a **phosphodiester** bond between the incoming nucleotide and the growing chain.

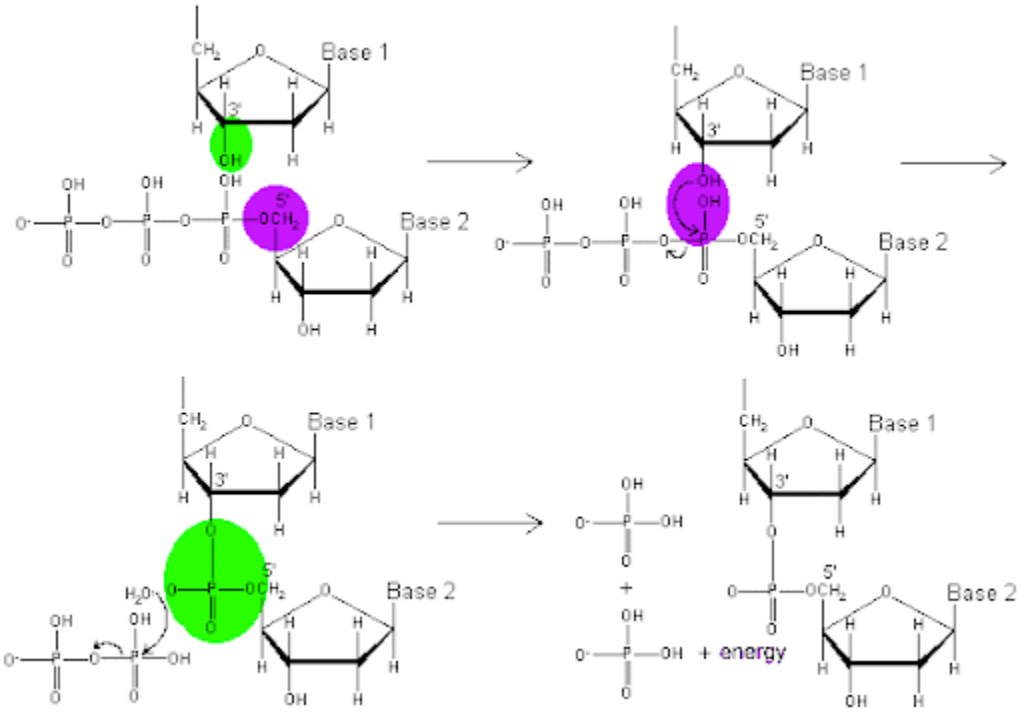


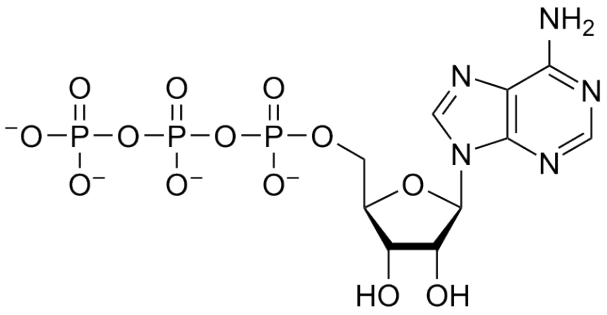
Figure 5. The polymerization of DNA nucleotides. The 3' hydroxyl group attacks the triphosphate group on the incoming nucleotide. A new phosphodiester bond is formed (large green circle), leaving behind two phosphates from the original nucleotide triphosphate. The final product is a daughter that is one nucleotide longer, with a new 3'OH end.

The mechanism of the reaction is shown in **Figure 5**. The green OH group represents the 3' end of a newly synthesized daughter strand. The pink highlights the 5' carbon of the incoming nucleotide triphosphate, from which extend three phosphate groups. The formation of the phosphodiester bond between the 3'OH and the phosphorus closest to the 5' carbon results in the release of two of the phosphates from the nucleotide. The release of the diphosphate, and the subsequent separation of the two released phosphate groups, releases enough energy to drive the formation of the phosphodiester bond. Note that, although DNA is shown in Figure 5, the only difference for RNA would be an OH group at the 2' position of each nucleotides.

Of course, there are significant differences between replication and transcription. We will start with replication, which requires that both strands of DNA are read simultaneously to create two new complementary strands, eventually resulting in a complete and nearly perfect copy of an entire organismal genome. Much of what we know about replication comes from studying prokaryotes, in particular the model organism *E. coli*. The next sections focus on replication in prokaryotes, with notes about the differences in eukaryotes.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=123#h5p-9



Test your understanding A.

Media Attributions

- Structure of a nucleotide © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- The chemistry of DNA and RNA synthesis © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- DNA replication © Ed Vitz, John W. Moore, Justin Shorb, Xavier Prat-Resina, Tim Wendorff, & Adam Hahn is licensed under a CC BY-SA (Attribution ShareAlike) license
- Chemical structure © Wikipedia is licensed under a CC0 (Creative Commons Zero) license

MECHANISM OF REPLICATION

The basic chemical reaction of DNA synthesis is the addition of a nucleotide residue to the 3' end of a polypeptide chain. The reaction is catalyzed by a DNA polymerase. But DNA polymerases have a biochemical limitations: they can only add nucleotides to the 3' end of a strand that already exists. They cannot add nucleotides to the 5' end of a molecule, and they cannot synthesize DNA *de novo*, or from the beginning by linking together two nucleotides. Therefore, to perform DNA synthesize with fidelity on both strands of the template DNA requires multiple polymerases and a collection of other proteins and enzymes. These include:

1. **DnaA:** Recognizes the origin of replication, where the process begins
2. **Helicase:** Unwinds the double helix, allowing the replication machinery to access the template DNA
3. **SSB:** Holds the unwound strands of the double helix apart long enough for replication to occur
4. **Primase:** Seeds the replication process, by synthesizing a short RNA molecule called a primer complementary to the template strand, since DNA polymerase can't synthesize DNA de novo.
5. **DNA Polymerase III:** The main replicative polymerase
6. **DNA Polymerase I:** Removes the RNA primers and fills in DNA across the gap in the daughter strands
7. **Clamp**: Keeps the polymerase from dissociating from the template DNA, allowing replication to continue for a long period of time without reloading the polymerase onto the DNA
8. **Ligase:** Seals nicks in the phosphodiester bond that remain after the removal of RNA primers

Together, these proteins and enzymes make up the **replisome**, or replication machinery. Let's look at each of them in turn. This next section is organized with a series of static figures that highlight what is happening to the DNA at each step of the process. The section concludes with a link to an animated GIF that shows the action of each of the proteins of the replisome.

Replication begins at specific sites in the DNA called origins of replication. Prokaryotes, which have a single circular chromosome, have a single chromosomal origin of replication. The origin includes a specific DNA sequence called oriC. In *E. coli,* this origin of replication is approximately 245 base pairs long and is rich in AT sequences. The oriC sequence is recognized by an initiator protein called **DnaA**, which binds to the origin in a sequence-specific manner. DnaA uses the energy of ATP hydrolysis to introduce torsional pressure that partially unwinds the double helix at the origin.

From the origin of replication, **helicase** enzymes extend the unwinding, again using the energy from the hydrolysis of ATP to break the hydrogen bonds connecting the strands of the double helix. Helicases move bidirectionally away from the origin, separating the two strands of the double helix into what is called a

replication bubble. The Y-shaped edges of the bubble, where the DNA transitions from single-strand to double-strand, are called replication forks. Each replication bubble includes two replication forks (**Figure 6**). As replication proceeds, the two replication forks move continuously away from each other as the helicases continue to unwind more of the chromosome.

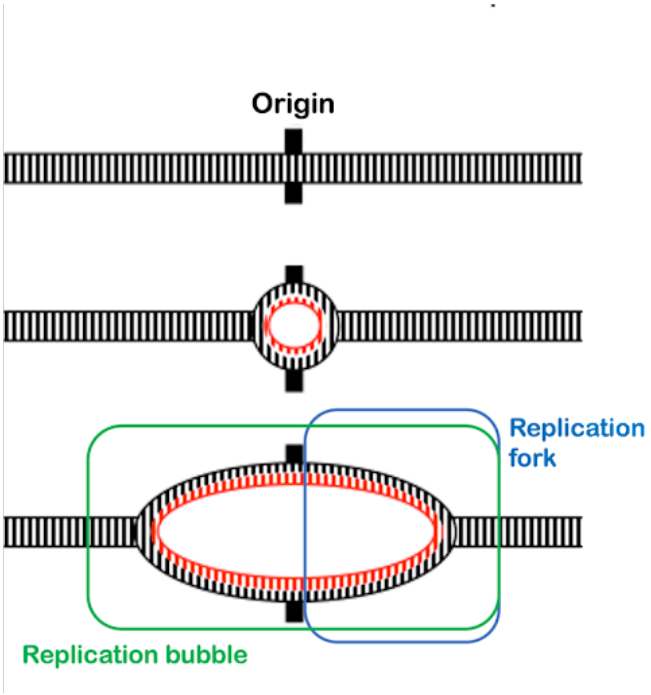


Figure 6. The firing of an origin of replication creates a replication bubble. Each replication bubble is made up of two replication forks.

Re-pairing of the bases is energetically favorable and would happen spontaneously in an aqueous environment like a cell. But the replication machinery needs access to single-stranded DNA template. **Single-stranded binding protein** (SSB) coats the single-stranded DNA, holding the strands apart and preventing them from re-pairing.

Because DNA polymerases cannot synthesize DNA *de novo*, an additional enzyme is necessary to begin synthesis. **Primase** synthesizes short RNAs called **primers** complementary to the DNA near the origin. These short RNA molecules provide a free 3' end for the DNA polymerase.

There are multiple DNA polymerases, but **DNA polymerase III (Pol III)** is the main replicative polymerase in prokaryotes. It performs most of the DNA synthesis at the replication fork. DNA pol III extends the primers outward from the origin toward the replication forks by adding nucleotides to the 3' end, as shown in **Figure 7**.

The continuous synthesis from the 3' end of the original primers form what are called the **leading strands**. Leading strand synthesis will continue for millions of bases! But Pol III doesn't hold on to DNA very well: by itself, it would fall off within a handful of nucleotides. So a ring-shaped protein called the **sliding clamp** holds the polymerase into place on the template strand, making DNA synthesis highly efficient. The sliding clamp works almost like a carabiner, opening and closing to encircle the DNA. An enzyme called the **clamp loader** opens and closes the ring to engage clamp on the DNA. (Clamp loader is not shown in the images in this text.)

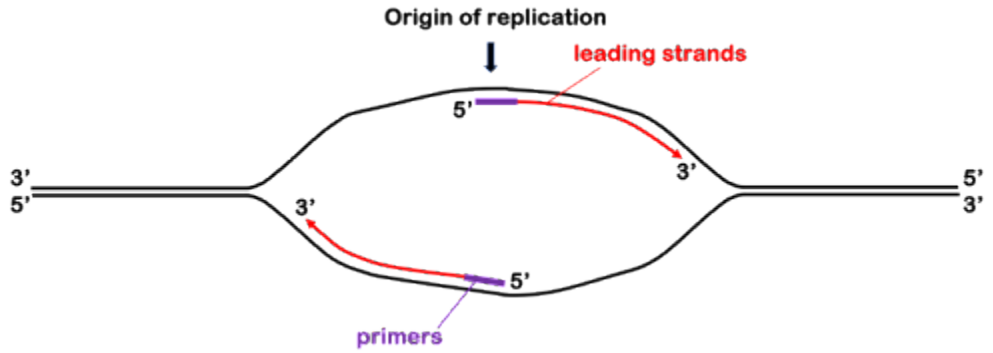


Figure 7. The leading strands are formed by the extension of the original primers at the origin of replication. Two leading strands are begun at each origin, extending in opposite directions.

Prokaryotes can replicate about one thousand nucleotides of template DNA per second! Leading strand polymerization will continue until the polymerase runs out of unreplicated template, either because it bumps against another replication fork proceeding in the opposing direction or because it comes to the end of a linear chromosome (in eukaryotes).

In **Figure 7**, you can see that the synthesis of the leading strands leaves half of the replication bubble unreplicated: DNA polymerase cannot extend the original primers on the 5' end. Thus, an additional primer is placed on each strand nearer the replication fork. The polymerase synthesizes the **lagging strand** from this second primer, extending the 3' end back toward the original RNA primer. This is shown in **Figure 8**. The 3' ends of the daughter strands are indicated by arrowheads. You can see that the leading strand "points" toward the replication fork, while the lagging strand points away from the fork.

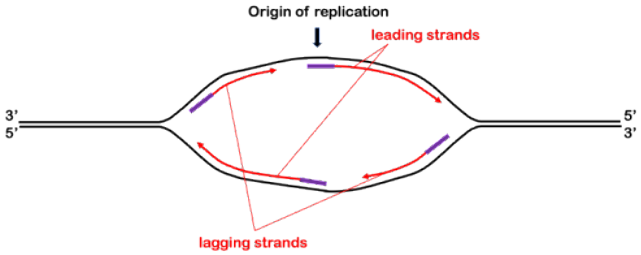


Figure 8. The lagging strands are begun with additional primers synthesized close to the replication fork, past the 5' end of the original primer at the origin. The lagging strand is synthesized by adding nucleotides to the 3' end of the lagging strand primer, extending the lagging strand away from the fork.

As the helicases continue to open the replication bubble, additional primers are created, and additional fragments of the lagging strand are synthesized. These lagging strand fragments are called **Okazaki fragments** after the married researchers who first hypothesized and confirmed their existence. **Figure 9** shows a replication bubble with leading and lagging strands. All 5' and 3' ends are labeled, and the 3' ends of the daughter DNA strands are indicated with arrowheads.

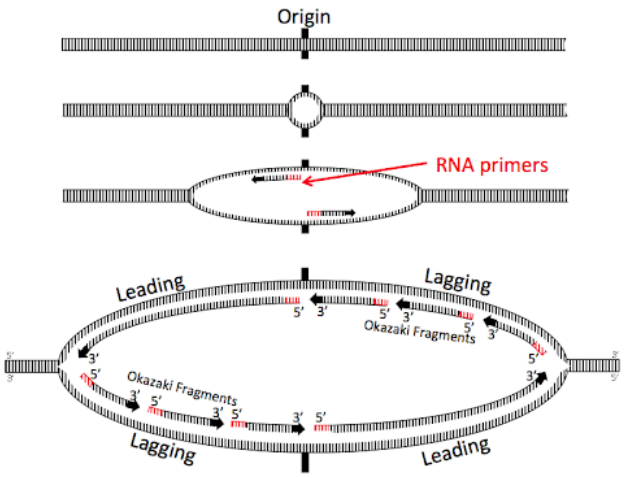


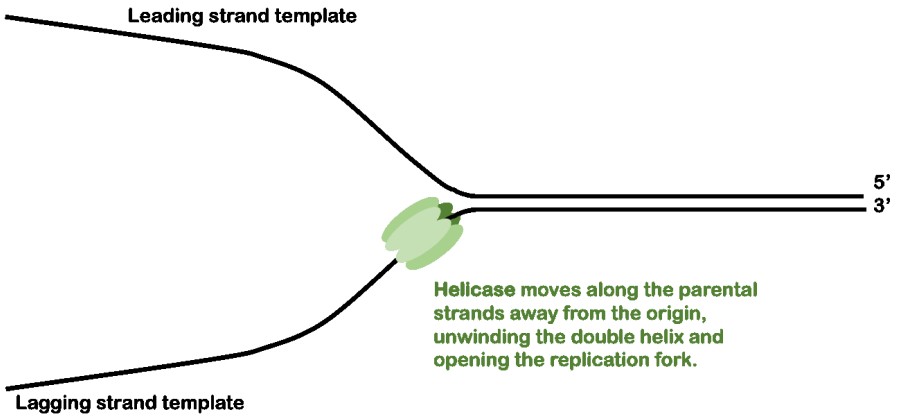
Figure 9. An origin of replication. The DNA duplex is melted then the primase synthesizes RNA primers on each strand. DNA replication begins with the extension of the primers and proceeds bidirectionally as the two replication forks head off in opposite directions. The leading and lagging strands are shown along with Okazaki fragments. Note the 5' and 3' orientation of all strands.

You'll see the original RNA primers at the origin near the center of the replication bubble. These original primers form the 5' end of two leading strands, synthesized continuously away from the origin.

The top right and bottom left of the replication bubbles are the lagging strands, with three Okazaki fragments shown for each lagging strand. The oldest Okazaki fragments are closest to the origin; the newest are closest to the forks.

Note the diagonal symmetry of the bubble: one leading strand is on the top left of the bubble, and the other is on the bottom right. The 3' ends of all of the daughter strands on the top half of the bubble are oriented to the left, antiparallel to the 3' end of the parent strand. The opposite is true for the bottom of the bubble. This geometry happens because of the antiparallel orientation of the strands of the double helix.

The lagging strand polymerase extends the 3' end of each new RNA primer toward the previous Okazaki fragment, but the RNA is not retained in the genome. **DNA polymerase I** acts to remove the RNA and patch the space with DNA nucleotides. **DNA polymerase I** has a 5' to 3' exonuclease function that removes the RNA primer while it at the same time extends the 3' end of the newest Okazaki fragment into the newly created gap. Fragments are linked together through the action of **DNA ligase.**



Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=129#h5p-10

Media Attributions

- Replication bubble © Online Open Genetics is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- The leading strands © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- The lagging strands © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Origin of replication © Online Open Genetics is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Stepped replication fork images © Amanda Simons is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license

THREE DIMENSIONAL DNA STRUCTURE DURING REPLICATION

Prokaryotic chromosomes are circular and typically have one origin of replication. The whole replication process continues in both directions away from the origin until the whole chromosome has been replicated (Figure 10A). The intermediate structures formed by the replication of a circular chromosome are sometimes called theta structures, since they look like the Greek letter theta θ. Another view of the process can be seen in **Figure 10B,** with an animated version online at Wikimedia Commons.

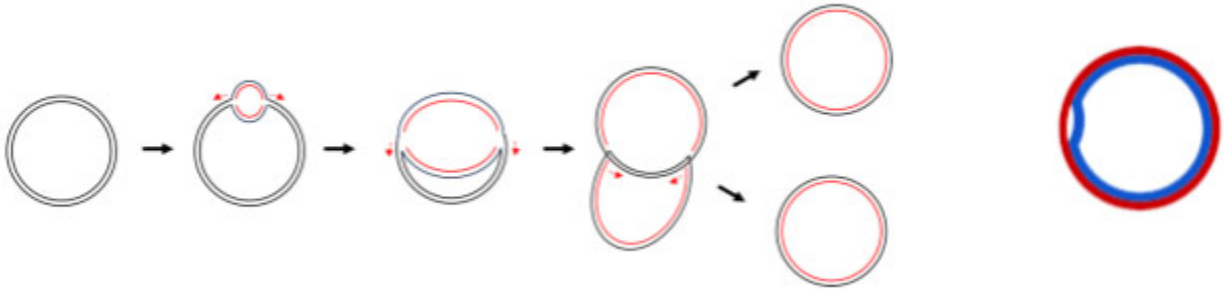


Figure 10. Bacterial chromosomes are replicated from a single origin of replication. **A (left)**. Stepped image showing the replication of a circular chromosome, as the replication bubble expands from the origin. The parent strands are shown in black, and the daughter strands are shown in red. Image source: AS CC 4.0 BY SA. **B (right)**. Animated gif showing the synthesis of daughter strands and the separation of the parent strands can be found at Wikimedia Commons. The two original strands separate from each other and serve as templates for the synthesis of new strands. Replication is terminated when the forks meet and the two chromosomes separate. Each new identical DNA molecule contains one template strand from the original molecule, shown as a solid line, and one new strand, shown as a dotted line.

In this description and in the figures, we've presented this as if the leading strand is synthesized first and the lagging strand second. Even the names give this impression! But the synthesis of leading and lagging strands happens simultaneously. This requires the proteins involved in this process to work in concert. While the leading strand polymerase acts continuously on the leading strand template, the lagging strand polymerase dissociates after each Okazaki fragment, rebinding to each new primer. Throughout this process, the two polymerases stay linked so that as the replication fork moves away from the origin, both strands are replicated at once. To accomplish this, all of the replication participants must be organized very specifically in three dimensional space.

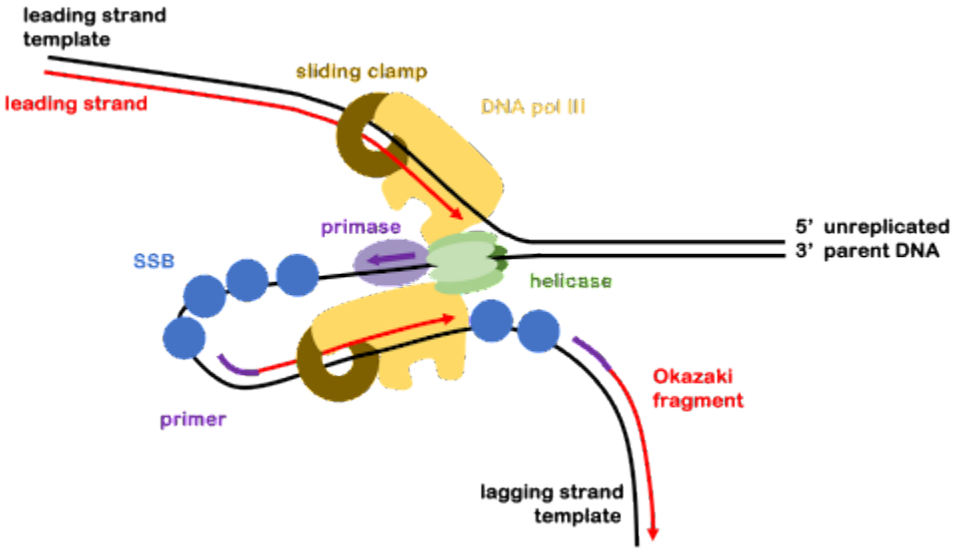


Figure 11. Trombone model of replication. In this image, the replication fork is moving toward the right. The lagging and leading strand polymerases move together, linked as part of the replisome. The lagging strand template is looped out and around, allowing the lagging polymerase to move in the same direction as the leading polymerase. The loop gets bigger as more parent DNA is unwound and the lagging polymerase extends an Okazaki fragment. The loop is released when one Okazaki fragment is completed, and a new loop forms when synthesis of a new fragment is begun.

As the replication fork opens, the lagging strand template becomes looped around, as shown in **Figure 11**, where the lagging strand has been folded under itself to bring the two polymerases closer together. This is called the trombone model of replication because as the fork progresses away from the origin, this loop appears to grow and shrink as the DNA template moves in relation to the polymerase. The entire process is shown in this animation of the replication process, produced by HHMI Biointeractive.

Note: Although they are not shown in the figures of this text, additional subunits of the replisome also participate in the replication process. Additional components of the replication machinery help perform functions like loading and unloading the clamp for new Okazaki fragments and keeping the leading and lagging polymerases together so that both template strands are replicated in concert.

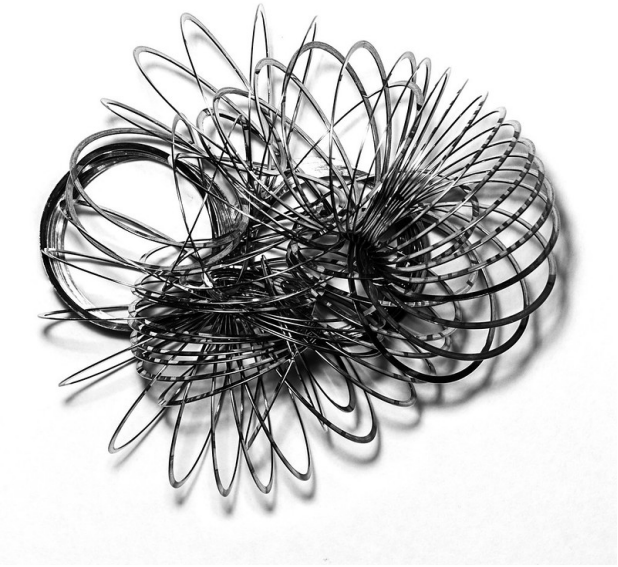


Figure 12. A tangled spring toy.

The three-dimensional structure of the replisome and the length of the DNA strands also cause certain difficulties. Have you ever tried to untangle a spring toy (**Figure 12**), comb through long tangled hair, or wrestle with the power cord on a hand-held appliance? Just like those examples, the parent and daughter DNA strands become twisted around one another in a way that makes it difficult to both melt the template DNA and separate the two daughter duplexes after replication is complete.

A class of enzymes called **topoisomerases** relieve the torsional strain caused by melting the double helix and untangle the daughter DNA.

Media Attributions

PROKARYOTIC VS EUKARYOTIC REPLICATION

Replication in eukaryotes uses the same chemistry of 5' to 3' synthesis as in prokaryotes, but there are notable differences that stem from the differences in prokaryotic and eukaryotic genome structure.

First, bacterial chromosomes are circular and relatively small compared to linear eukaryotic chromosomes. For example, the *E coli* genome is about 4,600,000 base pairs, or 4.6 megabases. Most bacterial chromosomes have a single origin of replication, and replication proceeds bidirectionally around the chromosome until the two replication forks meet. At 1000 basepairs per second, the whole bacterial chromosome can be replicated in a little more than half an hour.

In comparison, eukaryotic replication forks are much slower, moving at about 50 bases per second. The slower speed is at least in part because of the structure of the chromatin packaging. And eukaryotic chromosomes are much larger. Human chromosomes, for example, range from 48 megabases (Chromosome 21) to 249 megabases (Chromosome 1). If each eukaryotic chromosome were replicated only from one origin, at 50 base pairs per second, it would take about 5.5 days to replicate the smallest chromosome and about a month to replicate the longest one!

But eukaryotic cells don't take that long to replicate a chromosome. Instead, eukaryotic chromosomes typically activate multiple origins along the length of each chromosomes, as shown in **Figure 13**.

Eukaryotic cells also use different enzymes, although they perform functions like those in prokaryotes. For example, while DNA pol III is the primary DNA polymerase in prokaryotes, in eukaryotes Pol epsilon and Pol delta perform the majority of leading strand and lagging synthesis, respectively. The sliding clamp function is performed by the eukaryotic protein PCNA. And rather than Pol I to replace the RNA primer with DNA, in eukaryotic systems RNAseH breaks down the RNA primers as Pol epsilon and delta continue their synthesis across the gap.

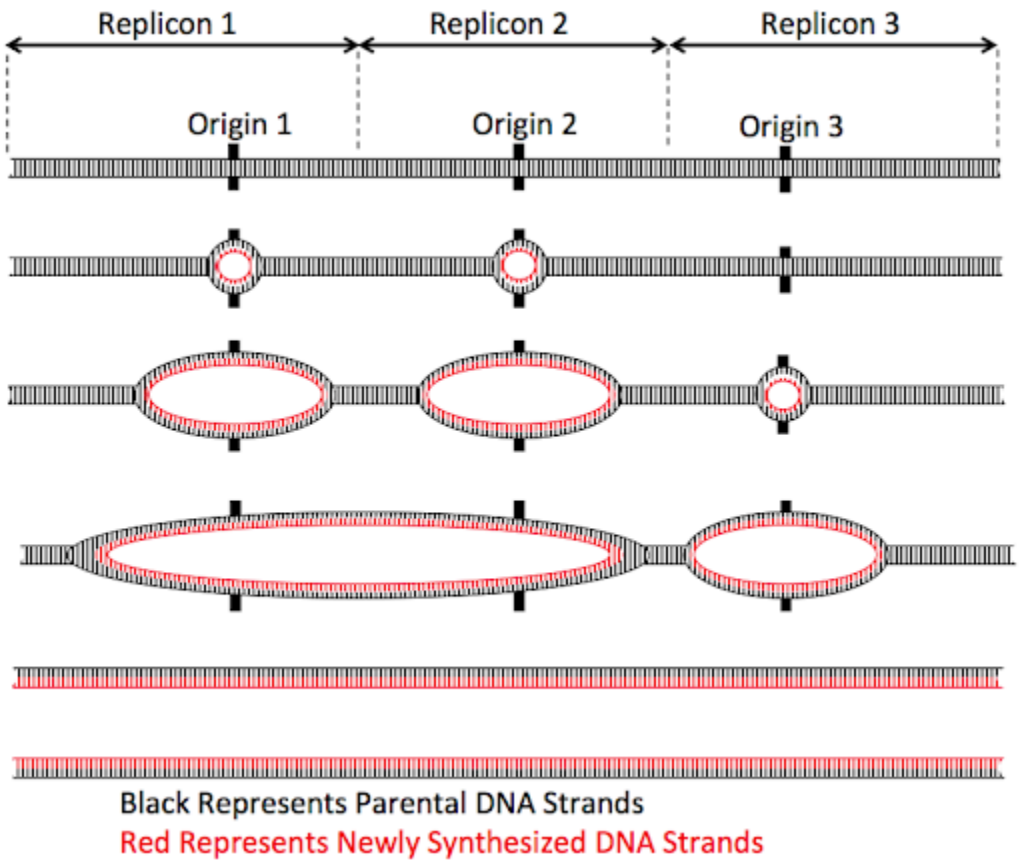


Figure 13. Eukaryotic chromosomes have multiple origins. From Online Open Genetics, Chapter 2. Part of a eukaryote chromosome showing multiple Origins (1, 2, 3) of Replication, each defining a replicon (1, 2, 3). Replication may start at different times in S-phase. Here #1 and #2 begin first then #3. As the replication forks proceed bi-directionally, they create what are referred to as "replication bubbles" that meet and form larger bubbles. The end result is two semi-conservatively replicated duplex DNA strands.

But perhaps the biggest difference between eukaryotic and prokaryotic replication comes from the fact that eukaryotic chromosomes are linear rather than circular. Because DNA replication requires a primer to begin, and because synthesis can only proceed in a 5' to 3' direction, eukaryotes have an "end replication problem": they cannot replicate the 5' most end of a linear chromosome. The 5'-most end of a newly synthesized daughter strand must always begin with an RNA primer. But if the 5' RNA primer were degraded, there is no upstream 3' end to extend. This leaves a gap at the 5' end of the daughter strand, as shown in the top panel of **Figure 14**.

For most eukaryotic cells, this means that in every round of cell division, the ends of chromosomes, called **telomeres**, will get a little bit shorter. Older individuals typically have shorter telomeres than younger individuals of the same species. This shortening of telomeres is thought to play a role in cellular aging.

Telomeres are comprised of a repeating DNA sequence. In humans, the sequence is TTAGGG, but other species have slightly different sequences. The telomeric sequence can be repeated 100-1000 times at the ends

of chromosomes. The telomeric sequences thus serve as a "cap" of sorts to the linear chromosomes: There are no genes in telomeric regions of the chromosome, so the DNA lost from the ends is not needed for gene expression.

However, some specialized cells – germ cells and stem cells – have an enzyme called telomerase that counteracts this problem. It is tempting to hypothesize that telomerase just works to fill in the 5' gap. However, what telomerase actually does is make the template strand – already longer than the daughter strand – even longer!

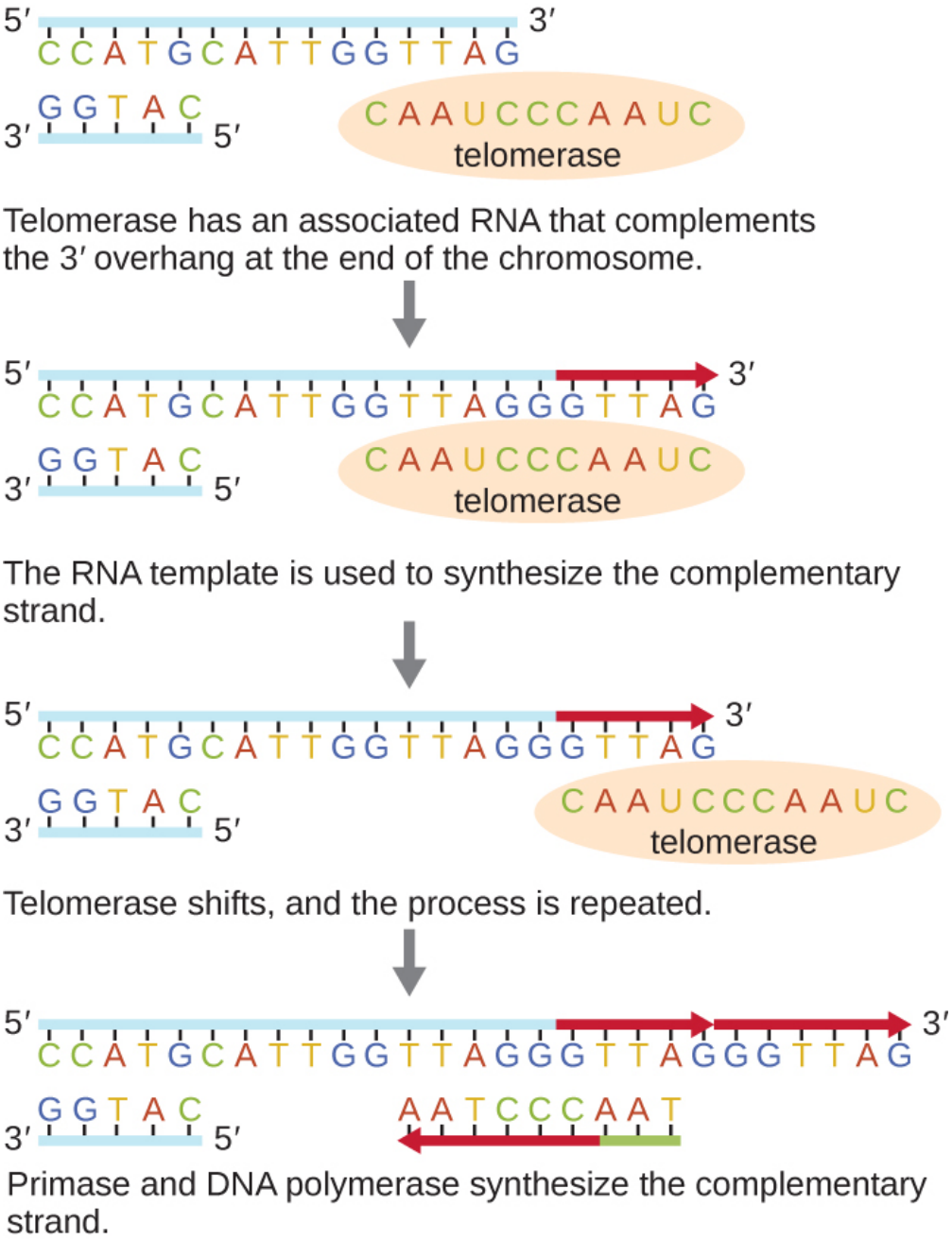


Figure 14. Telomerase carries an RNA template that is used to extend the 3' end of the template strand beyond the original length of the parent DNA.

Telomerase does this by carrying its own template for DNA replication: an RNA molecule that is a component of the telomerase enzyme. The RNA molecule is complementary to the telomeric sequence. Telomerase binds to the 3' end of the parent strand, with the RNA component base-pairing with the telomeric sequence in such a way that the telomerase RNA overlaps past the end (**Figure 14**). It supplies its own template for replication! The protein component of telomerase then adds additional nucleotides to the 3' end of the parent strand, making it even longer. Telomerase then shifts toward the new 3' end, and the process is repeated.

The additional length on the parent strand permits additional lagging strand fragments to be synthesized, thus lengthening the daughter strand as well. Note, however, that the daughter strand will still be shorter than the parent strand since that additional fragment also required a primer to initiate synthesis.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=138#h5p-11

Why do we care about telomerase? Telomerase, aging, and cancer

The 2009 Nobel Prize in Physiology or Medicine was awarded jointly to Elizabeth Blackburn, Carol Greider , and Jack Szostak for their work in understanding the mechanism of telomerase[1]. This reflects the profound implications of telomeres and telomerase for human medicine.

Telomerase may seem like an interesting trivia tidbit: after all, it's only active in a very, very limited number of cells in an adult body. But telomerase is very important in "resetting" the genome to a youthful, long-telomeric state during gametogenesis – the production of egg and sperm for reproduction. Without the action of telomerase, each generation of offspring would have shorter and shorter telomeres. Telomerase is also active in stem cells. Stem cells have the potential to develop into other cell types, and they are self-renewing. Stem cells play an important role in replenishing tissues like blood and immune cells, as well as repairing tissues damaged by injury.

Mutations in the genes that encode telomerase (both the enzymatic component and the structural RNA

1. The Nobel Prize in Physiology or Medicine 2009. NobelPrize.org https://www.nobelprize.org/prizes/medicine/2009/illustrated-information/.

component) are associated with phenotypes of premature aging. In humans, people who inherit rare mutations in these genes may show premature gray hair and other signs of aging, fibrosis of the lung and liver, immune dysfunction, and bone marrow failure. Other studies show that people with shorter telomeres than their same-aged peers are at higher risk of dying from certain causes, including cardiovascular, respiratory, and digestive diseases[2].

Because of these observations, medical researchers who study aging have hypothesized that telomerase-*activating* drugs might be useful in combating aging and age-related diseases. But it's not so straightforward. One more type of cell often expresses high levels of telomerase: cancer cells.

Cancer cells cause disease because they divide uncontrollably, outgrowing their surrounding cells and destroying the function of healthy tissue. They also have the potential to **metastasize**, migrating throughout the body and colonizing secondary tumors in other tissues. Because telomeres get shorter with every round of cell division, most healthy cells can undergo a finite number of cell divisions before they **senesce,** or stop dividing. But many cancers express telomerase, allowing them to bypass this limit, continue to divide, and out-compete healthy cells in the body. This means that an anti-aging drug that activates telomerase also has the potential to feed cancer growth – not quite the healthful result that one might desire! On the flip side, telomerase *inhibitors* may prove to be a useful treatment for cancer. As of the writing of this textbook, at least one telomerase inhibitor is in clinical trials for certain types of cancer.

Media Attributions

2. Schneider, C. V. et al. Association of Telomere Length With Risk of Disease and Mortality. JAMA Intern. Med. 182, 291–300 (2022).

WRAP-UP QUESTIONS

1. One of the best ways to test how much you recall and understand from a reading is to close your notes and try to draw out a concept from memory.

a. Draw a replication bubble and/or replication fork, with leading and lagging strands and all 5′ and 3′ ends labeled.
b. Draw the ends of a newly-replicated daughter strand of DNA, before and after telomerase has acted. Be sure to label the 5′ and 3′ ends.

2. Who do you expect to have longer telomeres: an infant or their parent? Explain your reasoning.

3. Chemical substances found in certain foods have been found to activate telomerase in laboratory experiments. Do you think those substances could be more useful in reversing aging or in treating cancer? Explain your reasoning.

PART V
TRANSCRIPTION

Objectives

After completing this module, students should be able to:

1. Recognize that not all RNA molecules are used to translate protein.
2. Compare and contrast transcription and replication.
3. Describe the process of transcription, and describe the DNA elements and protein factors required for transcription.
4. Identify template vs nontemplate strand of DNA, given an RNA molecule.
5. Predict RNA sequence when given a DNA sequence.
6. Define: promoter, terminator, sigma factor, -10, -35, upstream, downstream, hairpin, factor, element, consensus, RNA polymerase, RNA polymerase II, holoenzyme, transcription factor, mediator

Introduction: The concept of a gene

If you ask 10 biologists to define the word "gene", you will likely get slightly different answers from all of them. Some definitions might focus on the concept of a gene as the simplest unit of heredity, while others might focus on the physical properties of a gene as a segment of DNA.

The concept of a gene has changed throughout the history of genetics, beginning with scientists in the 1800's. Early geneticists like Gregor Mendel, Thomas Hunt Morgan, and Barbara McClintock understood that there was a hereditary element that specified traits, although many of them did not use the word "gene" to describe this concept, and they did not know which biological molecule(s) might be the hereditary elements. Although Mendel did not use the word gene, his work on pea plants did give us the concept of a hereditary unit that specified a trait, as well as the concept of dominant and recessive alleles of a gene.

Mendel's work did not receive much attention for until almost forty years after it was first published. But by the early 1900's, additional researchers like Thomas Hunt Morgan and Barbara McClintock were exploring

the concept of a gene. Thomas Hunt Morgan's work in the fruit fly *Drosophila melanogaster* showed that the white-eyed trait was associated with sex, and he suggested that the trait might therefore be carried on the sex chromosomes that could be visualized microscopically. He further suggested that other traits might therefore be carried on other chromosomes. He later showed that certain genes were linked together in a linear fashion, with measurable distances separating linked genes.

Barbara McClintock's work in maize (corn) further showed that distinctive traits were carried on physical chromosomes: she observed unique structural differences in the chromosomes of certain strains of maize, and she showed that the presence of these structurally unique chromosomes was associated with particular phenotypes.

In the early 1900's, Archibald Garrod showed that hereditary traits were linked with biochemical activity. He was studying a hereditary illness called alkaptonuria, which causes dark urine starting as a newborn and a host of health problems later in life. He found that the disease is caused by a loss of function in the ability to metabolize the chemical alkapton. (We know now that alkapton is a breakdown product of the amino acids tyrosine and phenylalanine.) He later expanded his studies to include other **inborn errors of metabolism**, hereditary diseases that were caused by defects in enzymes important for metabolism. These studies showed a link between genes and enzymes.

By the 1940's, George Beadle and Edward Tatum, studying amino acid biosynthesis in the slime mold *Neurospora crassa*, proposed the **"One gene, one enzyme"** hypothesis, which stated that each gene provided the instructions for one enzyme, and the function of the enzyme was what contributed to the phenotype of an organism. This was later revised to **"one gene, one polypeptide"** and when it was understood that not all proteins were enzymes and that many functional proteins were assembled from multiple polypeptide chains.

This understanding is reflected in the Central Dogma of molecular genetics, which states that information stored in DNA is used to transcribe (synthesize) RNA, that information in that RNA is used to synthesize protein, and that proteins (among them, enzymes) complete many of the biochemical and molecular functions of the cell.

Today, it is understood that even "one gene, one polypeptide" is an incomplete picture of what a gene can be. As we will see the section on RNA processing, one gene can often produce multiple distinct polypeptides. And many genes encode functional RNAs: RNA molecules that perform molecular functions themselves and are not used as a template for protein synthesis. A partial list of RNAs and their functions is listed in **Table 1**.

Table 1 Selected types of RNA

Type of RNA	Function
mRNA (messenger RNA)	Encodes proteins
tRNA (transfer RNA)	Adaptor between mRNA and amino acids during translation
rRNA (ribosomal RNA)	Enzymatic component of the ribosome
TERC (Telomerase RNA component)	Extension of telomeres during replication
snRNA (small nuclear RNA)	Component of splicing machinery
miRNAs (micro RNA)	RNA interference and post-transcriptional gene regulation
snoRNA (small nucleolar RNA)	Modification or rRNA and tRNA and regulation of alternative splicing

In this module, we will define genes as a segment of DNA that encodes an RNA molecule, although this too is considered an over simplification by many scientists[1].

In this module, we will:

1. Compare the chemical nature of DNA and RNA.
2. Explore the DNA elements that surround a gene and provide information for the control of transcription in eukaryotes.
3. Compare and contrast this with the process of transcription in eukaryotes
4. Look at the ways eukaryotic mRNA are processed during and after transcription: polyadenylation, capping, and splicing.

In later modules, we will look at how some of these RNA molecules are eventually translated into protein and how gene expression is regulated.

1. Lee, H., Zhang, Z. & Krause, H. M. Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners? Trends Genet. 35, 892–902 (2019).

OVERVIEW OF TRANSCRIPTION CHEMISTRY

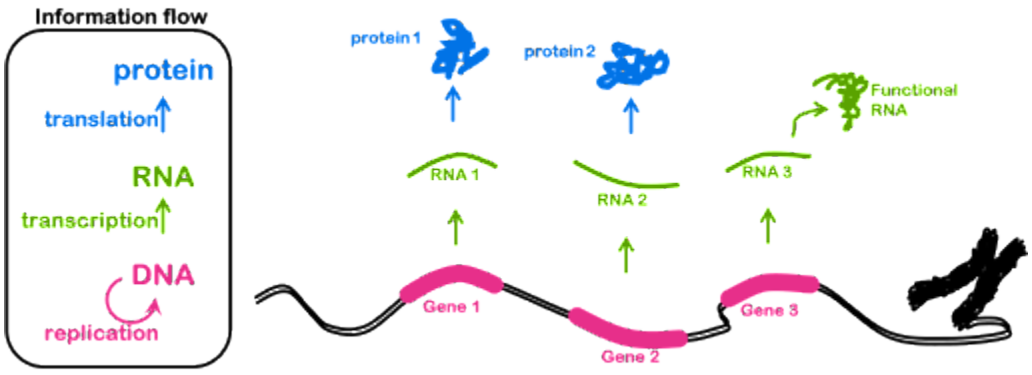


Figure 1. The Central Dogma of molecular genetics states that genetic information stored in DNA can be used to make more DNA through the process of replication and can be used to make RNA through transcription. RNA can be used as is by the cell, or it can be used to make protein through the process of translation. Here, genes (pink) are interspersed with non-gene sequence (black). Genes are the regions of the genome that are transcribed into RNA (green). RNAs may be translated into protein (blue), or they may serve other purposes in the cell (functional RNA).

Transcription is the process of RNA synthesis, and RNA molecules are also called **transcripts**.

Within a chromosome, gene sequences are interspersed with intergenic sequence (in-between sequences that are not transcribed into RNA), as shown in **Figure 1**.

The relative abundance of gene to intergenic sequence varies among organisms. In bacteria, most of the genome encodes protein. But in eukaryotes this is much lower: for humans, only about 2% of the genome encodes protein. Some of the remaining 98% encodes functional RNAs, but the function of the remaining sequence is not well understood. However, some of the intergenic regions include sequences that are important for the regulation of **gene expression**. When a gene is transcribed, it is said to be expressed. Not all genes are expressed in every cell type.

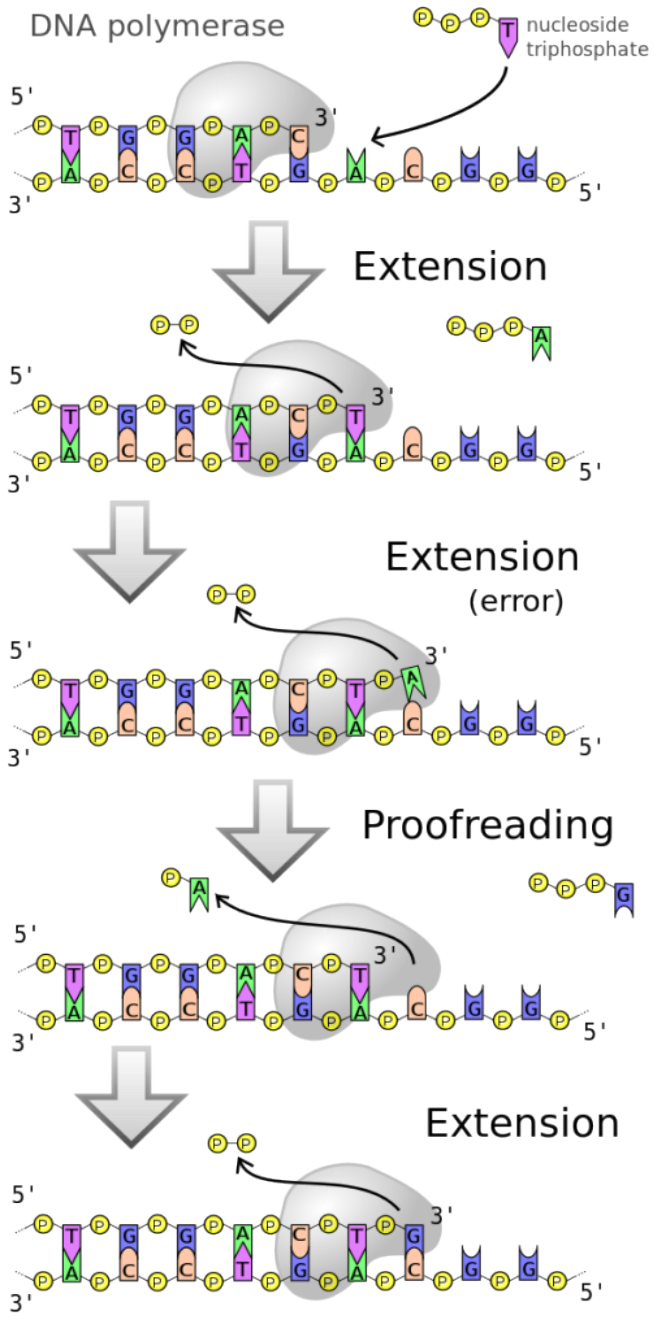


Figure 2. The chemistry of DNA and RNA synthesis. Both replication and transcription use a single-stranded DNA template. New strands are synthesized by from nucleotide triphosphates. The beta and gamma phosphates are lost as a phosphodiester bond is formed between the 3' end of the growing chain and the incoming nucleotide.

The biochemistry of transcription is similar to replication: one strand of DNA is used as a template to synthesize a daughter strand with a complementary sequence. In both replication and transcription, nucleotide triphosphates are used as the building block for the new polymeric daughter strand. In both replication and transcription, synthesis proceeds from the 5' to 3' end, with nucleotides added on to the 3' end of a growing polymer. Once nucleotides are incorporated into the polymer, they are called nucleotide residues. The chemistry of synthesis is discussed in more detail in the Replication module and is illustrated in **Figure 2**.

However, there are notable differences between replication and transcription: there are differences in chemical composition of DNA and RNA, replication results in a double-helical DNA molecule while transcription produces a single-stranded RNA, replication must faithfully replicate the entire genome in every dividing cell but transcription copies only a small part of the genome, and transcription is regulated so that only a subset of genes are transcribed at a given time in any cell.

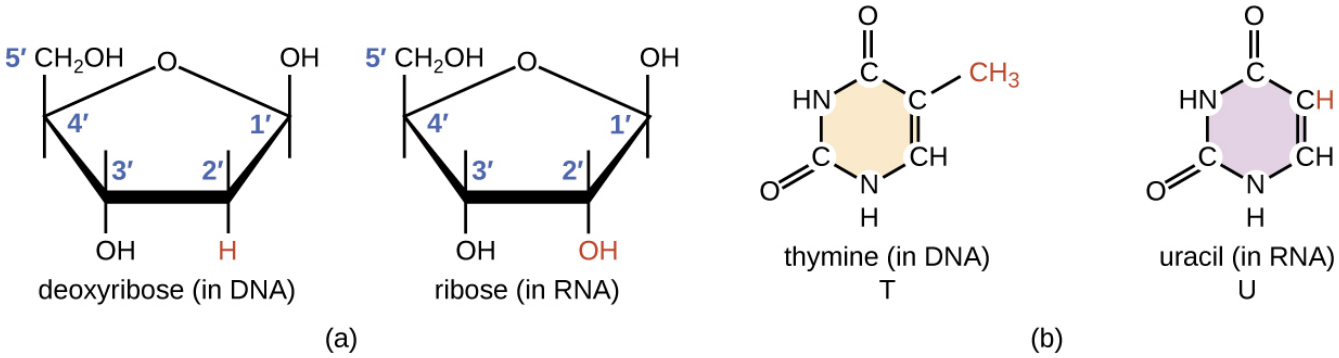


Figure 3. Chemical differences between DNA and RNA. A (left): DNA contains the sugar deoxyribose, while RNA contains the sugar ribose. Ribose has one additional OH on the 2' carbon. B (right) DNA contains the base thymine, while RNA contains the base uracil. Thymine has one additional methyl (-CH_3) group.

Chemical differences between DNA and RNA

First, the chemical structure of DNA and RNA nucleotides differs: DNA (deoxyribonucleic acid) nucleotides use deoxyribose as the sugar. RNA (ribonucleic acid) nucleotides use ribose. DNA nucleotides include the bases adenine, thymine, guanine, and cytosine, but RNA nucleotides include the base uracil instead of thymine. Uracil and thymine are structurally similar and differ only by one additional methyl group in thymine (**Figure 3**). This methyl group does not affect base pairing: Both thymine and uracil base pair with adenine.

Three dimensional structure of RNA

Second, DNA usually exists as a stable double-helical structure with two complementary strands. But most functional RNA is single-stranded. The bases can – and do! – form base pairs with one another. Because of the single-stranded nature of RNA, intra-strand base pairs often form, and the three-dimensional structure of folded RNA molecules can be quite variable. An example is shown in **Figure 4**.

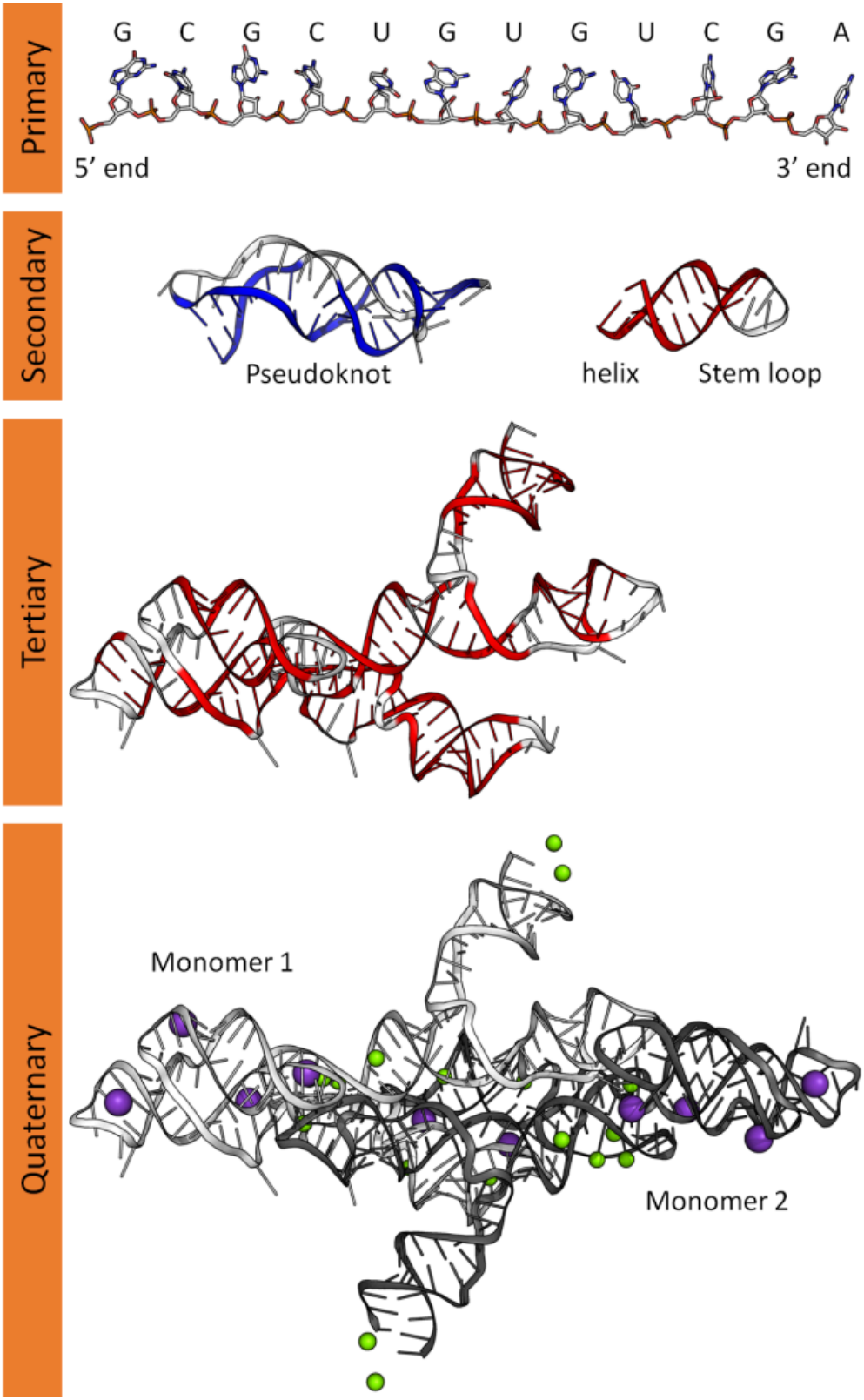


Figure 4. Illustration of levels of three-dimensional structure of RNA molecules: primary, secondary, tertiary, and quaternary. The tertiary and quaternary structures shown here are of the VS ribozyme, and RNA enzyme responsible for a particular mode of replication in the mitochondria of Neurospora crassa.

The structure of RNA molecules, like that of proteins, can be described in terms of primary, secondary, tertiary, and quaternary structures. The primary structure (1^{o}) is the sequence of the nucleotide residues. The secondary structure (2^{o}) is a collection of recognizable structural elements, including double-helices and stem loops that can form when one RNA molecule folds and basepairs internally. The tertiary structure (3^{o}) is the three-dimensional structure of the whole, folded RNA molecule, which can include elements of secondary structure. Some higher-order complexes can form from multiple RNA molecules. In those cases, quaternary structure (4^{o}) refers to the resulting multimeric structure.

Only parts of the genome are transcribed

Finally, although both strands of the entire genome must be replicated in each cell cycle, only part of the genome is transcribed into RNA: genes are interspersed between long stretches of intergenic DNA. In addition, usually only one strand of the double-helix will be used as an RNA template (Figure 5). Within and around a gene, the strands of the DNA double helix are therefore referred to as the **template** and **nontemplate** strands.

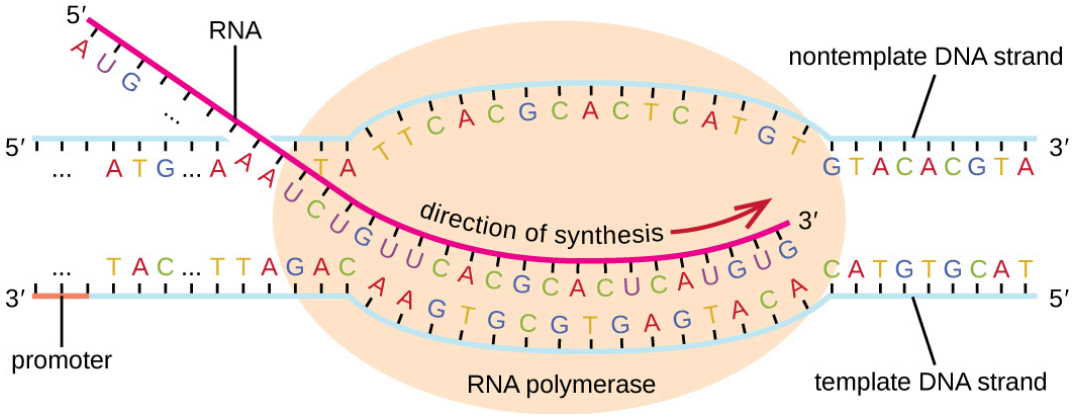


Figure 5. Diagram of a transcription bubble. During elongation, the bacterial RNA polymerase tracks along the DNA template, synthesizes mRNA in the 5' to 3' direction, and unwinds and rewinds the DNA as it is read. In this image, the ellipsis "..." indicates additional bases not shown.

Because only certain parts of the genome are destined to be transcribed, RNA synthesis depends on sequences in the DNA that signal where the transcription machinery should bind (and which strand to use as a template), where transcription should begin, and where transcription should end. These signal sequences, called **DNA elements**, are important to the function of a gene, despite not being incorporated into the RNA itself. They often serve as binding sites for components of the transcriptional machinery.

The process of transcription can be broken into three stages: initiation, elongation, and termination. The rest of this module looks at the mechanism of transcription and the DNA elements controlling RNA production.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=149#h5p-12

Media Attributions

- Dogma of molecular genetics © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- DNA polymerase © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- 3D structure of RNA © Wikipedia is licensed under a CC0 (Creative Commons Zero) license

TRANSCRIPTION IN BACTERIA

The main enzyme that catalyzes the RNA synthesis reaction is RNA polymerase. In prokaryotes, there is only one RNA polymerase, but in eukaryotes there are multiple RNA polymerases that are each responsible for a subset of RNA types. Transcription depends on the interaction between **DNA elements** and **protein factors** involved in transcription.

Initiation

The transcription machinery binds to sequence elements called **promoters**. In bacteria, the core promoter sequence is usually found about 10 base pairs **upstream** of the actual transcriptional start site, or toward the 5' end of the nontemplate strand of the DNA. The bases around a gene are numbered to give geneticists a nomenclature for discussing bases within a gene. The first base that is incorporated into the RNA molecule – the start site of transcription — is called the **+1 site.** The nontemplate strand of the DNA is treated as sort of a number line, counting negative (-1,-2,-3) upstream from the +1 site into the untranscribed region and counting positive into the transcribed region of the gene (+1+2,+3). This is shown in **Figure 6**.

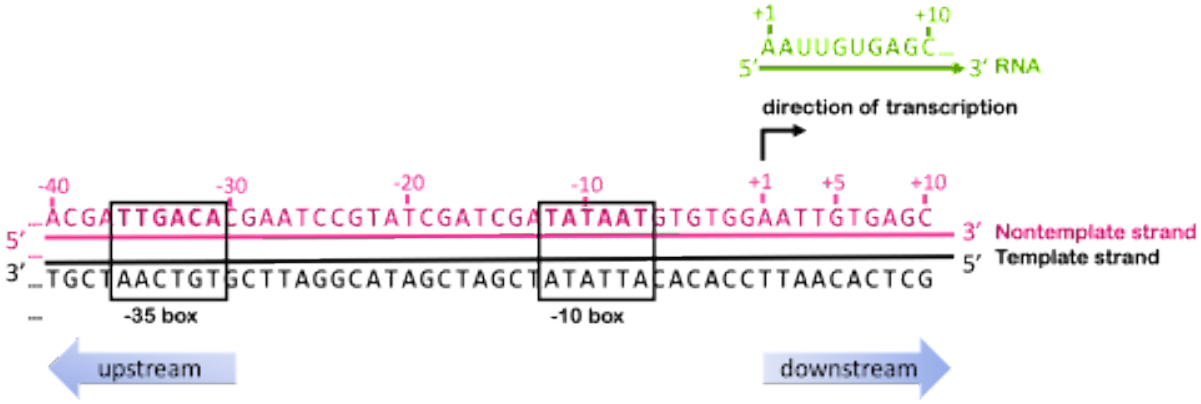


Figure 6. Bases around a gene are numbered relative to the start site of transcription (+1). The bases incorporated into the RNA molecule are numbered with positive numbers (+1,+2,+3) relative to the start site. The bases upstream of the transcriptional start site (toward the 5' end of the nontemplate strand) are numbered with negative numbers (-1,-2,-3). The -10 and -35 boxes are indicated.

Although promoters vary from one gene to another, certain sequence elements are **conserved**, or similar in many promoters. In prokaryotes, two such DNA elements are the -10 box and the -35 box, so named for their position relative to the +1 transcriptional start site. These are also shown in **Figure 6**.

The prokaryotic RNA polymerase is made up of five core subunits: two α ("alpha") subunits, a β ("beta") subunit, a β' ("beta prime") subunit, and an ω ("omega") subunit. During initiation of transcription, the core polymerase becomes associated with a sixth subunit called σ ("sigma"), forming a **holoenzyme**.

The role of the **σ factor** is to recognize promoter sequences, and σ factor disassociates from the polymerase after initiation of transcription. There are several different σ factors, and each of the sigma factors recognize slightly different promoter sequences. One common σ factor in *E. coli* is called sigma70 (σ^{70}). Sigma70 is sometimes referred to as a "housekeeping" or general purpose sigma factor, but other σ factors have specialized functions like the transcription of stress-response genes, genes involved in nutrient uptake, or genes required for sporulation.

Sigma factors bind to the -10 and -35 boxes. The **consensus sequence**, or most commonly recognized sequence, for σ^{70} is 5'TTGACA3' (-35) and 5'TATAAT'3 (-10), although the actual sequences of individual σ^{70}-regulated genes can vary slightly from the consensus. It's important to note that the -10 and -35 boxes are bound by sigma as double-stranded DNA, but by convention only the sequence of the nontemplate strand is given when describing elements of a gene.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=154#h5p-13

Elongation

The sequence and orientation of the -10 and -35 boxes specify which strand is to be used as the transcriptional template. RNA polymerase holoenzyme binding leads to the unwinding of the double helix around the promoter. Transcription initiates at the +1 site and synthesis continues with polymerase adding nucleotides to the 3' end of the growing RNA. We thus say that **transcription proceeds in a 5' to 3' direction**.

Be aware, however, that the polymerase is moving along the template strand toward the 3' end since the RNA is oriented antiparallel to the template strand. In Figure 5, you can see the 3' end of the RNA is paired with the template DNA strand. The red arrow indicates the direction of transcription, pointing toward the 3' end of the RNA and the 5' end of the template strand.

Unlike replication, transcription does not require synthesis of a primer first. Although DNA polymerases

cannot synthesize "*de novo*", RNA polymerases can. The RNA polymerase will move toward the 5' end of the DNA template strand, adding nucleotides to the 3' end of the RNA molecule as it goes.

Also unlike with replication, the transcription bubble closes from behind and opens ahead as transcription proceeds. This means that the 5' end of the RNA molecule is displaced from the template strand, extending outward from the transcription bubble. You can see this in **Figure 5** as well.

You'll also note in both **Figure 5** and **Figure 6** that the RNA molecule is complementary to the template strand – because that is how it is synthesized! The template and the RNA are antiparallel to one another, with the 5' end of the RNA oriented toward the 3' end of the template strand. But the sequence of the RNA molecule is *identical* to the *nontemplate* strand (just with U's instead of T's): same sequence, same 5' to 3' orientation.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=154#h5p-14

Genes are usually transcribed more than once at a time, producing many RNA molecules. If the RNA will be used for protein coding, having many RNA molecules allows the cell to produce many protein molecules at the same time. Multiple RNA polymerases can act on the same gene, one after another, with additional polymerases beginning transcription before the first polymerase is finished. This is illustrated in **Figure 7**.

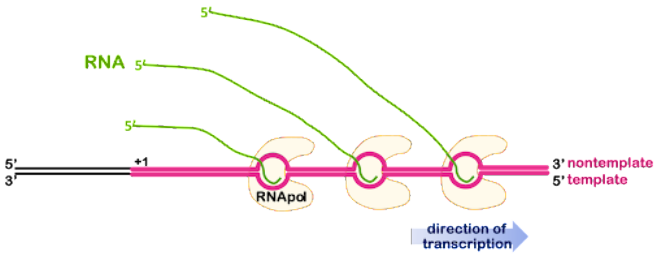


Figure 7. A new RNA polymerase can begin transcribing a gene before the previous one has finished. In this image, three polymerases are

Termination of transcription

RNA polymerase begins synthesis of the RNA at the +1 site and continues until it reaches a **terminator** element in the DNA. Terminators are the signal sequence in DNA that mark the end of transcription in prokaryotes. There are two types of terminators: Rho-dependent and Rho-independent or intrinsic terminators. Rho-dependent terminators require the action of an RNA helicase called Rho and a sequence in

the DNA template that causes the RNA polymerase to pause its elongation of the RNA. Rho unwinds the RNA paired with the template and the RNA is released from the transcription bubble. The polymerase can be recycled and used again to transcribe another RNA molecule.

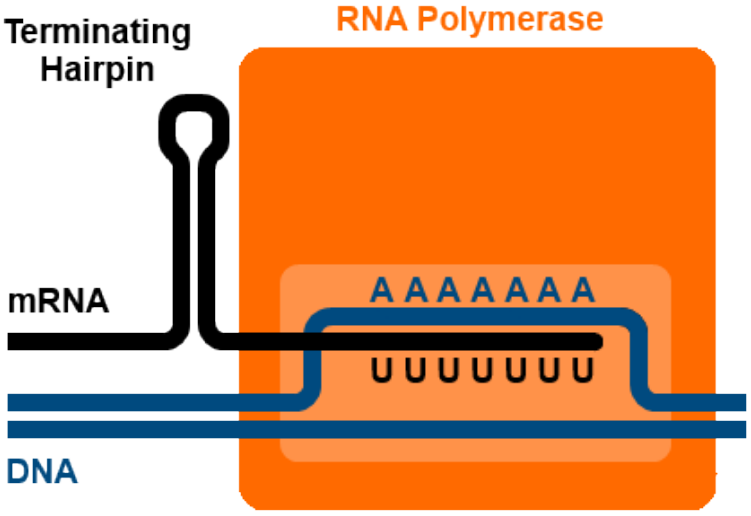


Figure 8. Rho-independent termination occurs in RNA molecules with a GC-rich sequence that can form a hairpin followed by a long string of U's (A's in the template). The hairpin causes a pause and destabilization of the RNA:template interaction. A:U base pairs are inherently weaker than G:C base pairs, and the destabilization leads to the release of the RNA from the transcription complex.

Rho-independent terminators are also called intrinsic terminators. Termination of transcription happens because the RNA structure itself forces the polymerase to release the RNA. RNA molecules with Rho-independent terminators have a GC-rich stretch of base pairs that can fold into a hairpin, followed by a length of U bases. A:U base pairs, like A:T basepairs, are somewhat weaker than G:C base pairs because they form only two hydrogen bonds rather than three. The RNA likely pauses transcription at the A:U base pairs, and the formation of the hairpin destabilizes the RNA:template strand complex enough that the A:U base pairs are not strong enough to hold the RNA and template strands together[1]. This is illustrated in **Figure 8**.

Media Attributions

- Start site of transcription © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike)

1. Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial Transcription Terminators: The RNA 3′-End Chronicles. J. Mol. Biol. 412, 793–813 (2011).

EUKARYOTIC TRANSCRIPTION

In prokaryotes, there is a single RNA polymerase responsible for transcription of all genes. But eukaryotes have three RNA polymerases. RNA Polymerase I (RNA Pol I) transcribes most rRNA; RNA Pol II transcribes mRNA; RNA Pol III transcribes tRNA. Pol II and Pol III each transcribe some snRNA and miRNA. Many of these RNA molecules are also processed post-transcriptionally before they become active.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=160#h5p-15

Like the prokaryotic RNA polymerase, all three eukaryotic polymerases require DNA elements to recognize a promoter, synthesize the transcript, and release the RNA from the transcription machinery. But the prokaryotic RNA polymerase holoenzyme including sigma factor can recognize promoters and begin transcription without other cofactors. The eukaryotic polymerases cannot: they require the action of additional general and regulatory transcription factors to transcribe. Because of this, eukaryotic promoters tend to be more complicated than bacterial promoters.

Although there are similarities in the mechanism of action for all three eukaryotic polymerases, each recognizes a different type of promoter, and each interacts with its own set of transcription factors.

RNA polymerase II transcription initiation

Let's look in detail at RNA Pol II transcription of mRNAs, the protein-coding RNAs. RNA Pol II requires the action of **general transcription factors** to initiate transcription. Most general transcription factors are named by convention TFII_, where TFII indicates they are transcription factors for RNA Pol II, and "_" is a letter that specifies individual factors. Some examples are TFIIB, TFIID, and TFIIH. (The other RNA polymerases also have designated transcription factors. Theirs are named TFI_ and TFIII_.)

Collectively, the general transcription factors and RNA polymerase are called the core or **basal** transcription

machinery. The word basal has the same root word as "basic" or "basis": it is the basic transcription machinery for all RNA Pol II-transcribed genes.

The general transcription factors assemble onto a **core promoter** about 30 base pairs from the start site of transcription. RNA Pol II-transcribed genes have a common sequence called the TATA box within the core promoter. The TATA box is so named because it has lots of T's and A's: the consensus sequence for the TATA box is "TATA(T/A)A(T/A)", although individual genes vary slightly. Like the sequences of the -10 and -35 boxes, the consensus sequence is found in the nontemplate strand upstream of the transcriptional start site.

The TATA box serves a similar purpose as the -10 box in prokaryotes. It is bound by a protein called TATA binding protein (TBP). TBP is part of a larger complex called TFIID. Additional general transcription factors bind after TFIID, recruiting RNA Pol II to the promoter **(Figure 9)**

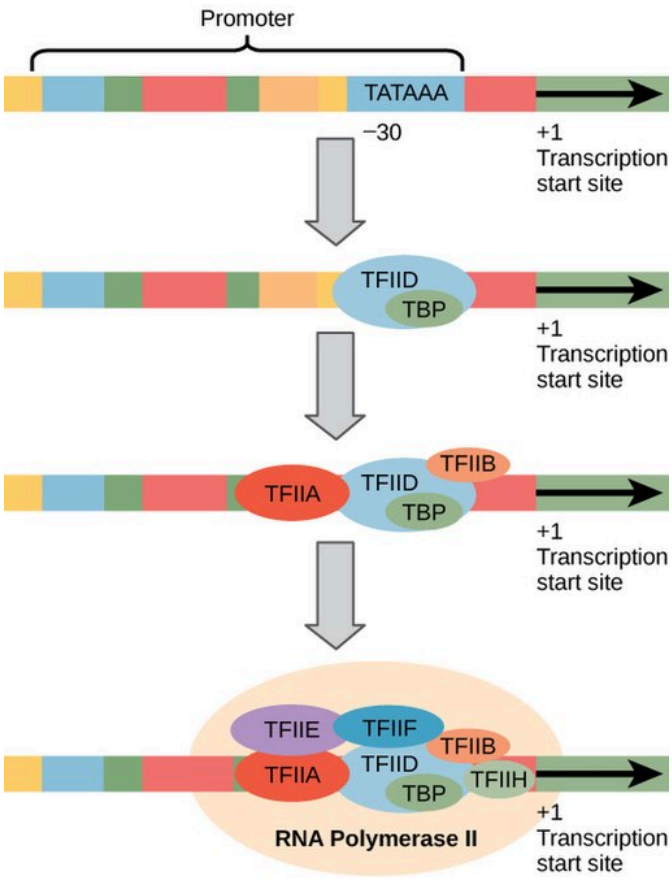


Figure 9. The core transcription machinery assembles on the core promoter, beginning with TBP/TFIID. TFIIB and TFIIA bind next, then recruit the polymerase and additional general transcription factors. Regulatory transcription factors may assemble on other binding sites within the promoter.

In addition to the TATA box, the promoter often contains other common sequence elements. Two such elements are GC boxes and the CAAT box, again named for the nucleotides of the consensus sequence. The CAAT box is usually found about 60-100 bases from the +1 site, and GC boxes about 80-110 bases from the +1 site. Eukaryotic promoters are highly variable! Not all genes have CAAT boxes or GC boxes. In fact, some genes don't have TATA boxes, either! In some TATA-less promoters, the presence of multiple GC boxes is sufficient to assemble the transcription machinery and initiate transcription.

About 100-250 base pairs from the +1 site of transcription are additional proximal promoter elements that play a role in gene regulation: they may be bound by general transcription factors or **regulatory transcription factors.** Regulatory transcription factors are also called **specific** transcription factors: they are specific to the adjacent gene, in contrast to the general transcription factors that facilitate transcription of all RNA Pol II-transcribed genes. Regulatory elements and regulatory transcription factors are discussed more in detail in the module on Gene Regulation.

Thousands of base pairs away – upstream or even downstream! – eukaryotic promoters may have **enhancers**

or **silencers**: DNA elements with binding sites for multiple regulatory transcription factors. The core promoter, proximal elements, and far-away enhancer/silencer elements are shown in **Figure 10**.

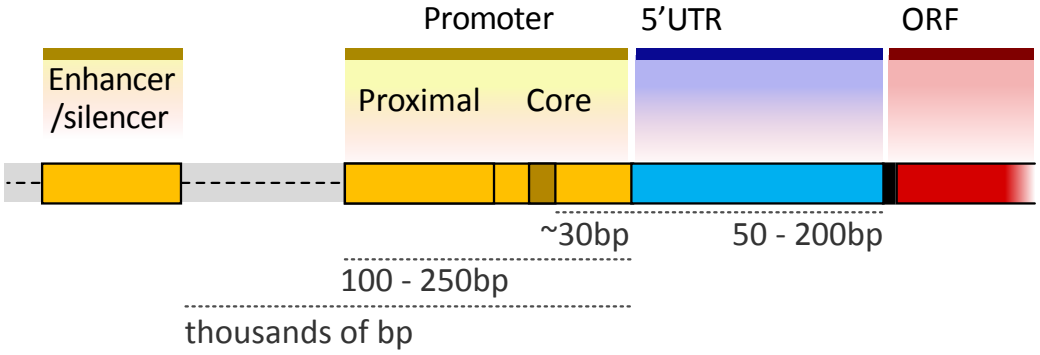


Figure 10. Eukaryotic RNA Pol II promoters consist of a core promoter, proximal elements, and more distant enhancers or silencers.

These distal regulatory elements help the polymerase bind to the promoter when conditions are right for the gene to be active. Although they can be tens of thousands of base pairs away from the core promoter when measured linearly, DNA is a flexible molecule! It bends, twists, and loops in the nucleus of the cell. Elements that are far apart on the chromosome can actually be adjacent in three-dimensional space if the DNA loops around.

This is shown in **Figure 11**. Generally the proteins that bind to enhancers interact with other parts of the transcription machinery to help stabilize the polymerase at the promoter enough to initiate transcription. Like the proximal regulatory elements, enhancer elements will be discussed in more detail in the module on Gene Regulation.

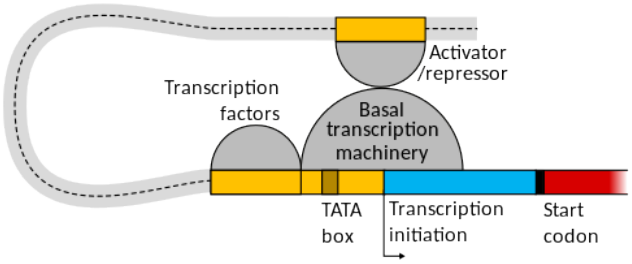


Figure 11. DNA is flexible and can loop to bring distal elements into close proximity of the promoter in three dimensional space.

RNA polymerase II termination

Once the polymerase and supporting transcription factors have bound to the promoter, the polymerase will synthesize the first nucleotides of the RNA, beginning at the +1 site and continuing downstream from the +1 site. But there are no terminators like those seen in prokaryotes. Instead, each of the three eukaryotic RNA polymerases terminates transcription via different DNA elements and/or protein factors.

RNA Pol II-transcribed genes are terminated via **polyadenylation**. Polyadenylation adds up to 250 adenosine nucleotides to the 3' end of an mRNA molecule. In the process, transcription is terminated. These

A nucleotides are not templated: there are no corresponding T's in the template strand of the DNA. The length of A's is called the **poly-A tail**.

RNA pol II transcription machinery includes polyadenylation factors CPSF (Cleavage and Polyadenylation Specificity Factor) and CstF (Cleavage Stimulating Factor), which ride along with the polymerase as it transcribes. Once the polyadenylation signal sequence AAUAAA is transcribed, CPSF and CstF are transferred to the RNA molecule. CstF cleaves, or cuts, the RNA molecule about 35 nucleotides downstream of the polyadenylation signal, even while transcription continues at the 3' end of the RNA. PAP, or Poly-A Polymerase, adds adenosine nucleotides to the new 3' end of the cut RNA. This is shown on the left side of **Figure 12**.

Transcription of the original RNA molecule continues beyond the polyA signal sequence. Rat1 "torpedo" exonuclease then binds to the new 5' end created by the cleavage. An exonuclease is an enzyme that breaks down nucleic acids starting at one end. Rat1 breaks down the remnant RNA from 5' toward the 3' end. It is called a "torpedo" exonuclease because one hypothesis is that this movement toward the 3' end acts like a torpedo to break apart the DNA/polymerase complex and terminate transcription. This is shown on the right side of Figure 12. The dismantling of the transcription machinery is not well understood and likely requires the action of additional DNA elements and protein factors[1].

The transcripts of the other RNA polymerases, other methods of termination, and transcripts of RNA Pol I and RNA Pol III are not typically polyadenylated[2]. In addition to participating in transcriptional termination, Poly-A tails are important for the export of mRNAs from the nucleus and to the stability of mRNAs over time.

1. Rodríguez-Molina, J. B., West, S. & Passmore, L. A. Knowing when to stop: Transcription termination on protein-coding genes by eukaryotic RNAPII. Mol. Cell 83, 404–415 (2023).

2. Arimbasseri, A. G., Rijal, K. & Maraia, R. J. Transcription termination by the eukaryotic RNA polymerase III. Biochim. Biophys. Acta 1829, 318–330 (2013).

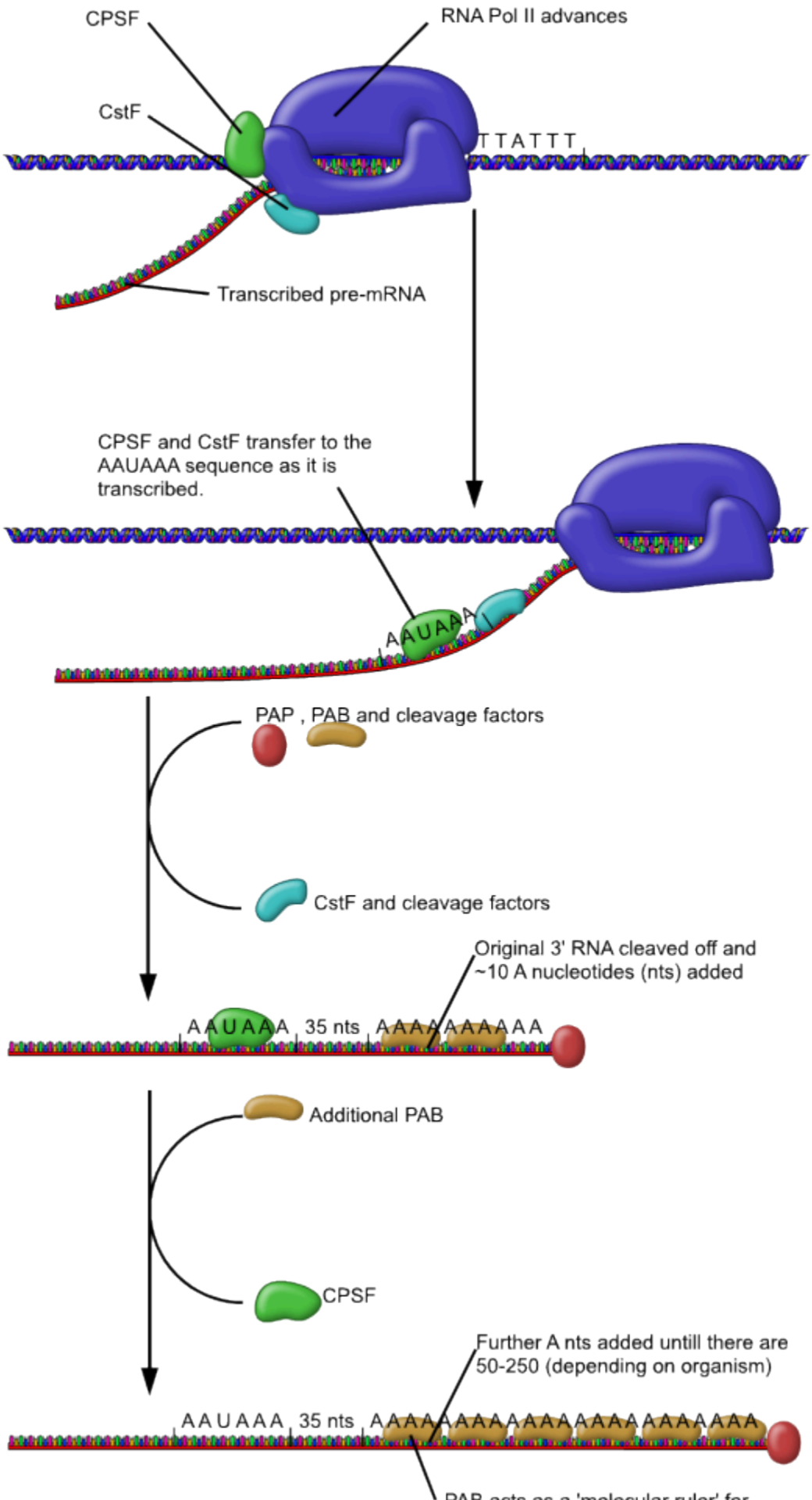

CPSF
RNA Pol II advances
CstF
TTATTT
Transcribed pre-mRNA
CPSF and CstF transfer to the AAUAAA sequence as it is transcribed.
AAUAAA
PAP , PAB and cleavage factors
CstF and cleavage factors
Original 3' RNA cleaved off and ~10 A nucleotides (nts) added
AAUAAA 35 nts AAAAAAAAAA
Additional PAB
CPSF
Further A nts added untill there are 50-250 (depending on organism)
AAUAAA 35 nts AAAAAAAAAAAAAAAAAAAAAAAA

Figure 12. Process of polyadenylation. The eukaryotic transcription machinery includes polyadenylation factors like CPSF and CstF, which ride along with the polymerase until the Poly-A signal sequence, AAUAAA, is transcribed. CPSF and CstF are transferred to the growing RNA molecule and cleave the molecule about 35 nucleotides downstream of the pA signal. PAP adds A nucleotides to the new 3' end of the RNA – up to 250 nucleotides, depending on the organism. The 3' remainder of the RNA stays associated with the polymerase and transcription continues beyond the signal sequence. The exonuclease rat1 binds to the new 5' end of the RNA, degrades the RNA toward the 3' end, and triggers the release of the transcription machinery from the transcription bubble.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=160#h5p-16

Media Attributions

- Core promoter © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Transcription initiation eukaryote B © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Polyadenylation © Wikipedia is licensed under a CC0 (Creative Commons Zero) license

RNA PROCESSING OF RNA POL II TRANSCRIPTS

The poly-adenylation of RNA Pol II transcripts is one example of the **RNA processing** that occurs in eukaryotes. Before processing, the RNA Pol II transcript is called the primary transcript, or **pre-mRNA.** Only after processing is complete is it described as an mRNA.

Primary transcripts produced by RNA Pol II undergo three main types of processing:

1. 5' capping
2. Splicing
3. Poly-adenylation (discussed in the previous section)

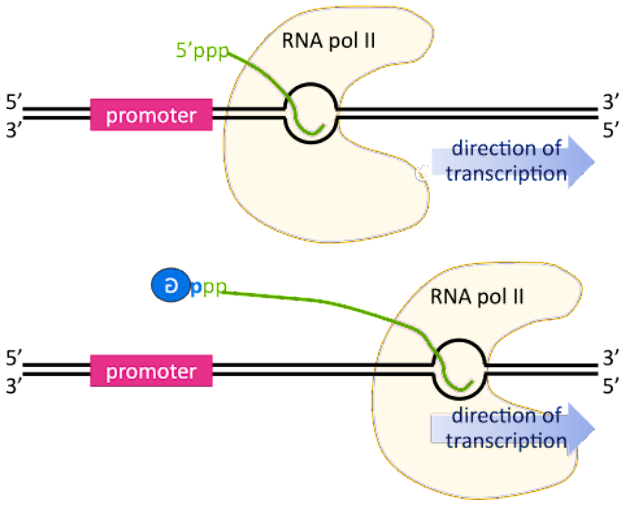


Figure 13. Capping. As the 5' end of the growing pre-mRNA extends outward from the RNA polymerase, it is capped by the addition of an inverted guanosine nucleotide. This connects the 5' alpha phosphate of the guanosine cap with the beta and alpha phosphates of the 5'-most nucleotide of the pre-mRNA.

These modifications of the pre-mRNA are sometimes called post-transcriptional modifications, although this is not a very good description because all three types of modifications are made to RNA as it is still being transcribed. Modification of the transcripts happens in the nucleus, but after processing the mRNA molecules will be exported to the cytoplasm for translation/protein synthesis.

5' capping

As transcription transitions to the elongation phase, the transcription bubble is translocated down the length of the gene, and the 5' end of the RNA is displaced from the template strand and extends outward from an exit channel in the RNA polymerase.

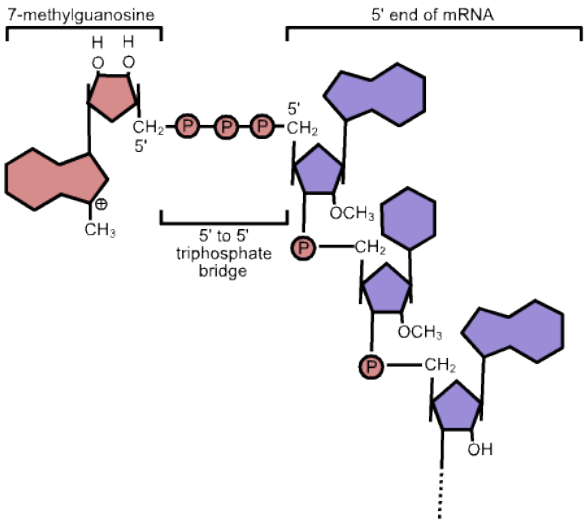


Figure 14. Structure of the cap.

The pre-mRNA initially has a triphosphate at the 5' end, since the beta and gamma phosphates are only lost from a nucleotide when it is added to the 3' end of a growing chain (review Figure 2 for a reminder!).

The 5' end is **capped** by the addition of a G nucleotide. The capping reaction connects the 5' end of the RNA to the 5' position of the capping guanosine, so it is sometimes described as "inverted" or "backward". In **Figure 13** this is indicated by the backward "G" in the blue circle.

The capping reaction retains the alpha and beta phosphates from the 5' end of the RNA and the alpha phosphate from the capping guanosine, so there is an unusual linkage of 3'HO-sugar-phosphate-phosphate-phosphate-sugar at the capped end. After the guanine is linked to the RNA, it is then methylated at position 7 (a -CH_3 group is added). This chemical structure of the cap is shown in **Figure 14**.

The 5' cap serves multiple purposes. First, it participates in the translocation of the mRNA out of the nucleus to the cytoplasm for translation. Second, it protects the mRNA from degradation. And third, it serves as a ribosome binding site during translation.

Interestingly, the 5' cap is exploited by some viruses that infect eukaryotic cells. Viruses like Influenza A have an RNA genome, and specialized viral RNA-dependent RNA polymerases transcribe using the RNA genome as a template. But the products of viral transcription are not capped by the virus. Instead, the virus "snatches" caps from cellular RNA molecules, along with about 10-15 base pairs of the cellular RNA. These stolen caps are used to prime RNA synthesis to make viral RNAs. Because cap-snatching is unique to viruses – this is

not a function uninfected cells perform – the cap-snatching mechanism is a drug target for treatment of viral infections. Inhibiting cap-snatching blocks the viral life cycle without harming the host[12].

Splicing

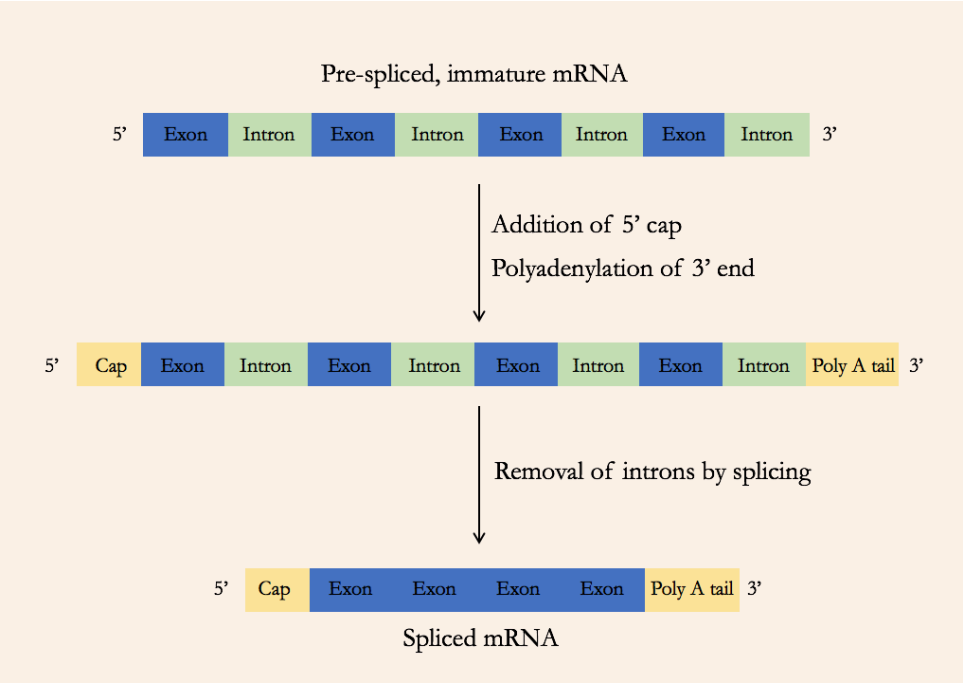


Figure 15. Introns are removed from pre-mRNA via splicing.

In bacteria, the protein-coding sequence of a gene is present in the genome in one long continuous stretch. But in eukaryotes, the protein-coding sequence can be broken up into many segments within the genome, interspersed with noncoding DNA. The pre-mRNA includes both the protein-coding sections (called **exons**) and the intervening noncoding sections (called **introns**). But the mature RNA includes only the exons: the introns are removed during a process called splicing (**Figure 15**).

1. Samji, T. Influenza A: Understanding the Viral Life Cycle. Yale J. Biol. Med. 82, 153–159 (2009).

2. Noshi, T. et al. In vitro characterization of baloxavir acid, a first-in-class cap-dependent endonuclease inhibitor of the influenza virus polymerase PA subunit. Antiviral Res. 160, 109–117 (2018).

The most common splicing mechanism uses a complex called the **spliceosome** to remove the introns and connect exons together. The general mechanism of splicing is shown in **Figure 16.**

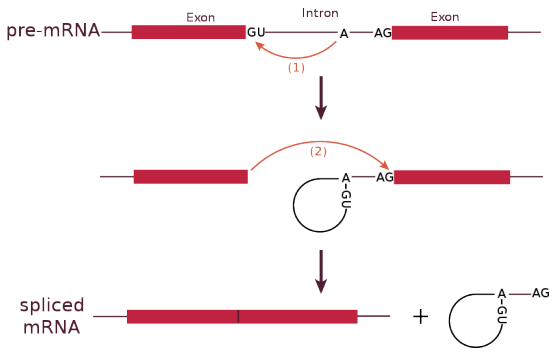


Figure 16. Mechanism of splicing:
1. A loop is formed by a bond between the branch point A and the 5' end of the intron. This separates the 5' end of the intron from exon 1.
2. The 3' end of exon 1 is connected to the 5' end of exon 2, releasing the looped intron from the mRNA.

Within the intron is a sequence called the branch point. The spliceosome forms a covalent bond between an A nucleotide at the branch point and the first (5') nucleotide of the intron. This breaks the sugar-phosphate backbone at the 5' intron/exon boundary and forms a loop, or a **lariat**, in the intron.

A new phosphodiester bond is then formed between the newly exposed 3'-OH group of the first exon and the 5' phosphate at the beginning of the second exon, at the 3' intron/exon boundary. This releases the intron from the second exon as a lariat structure. The lariat is later degraded in the cell.

Like the other processes of molecular genetics, splicing is regulated by the interaction between sequence elements and regulatory factors.

The spliceosome is a large multi-subunit complex assembled from multiple **snRNPs**, pronounced "snurp". snRNP stands for "small nuclear ribonucleoproteins." The spliceosome snRNPs are named U1, U2, U4, U5, and U6. (There is also a U3 in the cell, but it does not participate in splicing).

Each snRNP has an RNA and a protein component. The snRNPs bind to consensus sequences at the intro/exon boundaries via complementary base pairing between the snRNAs and the pre-mRNA. These sequences are called the 5' and 3' splice site, after their orientation at the 5' and 3' ends of the intron. snRNPs also bind to the region of the RNA around the branch point A. The three sequences are shown in Figure 17[3].

The sequence at the 5' splice site, the branch point, and the 3' splice site can vary from intron to intron. Even multiple introns from the same gene will not have identical sequences. But some of the bases are highly conserved: nearly every eukaryotic intron begins with the bases "GU" and ends with the bases "AG". These bases are bolded in **Figure 17**.

3. Book: Cells – Molecules and Mechanisms (Wong). Biology LibreTexts https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Cells_-_Molecules_and_Mechanisms_(Wong) (2018).

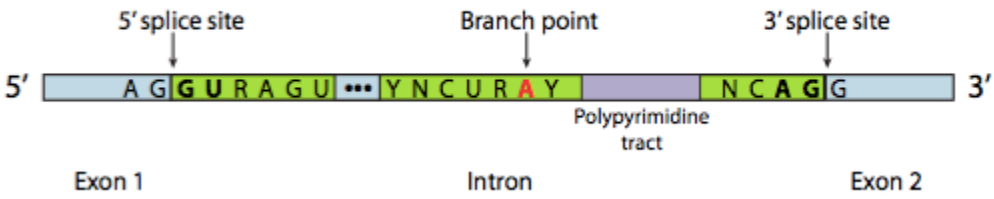


Figure 17. Consensus sequences of the 5' splice site, the branch point, and the 3' splice site.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=168#h5p-17

The spliceosome-mediated splicing described here is a common mechanism of splicing. But it is worth mentioning that some introns are spliced via other means. Some introns are spliced via another spliceosome that uses different snRNPs. And some introns are self-splicing! Remember that RNA can act as an enzyme. Self-splicing introns catalyze their own removal from the pre-RNA.

Alternative splicing

Most eukaryotic genes have at least one intron, but some have many more. The human dystrophin gene has 78 introns!

In some cases, a single gene can be used to produce multiple RNAs. This is accomplished through **alternative splicing**. With alternative splicing, not all exons are incorporated into a mature mRNA. Which exons are included are regulated by a collection of splicing factors, which block the use of some 5' and 3' splice sites and promote the use of others. Alternative splicing is illustrated in **Figure 18**. In **Figure 18**, a gene with five exons is transcribed into pre-mRNA. The pre-mRNA can be spliced to include all five exons or alternatively spliced to skip one or more exons. The alternatively spliced mRNA will encode different forms of the translated protein.

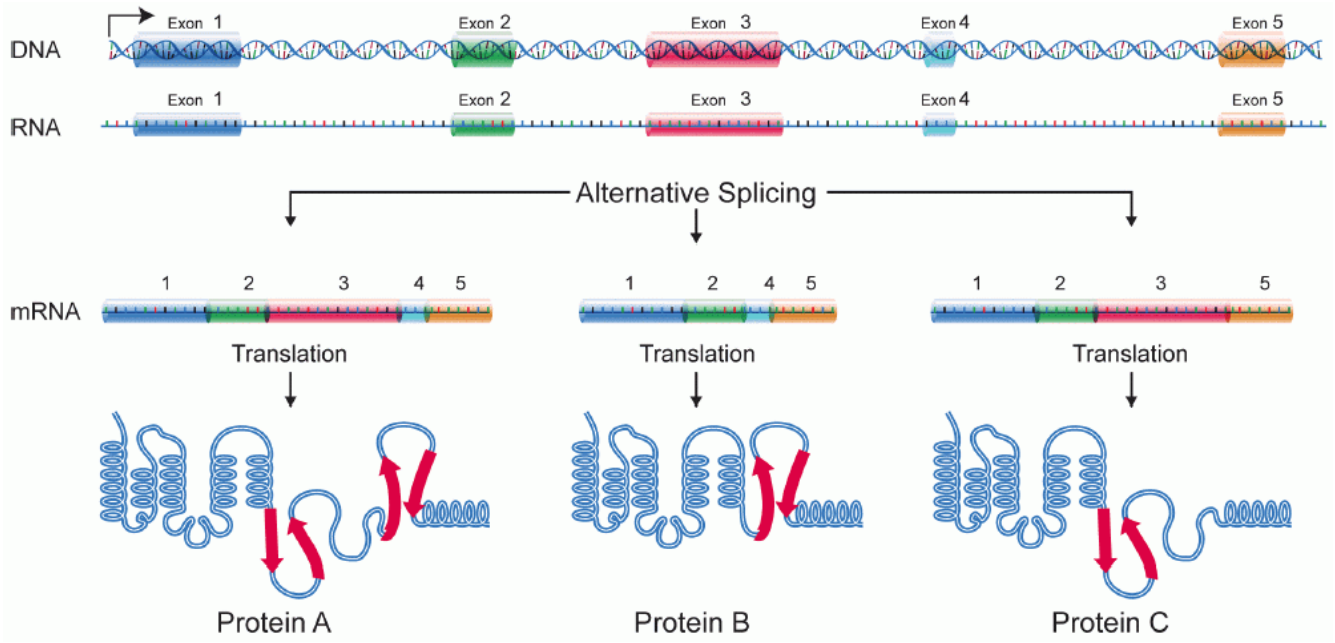


Figure 18. Alternative splicing produces three different mRNAs from the same pre-mRNA. The translation of the mRNA will produce three different protein isoforms. Protein A results from the translation of all the exons, whereas Proteins B and C result from exon skipping.

Alternative splicing builds versatility and efficiency into the genome: only one gene is needed to produce multiple proteins. Alternative splicing is also an exception to the "one gene, one polypeptide" rule discussed at the beginning of this module: with alternative splicing, one gene can produce several different polypeptides.

Additional forms of RNA processing

Although not discussed extensively in this text, RNA can undergo additional forms of post-transcriptional modification. RNAs can be edited through the addition or deletion of bases or by converting one base to another. This includes the modification of bases to incorporate non-canonical (unusual) bases. tRNAs tend to have many modified bases, including the non-canonical structures like inosine and pseudouracil. Inosine is formed by the modification of adenosine. Psuedouracil is formed by modification of uracil.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it

online here:
https://rotel.pressbooks.pub/genetics/?p=168#h5p-18

Media Attributions

- Capping © Amanda Simons is licensed under a CC0 (Creative Commons Zero) license
- Structure of the cap © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Messenger RNA maturation © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- RNA splicing reaction © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Consensus sequences © LibreTexts is licensed under a CC0 (Creative Commons Zero) license
- DNA alternative splicing © Wikipedia is licensed under a CC0 (Creative Commons Zero) license

SUMMARY

- The process of transcription synthesizes an RNA molecule that is complementary to the template strand of the gene. But it is identical to the nontemplate strand in both 5' to 3' polarity and sequence, with the substitution of U in the RNA for T in the DNA.
- Some RNA molecules encode protein sequences and will later be used in translation (mRNAs). Other RNA molecules function directly in different cellular processes.
- Transcription of a gene requires DNA elements that signal where the transcription machinery binds to the DNA, where transcription begins, and where transcription ends. These sequences are summarized in **Table 2**. Many of these DNA elements are bound by protein factors that facilitate transcription.

Table 2: Summary of elements required for transcription

	Transcription machinery is first recruited to	**The RNA molecule begins with**	**The RNA molecule ends with**
Prokaryotes	-10 and -35 boxes in the promoter (bound by σ factor)	**+1** nucleotide	The last base of the **terminator**
Eukaryotes	TATA box in the promoter (bound by TATA-binding protein, part of TFIID)	**+1** nucleotide	The **poly-A cleavage site** followed by a long stretch of untemplated A's

- In eukaryotes, the primary transcript produced by RNA polymerase II is processed to become a mature mRNA. Processing includes the addition of a 5' cap, the addition of a poly-A tail, and splicing to remove introns.
- Alternative splicing allows multiple mature mRNAs (and, later, multiple distinct proteins) to be produced from a single gene.

WRAP-UP QUESTIONS

1. Give examples of two types of RNA that are not used to make protein.

2. Sequence relationships:

a. If an RNA molecule has the sequence 5'AGGCCU3', give the sequence of the template and nontemplate DNA strands.
b. If part of a DNA template strand has the sequence 5'TTCGAA3', give the sequence of the RNA and nontemplate DNA strands.
c. If part of a DNA nontemplate strand has the sequence 5'CCTGAG3', give the sequence of the RNA and the template DNA strands

3. Compare and contrast the processes of replication and transcription: Create a table that could be used as a study tool. In your table, be sure to include the chemical composition of DNA and RNA, the biochemistry of DNA and RNA synthesis, and the templates and the products of the reaction including the parts of the genome used.

4. Draw a diagram of a transcription bubble. Label 5' and 3' ends of the template, nontemplate, and RNA strands, and indicate the direction of transcription.

5. A simplified gene sequence is diagrammed below. The underlined sequences are, in order, the -35, -10, +1, and terminator sequences.

5'aaTTGACAactgtgtgatgtagcttaTATAATatgatgcta**C**gttaaaggaggtgggtaatgcctaaatggcggtaacgtgaGGCCGC GGCCTTTTcccc3'

a. What role does each of these underlined elements play in transcription?
b. Is this the template strand or nontemplate strand of the DNA?
c. What is the sequence of the RNA that is produced from this gene?
d. Is this a eukaryotic or prokaryotic gene?

6. Draw a diagram of a prokaryotic gene and the DNA elements necessary for transcription. Include the -10 and -35 boxes, the site at which transcription begins, and the site at which transcription ends.

7. Draw a diagram of a eukaryotic gene with two introns. Include the DNA elements necessary for transcription and RNA processing. Include the 5' and 3' splice sites, the TATA box, a CAAT box, the polyadenylation signal, the start site of transcription, and the approximate location where transcription would end.

8. Here are some links to additional practice questions for splicing:

   a. Diagram the meg-1 gene.
   b. Alternative splicing of the egl-15 gene

Science and Society

9. Choose one of the researchers mentioned in the modules on Replication or Transcription. Ask ChatGPT or another generative AI to write a 500-1000 word biography about them. Repeat this for the same researcher multiple times. What common themes appear? How does that researcher's identity and background appear to influence their work?

PART VI

TRANSLATION AND GENETIC CODE

Objectives

By the end of this module, you should be able to:

1. Describe the levels of organization of protein structure and recognize that the amino acid sequence influences protein structure.
2. Describe the stages of translation (initiation, elongation, and termination).
3. Identify the location of signal sequences necessary for translation (Shine-Dalgarno/ribosome binding sequence, start codon, stop codon) within DNA or RNA sequence.
4. Find an open reading frame within a sequence of DNA or RNA and predict the polypeptide sequence encoded by an RNA or DNA.
5. Define the following terms: codon, anticodon, ribosome, tRNA, N-terminus, C-terminus, amino acid, polypeptide, Shine-Dalgarno sequence, large subunit, small subunit, start codon, stop codon, open reading frame.

Introduction

Translation is the process of protein synthesis. In many ways, proteins can be thought of as the molecular workers of a cell. Proteins are used as structural components of the cell, to import and export materials from the cell, to metabolize nutrients, to build macromolecules, to communicate with other cells, to perform DNA replication and transcription, and for many other purposes. Nearly every cellular function requires the action of proteins.

Proteins are polymeric macromolecules assembled from amino acid subunits, per instructions encoded in genes. In the Transcription module, we saw how the cell transcribes genes into RNA. In this module, we will look at how the large multicomponent **ribosomes** translate mRNA (messenger RNA) to synthesize proteins.

This requires tRNAs (transfer RNA) that act as adaptors, carrying amino acids into the ribosome to match the mRNA sequence, as well as additional protein factors.

This module includes the following sections:

- An overview of protein structure
- A comparison of the prokaryotic and eukaryotic ribosomes
- A look at the genetic code and the structure of tRNAs
- A description of the steps of translation, including initiation, elongation, and termination.
- A review of gene structure, with elements required for both transcription and translation

PROTEIN STRUCTURE

Amino acids are the molecular building blocks used to assemble proteins. The basic structure of an amino acid is shown in **Figure 1**. Every amino acid consists of a central carbon around which four functional groups are arranged: an amino group ($-NH_3^+$), a carboxylic acid ($-COO^-$), a hydrogen (H), and an "R" group that is variable from amino acid to amino acid. Under physiological conditions, the amine, the carboxylic acid, and many of the R groups are charged due to ionization in the aqueous environment of the cell.

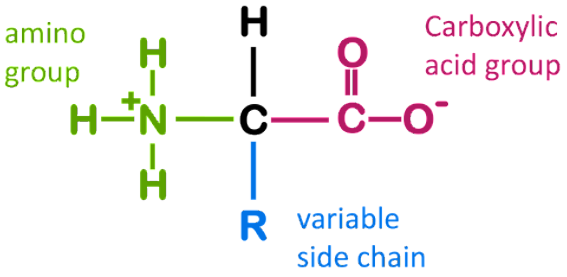


Figure 1. Structure of an amino acid.

A polypeptide is a polymer of amino acids linked via peptide bond (**Figure 2**). Figure 2 shows the structure of a dipeptide, but a polypeptide can have hundreds or thousands of amino acids linked together. Once the amino acids are incorporated into a polypeptide chain, they are called amino acid **residues**.

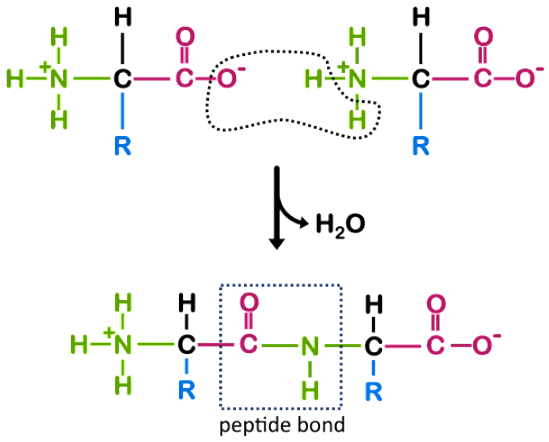


Figure 2. Peptide bond between two amino acids is a condensation reaction that removes a water molecule.

In **Figure 2**, you'll see that the backbone of the dipeptide has the repeating structure -N-C-C-N-C-C-, with each of the amino acid residues contributing one "-N-C-C-" unit to the repeat. Note that the backbone has polarity: at one end of the structure is a free amino group, and at the other is a carboxylic acid group. All polypeptides, regardless of length, have this same polarity: an amino group at one end and a carboxylic acid at the other. The ends of the polypeptide are therefore called the amino- or N-terminus and carboxyl- or C-terminus. During translation, the ribosome adds subsequent amino acids to the C-terminus of a growing polypeptide chain, so we say that translation proceeds from the N- to C- terminus just like DNA and RNA synthesis proceed from 5' to 3'.

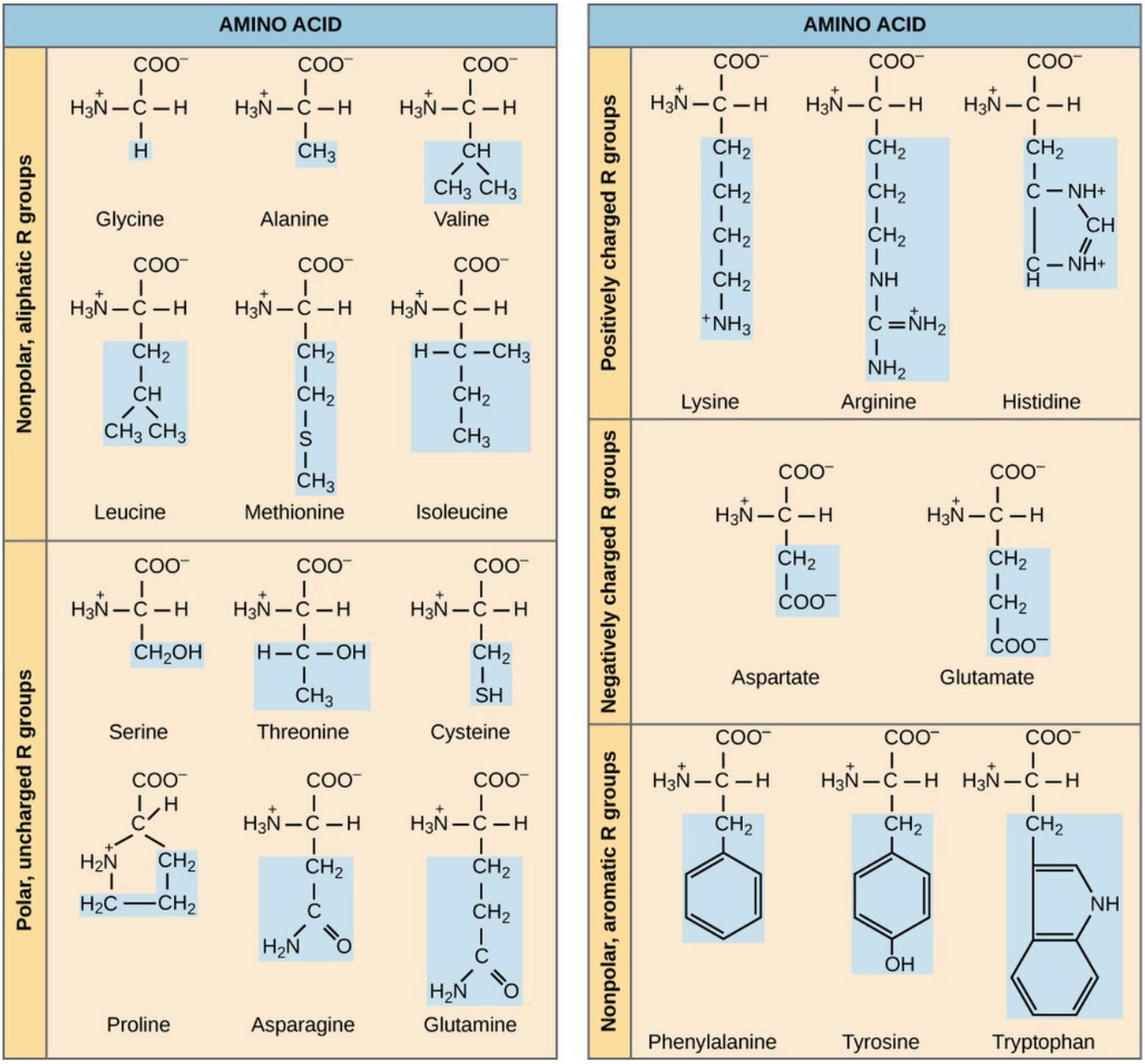


Figure 3. Amino acid structures.

There are twenty amino acids commonly used for proteins in the cell. The names and structures of these are shown in **Figure 3**. **Figure 3** sorts the amino acids based on the chemistry of their side chains. You'll notice that there are nonpolar, polar uncharged, positively charged, negatively charged, and uncharged aromatic side groups. The chemistry of these side chains affects the structure and function of the protein.

The polypeptide will fold up into a three dimensional structure as it is synthesized, depending on the interactions among the R-groups of the amino acid residues and the backbone of the polypeptide. This is dependent on ionic bonds, hydrogen bonds, covalent bonds, and hydrophobic interactions. Examples of these bonds are shown in **Figure 4**.

In Figure 4, the backbone of the polypeptide is depicted as a ribbon, with selected amino acid side chains

shown participating in folding. You can see an ionic bond between positively charged lysine and negatively charged aspartate, a hydrogen bond between serine and and asparagine, hydrophobic interactions between two valines, and a covalent disulfide linkage between two cysteine side chains. These types of bonds hold the folds of the backbone in place, and the sum of these intramolecular bonds results in a distinctive three dimensional structure for each protein.

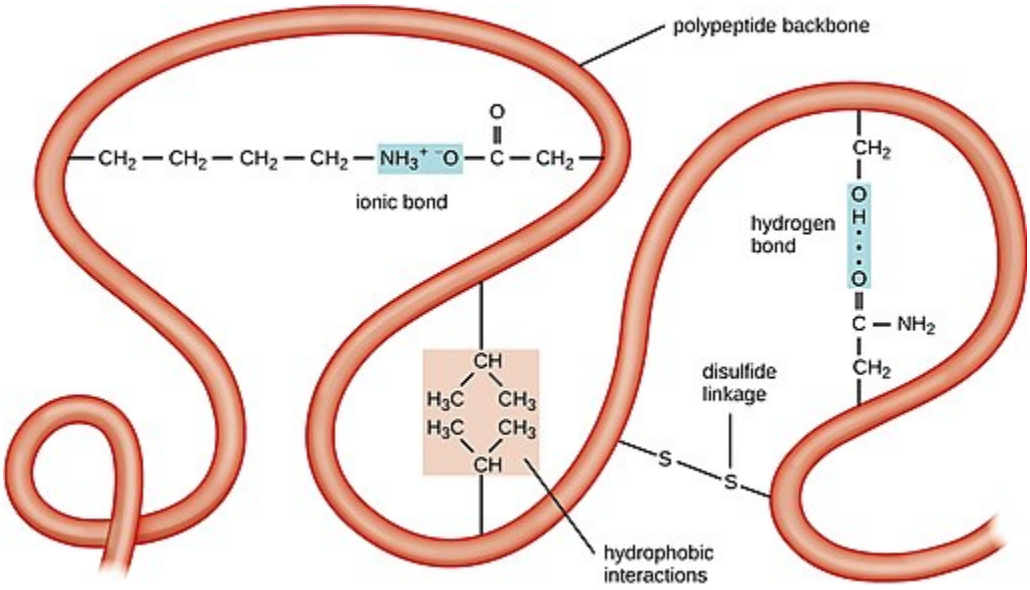


Figure 4. Types of bonds in protein folding. The backbone of a polypeptide is shown in orange. Selected amino acid side chains are shown, extending from the backbone to participate in protein folding.

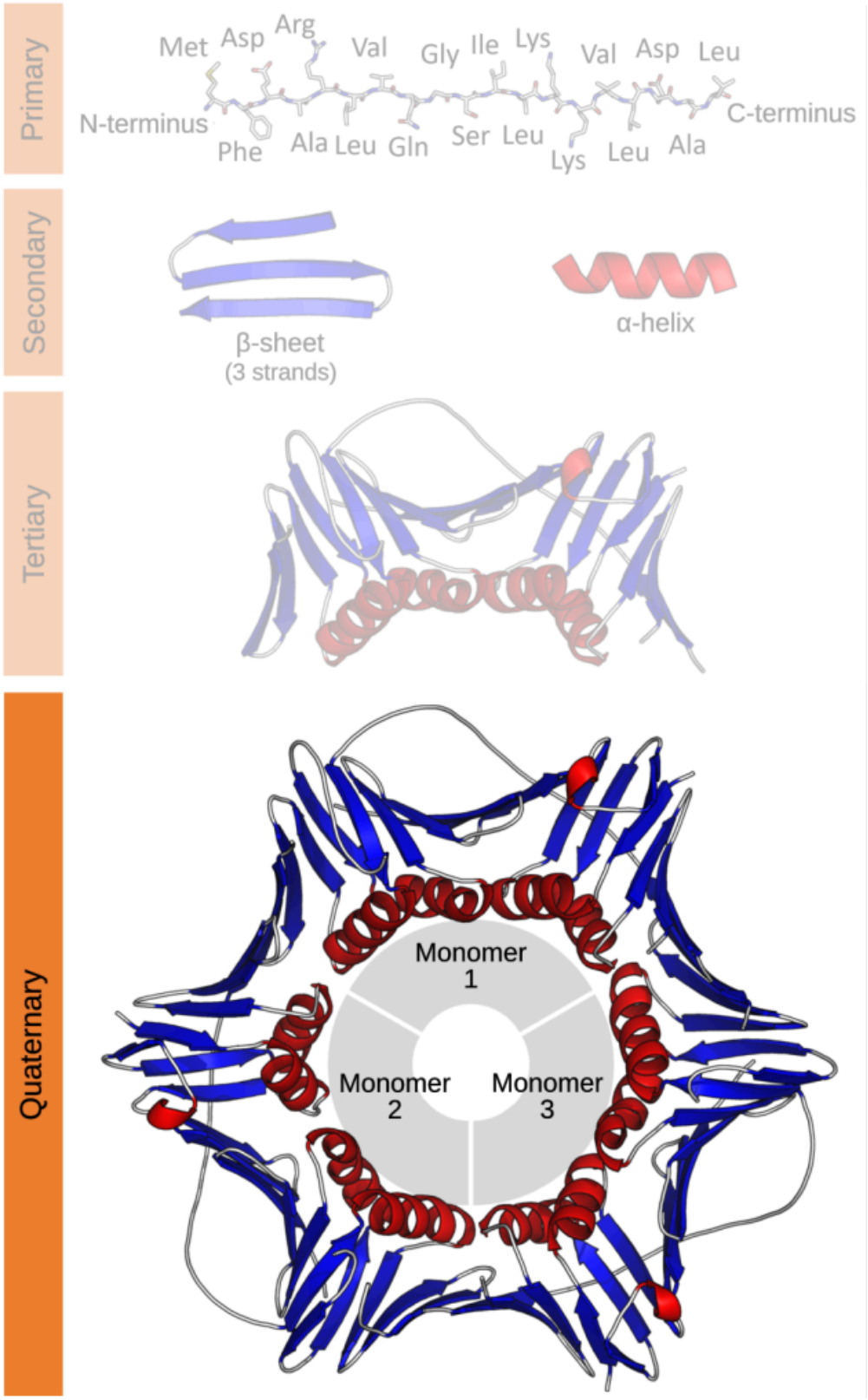


Figure 5. Illustration of the levels of three-dimensional structure of protein molecules: primary, secondary, tertiary, and quaternary.

The structure of a protein can be described in terms of its primary, secondary, tertiary, and quaternary structure. The **primary (1^o)** structure is simply the order of the amino acids in the polypeptide. Convention is to list them in order from N-terminus to C-terminus. The **secondary structure (2^o)** refers to the recognizable elements **alpha helix** and **beta sheet** within the larger structure. An alpha helix is a region of the protein that folds into a coil, while a beta sheet is a region where the polypeptide backbone folds back and forth in a pleated structure.

Tertiary structure is the full three-dimensional structure of the folded polypeptides. Most proteins have both alpha helices and beta sheets within their **tertiary (3^o)** structure. However, it is not uncommon for a protein to be primarily alpha helices or beta sheets. Some functional proteins are made up of multiple polypeptide chains. These proteins have **quaternary (4^o)** structure. Not all proteins have quaternary structure.

Examples of this hierarchy of protein structure are shown in **Figure 5**.

Media Attributions

AN INTRODUCTION TO THE GENETIC CODE

The sequence of bases in the mRNA molecule specifies the order of the amino acids to be assembled by the ribosome. But it's not a 1 to 1 translation of base to amino acid. There are only four nucleotides, but there are 20 amino acids to encode. So early on, molecular biologists hypothesized that the **genetic code** must read multiple bases together as a unit, called a **codon**.

The code is degenerate.

The genetic code uses three-base codons, reading 5' to 3' along the RNA. There are 64 possible combinations of three bases and 64 codons. 61 of these codons specify amino acids, and 3 are used as stop codons to mark the end of the coding sequence of a protein. Because there are 64 codons but only 20 amino acids, there are extra codons to go around and some amino acids are specified by more than one codon. Because of this, the genetic code is said to be **degenerate**.

The code is universal.

The code is also **universal** – with only a few very rare exceptions, the meaning of these codons is consistent among all living organisms. A **codon table** is shown in **Figure 8**. All 64 possible combinations of bases are listed, organized by the first base of the codon in four rows, the second base of the codon in four columns, and the third base of the codon on different lines in each of the 16 boxes of the table. The codons are read 5' to 3', left to right, so a GUU valine codon could be more completely described as 5'GUU3'.

There are four special codons to notice: First, the codon AUG is labeled "Met or Start". AUG specifies the amino acid methionine. AUG is also the first codon to be translated in nearly every protein, which is why it is called the start codon. Methionine is thus the first amino acid incorporated into nearly every protein, although in prokaryotes a specially modified methionine called formyl-methionine (or f-Met) is used as the initiating amino acid. AUG codons can also be found in the middle of a coding sequence, where they also specify methionine.

Second, the codons UAA, UAG, and UGA are labeled "stop". These codons do not specify an amino acid. Instead, they are used to signal the end of a protein-coding sequence.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=185#h5p-19

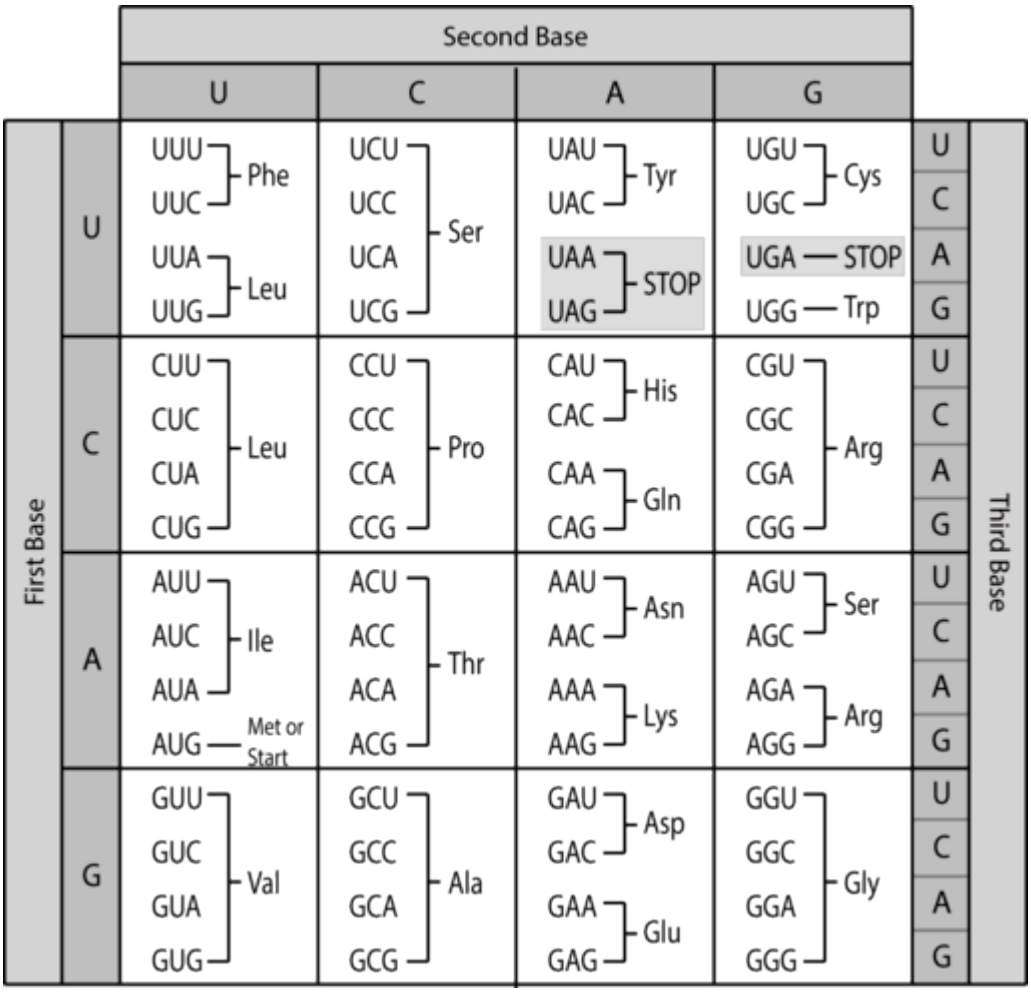


Figure 8. The genetic code in a condon table.

The code is non-overlapping

The genetic code is read three bases at a time, with each codon immediately adjacent to the next. There are no "spacer" bases, and codons do not overlap. This is illustrated in **Figure 9**. The protein is synthesized from N- to C- terminus. So the RNA sequence AAAUUUGGG might be read in non-overlapping codons of AAA-UUU-GGG, encoding the protein N-lys-phe-glu-C.

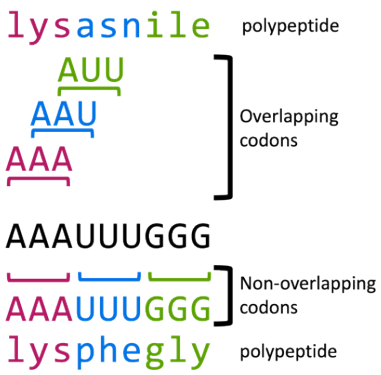


Figure 9. The genetic code is non-overlapping.

Why "might" be read? Because the codons are non-overlapping, it also means that for every RNA, there are three potential **reading frames**, depending on which base you start reading with. For the sequence of AAAUUUGGG shown in **Figure 9**, the three potential reading frames would be:

AAA-UUU-GGG (lys-phe-gly)
A-AAU-UUG-GG (-asn-leu)
AA-AUU-UGG-G (-lle-trp)

These reading frames encode very different polypeptides!

Note that an mRNA is *not* read beginning with the first amino acid. There is always an **untranslated region (UTR)** at the 5' end of the molecule. It is the start codon that establishes the reading frame for an RNA. If you do not know the reading frame for an RNA, then you cannot determine the sequence of the protein, even if you know the RNA sequence.

The amino acid-specifying codons are read in frame to specify the sequence of amino acids. The coding sequence ends with a stop codon, but the RNA continues past that point. Just as every RNA molecule has a **5' UTR**, every RNA also has a **3'UTR**.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=185#h5p-20

Media Attributions

- Genetic Code © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Non-overlapping © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

THE RIBOSOME IS THE TRANSLATION MACHINERY

The ribosome catalyzes peptide bond formation. The ribosome is a large complex assembled from many different components, including ribosomal RNA (rRNA) and protein. The rRNAs are functional RNAs that are never translated into protein.

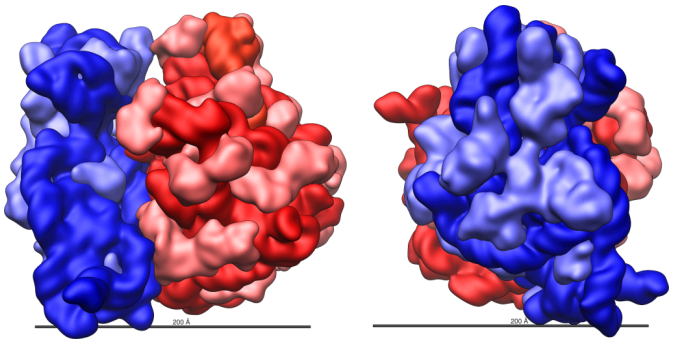


Figure 6. Structure of the E.coli ribosome seen from two different angles, with large 50S subunit in red and small 30S subunit shown in blue. RNA components are darker in color and protein components are lighter.

The ribosome has a large and small subunit. The ribosomal components are described in terms of Svedburg units (S). These are units that indirectly approximate size based on sedimentation during centrifugation. The units are not additive, because sedimentation depends on both the size and shape of a complex. So for example, in prokaryotes the large subunit is 50S, the small subunit is 30S, and the whole ribosome is 70S.

The structure of the *E.coli* ribosome is shown in **Figure 6**, with the large subunit colored red and the small subunit colored blue. Prokaryotic ribosomes have 3 ribosomal RNAs (rRNA): 23S rRNA and 5S rRNA in the large subunit, and 16S rRNA in the small subunit. They also have about 50 proteins.

Eukaryotic ribosomes also have a large and small subunit and are structurally very similar. In eukaryotes, the large subunit is 60S, the small subunit is 40S, and the whole ribosome is 80S. Eukaryotic ribosomes have 4 rRNAs and about 80 proteins. The 28S and 5.8S, and 5S rRNAs are part of the large subunit, and the 18S rRNA is part of the small subunit. It is the rRNA components that are catalytically active: they catalyze the peptidyl transferase reaction that builds the RNA molecule. The ribosome is therefore an example of a **ribozyme**, an RNA molecule that acts as an enzyme.

A comparison of the eukaryotic and prokaryotic ribosomes is summarized in **Table 1**.

Table 1 Comparison of eukaryotic and prokaryotic ribosomes

	Eukaryote Ribosome	**Prokaryote Ribosome**
Overall size	80S	70S
Large subunit	60S	50S
Large subunit components	28S rRNA 5.8S rRNA 5S rRNA About 50 proteins	23S rRNA 5S rRNA About 30 proteins
Small subunit	40S	30S
Small subunit	18S rRNA About 30 proteins	16S rRNA About 20 proteins

The large and small subunits assemble during translation with the mRNA sandwiched between them. The bases of the mRNA are being positioned within three adjacent sites in the ribosome, called the E, P, and A sites.

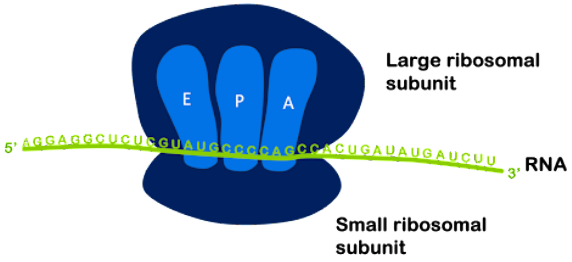


Figure 7. EPA sites of the ribosome.

A stands for **aminoacyl**: this is the acceptor site for aminoacyl tRNAs (those carrying an amino acid) to enter the ribosome. P stands for **peptidyl**: this is the site of the peptidyl transferase reaction that forms peptide bonds between a growing polypeptide and an incoming amino acid. E stands for **exit**: tRNAs exit from the ribosome via the E site after they have donated their amino acid to the peptide.

Media Attributions

- Ribosome shape © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- EPA sites © Amanda Simons is licensed under a CC0 (Creative Commons Zero) license

TRNAS ACT AS AN ADAPTOR BETWEEN MRNA AND AMINO ACID

The ribosome assembles a polypeptide to match the mRNA according to the genetic code. To "read" the mRNA requires the use of a **tRNA**. tRNAs act as an adaptor between the mRNA and an amino acid. tRNAs form complementary base pairs with the codons of an mRNA, and they carry with them a cognate amino acid.

All tRNAs fold into a characteristic structure. In two dimensions, they are often depicted as a cloverleaf structure, because the backbone of the RNA molecule folds into three **stem loop** structures as shown on the left in **Figure 10**. However, their three-dimensional structure is more L-shaped as shown on the right in **Figure 10**.

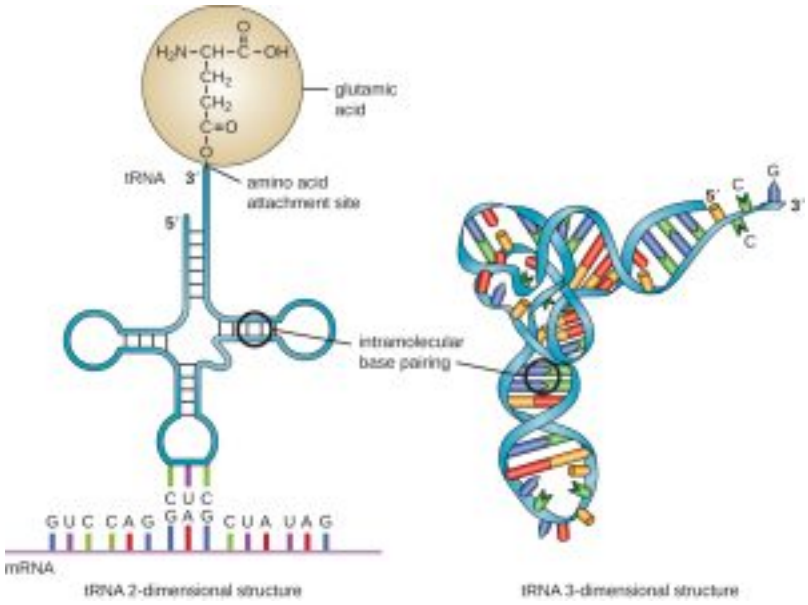


Figure 10. tRNA structure.

At the bottom of the structure shown in **Figure 10** are unpaired bases that make up the **anticodon.** Each tRNA has an anticodon that can form complementary base pairs with one of the 61 amino acid-specifying codons of the genetic code. In this figure, the tRNA anticodon is CUC, base pairing with the codon GAG, which codes for glutamic acid.

At the top of the structure is the acceptor stem: the region of the tRNA that links to its cognate amino acid. On the left in **Figure 10**, you can see the tRNA is depicted as "charged" with glutamic acid. The carboxylic

acid functional group of the amino acid is covalently linked with either the last 2'OH or 3' OH group of the tRNA. Once a tRNA has been charged with its matching amino acid, it is more properly called an aminoacyl-tRNA. Enzymes called **aminoacyl tRNA synthases** perform the charging reaction. Each aminoacyl tRNA synthase is specific for both the tRNA and the amino acid.

Nomenclature: Each amino acid has one or more designated tRNAs. The tRNAs are named for their amino acid, so the one shown in **Figure 10** is called $tRNA^{glu}$. Once it is charged with an amino acid, the amino acid linkage is indicated as well: glu-$tRNA^{glu}$ is glutamic acid tRNA that is carrying a glutamic acid. (Why the double listing of the glu? Under some conditions, tRNAs can be charged with other amino acids.)

The codon-anticodon pairing is antiparallel, just like all base-pairings. And, remember: the codons of a codon table are all written from 5' to 3', by convention. So, for example, a UGC codon (Cys) would be recognized by anti-codon 3'ACG5'.

tRNAs have some unusual properties. First, they include **non-canonical** (unusual) nucleotide residues like inosine and pseudouridine. Remember that bases and nucleoside/nucleotides are named differently; for example, the base adenine is a component of the nucleotide adenosine. Pseudouracil is the base in the nucleotide pseudouridine, and hypoxanthine is the base in the nucleotide inosine. Inosine/hypoxanthine is abbreviated I, and pseudouracil is abbreviated with the symbol psi: Ψ.

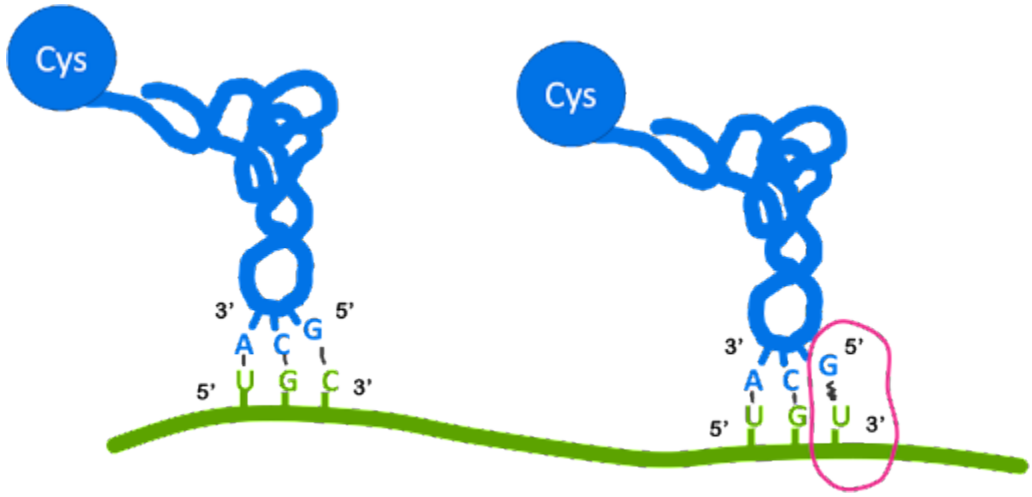


Figure 11. Wobble pairings at the 3' end of the codon can use noncanonical base pairs. This allows one tRNA to recognize multiple codons. Here, $tRNA^{cys}$ can recognize both cysteine codons UGC and UGU. These non-canonical bases are incorporated in tRNA through post-transcriptional modifications. The deamination of adenine produces hypoxanthine, and the isomerization of uracil produces pseudouracil.

Second, the codon-anticodon pairings do not have to be a perfect match to one another at the third position of the codon-anticodon pairing (3' codon/5' anticodon). At this position, noncanonical pairings called **wobble** base pairs may be used. This allows a single tRNA to recognize multiple codons. The example illustrated in

Figure 11 shows that a $tRNA^{cys}$ with anticodon 3'ACG5' can recognize the codon 5'UGC3' via canonical base pairing. It can also recognize the codon 5'UGU3' via wobble base pairing.

Only the 3' end of the codon, called the wobble position, can utilize wobble pairing. **Table 2** lists the base pairings observed at the wobble position. The noncanonical nucleotide inosine is used at the wobble position of some tRNAs, which allows the tRNA to pair with three different codons.

If you look carefully at the codon table shown in Figure 8, you'll see that the degeneracy of the code is not random. When two codons specify the same amino acid, the codons usually share the same base at the first two positions with either A/G or C/U at the third position. The codons "XXG" and "XXA" can both be recognized by a "XXU" anticodon. And codons "XXC" and "XXU" can both be recognized by a "XXG" anticodon.

Table 2 Wobble pairings at the 3′ end of codons

tRNA 5' end of the anticodon	mRNA 3' end of anticodon
A	U
C	G
G	C or U
U	A or G
I	U, C, or A

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=193#h5p-21

Media Attributions

- tRNA structure © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Wobble pairings © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

TRANSLATION IN PROKARYOTES

Like transcription, translation has three stages: initiation, elongation, and termination. In the initiation step, the ribosome must be recruited to the mRNA and positioned over the correct part of the RNA to begin translation in the correct reading frame. The 5' end of the RNA is not typically translated. During elongation, the ribosome slides along the mRNA from 5' to 3', synthesizing a polypeptide to match the codons. During termination, the growing polypeptide is released from the ribosome, and both the ribosome and the mRNA can be reused in another round of translation.

Initiation

In prokaryotes, initiation of translation can begin even before transcription has been completed: Once the 5' end of the RNA is free of the RNA polymerase, the ribosome can contact the RNA and begin translation as shown in **Figure 12**. An image of this process as visualized via electron microscope of this process can be seen at Scitable.com.

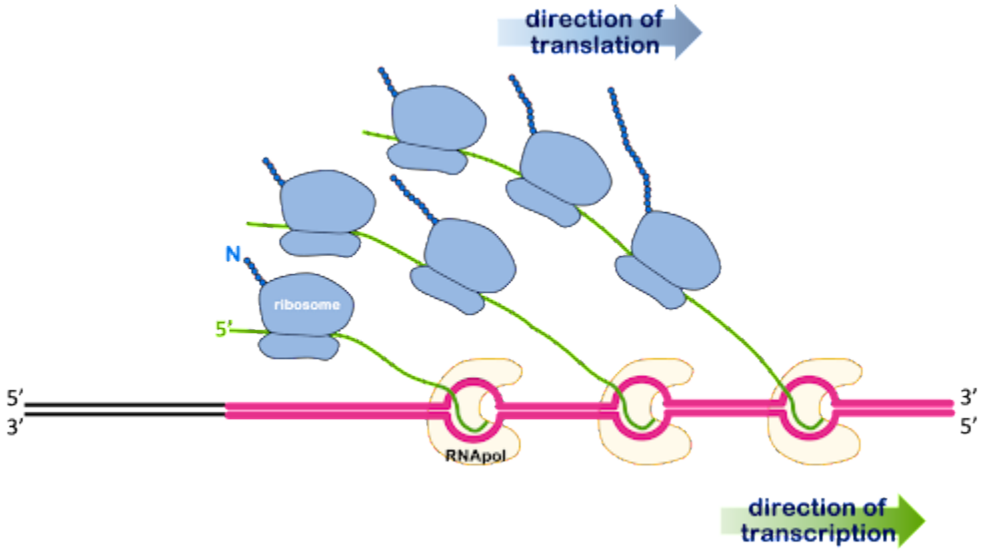


Figure 12. In prokaryotes, translation can begin even before the RNA polymerase has terminated transcription. Multiple ribosomes can transcribe in tandem along one RNA molecule.

Near the 5' end of prokaryotic mRNA molecules is a sequence called the Shine-Dalgarno site, or **ribosome binding site**. This has the consensus sequence of 5'AGGAGG3'. Complementary base pairing between the

Shine-Dalgarno site and the sequence 3'UCCUCC5' within the 16S rRNA of the small subunit is what brings together the mRNA and the ribosome. The Shine-Dalgarno site is positioned to bring the small subunit into proper alignment with the AUG start codon.

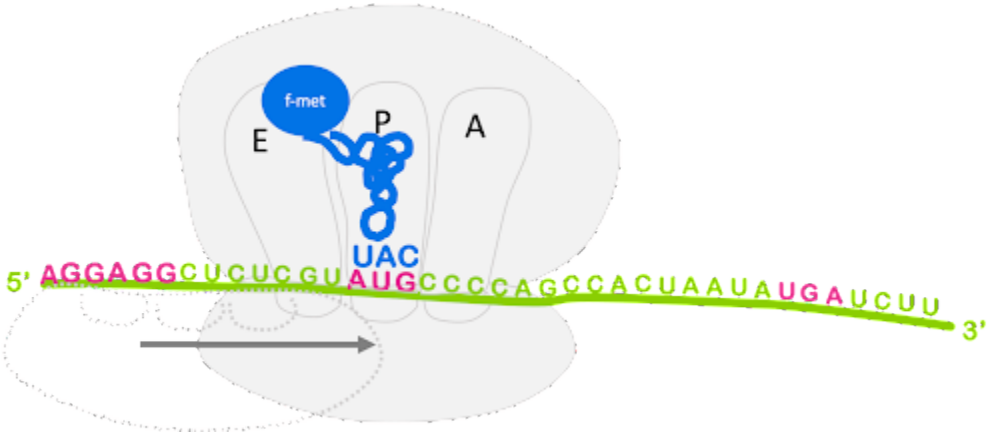


Figure 13. Translation initiation begins with the small ribosomal subunit binding to the ribosome binding site. This positions the small subunit over the start codon. The start codon is bound by an initiator tRNA charged with fMet. The large ribosomal subunit then binds, sandwiching the mRNA.

An initiator tRNA charged with f-Met binds to the start codon, and last the large subunit joins, sandwiching the mRNA and fMet-tRNA between the large and small subunit with the initiator tRNA positioned in the P site of the ribosome. This is shown in Figure 13.

Although not shown in **Figure 13**, the process of translation initiation requires translation initiation factors (IFs) that facilitate binding of the small subunit to the ribosome binding site, prevent premature association of large and small subunit, and help position fMet-tRNA^{fMet}. Translation initiation uses energy from the hydrolysis of GTP to power these steps[1].

Elongation

After the large subunit binds, the elongation stage begins. A charged tRNA with anticodon complementary to the next codon will enter the A site of the ribosome, escorted by elongation factor EF-Tu. Again, GTP hydrolysis provides the energy for this process. The ribosome catalyzes the formation of a peptide bond

1. Rodnina, M. V. Translation in Prokaryotes. Cold Spring Harb. Perspect. Biol. 10, a032664 (2018).

between the first fMet and the second amino acid. In this process, the bond between fMet and tRNA is broken, and fMet is linked with the second amino acid. The second tRNA becomes a peptidyl-tRNA, as it is now linked with a dipeptide instead of a single amino acid. This is shown in **Figure 14**.

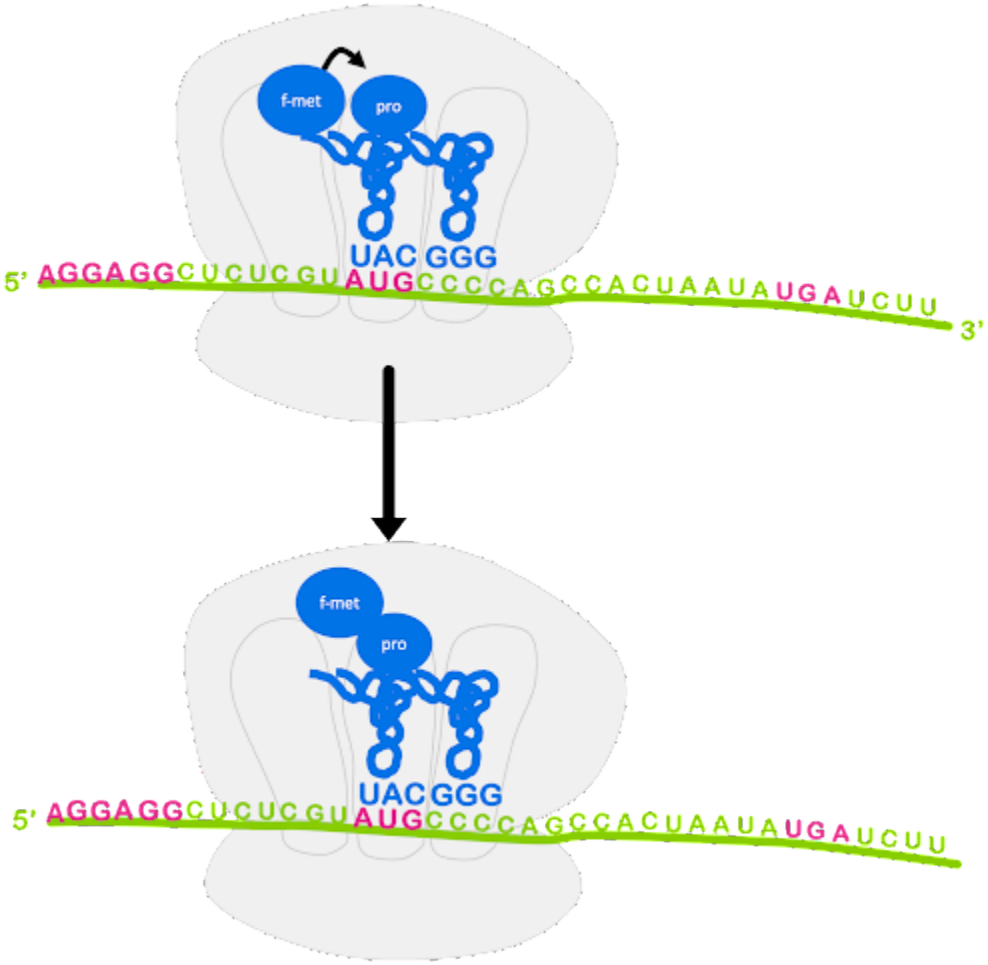


Figure 14. The Peptidyl transferase reaction transfers the fMet amino acid from the initiator tRNA to the amino acid on the second, incoming tRNA.

The ribosome translocates, or moves, along the mRNA so that the start codon is positioned in the E site and the original initiator tRNA can exit. The peptidyl-tRNA is in the P site, and the A site is open and can accept a new tRNA. This is shown in Figure 15. Although not shown, GTP hydrolysis by elongation factor EF-G is required for translocation[2].

2. ibid

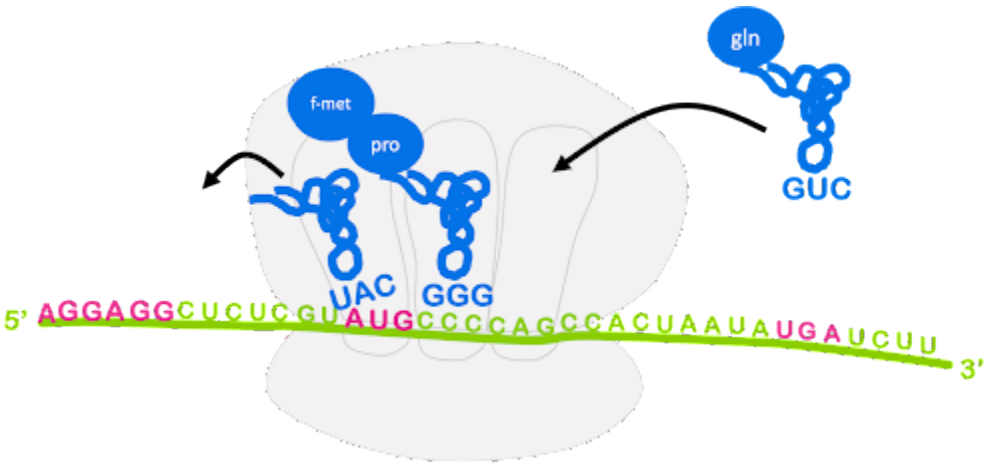


Figure 15. Translocation of the ribosome. The ribosome translocates along the mRNA from 5' to 3'. As a spent tRNA shifts into the A site, it can exit the ribosome. Incoming charged tRNAs enter via the E site.

Termination

The process of bringing in a new charged tRNA, catalyzing peptide bond formation, translocation, and release of the spent tRNA repeats until the entire coding sequence has been translated and the ribosome encounters a stop codon in the A site. There are no tRNAs that recognize the stop codons.

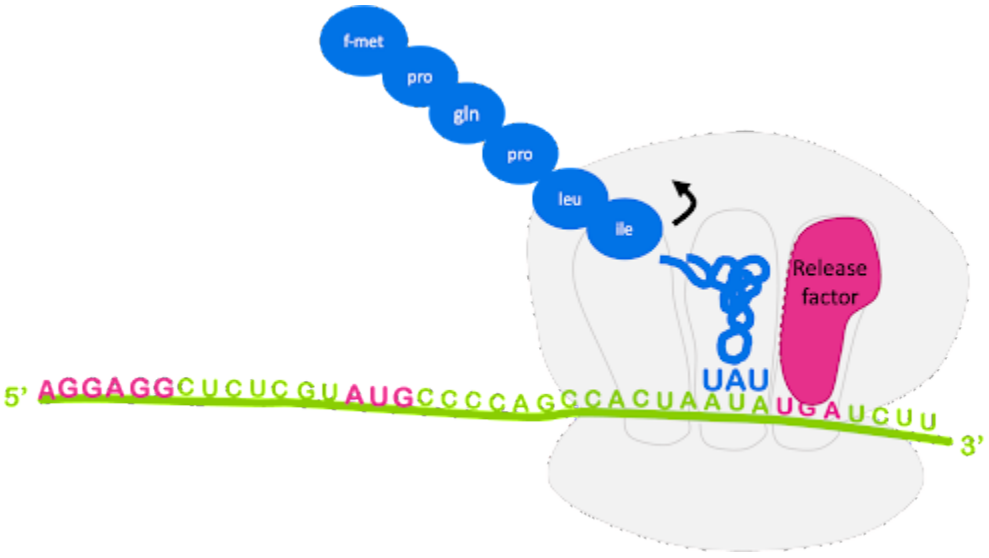


Figure 16. Termination of translation requires release factors. When a ribosome encounters a stop codon in the A site, there is no corresponding tRNA. Instead, a release factor binds the A site and catalyzes the release of the polypeptide and the disassembly of the ribosome.

Instead, release factors (RFs) recognize the stop codons in the A site. Release factor binding to the A site

promotes the release of the polypeptide from the peptidyl tRNA in the P site. Ribosome recycling factor (RRF) prompts the separation of mRNA, small subunit, and large subunit and recycling of the ribosome (**Figure 16**). Like initiation and elongation, termination is coupled to GTP hydrolysis[3].

The whole process is shown in the animated gif in **Figure 17**.

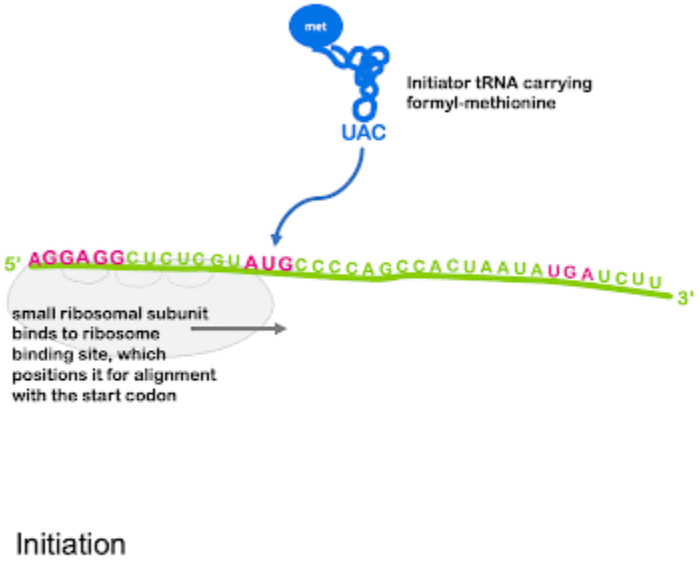


Figure 17. Transcription initiation, elongation, and termination.

Media Attributions

- Prokaryotes © Amanda Simons is licensed under a CC0 (Creative Commons Zero) license
- Small ribosomal subunit © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Peptidyl transferase reaction © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Translocation of the ribosome © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Termination of translation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike)

3. ibid

license

- Transcription animation

TRANSLATION IN EUKARYOTES VS PROKARYOTES

Translation in eukaryotes is very similar to that in prokaryotes. The biggest differences are seen in initiation. Prokaryotes have no internal membrane-bound organelles and no nucleus, so both transcription and translation happen in the same compartment. This allows for the simultaneous transcription and translation illustrated in **Figure 10**. But in eukaryotes, transcription happens in the nucleus, and translation happens on ribosomes in the cytoplasm – transcription and translation, therefore, can't both act on the same mRNA at the same time.

Eukaryotes also do not have a Shine-Dalgarno sequence. Instead, the 5' cap serves as a ribosome binding site, and the ribosome scans toward the 3' end of the mRNA until a start codon is encountered. fMet is not used as an initiator amino acid; all eukaryotic polypeptides begin with methionine[1]. However, there is a dedicated initiator $tRNA_i^{Met}$ that is distinct from the methionine tRNA used during elongation[2].

The prokaryotic initiation factors, elongation factors, and release factors all have eukaryotic homologs that make prokaryotic and eukaryotic translation very similar mechanistically.

1. Hinnebusch, A. G. & Lorsch, J. R. The Mechanism of Eukaryotic Translation Initiation: New Insights and Challenges. Cold Spring Harb. Perspect. Biol. 4, a011544 (2012).

2. Kolitz, S. E. & Lorsch, J. R. Eukaryotic Initiator tRNA: Finely Tuned and Ready for Action. FEBS Lett. 584, 396–404 (2010).

GENE STRUCTURE

The codons make up the **coding sequence** of a gene, but it's important to remember that the coding sequence is not the only important part of the gene. All the regulatory elements we've discussed in the transcription and translation modules play a role in the function of the gene (**Figure 18**).

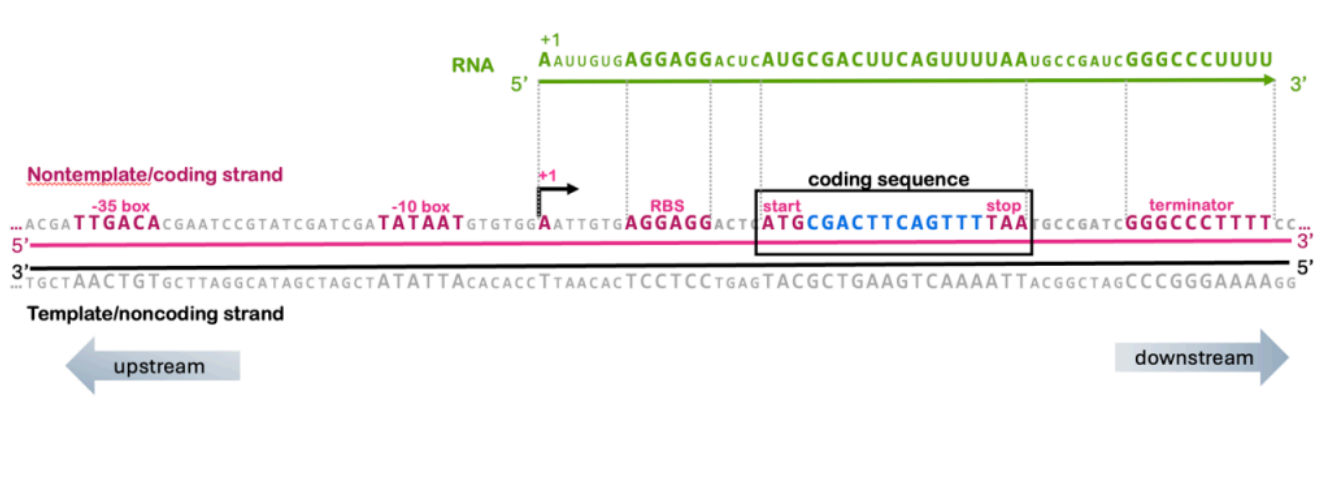


Figure 18. Diagram of a prokaryotic gene with elements required for transcription and translation.

You'll remember from the unit on transcription that the RNA molecule has the same sequence as the nontemplate strand of the gene. The nontemplate strand is also called the **coding strand** because it contains the **coding sequence** of the gene, and the *template* strand is also called the *noncoding* strand.

The translation control elements discussed in this module are also therefore found in the coding strand of the DNA: the ribosome binding site (AGGAGG), the start codon (ATG), and the stop codons (TAA, TAG, TGA) are all recognizable elements in the coding strand of a gene, as shown in **Figure 18**, you can even find codon tables that use DNA codons rather than RNA codons.

The coding sequence is also called an **open reading frame (ORF)** because it is a long stretch of codons that lack a reading frame. For simplicity, the ORF in **Figure 18** is only 6 codons long, but most ORFs are hundreds or thousands of codons.

A note on terminology: What's the difference between the ORF and the coding sequence? The term ORF is used when searching genomic DNA sequences for potential "new" genes. Because 3/64 codons are stop codons, if reading random, non-coding genomic sequences, you'd expect to find a codon about every 20 codons. If a long stretch of DNA – say, hundreds or thousands of codons – does not have a stop, that is called an "open" reading frame, and it is potentially part of a protein-coding gene. All coding sequences are ORFs, but not all ORFs end up being genes.

With this understanding of the relationship between coding strand, template strand, and RNA, we can therefore build a more complete understanding of the structure of a gene. In **Figure 18**, you can see that the RNA includes the sequence from the +1 to the end of the terminator, and all of the translation control elements (ribosome binding, start, and stop) are also within those boundaries.

Figure 19 shows the structure of a typical eukaryotic gene[1], reading from 5' to 3' on the coding strand. The transcribed region is shown between the two light blue regions, starting with the +1 site (labeled) and ending at the pA cleavage site (not labeled). The promoter sequences are shown upstream of the +1 site, with TATA box, CAAT box, and GC box drawn for this gene. The yellow boxes are enhancers, and you can see multiple enhancers upstream, downstream, and within the gene, as is typical for enhancers.

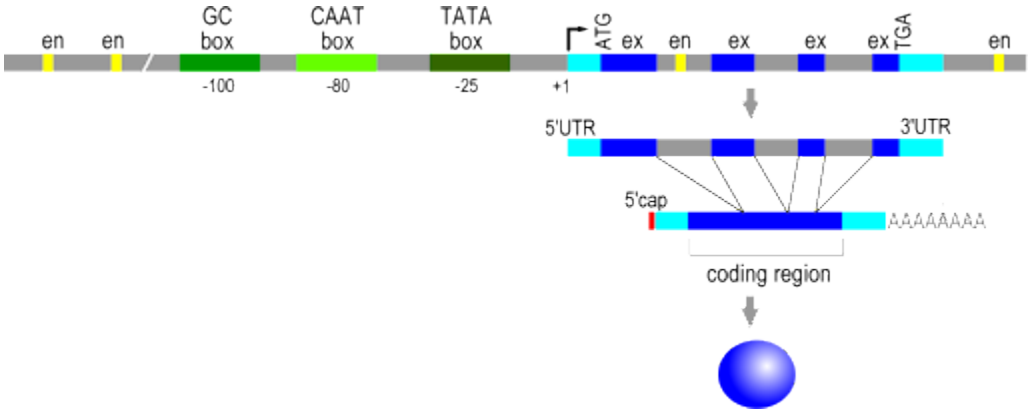


Figure 19. Structure of a typical eukaryotic gene. Shown are enhancers (en, yellow), GC box, CAAT box, and TATA box, the 5' and 3' UTRs (turquoise), and exons (ex, dark blue). The pre-mRNA includes the introns, but in the mature mRNA, they have been spliced out.

For both the prokaryotic and eukaryotic genes, notice that the control elements for *translation* are all within the boundaries of the *transcribed* region, that is, between the +1 site and the polyadenylated sequence, or between the +1 site and the terminator. This makes sense! A codon must be part of the RNA to be useful, so the start codon and stop codons must be between the +1 and the polyA site.

ORFs vs coding sequence revisited: In prokaryotes, the coding sequence is continuous, so there is one ORF for the gene. However, in eukaryotes, the coding sequence is discontinuous in the DNA sequence. The ORFs do not extend through the introns, from exon to exon, because the intron sequence will contain stop codons in frame with the coding sequence. Eukaryotic genes, therefore, include multiple ORFs.

1. Locke, J. 'Open Genetics Lectures' textbook for an Introduction to Molecular Genetics and Heredity (BIOL 207). (2017) doi:10.7939/DVN/XMUPO6.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=207#h5p-22

Media Attributions

- Diagram of a prokaryotic gene is licensed under a CC BY-SA (Attribution ShareAlike) license
- Structure of a typical eukaryotic gene © Open Genetics Lectures (OGL) Fall 2015 is licensed under a CC0 (Creative Commons Zero) license

SUMMARY

- Translation is the process of protein synthesis.
- Proteins, or polypeptides, are assembled from amino acid monomers. There are 20 amino acids commonly used for protein synthesis in biological systems. Amino acids are connected via peptide bonds during protein synthesis. Codons in an mRNA specify the order of amino acids in a protein.
- The genetic code is universal, degenerate, and nonoverlapping.
- The ribosome is the molecular machine that catalyzes the peptide bond formation. It is built from a large and small subunit, each of which is built from multiple RNA and protein components.
- tRNA acts as an adaptor, binding to mRNA via codon/anticodon interactions and escorting amino acids to the ribosome. tRNAs contain modified bases and can participate in wobble pairing with codons.
- Translation occurs in three stages: initiation, elongation, and termination. During translation initiation, the small subunit of the ribosome interacts with the ribosome binding site (or 5' cap). The P site is aligned with the start codon, where an initiator codon binds with a cognate Methionine (or fMet in prokaryotes). During elongation, the ribosome catalyzes peptide bond formation. During termination, release factors release the polypeptide from the ribosome and recycle the ribosome and mRNA.

WRAP-UP QUESTIONS

1. **Figure 4** (reproduced below) shows the folding of a hypothetical protein, with noncovalent bonds holding the folds in place. An ionic bond between lysine and aspartate is shown.

Mutations can change the amino acid composition of a protein. What do you think might be the consequence of folding of a mutation that changed the negatively charged aspartate to a positively charged arginine?

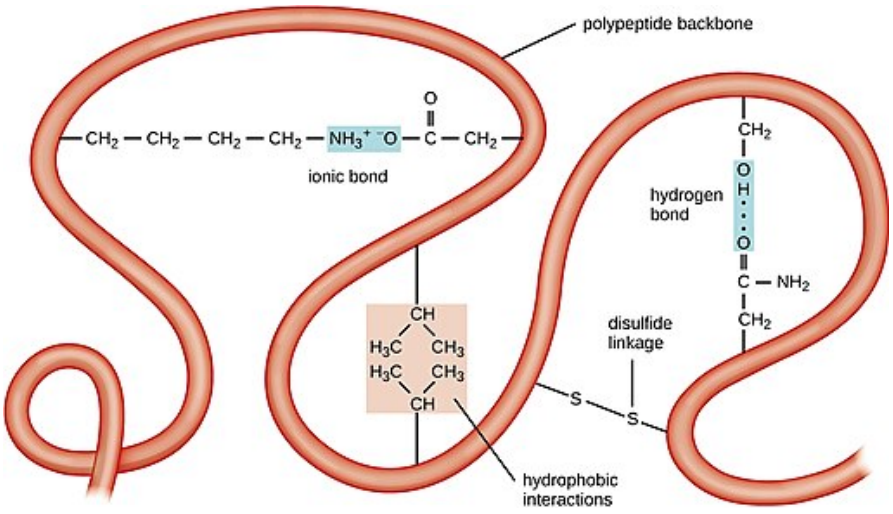


Figure 4. Types of bonds in protein folding. The backbone of a polypeptide is shown in orange. Selected amino acid side chains are shown, extending from the backbone to participate in protein folding.

2. Early on, molecular biologists hypothesized that, with 20 amino acids found in proteins and only four bases found in nucleic acids, the genetic code must use triplet codons. Triplet codons give 64 possible sequence combinations, which is more than sufficient to encode 20 amino acids.

a. What is the maximum number of amino acids that could be specified by a two-base code? What is the maximum number of amino acids that could be specified by a four-base code?
b. The process of transcription is energetically demanding for the cell. Explain why it would not be advantageous for cells to use a four-base code to specify 20 amino acids.

3. **Figure 8** gives one version of a codon table, but other ways of presenting the information exist. Do a web search for "codon table" images. Compare 3-4 different versions of a codon table. What are some similarities

and differences with how the information is presented? Which version would you personally find easiest to use?

4. There are 6 arginine codons, as shown in **Figure 8**. What is the minimum number of anti-codons necessary to recognize all six codons? What are their sequences?

5. **Figure 16** shows the structure of a typical eukaryotic gene. Draw a similar diagram for a prokaryotic gene. In your diagram, include the -10 and -35 boxes, the +1 site, the terminator, the ribosome binding site, the start codon, and the stop codon. Which of these sequences is important for transcription? Which of these sequences is important for translation?

6. A simplified prokaryotic gene sequence is diagrammed below, with the -35, -10, +1, ribosome binding site, and terminator sequences underlined.

5'aaTTGACAactgtgtgatgtagcttaTATAATatgatgcta**C**gttaaAGGAGGtgggtaatgcctaaatggcggtaacgtgaGGCCGCGGCCTTTTcccc3'

a. Is this the coding or noncoding strand of the DNA?
b. Circle the start and stop codons for this gene.
c. What is the sequence of the polypeptide that is produced from this gene?

7. The genes encoding tRNA can be mutated, just like any other gene. In some rare cases, the tRNA for Leucine has its anticodon mutated so that the AAU anticodon sequence is changed to AUU. What codon(s) will the new, mutated tRNA recognize? What do you think the consequence of the mutation might be to the process of translation?

Media Attributions

PART VII
MUTATION

Objectives

1. Describe mechanisms by which mutations arise.
2. Classify mutations based on the affected cells, the change to DNA, the effect on the coding sequence, the effect on the protein, and the effect on the cell or organism.
3. Describe the difference between gain-of-function and loss-of-function mutation
4. Distinguish between germline and somatic mutation
5. Distinguish between indel and base substitution mutations
6. Classify base substitution mutations as missense, nonsense, silent, or none of the above
7. Recognize that not all indel mutations are frameshift mutations
8. Recognize that not all mutations occur in the coding sequence of a gene, and that mutations in noncoding sequence may also affect phenotype.
9. Recognize that not all mutations are bad.
10. Predict the effect of a mutation on the function of a protein.
11. Explain the connection between DNA damage and mutation.
12. Describe the mechanisms by which DNA damage leads to mutation, including replication errors and damage to bases.

Source material

- Selected images and text remixed from *Online Open Genetics* (Nickle and Barrette-Ng), available through Biology LibreTexts[1]. Chapter 4: Mutation and Variation
- Selected images and text remixed from Open Genetics Lectures[2]. Chapter 11: Mutation Causes and 12: Mutation Consequences.

Introduction

Within any species, you will see normal phenotypic variation. For example, tomatoes may produce fruits of different size, shape, and color. Dogs can weigh five pounds, one hundred and fifty pounds, or any weight in between. Humans can have hair color ranging from blond to red to black. Those differences in phenotype are caused by differences in DNA sequence among individuals in the population.

Previous modules looked at how DNA acts as an information storage system and how that information is used to build RNA and proteins that perform many cellular functions. The structure and function of those RNA and protein molecules is specified by the sequence of the DNA, so any changes to the DNA sequence can affect the structure and function of those gene products. Changes to the function of gene products can impact the behavior of a cell or even a whole organism.

Changes to DNA sequence are called **mutations**. In general (non-science) usage, the word can have very negative connotations (when it is not associated with superheroes, of course!). But to a geneticist, the word mutation just means a change. Mutations may be harmful to an organism, beneficial to an organism, or neutral. Mutation also drives evolution: It is the source of variation in a population, and without variation, there is no evolution. (More on this in other modules!)

Some terminology regarding variations in DNA: A mutation is a change in DNA sequence from a reference organism. For model organisms in the lab, mutations are in relation to the most common, "wild-type" phenotype. But when talking about populations of organisms outside of the lab, including humans, it can be difficult to decide what is "normal". For example, some people have blood type A, some have type B, some have type AB, and some have type O. None is more normal than another.

In these cases, the term **polymorphism** is used to describe the variation we see in a population. If an allele is seen in more than 1% of the population and is not associated with disease, we'd use the word **polymorphism**. If fewer than 1% of the population has a variation or if the variant is associated with disease, the word **mutation** would describe the sequence[3].

When we are describing these unusual variations in genetics, it can be commonplace for geneticists who study model organisms to describe an individual as mutant – as in "a mutant fruit fly with white eyes". But this

---

1. Nickle and Barrette-Ng. Open Online Genetics. in *Open Online Genetics* (2016).

2. Locke, J. 'Open Genetics Lectures' textbook for an Introduction to Molecular Genetics and Heredity (BIOL207). Borealis https://doi.org/10.7939/DVN/XMUPO6 (2017).

3. Karki, R., Pandya, D., Elston, R. C. & Ferlini, C. Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Med. Genomics* **8**, 37 (2015).

is not appropriate in human genetics, where the words "mutation" and "mutant" are limited to a description of gene or protein sequence, not people.

In this module and in others throughout this text, we see examples of mutations in human genes linked with genetic disorders. We can learn a lot about human genetics from studying rare differences in the population – and what we learn can help patients with a genetic disorder as well as further our understanding of human biology. But as you read, please keep in mind that the examples and images that we discuss are of real people who are deserving of respect. People who have consented to contribute their stories to education and research also deserve our gratitude. We should also remember that those differences do not indicate social worth, even when a genetic difference may cause health disparities.

MUTATIONS RESULT WHEN THE GENOME IS NOT PASSED PERFECTLY INTACT TO OFFSPRING

In order to pass on genetic information, a parent cell copies its DNA and divides the copies among daughter cells in meiosis or mitosis. The process must be completed with fidelity, ensuring that the genomic sequence of daughter cells is the same as the parent. But in practice, errors occur: there is about one mutation with every round of cell division, although this varies by both species and cell type[12]. This mutation in cell division is illustrated in **Figure 1**.

One mutation per cell division doesn't sound like a lot! But to grow a mature human body requires about 3.5×10^{13} (35 trillion) cells. We could therefore hypothesize that it takes about 45 cycles of cell divisions to get from a one-celled zygote to a mature human body, assuming every daughter cell survives and divides again. That means that any individual cell in your body may have about 45 mutations that are different from the fertilized egg! But those mutations are not necessarily the same 45 mutations as in other cells of the body. In **Figure 1**, even after just two rounds of cell division, the four daughter cells have different combinations of mutations.

To note: this also means that any egg or sperm that your body produces also has mutations compared to your original zygotic genome. These would give rise to *de novo* mutations in offspring: mutations that were not seen in previous generations.

1. Werner, B. et al. Measuring single cell divisions in human tissues from multi-region sequencing data. *Nat. Commun.* **11**, 1035 (2020).

2. Lynch, M. Evolution of the mutation rate. *Trends Genet. TIG* **26**, 345–352 (2010).

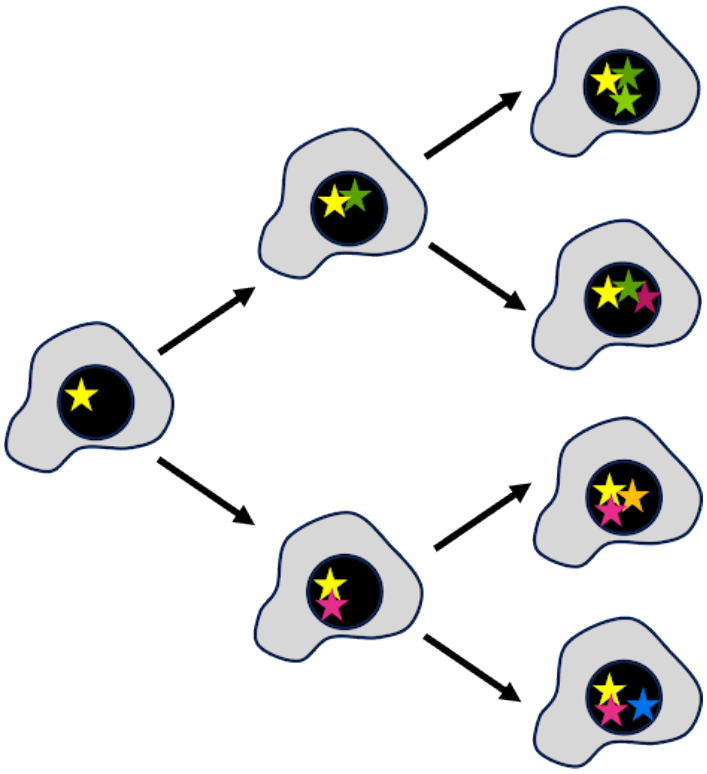


Figure 1. Mutations accumulate with every cell division. As cells divide, mutations (represented by stars) accumulate in daughter cells. Cells gain about one additional mutation per cell division, although the exact number varies among cell types.

By some estimates, over the course of a human lifetime, there are cumulatively about 10^{16} cell divisions[3]. This means that individual cells in an individual can be different from one another, with as many as 10^{16} differences collectively found in the human body! But most of these differences that accumulate over the course of a lifetime do not dramatically impact phenotype.

This module is divided into two parts. Part I looks at types of mutations and the effects of mutations. Part II looks at the mechanisms by which replication errors and DNA damage lead to mutation.

A preview of how mutations can be classified is listed below in Table 1.

3. Number of cell divisions in an average human – Human Homo sapiens – BNID 100379. https://bionumbers.hms.harvard.edu/bionumber.aspx?id=100379.

Table 1 Mutation Classifications

Ways to classify	Examples
Type of cell affected	Germline Somatic
Change to DNA Sequence	Base substitution (transition or transversion) Insertion Deletion Chromosomal rearrangement
Change to gene function	Neutral Gain of function Loss of function
Effect on phenotype	Dominant Recessive
Change to protein coding sequence	Silent Missense Nonsense Frameshift
Effect on other mutations	Intragenic suppressor Intergenic suppressor

Media Attributions

- Mutations accumulate © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

PART I: TYPES OF MUTATIONS

Geneticists use different terms to describe mutations, depending on context. In this section, we will look at ways to define mutations by the type of cell affected, the change made to DNA, and the effect on phenotype. If the mutation affects a gene, the mutation may also be described by the effect on gene function. If the mutation occurs within the protein-coding sequence of a gene, the mutation can be described by its effect on protein structure.

Type of Cell: Germline vs somatic mutations

All mutations can be described by the type of cell that is mutated. Multicellular organisms that reproduce sexually use meiosis to produce haploid cells (egg and sperm) that come together to form a diploid zygote. That zygote undergoes mitosis, and successive cell divisions accumulate the cells that make up the embryo and, eventually, the full-grown organism.

If a cell in the embryo or full-grown organism sustains a mutation, that mutation can be passed along to all other cells that arise from mitosis of that cell. This is called a somatic mutation because it happens in somatic (nonreproductive) cells. A somatic mutation can result in just one or two cells that harbor the mutation, but if the mutant cell undergoes frequent mitosis, it might create a patchwork of mutant cells. This is shown on the left in **Figure 2**, where the mutation occurs in the embryo. A somatic mutation cannot be passed to offspring because the germ cells are not mutated.

If a mutation occurs in a germ cell (egg or sperm), the mutation will become part of the genome in the zygote. As the zygote undergoes mitosis, the mutation will be passed along to all daughter cells. Every cell in the mature organism will have the mutation, including its germ cells. So the mutation is called a **germline** mutation, and the mutation can be passed to the next generation of offspring. This is shown on the right in **Figure 2**. Germline mutations can also arise if a mutation occurs in the egg, the zygote, or in an early enough stage of the embryo that the reproductive cells of the organism contain the mutation.

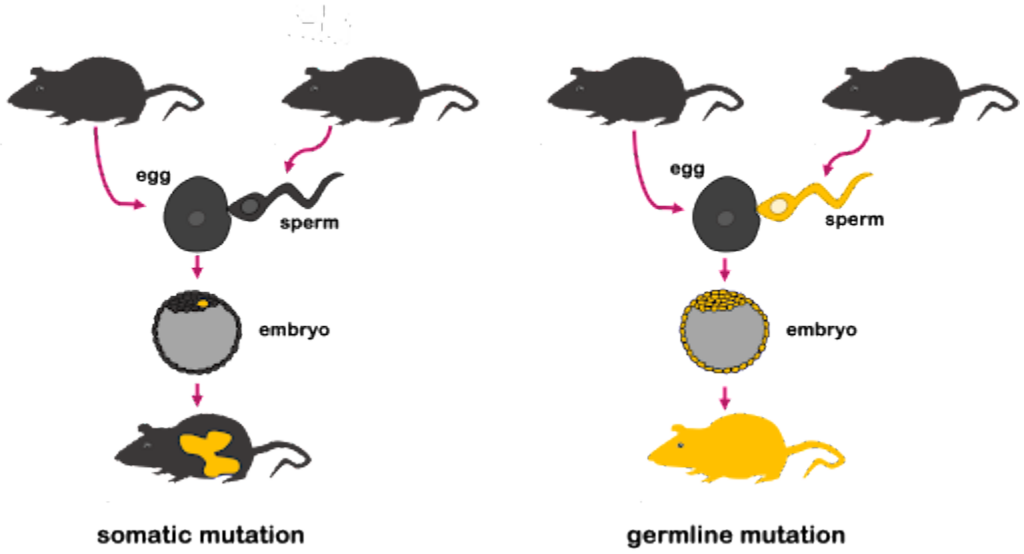


Figure 2 Germline vs somatic mutation. A somatic mutation is shown on the left, with a mutation (orange) occurring in one cell in the embryo. That embryo grows into a mouse with both mutant and wild-type cells. A germline mutation is shown on the right, with a mutation (orange) occurring in the sperm cell. Every cell in the embryo and full-grown mouse has the same mutation. A germline mutation can be passed to offspring, but a somatic mutation cannot.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=231#h5p-23

Changes to DNA: Insertions, deletions, base substitutions, and large-scale rearrangements

All mutations can be described by what happens to the DNA: additional bases can be added (an **insertion**), removed (a **deletion**), or one or more bases can be changed for different bases (a **base substitution**). These are illustrated in **Figure 3.**

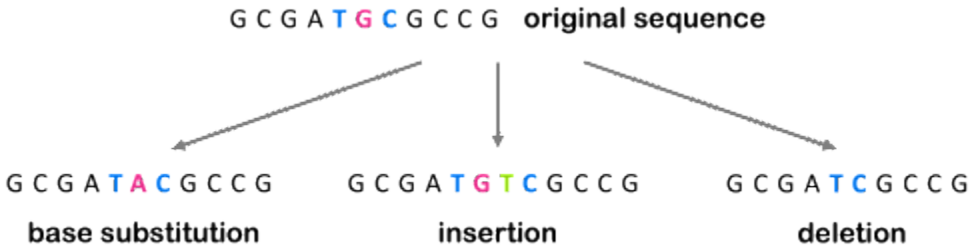


Figure 3. Mutations defined by change to DNA.

All three of the examples shown in **Figure 3** are described as **point mutations** because they affect one point within the genome. But we also see mutations that affect a larger portion of the genome, called **structural variants**. These larger-scale chromosomal rearrangements can include the gain or loss of entire chromosomes, which is a change in chromosome number called **aneuploidy**. On the other hand, part of a chromosome can be lost, part of a chromosome can be duplicated, or part of a chromosome can be **translocated** to another. Examples are illustrated in **Figure 4**. Although not shown in **Figure 4**, **inversions** are segments of the chromosome that have been reversed or flipped in orientation within the rest of the sequence.

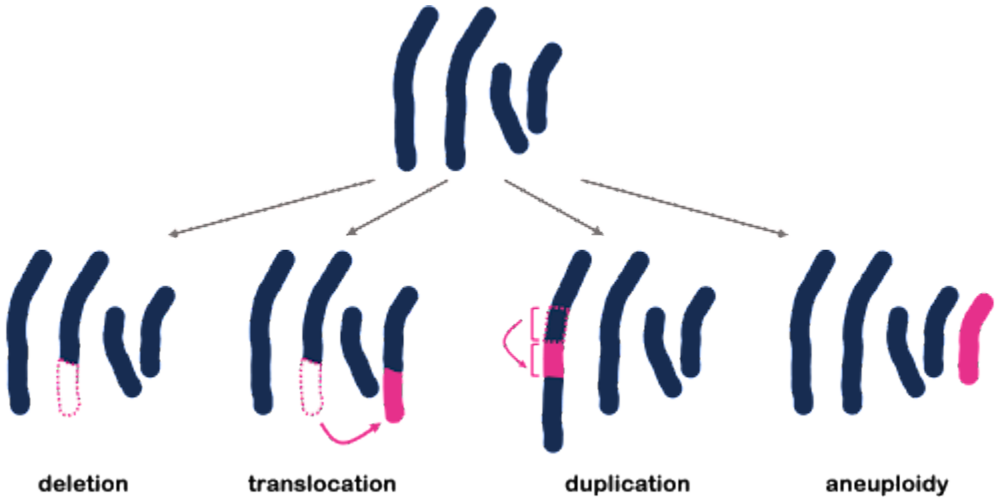


Figure 4 Large-scale chromosomal rearrangements affect large parts of the genome. Examples of large-scale rearrangements include deletion of part of a chromosome, translocation of part of a chromosome to another, duplication of a segment of a chromosome, and aneuploidy, which is the gain or loss of a chromosome.

Changes to protein sequence: Frameshift, missense, nonsense, and silent mutations

Most mutations are not found within the protein-coding sequence. After all, in humans, only about 1% of the

genome is protein-coding sequence! But if a point mutation occurs in the coding sequence for a protein, we can also describe the mutation with respect to how it changes protein sequence.

Base substitution mutations fall into one of three categories, illustrated in **Figure 5**.

- A **silent** mutation changes the DNA sequence but does not change the protein sequence. This is possible because most amino acids are specified by more than one codon. It is also called a **synonymous** mutation.
- A **nonsense** mutation changes an amino acid-specifying codon to a stop codon. This creates a protein that is shorter than normal because translation can't continue past the stop codon.
- A **missense** mutation changes an amino acid for a different amino acid. Missense mutations can be conservative (changing an amino acid for one that is chemically similar) or nonconservative (like exchanging an acidic base for a basic base or a polar base for a nonpolar one).

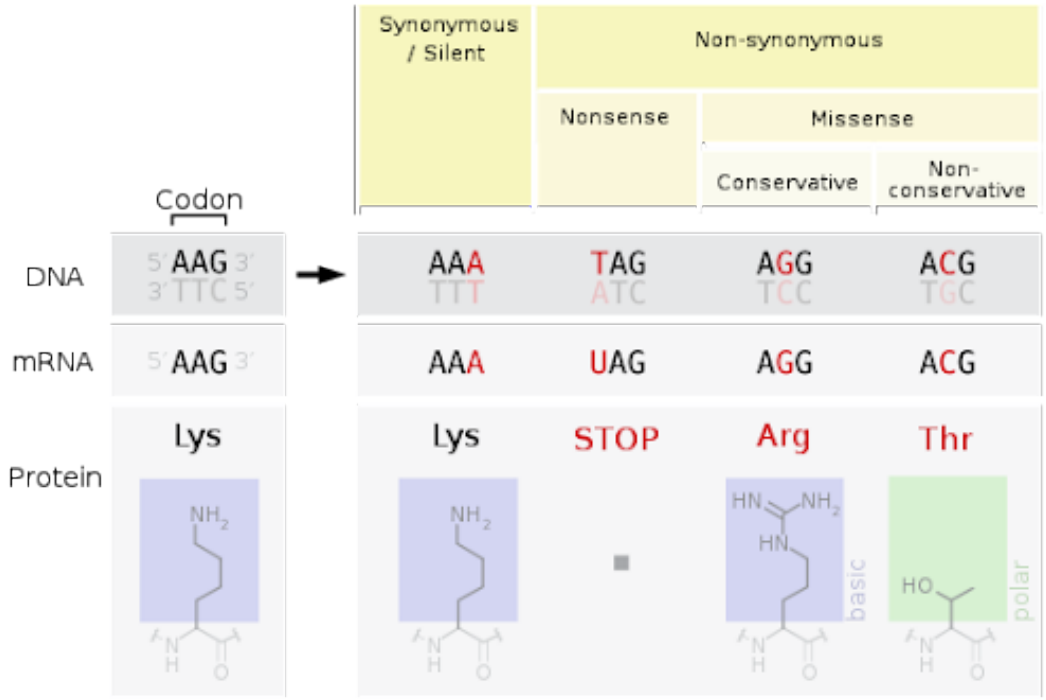


Figure 5. Base substitution mutations that happen in the coding sequence of the gene can be further characterized by their effect on protein sequence. On the left, an AAG codon is shown, with the noncoding strand shown underneath. This codes for lysine. If the base change creates another lysine codon, that is a silent mutation. A nonsense mutation creates a stop codon from an amino-acid specifying codon, and a missense mutation changes one amino acid for another. Missense mutations can be conservative – changing an amino acid for a similar one – or non-conservative, changing an amino acid for a chemically different one, as in this image where a basic side chain is changed for a polar side chain. Base substitution mutations that happen in the coding sequence of the gene can be further characterized by their effect on protein sequence. On the left, an AAG codon is shown, with the noncoding strand shown underneath. This codes for lysine. If the base change creates another lysine codon, that is a silent mutation. A nonsense mutation creates a stop codon from an amino-acid specifying codon, and a missense mutation changes one amino acid for another. Missense mutations can be conservative – changing an amino acid for a similar one – or non-conservative, changing an amino acid for a chemically different one, as in this image where a basic side chain is changed for a polar side chain.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it

online here:
https://rotel.pressbooks.pub/genetics/?p=231#h5p-24

Insertion or deletion mutations may simply add or subtract amino acids from the protein, but only if the insertion is a multiple of three nucleotides. If the number of inserted or deleted nucleotides is not divisible by three, this will result in a **frameshift.** Remember that since the coding sequence is translated in codons of three bases, each DNA strand has three potential reading frames. Inserting or deleting one, two, four, or any number of bases that are not a multiple of three will cause the ribosome to translate the wrong frame downstream of the mutation. Every amino acid after the insertion or deletion point will be different. This is shown in **Figure 6**.

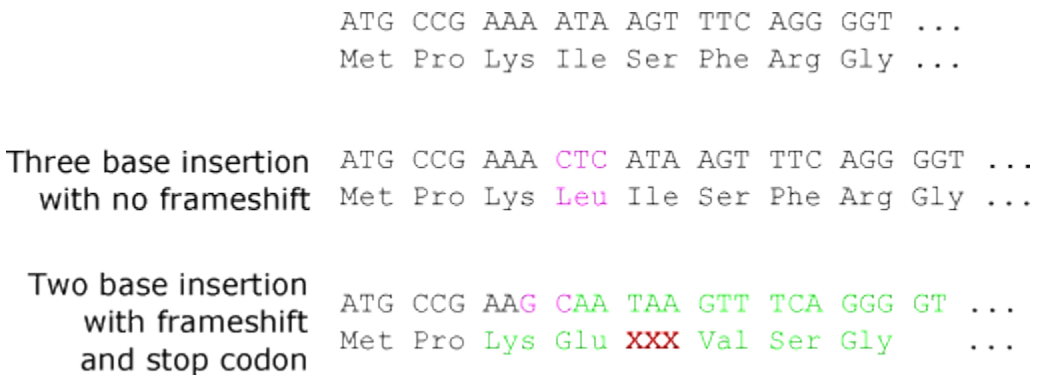


Figure 6. A frameshift mutation results if an insertion or deletion does not include a multiple of three nucleotides. The reference sequence at the top is modified by the insertion of three nucleotides (middle). This adds an amino acid to the middle of the protein (leucine, in purple), but it does not affect the rest of the amino acid sequence. On the bottom, an insertion of two bases disrupts the reading frame, affecting not only the amino acid at the insertion point but all amino acids that follow. Insertion or deletion of any number of nucleotides that is not divisible by three will result in a frameshift.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=231#h5p-25

Changes to gene function: Gain of function vs loss of function mutations

The terms silent, nonsense, missense, and frameshift only apply to mutations in the coding sequence of a gene. But mutations in DNA can occur anywhere in the genome – they can occur within genes, they can occur between genes, they can occur in coding sequences or in regulatory elements or in other noncoding regions. In fact, most mutations occur in non-coding sequences simply because the coding sequence is such a very small fraction of the genome. As a result, most mutations don't have much effect – we call them **neutral** mutations.

Even mutations within a coding sequence can be neutral: Silent mutations and conservative missense mutations are often neutral mutations.

But sometimes mutations can affect the function of a gene – either because the coding sequence is altered or because an important regulatory element is affected. For example, a mutation might delete the promoter or change an intron sequence in a way that prevents proper splicing. In all of those cases, we describe the mutation in terms of how the mutation affects the function of a gene.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=231#h5p-26

A **gain of function** mutation is a mutation that causes a gene to do something extra. For example, in Figure 4, duplications that result in an extra copy of a gene are often gain of function mutations because an extra copy of a gene means more protein is produced. A mutation that prevents transcription from being turned off would be another example of a gain of function mutation. Gain of function mutations can also affect protein sequence: for example, if an inhibitory domain of a protein is altered, the whole protein could be active even if it is meant to be inactive. We see examples of this in the Gene Expression chapters of this text.

You might think that doing something extra is good! And some gain of function mutations may be beneficial to an organism. But not necessarily. Think of an analogy to a gas stove (**Figure 7**). If one of the burners makes an extra big flame and cannot be turned off, that could start a kitchen fire! As an example in human biology: gain of function mutations in certain proteins that trigger cell division are associated with cancer because the mutation causes cells to divide uncontrollably. (The cancer-causing gain of function mutations are somatic mutations, not germline.)

Figure 7. Two gas flame burners on a stove.

A **loss of function** mutation is a mutation that lessens the activity of a gene. This could be a complete loss of function – say if no RNA or protein is produced. If no functional protein is produced, it is also called a **null** mutation – "null" is a word that simply means "zero". The chromosomal deletion in Figure 4 would cause a null mutation of all deleted genes, as would any mutation that destroyed the function of a gene's promoter. To go back to the stove analogy in **Figure 7**, this would be like a burner that can't turn on.

Some loss of function mutations reduce but don't completely block the action of the gene. These partial loss of function mutations can still affect the phenotype of an organism.

For example, in humans, the disease cystic fibrosis is caused by a loss of function in the gene CFTR, which is important for Cl^- transport across the cell membrane. Without functional CFTR protein, salt and water balance is impacted, and as a result, people with cystic fibrosis produce thick, sticky mucous that causes problems in the lungs and digestive tract.

Interestingly, some disease-associated CFTR mutations are not null: protein is produced, although it does not function. Medical researchers have developed drugs that restore some function to the altered protein, relieving CFTR symptoms. The first of these drugs, trade name Kalydeco (ivacaftor), was released in 2012. These drugs have improved the quality of life and dramatically increased life expectancy for people with that mutation. However, they only work for patients with a subset of CFTR mutations. These drugs do not work for patients with null mutations, for example.

Loss of function does not imply that it's good or bad! Loss of function mutations can be beneficial to an organism, too. For example: loss of function mutations in the CCR5 gene make people resistant to HIV infection. The normal CCR5 protein acts as a co-receptor, or attachment site, for the HIV virus to infect a

cell, as shown in **Figure 8**. Without CCR5 protein in the membrane, there's no infection. The mutations that confer resistance to HIV are found in the promoter (reducing transcription of the gene) and in the coding sequence of the gene (nonsense, missense, and frameshift mutations).

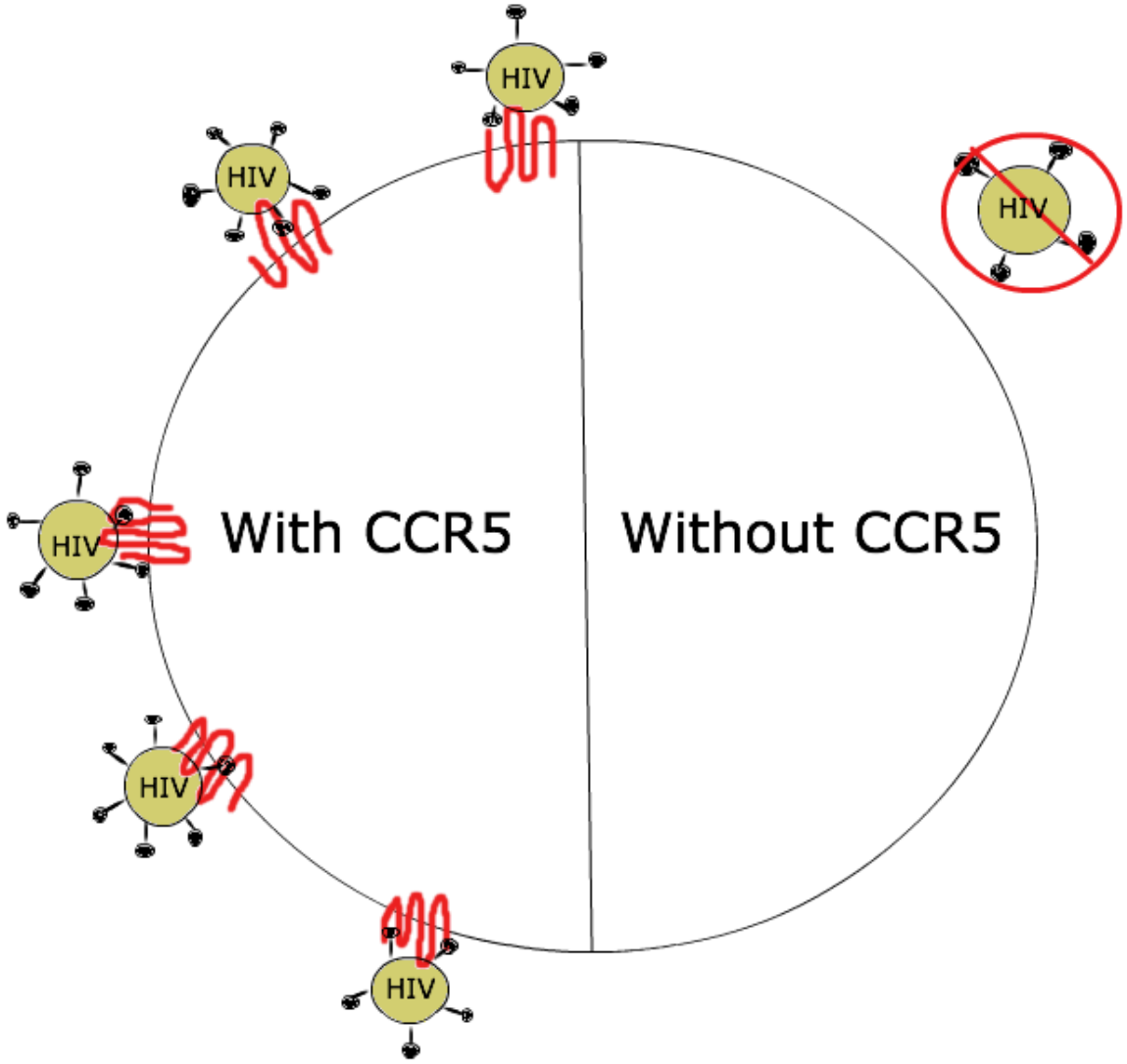


Figure 8. A loss of function mutation in the HIV coreceptor protein CCR5 makes certain lucky people resistant to HIV infection.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=231#h5p-27

In diploid organisms, gain of function mutations are usually, but not always, **dominant:** only one mutant allele is necessary to cause a phenotype. To use the stove analogy again: if one burner is on fire, it doesn't matter what the backup burner is doing; the kitchen is still on fire.

Loss of function mutations are usually, but not always, **recessive**. Two mutant alleles are needed to produce the mutant phenotype. Back to the stoves again: if one burner isn't working, you can still cook dinner on the other. But if both burners are out, no dinner is heated.

An exception to this dominant/recessive rule is seen in **haploinsufficient** genes. The word haploinsufficient means "half is insufficient" or "half is not enough." These are genes for which one copy is not enough to produce a normal phenotype, usually because the quantity of protein is important. Note that in this case, the word haploinsufficient applies to the gene, not the mutation!

An example of a human trait caused by haploinsufficiency is Ehlers-Danlos Syndrome (EDS). People with EDS have overly loose joints prone to dislocation along with stretchy, elastic skin, which you can see in **Figure 9**[1]. EDS can be caused by a mutation in one of several genes, but the classic form of the syndrome is caused by a loss of function in the COL5A1 gene. This gene encodes a form of collagen, a fibrous protein important in connective tissue. With one less functional copy of the gene, less collagen is produced, leading to the phenotype[2].

1. Whitaker, J. K., Alexander, P., Chau, D. Y. & Tint, N. L. Severe conjunctivochalasis in association with classic type Ehlers-Danlos syndrome. *BMC Ophthalmol.* **12**, 47 (2012).

2. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number: *120215 COLLAGEN, TYPE V, ALPHA-1; COL5A1. last edit 05/18/2021. World Wide Web URL: https://omim.org/.

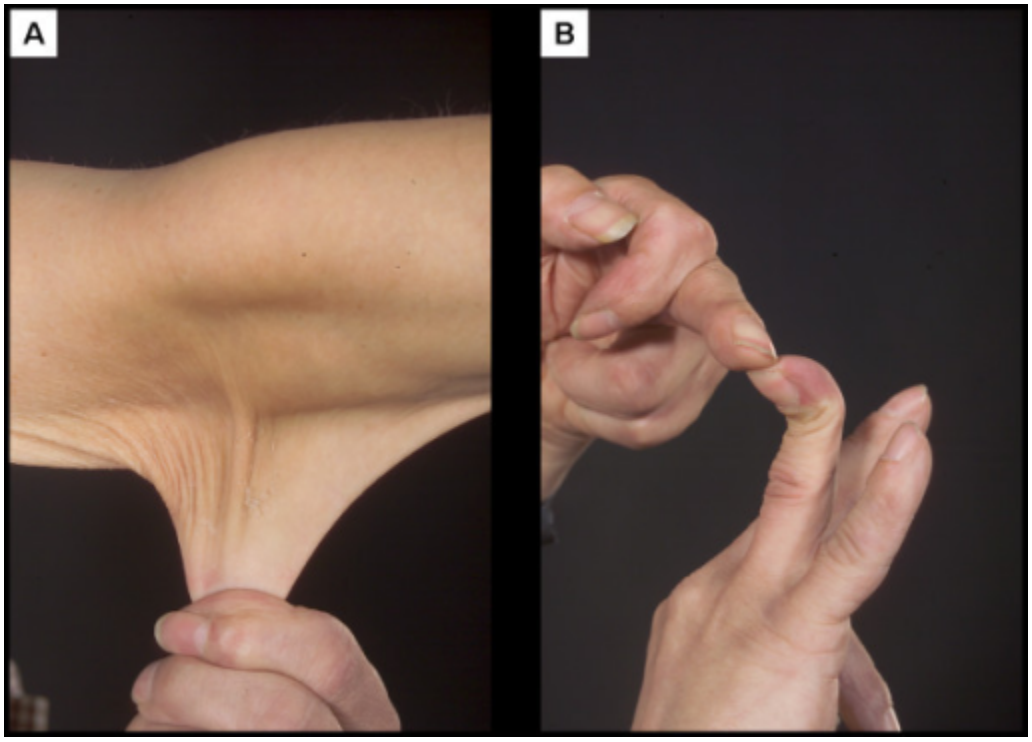


Figure 9. Hyper-elasticity of the skin (A) and hypermobile joints (B) in Ehler-Danlos syndrome.

Mutations in both coding and non-coding regions can affect the function of a protein. For example, if the sequence of the promoter is changed or deleted so the transcription machinery can't bind to the promoter, no RNA or protein will be produced. This is a loss of function. If the sequence of the promoter is changed to increase the binding of the transcription machinery, the rate of transcription may be increased, leading to extra protein production. This is a gain of function.

Even intron sequences can affect phenotype: the module on Transcription and RNA Processing explains how the removal of introns depends on consensus sequences at the intron/exon boundaries plus a "branch point site", shown below in **Figure 10**. If those sequences are altered, the intron won't be removed from the RNA. If the mRNA retains introns, it can't be properly translated.

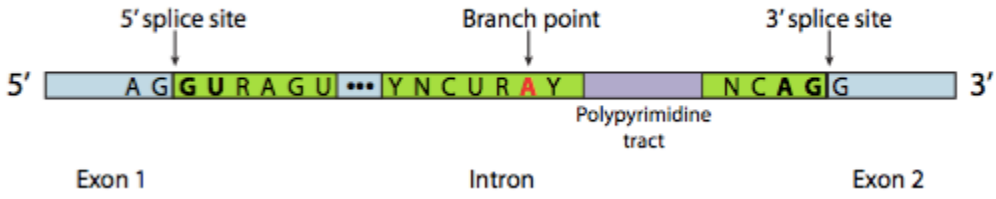


Figure 10. Consensus sequences of the 5' splice site, the branch point, and the 3' splice site.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=231#h5p-28

Does the mutation change the effect of other mutations? Revertant and suppressor mutations

The examples so far have focused on single mutations. But in some cases, mutations can even undo the effects of other mutations! Such mutations are called **suppressor** mutations because they suppress, or block, the first mutation.

Suppressor mutations can be i**ntragenic**, undoing the effect of another mutation in the same gene, or **intergenic**, undoing the effect of a mutation in another gene.

Although there are many mechanisms by which suppressor mutations can act, one example is through "matching" mutations in molecular binding partners. For example, if two proteins interact to form a larger complex, as shown in **Figure 11**, a change to one protein might disrupt binding with its partner. But if a second mutation changes the partner in exactly the right way, this could allow binding to occur again.

Figure 11. An intergenic suppressor cancels out the effects of a mutation in another gene. Here, a mutation in the gene encoding protein A disrupts binding between protein A and protein B. However, a complementary mutation in the gene for protein B would restore binding.

An example of this are the so-called **nonsense suppressors.** As discussed earlier in this module, a nonsense mutation changes an amino acid-specifying codon to a stop codon, resulting in a truncated (short) protein. A nonsense suppressor allows the translation machinery to read through a stop codon and continue synthesizing the rest of the protein! How does this work? Remember, the stop codon usually signals translational

termination because it can't be recognized by a tRNA. **Nonsense suppressors** are mutations in a gene encoding a tRNA! The tRNA sequence is mutated so that the anticodon is complementary to the stop codon.

Although nonsense suppressors occur naturally, engineered nonsense suppressors are an interesting possibility for treating genetic diseases caused by nonsense mutations (**Figure 12**). Because many diseases are caused by nonsense mutations, it is hoped that a treatment using nonsense suppressor tRNA could be used to treat multiple diseases[3]. As of 2023, several companies are working to develop tRNA-based therapies for diseases caused by nonsense suppression[4].

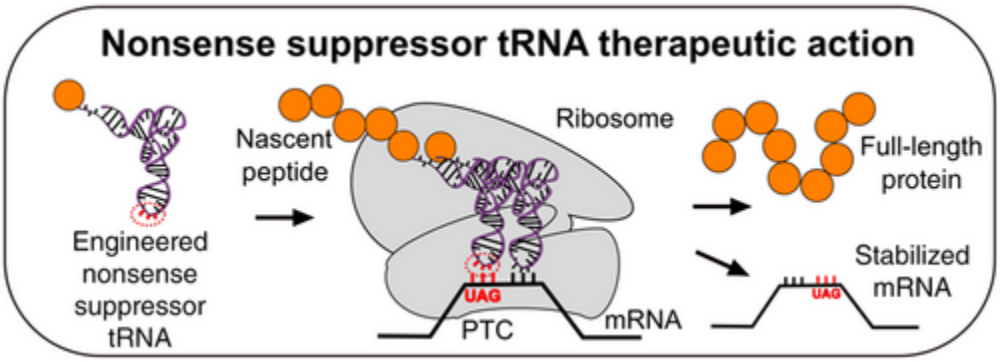


Figure 12. Nonsense suppressors may be useful to treat genetic diseases caused by nonsense mutations. In this image, a tRNA (left) is engineered so that the anticodon can bind to the codon UAG. In the center of the image, this engineered tRNA is shown bound to the UAG codon, carrying an amino acid to the ribosome. This allows read-through past the stop codon, so a full-length protein can be produced. Because mRNAs with premature stop codons are also degraded quickly by the cell, this also stabilizes the mRNA.

Intragenic suppressor mutations are a second mutation that undoes the effects of a first mutation. An example of this is shown in **Figure 13**, where an insertion of two bases causes a frameshift, but an insertion of two bases *and* a deletion of two bases results in a short change to the sequence but no frameshift.

The type of change to a protein determines the impact to protein function, as well as the location of the mutation. For example, a silent mutation is predicted to have no impact on the function of a protein. A conservative missense mutation is likely to have far less impact on the function of a protein than a frameshift, since only one codon is affected rather than many. And a frameshift early in a gene would likely to have greater impact to the function of the protein than would a frameshift that happens at the end of a gene, simply because more codons are affected.

3. Porter, J. J., Heil, C. S. & Lueck, J. D. Therapeutic promise of engineered nonsense suppressor tRNAs. *WIREs RNA* **12**, e1641 (2021).

4. Elie, D. tRNA therapeutics burst onto startup scene. *Nat. Biotechnol.* **40**, 283–286 (2022).

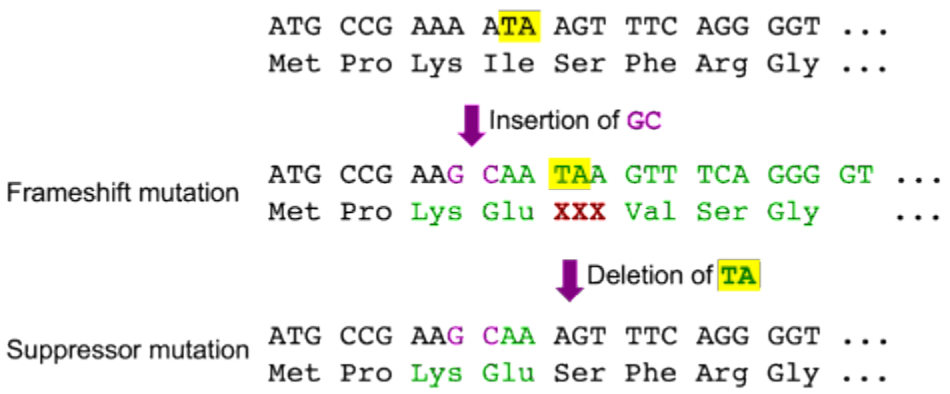


Figure 13. An intragenic suppressor is a second mutation in the same gene that undoes the effects of the first. Here, an insertion mutation followed by a deletion resets the reading frame back to the original. Although there is an alteration of two codons, there is no frameshift.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=231#h5p-29

Media Attributions

- Germline vs somatic mutation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Mutations and DNA © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Large-scale rearrangements © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Base substitution mutations © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Frameshift-mutation © L. Canham is licensed under a CC BY-NC (Attribution NonCommercial) license
- Two-gas-burners © By BogTar201213-Own work is licensed under a CC BY-SA (Attribution ShareAlike) license

PART II: DNA DAMAGE CAUSES MUTATIONS

So how do these changes to DNA arise? Mutations begin with damage to DNA.

DNA damage can be endogenous (from within) or exogenous (caused by external forces). Endogenous sources of DNA damage include mistakes during replication and exposure of the DNA to certain natural byproducts of metabolism. Exogenous sources of DNA damage include exposure to UV light, carcinogens, and radiation (like X-rays). Most DNA damage is repaired by the cell, discussed in the module on Cancer Genetics. But if damage is not fixed before replication occurs again, the error can become fixed in the genome as a mutation. The next section looks at sources of DNA damage and mutation.

Replication errors can cause single base changes

Many mutations begin as replication errors, with an incorrect base inserted opposite the parent template during elongation of the daughter strand.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=240#h5p-30

Figure 14. Tautomeric shift of guanine from the common keto form to the uncommon enol form.

One way this is hypothesized to occur is due to the chemistry of the nucleotide bases. Although we draw the bases as static structures, the bases undergo rare **tautomeric shifts**, a spontaneous rearrangement of hydrogens within the structure. These changes happen spontaneously but infrequently. You can think of this almost like the bases briefly flickering back and forth from one form to another, as is shown for guanine in the animated

slides in **Figure 14**. The changes are reversible, but the shift appears to last long enough for the replication to be affected if the replication machinery encounters the wrong tautomer[1].

Tautomeric forms of all four bases are shown in **Figure 15**. The common tautomers are shown left in the figure. Those are in equilibria with the less common variants shown on the right.

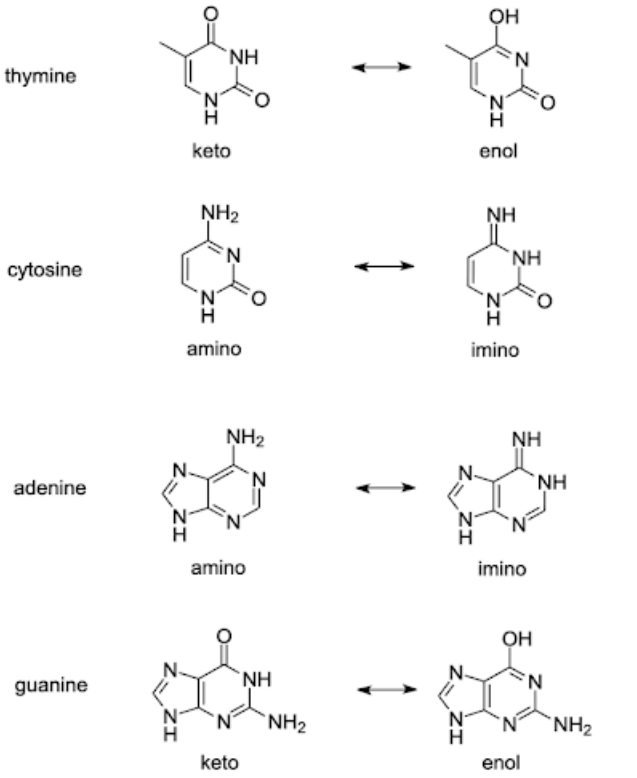


Figure 15. Tautomers of the four DNA bases. The common forms of the bases are shown on the left, and the rare forms are on the right. Note that the base-pairing functional groups are altered, so tautomers pair with different bases.

The arrangement of hydrogen bond donors and acceptors is changed in the **base-pairing** part of the tautomers. Although the common keto form of thymine base pairs with adenine, the rare enol form of thymine pairs with guanine. Likewise, the rare tautomers of other bases the rare enol form of guanine pairs with thymine, the rare imino form of adenine pairs with cytosine, and the rare imino form of cytosine pairs with adenine. If a base in an unwound template strand undergoes a shift, the wrong base may be incorporated in the daughter strand[23].

After replication, the base can shift back to the preferred amino or keto state, but this leaves behind a mismatch in the DNA. This is not a mutation yet! But it is a lesion in the DNA. Most mismatched DNA is repaired, replacing the wrong base for the right one. But if the lesion is not fixed before the next round of replication, the mis-incorporated base will be used as a template, and the daughter double helix will have a single base mutation compared to the original parent. **(Figure 16)**

1. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number: *120215 COLLAGEN, TYPE V, ALPHA-1; COL5A1. last edit 05/18/2021. World Wide Web URL: https://omim.org/.

2. Slocombe, L., Winokan, M., Al-Khalili, J. & Sacchi, M. Proton transfer during DNA strand separation as a source of mutagenic guanine-cytosine tautomers. *Commun. Chem.* **5**, 144 (2022).

3. Fedeles, B. I., Li, D. & Singh, V. Structural Insights Into Tautomeric Dynamics in Nucleic Acids and in Antiviral Nucleoside Analogs. *Front. Mol. Biosci.* **8**, 823253 (2022).

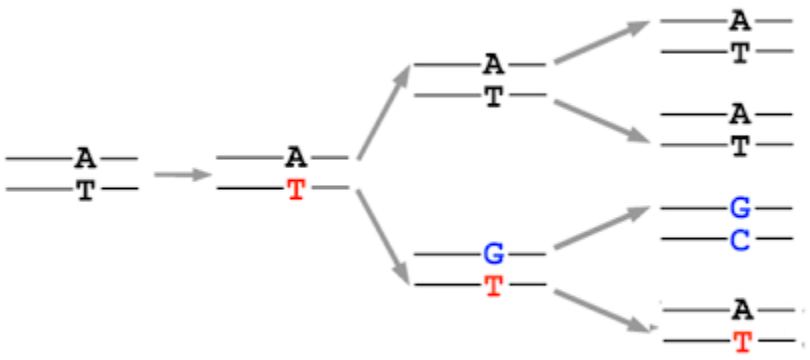


Figure 16. A mis-incorporated base will become a mutation if it is not corrected before the next round of replication. Mispairing of bases (e.g. G with T) can occur due to tautomerism, alkylating agents, or other causes. The mispaired GT basepair will likely be repaired or eliminated before further rounds of replication. But in this example, if it is not repaired, the AT base pair in the original DNA strand will become permanently substituted by a GC based pair in some progeny.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=240#h5p-31

The types of mutations that might be caused by tautomeric shifts are called **transition mutations.** These are single base changes that switch a purine for a purine or a pyrimidine for a pyrimidine. Other types of DNA damage, discussed in a later section, may cause **transversions**, which switch a purine for a pyrimidine or vice versa. This is listed in **Table 2**. We will see examples of DNA damage that cause transversion mutations later in the chapter.

Table 2. Types of DNA Damage

Type of mutation	Base Change
Transition	A⇔T / purine ⇔ purine C⇔G / pyrimidine ⇔ pyrimidine
Transversion	A⇔C / purine ⇔ pyrimidine G⇔T / purine ⇔ pyrimidine T⇔A / pyrimidine ⇔ purine C⇔G / pyrimidine ⇔ purine

Replication errors can cause insertions and deletions

Replication errors can also result in insertions and deletions of bases. This is due to **strand slippage**: During replication, the template and daughter strands sometimes dissociate (unpair) from one another temporarily. Replication cannot continue until the 3' end of the growing daughter strand re-pairs with the template. But sometimes, re-pairing is misaligned, as shown in **Figure 17.**

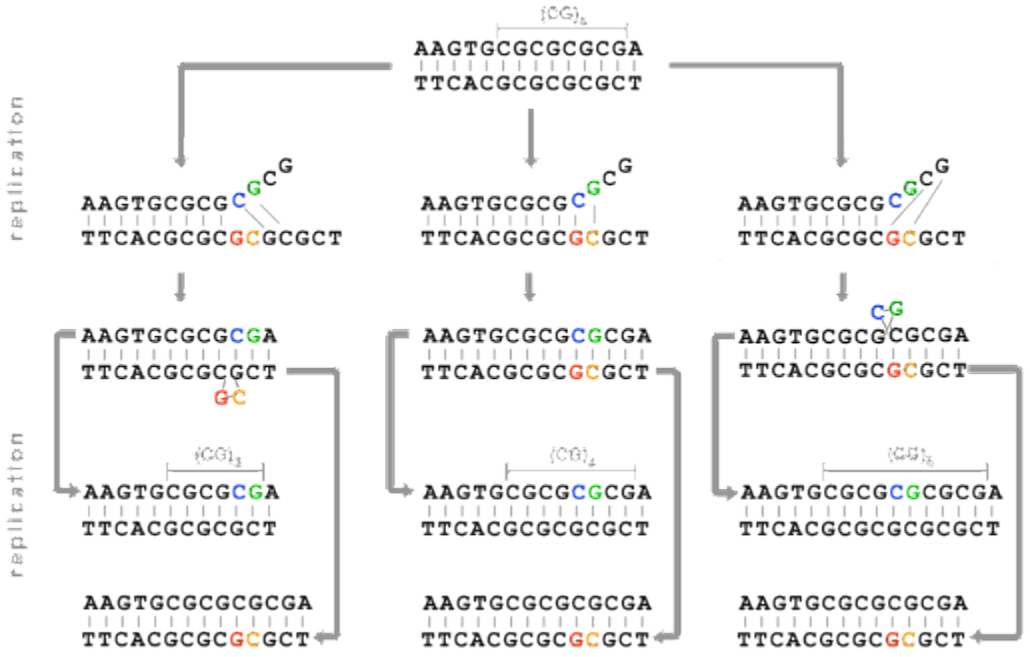


Figure 17. Strand slippage can cause insertions and deletions. Strand slippage can occur occasionally during replication, especially in regions with short, repeated sequences. This can cause a "looped out" section of the DNA on either the parent or daughter strand. If the looped-out DNA is replicated again, this can lead to either deletion (left) or insertion (right) of sequences compared to the products of normal replication (center), depending on whether the template strand or daughter strand was "looped-out".

This strand-slippage happens because one or more bases are looped out when the strands re-pair. If the parent strand forms a loop, the daughter strand will be missing the looped bases, as shown on the left in **Figure 17**. If the daughter strand forms a loop, re-pairing the 3' end along a segment that has already been replicated, the strand will end up with extra bases compared with the parent. This is shown on the right in **Figure 17**.

Strand slippage is especially common in regions of the genome with repeated sequences – like multiple CGCGCGC repeats, or triplet repeats like CAGCAGCAG. These types of repeated sequences are called **microsatellites.** These parts of the genome show more frequent germ-line changes than other areas of the genome, with expansion (insertion of additional repeats) or contraction (deletion of repeats) common.

Because of this, they're said to be unstable. Some studies suggest a mutation rate of about 1 in every 1000 parent-offspring transmissions, which is much higher than what is observed for the rest of the genome[4].

Many (but not all) microsatellites are in noncoding regions of the genome, and their expansion or contraction has little effect on an individual's phenotype. Microsatellite regions in noncoding DNA do not typically affect an individual's reproductive fitness, and microsatellite regions of the genome tend to be very variable within a population.

Differences in microsatellite length can easily be detected using PCR analysis followed by gel electrophoresis. [Although PCR is not discussed extensively here, you can see an overview of PCR from the National Human Genome Research Institute.] This made microsatellites useful in **DNA fingerprinting**: using DNA in forensic analysis to identify perpetrators or victims of crime, for paternity testing, or for clinical applications like tracking recovery after a bone marrow transplant. DNA fingerprinting is not limited to human populations, either. Ecologists can use DNA fingerprinting in much the same way to track familial relationships among individuals in a population.

Consequences of microsatellite expansion: Huntington's Disease

Sometimes, though, short sequence repeats *can* affect phenotype. A family of genetic disorders called **triplet repeat disorders** results from the expansion of three nucleotide repeats. Most of the triplet repeat disorders are neurodegenerative diseases, causing neurological problems that increase in severity with age. One of the best-known is Huntington's disease. Huntington's disease is an autosomal dominant disorder – only one copy of the disease-associated allele is necessary to cause symptoms. Huntington's disease causes profound neurological symptoms that appear in mid-life. These include involuntary movements, cognitive decline, behavioral problems, and dementia. These symptoms are associated with neuronal cell death. Life expectancy after diagnosis is typically around 17 years[5].

Huntington's disease is caused by an expansion of a CAG repeat within the gene HTT, which encodes the protein huntingtin. These are insertion mutations, but they are not frameshifts since the insertion is always a multiple of three bases. CAG is the codon for glutamine, abbreviated Q. In healthy individuals, alleles have fewer than 35 CAG repeats, and the resulting protein has fewer than 35 glutamine residues in a row.

4. Xu, X., Peng, M., Fang, Z. & Xu, X. The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **24**, 396 (2000).

5. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number: #143100: Huntington Disease. last edit 12/02/2022. World Wide Web URL: https://omim.org/.

Individuals with >40 repeats in at least one allele develop profound neurological symptoms between ages 30-40, with more repeats correlating with earlier onset[6].

The polyglutamine tracts tend to aggregate, or form clumps, in degenerating neurons in the brain, although it is not clear whether it is the protein or another aspect of the CAG expansion that causes neuronal toxicity. What is apparent, though, is that more than 35 CAG repeats causes the allele to be unstable through somatic cell divisions, so the repeat expands even more through multiple somatic cell divisions. The dying neuronal cells can have far more repeats than the inherited germline allele.

This is shown in **Figure 18A**, reprinted from Donaldson et al (2021)[7]. A healthy allele has between 5 and 35 CAG repeats and produces HTT protein with a non-expanded, non-pathogenic number of glutamine residues (dark blue in the figure). A disease-associated allele has 36-100+ CAG repeats, producing a protein with many more glutamines. In **Figure 18B**, the steps to disease progression are illustrated. An individual may inherit an expanded allele with 36 or more repeats that results in toxicity to those cells[8].

6. ibid

7. Donaldson, J., Powell, S., Rickards, N., Holmans, P. & Jones, L. What is the Pathogenic CAG Expansion Length in Huntington's Disease? *J. Huntingt. Dis.* **10**, 175–202.

8. ibid

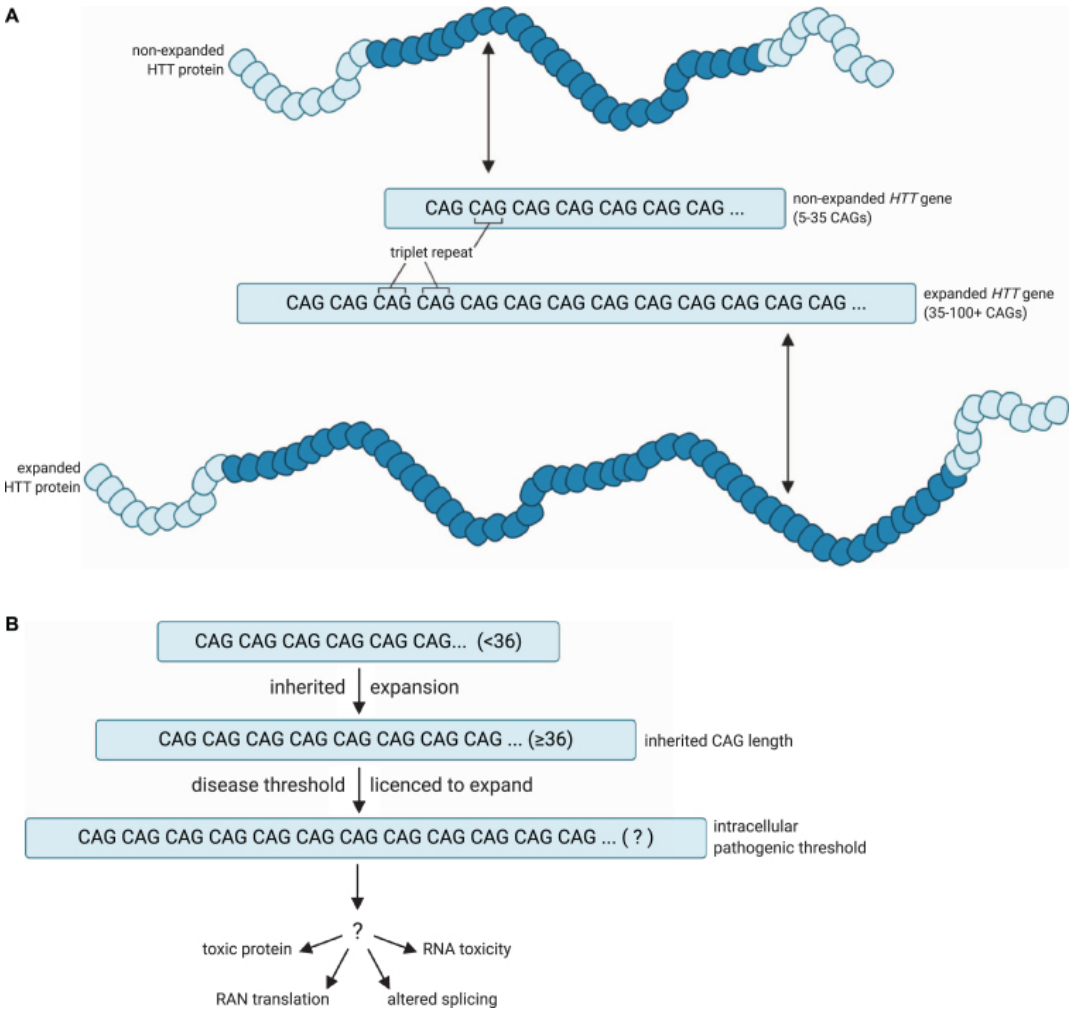


Figure 18. Reprinted from Donaldson et al (2021), CC BY NC 4.0: A model for the pathogenic threshold in HD. A) HD pathogenesis is largely determined by an expanded cytosine-adenine-guanine (CAG) trinucleotide repeat within exon 1 of the huntingtin (HTT) gene, which is translated into an expanded polyglutamine tract in the corresponding HTT protein. Wild-type HTT possesses 5– 35 CAG repeats (non-expanded HTT gene) and can undergo expansion into the disease range in the germline to create apparent de novo HD subjects, but ≥36 + repeats are associated with a significantly increased risk of developing HD (expanded HTT gene). B) An expanded HTT allele with 36 or more repeats is unstable and licensed to further expand in cells over the lifespan of the HD at-risk individual. HD symptoms would

manifest and progress as increasing numbers of disease-relevant cells undergo somatic expansion beyond an unknown intracellular pathogenic threshold that renders the gene toxic in those cells. Figures created using BioRender.com, adapted from a figure by the National Institute of General Medical Sciences, National Institutes of Health.

Repeat length is correlated with greater instability. Although people with a high healthy number of repeats (27-35) typically will not develop symptoms of H.D., their children are at greater-than-normal risk for inheriting a germline mutation with additional repeats[9]. Larger numbers of repeats are also correlated with earlier onset and increased severity of disease.

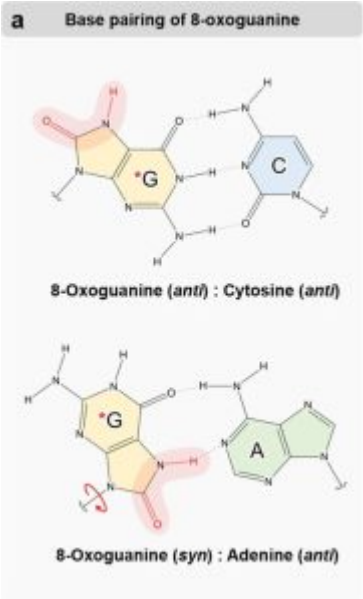
Byproducts of metabolism can damage DNA

Tautomeric mispairings and strand slippage are examples of replication errors. But in addition to replication errors, endogenous DNA damage can occur due to exposure to byproducts of metabolism, including **reactive oxygen species (ROS)** like hydrogen peroxide (H_2O_2) and superoxide (O^{2-}). Some examples of damaged bases occurring during normal cell metabolism are shown in **Figure 20**.

9. Migliore, S., Jankovic, J. & Squitieri, F. Genetic Counseling in Huntington's Disease: Potential New Challenges on Horizon? *Front. Neurol.* **10**, (2019).

Oxidative DNA damage results in base substitutions

ROS contact can cause oxidative damage. One of the most common forms of oxidative damage is the formation of 8-oxoguanine. 8-oxoguanine can still form a normal base pair with cytosine, S, as shown at the top of **Figure 19**[10]. However, 8-oxoguanine can also rotate around the glycosidic bond connecting the base to the sugar, as shown in the bottom panel. If this happens, 8-oxoguanine can mispair with adenine via functional groups on the modified part of the molecule. If this mispairing occurs during replication and is not repaired, it will convert a GC base pair to a TA base pair after a second round of replication.



Part I9: 8-oxoguanine can rotate around the glycosidic bond to mispair with adenine.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=240#h5p-32

Deamination of bases and abasic sites

Other base damage that can occur includes deamination of bases and loss of bases to form an **abasic** site. Deamination of cytosine produces uracil, as shown in **Figure 20A**. The deaminated cytosine pairs with A instead of G, potentially introducing transition mutations if the damage is not corrected.

A nucleotide residue can also lose its base entirely, resulting in an abasic site, as shown in **Figure 20B.** Abasic sites are also called apurinic or apyrimidinic sites (abbreviated AP sites). This occurs because the glycosidic bond that connects the base to the sugar can undergo hydrolysis, a chemical reaction that breaks the bond with the addition of a water molecule. Purines are about 20 times more susceptible to hydrolysis than pyrimidines.

10. Hahm, J. Y., Park, J., Jang, E.-S. & Chi, S. W. 8-Oxoguanine: from oxidative damage to epigenetic and epitranscriptional modification. *Exp. Mol. Med.* **54**, 1626–1642 (2022).

The rate of base hydrolysis is increased when DNA is single-stranded during replication and with exposure to ROS[11].

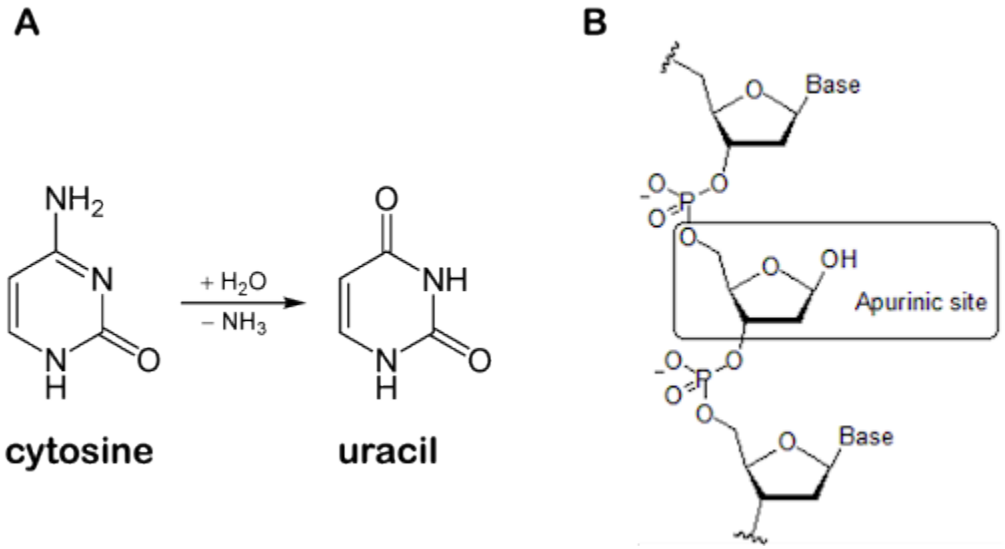


Figure 20. Damage to bases results during normal cellular metabolism. A: Deamination of cytosine creates uracil, which base pairs with adenine. This can ultimately cause CG>TA mutations. B: Hydrolysis of the glycosidic bond connecting a base to a sugar can result in an abasic site. This is also called an apurinic or apyrimidinic site (AP site), although the loss of purines happens about 20x more frequently than the loss of pyrimidines. Image source: Modified from Wikipedia-Yikrazuul-PD (left) and Wikipedia-Chemist234- CC BY-SA 3.0 (right), via Open Genetics Lectures.

Exogenous causes of DNA damage and mutations

Although DNA damage and a subsequent mutation can occur spontaneously or as a result of normal cellular metabolism, DNA damage can also be induced through exposure to particular kinds of chemicals. Such chemicals are called **mutagens.** Alkylating agents and oxidizing agents are known to cause damage to bases which is seen above. Other mutagens work differently, introducing insertions or deletions or even breaking the backbone of DNA itself.

Other causes of DNA damage include exposure to environmental radiation like X-rays or UV light. Exposure to X-rays and other forms of ionizing radiation can break the sugar-phosphate backbone of DNA, causing breaks in one or both strands of the molecule. These must be repaired before the cell replicates the DNA or undergoes mitosis or meiosis; otherwise, parts of a chromosome can be lost.

11. Chastain, P. D. *et al.* Abasic sites preferentially form at regions undergoing DNA replication. *FASEB J.* **24**, 3674–3680 (2010).

UV light causes **intrastrand crosslinks**, which are covalent bonds that form between adjacent pyrimidines in the backbone of DNA. They are also called **pyrimidine dimers**.

The structure of a thymine dimer is shown in **Figure 21**, but there are other ways adjacent pyrimidines can become linked as well. If not repaired before DNA is replicated, these lesions can cause frameshift mutations because the crosslinked bases may be interpreted as a single base rather than two bases.

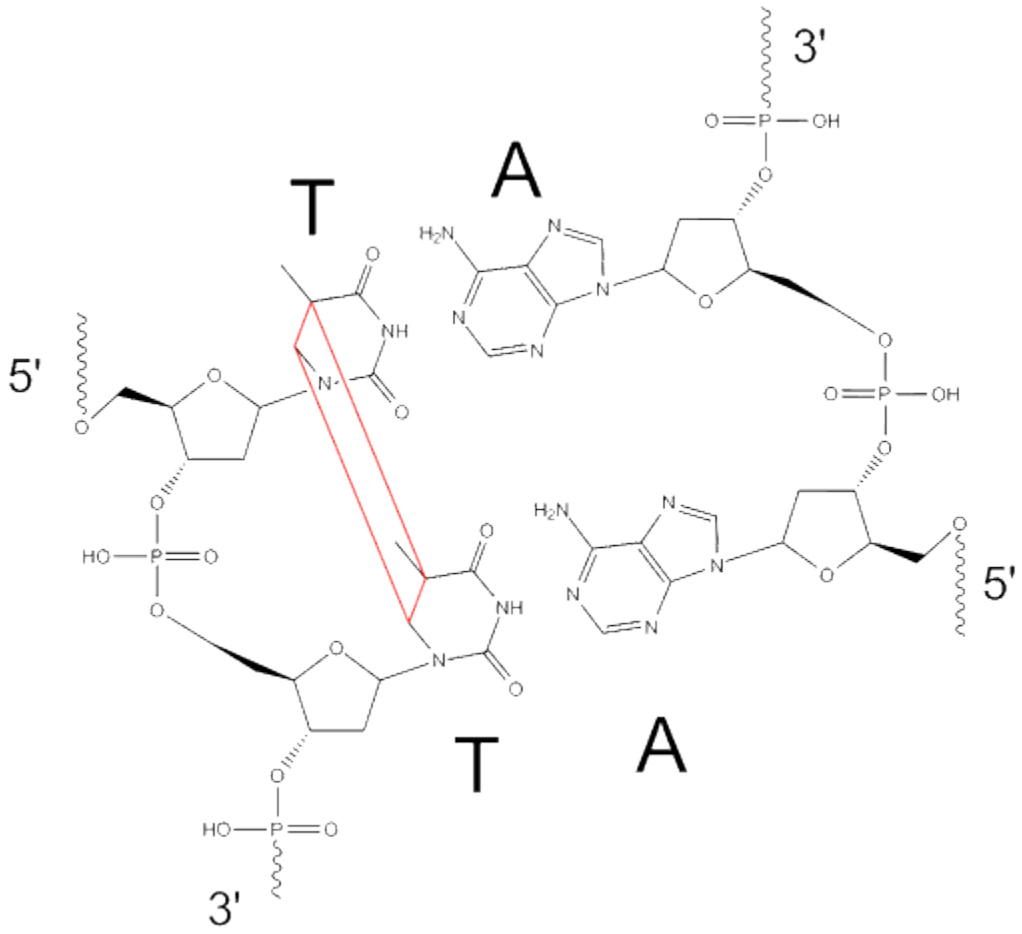


Figure 21. Structure of a thymine dimer caused by UV light.

UV light – which is part of sunlight – is a common source of somatic mutations in skin cells. This contributes to the development of many skin cancers! When you are cautioned to wear sunscreen and avoid tanning, it is because cancer-causing mutations can accumulate in skin cells over the course of a lifetime of exposure to sunlight. This is discussed more in the module on Cancer.

Cells constantly sustain damage – but most of it is repaired

You might think that DNA damage is a pretty rare event – after all, we started this chapter by saying that there is about one mutation per cell division. But, in fact, DNA damage happens far more often than that! By one estimate, in humans, around 10,000 apurinic sites (the most common DNA lesions) occur *per cell* every day[12]! So why are there not more mutations?

DNA damage is repaired by DNA damage response proteins. There are multiple pathways for DNA repair in cells, each of which recognizes a different form of DNA damage. Many of the DNA damage response proteins are **tumor suppressor** proteins, the accumulation of somatic mutations that can lead to an individual cell becoming cancerous. This is discussed more in detail in the module called DNA Repair and Cancer.

It is only when a DNA lesion escapes repair long enough to be replicated that a lesion becomes a mutation.

Media Attributions

- Tautomers © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Mis-incorporated base © Amanda Simons modified from Open Genetics Lectures: Original-Deyholos is licensed under a CC BY-NC (Attribution NonCommercial) license
- Strand slippage © Amanda Simons modified from Open Genetics Lectures: Original-Deyholos is licensed under a CC BY-NC (Attribution NonCommercial) license
- Pathogenic threshold in HD © Donaldson et al (2021) is licensed under a CC BY-NC (Attribution NonCommercial) license
- 8-Oxoguanine-e1708150331221-627×1024-1 © Hahm, J. Y., Park, J., Jang, E.-S. & Chi, S. W. 8-Oxoguanine: from oxidative damage to epigenetic and epitranscriptional modification. Exp. Mol. Med. 54, 1626–1642 (2022).
- Damage-cellular-metabolism © Modified from Wikipedia-Yikrazuul-PD (left) and Wikipedia-Chemist234 is licensed under a CC BY-NC (Attribution NonCommercial) license
- Thymine-dimer © Open Genetics Lectures, Original-Deyholos is licensed under a CC BY-NC (Attribution NonCommercial) license

12. Thompson, P. S. & Cortez, D. New Insights into Abasic Site Repair and Tolerance. *DNA Repair* **90**, 102866 (2020).

SUMMARY

DNA damage can cause changes to the DNA sequence. This can happen through errors during replication, exposure to normal byproducts of metabolism, or exogenous causes like exposure to carcinogens or radiation. The changes to the DNA sequence are called mutations. Most mutations do not impact the function of genes and therefore do not affect the phenotype of an organism. These are neutral mutations. However, other mutations do affect the function of a gene and the phenotype of the organism.

Mutations can be classified depending on the cell affected, the change to DNA sequence, the change to protein sequence, or the change to protein function.

WRAP-UP QUESTIONS

Amanda Simons

Questions 1-4 are from *Open Online Genetics* (Nickle and Barrett-Ng), Chapter 4[1].

1. How are polymorphisms and mutations alike? How are they different?

2. What are some of the ways a substitution can occur in a DNA sequence?

3. What are some of the ways a deletion can occur in a DNA sequence?

4. What are some of the ways an insertion can occur in a DNA sequence?

5. The chapter on Translation has a figure that shows the structures of amino acids. Use that figure to predict which mutation would have a bigger effect on the function of a protein.

a. An alanine-to-glycine missense or an alanine-to-asparagine missense.
b. An alanine-to-asparagine missense or a nonsense mutation in the second-to-last codon.
c. An alanine-to-asparagine missense or a nonsense mutation in the 10th codon.
d. A nonsense mutation in the second-to-last codon or a frameshift mutation in the 10th codon.

6. Choose your favorite superhero or fictional character who has extraordinary abilities due to a mutation. How would you describe the mutation that gives them those abilities, using the terms in this chapter? Justify your reasoning with character traits. If you need some ideas, you could use Spiderman, Incredible Hulk, Teenage Mutant Ninja Turtles, or X-men, but you're not limited to these characters. (My personal favorites are from Kipo and the Age of Wonderbeasts.)

7. If you sequenced your genome, you'd find that you probably have between 40-80 germline mutations compared to your parents' DNA. The differences aren't evenly distributed between your maternal and paternal half-genomes. On average, you will have about 15 mutations in your maternal chromosomes, regardless of your mom's age when she had you. But the number of mutations carried by sperm increases with paternal age: if

1. Nickle and Barrette-Ng. Open Online Genetics. in *Open Online Genetics* (2016).

your dad had kids at age 20, he likely passed on an average of 25 mutations, but if your dad had kids at age 40, the average number of mutations per sperm is 65[2].

a. What does that suggest about the number of cell divisions needed to produce an egg or sperm?
b. Why do mutations accumulate with the age of the father but not the mother? Compare and contrast oogenesis and spermatogenesis during development, using outside resources as necessary.

8. Microsatellite instability is frequently seen in certain kinds of colorectal cancer. The tumor cells have acquired somatic mutations in the MSH family of proteins, which repair the kinds of lesions generated by strand slippages.

What kinds of mutations would you expect to see if you looked at the sequence of MSH genes in the tumor cells? Would you also expect to see the same mutations in healthy cells from the same patient?

The next series of questions use the Online Mendelian Inheritance in Man (OMIM) database to look at examples from human genetics. OMIM is an online database that compiles information about genes and phenotypes. It has entries for phenotypes (labeled with "#" and a number) and genes (labeled with "*" and a number).

There will be more information in each entry than you will be able to easily understand – and that's ok! Part of this exercise is to weed through an over-abundance of information to find what you need.

9. Coagulation Factor IX (F9) is an important factor in blood clotting. The OMIM entry for Factor IX is *300746. Use this entry to answer the following questions.

a. The **Gene-Phenotype Relationships Table** at the top of the entry lists multiple phenotypes associated with mutations in this gene, including Hemophilia B and Thrombophilia 8.
b. Click the links in the table to explore the **phenotype listings** for these traits.
    i. Describe Hemophilia B in terms a non-biologist would understand. You may need to look up some vocabulary words in the entry to answer this question! Is it caused by a gain or loss of function in the F9 gene? Is it a dominant or recessive disorder?

2. Philips, R. M. & R. » How many chromosome replications occur per generation? https://book.bionumbers.org/how-many-chromosome-replications-occur-per-generation/.

   ii. Describe Thrombophilia 8 in terms a non-biologist would understand. Again, you will need to look up some vocabulary to answer this question! Is it caused by a gain or loss of function in the F8 gene? Is it a dominant or recessive disorder?

c. Can mutations in a single gene cause different phenotypes? Explain your reasoning based on what you've learned from this exercise.

10. There are many variations in the gene Factor XIII (F8) that cause the excessive bleeding disorder Hemophilia A.

a. Open OMIM entry *300841. The **Table of Contents** menu at the left of the page has a link to **Allelic Variants**. Click on the **Table View** to bring up a list of variants that have been observed in humans. Scroll through the table.
   i. About how many alleles are listed?
   ii. Are all of the alleles mutations, or are some polymorphisms?
b. How many alleles of a gene is any one individual expected to have?
c. Do you expect all alleles to be evenly represented in the human population, or do you expect that some are more common than others? Explain your reasoning.

11. Many mutations in the Factor XIII gene cause Hemophilia A. Click on the **Allelic Variants** link in OMIM entry *300841 to bring up a detailed listing of alleles associated with this disorder.

a. What type of mutation is .0208?
b. What type of mutation is .0209?
c. What type of mutation is .0210?
d. What kind of mutation is .0079?
e. Explain why all four of these mutations give a similar phenotype.

12. Lists of single-gene disorders can be found online. Choose one, and look up the disorder in the database Online Mendelian Inheritance in Man (www.omim.org).

a. Describe the phenotype associated with the disorder.
b. What gene is linked with the disorder? This information is found in the **Gene-Phenotype Relationships** table at the top of the page and in the text of the entry. [Look carefully! *Other genes may be listed that modify the main phenotype*.]
c. The **Gene-Phenotype Relationships** table has a link to the gene page. Click through to that reference.
d. What is the function of the gene?
e. Look at the **Allelic Variants** for this gene. What types of mutations cause the disorder you initially

chose? Describe each mutation by the effect on gene function, change to DNA sequence, and impact on protein sequence.

Science and Society

13. In genetics, and science in general, we often learn a lot from exceptional situations: what is different genetically, molecularly, and cellularly in people with genetic disorders like Ehlers-Danlos Syndrome can teach us about the workings of a cell. And people with Ehlers-Danlos Syndrome can benefit from research into their conditions. But how do we balance the benefit of having the information with the risks to those who participate in medical studies and whose stories are told in textbooks like this? These risks might include physical health risks from participating in medical research, but they also can include risks to privacy, self-esteem, and mental health. Would you want your image to be used in a medical journal or textbook? Why or why not?

14. One of the challenges of modern genetics is distinguishing between rare polymorphisms and pathogenic (disease-causing) mutations. The majority of the original human reference genome was assembled primarily from one individual's DNA. However, later projects have looked to catalog variations. Why is it important that such projects look at genomes from individuals of varied ancestry?

PART VIII

OVERVIEW OF GENE REGULATION

Objectives

After completing this module, you should be able to:

1. Describe levels of gene regulation.
2. Describe how activators and repressors can act to increase or decrease transcription of a gene.
3. Explain how activators and repressors can work together to regulate gene expression in response to changing conditions, using the lac operon and bacteriophage lambda as an example.
4. Explain how translation of the trp leader results in transcriptional termination.
5. Predict how mutations in the lac operon would affect its expression.
6. Compare and contrast mechanisms of gene regulation in prokaryotes and eukaryotes.
7. Define: gene expression, gene regulation, transcriptome, proteome, activator, repressor, positive regulation, negative regulation, constitutive mutation, attenuation

Introduction to gene regulation

In the chapters on replication, transcription, and translation, we looked at the process by which a cell uses the information stored in the genome to produce RNA and protein. But the process of RNA and protein production is tightly controlled, so only some genes are **expressed**, or active, at any time.

The genome is the collection of DNA that is present in a cell or organism. Likewise, the **transcriptome** is the collection of RNA present in a cell or organism. The transcriptome includes protein-coding mRNAs as well as functional non-coding RNAs. The transcriptome can also include multiple forms of an RNA for each gene: edited and alternatively spliced RNA increases the diversity of the transcriptome. The **proteome** is the collection of proteins in a cell or organism. Just like for RNA in the transcriptome, one gene can contribute many proteins to the proteome, since multiple protein molecules can be generated through edited or spliced

RNA molecules. The proteome also includes proteins modified post-translationally, through covalent linkage of small molecules like phosphates, acetyl groups, sugars, or lipids.

This is illustrated in **Figure 1.**

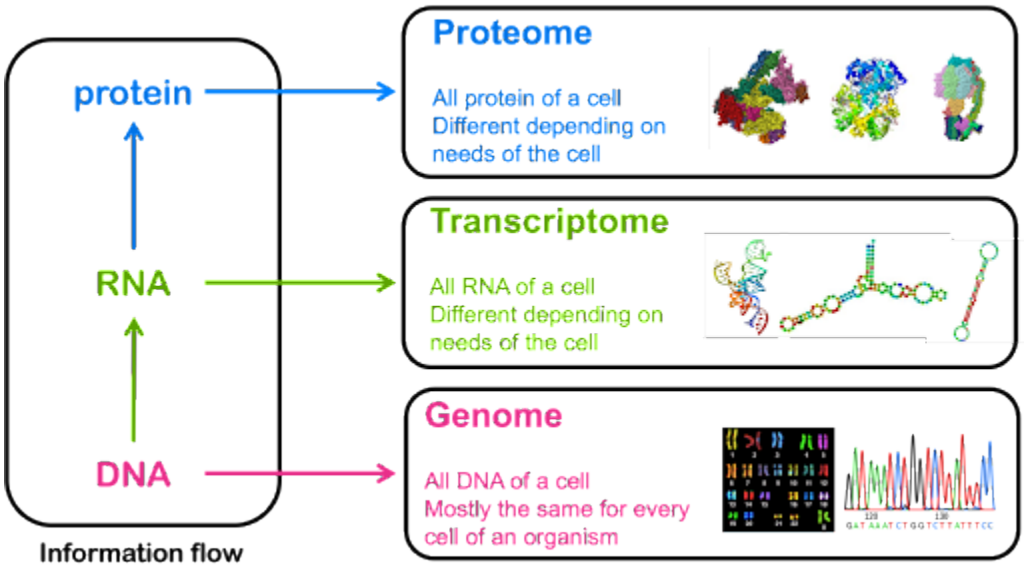


Figure 1 Genome, transcriptome, proteome. The genome encompasses all the DNA in a cell. The transcriptome is all of the RNA in a cell, including functional RNA and coding RNA. The proteome is all the protein of a cell, including post-translational modifications.

Although the genome is mostly the same in every cell of a multicellular organism, this is not true of the transcriptome and proteome. For many (but not all) genes, the ultimate gene products are protein molecules. Different proteins are needed for different cellular functions, but not all proteins are always needed by all cells. It's inefficient – and, sometimes, counterproductive – for cells to produce proteins not appropriate for the cell's function. **Gene regulation** refers to the mechanisms by which cells control **gene expression**.

The presence or absence of a protein, as well as the amount of protein produced, affects cell function. For example, in humans and other mammals, red blood cells transport oxygen throughout the bloodstream. Red blood cells contain large amounts of hemoglobin, a protein which binds oxygen in the lungs and releases it in oxygen-poor tissues. Hemoglobin is not expressed in other cells. B cells are specialized white blood cells that participate in the immune system. B cells produce antibodies, which are proteins used to launch an immune response against pathogens like viruses or bacteria. Antibodies are not produced by other cell types not involved in the antibody-mediated immune response.

On the other hand, some proteins, called **housekeeping** proteins, are needed in all cell types. For example, in both eukaryotes and prokaryotes nearly all cells require the enzymes used in glycolysis, the metabolic pathway that couples the breakdown of glucose to the production of ATP. Housekeeping genes like those of glycolysis – fundamental to the function of the cell – are always active.

Even a single-celled organism will not express every gene all at once. RNA and protein production are

energetically demanding, so genes that are not needed will not be transcribed or translated. For example, sometimes glucose is not available as a food source, and cells would need to use a different set of enzymes to break down alternative sugars. Those enzymes are only needed sometimes, so the genes are only expressed under those specific conditions. Likewise, genes required to synthesize certain amino acids are only active when the cell is short of those amino acids. Genes involved in stress response are only active when the cell is stressed.

This module begins with an overview of the mechanisms of gene regulation, with additional focus on regulation at the level of transcription. We discuss the action of transcriptional activators and repressors, including how multiple genes can be co-regulated to coordinate expression. Finally, this module ends with three well-studied examples of gene regulation in bacteria: the lac operon illustrates how activators and repressors can work together to regulate a gene, the lambda repressor demonstrates how a single protein factor can act as both an activator *and* a repressor, and the trp attenuator shows how, in prokaryotes, the process of translation can affect whether or not a gene is transcribed.

Media Attributions

LEVELS OF GENE REGULATION

In previous modules, we discussed the processes by which proteins are produced in a cell. Recall that protein production requires transcription of the gene, processing of the RNA, and translation of the RNA.

All these steps and others can be controlled by the cell in some fashion: are many RNAs produced from a gene, just a few, or none? Is an RNA molecule translated many times, producing many protein molecules from a single RNA? Or is it translated only a few times? Each step represents a potential control switch for the cell.

Levels of gene regulation are illustrated **Figure 2**:

- Chromatin compaction: How tightly the DNA is packaged affects how easily transcription machinery can access a gene.
- Transcription: Is the gene transcribed? How frequently?
- RNA processing: Is the RNA alternatively spliced? Is it edited?
- Translation: How frequently is the RNA translated?
- Post-translational modification: Does covalent addition of small molecules change protein activity?
- RNA and protein stability: How long do the RNA and protein persist in the cell before it is degraded?

Whether or not a gene is transcribed serves as the first level of gene regulation. Transcription factors can activate or repress transcription in both prokaryotes and eukaryotes. In eukaryotes, whether a gene is transcribed also depends on chromatin structure: tightly compacted chromatin cannot be transcribed because genes are not accessible to transcription machinery. This is discussed in more detail in the module on Epigenetics.

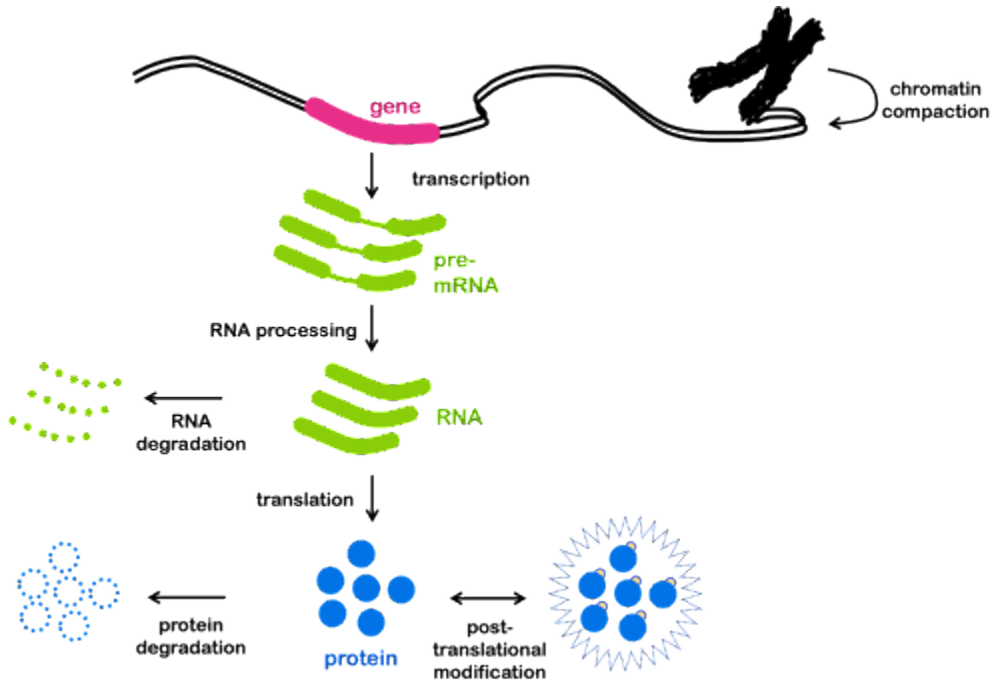


Figure 2 Levels of Gene Regulation. Expression of a gene and production of an active protein requires transcription, RNA processing, translation, and often post-translational modification.Each of those steps can be controlled by the cell to regulate gene expression Both RNA and protein have a limited lifespan in the cell, so their degradation represents additional points of control by the cell.

In eukaryotes, primary RNA transcripts are processed to become mature mRNA. In eukaryotes, processing serves as an additional level of regulation, since RNA splicing and/or RNA editing can result in different mRNAs produced under different cellular conditions. In both prokaryotes and eukaryotes, whether a mRNA is translated – and how many times each RNA is translated – can also be controlled. Post-translational modification of proteins can change the activity of a protein. Finally, although it is not shown in **Figure 2**, remember that although transcription occurs in the eukaryotic nucleus, translation occurs in the cytoplasm. Export of RNA and import and export of protein from the nucleus also can be controlled.

Transcriptional regulation of protein-encoding genes serves as the focus of this chapter, and, indeed, tends to get the most attention in genetics textbooks. However, be aware of these additional levels of regulation. If you are interested in reading more about the topic, you can read review articles about about translational regulation here, alternative splicing, and RNA transport.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=250#h5p-33

Media Attributions

- Levels of Gene Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

TRANSCRIPTIONAL GENE EXPRESSION: ACTIVATORS AND REPRESSORS

Whether a gene is transcribed, and how much RNA is produced from a gene, is called **transcriptional regulation** of a gene. How much RNA is produced is correlated with how readily transcription can be initiated. This, in turn, is correlated with how well the transcription machinery can bind to the promoter of a gene. Cells can "turn on", "turn up", "turn down" or "turn off" genes to adjust how much of a particular protein is produced. This allows a cell to react to different environmental stimuli and can allow cells to **differentiate**, or become specialized.

Recall from the module on transcription: in prokaryotes, transcription is initiated by sigma factor binding to the -10 and -35 boxes, which recruits the RNA polymerase. The "strength" of a promoter is correlated to how well the -10 and -35 boxes match the consensus sequence.

A promoter with a perfect match to the -10 and -35 consensus sequences is called a "strong" promoter. A perfect match to the -10 and -35 consensus sequence (shown in **Figure 3**) means the promoter can be bound readily by sigma factor. Binding between sigma factor and the -10 and -35 sequences depends on noncovalent bonds between the nucleotide residues and amino acid side chains within the protein. Sigma factor brings RNA polymerase to the promoter, helping it bind to initiate transcription.

A strong promoter is highly transcribed, even without any additional help from other transcription factors. Many RNA molecules are produced from such genes, no additional factors needed (**Figure 3**, left). For genes with strong promoters, the default expression is "on". In bacteria, the so-called housekeeping genes, which are always needed regardless of cell conditions, are controlled by strong promoters.

Weak promoters, however, have promoters that are a poor match to the consensus. Changes to the -10 or -35 sequence mean that the sigma factor forms fewer noncovalent bonds with the promoter, and the overall binding between the two is more prone to dissociate, or fall apart, before transcription can begin[1]. Therefore, the gene is transcribed less frequently and the default state is for only low levels of RNA and subsequently protein to be produced (Figure 3, right).

1. Davis, M. C., Kesthely, C. A., Franklin, E. A. & MacLellan, S. R. The essential activities of the bacterial sigma factor. *Can. J. Microbiol.* **63**, 89–99 (2017).

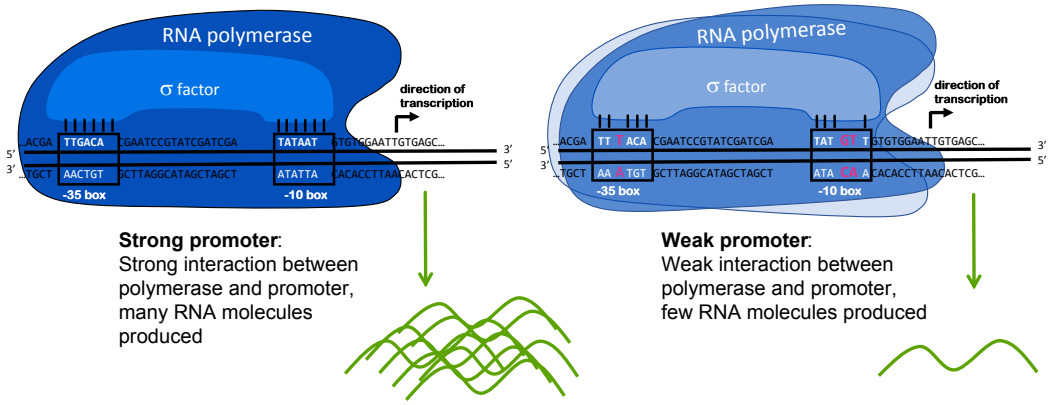


Figure 3. Sigma factor binds to the -10 and -35 boxes via noncovalent interactions with the nucleotide residues of the sequence. A strong promoter has a close match to the consensus sequence. Sigma factor binds tightly, and lots of RNA is produced (left). A weak promoter has an imperfect match to the consensus sequence, and sigma factor binds more weakly resulting in less RNA produced.

Transcription factors can increase transcription from a weak promoter (in which case they are described as transcriptional **activators**). When a gene is regulated by an activator,the process is called **positive regulation**. Also, transcription factors can decrease or block transcription (in which case they are described as **repressors**). When a gene is regulated by a repressor, the process is called **negative regulation**. Both strong and weak promoters can be regulated by repressors. A single gene can be regulated by both activators and repressors.

In prokaryotes, activators can stabilize the transcription machinery at a weak promoter by binding both DNA and either polymerase or sigma factor. Repressors commonly bind to DNA and block access of the RNA polymerase holoenzyme to the promoter, preventing transcription. This is often accomplished via the simple mechanism of taking up space! To use an analogy: if my dog is sitting on the couch, there is no room for me. If a repressor is bound to the DNA near the promoter, there is no room for the polymerase.

This is shown in **Figure 4**. Many activators and repressors bind specific DNA sequences in the promoter of the genes they regulate. Differences in the regulatory regions of promoters plus the activity of different transcription factors are what allow genes to be differently expressed.

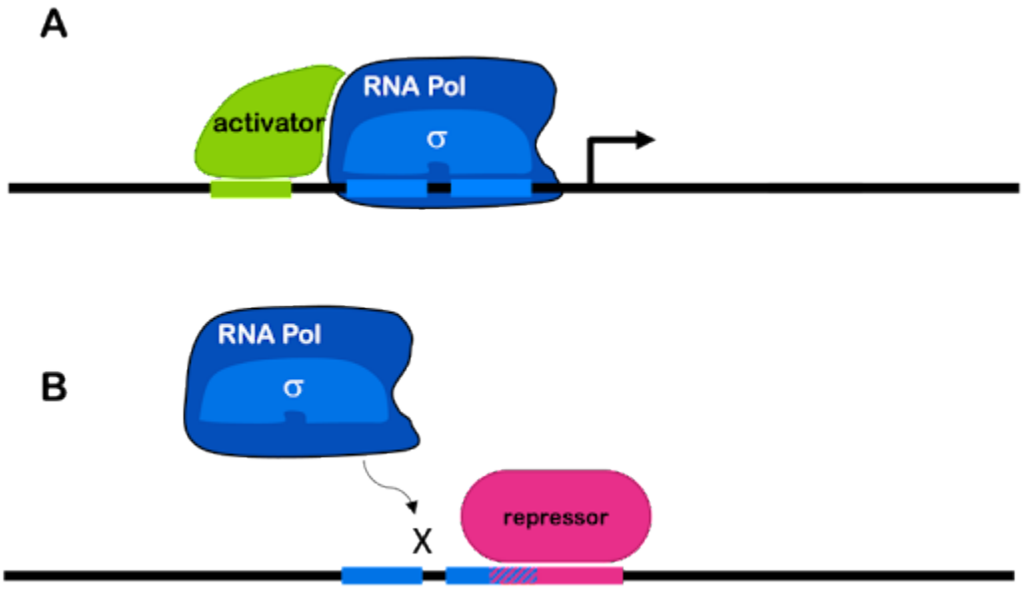


Figure 4. Positive and negative regulation of transcription. In bacteria, activators typically bind both DNA and either the polymerase or sigma factor, stabilizing the interaction between polymerase holoenzyme and promoter. Repressors typically bind to DNA elements called operators near the promoter and physically get in the way of polymerase binding.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=258#h5p-34

Activators and repressors in eukaryotes

Although both prokaryotes and eukaryotes use activators and repressors to control gene expression, the situation is more complex in eukaryotes.

There are no strong eukaryotic promoters: All promoters require additional transcription factors to initiate transcription. In eukaryotes, a large protein complex called the **mediator** helps assemble the transcription machinery on eukaryotic promoters. Looping of the DNA brings distal elements into proximity with the core promoter. This is shown in **Figure 5**.

Eukaryotic activators may interact directly with the core polymerase – as is most often true in prokaryotes

– but they may also interact with the polymerase indirectly via the mediator or other transcription factors. Activators may bind via proximal elements near to the core promoter, or they may bind to distal elements as part of an **enhancer** farther away from the promoter.

Each enhancer typically has binding sites for multiple transcription factors, often including both activators and repressors. This allows a combination of factors to coordinate to regulate gene expression under very precise conditions. In **Figure 5**, you can see that the light-blue element is shared between the enhancers of both genes, but other elements of the enhancer are different. This reuse of sequence elements (and their corresponding protein factor) in multiple genes is common in eukaryotic regulation.

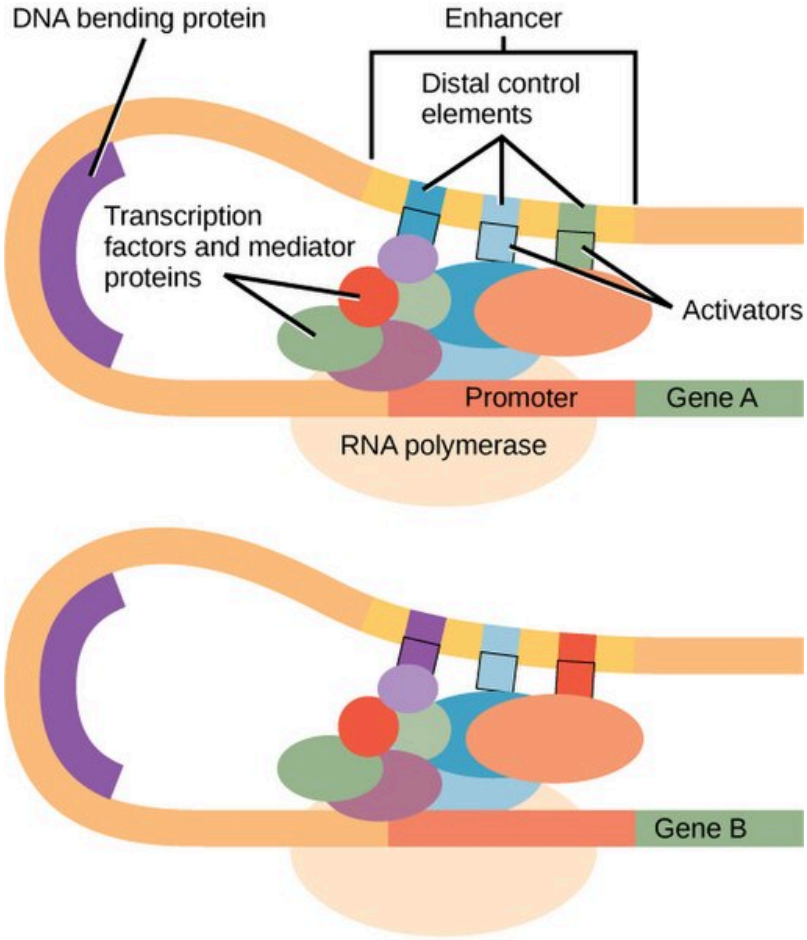


Figure 5. DNA is flexible and can loop to bring distal control elements into close proximity of the promoter in three dimensional space. This is helped by the action of DNA bending proteins and the mediator complex, which acts as a scaffold between proteins bound to the core promoter and distal control elements. Different genes have different distal elements, allowing for gene expression under different conditions.

Like their prokaryotic counterparts, eukaryotic repressors may directly block access of the polymerase to the promoter as in eukaryotes. But they may also block an activator from binding to a distal control element in

an enhancer, they may interfere with the mediator, or they can alter chromatin compaction in a way that the chromatin itself blocks the transcription machinery's access to DNA.

Enhancers can be found upstream of a promoter as shown in **Figure 5**. But some genes are regulated by enhancers found downstream of the coding sequence or even in an intron!

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=258#h5p-35

Coordinated expression of multiple genes

In many cases, multiple proteins are needed to work together as a team in a biological pathway. Prokaryotes and eukaryotes use different mechanisms to ensure that all needed proteins are produced at the same time and at appropriate amounts in the cell.

In prokaryotes, functionally related genes are often linked in an **operon**. These genes are arranged one after the other on the bacterial chromosome, but they are under the control of one single promoter, as shown in Figure 6. The promoter can include binding sites for activators and/or repressors (an activator is shown in Figure 6).

The promoter controls the production of a single mRNA with the coding sequences for multiple **structural genes** one after another. This is called a **polycistronic** RNA (a cistron is equivalent to a gene). Each protein is translated independently: each open reading has its own ribosome binding site, start codon, and stop codon. Thus, the genes in an operon are transcribed together, but translated separately. We will see an example of this in the lac operon, described later in this module.

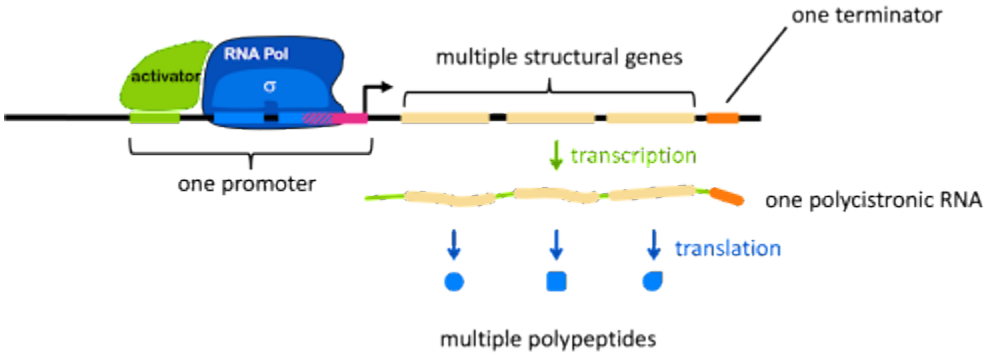


Figure 6. operon structure. An operon is a single transcriptional unit that contains multiple co-regulated genes. It has a single promoter and a single terminator and one polycistronic RNA is produced. Multiple ribosome binding sites allow multiple polypeptides to be translated from the single RNA> This mechanism ensures that if one gene is transcribed, they all are, allowing related genes to be co-regulated.

In addition to the structural genes and the promoter, an operon usually has one or more cis-acting **elements** that are important for gene regulation. **Figure 6** shows two such elements: the green element bound by an activator, and a pink element overlapping with the -10 box. Each element is typically recognized by a protein **factor** that either blocks or facilitates transcription. Note the vocabulary: an element is a segment of DNA, and a factor is a protein.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=258#h5p-36

Note: Eukaryotes typically do not use operons to co-regulate genes. You'll recall from the module on translation that prokaryotes use a ribosome binding site to bring the ribosome to an RNA, and the ribosome binding site positions the ribosome near the AUG start codon for translation initiation (**Figure 7A**). So it doesn't matter much whether the ribosome binding site is at the 5' end of the RNA, or somewhere in the middle: translation will always begin with the nearby AUG, and an RNA molecule can have multiple ribosome binding sites!

But eukaryotes do not generally use ribosome binding sites to attract the ribosome to the RNA. Instead, the

ribosome binds to the 5' cap (**Figure 7B**). Each RNA only has one 5' cap, so typically only one open reading frame can be translated per RNA (there are some exceptions to this general rule).

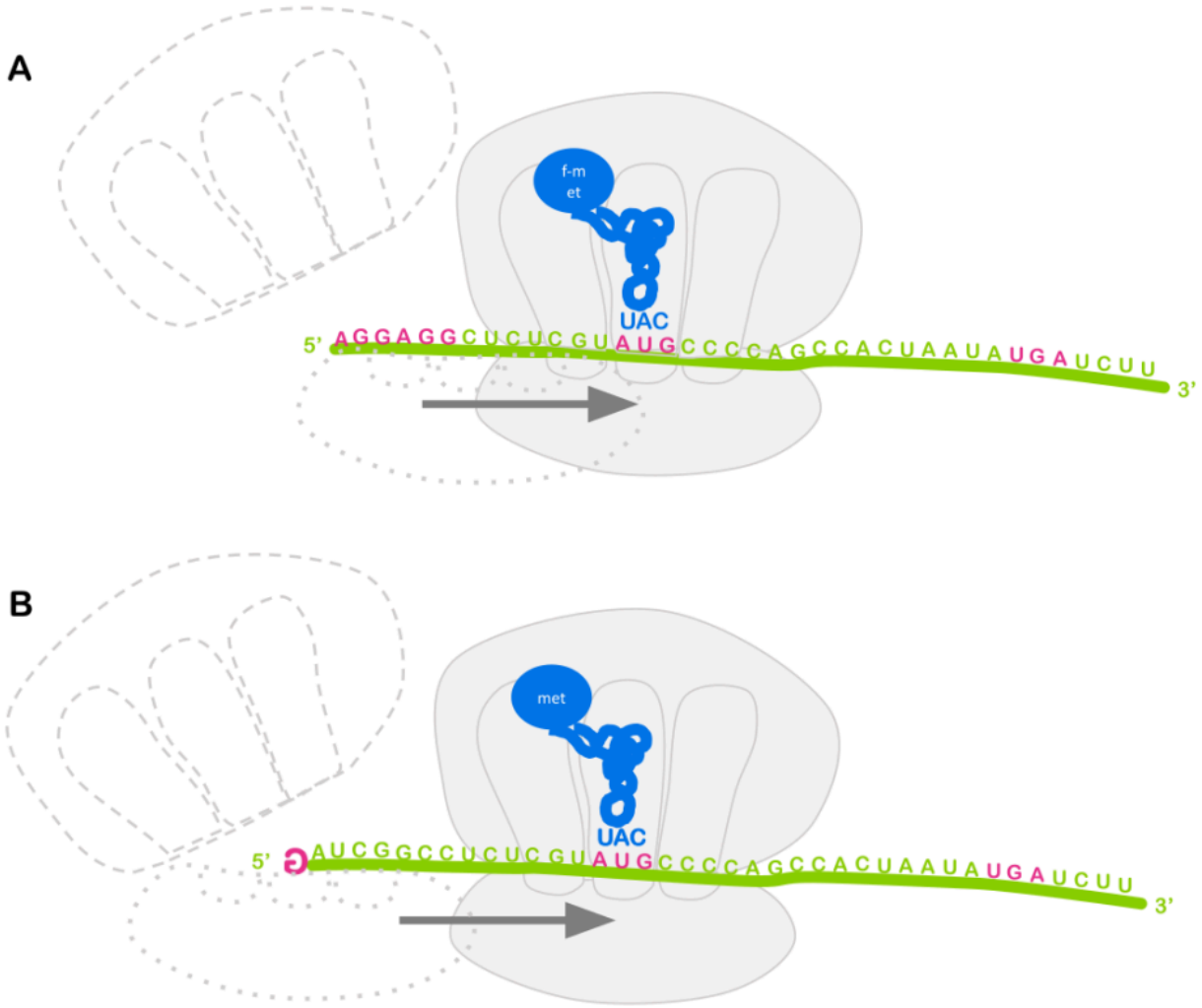


Figure 7. Translation initiation. In prokaryotes (A) the ribosome binding site (pink AGGAGG) positions the ribosome in place near the AUG start codon. In eukaryotes (B) the ribosome binds to the 5' cap and slides downstream on the RNA until it encounters a start codon.

Instead of using operons, eukaryotes use regulatory sequences found in multiple places in the genome. Co-regulated genes have similar elements in their regulatory promoters, which allow them to be activated (or repressed) simultaneously whenever a single transcription factor is present.

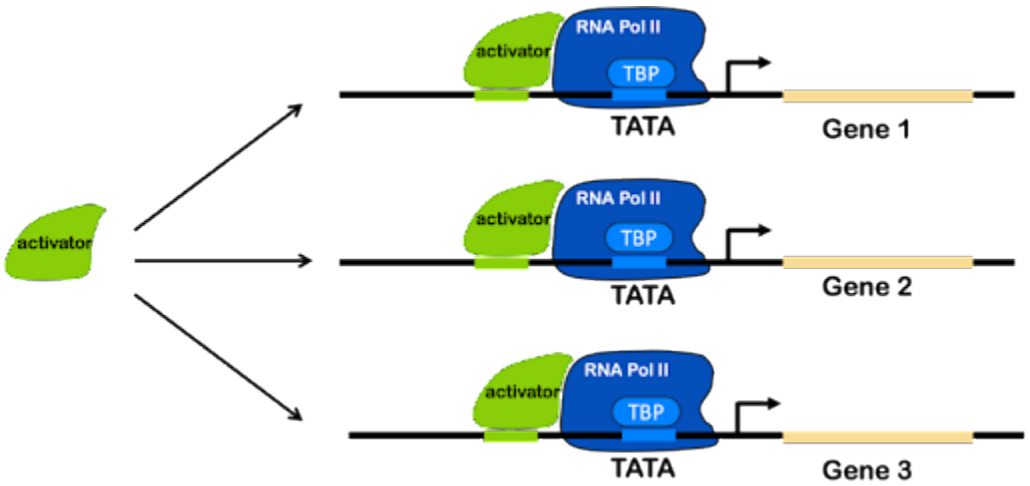


Figure 8. One activator controls many genes simultaneously by binding to similar DNA elements in the regulator regions for each gene.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=258#h5p-37

Media Attributions

- strong promoter weak promoter
- Regulation of transcription © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- DNA bending protein © CNX OpenStax is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Operon structure © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Translation initiation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- One activator © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

ONE PROMOTER CAN BE REGULATED BY BOTH ACTIVATORS AND REPRESSORS: THE LAC OPERON

One of the first- and best-studied systems of gene regulation is the lac operon in *E. coli.* In a series of experiments conducted in the late 1950's, Francois Jacob, Jacques Monod, and Arthur Pardee collaborated to determine how *E coli* lactose-metabolizing enzymes are regulated. These experiments are called the PaJaMa (or sometimes PaJaMo) experiments, a name created from the first letters of the researchers' last names. The specifics of the PaJaMa experiments are discussed later in this chapter.

E coli can use many different sugars as a food source although glucose is usually the most efficient option if it is available. Other sugars like **lactose** can be used as an energy source, but additional enzymes are needed for their metabolism. Lactose is a disaccharide comprising glucose and galactose monosaccharides **(Figure 9**).

In *E. coli,* metabolism of lactose begins with two proteins: lac permease, which is a transporter that brings lactose into the cell, and β-galactosidase, which breaks the glycosidic bond connecting the monosaccharides, producing glucose and galactose. Lac permease is encoded by the *lacY* gene, and β-galactosidase is encoded by the *lacZ* gene.

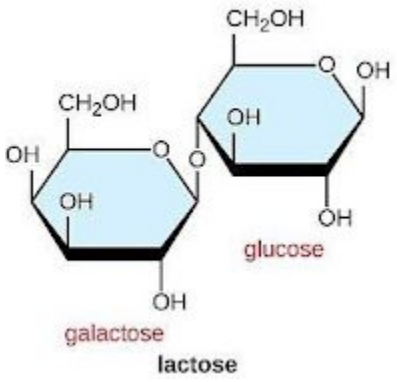


Figure 9. Structure of lactose.

The metabolism of lactose requires the proteins to work together, so the genes must be co-regulated. *LacZ*, *lacY*, and a third gene, *lacA*, are arranged in an operon. *LacA* encodes trans-acetylase, the role of which isn't clear in lactose metabolism.

The gene of the **lac operon** are grouped together under the control of the **lac promoter**, called ***lacP*** (**Figure 10**). The -10 and -35 sequences of the lac promoter are not a perfect match to consensus, so *lacP* is a weak promoter.

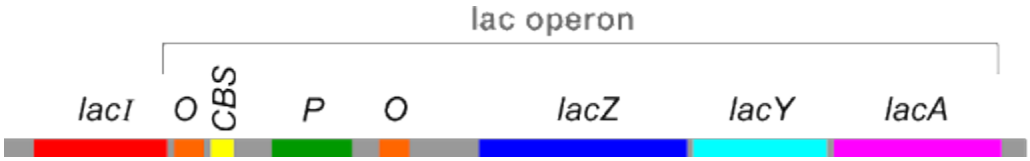


Figure 10. Diagram of a segment of an E. coli chromosome containing the lac operon, as well as the lacI coding region. The various genes and cis-elements are not drawn to scale.

Transcribing and translating these genes takes up a lot of energy, so these genes are only expressed when they are needed: i.e. when lactose is available but the preferred energy source, glucose, is not. To accomplish this control, transcription is regulated by two transcription factors: the **lac repressor** (encoded by the gene *lacI*, separate from the operon), and **catabolite activator protein**, or **CAP,** which is not shown in this image.

LacI is a lactose sensor: LacI blocks transcription if there's no lactose

The lac repressor, as the name suggests, represses expression – it turns the operon "off"unless lactose is present. It is a lactose sensor. The lac repressor protein binds to sites called **operators**, which surround the promoter and overlap with the +1 site of the operon. The operators are labeled "O" in **Figure 10.** When there is no lactose present, the repressor binds to the operators, blocking transcription. When there is lactose present, the repressor releases the operators and transcription can occur. This is shown in **Figure 11**.

Note that the repressor acts as both a dimer and a tetramer: two repressor polypeptides bind to each operator, and the operator-bound dimers can come together to form a tetramer that further blocks access of RNA polymerase to the promoter.

Be careful not to confuse the terms **operon** and **operator**! An *operon* is a single transcriptional unit that includes multiple genes, while the *operator* is a DNA element that is recognized by a repressor.

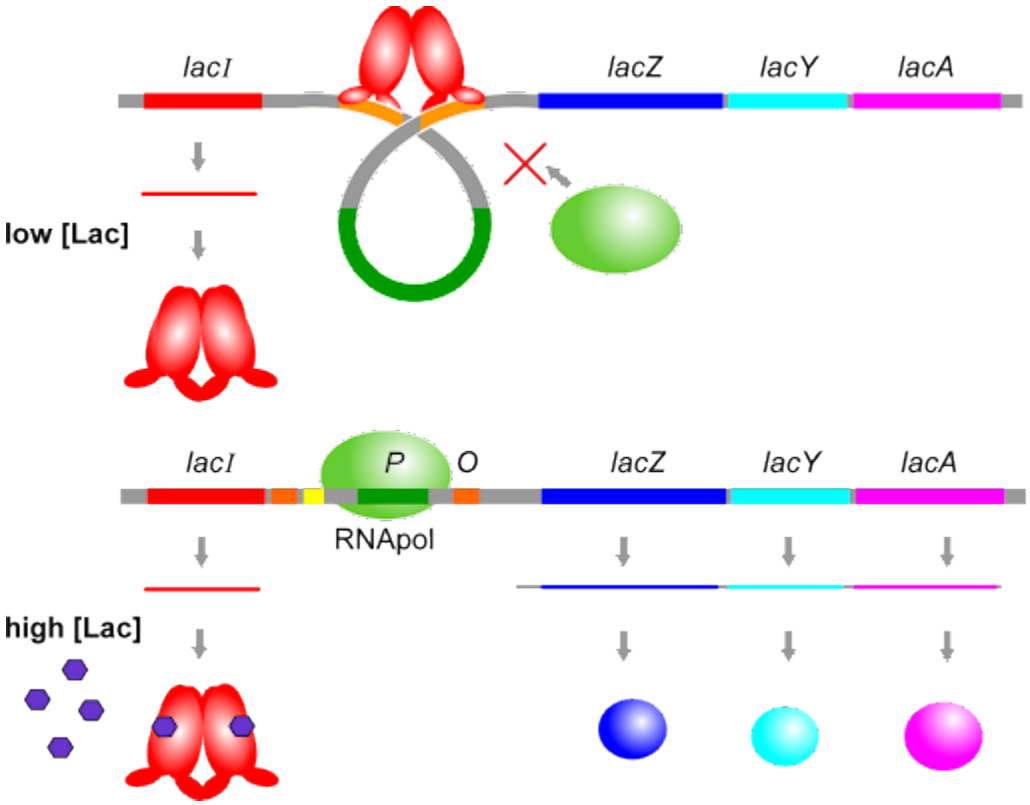


Figure 11. When lactose is not present, the lac repressor (red) binds to the lac operators. The repressor molecules on each operator can also bind to each other, forming a loop that blocks access of the polymerase to the promoter. When lactose is present, the lac repressor releases the operator, allowing the genes of the lac operon to be transcribed and translated.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=267#h5p-38

CAP is a glucose sensor: CAP activates transcription only if there's no glucose

The lac operon is not expressed at high levels when glucose is present, via a mechanism called **catabolite**

inhibition. But this is not negative regulation! Instead, the low expression is actually the absence of a positive regulator.

The lac operon can be negatively regulated (or turned off) by the action of the lac repressor. But the absence of the repressor is not enough to produce great numbers of lac mRNA. This is because the lac promoter is not a very good match to the consensus -10 and -35 sequences. In fact, the weak promoter shown back in **Figure 3** is actually the sequence of the lac promoter. The sigma factor and the RNA polymerase bind inefficiently to the imperfect sequence, which in turn leads to inefficient transcription of the operon.

In order to produce high amounts of the lac RNA, the operon needs an activator, CAP. CAP stands for catabolite activator protein. CAP binds to the CAP binding site (CBS in **Figure 10**) upstream of the -35 box. CAP also binds to RNA polymerase, stabilizing it on the promoter and increasing transcription.

Like the lac repressor, CAP is an allosteric protein that exists in both active and inactive conformations. CAP is not always active: CAP can only bind to DNA when it is complexed with cyclic AMP (cAMP). For this reason, CAP also sometimes goes by the name cAMP receptor protein, or CRP for short.

cAMP levels are high typically only when there is not much glucose in the environment, because low levels of glucose trigger an enzyme called adenylate cyclase to produce cAMP from ATP, as shown in **Figure 12**. So CAP only binds to the CBS to activate the lac operon when glucose is low. High levels of glucose inhibit the action of adenylate cyclase, inactivating CAP, so the operon will not be activated.

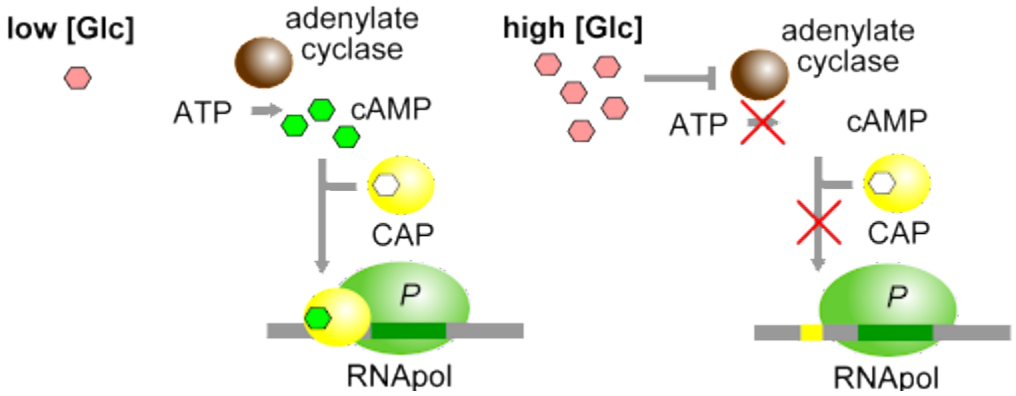


Figure 12. CAP, when bound to cAMP, helps RNApol to bind to the lac operon. cAMP is produced only when glucose [Glc] is low.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=267#h5p-39

CAP and the lac repressor work together to react to changing conditions

Remember, glucose is an optimal sugar source for *E.coli*, but lactose is acceptable if that's what is available. The lac repressor and CAP allow the cell to react to changing environmental conditions. Lac essentially acts as a lactose sensor, and CAP essentially acts as a glucose sensor, although neither factor binds directly to the sugar. The cell can therefore react to four separate combinations of environmental conditions: with glucose alone, with both glucose and lactose, with lactose alone, and neither.

When glucose alone is present, the repressor is bound to the operator and CAP is absent. This leads to no transcription.

When both glucose and lactose are present, the repressor is *not* bound, but neither is CAP. This leads to low (**basal**) levels of transcription since the operon is neither repressed nor activated.

When lactose alone is present, the repressor is *not* bound, and CAP is! This leads to high levels of transcription.

And when neither sugar is present, both the repressor and activator can bind. Who wins? The repressor wins: the operon is not transcribed. This is because of the way the activator and repressor function: with the repressor bound to the operator, there is no room for the polymerase to bind, even with the activator to assist. Two things cannot be in the same place at the same time.

This is summarized in the flow chart in **Figure 13**.

CAP and the lac repressor

Sugars present	Is the repressor bound?	Is CAP active?	Expression
Glucose	Yes	No	None
Glucose + Lactose	No	No	Basal (low)
Lactose	No	Yes	High
Neither glucose nor lactose	Yes	Yes (but doesn't matter, since repressor blocks transcription)	None

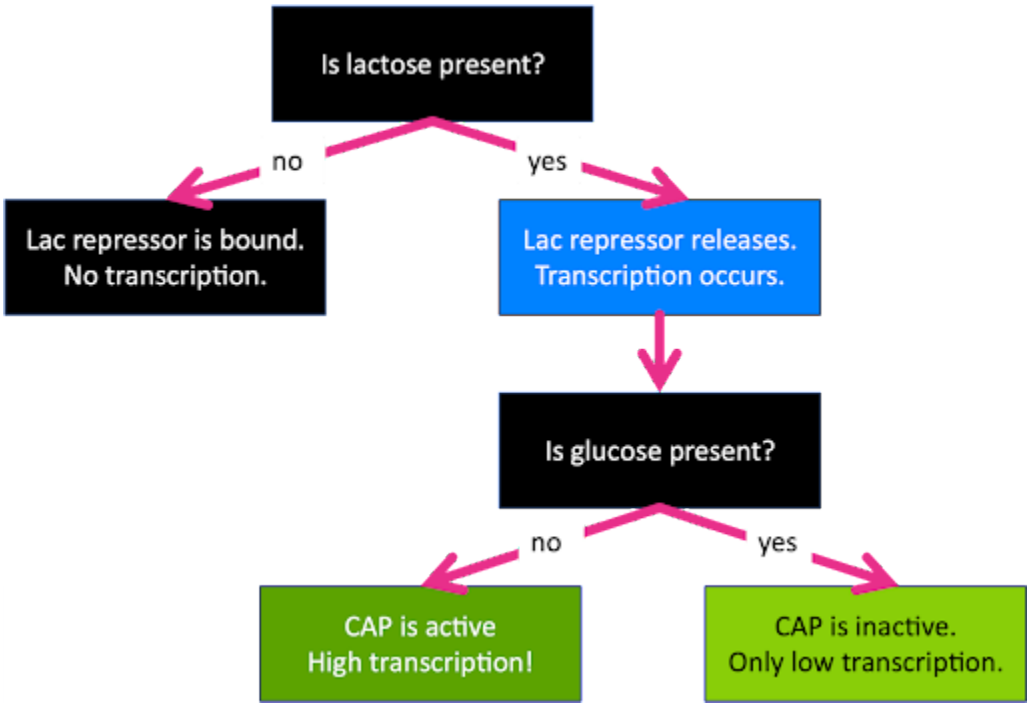


Figure 13. Lac operon flow chart. If lactose is not present, the repressor blocks transcription of the operon. If lactose is present, the repressor releases the operator and transcription can occur. If glucose is present, CAP is inactive, so only basal (low) levels of transcription occur. If glucose is not present, CAP activates transcription of the operon and high levels of transcription occur.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=267#h5p-40

Lac repressor mechanism of action

The lac repressor is a great example of the modular nature of many transcription factors: the protein has separate domains that perform different tasks. These can be seen in **Figure 14**. If you would like to explore a 3D animated version of this structure that you can drag and turn, you can find one at Proteopedia.

At the top of **Figure 14** you can see the DNA binding domain of the protein, bound to operator DNA (in yellow). The repressor binds as a dimer to the operator consensus sequence

5′TGGAATTGTGAGCGGATAACAATT3′[1]. One operator is located between the -10 and +1 sites of the gene, and the other is farther upstream, as shown in **Figure 11**. When the repressor binds to the operator, the repressor takes up enough physical space that there is no room for sigma/polymerase to bind: the bound repressor blocks transcription, just like the repressor example in **Figure 4**.

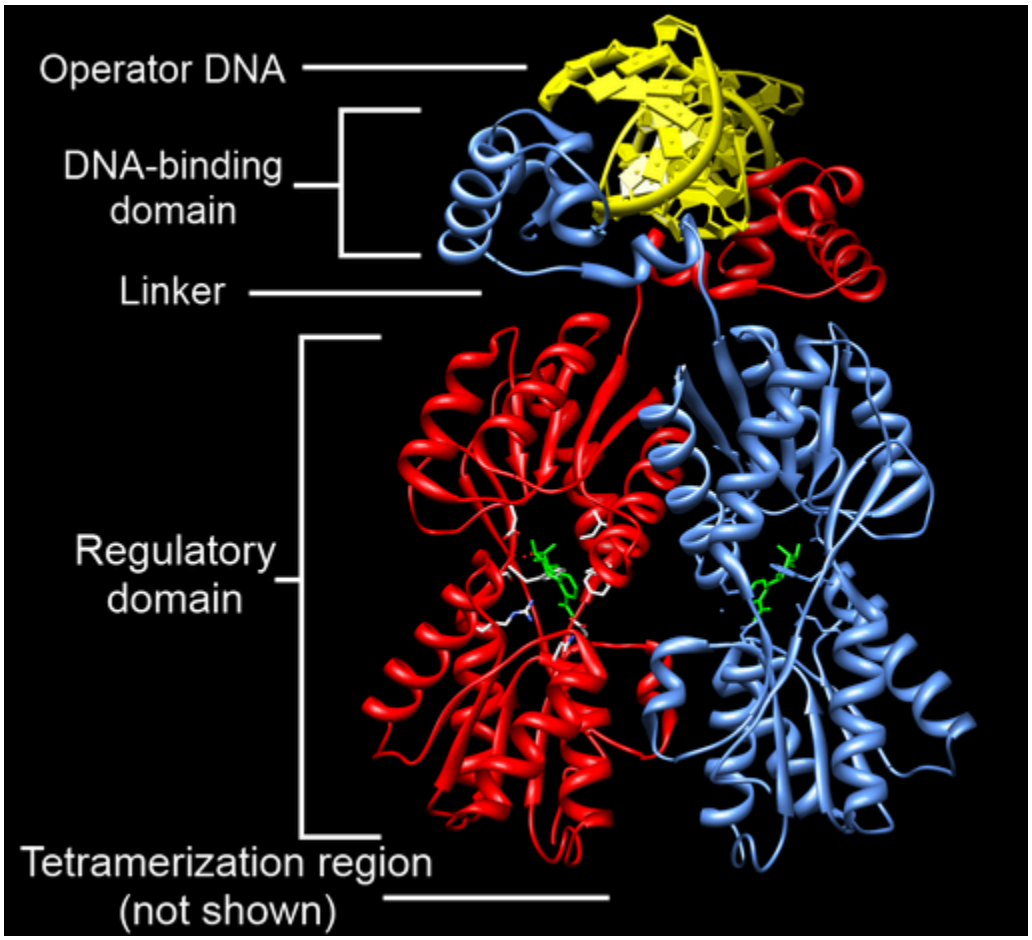


Figure 14. Structure of the Lac repressor. The lac repressor binds to the lac operator as a dimer. The DNA binding domain interacts with the lac operator (yellow). The regulatory domain can interact with allolactose. Binding to allolactose introduces a conformational change that leads to the release of the operator.

Although the lac repressor acts as a lactose sensor, it does so indirectly. The repressor only binds to the operator when there is no lactose present. When there *is* lactose present, a metabolite of lactose called **allolactose** binds to the repressor via the regulatory domain of the protein. The allolactose-bound repressor changes conformation and can no longer bind to the operator. This frees up the promoter, making transcription of the lac operon possible. Because the repressor exists in multiple conformations, it is called an **allosteric** protein.

1. Gilbert, W. & Maxam, A. The Nucleotide Sequence of the lac Operator. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3581–3584 (1973).

The presence of allolactose leads to increased expression of the operon, so allolactose is said to be an **inducer** of the operon. Note that other molecules can also induce the operon. For example, a molecule called IPTG is often used as an inducer of the lac promoter in a laboratory setting.

Although it is not shown in **Figure 14**, the repressor has an additional domain that allows two DNA-bound dimers to come together to form a tetramer. This allows the repressors on both operators to interact, forming a loop of the intervening sequence between the two operators. This larger complex likely further inhibits transcription. **Figure 15** shows the tetramerization between two DNA-bound dimers.

This modular nature of the repressor is characteristic of other transcription factors as well. It is common for a transcription factor to have separate DNA binding domains and regulatory domains.

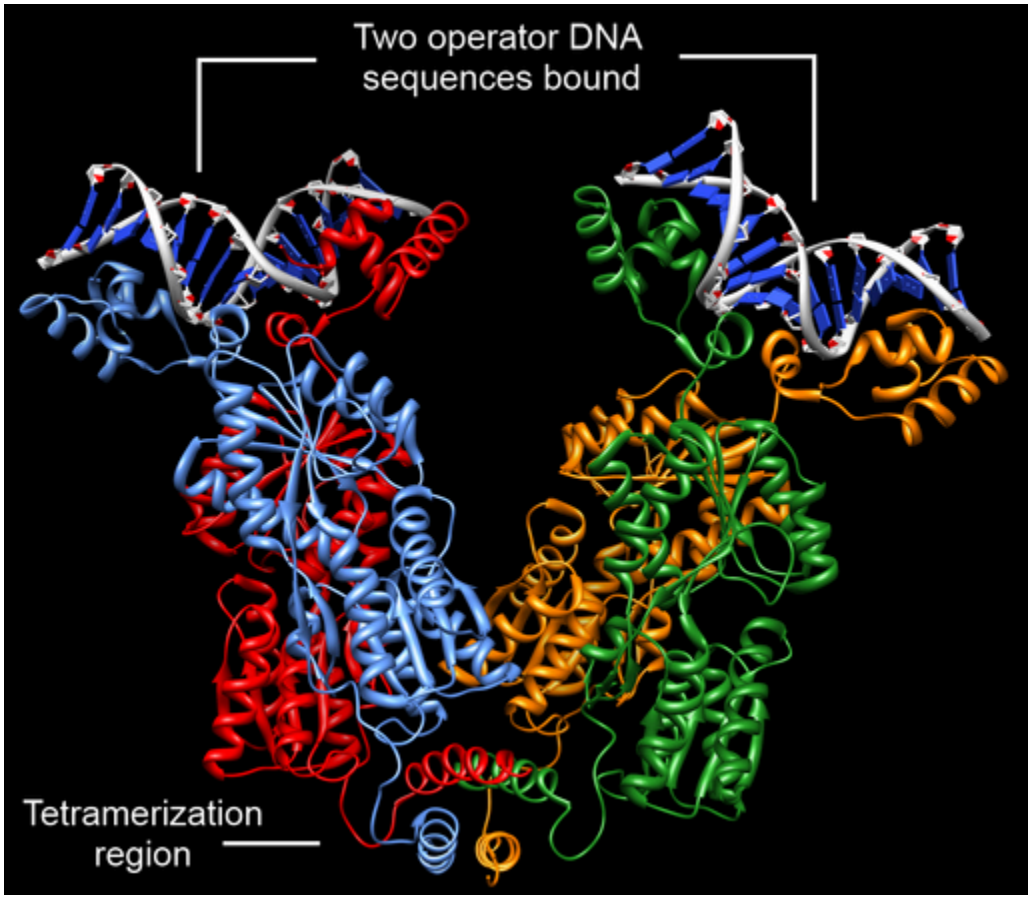


Figure 15. Two operator-bound dimers of lac repressor interact via the tetramerization domain. This interaction causes the looping of lac promoter DNA, blocking access of the RNA polymerase and preventing transcription. Image source: By SocratesJedi – Own work, CC BY-SA 3.0,

Genetic mutations of the lac operon

Experiments like the PaJaMa experiments allowed geneticists to understand the interactions between DNA

elements and protein factors. Such experiments typically rely on a collection of bacteria that have mutations in one or more of the components of the lac operon:

- **Null mutations in the structural genes**, called Z^- and Y^-. These mutations cannot produce β-galactosidase or lac permease depending on the mutation. (Null means "none" or "zero".)
- **Constitutive mutation of the operator**, O^c. These mutations are in the operator sequence that prevents the repressor from binding so that the proteins of the operon are always produced, regardless of environmental conditions. (The word constitutive means "always on".)
- **Loss of function mutation of the lac repressor**, called I^-. These may be null mutations with no protein produced, or they may have a mutation in the DNA binding domain of the repressor so that it cannot bind to the operator. These mutations, like the O^c, prevent the operon from being turned off, so it is also considered a constitutive mutation.
- **Gain of function mutations in the lac repressor**, called I^S. These so-called "super repressors" have a mutation in the allolactose binding domain, so that sugar-sensing domain does not function. These super repressors never release the operator, so the operon is never expressed.

These mutants are compared to **wild-type versions** of all genes, usually indicated with a superscript "+": I^+, O^+. Z^+, Y^+.

These experiments relied on the use of **partial diploid** bacteria: bacteria that carried an extra copy of the lac operon on an extrachromosomal piece of DNA called the F' episome.

F' is also called the fertility factor: copies of F' are transferred between bacteria via **conjugation**, or bacterial mating. For more information on conjugation, see this video from HHMI Biointeractive.

In this case, the extra copy of the lac operon allowed researchers to determine which genetic elements act in cis (that is, only upon the chromosome on which they reside) or trans (upon other chromosomes, too). Those mutations that act in cis are the mutations in DNA elements – loss of function in a promoter or operator, for example, affects transcription of adjacent genes, but has no effect on the transcription of genes on another chromosome. On the other hand, mutations that act in trans are mutations in a protein factor, having the potential to affect the transcription of both copies of the lac operon.

The extra copy of the chromosome also allowed these bacterial geneticists to determine whether a mutation was dominant or recessive. For example, the gain-of-function super repressor I^S is a dominant, trans-acting mutation: even if a second normally functioning repressor is present, the super-repressor can bind to both operators and repress transcription of both copies of the operon. The constitutive operator O^C is a dominant cis-acting mutation: it only affects the expression of adjacent genes, but the always-on expression of one copy of the operon means that the cell always has β-galactosidase and permease, regardless of whether the second copy of the operon is properly regulated.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=267#h5p-41

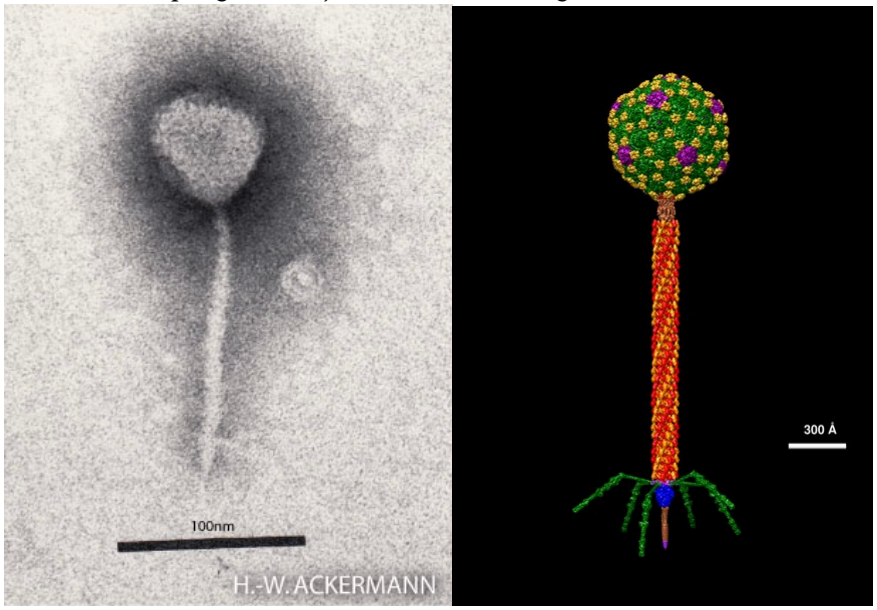
Media Attributions

- Structure of lactose © CNX OpenStax is licensed under a CC BY-SA (Attribution ShareAlike) license
- E. coli chromosome © Online Open Genetics (Nickle and Barrett-Ng) is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Lactose is not present © Original-Deyholos-CC:AN, via Online Open Genetics. is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- CAP bound to cAMP © Nickle and Barrett-Ng via. Online Open Genetics ().
- Lac operon flow chart © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- LacI Dimer Structure © SocratesJedi at English Wikipedia is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Bound Tetrameric Lac Repressor © SocratesJedi at English Wikipedia is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license

ONE TRANSCRIPTION FACTOR CAN BE BOTH AN ACTIVATOR AND REPRESSOR: THE LAMBDA PHAGE LIFE CYCLE

A second example of prokaryotic gene regulation is the switch between lytic and lysogenic growth in the bacteriophage λ (lambda). **Bacteriophages**, or phages for short, are viruses that infect bacteria: yes, even bacteria can catch a virus!

The structure of phage λ is shown in **Figure 16**. The left and center image show an electron micrograph and artist's interpretation, respectively. On the right in **Figure 16** is a cartoon of basic phage structure. Every phage has a "head" structure, which encases the genome. A phage also has a long sheath and a base plate with tail fibers extending out and down. The tail fibers help the phage attack and attach to a target bacterium. After attachment, the phage will inject its DNA through the shaft into the host cell.



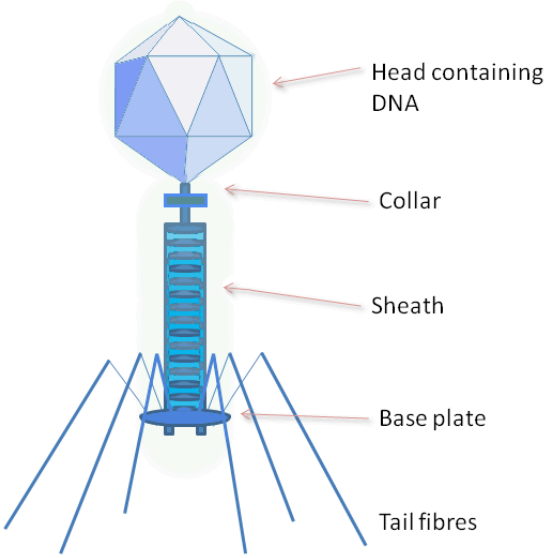


Figure 16. Bacteriophage lambda. Upper-Left: Electron micrograph of phage lambda. Upper-Right: interpretation of molecular data. Middle: cartoon of basic phage structure, showing a cube-shaped head that contains DNA, a cross-shaped collar beneath the head attached to a sheath that supports the collar and head and is held up by a base plate with tail fibers holding up the base plate.

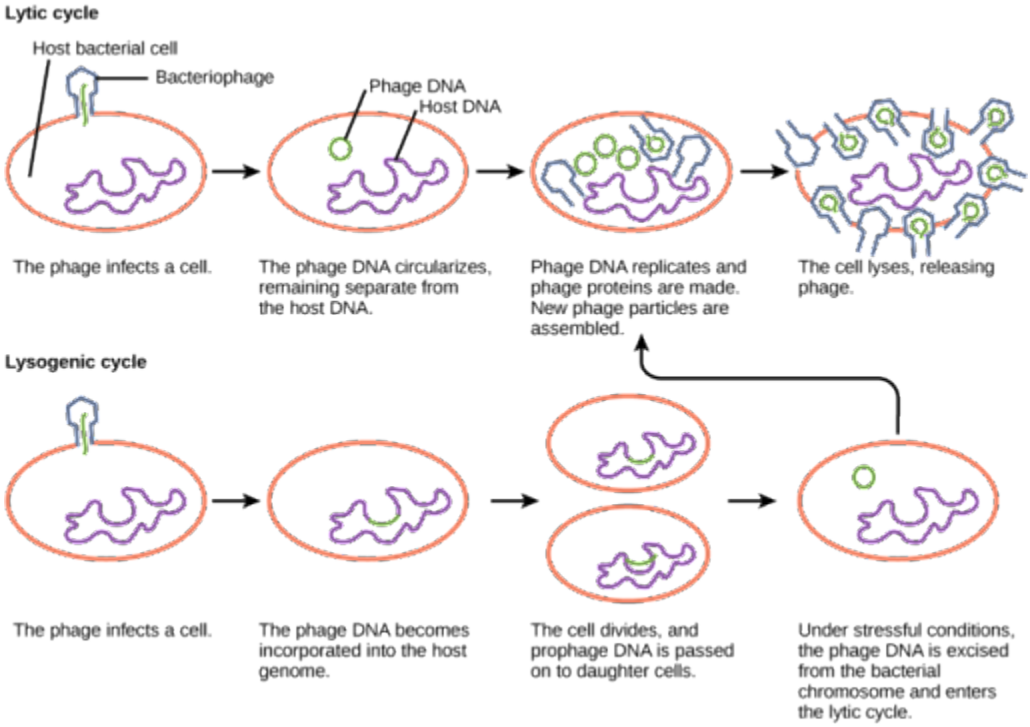


Figure 17. Lytic vs lysogenic growth. In the lytic cycle, phage DNA transcribed and translated to produce the RNA and protein needed to package new phage particles. The new phage particles are released by the lysis of the host cell. In the lysogenic cycle, the phage DNA is integrated into the host genome and replicated and passed to daughter cells as the host cell divides. Under conditions of stress, the phage DNA is cut out of the host genome and transitions to lytic growth and release of new phage particles.

After the viral genome enters the host cell, the bacteriophage may switch between either **lytic** or **lysogenic** growth. During lytic growth, shown on the top in **Figure 17**, the phage genome is replicated, transcribed, and translated by the host cell, which has been commandeered by the virus to make more viruses. Once new virus particles have been assembled, the cell is **lysed**, or broken open, releasing the new viral particles to infect other cells.

During lysogenic growth, shown on the bottom in **Figure 17**, the viral genome is injected into the cell and integrated into the host genome. The integrated viral DNA sequence is called the prophage. It is replicated and passed to daughter cells along with the host genome as the cell divides. The prophage can persist in this fashion indefinitely, but conditions of stress in the host cell can switch the phage to lytic growth.

The switch between lysogenic and lytic growth is controlled by expression three promoters in the phage genome, called P_L, P_R and P_{RM}, diagrammed in **Figure 18A**. P_L and P_R drive expression of genes needed for early stages of lytic growth, and P_{RM} drives expression of the **lambda repressor, λCI**. The promoters are oriented so they use different strands as template for transcription. In **Figure 18,** P_R uses the bottom strand as template and transcribes to the right in this image. P_{RM} and P_R use the top strand as template and transcribe to the left in this image.

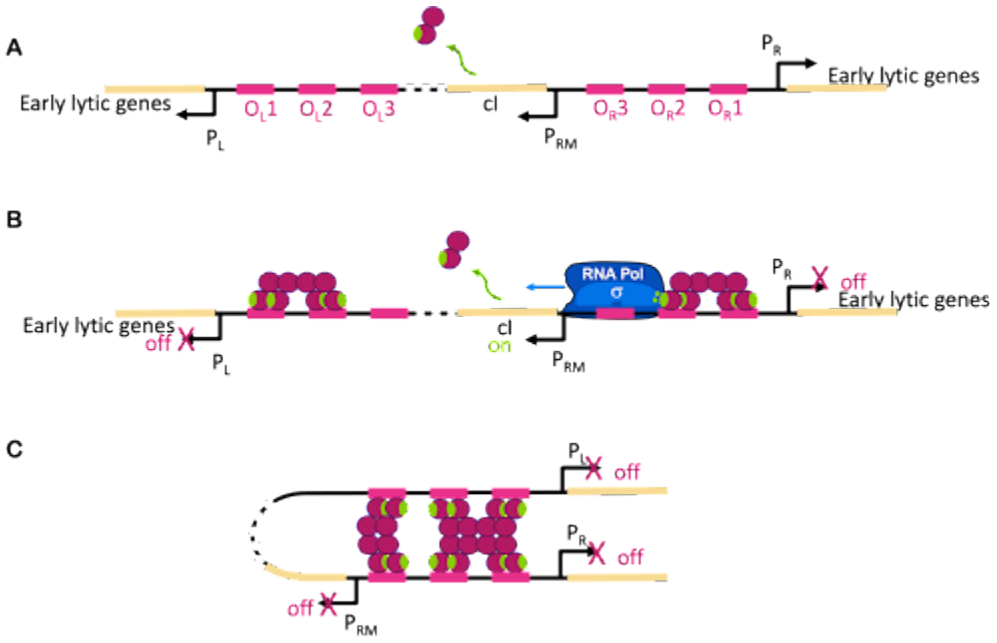


Figure 18. The lambda repressor (λCI) controls a switch between lysogenic and lytic growth of bacteriophage lambda.A. The switch between lytic growth and lysogeny depends on expression from three promoters, P_L, P_{RM}, and P_R. P_{RM} controls expression of the lambda repressor, (λCI). B. λCI dimers bind cooperatively to operators O_L1, O_L2, and O_R1, $O_2$2 to repress transcription from P_R and P_L. λCI also activates transcription from P_{RM} to positively autoregulate. C. As λCI accumulates in the cell, it reaches a great enough concentration that it can also bind to operators O_L3 and O_L3. This facilitates the looping of DNA around the operators, with higher-order complexes between λCI dimers holding the loop in place and negatively autoregulating expression from P_{RM}. This maintains λCI in a limited window of concentration in the cell.

λCI is called the lambda repressor, but this is not an accurate name: λCI can act as both a repressor *and* an activator!

Like the lac repressor, λCI has multiple domains: the N-terminal domain binds DNA. The N-terminal domain also has a patch on the surface that interacts with RNA polymerase. The C-terminal domain is a multimerization domain that allows each molecule of λC to interact with others. λCI functions as a dimer, but each dimer can interact with other dimers to form tetramers or octamers. You can drag, spin, and magnify the structure of λCI at Proteopedia.

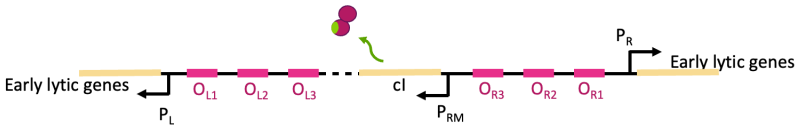
λCI binds to operators near these three promoters, but it does not bind equally well to all of them. Dimers of λCI bind best to O_L1 and O_R1, and this helps additional dimers bind to O_L2 and O_R2 in what is called cooperative binding. This binding blocks polymerase access to P_L and P_R and blocks transcription of lytic genes.

But wait, there is more! Although it is called the lambda repressor, the protein can also act as an activator for its own promoter. Part of the lambda repressor binds to RNA polymerase, recruiting it (helping it to bind) to

P_{RM}. The lambda repressor therefore acts as its own activator. This is **positive autoregulation**, where "auto" means it regulates its own gene. This is shown in **Figure 18B**.

But there is even more! λCI must be carefully regulated so there is enough protein to repress the lytic genes, but not so much that the phage can't undo this repression when needed. So λCI can act as its own repressor, too! At high concentrations of λCI, O_L3 and O_R3 will also be occupied by dimers of λCI. This leads to looping of the DNA as shown in **Figure 18C**, with λCI holding the loop in place and repressing transcription from P_{RM}, too. This is **negative autoregulation**.

Stress response in the host cell can trigger the switch to lytic growth and release of new phage particles. This happens through cleavage of the λCI protein, which de-represses P_R and P_L and leads to the production of lytic proteins. One protein produced from P_R is called cro. Cro binds to O_R3 and blocks transcription of the cI gene. In this way, either lytic genes or lysogenic genes are transcribed, but not both at the same time[1].



The switch between lytic and lysogenic growth is controlled by promoters P_L, P_R, and P_{RM}. P_L and P_R control genes important for early stages of lytic growth. P_{RM} drives expression of the lambda repressor gene, cI.

Animated gif showing the the Lambda Repressor switch in action.

So why does the λCI sometimes act as a repressor and sometimes an activator? It has to do with the geometry of binding. O_R1 and O_R2 are positioned over the -10 and -35 boxes for P_R. Remember, repressors can repress just by taking up space around a promoter. Two things cannot be in the same place at the same time, so if λCI is bound then RNA polymerase cannot. On the other hand, when λCI is positioned in O_R2 the RNA polymerase-interacting domain is positioned perfectly to stabilize a polymerase bound to P_{RM}. λCI can only activate transcription if it is properly aligned with the promoter.

1. Hochschild, A. & Lewis, M. The bacteriophage λ CI protein finds an asymmetric solution. *Curr. Opin. Struct. Biol.* **19**, 79–86 (2009).

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=274#h5p-42

Media Attributions

- Electron micrograph of Lambda © Photographer: Hans-Wolfgang Ackermann via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Molecular data of Lambda © Dr. Victor Padilla-Sanchez, PhD via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Phage structure, © GrahamColm at English Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Lytic vs lysogenic © LibreTexts is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Lambda repressor © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

ATTENUATION OF TRANSCRIPTION: THE TRP OPERON

Our last example of gene regulation in prokaryotes uses a mechanism of action that is unique to prokaryotes: transcriptional **attenuation**. In attenuation, instead of a full RNA being transcribed, a shortened (or attenuated) version is transcribed instead.

The trp operon is an example of attenuation. The **trp operon** includes five structural genes that are needed for the synthesis of the amino acid tryptophan. But the enzymes encoded by these genes are not needed if there are already high levels of tryptophan in the cell. The trp operon is negatively regulated by a repressor. But it uses attenuation as a second method of negative regulation**.**

Trp operon attenuation uses the action of the *ribosome* to terminate transcription early! That's pretty unusual since we think of translation happening after transcription. So let's look at how that's possible.

Recall two things from our discussion of transcription in prokaryotes:

1. Transcription and translation can happen simultaneously, with the 5' end of the RNA being translated even as the 3' end is still being transcribed.
2. Rho-independent terminators work because a hairpin structure forms in the RNA and causes the transcription machinery to dissociate, or fall apart.

The trp operon RNA has a "leader" region upstream of the first structural gene. The leader region has four important regions as shown in **Figure 19**. Regions 2, 3, and 4 are complementary and can base pair in various ways: Regions 2 and 3 can base pair, *or* regions 3 and 4 can base pair. When regions 3 and 4 base pair, they form a terminator structure and transcription ends prematurely – before the structural genes are transcribed! The terminator structure is called the **trp attenuator.**

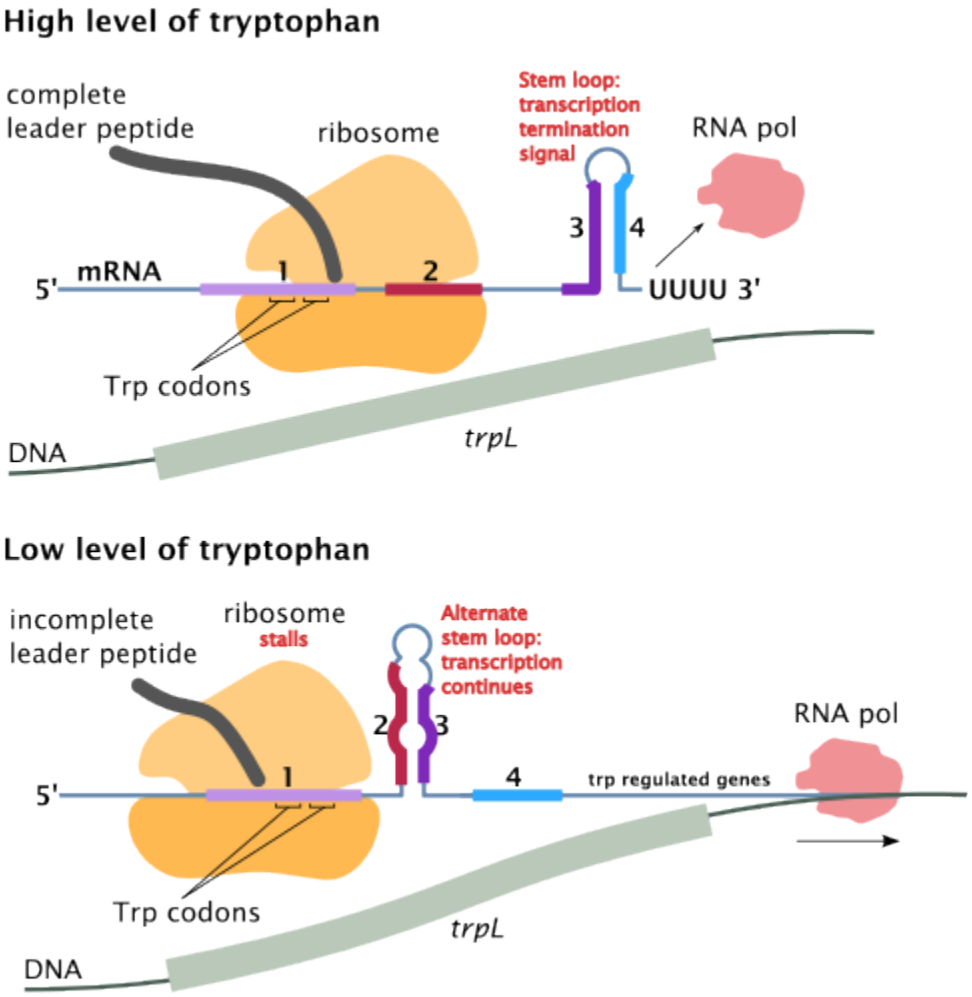


Figure 19. The trp operon is attenuated by a leader sequence (trpL) upstream of the structural genes. The leader RNA has four regions 1, 2, 3, and 4. When tryptophan levels in the cell are high (top), the leader sequence is translated, with the ribosome occupying sites 1 and 2. This allows regions 3 and 4 to base pair, forming a terminator hairpin.This results in a shortened, or attenuated, RNA. When tryptophan levels are low (bottom), tryptophan codons in the leader RNA are slow to recruit trp-tRNA, and the ribosome gets stuck at region 1. This allows regions 2 and 3 to base pair, leaving region 4 unpaired. No terminator forms, and transcription continues into the structural genes. Histidine

So what about region 1? Region 1 includes a short open reading frame with multiple tryptophan codons. When there is lots of tryptophan in the cell, the ribosome translates the 5' leader just fine. The presence of the ribosome ties up region 2, leaving 3 and 4 to pair and form the terminator. That means lots of tryptophan = short RNA. This is shown in the top of **Figure 19.**

When there is little tryptophan in the cell (**Figure 19**, bottom), the ribosome gets stuck at the tryptophan codons because there aren't enough tryptophan-charged tRNAs for translation to proceed. This parks the

ribosome squarely on top of region 1. Region 2 is freed up to pair with region 3, so the terminator hairpin does not form. Low tryptophan = long RNA = structural genes can be translated.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=277#h5p-43

Ribosome-mediated transcription attenuation is a mechanism of gene regulation that is unique to prokaryotes, because it depends on simultaneous transcription and translation. Ribosome-mediated attenuation does not happen in eukaryotes since transcription occurs in the eukaryotic nucleus and an RNA must be exported to the cytoplasm to be translated. The ribosome is not present when transcription is occurring.

In trp operon example, the ribosome mediates transcriptional termination, but there are other mechanisms of attenuation in bacteria as well. For example, in some versions of attenuation, the presence of a small molecule affects the formation of a terminator hairpin instead of the ribosome.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=277#h5p-44

Media Attributions

- Trp operon attenuation © Histidine via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

SUMMARY

Gene expression refers to whether a product is produced from a gene. Gene regulation refers to the control of gene expression, whether a gene is expressed under appropriate conditions.

Genes can be regulated at multiple levels, although this chapter focuses primarily on transcriptional regulation. Proteins called activators positively regulate gene expression by increasing the rate of transcription. Proteins called repressors negatively regulate gene expression by decreasing the rate of transcription.

Activators and repressors can work together on a single promoter to tightly control expression of a gene. The lac operon is an example of this, regulated by both an activator and a repressor. For proper gene regulation, a gene must be expressed at appropriate times and also turned off at appropriate times – it's not just a matter of being able to turn the gene on! Constitutive mutants of the operator – O^C result in an operon that is always on, but this is not properly regulated because it cannot be turned off when there is no lactose.

One single protein can be both an activator and a repressor, depending on where it binds in the genome. An example of this is the lambda repressor, which controls the switch between lytic and lysogenic growth of bacteriophage λ.

In prokaryotes, but not eukaryotes, the action of the ribosome during translation can affect whether or not a gene is fully transcribed. The trp operon is an example of that.

WRAP-UP QUESTIONS

Questions 1-5 are from *Online Open Genetics (Nickle and Barrett-Ng)*, Chapter 12[1].

1. List the mechanisms that can be used to regulate gene expression in eukaryotes.
2. For each of the following lac operon genotypes, would β-galactosidase be expressed?

a. I^+, O^+, Z^+, Y^+ (no glucose, no lactose)
b. I^+, O^+, Z^+, Y^+ (no glucose, high lactose)
c. I^+, O^+, Z^+, Y^+ (high glucose, no lactose)
d. I^+, O^+, Z^+, Y^+ (high glucose, high lactose)
e. I^+, O^+, Z^-, Y^+ (no glucose, no lactose)
f. I^+, O^+, Z^-, Y^+ (high glucose, high lactose)
g. I^+, O^+, Z^+, Y^- (high glucose, high lactose)
h. I^+, O^c, Z^+, Y^+ (no glucose, no lactose)
i. I^+, O^c, Z^+, Y^+ (no glucose, high lactose)
j. I^+, O^c, Z^+, Y^+ (high glucose, no lactose)
k. I^+, O^c, Z^+, Y^+ (high glucose, high lactose)
l. I^-, O^+, Z^+, Y^+ (no glucose, no lactose)
m. I^-, O^+, Z^+, Y^+ (no glucose, high lactose)
n. I^-, O^+, Z^+, Y^+ (high glucose, no lactose)
o. I^-, O^+, Z^+, Y^+ (high glucose, high lactose)
p. I^s, O^+, Z^+, Y^+ (no glucose, no lactose)
q. I^s, O^+, Z^+, Y^+ (no glucose, high lactose)
r. I^s, O^+, Z^+, Y^+ (high glucose, no lactose)
s. I^s, O^+, Z^+, Y^+ (high glucose, high lactose)

1. Nickle and Barrette-Ng. Open Online Genetics. in *Open Online Genetics* (2016).

3. In the *E. coli* strains listed below, some genes are present on both the chromosome and the extrachromosomal F-factor episome. The genotypes of the chromosome and episome are separated by a slash. Would β-galactosidase be expressed in each of these strains? All of the strains are grown in media that lack glucose.

a. $I^+, O^+, Z^+, Y^+ / O^-, Z^-, Y^-$ (high lactose)
b. $I^+, O^+, Z^+, Y^+ / O^-, Z^-, Y^-$ (no lactose)
c. $I^+, O^+, Z^-, Y^+ / O^-, Z^+, Y^+$ (high lactose)
d. $I^+, O^+, Z^-, Y^+ / O^-, Z^+, Y^+$ (no lactose)
e. $I^+, O^+, Z^-, Y^+ / I^-, O^+, Z^+, Y^+$ (high lactose)
f. $I^+, O^+, Z^-, Y^+ / I^-, O^+, Z^+, Y^+$ (no lactose)
g. $I^-, O^+, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (high lactose)
h. $I^-, O^+, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (no lactose)
i. $I^+, O^c, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (high lactose)
j. $I^+, O^c, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (no lactose)
k. $I^+, O^+, Z^-, Y^+ / I^+, O^c, Z^+, Y^+$ (high lactose)
l. $I^+, O^+, Z^-, Y^+ / I^+, O^c, Z^+, Y^+$ (no lactose)
m. $I^+, O^+, Z^-, Y^+ / I^s, O^+, Z^+, Y^+$ (high lactose)
n. $I^+, O^+, Z^-, Y^+ / I^s, O^+, Z^+, Y^+$ (no lactose)
o. $I^s, O^+, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (high lactose)
p. $I^s, O^+, Z^+, Y^+ / I^+, O^+, Z^-, Y^+$ (no lactose)

4. What would be the effect of the following loss-of-function mutations on the expression of the lac operon?

a. loss-of-function of adenylate cyclase
b. loss of DNA binding ability of CAP
c. loss of cAMP binding ability of CAP
d. mutation of CAP binding site (CBS) *cis*-element so that CAP could not bind

5. How are eukaryotic and prokaryotic gene regulation systems similar? How are they different?

6. The trp operon leader has multiple tryptophan codons in a row.

a. What do you think would be the impact to transcription of the trp operon, if a mutation changed those codons from UGG to UGA?
b. What do you think would be the impact to transcription of the trp operon, if a mutation changed the tryptophan codons from UGG to UGC? Explain your reasoning.

PART IX

EUKARYOTIC GENE REGULATION IN ACTION: EXAMPLES FROM DEVELOPMENT

Learning Objectives

1. Define homeotic mutation, differentiation, enhancer, enhanceosome, silencer, insulator, TAD, maternal effect gene, Hox gene.
2. Explain how multiple enhancers can work to generate complex patterns of gene expression, using *eve* and *Shh* regulation as examples.
3. Describe mechanisms by which chromatin structure can influence gene expression
4. Describe the difference between a forward and reverse genetic screen.
5. Describe how maternal effect, gap, pair-rule, segment polarity, and segment identity genes interact to direct body patterns in the developing embryo.
6. Recognize that proper regulation of a gene doesn't just mean the gene is turned on! It means that it is only turned on under proper conditions.
7. Recognize that mutations in non-coding regulatory sequences can have profound effects on phenotype.

Introduction: How does a body get built? Homeotic mutations rearrange body structures

In the 1970's, researchers like Edward Lewis were interested in mutations that resulted in a rearrangement of body structures in *Drosophila,* where one body part was replaced with another. The mutated genes were called **homeotic** genes. One such example was a mutation in the gene *ultrabithorax* (*Ubx). Ubx* mutants had an extra set of wings in place of the halteres, which in wild type flies are short structures important for stability

(**Figure 1a**). Other mutants were missing expected body parts in particular segments. One mutant even had legs growing where antennae should be (**Figure 1b**)!

Figure 1a. A fruit fly with an extra set of wings.

Figure 1b. Antennapedia – A fruit fly with legs growing where antenna are expected.

Around the same time, geneticists Eric Wieschaus and Christiane Nüsslein-Volhard were beginning to look at developmental mutant phenotypes, too. However, they were not interested in mutant phenotypes in the adult fly. They reasoned that the genes important for body patterning were likely to be so important that mutations would be embryonically lethal. So, instead, Wieschaus and Nüsslein-Volhard looked for mutant phenotypes in the Drosophila embryo.

About 24 hours after fertilization of the egg, a wild-type Drosophila embryo looks like **Figure 2**, with visible segments. In this image, the anterior (head) part of the embryo is to the left, and the posterior (tail) is to the right. You can watch a video of the early stages of *Drosophila development* on YouTube here and here too. You can see that each segment is slightly different from the others. As development progresses, the segments each

eventually assume a different identity. For example, one segment will produce antennae, one will produce a wing, etc.

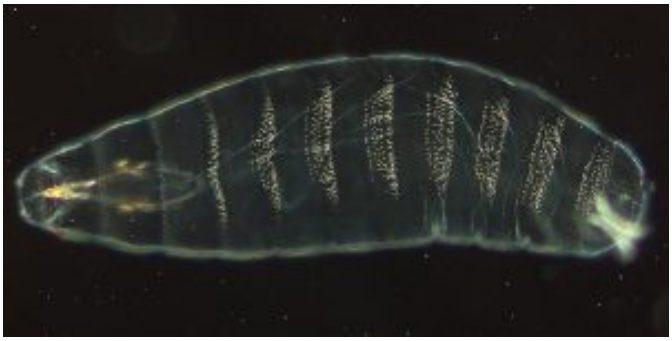


Figure 2. Drosophila embryos have a segmented appearance. The light-colored spots are denticles, short projections that extend from the ventral side of the embryo. Each segment has a slightly different denticle pattern.

How does a single-cell zygote develop into a multicellular organism, with specialized cells, distinct tissues, and organ systems? And how do those processes go wrong in the homeotic mutants? The differences in cell and tissue types largely come down to differences in gene regulation.

In this module, we look at how genes are regulated in eukaryotes, using embryonic development in animals as an example. We will review the basics of transcription and gene regulation, and then dive in to see how gene regulation drives cell differentiation during embryonic development. We'll look at how multiple enhancer elements can regulate gene function, using the *Drosophila* gene *eve* as an example.

Media Attributions

REVIEW OF TRANSCRIPTIONAL REGULATION

In the Overview of Gene Regulation chapter, we saw examples of the levels at which genes can be regulated, from accessing the DNA, transcription of a gene, processing of RNA, and protein synthesis. We ended the chapter with examples of how transcription factors – both activators and repressors – work together to control the expression of prokaryotic genes. As in prokaryotes, eukaryotes use a combination of activators and repressors to control gene expression according to the needs of the cell, although this process is usually much more complicated than in prokaryotes. But there are some other differences as well.

In contrast to prokaryotic genes that might be controlled by one or two regulatory transcription factors, eukaryotic gene regulation is much more complicated. A reminder that the word **factor** refers to a protein, and the word **element** refers to a DNA segment.

Remember that prokaryotes can use operons to co-regulate genes in one long RNA transcript with multiple internal ribosomal binding sites and multiple coding sequences. However, this is not possible in eukaryotes where the ribosome binds to the 5' cap. So instead, eukaryotes co-regulate genes by reusing the binding sites for a specific transcription factor in multiple places in the genome, allowing those genes to be co-regulated. These mechanisms are compared in **Figure 3**, reprinted from the Overview of Gene Regulation chapter.

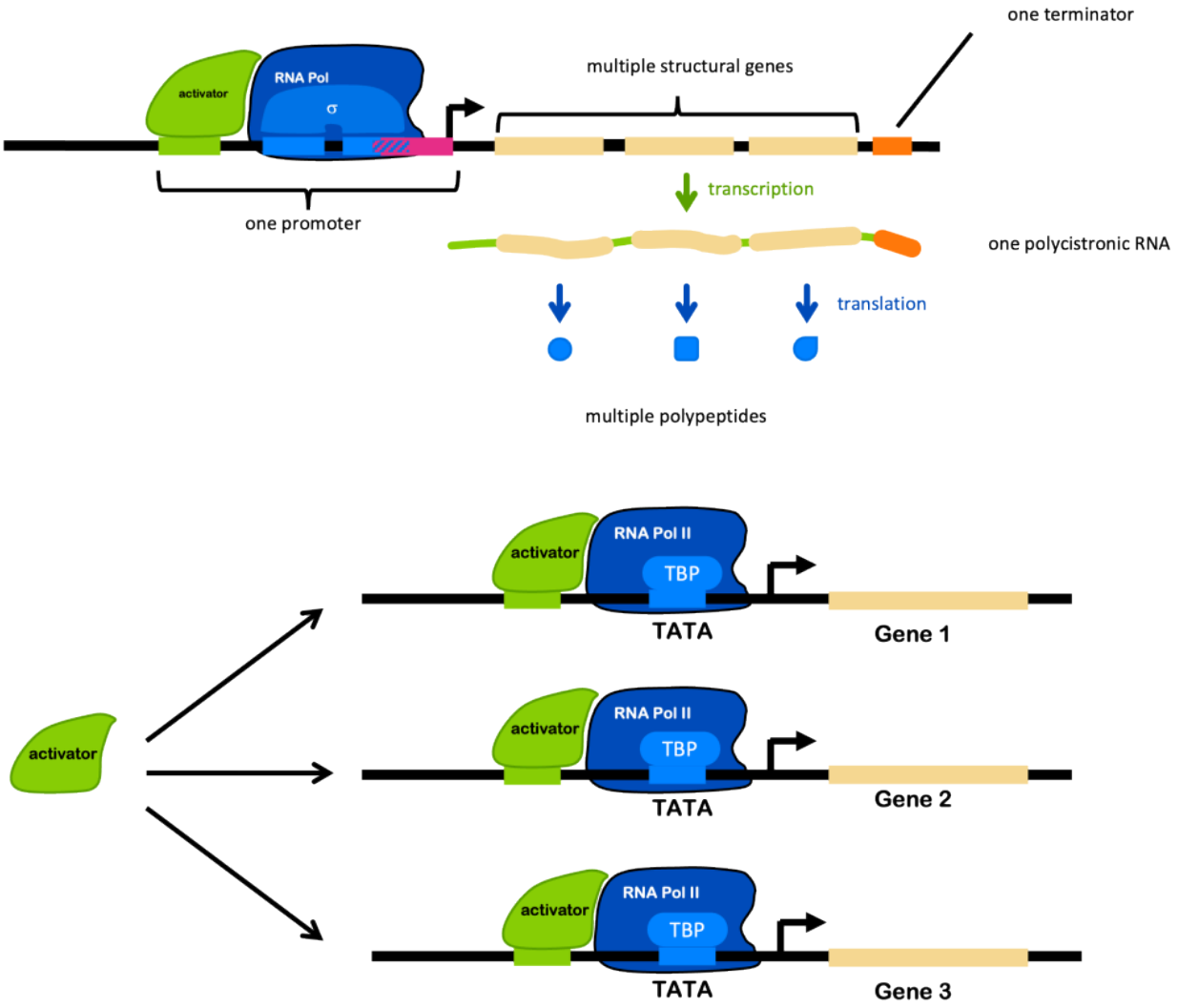


Figure 3. Coregulation of genes happens by different mechanisms in eukaryotes and prokaryotes.

An example: in humans and other mammals, many genes are turned on in response to the hormone estrogen. Those genes have an **estrogen response element** in the regulatory region of their promoter, which is bound by the estrogen receptors to activate transcription. The estrogen response element would be represented by the green DNA element in the **Figure 3** (bottom panel). The green activator in the image represents the estrogen receptor, which acts as a transcription factor.

Enhancer elements regulate transcription in eukaryotes

A eukaryotic gene usually has a core promoter directly upstream of the gene (**Figure 4**). The core promoter is bound by the **general transcription factors**, which participate in the transcription of all genes transcribed by a polymerase. As discussed in the Transcription module, these are often named for the RNA polymerase that

they assist. For example, the general transcription factors that work with RNA Polymerase II are named TFII_, distinguished from one another by letters.

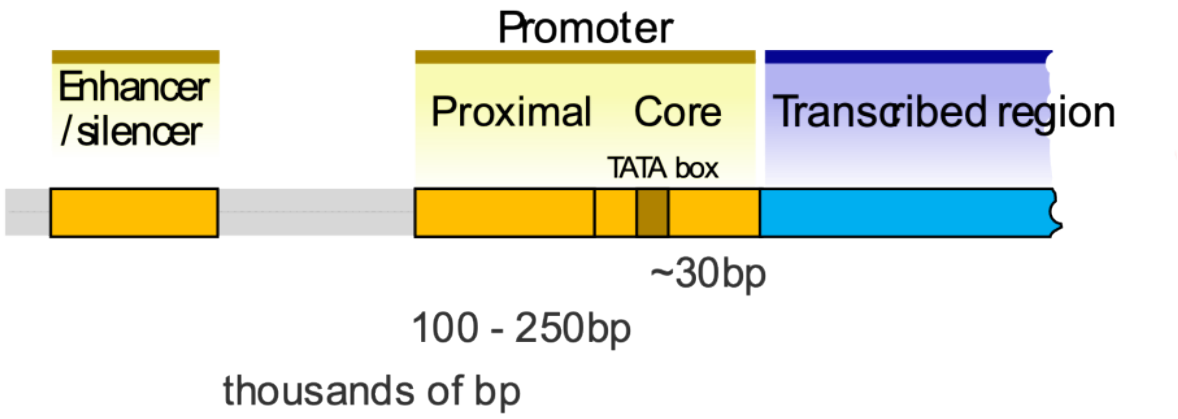


Figure 4. Eukaryotic RNA Pol II promoters consist of a core promoter, proximal elements, and more distant enhancers or silencers.

For genes transcribed by RNA Polymerase II, which includes the mRNAs to be translated into protein, the most recognizable feature of the core promoter is the TATA box. The TATA box is bound by transcription factor TFIID in an early stage of transcription initiation. However, additional elements may be present as well, including BRE (TFIIB recognition element), Inr (Initiator), and other elements like the CAAT box and GC box (named for the sequence of bases in the element). Many promoters don't even have a TATA box!

In prokaryotes, a strong core promoter (-10 and -35 box elements, for example) is sufficient to recruit polymerase and drive expression of an adjacent gene. But in eukaryotes, a core promoter and general transcription factors are not sufficient to drive the expression of eukaryotic genes. Eukaryotic genes require additional regulatory elements. These can include nearby **proximal promoter elements**, which are bound by **specific transcription factors**, which are so-named because they are specific to individual genes (or families of genes). They also include **distal** elements like **enhancers** and **silencers**, discussed more in the next section.

These distal elements can be thousands of base pairs from the transcriptional start site, but they may be brought into proximity to the core promoter through the bending of the DNA, as shown in **Figure 5,** reprinted from the Transcription and Overview of Gene Regulation chapters.

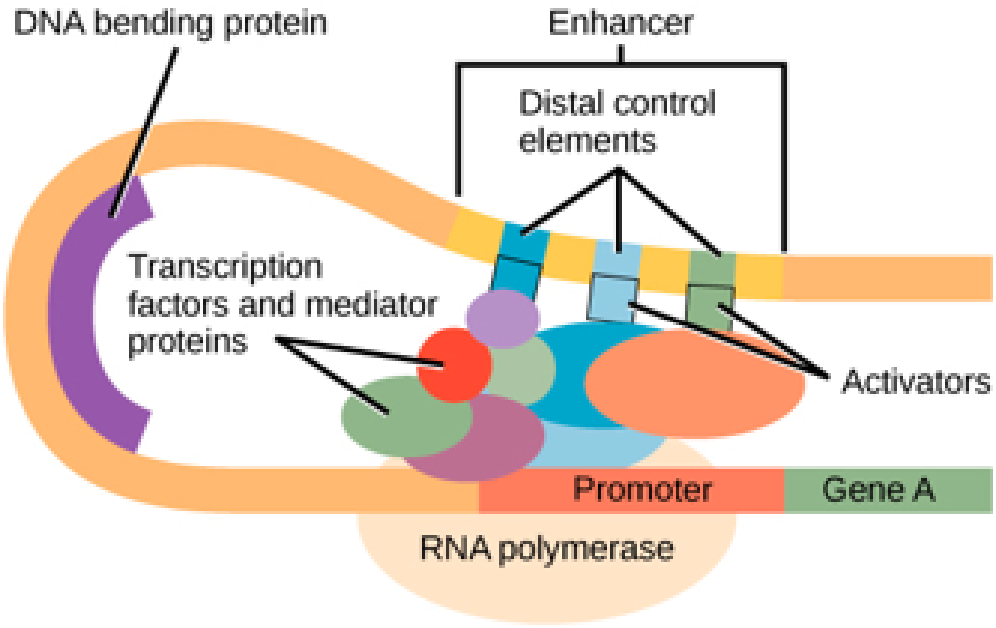


Figure 5. DNA is flexible and can fold or loop to bring distal control elements into close proximity to the promoter in three-dimensional space. This is helped by the action of DNA bending proteins and the mediator complex, which acts as a scaffold between proteins bound to the core promoter and distal control elements. Different genes have different distal elements, allowing for gene expression under different conditions.

As suggested by the names, enhancers typically function to enhance the transcription of a gene when bound by appropriate factors. At the same time, silencers reduce the transcription of a gene when bound by appropriate factors. A single enhancer or silencer typically has binding sites for multiple factors.

Figure 5 shows three activator-binding elements in one enhancer, but enhancers can include dozens of binding sites for both activators *and* repressors. The repressors may block the assembly of the polymerase, the mediator, or even other transcription factors. The factors that bind cooperatively to an enhancer are collectively called the **enhanceosome**.

For some genes, regulatory promoters and enhancers may be pretty straightforward: for example, a gene that is always expressed in one particular cell type (and only that cell type) might have a small number of regulatory elements that control its expression.

But other genes are expressed in complex patterns: for example, they might be needed in multiple tissues (but not all tissues) throughout the adult organism and/or needed at particular times during development. Such genes may have multiple enhancers that control expression. Usually, only one enhancer needs to be active for the gene to be transcribed, but each enhancer might respond to different conditions. In this way, enhancers can act like a network of switches so that a gene might be expressed under condition A, condition B, *or* condition C.

A well-studied example of this is the *eve* gene from the fruit fly *Drosophila melanogaster*. The *eve* gene is

expressed in multiple locations along the length of the embryo, as shown in Figures 6 and 7. **Figure 6** [1]shows an example of an embryo that has been stained for eve mRNA. The anterior of the embryo is oriented to the left, the posterior is to the right, the dorsal side is on top, and ventral on bottom. Purple marks where eve RNA is produced.

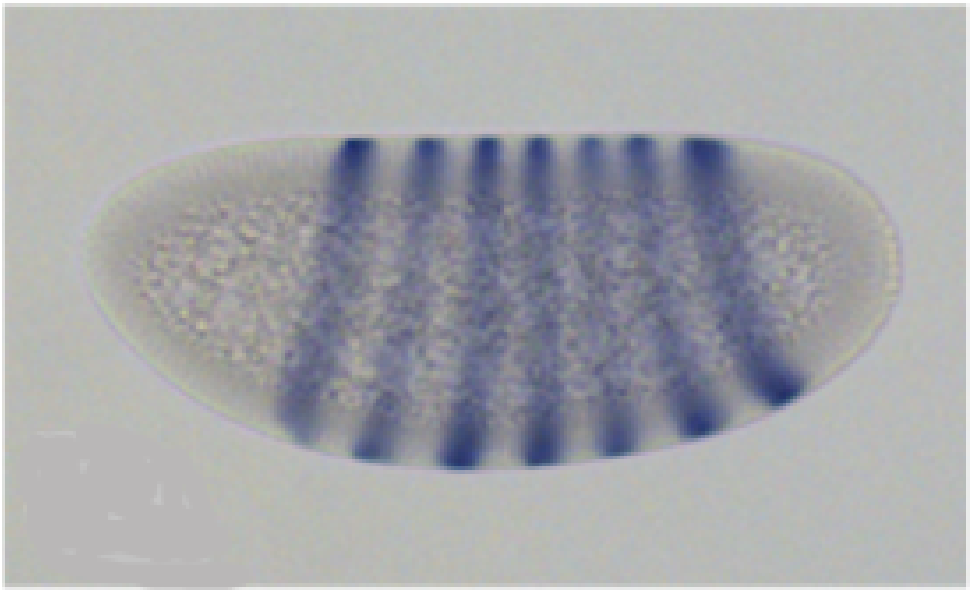


Figure 6. A fruit fly embryo stained for expression of the eve gene. The Eve gene is expressed in seven stripes along the length of the embryo from head to tail.

Figure 7. Diagram of the genomic region around the eve gene. The eve gene (blue) is regulated by five enhancers (green), each of which drives expression in one or two "stripes" along the anterior-posterior axis of the embryo. The enhancers are numbered according to the stripe they control.

In **Figure 7**, the genomic region around the *eve* gene is diagrammed with the transcriptional start site indicated

1. Modified from Peterson et al, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004688

by a right-facing arrow and the enhancers numbered according to the stripe(s) that each controls. The *eve* gene has five enhancers, each of which "turns on" expression in one or two stripe-like regions of a fruit fly embryo.

The mechanism for their activation is discussed more in detail in later sections. But one interesting thing to note is that the enhancers are found both upstream and downstream of the gene. This is typical of many eukaryotic genes, where enhancers do not need to be directly proximal to the core promoter due to the looping of DNA during transcriptional activation. Sometimes enhancers can be in introns. There are even rare examples of enhancers within exons for other genes!

Later sections of this chapter look at additional examples of how enhancers work in combination to drive complex patterns of gene expression in a multicellular organism. We will look at genes that, like *eve,* are expressed during embryonic development. We will look first at the expression of genes important for body patterning in *Drosophila melanogaster* and then at related genes in other organisms (including humans).

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=857#h5p-102

Media Attributions

- Figure 3 EG Regulation Operon vs eukaryotic coregulation 2 © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 4 EG Regulation © Thomas Shafee is licensed under a CC BY (Attribution) license
- Figure 5 EG Regulation © CNX OpenStax (Biology) is licensed under a CC BY (Attribution) license
- Figure 6 EG Regulation © Modified from Peterson et al is licensed under a CC BY (Attribution) license
- Figure 7 EG Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

THREE-DIMENSIONAL STRUCTURES MATTER

If enhancers are separated from their genes by, in some cases, a million base pairs, and if enhancers may be upstream or downstream of a gene's transcriptional start site, how do enhancers specifically enhance one gene over another nearby gene?

The rules for this quirk in enhancers are still poorly understood. Still, they are likely strongly influenced by the three-dimensional chromatin structures that assemble in the nucleus of a cell. Although individual chromosomes are not visible under the light microscope except when they are condensed during mitosis, the nucleus of the cell is still highly organized, and the looping of DNA that occurs to regulate transcription is carefully orchestrated. The organization has been described as chromosome "neighborhoods," with regions of the genome collected in **topologically associated domains**, or TADs. The word **topology** refers to shape and structure. For example, a topological map or topo map shows information about the hills and valleys of a region, so this is a fancy way of saying "associated in 3D space".

Figure 7 shows how those TADs might be organized. Panel A shows a cartoon diagram of how chromosomes tend to stay partitioned in their territory in the nucleus, with loops of DNA in proximity to each other, as shown in Panel B. Those loops allow interactions between genes and their regulatory elements. The dotted lines in Panel C show which DNA elements are brought together through folding of the DNA.

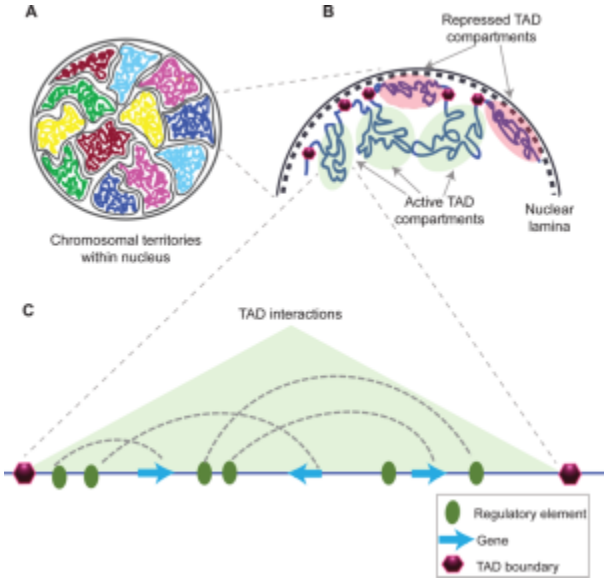


Figure 7. TADs are regions of a chromosome that are associated in three-dimensional space.

DNA elements called **insulators** form barriers between genes and nearby unrelated enhancers**.** A simple model for how insulators function is shown in **Figure 8**. Looping between the enhancer and the adjacent genes can occur, but the insulators form a boundary beyond which the enhancer cannot act. Insulator sequences are enriched at the boundaries between topologically associated domains and are bound by proteins like CTCF, which helps block the cross-talk of enhanceosomes beyond TAD boundaries.[1]

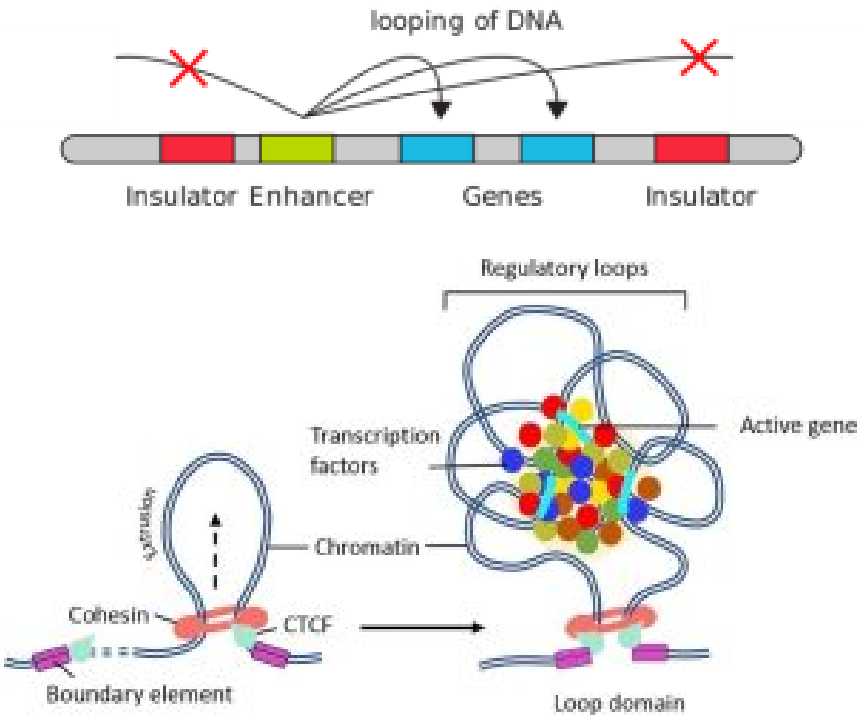


Figure 8. Insulators block enhancers from looping to distal promoters. (Left) Looping between the enhancer and the adjacent genes can occur, but the insulators form a boundary beyond which the enhancer cannot act. (Right) Insulator/boundary elements may be bound by proteins like CTCF. A cohesin ring is wide enough to allow a loop of DNA to extend through, but the loop cannot extend past the CTCF boundary elements. An enhanceosome can assemble with transcription factors binding to loops of DNA brought together in three dimensional space.

1. Pongubala, J. M. R. & Murre, C. Spatial Organization of Chromatin: Transcriptional Control of Adaptive Immune Cell Development. Front. Immunol. 12, (2021).

Chromatin proteins can interfere with gene expression

In addition, even though we've drawn diagrams of genes and enhancers as straight lines and boxes, in the nucleus of a eukaryotic cell DNA is almost always packaged into chromatin and surrounded by proteins. Even during transcription, DNA stays associated with histones, although they must be moved or navigated around for RNA pol to access the DNA.

Just as a repressor protein can block access of an activator to an enhancer, so can chromatin proteins. Two things can't occupy the same place at the same time, whether an activator is competing with a repressor or whether an activator is competing with a nucleosome.

Chromatin remodeling proteins slide nucleosomes out of the way, eject nucleosomes from the DNA, or loosen the association of DNA with histones to allow a segment of DNA to be accessed by the transcription machinery. Some mechanisms of this are shown in **Figure 9.**[2]

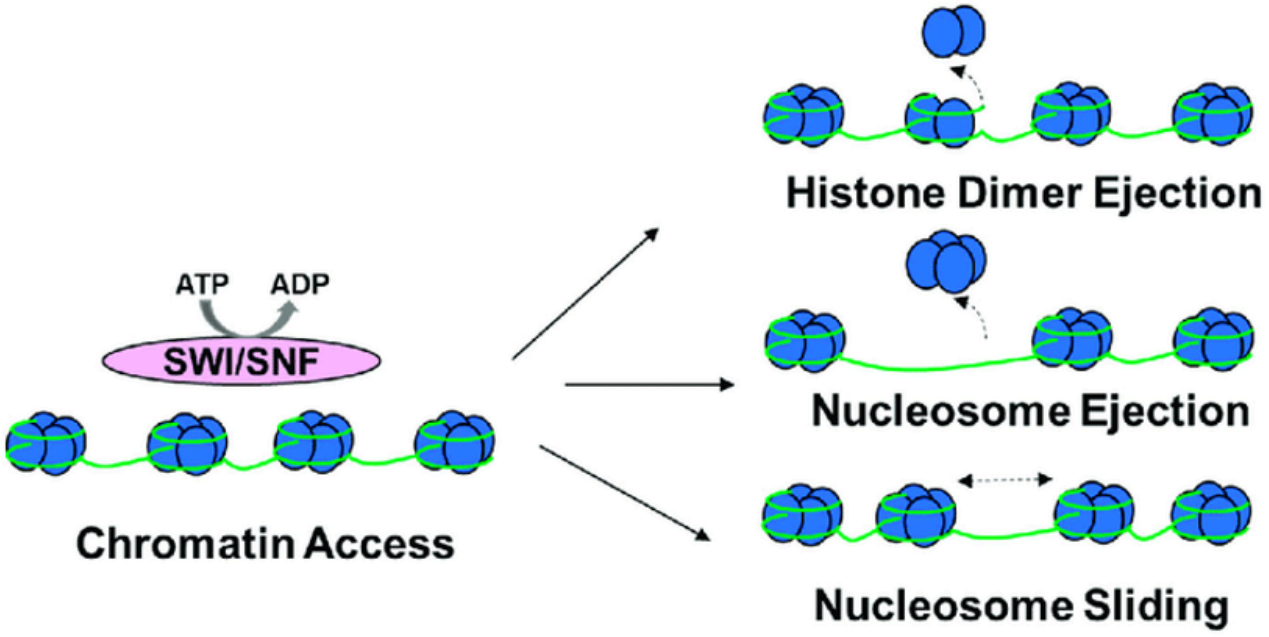


Figure 9. Mechanisms of chromatin remodeling. SWI/SNF is one chromatin remodeler that uses the energy of ATP hydrolysis to remodel nucleosomes. Nucleosomes may be remodeled by ejection of histone dimers, ejecting whole nucleosomes, or sliding nucleosomes along the DNA to expose an area of naked DNA.

Additional enzymes called **chromatin modifiers** can covalently modify histone proteins with small functional

2. Hasan, N. & Ahuja, N. The Emerging Roles of ATP-Dependent Chromatin Remodeling Complexes in Pancreatic Cancer. Cancers 11, 1859 (2019).

groups like acetyl or methyl groups. Enzymes called **histone acetyltransferases (HATs)** and **histone deacetylases (HDACs)** add and remove acetyl groups, respectively. Enzymes called **histone methyltransferases** add methyl groups. These modifications make it easier or more difficult to move histones out of the way depending on the modification, shown in **Figure 10**. [3]

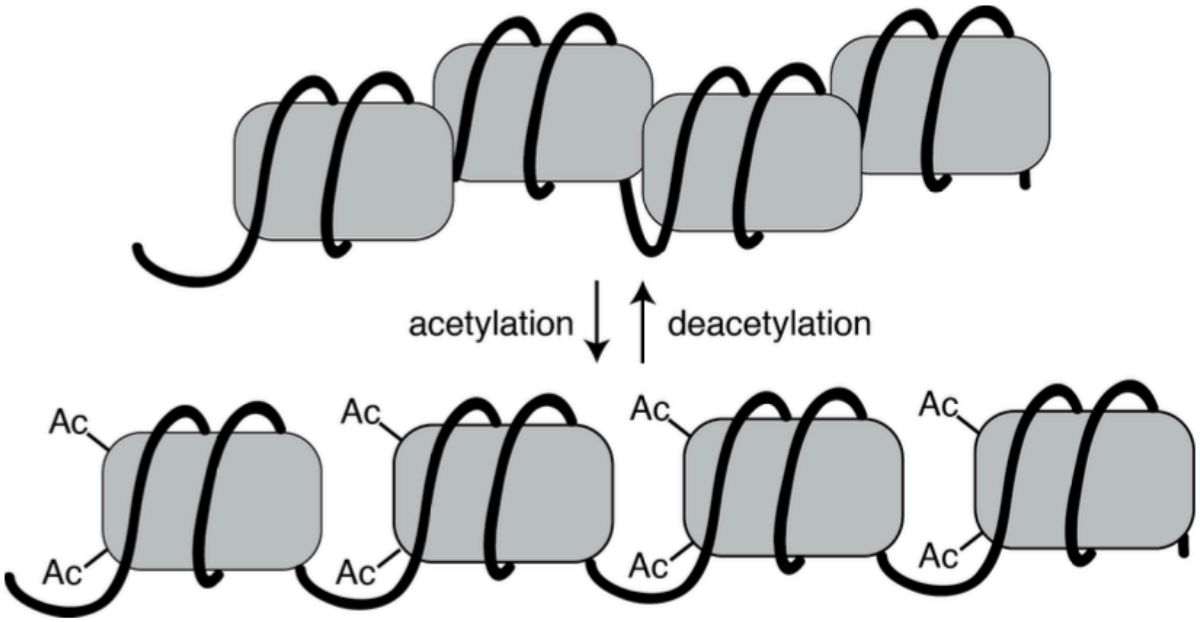


Figure 10. Histones can be modified covalently to change chromatin structure. Acetylation of histone proteins is often associated with more open chromatin configuration. Acetylation is a reversible process.

In some cases, these modifications "spread", with the histone modification state of one part of the genome influencing the modification of nearby nucleosomes. This type of regulation may allow temporal regulation (changes in gene expression over time) of genes that are arranged next to each other along the length of a chromosome.

Some modifications may result in such tight packing of chromatin that transcription cannot occur at all. Tightly packed, transcriptionally silenced chromatin is called **heterochromatin**, while loosely packed transcriptionally active chromatin is called **euchromatin**. Certain insulators may serve as barriers to the spread of chromatin modification so that chromatin in adjacent TADs can be in different states.

The influence of chromatin modification in gene expression is discussed in more detail in the chapter on *Epigenetics*.

3. Locke, J. 'Open Genetics Lectures' textbook for an Introduction to Molecular Genetics and Heredity (BIOL207). Borealis https://doi.org/10.7939/DVN/XMUPO6 (2017).

Media Attributions

- Figure 7 EG Regulation [TADS] © Navneet Matharu, Nadav Ahituv - Matharu, Navneet (2015-12-03). " Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease". PLOS Genetics 11 (12): e1005640. DOI:10.1371/journal.pgen.1005640. PMID 26632825. PMC: PMC4669122. ISSN 1553-7404. is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 8(Left) EG Regulation © Squidonius (talk) is licensed under a Public Domain license
- Figure 8 (Right) EG Regulation © Pongubala and Murre, Frontiers in Immunology is licensed under a CC BY (Attribution) license
- Figure 9 EG Regulation-Mechanisms of chromatin remodeling. © Modification of Hasan and Ahuja adapted by Amanda Simons is licensed under a CC BY (Attribution) license
- Figure 10 EG Regulation Histones © Deyholos - Open Genetics Lectures adapted by Amanda Simons is licensed under a CC BY-NC (Attribution NonCommercial) license

ADDITIONAL LEVELS OF GENE REGULATION

You'll recall from the chapter on Overview of Gene Regulation, we introduced the levels at which genes could be regulated, which is revisited here in **Figure 11.**

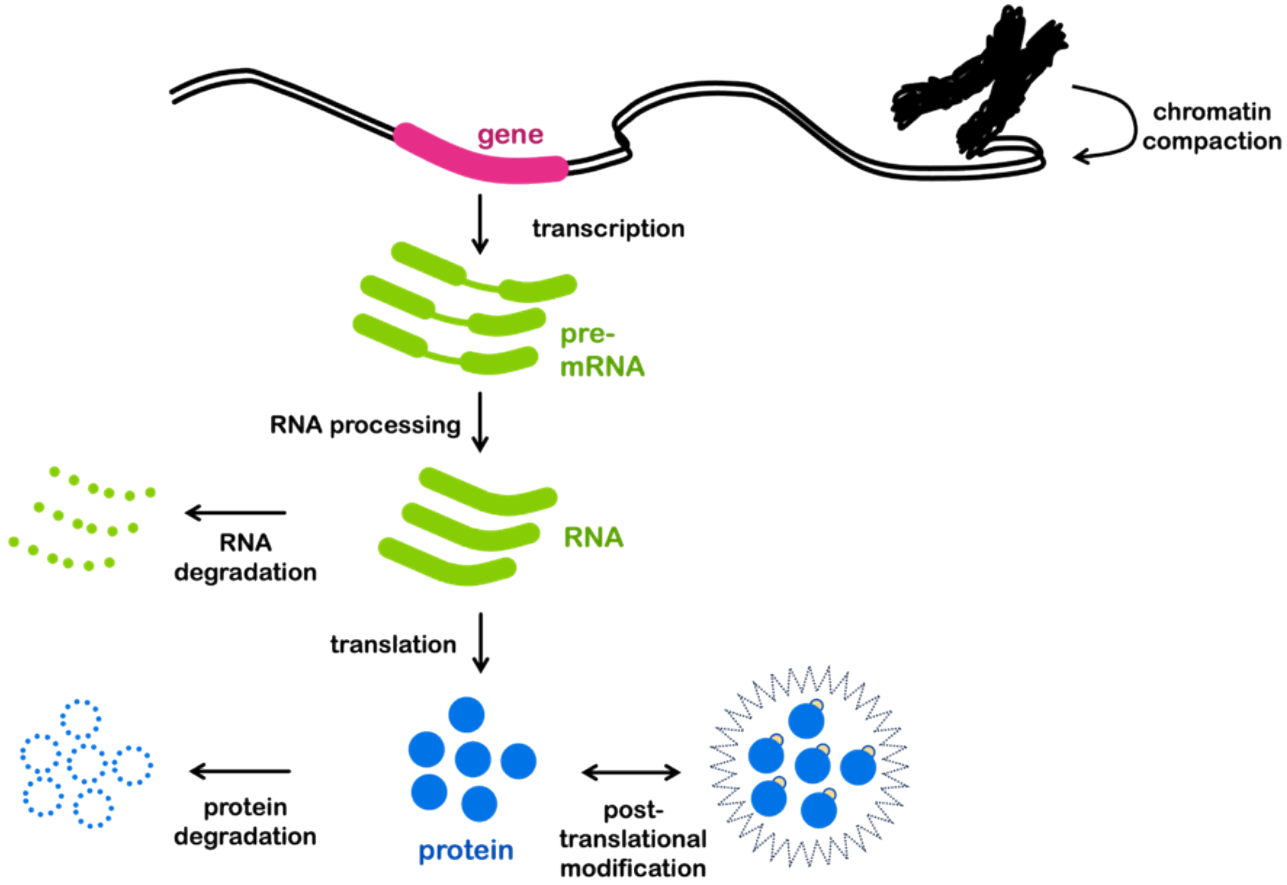


Figure 11. Levels of Gene Regulation. Expression of a gene and production of an active protein requires transcription, RNA processing, translation, and often post-translational modification. Each of those steps can be controlled by the cell to regulate gene expression Both RNA and protein have a limited lifespan in the cell, so their degradation represents additional points of control by the cell.

Both this chapter and the chapter on Overview of Gene Regulation have focused on regulation of gene expression at the level of transcription, focusing on the interactions of transcription factors that drive gene expression.

But keep in mind that the gene action also is affected by these other levels of control:

- RNA molecules are degraded after time, with long-lived RNA molecules typically being used to produce more protein than short-lived RNA.
- Translation can be increased or blocked by factors that interact with an RNA molecules.
- Post-translational modifications (either through covalent modification or non-covalent interactions with signaling molecules) can change the structure and behavior of proteins in the cell. This is a common method of regulation in cell-signaling, which allow a cell to respond to a chemical signal from the environment – kind of like the presence of lactose changes the behavior of the lac repressor, as we saw in the chapter on Overview of Gene Regulation.)
- And protein degradation is important in "turning off" the action of a gene as well

mRNA degradation via miRNA

Along those lines, another mechanism of control is the action of small regulatory RNAs.

One class of small regulatory RNA is microRNA (miRNA). miRNAs should not be confused with mRNA (messenger RNA). mRNA is translated into protein, but miRNA are functional RNAs that will never be translated.

Precursors to miRNA molecules are transcribed as longer RNA molecules that typically self-base-pair to fold up and form hairpins (**Figure 12**). Their respective genes are transcribed by either RNA Pol II or RNA Pol III. The long hairpin is cut into smaller fragments of double-stranded, or duplex, RNA. These fragments are typically only about 20 base pairs long.

One strand of the duplex RNA is typically loaded into a protein complex that mediates the interaction between the miRNA and a complementary mRNA. The interaction between the miRNA and its binding partner mRNA can lead to the dicing up of the mRNA; the degradation blocks the production of protein in the cell. In some cases, imperfect base pairing between the miRNA and mRNA does not trigger the dicing of the mRNA. Instead, the interaction between mi- and mRNA blocks translation of the mRNA. This process, called **RNA interference** (or RNAi) is illustrated in **Figure 12.**

Although the most well-studied examples of miRNA gene regulation down-regulate translation from target RNAs, there are also examples of miRNA increasing protein production or regulating transcription. And while most miRNAs appear to regulate mRNAs in the cells in which they are produced, some miRNAs may be exported from the cell to regulate mRNA stability in nearby cells.[1]

1. O'Brien, J., Hayder, H., Zayed, Y. & Peng, C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front. Endocrinol. 9, (2018).

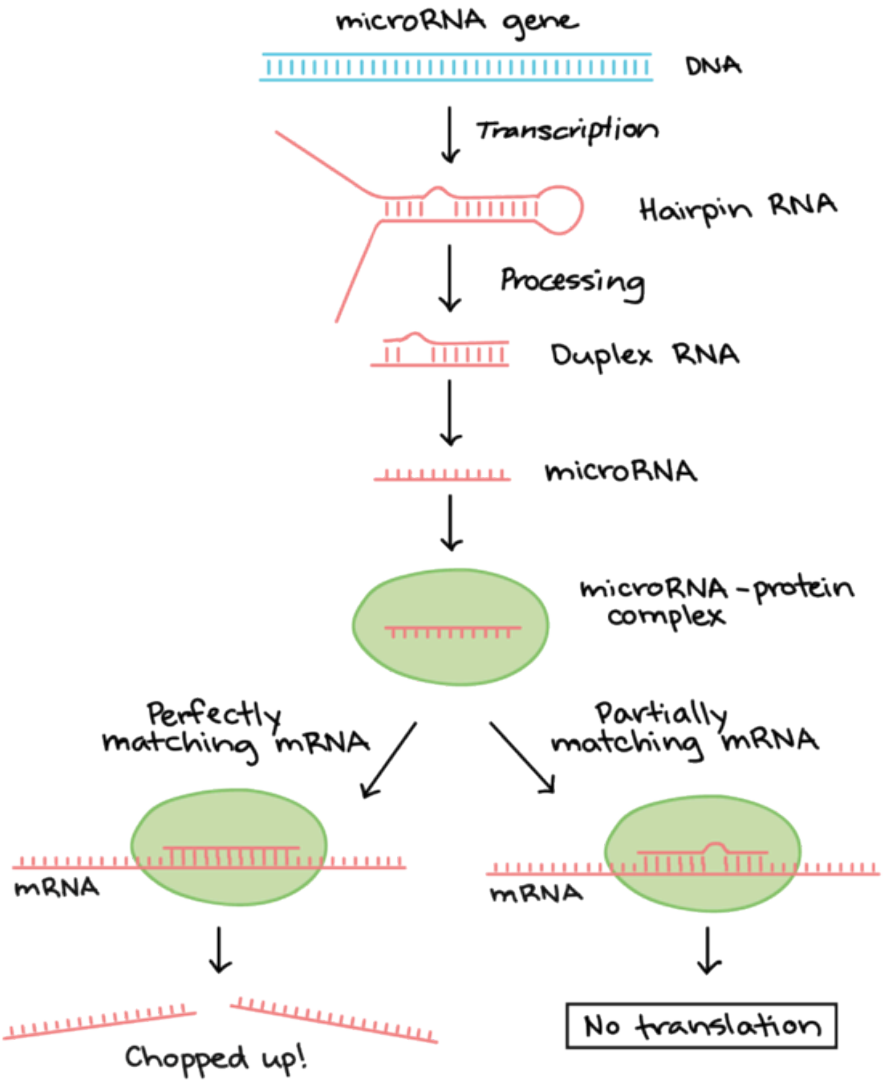


Figure 12. microRNA plays a role in post-transcriptional gene regulation by preventing translation of a target mRNA.

Andrew Fire and Craig Mello were awarded the Nobel Prize in Physiology or Medicine in 2006 for their discovery of RNA interference. Their work had a great influence on the tools of genetic manipulation in the lab: cellular mechanisms of RNA interference have been exploited by geneticists to experimentally silence the expression of target genes. These experiments are a form of reverse genetics, since the phenotype of cells or organisms can be monitored following gene silencing.

miRNAs (and related siRNAs) are also being exploited for therapeutic use: a sequence can be readily designed to target any gene product and delivered to target cells. They may be able to act on disease when traditional methods targeting cellular proteins do not. miRNA and siRNA drugs are in development for the treatment of diseases like cancer and hereditary metabolic disorders.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=953#h5p-103

Media Attributions

- Figure 11 GE REgulation Levels of Gene Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 12 GE Regulation microRNA © Narayanese adapted by Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

ENHANCERS IN ACTION: DROSOPHILA DEVELOPMENT AND THE EVE GENE

A classic example of eukaryotic gene regulation in action is seen during early development of the *Drosophila melanogaster* embryo. The *eve* gene described in the earlier section on enhancers is part of this process. The *eve* gene was first identified by Wieschaus and Nüsslein-Volhard in a search for embryonic mutants.

About 24 hours after fertilization of the egg, a wild type *Drosophila* embryo looks like **Figure 13** with visible segments. In this image, the anterior (head) part of the embryo is to the left, and the posterior (tail) is to the right.

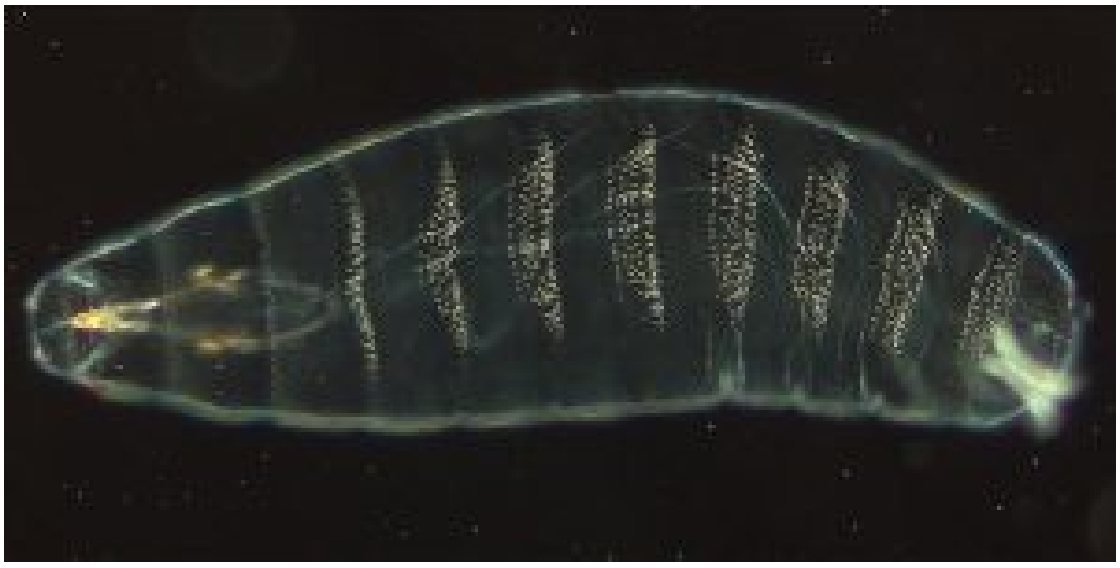


Figure 13. Drosophila embryos have a segmented appearance. The light-colored spots are denticles, short projections that extend from the ventral side of the embryo. Each segment has a slightly different denticle pattern.

Wieschaus and Nüsslein-Volhard searched for misshapen, mutant embryos that looked different to the wild type. Exposure to mutagens (chemicals that introduce mutations into DNA) leads to in increased frequency of mutant phenotypes. This type of experiment is called a **forward genetic screen**, where mutant embryos are screened for interesting phenotypes. See **Figure 14** for a comparison of forward and **reverse genetic screens**.

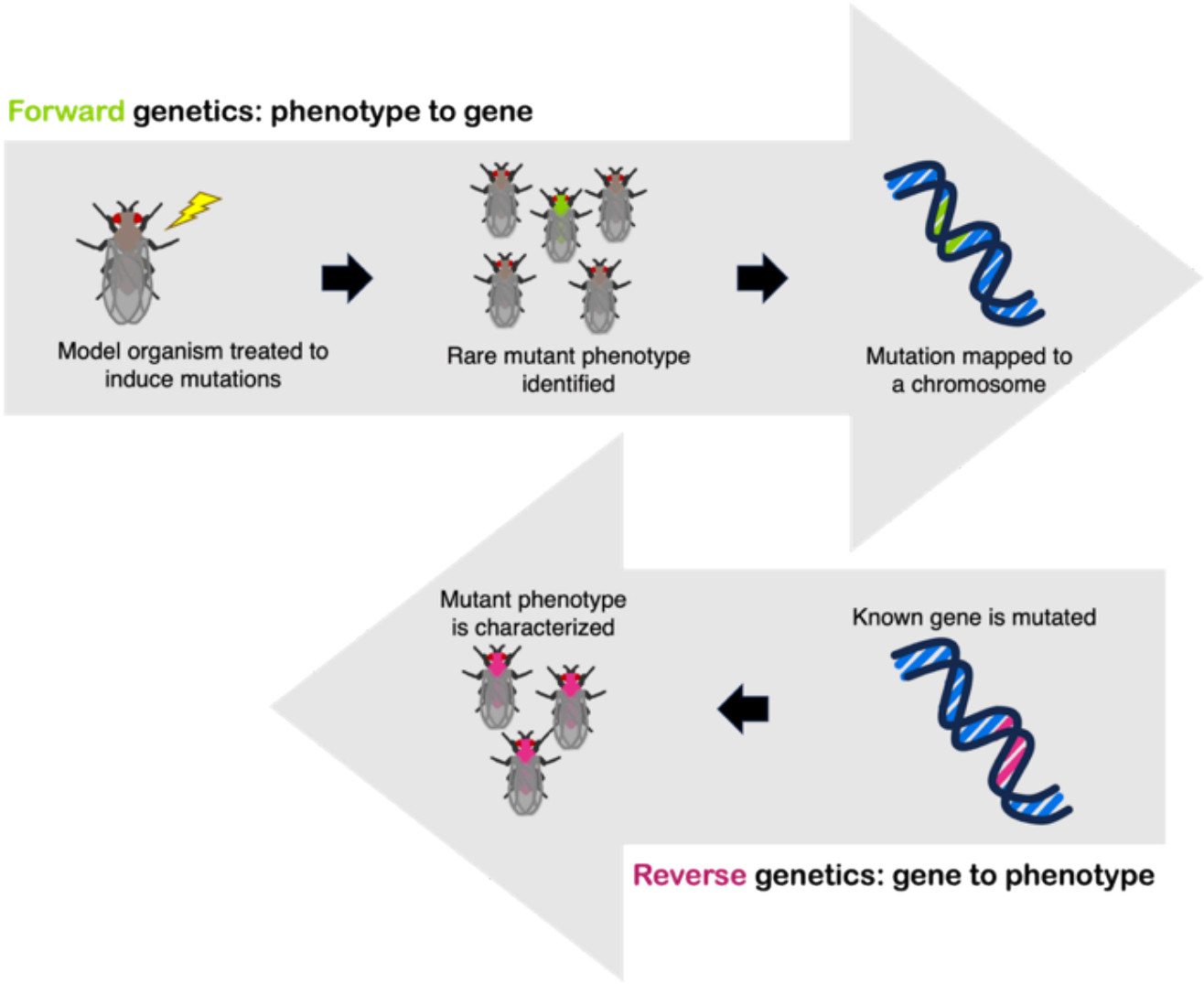


Figure 14. Forward and reverse genetics. A forward genetic screen begins by inducing mutations in a target population, searching for mutant phenotypes, and identifying causative genes. A reverse genetic screen begins by modifying a cell or organism to alter the function of a specific gene. Some experiments may add an extra copy of a gene to a cell, leading to over-production of the encoded product. Other experiments may "knock out" function of a gene. After modification, the cell or organism is screened for mutant phenotypes.

Nüsslein-Volhard and Wieschaus searched painstakingly through mutant *Drosophila* embryos, looking for those with mutations affecting development. And they did find misshapen embryos! Most of these embryonic phenotypes were lethal, so the embryo could not grow into an adult fly. But the early embryos had clearly visible mutant phenotypes, with the wild-type pattern of segmentation disrupted. You can see images of the mutant phenotypes in their 1980 paper.

Some mutant embryos were missing multiple segments in the middle of the embryo. Others were missing every other segment or showed structures reversed in orientation. (Yes, they looked very weird.) The *eve* mutants were missing every even-numbered segment, so the gene was named *even-skipped*, or *eve* for short.

Studies of these embryonic mutants revealed the underlying genetic rules that control development in *Drosophila*. Eric Wieschaus, Christiane Nüsslein-Volhard, and Edward Lewis were awarded the Nobel Prize in Physiology or Medicine in 1995 for this work.

Note: Although the examples given here are from *Drosophila*, these genetic mechanisms were later shown to be conserved in other organisms. Today, we know that human versions of genes important for patterning in the *Drosophila* embryo also are important in human development. The regulation of those genes is what makes it possible to build a multicellular organism with differentiated tissues.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=869#h5p-104

Eve expression during development

For multicellular eukaryotic organisms that reproduce sexually, all cells of the body originate from a single cell: the **zygote** is the fertilized egg that contains the genetic information from both parents. The zygote undergoes subsequent rounds of cell division as it develops into an embryo. As the cells divide, they become specialized so that different tissues can perform different tasks within the body. This process of specialization is called **differentiation.** At each step of the differentiation process, changes in gene expression cause a different set of proteins to be produced, which affects cell phenotype. To use vocabulary of other modules, each cell of a developing embryo has the same genome but a different transcriptome and proteome.

Remember from earlier in this chapter: the gene *eve* is expressed in multiple stripes along the length of an embryo. Embryos mutant for *eve* are missing every other segment – the ones where *eve* is normally expressed. Expression of eve in this stripe-like pattern is controlled by five enhancers, as shown in the top panel of **Figure 15**.

Each of the five eve enhancers has binding sites for multiple factors. The stripe 2 enhancer is diagrammed in **Figure 15.**

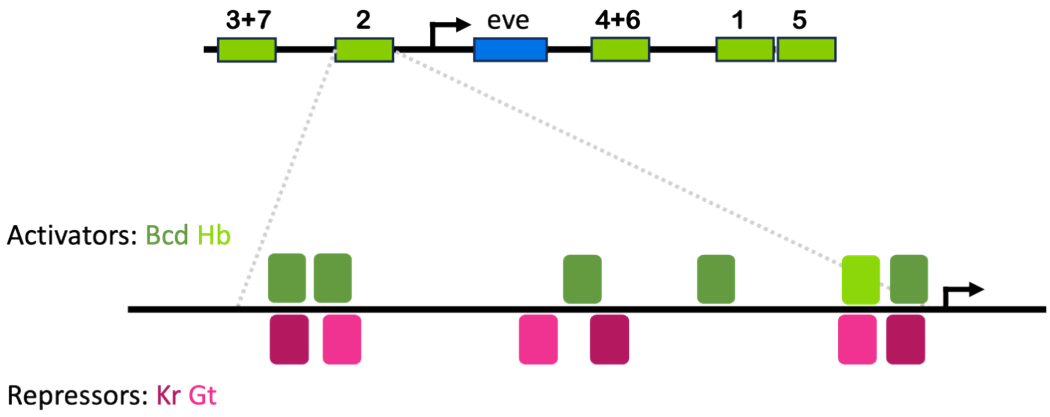


Figure 15. The Eve stripe 2 enhancer. The minimal stripe 2 enhancer has multiple binding sites for activators Bicoid and Hunchback and repressors Giant and Kruppel (not drawn to scale). Note that there are additional transcription factor binding sites that are not shown in this image.

The stripe 2 enhancer drawn in **Figure 15** is about 500 base pairs long. It has binding sites for activators Bicoid (Bcd, dark green) and Hunchback (Hb, light green), as well as repressors Giant (Gt, light pink) and Kruppel (Kr, dark pink)8. These funny-named factors are all named for the phenotypes observed in mutant embryos – more on that later.

You'll notice that the activator binding sites overlap with many of the repressor binding sites. If repressors are bound to these elements, activators presumably cannot bind. In this way, the enhancer drives expression of the *eve* gene in any tissue that has both activators (Bcd and Hb) but neither of the repressors (Gt and Kr).

This is important: it allows the *eve* gene to be turned on, but also to be turned off! Proper gene regulation means that a gene is turned on when appropriate *and* turned off when appropriate. Not all genes are expressed in all tissues!

Bicoid, Hunchback, Giant, and Kruppel are expressed in varying patterns along the length of the *Drosophila* embryo, from anterior to posterior, as shown in **Figure 16**. About one third of the way from the anterior of the embryo, there is one area several cells wide where both Bicoid and Hunchback proteins are present and where Kruppel and Giant proteins are not. This is where *eve* stripe 2 is expressed.

Figure 16. Localization of Bicoid, Hunchback, Kruppel, and Giant expression, relative to eve stripe 2. The stripe 2 enhancer drives Eve expression where Bicoid and Hunchback are present, and Kruppel and Giant are not, indicated by the blue dotted line. At the posterior end of the embryo, a lack of Bicoid prevents the eve stripe 2 enhancer from being active. At the posterior end of the embryo, both activators are present. However, eve expression is blocked by additional factors not shown here. Not drawn to scale.

You can think of Bicoid and Hunchback as activating *eve* expression everywhere in the anterior end of the embryo, but Giant and Kruppel block the edges of that expression. (Note: Additional influences not discussed here limit expression of eve at the anterior-most end of the embryo.)

The four other enhancers are regulated by these factors and others, with each enhancer acting as its own independent switch. For example, the stripe 3+7 enhancer is activated by proteins found everywhere in the embryo, but it is inhibited by Hunchback and a protein called Knirps, as shown in **Figure 17**. (Knirps is pronounced "nerps").

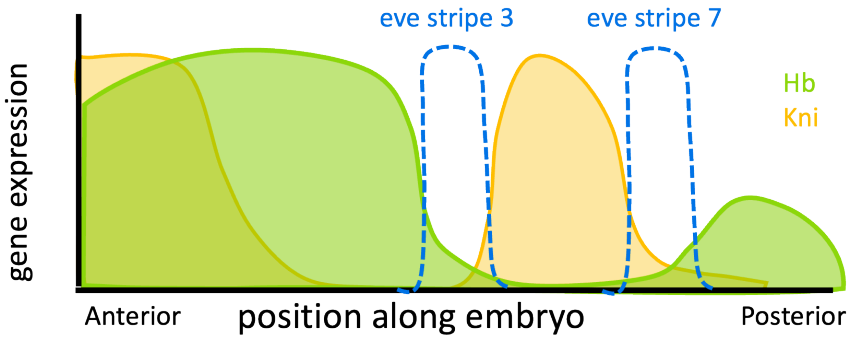


Figure 17. Stripe 3+7 enhancer is activated by proteins expressed everywhere in the embryo, but it is repressed by Hunchback (Hb) and Knirps (Kni). Eve Stripes 3 and 7 therefore form between the boundaries marked by these two repressors. Not drawn to scale.

Because of the different arrangements of binding sites, Hunchback acts as an activator of stripe 2, but a repressor on stripes 3 and 7! Many transcription factors act to both activate and repress transcription, depending on context. This is true in both prokaryotes and eukaryotes – the lambda repressor described in the Overview of Gene Regulation chapter was another example.

Note that only one enhancer needs to be active for the gene to be expressed. Conditions are right for the stripe 3+7 enhancer to drive eve expression in stripes 3 and 7 of the embryo. Conditions are right for the stripe 2 enhancer to be active in stripe 2. The same is true for other enhancers in other regions of the embryo.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=869#h5p-105

A note on Drosophila naming conventions

Most of these genes were initially discovered in forward genetic screens for homeotic or

embryonic mutants, and the names that reflect the original researcher's description of the mutant phenotype. For example, the name *bicoid* is derived from the fact that mutant embryos lacked anterior (head) structures. These mutants were described as "bicaudal", which means having two tails. An illustration of a bicoid mutant embryo is shown in **Figure 18**, and you can see a micrograph of a *bicoid* mutant embryo at the CSH DNA Learning Center. Some of the names are quite whimsical: the name *gooseberry* reflects the round mutant embryo that sort of resembled a gooseberry **(Figure 19**), and *armadillo* mutants resembled a little armadillo.

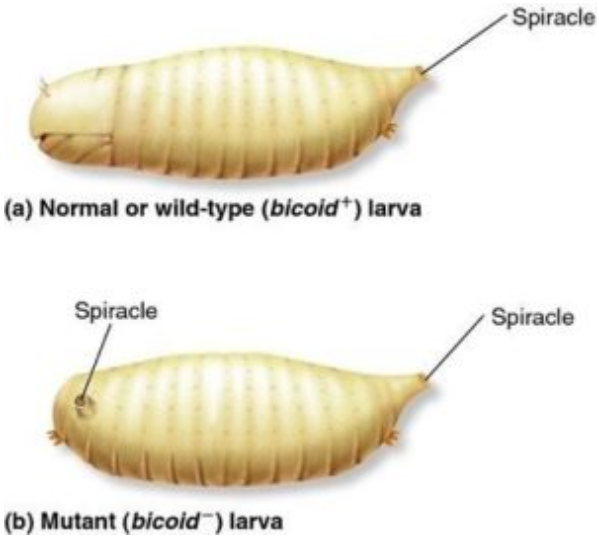


Figure 18. A bicoid mutant lacks anterior structures.

Figure 19. Gooseberries.

It should be acknowledged, though, that some of the names reflect their times: *Kruppel*, for example, is derived from a German word meaning cripple, and *hunchback* was named for the shape of the unusual shape and distribution of denticles. A fruit fly doesn't care if it's called a

cripple, but this becomes relevant to people because many of these genes have human orthologs that have implications for human health and disease. A gene name like "lunatic fringe" or "Indy" ("I'm not dead yet") or "roadkill" can make even more difficult conversations with patients who are already facing challenges.

In the chapter on Mutation, we discussed the rationale for a shift from the word "mutant" to the word "variant" when describing human genetic variation. Here is another example of how terminology has changed over time as scientists and medical professionals recognize the impact of the vocabulary. The human versions of some of these genes have been renamed,[1] often as abbreviations that become the more common nomenclature (e.g. *Kruppel-like Factor 1* is KLF1).

Media Attributions

- Figure 2 EG Regulation is licensed under a Public Domain license
- Figure 14 EG Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 15 EG Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 16 EG Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 17 EG Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 18 EG Regulation © Charodeika ya is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 19 EG Regulation © neurovelho is licensed under a CC BY-SA (Attribution ShareAlike) license

1. Hopkin, M. Troublesome gene names get the boot. Nature news061106-2 (2006) doi:10.1038/news061106-2.

SPATIOTEMPORAL GENE REGULATION DURING DEVELOPMENT

So, the eve gene is regulated by bicoid, hunchback, kruppel, and giant – but how do those transcription factors end up expressed in exactly the right place? And how does it relate to the homeotic mutants we saw at the beginning of this chapter? They are all interdependent.

In the next section, we will look at five different classes of genes that play a role in building body structures along the anterior-to-posterior axis of the embryo.

1. Maternal effect genes (like *bicoid*)
2. Gap genes (like *hunchback*, *kruppel*, and *giant*)
3. Pair-rule genes (like *eve*)
4. Segment polarity genes
5. homeotic genes (like *antennapedia* and *ultrabithorax*)

These genes are all part of a larger regulatory network that leads to the differentiation of cells in the embryo. Each gene encodes a protein factor that influences the expression of other downstream genes, either directly or indirectly. Many are transcription factors, but others are signaling molecules that indirectly affect transcription. Some of these mechanisms are discussed later in this chapter. We see that expression is controlled over time and in different spaces within the embryo – **spatiotemporal** gene regulation. Although this section focuses on anterior-posterior patterning, similar factors influence dorsal-ventral patterning as well.

SUCCESSIVE ACTION OF MATERNAL EFFECT, GAP, PAIR-RULE, SEGMENT POLARITY, AND HOMEOTIC GENES

You may recall that, during fertilization, the sperm mostly just contributes DNA to the zygote, while the egg contributes both DNA and cytoplasmic components.

The cytoplasm of the fertilized egg contains maternal mRNA. After fertilization, the maternal mRNA can be used to produce protein quickly before gene expression begins from the zygotic genome. Because the genotype of the mother influences the phenotype of the zygote (not the genotype of the zygote itself!), these are called **maternal effect genes**. *Bicoid* is an example of a maternal effect gene. As shown in **Table 1**, maternal effect genes show a non-Mendelian pattern of inheritance because the phenotype is not dependent on the genotype of the individual.

Table 1. Maternal effect genes show non-Mendelian inheritance patterns. In this image, if a female with at least one wil crossed to a male of any genotype, her offspring will have a wild-type phenotype. A female with two mutant alleles (m) w offspring, regardless of their genotype.

All offspring show the wild-type phenotype	All offspring show the mutant phenotype
♀ m/+ x ♂ m/+ ♀ m/+ x ♂ m/m ↓ ↓ ● m/m ● m/+ ● +/+ ● m/m ● m/+	♀ m/m x ♂ m/+ ♀ m/m x ↓ ↓ ▲ m/m ▲ m/+ ▲ m/-

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=910#h5p-106

The *Drosophila* embryo is different from a mammalian embryo in that the earliest cycles of mitosis happen

very rapidly – as fast as 8 minutes per cycle – but do not include cytokinesis (the separation of the cytoplasm)(**Figure 20)**. There is also not much growth in size. As a result, the early embryo consists of a **syncytium** (pronounced "sin-**sish**-um") multiple nuclei together in a continuous cytoplasm. Within a few hours of fertilization, the nuclei migrate to the outer edges of the syncytium. The cytoplasm is compartmentalized to form individual cells.

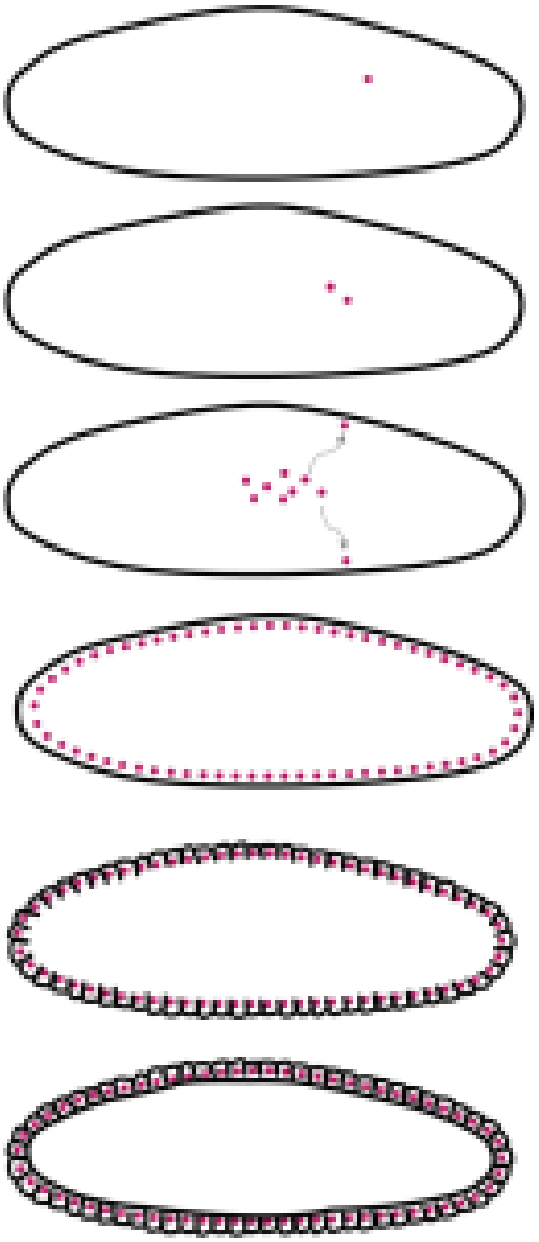


Figure 20. The Drosophila embryo begins as a syncytium, a single continuous cytoplasm with multiple nuclei. Immediately after fertilization, the zygote undergoes multiple rounds of nuclear mitosis without cell growth or cytokinesis. This results in many nuclei in one cytoplasmic compartment. Within a few hours after fertilization, the nuclei migrate to the outer edges of the embryo. The cell membrane is pulled inward, and individual cells are separated from one another. Note that this is different from early mammalian development. Mammalian zygotes fully

Bicoid mRNA is localized to the anterior end of the syncytium. As the embryo translates the mRNA, Bicoid protein diffuses throughout the syncytium, forming a gradient with the highest concentration at the anterior end of the embryo. Other maternal effect genes have different distribution patterns throughout the embryo. For example, Nanos protein is distributed in a gradient with the highest concentration at the posterior end. And other proteins are important for dorsal-ventral patterning.

The bicoid protein is a **morphogen** that triggers the development of structures within the embryo through a cascade of gene expression events. Bicoid protein can activate zygotic genes where it contacts the zygotic nuclei within the syncytium. As individual cells pinch off parts of the shared cytoplasm to become separated from one another, they end up with decreasing amounts of Bicoid protein toward the posterior of the embryo. The different concentrations of bicoid in the separated cytoplasms lead to differential expression of bicoid-regulated genes (like eve).

Bicoid directly or indirectly regulates genes involved in the segmentation of the embryo, including *hunchback*, which is one regulator of the *eve* gene discussed earlier. Although maternal *hunchback* RNA is distributed throughout the embryo, it is only translated toward the anterior end of the embryo due to translational repression by the Nanos protein in the posterior. Soon after fertilization, though, the zygotic *hunchback* gene is transcribed, driven by multiple factors. Bicoid drives the production of Hunchback protein at the anterior of the embryo, while a second region of Hunchback expression occurs at the posterior end of the embryo, driven by bicoid-independent mechanisms.

Hunchback belongs to a class of genes called **gap genes** because they are expressed in broad swaths of the embryo, and loss-of-function mutations result in embryos with a gap of missing structures. *Bicoid* and *hunchback,* in turn, participate in regulating the expression of *kruppel*, *giant*, and *knirps*, which are three other gap genes.

divide into separate daughter cells and do not produce a syncytium

In the previous section, we looked at how Bicoid, Hunchback, Kruppel, and Knirps regulate the expression of *even-skipped* (*eve*). *Eve* belongs to a class of genes called **pair-rule genes**, along with other genes like *odd-skipped, hairy*, and *fushi-tarazu*. Like *eve*, these genes are expressed in alternating stripes along the length of the embryo. The expression of the pair-rule genes are dependent on maternal effect and gap genes, as well as interactions with each other.

The pair-rule genes, in combination with the maternal effect and gap genes, then influence the expression of segment-polarity genes, which establish a difference between the anterior and posterior edges of segments in the Drosophila embryo. Some examples of segment-polarity genes include the Drosophila genes hedgehog (Hh) and gooseberry (gsb). These genes were also named for the embryonic phenotypes linked with mutations in these genes. A drawing of a hedgehog mutant embryo is shown in **Figure 21**; rather than the orderly arrangement of denticles seen in a wild-type embryo, hedgehog mutant embryos had a lawn of denticles scattered across the surface of the embryo. Remember that denticles are small projections that extend from the surface of the embryo. Perhaps you can imagine why the mutation was named "hedgehog."

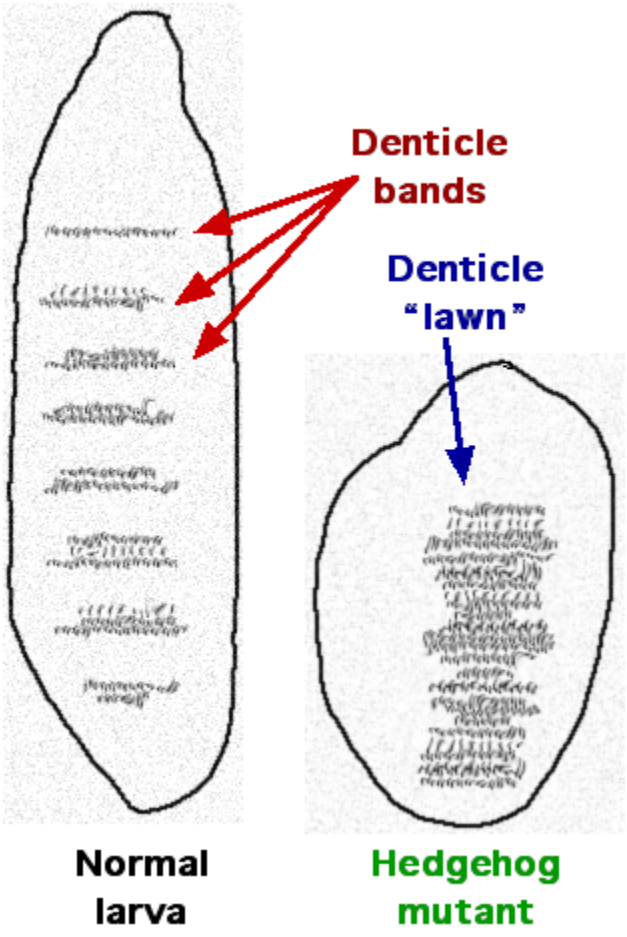


Figure 21. A wild-type drosophila embryo and a drawing of a hedgehog mutant embryo.

Finally, the segment polarity genes, pair-rule genes, gap genes, and maternal effect genes collectively regulate the **segment identity** genes. These genes include the homeotic genes like *antennapedia* (*Antp)* and *ultrabithorax (Ubx)* that were described at the beginning of this chapter. These genes do not affect segment number; they just affect the identity of the segment.

The homeotic genes (or Hox genes) also encode transcription factors. These factors share a common amino acid sequence called a homeobox. The homeobox is a DNA binding domain that allows these factors to recognize elements that regulate the development of body structures. They control the expression of the genes needed to build structures like wings, eyes, and antennae.

For example, the expression of *antennapedia* is important for the development of leg structures. But if the *antennapedia* is mis-expressed in the head, leg structures grow in place of the antenna. This is shown way back in in**Figure 2**, and is what gives the gene the name of *antennapedia*. Where ultrabithorax is expressed in thorax and abdominal segments, it represses wing formation. A loss of function in *ultrabithorax* results in the additional set of wings we see in **Figure 1**, while a gain of function mutation prevents the normal forewing structure from developing. Expression of a gene called *eyeless* in the leg leads to the formation of eye structures on the leg.

Together, the classes of genes described here exemplify **spatiotemporal gene regulation**, where the timing of gene expression and the position of a cell within an organism determine which genes are expressed. In early embryonic stages, each class of genes is dependent on the combinatorial effects of earlier-expressed classes. In later stages of development, these gene products together influence patterning in specific organ systems.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=910#h5p-107

Media Attributions

- Table 1-Wild-cropped © Celefin is licensed under a CC BY-SA (Attribution ShareAlike) license
- Table 1-Mutant-cropped © Celefin is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 20 EG Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 21 EG Regulation © JWSchmidt is licensed under a CC BY-SA (Attribution ShareAlike) license

FLIES ARE NICE, BUT WHAT ABOUT OTHER ORGANISMS? GENE EXPRESSION IN EVOLUTION AND DEVELOPMENT

The rules for establishing body patterns in other organisms are similar to those seen in *Drosophila*.

There are some differences in developmental processes, especially in the early stages of development (mammals do not have a syncytial embryo, for example). Mammals also do not have a *bicoid* homolog, but there are examples of other maternal effect genes. For example, some maternal effect genes appear to play a role in recurrent miscarriages and birth defects.[1]

There are many similarities, though. Hox genes, for example, have **orthologs** in other vertebrates. (An ortholog is a similar, or homologous, gene that is found in another organism.) A comparison of gene and genome structure between *Drosophila* and other vertebrates offers clues to evolutionary processes.

Although homeotic genes are scattered throughout the fruit fly genome, many of them are clustered in two regions on the *Drosophila melanogaster* Chromosome 3.

Interestingly, the genes are arranged on the chromosome in a similar order to their expression along the anterior-to-posterior axis of the organism. Each of those segments eventually develops into a different structure within the adult fly, as shown by the color-coding in **Figure 22**. The identity of each segment is driven by the expression of the Hox genes.

1. Mitchell, L. E. Maternal effect genes: Update and review of evidence for a link with birth defects. Hum. Genet. Genomics Adv. 3, 100067 (2021).

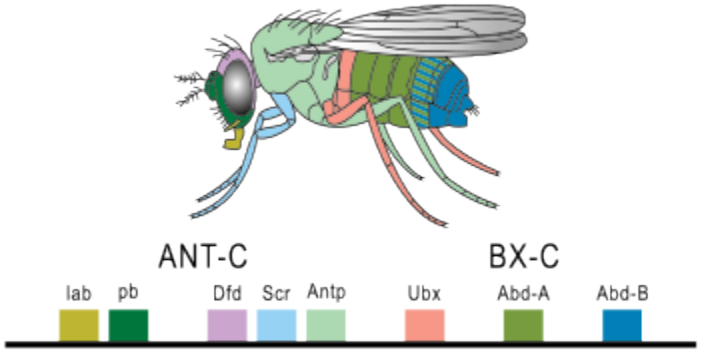


Figure 22. Two clusters of hox genes on Drosophila melanogaster chromosome 3. The Ant-C cluster includes 5 genes and the BX-C cluster includes 3. These genes are arranged along the chromosome in the same order that they are expressed along the anterior-posterior axis in the fly body.

The sequence homology between the *Drosophila* hox genes suggests that the cluster of genes arose through a duplication of an ancestral hox gene. The identification of Hox genes in *Drosophila* also made it possible to identify similar gene sequences in other organisms, including humans.

And it turns out that other organisms have hox genes, too. Humans and other mammals actually have four clusters of Hox genes, not two. This suggests that an ancestral cluster of Hox genes was duplicated during evolution of the mammalian genome. These clusters are named Hox-A, Hox-B, Hox-C, and Hox-D. Each cluster maintains the head-to-tail order of genes, with recognizable **paralogs** across clusters. (Paralogs are genes with similar sequences and function in the same organism. They result from gene duplication over evolutionary history.)

Although it is more difficult to imagine a human as a segmented organism, the human Hox genes nevertheless are expressed in segments in the body! This is illustrated in **Figure 23**, reprinted from *OpenStax Biology 2e.*[2]

The Hox genes are highly conserved among all species studied, meaning they have relatively few differences in sequence. This suggests most Hox gene mutations would not be compatible with survival: Sequences that are highly conserved throughout evolutionary history tend to be very important to organism function.

2. Rye, C. et al. 27.1 Features of the Animal Kingdom – Biology | OpenStax. https://openstax.org/books/biology-2e/pages/27-1-features-of-the-animal-kingdom (2016).

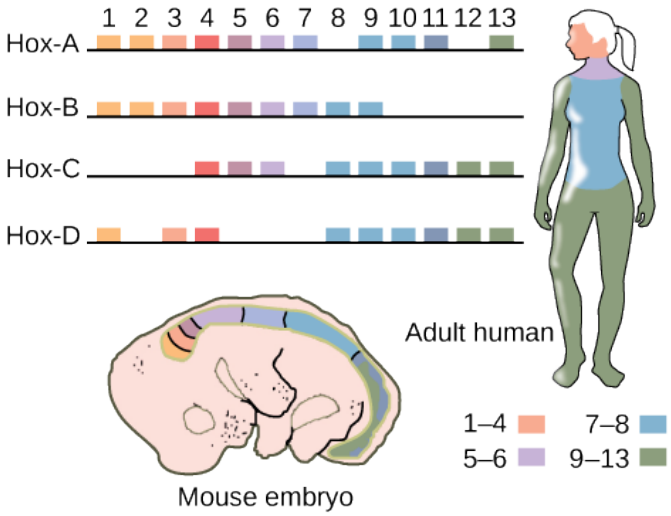


Figure 23. Hox genes are highly conserved genes encoding transcription factors that determine the course of embryonic development in animals. In vertebrates, the genes have been duplicated into four clusters: Hox-A, Hox-B, Hox-C, and Hox-D. Genes within these clusters are expressed in certain body segments at certain stages of development. Shown here is the homology between Hox genes in mice and humans. Note how Hox gene expression, as indicated with orange, pink, blue and green shading, occurs in the same body segments in both the mouse and the human.

As in *Drosophila,* all four Hox clusters have the genes arranged in the order in which they are expressed along the body axis. Like the gene sequence themselves, this organization is conserved across animal phyla, which suggests the order of the genes is also important for function.

This seems a little strange, right? Why would the order of genes on the chromosome match the order in which the genes are expressed?

The order of the genes may play a role in the timing of their expression, with the genes expressed temporally in the order in which they are organized on the chromosome.

Remember: chromatin and histone modification can influence gene expression, and histone modification can "spread" along chromosomes, with the modification of one area influencing the subsequent modification of an adjacent area. This process is seen in the hox gene cluster during development and is likely one factor contributing to why the linear arrangement of the genes has been conserved throughout evolutionary history.

hedgehog genes are conserved in vertebrates

Another interesting evolutionary story is seen with the *Drosophila* gene *hedgehog* and its vertebrate homologs.

In the previous section we mentioned the gene *hedgehog (Hh)* as a segment polarity gene. Like the Hox clusters, *hedgehog* has multiple orthologs in vertebrates. In keeping with the naming system begun by the *Drosophila* geneticists, the three vertebrate orthologs are named Indian hedgehog (*Ihh),* Desert hedgehog (*Dhh),* and Sonic Hedgehog (*Shh*). Yes, the gene was named after the video game character Sonic the Hedgehog.

In Drosophila, *Hh* is important for segment polarity and plays a role later in development: Hh is a secreted morphogen with target proteins that are important for the development of diverse organ systems, including the wing and nervous system.

The vertebrate orthologs of hedgehog are also morphogenic signaling molecules. *Sonic Hedgehog (Shh)* is probably the best-studied of the three and the closest in function to the *Drosophila* gene. *Shh* is important for

development in many species: in humans, variants in the *Shh* coding sequence are associated with anomalies in brain and facial structure.

Shh regulation is another good example of enhancers in action. As we saw with the *eve* gene, *Shh* is regulated by multiple enhancers,[3] diagrammed in **Figure 24.** These enhancers direct the expression of *Shh* in different tissue and organ systems. In panel A, a map of the mouse genome is shown, with enhancer locations marked by colored bars and genes in gray boxes. The human sequence is very similar.

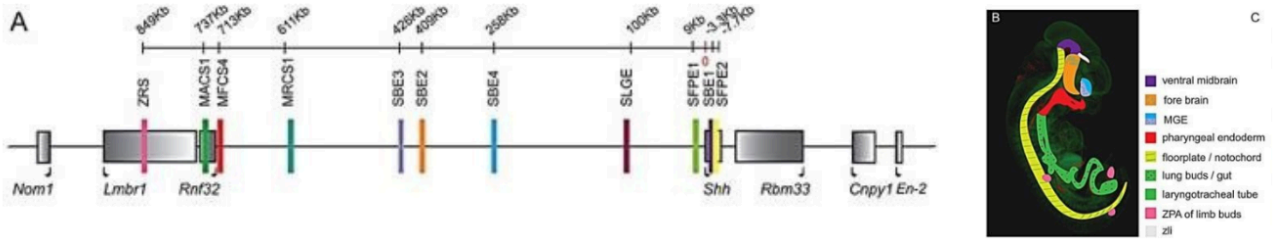


Figure 24. The Shh gene is regulated by multiple enhancers

The enhancers are scattered throughout a region of about 1000 kilobases (one million base pairs!), mostly upstream of the *Shh* gene. The enhancers are color-coded according to the region of the embryo where they drive *Shh* expression (Panel B). The enhancers contain binding sites for Hox proteins as well as other regulatory factors.

Although variants of the *Shh* coding sequence are primarily associated with brain and facial anomalies, mutations in enhancer sequences can have very different phenotypic effects, mostly confined to the part of the body in which the enhancer is active.

One of the most well-studied of these is the enhancer marked in pink and labeled ZRS in **Figure 24.** ZRS stands for Zone of Polarizing Activity Regulatory Sequence. The Zone of Polarizing Activity is a group of cells in the developing limb bud that coordinates limb digit formation and regulates the expression of Shh in limb buds. The ZRS is around one million base pairs away from the gene itself! It is an 800-base-pair sequence found within the intron of another gene.

Variants within the ZRS enhancer are associated with differences of limb development, like the preaxial polydactyly and triphalangeal thumb seen in **Figure 25.**

3. Anderson, E., Devenney, P. S., Hill, R. E. & Lettice, L. A. Mapping the Shh long-range regulatory domain. Dev. Camb. Engl. 141, 3934–3943 (2014).

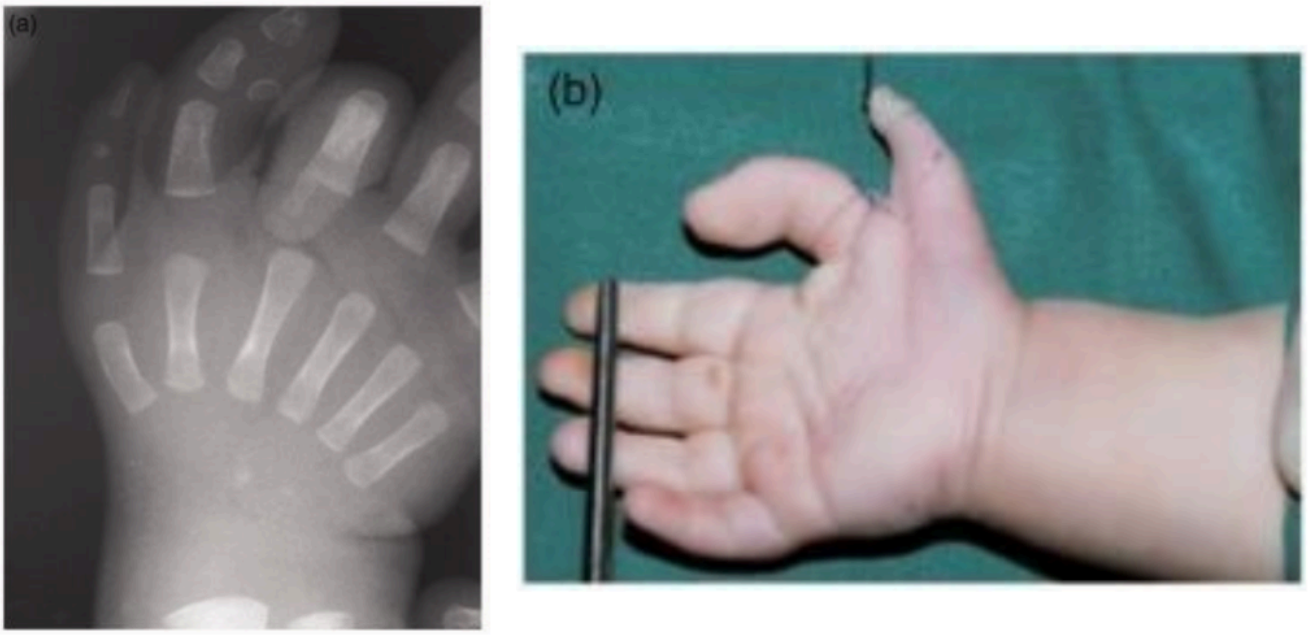


Figure 25. Preaxial polydactyly and triphalangeal thumb. Panel A (left): X-ray. Panel B (right): photograph.

It's important to note here that although *Shh* is expressed in many different tissues, variants of the ZRS enhancer are only associated with limb differences. Mutations in this enhancer do not affect the regulation of *Shh* in other tissues. Phenotypically these mutations are distinct from both mutations in the *Shh* coding sequence and mutations in other enhancers.

Shh enhancers in other species affect limb development, too!

When we compare *Shh* enhancer sequences among other species of vertebrates, we find that the ZRS enhancer is highly conserved – with some notable exceptions! Snakes are missing a 17-base pair segment within the ZRS enhancer. (**Figure 26)**

Snakes are unusual among their closest terrestrial relatives in that they do not have limbs. However, certain snakes including boa constrictors and pythons, do have vestigial internal limb structures, including rudimentary hindlimb bones. These structures suggest that snakes evolved from a legged ancestor and that leg structures were subsequently lost during evolution.

Although the snake ZRS sequence is quite variable from other vertebrates, all snakes seem to have this 17-base pair deletion in common.

Experiments to introduce targeted mutations into the mouse genome show that the loss of this 17-base pair

sequence prevents proper limb development in mice. This suggests that an ancestral deletion in ZRS is why snakes do not have legs.[4]

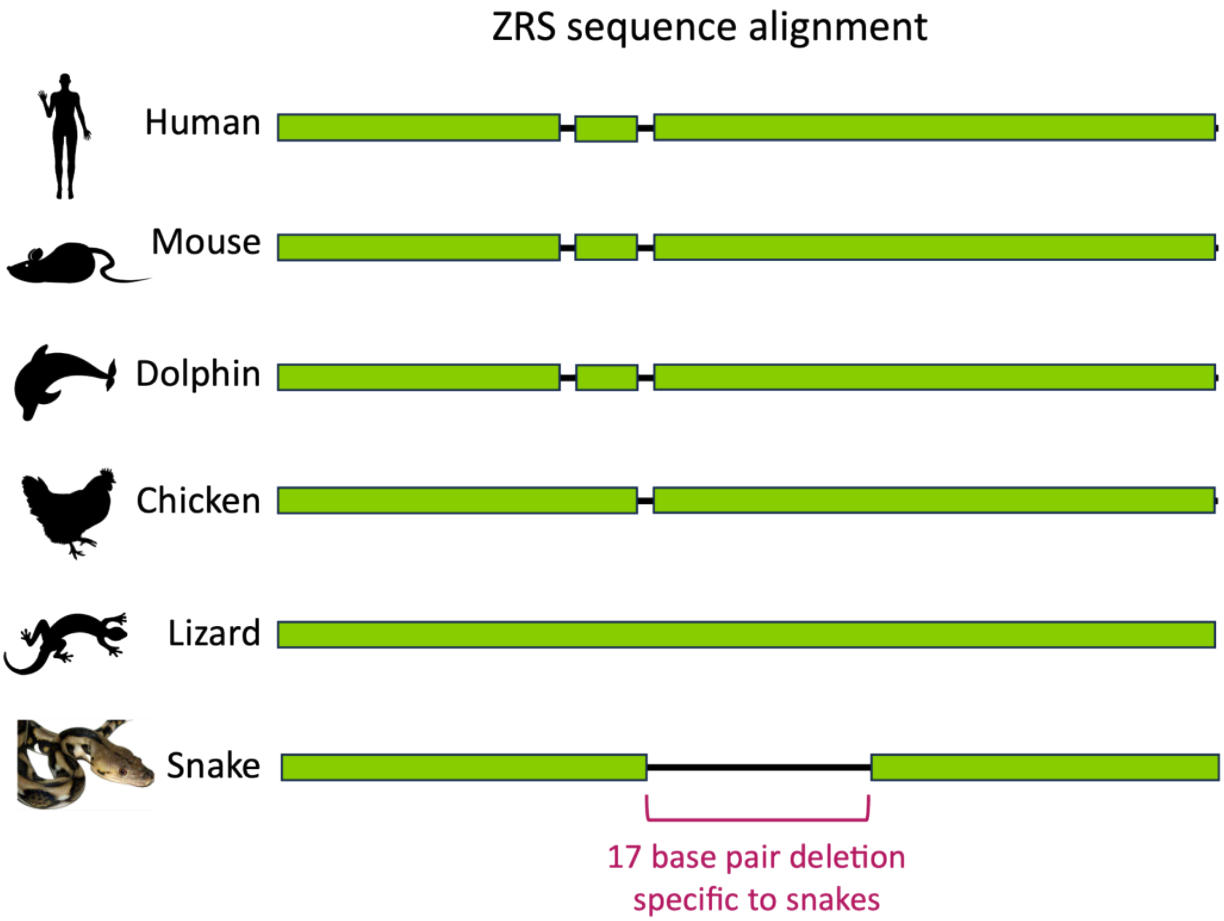


Figure 26. The ZRS enhancer of Shh is highly conserved among animals, but snakes are missing a 17-base pair segment compared to other animals.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=918#h5p-108

4. Kvon, E. Z. et al. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. Cell 167, 633-642.e11 (2016).

Media Attributions

- Figure 22 Fly with Hox gene locations identified is licensed under a Public Domain license
- Figure 23 Hox Genes in Mouse and Human © OpenStax Biology is licensed under a CC BY (Attribution) license
- Figure 24 EG Regulation The Shh gene © Anderson et al, 2014. is licensed under a CC BY (Attribution) license
- Figure 25 EG Regulation ZRS © Hovius et al, 2018 is licensed under a CC BY-NC (Attribution NonCommercial) license
- Figure 26 EG Regulation ZRS sequence alignment © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

HOW DO WE KNOW WHAT ENHANCERS DO? REPORTER GENES EXPERIMENTS DEMONSTRATE PROMOTER ACTIVITY.

How do we know what the enhancer does? One way is to introduce a modified version of the enhancer sequence into the organism's genome, fused to a **reporter gene**.

Reporter genes produce protein products easily visualized in the lab, often through a change in color. Some examples of reporter genes are Green Fluorescent Protein (which makes some jellyfish fluoresce green), luciferase (the enzyme responsible for fireflies "glowing"), and beta-galactosidase (which bacteria use to break down lactose, discussed with the lac operon in the Overview of Gene Regulation chapter). What's important is that the expression of these genes is easy to detect, and these genes are not normally present in the cell or system we are studying so there's no background signal to confuse.

When fused to a promoter or an enhancer we're interested in studying, the reporter gene can "report" whether the sequence is transcribed and translated by the factors normally present in the cell.

For many of the *eve* experiments, a eukaryotic promoter with the stripe 2 enhancer DNA was fused with the bacterial *LacZ* gene sequence. The engineered DNA construct was then reintroduced into *Drosophila* embryos, and the expression of LacZ was monitored. Because the promoter controlling the LacZ gene was fused just with the stripe 2 enhancer, LacZ is expressed in the same region of the embryo as the *eve* stripe 2 but not in the other stripes. This type of experiment confirms the role of this segment of DNA in regulating *eve* expression in stripe 2. An example of these types of experiments is illustrated in **Figure 12**.

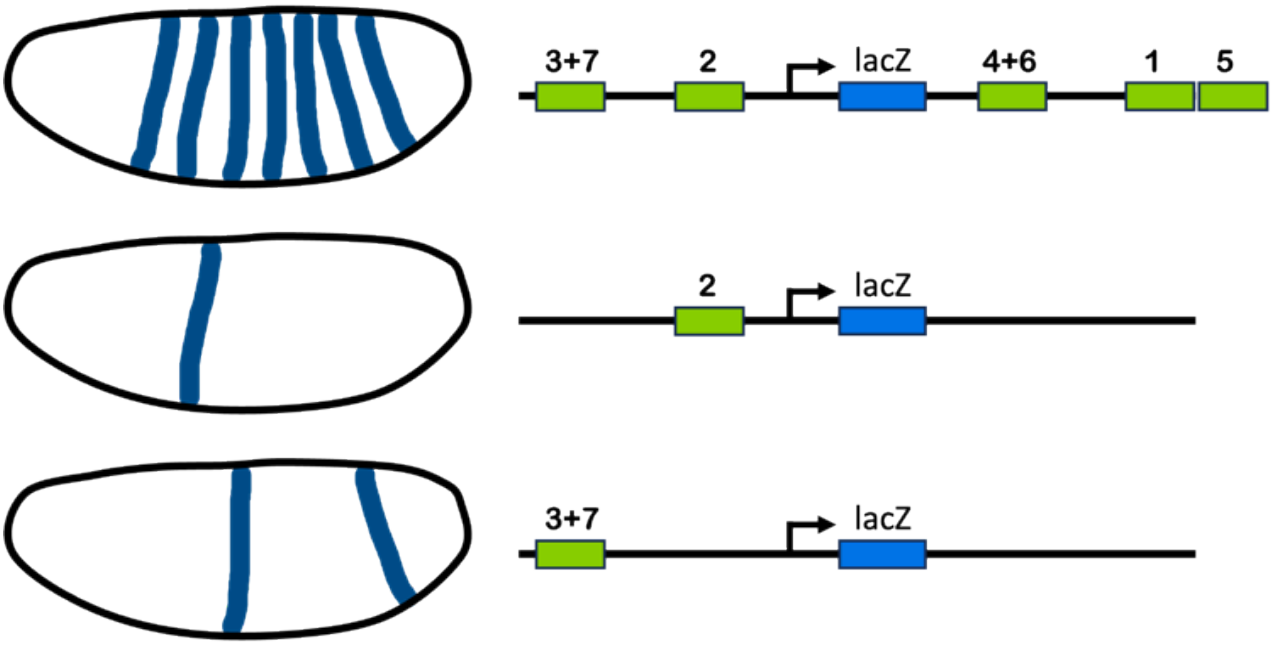


Figure 12. Reporter genes show which parts of the eve promoter are important for expression in each stripe. Top: LacZ reporter gene is linked with the eve regulatory sequence. LacZ expression is detected in the same seven stripes expected for the eve gene. Middle: LacZ reporter gene is linked just with the stripe 2 enhancer. LacZ is only expressed where eve stripe to is expected in the embryo. Bottom: LacZ reporter gene is linked just with the stripe 3+7 enhancer. LacZ is expressed where eve stripes 3 and 7 are expected in the embryo.

But it goes further than this! By engineering enhancers with parts missing or by crossing with flies that are mutant for other genes, we can break down the enhancer into its component parts and identify sequences important for function. For example, LacZ expression vanishes if the bicoid binding sites are mutated in the repressor. The width of the LacZ stripe expands toward the anterior pole if the enhancer is mutated to eliminate Gt binding sites if the reporter gene is expressed in an embryo that is mutant for *giant*.[1,2] These kinds of experiments are examples of **reverse genetics**, where the genome of an organism is specifically manipulated to see an effect on phenotype.

Media Attributions

- Figure 12 EG Regulation © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike)

1. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development 122, 205–214 (1996).

2. Goto, T., Macdonald, P. & Maniatis, T. Early and late periodic patterns of even-skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. Cell 57, 413–422 (1989).

license

CHAPTER SUMMARY

This chapter uses embryonic development as an example for mechanisms of gene regulation. Many of the genes studied here were identified as part of forward genetic screens for developmental mutants.

The *even-skipped* gene is an example of how multiple enhancers can regulate a single gene. There are five classes of genes controlling anterior-posterior patterning in *Drosophila*: maternal effect, gap, pair-rule, segment polarity, and segment identity. *Eve* is a pair-rule gene, regulated by maternal effect and gap genes.

Although much of the work in developmental genetics was done in the fruit fly, the lessons learned are applicable to other organisms as well. The mammalian Hox genes are structurally and functionally similar to those seen in *Drosophila*. The *hedgehog* ortholog *Shh* is as well.

Shh plays a role in the development of multiple organ systems, including the nervous system and limb development. This is controlled by multiple enhancers. A mutation in the ZRS enhancer causes anomalies of limb development in humans. Evolutionarily, a 17 base-pair deletion in the snake ZRS enhancer may explain why snakes do not have legs.

Chromatin structure plays an important role in gene regulation in eukaryotes, with enhancers needing to contact promoters in three-dimensional space, even though they can be hundreds of thousands of base pairs from their target sequence.

Other levels of gene expression are also important. miRNAs are small 20 base RNA molecules that influence the stability of RNA transcripts and can block translation of mRNA. Our understanding of miRNA function has led to the development of RNA techniques for reverse genetic screens as well as potential therapeutics to treat diseases like such as cancer and metabolic disorders.

WRAP UP QUESTIONS

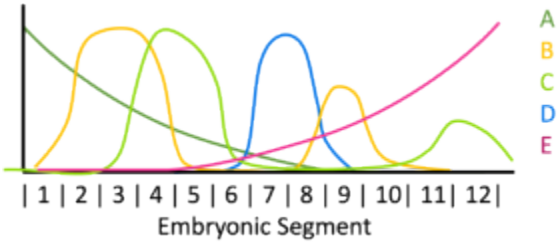
1. To observe a recessive maternal effect phenotype, the mother of an individual must be homozygous for the mutation. The individual's genotype does not affect its phenotype. Many forward genetic screens began by treating flies with a mutagen and screening offspring of the flies for phenotypes. The offspring would typically inherit only one copy of any germline mutant alleles. Dominant germline maternal effect mutations would cause a phenotype change in the F1 generation, but recessive and maternal effect mutations would not.

- Describe a series of crosses that could generate offspring with a recessive mutant phenotype.
- Describe a series of crosses that could generate offspring affected by a maternal effect mutation.

2. Mitochondrial inheritance shows a pattern where all offspring have the same phenotype as their mother. How is that different from the expression patterns you would see for a maternal effect gene?

3. A gene is regulated by the enhancers listed in the table below. In which segments would the gene be expressed?

Enhancer	Activators	Repressors
1	A+B	C
2	D+E	B+C



4. Compare and contrast eukaryotic and prokaryotic gene regulation. How are they similar? How are they different?

5. List four mechanisms by which genes can be regulated other than via transcriptional activators and repressors.

6. There are multiple ways to define a gene. Thinking about the characteristics that make up a gene, would you characterize the ZRS as a gene? Would you characterize the part of the genome that encodes an miRNA a gene? Explain your reasoning.

Science and Society

7. A discussion of the history behind some of the creative *Drosophila* names is found in this 2012 article from Harvard Magazine. Do you agree or disagree with the researchers quoted in the article? What are the benefits and disadvantages of such a system? Should insensitively-named genes be renamed?

Media Attributions

- Wrap Up Question Regulation by Enhancers © Amanda Simons is licensed under a CC BY (Attribution) license

PART X

MENDEL AND BASIC HEREDITY

Objectives

1. Recognize how a scientist's identity and experience affect their science.
2. Define gene, allele, locus, genotype, phenotype, dominant, recessive, monohybrid, dihybrid. Compare and contrast a testcross, a monohybrid cross, a dihybrid cross.
3. Define the law of Equal Segregation and explain how it is illustrated in the progeny of a cross between a homozygous recessive (aa) and heterozygous (Aa) individual.
4. Define the law of Independent Assortment and explain how it is illustrated in the progeny of a dihybrid cross (AaBb x AaBb) or a dihybrid testcross (AABB x AAbb).
5. Use a Punnett square to predict potential offspring from a given cross.
6. Use the Addition Rule and Multiplication Rule of probability to predict the likelihood of an outcome.

What makes you, you?

What do you look like? Do you have brown eyes or green eyes? Are you tall, short, or medium height, and do you have light skin or dark skin?

How do you behave? Do you run fast or are you very flexible? Do you have perfect pitch? Are you a thrill-seeker, or do you prefer more relaxing activities?

These types of measurable traits – whether they are easily visible, like physical attributes, or other measurable characteristics, like behavior or aptitude for a skill – make up your **phenotype**. The traits listed above all are likely to have an underlying genetic cause: a small variation in a part of your DNA (a **gene** or a **locus**), compared to other people, that may predispose you to a certain appearance or behavior. These different versions of a gene are called **alleles**, and your particular combination of alleles is called your **genotype**.

Classical genetics, **transmission genetics**, and **Mendelian genetics** are three terms that all refer to the same thing: genetic research that relies on tracking traits from parent to offspring. Classical genetic experiments

compare the phenotypes of parents and offspring to infer their genotype. Studying how traits are passed from one generation to the next was historically the earliest type of genetic research, beginning with the work of Gregor Mendel in the mid-1800's.

This module gives a brief biography of Gregor Mendel and his contribution to the field of genetics, including the laws of equal segregation and independent assortment. You'll practice using Punnett squares and the addition and multiplication rules of probability.

GREGOR MENDEL

Gregor Johann Mendel (1822–1884). (Courtesy of Professor William Bateson, London.)

Figure 1. Gregor Mendel.

Gregor Mendel, shown in **Figure 1**, is sometimes referred to as the father of genetics. He was one of the first researchers to use mathematics to systematically study patterns of inheritance, and his conclusions form the basis of our modern understanding of heredity.

Mendel has an interesting biography that demonstrates the importance of a scientific community with scientists from diverse backgrounds[1]. Science is meant to be impartial, and when we discuss the scientific method we often assume that a scientist's background is irrelevant because data analysis should be objective.

However, a person's background, identity, and experiences have a profound effect on the kinds of research questions they choose to pursue and the tools they choose for analysis. Their place in the scientific community also can impact how their work is considered by other scientists. Mendel's story demonstrates both the importance of diversity in science and the consequence of overlooking those perceived to be outsiders.

Mendel was essentially the 19th century equivalent of a first-generation college student, coming from a farming family with minimal formal education. He struggled to pay for school among wealthier classmates. He attempted (and failed at!) several jobs before finding his way to the discovery that forms the foundation of classical genetics. And then no one paid any attention to his discoveries for decades because he was very much an outsider to the scientific community. But it was the exact parts of his "outsider identity" that made it possible for him to make these discoveries. His identity drove his scientific curiosity, his approaches to research, and his ability to draw conclusions from his observations. In bringing his own unique experiences to spaces where his background was not common, he was able to make unique discoveries.

1. Corcos, Alain F., M., Floyd V. *Gregor Mendel's Experiments on Plant Hybrids: A Guided Study*. (Rutgers University Press, 1993).

Figure 2. The Czech Republic

Mendel lived from 1822-1884 in what is now the Czech Republic (**Figure 2**). Mendel was not a geneticist by training – genetics didn't even exist back then. He also did not set out to discover the laws of heredity. Mendel grew up on a family farm, lived in a farming community, and had a much more practical interest: crop production.

Mendel's local village did not have many educational opportunities, but after showing promise in early studies in a local grammar school and nearby gymnasium (secondary school), Mendel was sent to pursue more advanced studies in physics and philosophy at the University of Olmütz. Unlike many of his wealthier classmates, Mendel worked to earn the money to pay his way through school (both gymnasium and university). Mendel, by many accounts, was recognized as very smart but struggled in his academic career. At several points Mendel needed to take time off from his schooling to recover from health problems, likely exacerbated by overwork as he juggled jobs and schooling.

After the University of Olmütz, Mendel joined an Augustinian monastery in what is now Brno, in the Czech Republic (shown in **Figure 3**). The monastery's mission as a community and cultural center in Brno meant that many of the monks contributed to the community through teaching or medicine. Here, again, Mendel struggled.

Figure 3. St. Thomas's Abbey, Brno, Czech Republic.

Mendel first trained as parish priest. As part of such a position, Mendel would be expected to minister to the ill or infirm. However, he found he was not suited to such a role as he often became physically ill after working with suffering parishioners. The medicine part was not for him. Later, Mendel trained – and practiced – as a teacher. But Mendel failed the final licensure to become a teacher.

Despite this, the abbot of the monastery recognized Mendel's potential. He recommended that Mendel attend the University of Vienna to fill in the gaps in his education that were revealed by the licensure exams. In 1851, Mendel was sent to the University of Vienna to further his training.

At the University of Vienna, Mendel studied with two mentors who particularly influenced his future thinking: The first was Christian Doppler, who used mathematics to explain natural phenomena. (Doppler is most well-known for his work on the Doppler effect, which describes how the perception of sound waves change when the source of a sound moves toward or away from the listener.) Mendel's second mentor was Franz Unger, a botanist who applied the laws of physics and chemistry to studying plants. This kind of cross-disciplinary training would ultimately be exactly what brought Mendel to his work with plants and heredity.

When Mendel returned to the monastery after his studies at the University of Vienna, he resumed his teaching career. However, he also began a carefully controlled series of experiments on pea plants, intended to study the inheritance of traits in hybrid crops. Why hybrids? Remember that Mendel lived in a farming community. Because of his background and the role the monastery played in the agrarian community of Brno, Mendel was interested serving his community by studying a phenomenon called hybrid vigor. At the time of Mendel's work, farmers had long been aware that when two different strains of crops are crossed, the resulting offspring (called a **hybrid** of the two parents) often are more vigorous – with larger fruits and healthier plants –

than either of the parents. But hybrid plants rarely breed true: the offspring of two hybrid plants do not always share that vigor. He approached this problem with the intention of improving crop production.

Mendel's family upbringing, combined with his role at the monastery, gave him the insight, motivation, and resources to ask these research questions. In his work, he also applied lessons he learned from his education in physics, chemistry, and mathematics. His family, his education, his interests, and his vocation all converged, perfectly positioning him to discover rules of heredity.

So what information did Mendel start with? At the time Mendel began his work, scientists had only a limited understanding of how traits were passed from generation to generation, from parent to offspring. Biologists understood that offspring tend to look more like their biological parents than unrelated adults. They also knew that some characteristics (such as the size or color of fruit, or the length or color of a dog's coat) varied between individuals, and that crops and animals could be carefully bred to select for the most favorable traits.

One early hypothesis of how traits were passed from parent to offspring was that offspring had characteristics that blended those of the parents: a tall parent and a short parent might produce medium-height offspring, for example. And, in many individual cases, characteristics do appear to be inherited in this fashion: a tall and a short parent *might* have a medium-height child. However, there are also many examples where offspring do not show blending of parental characteristics; for example, a child of a brown-eyed parent and a blue-eyed parent will often have brown eyes, not an intermediate color halfway between brown and blue. Mendel recognized that hybrid vigor also ran counter to the blending hypothesis.

Through a series of carefully designed, carefully analyzed experiments, Mendel proposed two basic "laws" of heredity: one hundred and fifty years later, geneticists still use these laws as a foundation in understanding genetics. Mendel published his work in a paper titled, "Experiments on Plant Hybridization."

However, in part due to Mendel's role outside the scientific community, Mendel's work received very little attention. His paper was largely ignored until it was "rediscovered" in the early 1900's, stalling work on genetics for forty years and illustrating the consequences when established scientists ignore the work of outsiders.

Media Attributions

MENDEL'S EXPERIMENTS

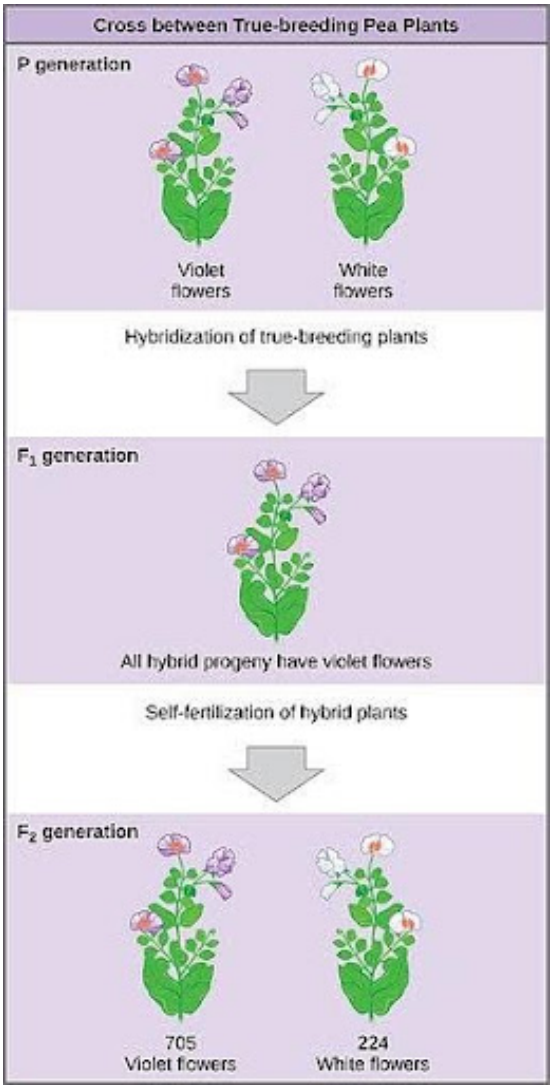


Figure 4. An example of Mendel's crosses. Mendel crossed purple-flowered pea plants with white-flowered pea plants. All of the F1 offspring had purple flowers. The F1 plants were allowed to self-fertilize, and the F2 generation included both purple and white flowered plants.

Mendel used pea plants as a **model organism** to study inheritance. A model organism is one that is studied to draw more general conclusions about biology. For example, although Mendel studied peas, his conclusions are applicable to all diploid organisms.

Why peas? Model organisms are usually chosen for practical reasons. Peas are easy to grow and have a short growing season. From seed to harvest takes about two months, so Mendel could grow many generations of plants in a year. There are many easily identifiable variations of pea plants, so Mendel could track the inheritance of these traits. And the flowers of the plants – the reproductive organs – are large enough to be easily manipulated to ensure controlled crosses between plants.

Mendel had several **true-breeding** strains of pea plants, each of which varied in one of seven characters: seed shape, seed color, flower color, pod shape, pod color, stem placement, and size. True-breeding means that, generation after generation, these traits breed true: all offspring look like their parents. Mendel carefully controlled breeding to cross these true-breeding strains, transferring pollen from the stamen of one plant to the stigma of another plant.

He recorded the appearance of thousands of offspring over several generations.

Because of Mendel's background in physics, chemistry, and math, he was prepared to recognize certain mathematical patterns. Just like chemical reactions consistently produced characteristic ratios of products, he noticed that crosses produced offspring with predictable traits, and those traits were present in characteristic ratios within the population of offspring. An example of one such cross is shown in **Figure 4**.

If he crossed plants that varied in a single character – purple flowers and white flowers, for example – he found that the first generation of offspring always matched the appearance of one of the parents. There was no blending!

In modern terms, we call the parents the **P generation** and the offspring the **F1 generation.** F1 stands for first filial generation – filial means child.

For the example of the P generation of purple crossed with white flowers, all the F1 offspring were purple. If two F1 offspring were crossed, though, both traits of the original parents would be seen in the next generation, called the F2 generation. For the purple flower x white flower cross, there were both purple and white F2 offspring. Mendel called purple the "dominating" trait; in modern terms we use the word **dominant.** Mendel called white the **recessive** trait, a term that we still use today.

This is where Mendel's math and chemistry background came in. Mendel was used to calculating the ratios of products to reactants in his studies of chemistry, just like you might have done when balancing reactions in your chemistry class. So, when Mendel observed a mixture of traits in the offspring, he calculated the ratios in which they appeared in the population. Because of his unique educational background, Mendel was one of the first researchers to think to apply such quantitative methods to studies of biology.

Mendel logged the traits of thousands of pea plant offspring, as shown in **Table 5**. Mendel realized that, in the F2 generation, the ratio of the dominant trait to the recessive trait was always approximately 3:1, or ¾ dominant and ¼ recessive.

Further self-fertilization of the F2 revealed that the recessive trait was always true-breeding. Some of the dominant F2 were also true-breeding, but other dominant F2 offspring were not. He thus revised his ratio to 1:2:1 true-breeding dominant: hybrid dominant (that did not breed true): true-breeding recessive.

Table 5 Mendel's results. For each of the seven characters Mendel tested, one trait was always dominant and one recessive. In the F2 generation, the dominant trait always outnumbered the recessive trait at a ratio of approximately 3:1.

Trait	Dominant vs Recessive	F2 Generation (Dominant)	F2 Generation (Recessive)	Ratio
Seed Shape	Round vs wrinkled	5474	1850	2.96:1
Seed Color	Yellow vs green	6022	2001	3.01:1
Flower color	Purple vs white	705	224	3.15:1
Pod shape	Inflated vs constricted	882	299	2.95:1
Pod color	Yellow vs green	428	152	2.81:1
Flower position	Axial vs terminal	651	207	3.14:1
Plant height	Tall vs short	787	227	2.84:1

Media Attributions

- Mendel's crosses © OpenStax is licensed under a CC BY-SA (Attribution ShareAlike) license

MENDEL'S FIRST LAW

The relative numbers of offspring of each phenotype that Mendel observed formed the basis for his Laws of Heredity. The first law addresses how alleles of a single gene are passed to offspring.

In particular, the ratio of 1 true-breeding dominant: 2 hybrid dominant: 1 true-breeding recessive offspring that Mendel observed from each F1 self-cross allowed Mendel to conclude that each individual organism has two hereditary elements that specify each trait. In a population, there are different versions, or **alleles**, of these hereditary elements, which contribute to the diversity of traits observed. For example, for the genetic element that specifies plant height, one allele specifies tall plants, while another specifies short. Each individual organism has two alleles that specify each trait, or, in modern terms, two copies of each gene.

Mendel concluded that, although an individual has two separate alleles controlling each trait, only one allele would be passed to offspring during the production of a gamete. Which of the two alleles inherited was random, so if all of the population from a cross were compared, 50% of gametes inherit one allele, and 50% the other. He summed this up in his first law of heredity, called **The Law of Equal Segregation**.

The Law of Equal Segregation: During the production of gametes, the two alleles of a gene are divided (segregated) among gametes, so that each gamete receives only one allele. This results in equal numbers of gametes with each allele.

What this means in simpler terms is that an individual has two copies of each gene. When that individual reproduces, half of its offspring get one allele, and half get the other allele.

TERMINOLOGY AND NOTATION

The form of genetic notation that we still use for describing genetic alleles comes from Mendel's original paper. Mendel used a capital letter (A) to describe a dominant trait, lowercase (a) to describe a recessive trait, and both together (Aa) to describe a hybrid that did not breed true. Mendel's F2 ratio of 1:2:1 could be more described by genotype as 1AA:2Aa:1aa.

Although Mendel's work focused on pea plants, his work is applicable for all diploid organisms. When an organism has two of the same allele (AA or aa, in Mendel's notation), the organism is said to be **homozygous** for that gene. When an organism has two different alleles (Aa in Mendel's notation), the organism is said to be **heterozygous**. For a trait controlled by dominant and recessive alleles of a single gene, heterozygous individuals will always show the dominant trait. The notation can also be used to describe alleles of more than one gene at a time. If so, each gene is indicated by a different letter. "AABb" would describe the genotype of an individual that is homozygous dominant for gene A and heterozygous for gene B.

Some additional terminology:

Homozygous individuals are often described as **true-breeding**, since a cross between two homozygous individuals with the same phenotype will always give offspring with that same phenotype. **Heterozygotes** – individuals who are heterozygous – are often referred to as hybrid. A **monohybrid** is heterozygous for a single gene (Aa); a **dihybrid** is heterozygous for two genes (AaBb); a **trihybrid** is heterozygous for three genes (AaBbCc), etc. Note that these terms are used to highlight the genes we are tracking in an experiment: a dihybrid individual actually has tens of thousands of genes! But we're just paying attention to two of them.

We can describe crosses by the types of parents involved. Some specific examples are listed below.

- A **monohybrid cross** is between two individuals both heterozygous for one gene: Aa x Aa.
- A **dihybrid cross** is between two individuals both heterozygous for two genes. AaBb x AaBb
- A **self-cross** is when the same individual contributes both gametes. This is not possible for sexually dimorphic organisms like humans, but it is possible for organisms like plants which can self-pollinate.
- A **back-cross** is when an individual is crossed "back" with a parent. In addition to happening in controlled lab settings (like crossing pea plants, fruit flies, or mice), it is also common in agricultural settings where plants or animals are being selected for desirable traits.

We will use this terminology – and add to this list – in later modules of this text.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=295#h5p-45

PUNNETT SQUARES

Although Mendel's work was largely ignored for forty years after it was published, its rediscovery in the early 1900's sparked a revolution in the study of heredity and the establishment of genetics as a distinct field of study. Among the early genetics researchers was Reginald Punnett, who expanded Mendel's studies to other organisms. Reginald Punnett devised what we now call the Punnett square, which is a tool to visually depict all possible progeny and their expected ratios from a controlled cross[1].

An example: In Mendel's crosses, each of Mendel's F1 offspring were monohybrid (Aa). The F1 self-cross Aa x Aa gave F2 progeny in a phenotypic ratio of 3 dominant: 1 recessive, or a genotypic ratio of 1 AA : 2Aa: 1aa. A Punnett square depicts this cross in a table.

Across the top row of the table, the possible gametes that an individual can produce are written. An individual with genotype Aa can produce a haploid gamete with either the A allele or the a allele:

	A	**a**

Down the first column of the table, the possible gametes of the second parent are written, each in its own row:

	A	**a**
A		
a		

Then, the remainder of the table is filled in by filling each box in with the gametes that correspond to each row and column:

1. Mendelism / by R.C. Punnett. *Wellcome Collection* https://wellcomecollection.org/works/u5wbhqtr/items.

	A	**a**
A	**A**	**a**
a	**A**	**a**

⇓

	A	**a**
A	**A A**	**A a**
a	**A a**	**a a**

These four boxes represent the combinations of alleles that are possible among offspring. These are the genotypes of potential offspring. You'll see that there is one box with genotype AA, two boxes with genotype Aa, and one box with aa. This is a visual representation of the 1AA : 2 Aa : 1 aa genotypic ratio Mendel observed. Note that, by convention, for a heterozygote we usually write the dominant allele first.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=297#h5p-46

The Punnett square gives us a tool to predict the possible offspring of any cross, given the genotypes of the parent. It also, however, can be used in reverse, to determine the genotype of a parent.

An example: Individuals with a dominant phenotype can either be homozygous dominant (AA) heterozygous (Aa). When the second allele is unknown, the genotype can be written A_, with the underscore indicating that uncertainty. How could we determine that unknown allele? One way to do this is to perform a **testcross**.

A testcross is a cross that includes a parent with the recessive trait. The recessive parent is called the **tester.** An individual with the recessive phenotype must be homozygous recessive, so the phenotypes of the offspring

will reveal the genotype of the unknown parent. If any offspring have a recessive phenotype, the unknown allele must be *a* since we see recessive offspring.

	A	_
a	dominant	recessive
a	dominant	recessive

Conclusion from the phenotypes:
the genotypes must be
⇓

	A	**a**
a	**Aa**	**aa**
a	**Aa**	**aa**

A testcross can be used with multiple genes at a time. So, for example, an individual with a phenotype of two dominant traits (A and B) and a genotype of A_B_ could be testcrossed with an individual of genotype aabb.

Testcrosses give characteristic offspring ratios. A **monohybrid testcross** (Aa x aa) is a testcross of a monohybrid individual. Note that this is different from a **monohybrid cross** (Aa x Aa).

A monohybrid testcross always gives offspring in a 1:1 ratio of dominant to recessive. Or, in other words, ½ of the offspring are dominant, and ½ are recessive. So, if a cross gives two different phenotypes in a 1:1 ratio, we can usually conclude that it was a monohybrid testcross and our unknown parent is heterozygous.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=297#h5p-47

A **dihybrid testcross** (AaBb x aabb) gives four potential combinations of phenotypes: AB, Ab, aB, and ab.

These combinations can be called **phenotypic classes**. For a dihybrid testross, the four phenotypic classes will all be present in roughly equal numbers. This is a 1:1:1:1 ratio. In the next section of the text, looking at Mendel's Second Law, we will look at the biological principles behind this ratio.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=297#h5p-48

MENDEL'S SECOND LAW

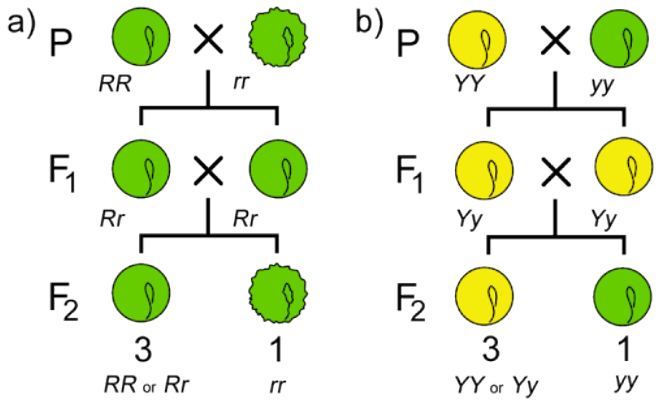


Figure 6. F1 Monohybrid crosses involving two distinct traits in peas. A) A parental cross between true-breeding round and wrinkled peas is followed by an F1 self-cross. ¾ of the F2 offspring are Round (RR or Rd) and ¼ are wrinkled (rr). B) A parental cross between true-breeding Yellow and green peas is followed by an F1 self-cross. ¾ of the F2 offspring are Yellow(YY or Yy) and ¼ are wrinkled (yy).

Mendel's second law addresses the inheritance of alleles of multiple genes.

Mendel's initial experiments with pea plants looked at what happens when plants that vary in one characteristic – for example, crossing plants with yellow seed and green seeds. He followed this up with experiments where plants varied in multiple characteristics – for example, crossing plants with round, yellow seeds and wrinkled, green seeds. He wanted to know if traits would segregate independently or whether traits that were coupled in the parents would be inherited as a single unit.

Mendel crossed peas that varied in two traits: seed color and seed shape. He crossed a true-breeding plant that had yellow and round seeds with a true-breeding plant that had green and wrinkled seeds. All the F1 offspring were Yellow and Round. The F1 offspring were dihybrids, meaning they were heterozygous at two genes (genotype YyRr). Mendel then self-crossed the F1. The F2 offspring could have any combination of traits: Yellow and round or green and wrinkled like the parents, or a **recombinant** phenotype of yellow, wrinkled or green, round.

For example, if he crossed parental peas with Yellow and Round seeds x green and wrinkled seeds, offspring could have any combination of traits: Yellow and round or green and wrinkled like the parents, or a **recombinant** phenotype of yellow, wrinkled or green, round.

He found that traits that were present together in the parents were not necessarily inherited together in the offspring. This led him to the **Law of Independent Assortment.**

Law of Independent Assortment: The distribution of alleles from one gene to offspring is not dependent on the distribution of alleles from another gene.

Or, in other words, when gametes are produced, the alleles from each gene are passed to the next generation independently of others.

Recall that in Mendel's first set of experiments, there was a 3:1 ratio of dominant: recessive traits from a monohybrid cross (Aa x Aa). This is shown in **Figure 6.** Also recall: a monohybrid cross indicates that we are

tracking the inheritance of one gene, but the organism itself still has tens of thousands of other genes that we are not paying attention to.

In a **dihybrid** cross (AaBb x AaBb), the 3:1 ratio still holds true for each of the genes. Yellow, Round seeds crossed with green, wrinkled seeds will give an F1 dihybrid with the Yellow and Round phenotypes. When the F1 dihybrid is self-crossed, ¾ of the offspring are Yellow, and ¼ of the offspring are green. ¾ of the offspring are Round, and ¼ are wrinkled. But how many of the Yellow seeds are also Round? How many of the Yellow seeds are wrinkled?

Mendel's results showed a 9:3:3:1 ratio of Yellow, Round: Yellow, wrinkled: green, Round: green, wrinkled. Or, in other words:

- 9/16 of the offspring have both dominant traits (Yellow and Round)
- 3/16 of the offspring have one dominant and one recessive trait (Yellow and wrinkled)
- 3/16 of the offspring have one recessive and one dominant trait (green and Round)
- 3/16 of the offspring have both recessive traits (green and wrinkled)

This is illustrated in **Figure 7.**

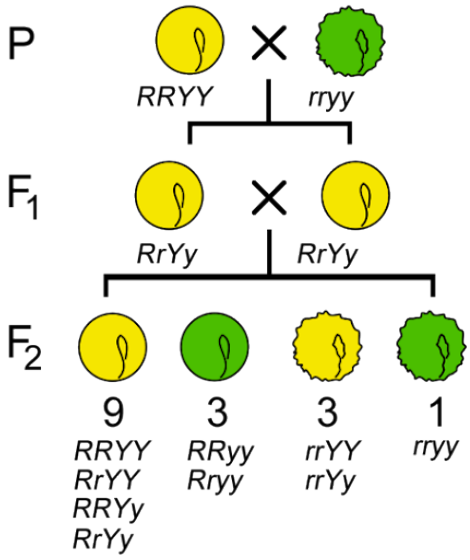


Figure 7. Pure-breeding lines are crossed to produce dihybrids in the F1 generation. The cross of these particular dihybrids produces four phenotypic classes, present in a 9:3:3:1 ratio. Image source: Open Online Genetics, by Nickle and Barrette-Ng. (Original-Deyholos-CC:AN)

These observed numbers are consistent with the probability of two independent events occurring together.

The **multiplication rule of probability** allows us to calculate the probability of two independent events **both** occurring, just by multiplying the two independent probabilities together. For example, if the round and the yellow genes are inherited independently, the probability of obtaining offspring that are both round **and** wrinkled is the probability of round offspring multiplied by the probability of wrinkled offspring. This is illustrated in the equation below, where P stands for Probability.

$$P_{\text{round and yellow}} = P_{\text{round}} * P_{\text{yellow}}$$

Using the multiplication rule:

In a monohybrid cross for seed color, we would expect ¾ of the progeny were yellow. In a monohybrid cross for seed shape, we would expect ¾ of the progeny were round. Therefore, for a dihybrid cross for seed color and shape, ¾ round × ¾ wrinkled = 9/16 of the progeny would be both yellow and round. Likewise, ¾ × ¼ = 3/16 of the progeny would be both round and green and so on. The calculations for all four potential combinations of traits are shown in **Table 1.**

Table 1 Potential Combinations

Phenotypic class	Probability of seed color	Probability of seed shape	Probability of combined phenotype
Yellow and Round	¾	¾	¾ * ¾ = 9/16
Yellow and wrinkled	¾	¼	¾ * ¼ = 3/16
green and Round	¼	¾	¼ * ¾ = 3/16
green and wrinkled	¼	¼	¼ * ¼ = 1/16

The multiplication rule of probability thus predicts a 9:3:3:1 ratio of phenotypic classes for a dihybrid cross if the alleles are indeed inherited independently. Mendel observed close to a 9:3:3:1 ratio for all combinations of traits he reported.

The Law of Independent Assortment means that we can treat genes independently, even if we are tracking more than one trait at a time.

Exceptions to the Law of Independent Assortment

Mendel did have a bit of luck: all of the traits he chose are encoded on different chromosomes in the pea genome. Since Mendel's time, it has become clear that genes located close together on the same chromosomes (so-called **linked genes**) do not follow the law of independent assortment. These exceptions to the rule will be discussed in the module on Linkage.

Media Attributions

- Monohybrid crosses © Nickle and Barrette-Ng via. Open Online Genetics is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Pure breeding © Nickle and Barrette-Ng via. Open Online Genetics is licensed under a CC BY-SA (Attribution ShareAlike) license

USING THE RULES OF PROBABILITY TO SOLVE PROBLEMS

The multiplication rule of probability states that the probability of one event **and** a second event **both** occurring is the product of the probabilities of each individual event occurring separately. This can be used in multiple ways in genetics.

Predict the probability of offspring with a particular combination of traits

In the previous section, the multiplication rule was used to explain the 9:3:3:1 phenotypic ratio observed from a dihybrid cross. In a similar fashion, we can use it to predict the probability of offspring with a particular combination of traits, regardless of how many genes we track.

For example, let's calculate the probability of an offspring with all recessive traits for the cross:

AaBbCc x AabbCc

We draw three separate Punnett squares, one for each gene, as shown in Figure 8. As seen by the Punnett squares, there is a ¼ probability of the recessive "a" phenotype. There is a ½ probability of the recessive "b" phenotype. And there is a ¼ probability of the recessive "c" phenotype. The probability of an offspring that has all three recessive traits is ¼ * ½ * ¼ = 1/32.

	A	a
A	AA	Aa
a	Aa	aa

	b	b
B	Bb	Bb
b	bb	bb

	C	c
C	CC	Cc
c	Cc	cc

Figure 8 Punnett squares for the cross AaBbCc x AabbCc. Genotypes that give the recessive phenotype are highlighted.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it

online here:
https://rotel.pressbooks.pub/genetics/?p=304#h5p-49

Additional applications of the rules of probability

In previous sections, the product rule of probability was used to calculate the probability of obtaining offspring with one trait **and** another independently assorting trait. But the probability rule can be used in any circumstance to calculate the probability of two or more events **all** happening. For example, the probability rule could be used to calculate the probability that, in a monohybrid cross, two offspring in a row will be heterozygous. If the probability of one heterozygous offspring is ½, the probability of two in a row would be:

½ * ½ = ¼.

A non-biology example: if the probability of heads on a coin flip is ½, the probability of two heads in a row is ½ * ½ , the probability of three heads in a row is ½ * ½ * ½, and the probability of four heads in a row is:

½ * ½ * ½ * ½.

The multiplication rule is used to calculate probability of one thing **and** another **both** occurring.

The **addition rule of probability**, on the other hand, is used to calculate the probability of one thing **or** another occurring. For example, in a monohybrid cross, we'd expect ¼ AA offspring, ½ Aa offspring, and ¼ aa offspring. The probability of any single offspring having a genotype of AA **or** aa is ¼ + ¼ = ½.

We will see additional uses for both rules of probability in later chapters.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=304#h5p-50

Media Attributions

- Punnett squares © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

USING AN EXPANDED PUNNETT SQUARE

The multiplication rule of probability offers a straightforward way to quickly calculate the probability of offspring with a particular combination of traits. However, sometimes it is nice to have a visual representation of all of the possible progeny from a particular cross. In this case, an expanded Punnett square can be constructed. In this section, we construct a two-gene Punnett square for a dihybrid cross, using the F2 offspring of Mendel's Yellow, Round x green, wrinkled cross as an example (RrYy x RrYy).

Note that the rows and columns do NOT list alleles individually, but possible combinations of alleles.

Although it is tempting to think of the Punnett square columns and rows as simply separating individual alleles, what the columns and rows actually represent is the *potential gametes* that a parent can produce. For example, a monohybrid individual (Yy) can produce gametes with Y or y. A dihybrid individual RrYy can produce gametes with RY, Ry, rY, and ry allelic combinations, so these are the column and row headers that we use for a dihybrid Punnett square. This is shown in **Figure 9A**. As in the single-gene Punnett square, the rows and columns are filled in according to the headers, as shown in **Figure 9B**.

When writing genotypes for two or more independently assorting genes, the convention is to write genotypes with both alleles of one gene, followed by both alleles of the next gene (*RrYy*, not *RYry*). By convention, any dominant alleles are written first (RrYy, not rRyY).

Figure 9B also labels the phenotypic classes associated with each genotype. Note that there are 16 possible offspring in the two-gene Punnett square. Of them, 9/16 are round and yellow. 3/16 are round and green, 3/16 are wrinkled and yellow, and 1/16 are wrinkled and green. This is the same 9:3:3:1 offspring ratio we saw when using the multiplication rule.

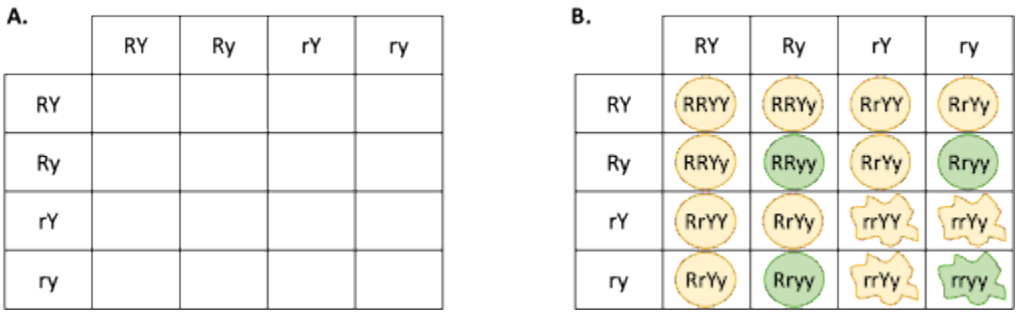


Figure 9. A two-gene Punnett square. **Panel A:** The parental gametes produced for the cross RrYy x RrYy. **Panel B:** The completed square, with genotypes of progeny. The phenotypes of the progeny are also indicated: 9 round and yellow, 3 round and green, 3 wrinkled and yellow, 1 wrinkled and green.

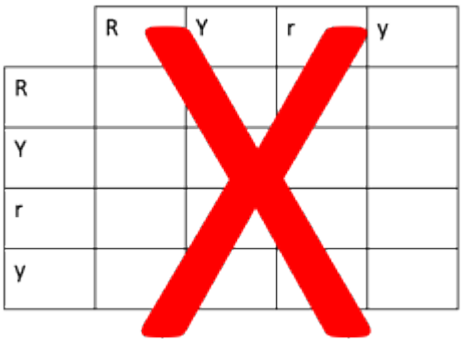


Figure 10. This is the wrong way to set up a two-gene Punnett square. Don't do this.

WARNING! A common mistake is to simply split up the four alleles from the dihybrid, as shown in **Figure 10**. But a dihybrid individual does NOT produce gametes of A, a, B, and b, and it is incorrect to set up a Punnett square with these headers for the columns and rows. You can check your work by making sure that each header has one copy of every gene, and each box representing offspring has two copies of each gene.

Larger Punnett squares can be used to track more genes at a time. The number of rows and columns in an expanded Punnett square is equal to 2^n, where n=# of genes. So a single-gene Punnett square has 2^1=2 rows and 2 columns, and a two-gene Punnett square has 2^2=4 rows and columns. A three-gene Punnett square fould have 2^3=8 rows and columns! You can see that constructing an error-free Punnett square becomes exponentially more challenging with additional genes. Although the expanded Punnett squares can be a useful visual representation under some circumstances, the multiplication rule is generally a faster (and less error-prone) way to solve multigene problems.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=308#h5p-51

Media Attributions

- A two gene Punnett square © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Wrong Punnett square © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

SUMMARY

Gregor Mendel's biography is a great example of how a scientist's background and identity influences their science. Because of his background coming from a farming community, with interests in botany, chemistry, and math, Mendel was able to develop two rules of genetics that we still use today: the law of equal segregation and the law of independent assortment.

The law of equal segregation gives us the principles we use to construct a Punnett Square and calculate the probability of offspring phenotypes. The law of independent assortment allows us to calculate offspring phenotypes derived from multiple genes at a time, using the multiplication rule of probability.

WRAP-UP QUESTIONS

Questions 1-4 modified From *Online Open Genetics* (Nickle and Barrette-Ng)[1].

1. Wiry hair (*W*) is dominant to smooth hair (*w*) in dogs.

   a. If you cross a homozygous, wiry-haired dog with a smooth-haired dog, what will be the genotype and phenotype of the F1 generation?
   b. If two dogs from the F1 generation mated, what would be the most likely ratio of hairphenotypes among their progeny?
   c. When two wiry-haired *Ww* dogs actually mated, they had a litter of three puppies, which all had smooth hair. How do you explain this observation?
   d. Someone left a wiry-haired dog on your doorstep. Without extracting DNA, what would be a way to determine the genotype of this dog?

2. An important part of Mendel's experiments was the use of homozygous lines as parents for his crosses. How did he know they were homozygous, and why was the use of the lines important?

3. Does equal segregation of alleles into daughter cells happen during mitosis, meiosis, or both?

4. A rare dominant mutation causes a neurological disease that appears late in life in all people that carry the mutation. If a father has this disease, what is the probability that his daughter will also have the disease?

5. In dogs, long hair is dominant to short and black fur is dominant to brown. A long-haired black dog and a long-hair brown dog have a short-haired brown puppy. Give the genotypes for both parents and the puppy.

6. Individuals with genotype AaBbCc and AaBBcc were crossed. What is the likelihood that the first offspring will have all three dominant traits?

1. Nickle and Barrette-Ng. Open Online Genetics. in *Open Online Genetics* (2016).

Science and Society

7. Although science is meant to be objective, a scientist's identity and background strongly informs the questions that are asked and the way data is analyzed. Mendel investigated hybrids because he came from a community where that was really important. He analyzed biological data with math, because he learned to do that from studies in chemistry/physics. What parts of your identity drive your scientific curiosity? What other strengths do you bring to your studies in biology and genetics?

PART XI
ALLELE INTERACTIONS

Objectives

1. Define gain of function, loss of function, and haploinsufficient. Explain why gain of function mutations are often dominant and loss of function mutations are often recessive.
2. Define the following terms, and give examples of each: complete dominance, incomplete dominance, codominance, penetrance, expressivity, lethal allele, pleiotropy
3. Predict offspring ratios from a cross involving any of the terms defined above
4. Use offspring ratios to predict the mode of inheritance and/or determine genotypes of parents
5. Explain how a single allele can be described in multiple ways (dominant, recessive, incompletely dominant, codominant, etc) depending on the phenotype that is observed.

Genes contribute to an organism's phenotype primarily through the function of the protein that the gene encodes. If the gene sequence – or the sequence of regulatory elements that affect gene expression – is changed, then the function of the protein might be affected. Changing protein function can sometimes affect phenotype.

When one phenotype predominates in a population, that trait is called **wild-type.** As an example, in wild populations of fruit flies – ones you might find circling your overripe bananas! – eye color is red, wings are longer than the body, and bristles are straight. For fruit flies, then, wild-type eye color is red, wild-type wings are long, and wild-type bristles are straight. Phenotypic differences are described as **mutant** in relation to those wild-type trait. Mutant fruit flies might have brown or scarlet eyes, short or curly wings, forked or bent bristles – or any combination of these traits. It's important to note that the word "mutant" does not have negative connotations to a geneticist: mutations are just differences .

Note that the term wild-type is not always appropriate to use: it may not be used if there is enough variation in a population that there is no single predominant phenotype. The term wild-type is also not typically used to describe human phenotypes. The term mutant is sometimes used to describe human allelic variants associated with disease (eg. a mutant CFTR allele) but never people themselves.

ALLELE FUNCTION: WHY ARE PHENOTYPES DOMINANT OR RECESSIVE?

The wild-type phenotype is often the result of fully functional alleles of a gene, while mutant phenotypes arise due to alterations in the function of one or more proteins. The variant alleles are often described in terms of how they affect protein function: **loss-of-function alleles** reduce the function of a protein (or sometimes make no functional protein at all). **Gain-of-function alleles** do something extra: perhaps a greater quantity of protein is produced, or a protein that is normally only active under certain circumstances is active in additional circumstances.

Loss-of-function mutations are usually (but not always) recessive mutations: having a back-up fully functional allele means that the organism still has some of the protein available. For example, the white eye color of certain mutant fruit flies is a recessive phenotype. It is due to a loss-of-function in a protein that contributes to pigment production.

Gain-of-function alleles often (but not always) result in a dominant phenotype. When an allele does something extra, the back-up copy can't block that extra function.

Figure 1. The white-eye fruit fly phenotype is recessive.

As an analogy, if you are making dinner but one of the burners of your stove is broken (a loss of function), you can still make dinner on another burner. The broken burner does not affect your dinner. But if all of the burners were broken – akin to homozygous loss-of-function mutations – your dinner will not be cooked. A gain of function mutation would be a burner that can't be shut off: Your kitchen might be in danger of catching fire, regardless of whether the other burners are functioning properly.

For the example of eye color mutations in the fruit fly: as long as there is one functional copy of the enzymes needed to produce eye pigments, eye pigments are produced and the fly's eyes are red. A homozygous mutation that knocks out the function of one protein blocks all pigment production, resulting in eyes that are white (**Figure 1**).

An example of a gain-of-function mutation is seen in the antennapedia gene (Antp) in fruit flies. Expression of the Antp gene triggers the development of leg structures. The gene is not normally expressed in the head of the insect – there is no antp protein in those cells of the body. However, Antp mutant flies mis-express the gene, so the protein is produced in embryonic cells that would normally develop into antenna. In these fruit flies, mis-expression of the antp gene causes leg-like structures to grow where wild-type flies have antenna[1]! A fly with this phenotype is shown in Figure 2. The extra production of the Antp protein makes it a gain of function mutation, and the antenna-leg phenotype is dominant. (Note: This is not a normal variation seen in wild fruit fly populations! This is a phenotype studied in the lab.)

1. Struhl, G. A homoeotic mutation transforming leg to antenna in Drosophila. *Nature* **292**, 635–638 (1981).

Although most gain-of-function alleles contribute to dominant phenotypes, and most loss-of-function alleles contribute to recessive phenotypes, there are exceptions. **Haploinsufficient** genes are one example. **Haploinsufficiency** means that one functional copy of the gene is not enough to produce a normal phenotype. Haploinsufficient genes are often genes for which the quantity of the protein product is important. For haploinsufficient genes, loss-of function alleles may be dominant because there is not enough protein produced to get the job done.

Figure 2. Antennapedia mutants have legs where wildtype flies would have antennae.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=318#h5p-52

So far in this text, we've considered dominance and recessiveness as if they are the only options for alleles. A diploid individual can only have two alleles, but in biological populations, when all individuals are accounted for, there may be far more than just two different variations of each gene. Those alleles can interact with each other in far more ways than simple dominance or recessiveness. Some of those types of interactions are described in the next sections and are summarized in Table 1.

Table 1 Allelic interactions

Terminology	Definition	Characteristic offspring ratios	Example
Incomplete dominance	Heterozygotes display a phenotype intermediate to that of the homozygous dominant and recessive	A monohybrid cross gives a 1:2:1 phenotypic offspring ratio, since the heterozygote displays a different phenotype than the homozygous dominant	Red snapdragon (flowers) crossed with white snapdragons give pink offspring, which are always heterozygous. Crossing two pink carnations gives a 1:2:1 ratio of red:pink:white offspring.
Codominance	Both the dominant and recessive phenotypes are displayed in the heterozygote	A monohybrid cross gives a 1:2:1 phenotypic offspring ratio, since the heterozygote displays a different phenotype than the homozygous dominant	Human ABO bloodtypes. Parents of blood type A and blood type B can have children with bloodtype AB.
Incomplete penetrance	The phenotype is not always as expected from the genotype	The ratio of offspring affected by the trait is lower than expected	70% of Individuals with BRCA1 mutations get breast cancer, but not all
Variable expressivity	Individuals with the same genotype may vary in the extent to which a phenotype is expressed.	Ratios are as expected for a dominant or recessive allele, but phenotypes vary.	Polydactyly can vary in the number of extra digits.
Pleiotropy	Pleiotropic alleles contribute to multiple phenotypes	Ratios are as expected, but multiple traits appear linked together	Patients with mutant ATM alleles have ataxia, telangiectasia, immune dysfunction, and high rates of cancer
Recessive lethal allele	Individuals homozygous for a recessive lethal allele are never seen in the population, because those individuals die during embryonic or fetal development.	2:1 ratio of offspring phenotypes	Tailless Manx cats are always heterozygous. Crossing two Manx cats gives 2:1 ratio of tailless to tailed kittens. T

Media Attributions

- Antennapedia © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

INCOMPLETE DOMINANCE

An incompletely dominant phenotype is one for which the heterozygote has a phenotype that is an intermediate between the two homozygous phenotypes. A classic example is pigment color in snapdragons and certain other flowering plants: a cross between a true-breeding plant with red flowers and a true-breeding plant with white flowers gives pink F1 offspring (Figure 3).

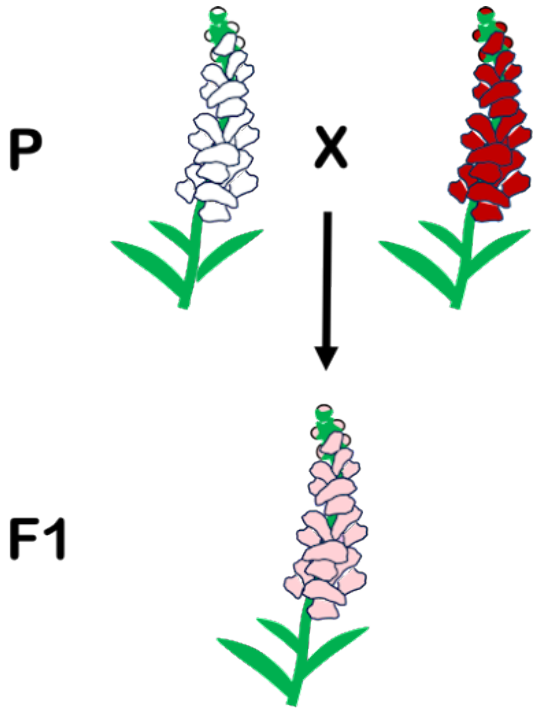


Figure 3 Red and white snapdragons produce pink F1 offspring.

The parental plants are both homozygous, but the F1 is heterozygous. (**Figure 4A**) Because neither allele is completely dominant, slightly different genetic notation is often used. Here, C^W indicates the white allele and C^R the red. The pink snapdragon will never breed true: a cross between two pink snapdragons will give a 1:2:1 ratio of white:pink:red, as shown by the Punnett square in Figure B.

To tie back to earlier vocabulary: this is due to the haploinsufficiency of the red allele. The enzyme product of this gene plays a role in the biosynthesis of pigment, but two copies of the gene together produce more enzyme (and more pigment). The enzyme expressed from one allele is not sufficient to produce the wildtype phenotype. Another way of thinking about this is that the alleles are additive: one allele of C^R produces light red (pink) petals. Two alleles of C^R produce darker red petals.

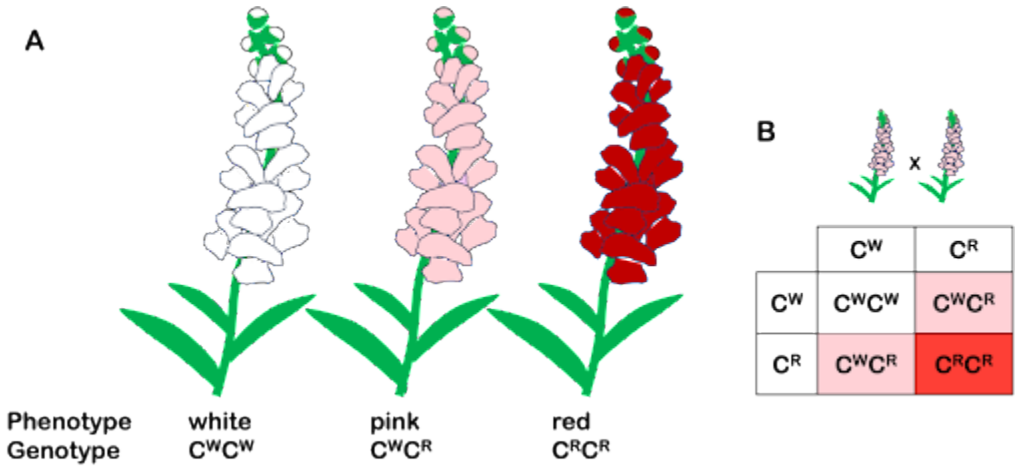


Figure 4. In snapdragon flowers, the red and white phenotypes are incompletely dominant, so the heterozygote appears pink. Panel A: Phenotypes and genotypes of white, pink, and red snapdragons. Panel B: Punnett square demonstrating that a cross between two pink snapdragons gives a 1:2:1 ratio of white:pink:red offspring.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=322#h5p-53

Media Attributions

- Incomplete Dominance © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Phenotype genotype © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

CODOMINANCE

Codominance is another example of a relationship between alleles that do not show complete dominance or recessiveness. It is often confused with incomplete dominance. However, for incompletely dominant alleles the heterozygote shows an intermediate phenotype. For codominant alleles, the heterozygote shows *both* alleles.

A classic example of codominance is the ABO blood type locus in humans. Whether an individual has blood type A, AB, B, or O is controlled by a single gene. The gene encodes a proteins (called an antigen) that is embedded in the surface of red blood cells. (A different gene controls whether you have a "positive" or "negative" blood type, eg. AB positive or O negative.) The surface antigens do not affect human health – they have no known function and cells without any of the ABO surface antigen function just fine. However, the antigens become important when a person receives a blood transfusion: some blood types are incompatible with one another, and receiving blood from someone with an incompatible blood type can trigger the recipient's immune system to attack the foreign cells, resulting in symptoms like fever, pain, red or brown urine, and renal failure.

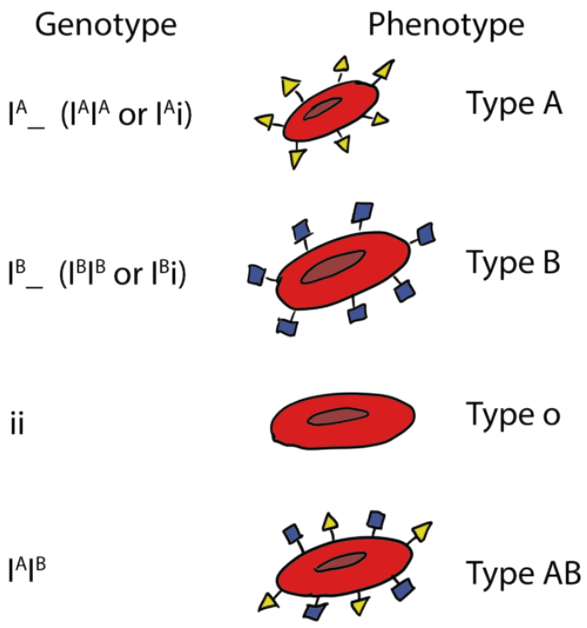


Figure 5 The ABO blood type locus is an example of incomplete dominance.

In the human population, there are three common alleles: I^A, I^B, and i. I^A and I^B are codominant to each other. i is recessive to both IA and I^B.

Individuals with genotypes I^AI^A or I^Ai make type A antigens and have blood type A. Individuals with genotypes I^BI^B or I^Bi make type B antigens and have blood type B. Individuals with genotype ii make no surface antigens and have blood type O. Individuals with genotype I^AI^B make *both* types of antigens and have blood type AB (Figure 5). Because *both* protein types are produced – not an intermediate phenotype – the alleles are codominant.

People with AB blood type are said to be universal acceptors for blood transfusions: they can receive blood of any type, since their immune system recognizes both type A antigens and type B antigens as "self". People with O blood type are universal donors: without any surface antigens, they don't trigger an immune response in anyone else. People with blood type A cannot receive blood from people with blood type B or AB, and people with blood type B cannot receive blood from people with blood type A or AB.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=325#h5p-54

Media Attributions

- Blood Type Codominance © Wikipedia is licensed under a CC0 (Creative Commons Zero) license

PENETRANCE AND EXPRESSIVITY

Sometimes, the phenotype of an organism may not match exactly to what is expected from their genotype. This can happen when a phenotype/trait is **multifactorial**. This means what it sounds like: multiple factors are influencing a trait, not just a single gene. This can mean that several different genes contribute to a phenotype, or it might mean that the environment might affect the phenotype.

However, because geneticists are often only tracking one gene at a time, sometimes it's not (yet) known why the genotype does not always yield the expected phenotype. Geneticists have vocabulary that describes the relationship between alleles and phenotype, even without considering the mechanism for the relationship.

Penetrance and **expressivity** are related concepts that describe the relationship between genotype and phenotype, when additional (known or unknown) factors affect the phenotype.

Penetrance

When an allele controls a phenotype, but that phenotype does not always appear, this is called **incomplete penetrance**. Penetrance is often described as a percentage. An allele that always correlates with phenotype – like blood type in humans or petal color in snapdragons, and most of the alleles we've described previously – is described as 100% penetrant. But many traits show less than 100% penetrance: only a subset of individuals with the genotype display the phenotype.

In many cases, the mechanism for the incomplete penetrance may not be known: geneticists might just know that the phenotype does not always match what's expected by the genotype. However, sometimes the mechanism of incomplete penetrance is known.

For example, certain variations in the human BRCA1 gene predispose patients to breast cancer. People heterozygous for these alleles have a very high occurrence of breast cancer: they have about a 70% chance of developing breast cancer in their lifetime. But it's not 100%. The phenotype is thus 70% penetrant.

In this case, the incomplete penetrance is due to random chance: for a cancer to develop, one (or more) cells within a heterozygous patient's body must sustain an additional DNA mutation in the second, *healthy* BRCA1 allele. This is a relatively rare occurrence, but given the number of cells in the human body and enough time, about 70% of heterozygous people do acquire that second mutation through random chance. Environmental factors can also play a role: Exposure to radiation or DNA damaging agents can increase the likelihood of mutation. Many other cancer-associated mutations are also incompletely penetrant for the same reason.

A second example of incomplete penetrance are the health problems and intellectual disabilities that can

result from the metabolic disease phenylketonuria. Phenylketonuria (PKU) is caused by a defect in a gene responsible for metabolizing the amino acid phenylalanine, which is part of a normal diet. Phenylalanine is even found in foods you might not expect: it is a component of the artificial sweetener aspartame.

Exposure to phenylalanine during infancy causes the intellectual disability characteristic of the disorder. However, most infants in the US are tested for this (and other) metabolic disorders at birth, and individuals with PKU can avoid most symptoms through a diet low in phenylalanine. Diet therefore prevents the symptoms of PKU and makes the phenotype incompletely penetrant[1]. In the United States, if you look at the nutritional labels for foods and beverages containing aspartame, you will see a statement that says, "Phenylketonurics: contains phenylalanine", which is a warning to people who need to adhere to a low-phenylalanine diet.

Expressivity

Figure 6. "Yellow" Labrador retrievers and other dog breeds can vary in color from light cream to red.

Variable expressivity, on the other hand, describes a phenotype that varies in the degree to which it is expressed.

An example of variable expressivity in dogs is the coat color seen in yellow Labrador retrievers (and other dogs, too!). Yellow Labs are all homozygous for the recessive allele of the E locus (the MC1R gene), which prevents the production of the pigment eumelanin in the hairs of the fur. But some "yellow" Labrador retrievers are more cream colored, while others are more golden with reddish undertones (Figure 6). This variation is due to differences in other genes that modify coat color in dogs.

1. Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* **132**, 1077–1130 (2013).

An example of variable expressivity in humans is polydactyly: the presence of extra digits on the feet or hands (Figure 7). Although alleles of several different genes can cause this phenotype, most are dominant, with variable expressivity. The phenotype is variable among patients with a polydactyly-associated allele, ranging from an extra underdeveloped partial finger or toe to multiple additional digits on both hands and feet. The phenotype can even vary for an individual, with left and right hands or feet showing differences.

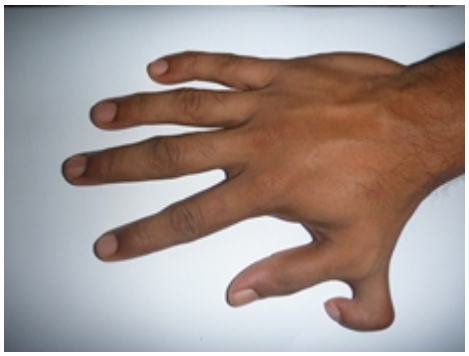


Figure 7. Polydactyly is a condition where extra fingers or toes are present.

Incomplete penetrance and variable expressivity can both be caused by additional genetic factors (as in the yellow Labradors), environmental conditions (as in the diet of patients with PKU), and even randomness (as in the acquisition of a somatic mutation causing breast cancer in patients homozygous for a mutant BRCA1 allele). And more than one of these can contribute! For example, the BRCA1 breast cancer phenotype is affected both by randomness and by environmental factors: exposure to DNA damaging chemicals or radiation increases the likelihood of acquiring a somatic mutation.

A trait can even be both incompletely penetrant and variably expressive. This is illustrated in Figure 8, where some individuals with a particular genotype do not have the trait at all, some have the trait to a small extent, and some have the trait to a large extent.

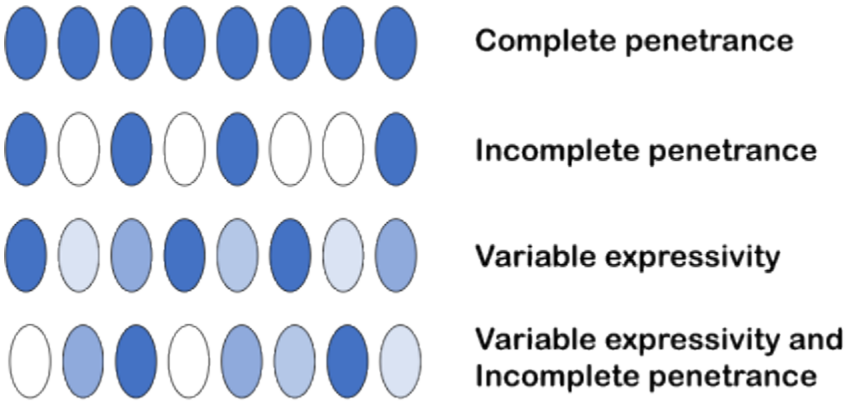


Figure 8. Penetrance and expressivity in individuals with identical genotypes. Penetrance describes whether a phenotype is expressed (dark blue) or not (white). Variable expressivity describes the degree to which a phenotype is expressed (light blue vs dark). Phenotypes can show both variable expressivity and incomplete penetrance.

Media Attributions

- Labrador Retrievers © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Polydactyly sreeakhil © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Penetrance © Amanda Simons is licensed under a CC0 (Creative Commons Zero) license

PLEIOTROPY

Some alleles contribute to multiple phenotypes that may appear to be unrelated.

An example of this are mutations in the gene ATM. In humans and other animals, mutations in ATM cause ataxia (irregular movement), telangiectasias (blood vessel abnormalities visible on the surface of the eye), immune dysfunction, and a predisposition to certain kinds of cancer[1]. These phenotypes all seem pretty different, but they're caused by mutations in the same gene.

Another example: in humans, mutations in the gene LMX1B cause Nail-Patella Syndrome. Individuals with Nail-Patella syndrome have abnormally shaped, underdeveloped, or absent fingernails or toenails. They also have skeletal abnormalities, including misshapen or missing patellae (kneecaps), and some people with Nail-Patella Syndrome develop kidney disease. The ATM and LMX1B disease-associated alleles are said to be **pleiotropic**, since the one allele causes many seemingly unrelated phenotypes.

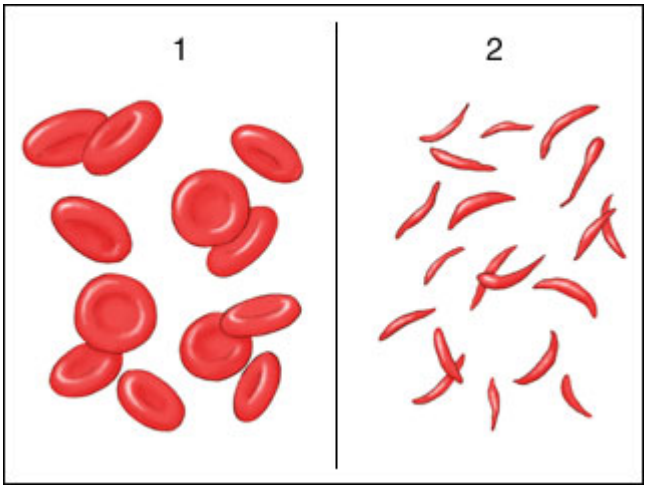


Figure 9. Healthy red blood cells are disk shaped with a indentation in the middle (1). Sickle cell disease causes red blood cells to assume a crescent or sickle shape (2) under conditions of low oxygenation.

A third example of pleiotropy is an allele of the gene that encodes the beta subunit of the protein hemoglobin. The HbS allele of the gene is linked with sickle cell disease (SCD). The healthy allele is called HbA. The HbS allele is most common among people with ancestors from sub-Saharan Africa and certain Mediterranean, Middle-Eastern, and Indian regions. In the United States, about 100,000 people are affected by sickle cell disease (about 3 in every 10,000)[2]. It affects 1 in every 365 Black or African American people.

Individuals who are homozygous for the HbS allele have sickle cell disease, which gets its name from the abnormal shape of patients' red blood cells, as seen under the microscope. Although healthy red blood cells are a flat disk shape with a depression in the center, under some conditions the red blood cells of patients with

1. Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number: #208900: last edit 03/22/2022. World Wide Web URL: https://omim.org/.

2. Data & Statistics on Sickle Cell Disease | CDC. *Centers for Disease Control and Prevention* https://www.cdc.gov/ncbddd/sicklecell/data.html (2022).

sickle cell disease form a crescent or sickle shape (Figure 9). The sickled red blood cells get stuck in capillaries, causing painful symptoms. Red blood cells of SCD patients are also fragile and short-lived compared to healthy cells, making patients anemic. SCD is also called sickle cell anemia.

The sickle shape is caused by a change in a single amino acid of the hemoglobin beta polypeptide: a single base difference in the DNA changes a glutamic acid in the healthy hemoglobin to a valine in the sickle cell hemoglobin. The hydrophobic valine makes the mutant hemoglobin somewhat insoluble in the aqueous environment of the cell, resulting in the clumping together of individual molecules of the mutant hemoglobin. This causes the cells to deform and sickle.

But the HbS allele also confers resistance to the disease malaria: individuals who are heterozygous for the HbS allele are far less likely to develop malaria if they are exposed to the infectious parasite that causes the disease. In this case, the two phenotypes of sickle cell disease and resistance to malaria make the allele pleiotropic.

As with penetrance and expressivity, the underlying mechanism for pleiotropy may not be understood. However, in many cases it is because one single gene has many downstream targets. In the example of ATM, the encoded protein is a kinase that adds a phosphate group to many target proteins. The phosphate group changes the activity of the target proteins, each of which is involved in a different cellular process. The altered activity of the downstream target proteins leads to the different mutant phenotypes.

For the HbS allele: resistance to malaria is likely a side effect of the sickling, since the malarial parasite *P. falciparum* infects red blood cells. Although individuals heterozygous for the HbS allele do not have sickle cell disease, their hemoglobin can clump and their cells can sickle under conditions of low oxygenation. One hypothesis is that the sickling that occurs in low-oxygen conditions interferes with the parasite's life cycle, preventing symptoms of malaria from occurring[3].

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=333#h5p-55

3. Archer, N. M. *et al.* Resistance to Plasmodium falciparum in sickle cell trait erythrocytes is driven by oxygen-dependent growth inhibition. *Proceedings of the National Academy of Sciences* **115**, 7350–7355 (2018).

Media Attributions

- Sickle cell anemia © Wikipedia is licensed under a CC0 (Creative Commons Zero) license

LETHAL ALLELES

Some mutations in essential genes may not be compatible with life. These are called **lethal alleles**. In this section we will focus primarily on lethal alleles that cause a failure of embryonic development, since those can impact expected offspring ratios. Other lethal alleles may cause death in later stages of life and are discussed toward the end of this section.

Diploid organisms have two copies of each gene, so inheritance of one non-functional allele still allows for a backup. However, if both alleles of an essential gene are knocked out, this can block the development of a growing embryo. This happens due to mutations in genes whose functions are essential for the development or survival of the organism. These mutations are called *recessive* lethal alleles because lethality requires two (homozygous) alleles. Heterozygotes may or may not have a phenotype that is distinct from the homozygotes for the healthy allele, but they survive.

Recessive lethal alleles are common in a laboratory setting, where geneticists use model organisms like mice or fruit flies to study gene function. Recessive lethal alleles affect expected offspring ratios from controlled crosses. The offspring ratios of a monohybrid cross (Aa x Aa) that involves a recessive lethal allele differ from the expected 3:1 phenotypic ratio because there are no individuals homozygous for an embryonic lethal allele detected among offspring.

If the heterozygote has the same phenotype as individuals homozygous for the healthy allele, the offspring will all be of one phenotypic class (essentially, 3:0). In this case, offspring ratios could not predict the presence of a lethal allele. However, if the heterozygote shows a mutant phenotype, the offspring of a monohybrid cross may show a phenotypic ratio of 2 (mutant):1 (wild-type). A 2:1 phenotypic offspring ratio therefore often indicates the presence of a lethal allele.

Some lab studies studying model organisms suggest that about 10-30% of genes may be essential[1], but naturally-occurring lethal alleles of these genes are rare outside of lab settings.

One example of a recessive lethal allele that is seen outside of the lab is the allele that causes the tail-less Manx phenotype in cats. Manx cats have no tails. A Manx cat is shown in Figure 10a.

All Manx cats are heterozygous for the Manx allele. There are no cats homozygous for the Manx allele, since homozygous embryos die early in development and are never born. There's no such thing as a true-breeding Manx cat: if two Manx cats are bred together, about 2/3 of the kittens will have no tail and 1/3 of the kittens will have a tail. This illustrated by the Punnett square shown in Figure 10b.

1. Zhan, T. & Boutros, M. Towards a compendium of essential genes – From model organisms to synthetic lethality in cancer cells. *Critical Reviews in Biochemistry and Molecular Biology* **51**, 74–85 (2016).

A

B

	A	a
A	AA (dead)	Aa (Manx)
a	Aa (Manx)	aa (normal tail)

Figure 10. Manx cats are all heterozygous for the Manx allele. A: A tailless Manx cat.

Note: The mutant Manx allele confers a *dominant* tailless phenotype (heterozygotes have a mutant phenotype!) and is indicated by a capital letter in the Punnett square in Figure 10b. But it is called recessive lethal because *lethality* is recessive.

An example of recessive lethality in humans is a particular form of dwarfism called achondroplasia. People with achondroplasia have shortened limbs and several other characteristic traits. Achondroplasia is caused by a heterozygous mutation in the gene FGFR3. Although there may be some associated health complications, people with achondroplasia typically can expect a normal lifespan.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=336#h5p-56

Other examples of lethal alleles

The previous example focused on alleles that cause embryonic lethality, but other lethal alleles may be lethal later in an organism's lifespan. Alleles that impact viability after birth or even later in life – reducing expected lifespan – can be considered lethal alleles, too, although they won't affect offspring ratios in the same way that embryonic lethal alleles do.

There are also rare examples of dominant lethal alleles. Dominant alleles causing *embryonic* lethality do not exist: Although dominant lethal mutations might arise spontaneously as a parent produces egg or sperm, any embryo who carried such a mutation would fail to develop to birth, so such alleles are not maintained in a population. But there are rare examples of dominant lethal alleles that don't impact viability until later in life, after reproductive maturity. In human populations, an example of dominant lethality allele is the allele that

causes Huntington's disease. Huntington's disease is a rare neurological disorder resulting from the death of neurons in the brain with associated cognitive decline, psychiatric symptoms, and uncontrollable movements. The onset of symptoms is in middle age, after many patients have had children and passed the dominant allele to their offspring. There is no cure for Huntington's disease, and patients ultimately die from the progressive degeneration of brain tissue.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=336#h5p-57

Media Attributions

- Manx cats © Michelle Weigold via Wikimedia is licensed under a CC BY-SA (Attribution ShareAlike) license

CLASSIFICATION OF ALLELES DEPENDS ON HOW THE PHENOTYPE IS DEFINED

In all the examples given in this module, the alleles are described based on the phenotype(s) they elicit in the organism.

Although we will commonly describe alleles as dominant or recessive (or incompletely dominant, codominant, etc), what we actually mean when we use these terms is that the resulting *phenotype* is dominant or recessive. We've seen this already in the case of recessive lethality. For example, the agouti phenotype in mice is dominant (heterozygotes for the agouti allele have the agouti phenotype), but the allele confers recessive lethality. The phenotypes associated with a pleiotropic allele may be all be dominant, all recessive, or a mixture.

Because of this, there may be multiple ways to classify an allele. A classic example of this is the HbS allele associated with sickle cell disease (Figure 9). Depending on how the phenotype is identified, the HbS allele can be considered completely recessive, dominant, incompletely dominant, and codominant. Ready for things to get complicated? Let's look:

- **The HbS allele can be considered recessive:** sickle cell disease only affects people with two copies of the HbS allele, which makes SCD a recessive disease.
- **The HbS allele can be considered dominant**: People heterozygous for the HbS allele are said to have Sickle cell *Trait* because their red blood cells can sometimes sickle under extreme conditions that lower blood oxygenation. They do not typically have the health problems associated with SCD. But if the ability of the blood to sickle is what "counts", then HbS is dominant because only one allele must be present.
- **The HbS allele can be considered incompletely dominant**: People with Sickle cell Trait do not have the severe health problems associated with Sickle cell Disease, but under rare conditions they may experience a health crisis because of sickling. This could be considered a much milder form of disease, in which case the allele could be classified as incompletely dominant since the heterozygote has an intermediate phenotype.
- **The HbS allele can be considered codominant**: The HbS and HbA alleles of the hemoglobin beta gene produce slightly different versions of hemoglobin. The different forms of the protein can be detected via lab techniques like chromatography or gel electrophoresis (Figure). People with sickle cell trait have the genotype HbS/HbA. *Both* versions of the protein are produced, so the HbS allele can be considered codominant.

It is common to describe an *allele* as dominant or recessive – geneticists use this shortcut in language all the

time. Instead of saying "the allele that causes the dominant phenotype", we'll just say "the dominant allele". The terms are even used that way in this textbook. But keep in mind that this shortcut of vocabulary is imprecise: it is the phenotype, not the allele, that is dominant or recessive. This is especially important when talking about pleiotropic alleles.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=338#h5p-58

WRAP-UP QUESTIONS

1. In dogs, certain alleles of the MITF gene are associated with extreme white spotting, although this is a bit of a misnomer as many of the dogs don't have white "spots" so much as they are just mostly white. The allele is also linked with deafness: Dogs with white coloration caused by the mutant MITF alleles are frequently, but not always, deaf. Deafness can affect one or both ears[1]. Based on these descriptions, how many ways could you classify this allele of MITF? Explain your reasoning.

2. A yellow-fur phenotype in mice is caused by a mutant allele of the agouti gene. Wild-type mice have a grey-brown fur color. In 1905, Lucien Cuénot observed that a cross between a yellow mouse and a true-breeding wild type mouse always gave a 1:1 ratio of yellow:wild-type offspring. But when he crossed two yellow mice, he observed a 2:1 ratio of yellow:wild-type. What does this tell you about the mutant allele? Explain your reasoning.

3. Familial hypercholesterolemia (FH) is caused by a loss of function allele of the LDLR gene: the protein product encoded by the loss of function allele cannot take uptake cholesterol from the blood, resulting in higher levels of LDL ("bad" cholesterol) in the bloodstream. Although loss of function alleles are usually recessive, FH is a dominant disorder. Explain why this loss of function allele confers a dominant phenotype.

4. A "dominant negative" allele produces a nonfunctional protein, but in a heterozygote, the nonfunctional protein gets in the way of the normal, healthy protein's function and confers a dominant phenotype. An example of this is the p53 tumor suppressor. Some p53 mutations act as dominant negative alleles because p53 normally functions as a tetramer (a protein made up of four subunits). The nonfunctional mutant protein binds to the healthy protein and prevents it from doing its job. Explain how a dominant negative allele is different from a loss of function allele in a haploinsufficient gene.

5. Some cat breeders specifically advertise themselves as Manx cat breeders. Is it possible for all of their kittens to be Manx?

Science and Society

6. Dominant lethal alleles are maintained in a population only if they are lethal after reproductive maturity. For example, someone with Huntington's Disease would not know they had the disease until after they had children, and their children have a 50% chance of inheriting the lethal allele. Huntington's Disease

1. Brancalion, L., Haase, B. & Wade, C. M. Canine coat pigmentation genetics: a review. *Animal Genetics* **53**, 3–34 (2022).

is a progressive neurological disorder associated with disorders in movement, cognitive impairment, and depression and other psychiatric complications. At the end of life, most HD patients are bedridden and unable to speak. Symptoms usually appear in middle age, although this is variable with juvenile and late-onset forms possible as well. Most patients die within 10-25 years of the onset of symptoms.

Modern genetic testing makes it possible to test whether someone has the allele. If one of your parents were diagnosed with HD, would you want to be tested? How would knowing you carried a lethal allele affect the way you approach your life?

PART XII

MULTIGENIC INHERITANCE

Objectives

1. Define the following terms: epistasis, complementation, quantitative trait locus
2. Recognize epistasis and complementation in a genetic cross
3. Predict offspring ratios from a cross involving epistasis
4. Explain how quantitative trait loci contribute additively to a phenotype
5. Use a complementation test to determine how many genes are involved in a phenotype

Introduction

In previous chapters, we've focused primarily on traits that are controlled by single genes. However, many characters are controlled not by one single gene, but by several genes working together to contribute to the phenotype! Several genes can work together additively to contribute to a phenotype, as in the case of traits like height or skin color. The action of individual genes can also cancel each other out, in some cases.

Because several genes can work in concert to control a single phenotype, the inheritance patterns we see when tracking these types of multigenic traits are different from what was discussed in the unit on Basic Heredity. In this module, we will discuss how **epistasis**, **complementation**, and **quantitative trait loci** affect expected offspring ratios.

The complex and variable coat or feather color in many animals is an example of multigenic inheritance. In dogs, for example, different genes specify hair length (long or short), coat texture, fur color (yellow or non-yellow), Pigment intensity (intense or pale/dilute) pigment, white fur or non-white fur, piebald (white) patches. Alleles for each of these genes can be present in many different combinations, giving many color combinations of dogs.

Figure 1. Dogs have different combinations of coat colors and textures, resulting from the contributions of many genes.

A list of some of the genes involved in controlling coat appearance in dogs is shown in Table 1. Historically, dog breeders came up with single-letter notations to describe the genes they tracked, as they intentionally bred dogs with specific features. For example, a single gene denoted as B seemed to distinguish between black and brown pigments, with Black (B) being dominant and brown (b) being recessive. These concepts of dog genetics were pretty abstract, just like in our unit on Basic Heredity: most dog breeders knew whether their dogs were homozygous black (BB), heterozygous black (Bb), or homozygous brown (bb), based on their parentage and/or the puppies they produced. They tracked traits through generations and counted offspring ratios just like you did in the Basic Heredity unit.

Figure 2. Tasha. Tasha's DNA was used for genome sequencing.

Modern tools of molecular genetics have allowed researchers to identify the part of the dog's genome that corresponds to those genes. This is called **mapping** a gene to a chromosome, and it is discussed more in detail in another module. In many cases, researchers found that traits map to genes that they already recognized. The names of those genes and their functions are also shown in Table 1. Why study dogs?

Interestingly, because the dog and human genome are about 95% identical, what we learn from dog genetics is directly applicable to human health. Dogs are a very powerful **model organism** – an organism that we can study to learn about the underlying principles that govern genetics in many species. Dogs are unique in that they have wide phenotypic variation – a huge Great Dane and a tiny Chihuahua are the same species – but centuries of inbreeding have created specific breeds that are genetically very homogenous. Comparing the genomes of dog breeds with different phenotypes can help identify loci linked with those phenotypes, and comparing the human and dog homologs of the genes can help medical researchers understand how changes in those genes affect human health.

It's important to recognize that molecular genetics is still a pretty young science – the Human Genome Project was completed in 2003[1], and a dog's genome was first sequenced in 2005. (Her name was Tasha, and

1. First complete sequence of a human genome. *National Institutes of Health (NIH)*https://www.nih.gov/news-events/nih-research-matters/first-complete-sequence-human-genome (2022).

she was a boxer. Her picture is shown in Figure 2.) Almost all the molecular and biochemical information about the genes listed in Table 1 is less than 20 years old. There's still a lot we don't know. Check out the Dog Genome Project from Elaine Ostrander's laboratory at the National Human Genome Research Institute for information about current research into dog genetics.

Table 1 Dog coat phenotypes and genotypes

Feature	Classical genetics notation	Phenotypes	Gene name
Fur texture	Cu/Cu^c	Curly/smooth fur	KRT71
Fur color	B/b	Black/Brown fur	TYRP1
Fur color	E/e and other alleles	Not yellow/Yellow fur (other alleles change pattern of dark coloration)	MC1R
Fur color	D/d	Full color/"dilute" pigment	MLPH
White patches	S/sp and other alleles	No white/ white spots or patches (an allelic series influences the extent of white)	MITF
Fur length	none	Short hair/long hair	FGF5

Media Attributions

- Dogs-1024×1024-1 © Karen Arnold is licensed under a Public Domain license
- Genome-Boxer © Nature Education is licensed under a All Rights Reserved license

REVIEW FROM BASIC HEREDITY

So how do the genes described in Table 1 (found here) interact to produce varied phenotypes of dogs? Let's start with a review from the Basic Heredity unit.

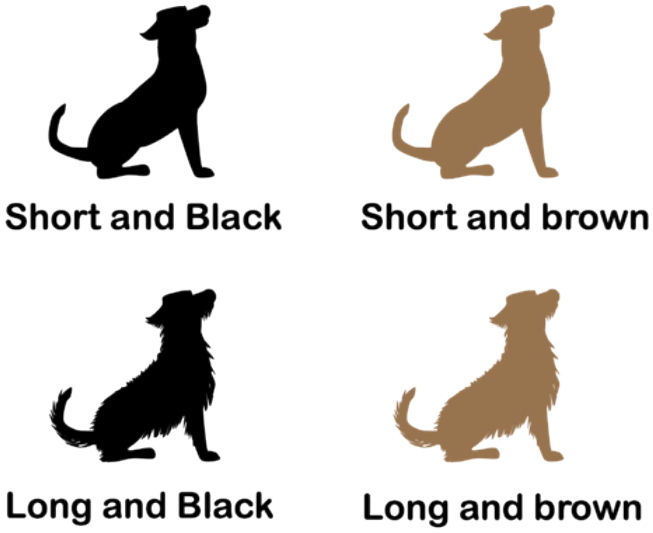


Figure 3. Two separate genes control fur length (black/brown) and fur color (black/brown) in dogs. These act independently, so a dog may have short and black fur, short and brown fur, long and black fur, or long and brown fur.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=350#h5p-59

Remember, the Law of Independent Assortment states that inheritance of alleles from one gene is independent of inheritance of alleles from another. For example, two different genes affect coat length and black/brown coat color: a dog may have long and black, long and brown, short and black, or short and brown fur (Figure 3). A dog that is heterozygous for both genes (dihybrid genotype) will likely have short, black fur.

If multiple genes affect the same phenotype, like the fur color loci shown in Table 1, things get more complicated. Those genes can interact in multiple ways. Some pairs of genes still produce distinct phenotypes in patterns that obey the laws of independent assortment. For example, the B locus determines whether a dog produces a black pigment or a brown pigment; black is dominant to brown. The D locus determines the saturation of color: dogs with a homozygous recessive genotype (dd) have "dilute" or lightened pigmentation compared to the dominant phenotypes. A black (B_) and dilute (dd) dog appears silvery blue-gray in color like the greyhound shown in Figure 4.1, and a brown (bb) and dilute (dd) dog appears silvery brown like the Weimaraner shown in in Figure 4.2. Black and brown dogs with the dominant D allele (D_) are darker.

Figure 4.1 A greyhound with genotype B_dd appears silvery gray, a color that is sometimes called "blue" by breeders.

Figure 4.2 A Weimaraner appears silvery-brown, with genotype bbdd. This color is sometimes called "Isabella" by breeders.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=350#h5p-60

Whether or not a dog is dilute does not affect whether the dog is black or brown. Because the B and D genes act independent of one another and independently assort, a dihybrid cross gives the expected offspring ratio for a dihybrid cross with independent assortment: 9 Black (black and undilute): 3 silver (black and dilute): 3 brown (brown and undilute): 1 silvery brown. This is shown in Figure 5, with a 4×4 Punnett square that shows all 16 possible offspring and their phenotypic class.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it

online here:
https://rotel.pressbooks.pub/genetics/?p=350#h5p-61

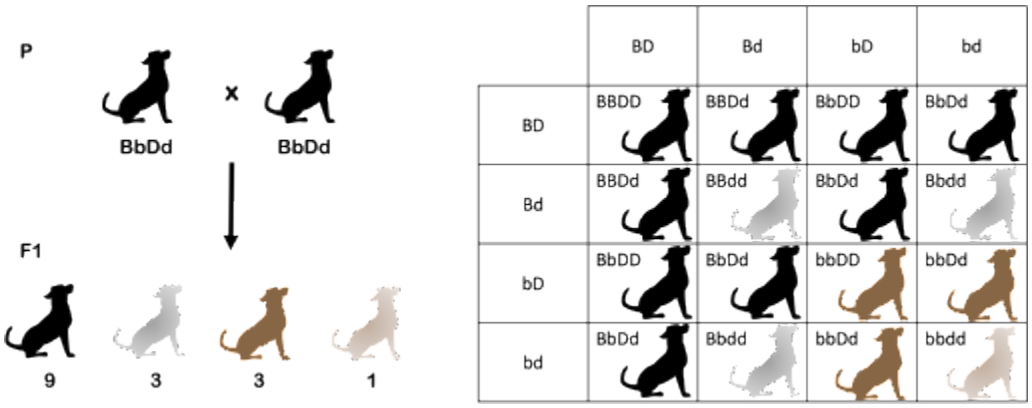


Figure 5. A Dihybrid cross gives a 9:3:3:1 phenotypic offspring ratio. For a dihybrid cross involving the B and D loci, this gives a ratio of 9 black : 3 silver : 3 brown : 1 silvery brown.

Media Attributions

- Black:brown dogs © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Weimaraner © Rufus46 via. Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Greyhound B_dd © Shaz91 via. Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Dihybrid cross © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

EPISTASIS

As discussed in the previous section, the canine B and D genes combine effects in a way that all four classes of offspring – BD, Bd, bD, and bd – have distinct phenotypes. But some gene pairs interact differently, with the alleles of one gene "masking" or overpowering the effects of another. This means that some classes of offspring may share the same phenotype.

This type of relationship is called **epistasis**, with the overpowering gene described as **epistatic**, and the hidden gene called **hypostatic.**

A classic example of epistasis is the relationship between the B and E loci in dogs. This relationship is most striking in the Labrador retriever breed although it is true for all other dog breeds as well. Standard colors for Labrador retrievers are black, yellow, and brown. These colors result from the combined effects of the B and E loci.

Remember, the B locus controls whether a dog produces black or brown pigment. But the E locus controls whether that pigment gets distributed to the fur. A dog with B_ genotype produces black pigment and a dog with bb genotype produces brown pigment. But if the dog *also* has the genotype ee, that dark pigment cannot be deposited in the fur, and the dog ends up with a yellow coat. The ee genotype overrides the B locus alleles, no matter what they are.

This also affects the expected offspring ratios. In a dihybrid cross, we'd expect offspring in a ratio of 9 B_E_ : 3 bbE_ : 3 B_ee : 1 bbee. But because the B_ee and bbee individuals are both yellow, we end up with a ratio of 9 B_E_ (black) : 4 __ee (yellow) : 3 bbE_ (brown). Two of the categories have combined into one! This is illustrated in Figure 6.

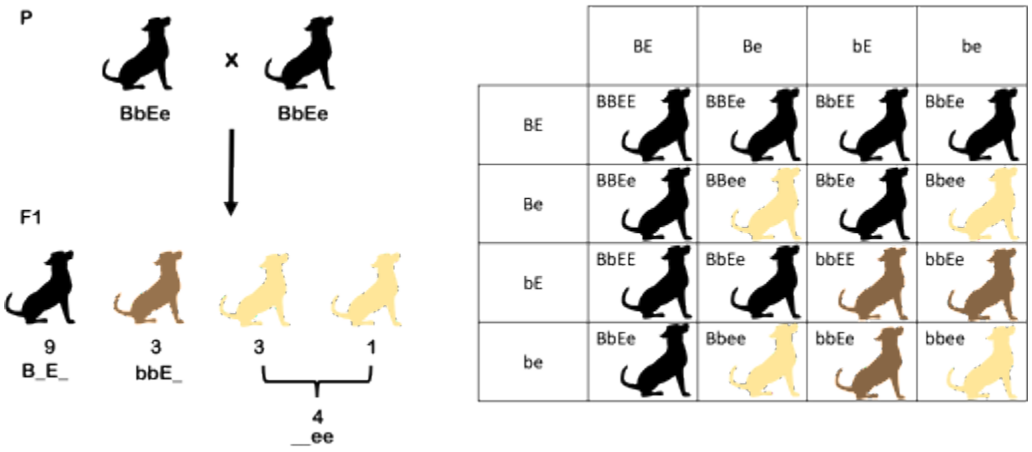


Figure 6. A dihybrid cross with epistasis gives a modified 9:3:3:1 phenotypic offspring ratio, with two or more of the phenotypic classes combining. For a dihybrid cross involving the B and E loci, this gives a ratio of 9 black : 4 yellow : 3 brown. This is a form of epistasis called recessive epistasis. Other types of gene interaction combine the classes of the 9:3:3:1 ratio differently.

For the interaction between the B and E loci, the recessive e allele is the one that covers up the B phenotype. This form of epistasis is therefore called **recessive epistasis**. Other forms of epistasis exist as well. For example, If a dominant allele overpowers the other gene, then it is called dominant epistasis, and a 12:3:1 ratio results. If the dominant alleles of both genes yield the same phenotype, this is called duplicate dominant epistasis, and a 15:1 ratio results.

Table 2 lists some other forms of epistasis, including expected offspring ratios for a dihybrid cross. You'll notice that in all cases, the 9:3:3:1 ratio collapses, so that there are fewer classes of offspring because one or more of the original ratio numbers have been combined together into one class. But the relative frequencies still all add up to 16.

It's important to note that the name of the interaction can vary from author to author: for example, complementary gene interaction is sometimes called "duplicate recessive epistasis". For the purposes of this textbook, we will group all these forms of interaction together under the collective term **epistasis.** It's also important to note that this is not a complete list. There are other forms of gene interaction that could result in any of the dihybrid classes (AB, Ab, aB, and ab) sharing a phenotype.

Table 2. Types of gene interaction and expected offspring ratios from a dihybrid cross, assuming complete dominance. Incomplete dominance, codominance, and incomplete penetrance will affect the expected ratios.

Type of interaction	Ratio	AB									Ab			aB			ab
No Interaction	9:3:3:1	AB	AB	AB	AB	AB	AB	AB	AB	AB	Ab	Ab	Ab	aB	aB	aB	ab
Dominant epistasis	12:3:1	AB	AB	AB	AB	AB	AB	AB	AB	AB	Ab	Ab	Ab	aB	aB	aB	ab
Recessive epistasis	9:3:4	AB	AB	AB	AB	AB	AB	AB	AB	AB	Ab	Ab	Ab	aB	aB	aB	ab
Complementary gene interaction	9:7	AB	AB	AB	AB	AB	AB	AB	AB	AB	Ab	Ab	Ab	aB	aB	aB	ab
Duplicate genes	15:1	AB	AB	AB	AB	AB	AB	AB	AB	AB	Ab	Ab	Ab	aB	aB	aB	ab
Duplicate genes, cumulative effect	9:6:1	AB	AB	AB	AB	AB	AB	AB	AB	AB	Ab	Ab	Ab	aB	aB	aB	ab

The offspring ratios indicate the type of molecular interaction.

The offspring ratios shown in **Table 2** give geneticists a way to study gene interactions using a controlled cross: a researcher can set up a dihybrid cross and draw conclusions from the offspring ratios about how genes affect phenotype. Some great examples of this come from agricultural research. Just like Mendel's original motivations, food production is a powerful motivator!

An example of this is awn length in barley, rice, and other cereals. Awns are the stiff bristle-like projections that give grasses a feathery appearance (**Figure 7**).

Figure 7. Barley

Barley can be awnless (no awns), have short awns, or long awns. Awns serve as sites of photosynthesis for the growing grain, so the presence of an awn can increase grain size and crop yield. But long awns also interfere

with harvest and processing of the grain[1], so the genetics of awn development is pretty important in commercial agriculture.

If you think about the modes of inheritance we discussed in the modules on Basic Heredity and Allele Interactions, a good initial hypothesis might be that the intermediate (short awn) phenotype is caused by incompletely dominant alleles of a single gene. But that turns out to be incorrect: short-armed barley can be true-breeding.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=356#h5p-62

Controlled crosses reveal the underlying mechanism. In some cases, if two different true-breeding strains of barley with short awns are crossed, the resulting F1 offspring have long awns. Self-cross of the long-awned F1 give a ratio of 9 long awns : 6 short awns : 1 awnless[2]. A careful look reveals that this is a modified 9:3:3:1 ratio, which indicates that this is a dihybrid cross with two genes controlling the phenotypes. The 9:6:1 ratio further indicates that the two genes have a cumulative effect: a dominant allele at just one gene makes a short awn, and dominant alleles at two genes make the awn longer. This is illustrated in **Figure 8**.

1. Ntakirutimana, F. & Xie, W. Unveiling the Actual Functions of Awns in Grasses: From Yield Potential to Quality Traits. *Int J Mol Sci* **21**, 7593 (2020).

2. Huang, B., Wu, W. & Hong, Z. Genetic Interactions of Awnness Genes in Barley. *Genes* **12**, 606 (2021).

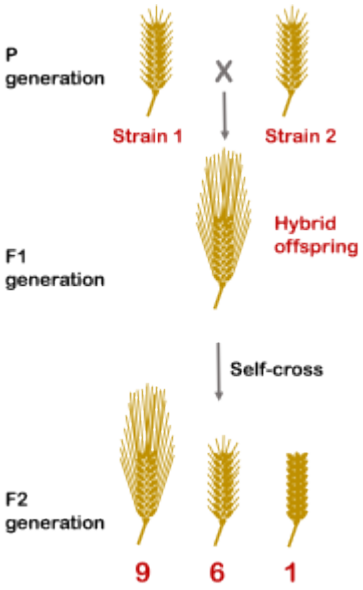


Figure 8. A cross between two strains of barley with short awns. The F1 offspring have long awns. F1 self-cross gives F2 offspring in a ratio of 9 long awns: 6 short awns: 1 awnless.

Link to Animation of **Figure 8**.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=356#h5p-115

The awn genes in barley work together additively to perform duplicate function: The action of one gene (or the other) gives a partial phenotype, but the action of both genes gives a full phenotype. This is sometimes called **duplicate gene interaction with cumulative effect**.

An example of a slightly different kind of interaction is seen in wheat pigmentation. Wheat grains can be either red or white. If two true-breed red strains are crossed, the offspring are also red. But in certain cases, if the red F1 wheat hybrid is self-crossed, a ratio of 15 red : 1 white is seen. A 15:1 ratio is a modification of the 9:3:3:1 ratio expected from a dihybrid cross, so this indicates that the red color is controlled by two genes. The 9, 3, and 3 classes have all been combined into one class. The presence of at least one dominant allele from either gene is sufficient to produce the red color; recessive homozygosity at both genes (genotype aabb) is the only way to produce the white color.

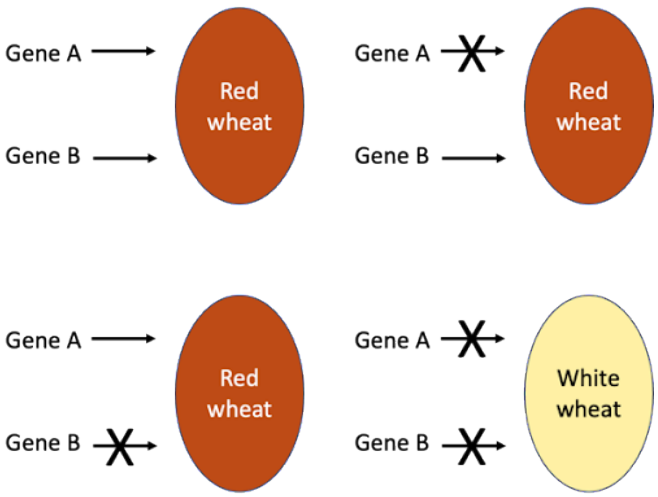


Figure 9. Multiple genes contribute to the red pigment of most wheat strains. A loss of function in one gene alone does not affect color. But a loss of function both genes together results in colorless ("white") wheat.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=356#h5p-63

The 15:1 ratio is often seen if the two genes act redundantly in similar pathways. In **Figure 9**, genes A and B work in duplicate: at least one dominant allele of one gene will elicit the Red phenotype. This is why the phenomenon is sometimes called duplicate genes. A homozygous loss of function of gene A still results in red wheat, since gene B acts as backup. Likewise, loss of function of gene B still results in red wheat. But a loss of function in *both* genes causes the wheat to be colorless (called "white").

The mutant phenotype will only be seen if there is a loss of function of both genes. The pigment-producing alleles are dominant, so to knockout function of both genes requires homozygous mutations: only wheat with a genotype of aabb will be colorless.

Media Attributions

- Awn dihybrid cross with epistasis
- Red wheat strains © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

MOLECULAR GENETICS OF EPISTASIS

In most cases, the genetic interactions listed in Table 2 occur because the interacting genes are participating in a single biochemical pathway. Let's look at a simple example of a pathway where two genes, A and B, act in sequence to produce a purple pigment (Figure 10).

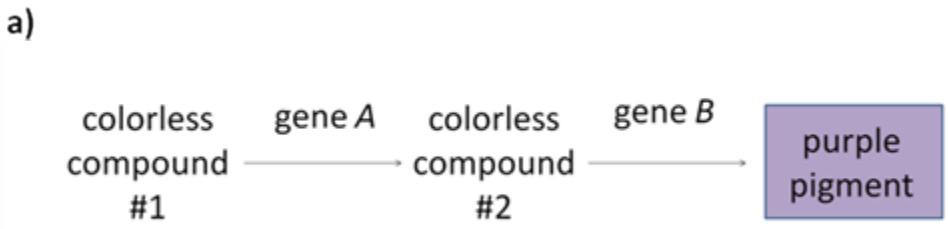


Figure 10. Simple reaction pathway showing production of purple pigment from colorless precursors in two steps

So how does this all work for fur color in dogs? It all comes down to the molecular function of the genes. Let's start with some basic biology.

Pigmentation in mammals (including humans!) is primarily a result of two different forms of melanin: eumelanin is black or brown in color, while pheomelanin is reddish gold. Both eumelanin and pheomelanin are produced in specialized organelles called melanosomes within cells called melanocytes. Melanocytes are specialized cells with long arm-like projections that extend through the epidermis and contact many other cells (**Figure 10**). Melanosomes are transported to other cells in hair follicles and skin, giving hair (or fur, in the case of dogs) and skin its pigment. The ratio of eumelanin to pheomelanin, the amount of melanin in each melanosome, and the number of melanosomes produced by each melanocyte all influence overall pigmentation.

Each of the genes in Table 1 contributes to a different step of the pigmentation process. The B locus corresponds to a gene called TYRP1. This gene encodes an enzyme that catalyzes one of the first steps in synthesizing eumelanin from the amino acid tyrosine. In dogs, the B and b alleles produce slightly different forms of the enzyme. The enzyme variants produce slightly different eumelanin structures, which then appear either black or brown.

The melanosomes are transported through the branches of the melanosomes toward target cells in the epidermis, including hair follicles. A protein called melanophilin, encoded by the D locus, plays a role in this process. Different forms of the protein affect how many melanosomes are transported. Black or brown dogs with a dd genotype appear lighter in color because fewer melanosomes are transported. Which form of

eumelanin (black or brown) is produced does not affect the action of melanophilin, which is why the B and D loci show no gene interaction in a dihybrid cross.

From the extensions of the melanocytes, melanosomes are targeted to other cells in the epidermis. The process is regulated by interactions between the melanocytes and the target cells. The E locus encodes a protein called MC1R. MC1R is a receptor protein found on the surface of melanocytes. In response to a signaling molecule called ASIP from hair follicles, MC1R triggers a switch between eumelanin and pheomelanin production in melanosomes destined for the fur. Variations in MC1R result in predominantly pheomelanin production (as can certain variations in the ASIP gene, also called the A locus)[1]. The reddish-color pheomelanin makes a dog appear yellow, cream, or red.

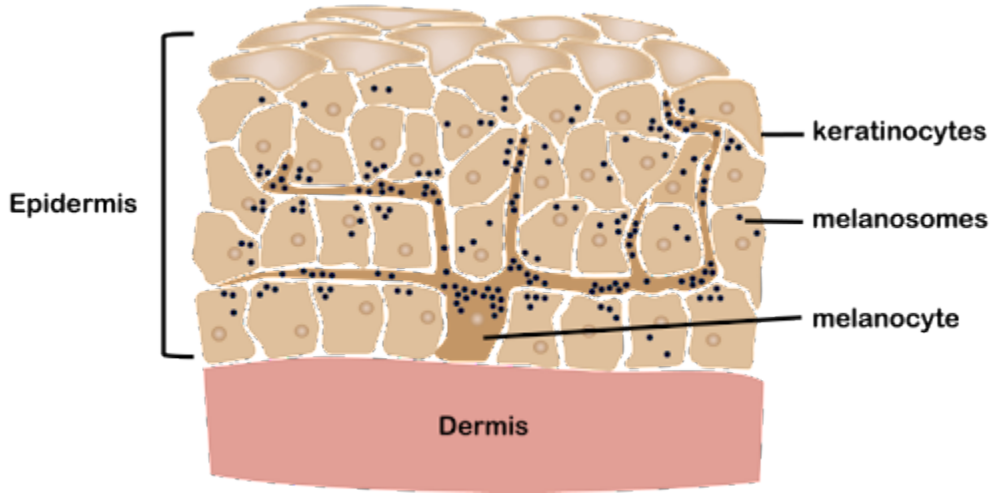


Figure 11. Melanocytes are specialized pigment-producing cells. Melanin is synthesized in organelles called melanosomes. Melanosomes are transported throughout the long, branching arms of the melanocytes and then to target cells. Target cells in the skin and hair follicles (not shown) take up the melanosomes from the melanocytes.

Do you have a yellow, cream, or red dog? Interestingly, you can tell your dog's genotype at the B locus by looking at their nose! Although the e allele prevents eumelanin from being deposited in the fur, the black or brown pigment is still present in the skin of the nose. If your dog has a brown or pink nose, they have a bb genotype. If your dog has a black phenotype, they have at least one B allele, genotype B_[2].

1. Bannasch, D. L. *et al.* Dog colour patterns explained by modular promoters of ancient canid origin. *Nat Ecol Evol* **5**, 1415–1423 (2021).

2. Schmutz, S. M., Berryere, T. G. & Goldfinch, A. D. TYRP1 and MC1R genotypes and their effects on coat color in dogs. *Mamm Genome* **13**, 380–387 (2002).

Figure 12.1 Draco is a yellow dog with black nose. Image courtesy Jennifer Sebalusky Williams.

Figure 12.2 Boo Radley is a yellow dog with a brown nose. Image courtesy Joel Walsh.

As mentioned previously, dogs are a very powerful model organism because they are one of the most visibly variable species, but individual breeds may have very little genetic variation due to inbreeding. This makes it easier to identify genetic variants that correspond to traits, and it gives geneticists a "way in" to understanding human health as well. Many of the genes mentioned in this chapter perform similar functions in other mammals: including humans! Table 3 lists some of the genes controlling appearance in dogs, along with gene function and known phenotypic variations in humans.

Gene name	Phenotypes in dogs	Phenotypes in humans	Gene function
KRT71 (Cu)	Curly/ smooth fur	One variant of KRT71 in humans affects hair texture. Variations in other types of keratin also affect hair texture in humans.	Encodes a protein from the keratin family that is a structural component of hair follicles.
TYRP1 (B)	Black/ brown	One variant of TYRP1 in humans is associated with blond hair in people of Melanesian ancestry. Other variants are associated with albinism[3].	The encoded protein interacts with the enzyme that performs the first step in the synthesis of eumelanin. The b allele results in eumelanin that is brown in color, while the B allele results in eumelanin that is black.
MC1R (E)	Not yellow/ yellow	Variations of MC1R in humans are associated with differences in skin and hair color[4].	Encodes a protein on the surface of pigment-producing melanocytes. The e allele prevents the deposition of the black/brown eumelanin in the fur, leaving fur reddish/yellow due to the presence of pheomelanin.
MLPH (D)	Full color/ faded or "dilute" pigment	A variation in MLPH is associated with Griscelli syndrome. People with Griscelli syndrome have hypopigmented skin and silvery hair[5].	Encodes a protein called melanophilin, which plays a role in the transport of pigment-containing organelles within melanocytes. The recessive allele reduces the amount of eumelanin (black/brown pigment) that is deposited in fur. These alleles seem to affect eumelanin more than pheomelanin, so they have less of an effect on the color of a yellow/red dog with an **ee** genotype.
MITF (S)	No white/ white spots or patches	Certain variants are associated with Waardenburg syndrome, which is characterized by areas of hypopigmentation and hearing loss.	Encodes a transcription factor that helps control the production of melanocytes (pigment-producing cells)
FGF5	Short/long hair	A variation in FGF5 is associated with excessively long eyelashes[6].	Encodes a growth factor that regulates the hair growth cycle in many mammals, including dogs and humans

3. TYRP1 gene: MedlinePlus Genetics. https://medlineplus.gov/genetics/gene/tyrp1/.

4. MC1R gene: MedlinePlus Genetics. https://medlineplus.gov/genetics/gene/mc1r/.

5. MLPH gene: MedlinePlus Genetics. https://medlineplus.gov/genetics/gene/mlph/.

6. Higgins, C. A. *et al.* FGF5 is a crucial regulator of hair length in humans. *Proceedings of the National Academy of Sciences* **111**, 10648–10653 (2014).

More than two genes

In most cases, the genetic interactions listed in Table 2 occur because the interacting genes are participating in a single biochemical pathway. But biochemical pathways often include far more than just two genes! The next sections look at how different forms of gene interaction influence phenotype when more than two genes are involved.

Media Attributions

- Purple pigment pathway © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Melanocytes © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Dog with brown nose © Joel Walsh is licensed under a CC BY-SA (Attribution ShareAlike) license
- Yellow dog black nose © Jennifer Sebalusky Williams is licensed under a CC BY-SA (Attribution ShareAlike) license

QUANTITATIVE TRAIT LOCI

Awn length in barley, discussed above, is an example of a trait where two genes are acting additively (or cumulatively) to contribute to a phenotype. The two awn genes described in that section have completely dominant alleles, so the two genes can contribute to three possible phenotypes. Two genes with a dominant allele (long awn), one gene with a dominant allele (short awn), no genes with a dominant allele (awnless):

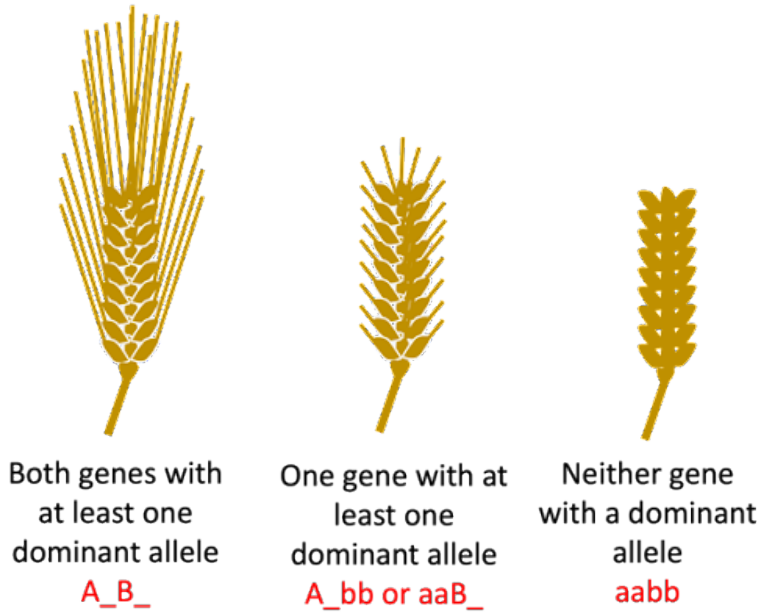


Figure 13 Awn length is determined by the cumulative effect of completely dominant alleles at two genes. A dominant allele at both loci will give the greatest extent of phenotype. A dominant allele at one locus will give an intermediate phenotypes, and dominant alleles at neither locus will give the least extent of the phenotype.

Awn length is an example of a **quantitative trait**: a trait that is a measurable phenotype, controlled by multiple genes acting cumulatively. We just looked at two genes, and a correspondingly limited number of phenotypes. But awn length is controlled by many more genes than just two, and there is more variation in awns than can be divided into simple phenotypic classes of long, short, and no awn.

One of the hallmarks of quantitative traits is that, rather than existing in easily distinguishable categories, the traits may vary continuously throughout a population. For comparison, you'll recall that Mendel observed round seeds and wrinkled seeds, yellow and green seeds. Those are examples of **discrete variation.**

But quantitative traits vary across a spectrum. There is no distinct separation of phenotypic classes. This is

called **continuous variation.** One example in humans is height, which can vary from very short to very tall. Other examples of continuously variable traits in humans are skin color and weight.

Quantitative traits are controlled by multiple genes acting cumulatively. These genes are called **quantitative trait loci**, commonly abbreviated **QTLs.** The more QTLs involved, the more phenotypic classes are possible. All loci contribute additively. To make it more complicated, in many cases the alleles of these genes are incompletely dominant, so the *alleles* contribute additively. This gives the potential for lots of phenotypic variation with a relatively few number of genes. Let's take a look.

You'll recall that a monohybrid cross gives a genotypic ratio 1:2:1 (AA Aa, and aa). For incompletely dominant alleles where the heterozygote has a distinct phenotype, this gives three distinct phenotypes.

Two cumulative effect QTLs with incomplete dominance can produce five discrete phenotypes, depending on how many of the alleles associated with are present. Note that the allele symbols used here are an oversimplification: for incompletely dominant alleles where neither is dominant over the other, we would not typically use capital and lowercase letters.

- Four phenotype-associated alleles – AABB (greatest extent of phenotype)
- Three phenotype-associated alleles – AaBB or AABb
- Two phenotype-associated alleles – AAbb, AaBb, or aaBB
- One phenotype-associated allele – Aabb or aaBb
- No phenotype-associated alleles – aabb (least extent of phenotype)

Three cumulative effect QTLs with incomplete dominance can produce seven discrete phenotypes:

- Six phenotype-associated alleles – AABBCC (greatest extent of phenotype)
- Five phenotype-associated alleles – AaBBCC, AABbCC, or AABBCc
- Four phenotype-associated alleles – aaBBCC, AaBbCC, AaBBCc, AAbbCC, AABbCc, or AABBcc
- Three phenotype-associated alleles – aaBbCC, aaBBCc, AaBbCc, AabbCC, AaBBcc, AAbbCc, or AABbcc
- Two phenotype-associated alleles – AAbbcc, AaBbcc, AabbCc, aaBBcc, aaBbCc, or aabbCC
- One phenotype-associated allele – Aabbcc, aaBbcc, or aabbCc
- No phenotype-associated alleles – aabbcc (least extent of phenotype)

Four cumulative effect QTLs would produce 9 phenotypes. The relationship between the number of QTLs can be summarized with the equation below, where n = the number of loci.

$$\#\ \text{phenotypes} = 2n + 1$$

But as the number of phenotypic classes increases, the differences between the classes get smaller. With a high enough number of genes, the discrete classes begin to blend together. This is especially true when

environmental influences are factored in. For example, even identical twins who share 100% of their DNA can have slightly different heights or skin color.

You'll note that in all cases there are multiple genotypes that can confer an intermediate phenotype. But the most extreme phenotypes are only seen in individuals who are homozygous recessive for all alleles, or homozygous dominant. This is why, for continuously variable traits, an intermediate phenotype is the most common. Very few individuals in a population will have the most extreme phenotypes. This is illustrated in Figure 14, which shows both the number of discrete phenotypes expected for 1, 2, 3, or many QTLs and the relative frequency of each phenotype in a population[1].

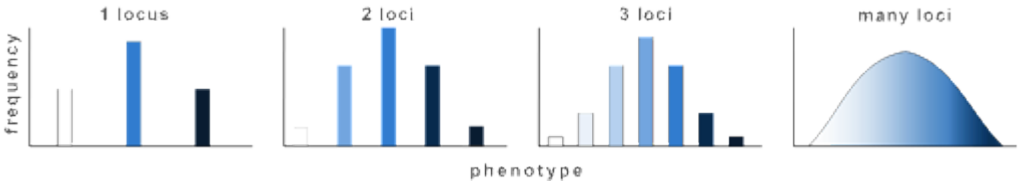


Figure 14. SEQ Figure * ARABIC 14 The more loci that affect a trait, the larger the number of phenotypic classes that can be expected. For some traits, the number of contributing loci is so large that the phenotypic classes blend together in apparently continuous variation.

Note: this assumes that each gene contributes equally to the phenotype. In practice, however, **quantitative trait loci** can also vary in the extent to which they influence a trait. For example, some genes that are involved in determining height have only a modest effect, while others play a much greater role in determining height.

How many genes can affect quantitative traits? As an example in humans, there are likely hundreds of QTLs contributing to the diversity of height and hundreds of QTLs contributing to the diversity of skin color in humans.

The cumulative effects of QTLs likely play a role in fruit size in polyploid commercial crops. Many of the cultivated fruits we consume, including strawberries and bananas, are much larger than their wild counterparts . The wild counterparts are often diploid (2n = 2x). Many cultivated strawberries are octoploid (2n = 8x) and the most common cultivated bananas are triploid.

1. Nickle and Barrette-Ng. Open Online Genetics. in *Open Online Genetics* (2016).

Figure 15. Wild vs domesticated strawberries. The sweet, flavorful, wild diploid is on the left, while the huge, cultivated octoploid is on the right.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=367#h5p-64

Media Attributions

- Awn length © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Phenotype © Deyholos via Open Genetics Lectures is licensed under a CC BY-NC (Attribution NonCommercial) license
- Wild vs domesticated strawberries © Wikipedia- Left: Ivar Leidus/Right: David Monniaux is licensed under a CC BY-SA (Attribution ShareAlike) license

COMPLEMENTATION

As discussed in the Basic Heredity module, recessive phenotypes are always "true-breeding": a cross between two individuals, both with a recessive phenotype, will always have recessive offspring. But in some cases, complementary genes give rise to a situation that appears to violate this principle. This happens when mutations in two different genes can produce similar phenotypes. Such genes are sometimes called complementary genes.

Because complementary genes give the same phenotype when mutated, complementary genes give a 9:7 phenotypic ratio in a dihybrid cross (Table 2). Complementary genes often participate in the same biochemical pathway: they give a 9:7 ratio in a dihybrid cross because a loss of function of either of the genes would block the function of the whole pathway. Several examples of how complementary genes might work together to perform a biochemical function – the production of purple pigment – are shown in Figure 15.

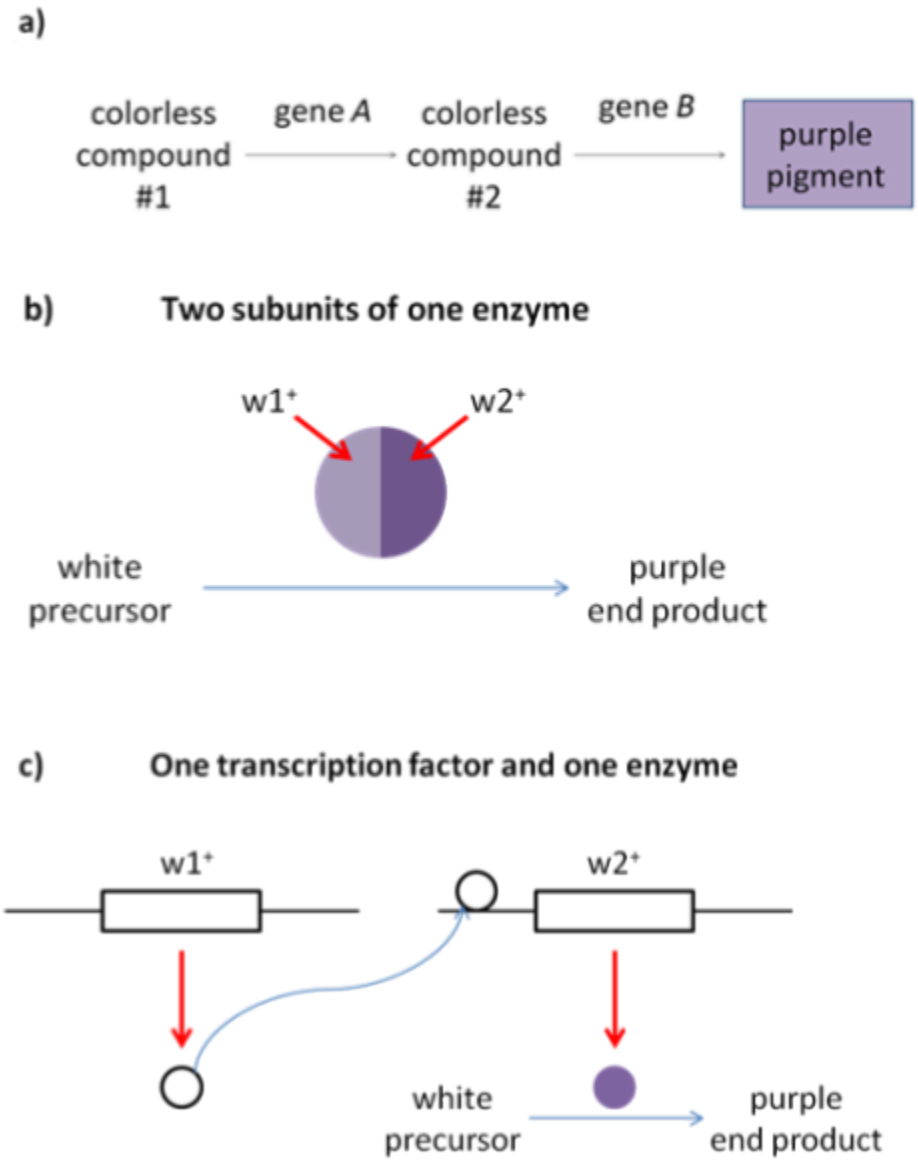


Figure 16. Molecular mechanisms behind complementary genes. a) A simplified biochemical pathway showing complementary gene action of A and B. Note that in this case, the same phenotypic ratios would be obtained if gene B acted before gene A in the pathway b) biochemical pathway showing two subunits of one enzyme c) biochemical pathway showing one transcription factor and one enzyme. (long description)
A simple graphic showing one transcription factor and one enzyme resulting in purple color

The word complementary comes from the word "complement", which shares a root word with "complete". (Not "compliment", which means to say nice things.) You can think of complementary genes as forming

a complete pathway. Without the complete whole, the pathway won't function, so a homozygous loss of function of any gene will give a mutant phenotype.

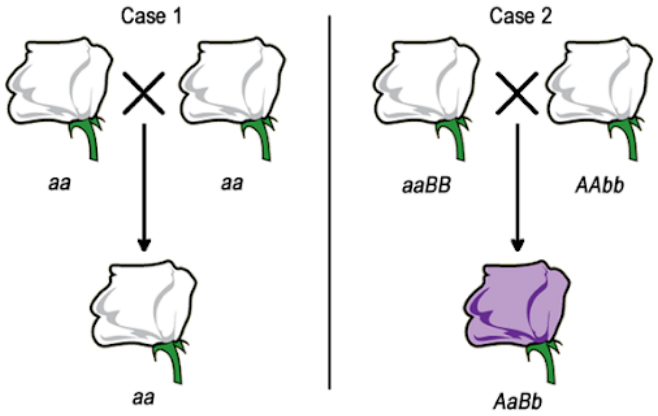


Figure 17. A cross between two white flowers is expected to produce white offspring (left, Case 1), since white is a recessive phenotype. However, sometimes a cross between two white flowers may produce wild type purple flowered offspring. This is due to complementation: the parents harbored mutations in two different genes, but the offspring inherited one wild type allele of each gene.

Consider a hypothetical situation where purple flowers are the wild type, and white flowers are a recessive mutation. Because white flowers are recessive, we'd expect a cross between any two white flowers to produce white offspring, as shown in Case 1 in Figure 16. But if the white color is caused by recessive mutations in different genes in each parent, the dihybrid offspring will have a wild type phenotype.

An example of this in humans: a number of genes contribute to a recessive phenotype of deafness. The phenotype requires homozygosity for the deafness-associated alleles, so we might therefore expect that two Deaf parents would always have a Deaf child (aa x aa). However, in some cases two Deaf parents may have a hearing child. This happens when the parents are homozygous for deafness-associated alleles of different genes (AAbb x aaBB).

How many genes? Complementation tests

Biochemical pathways can get pretty complicated. In the previous examples, we gave examples where two genes are controlling the same phenotype, but most biochemical pathways involve far more than just two genes.

A **complementation test** can be used to figure out if two related traits (like the hypothetical situation of purple vs white flowers, for example) are a result of different alleles or different genes. Complementation tests can also be used to figure out how many genes are contributing to a particular trait.

In its simplest form, a complementation test is a cross between two individuals with the same recessive phenotype. The F1 offspring indicate whether the parents have mutations in the same gene or mutations in two different genes. As shown in Figure 16, there are two possibilities for the results of a complementation test. If two white parental flowers produce white offspring, we say that the parental mutations **fail to complement** one another. The F1 offspring does not have a complete set of the wild-type alleles needed to produce pigment. But if two white flowers are crossed to produce wild type (purple) offspring, the parental mutations **complement** one another. The F1 offspring must have a complete set of the alleles needed to produce the wild type phenotype.

In the previous example, we considered a situation where two genes controlled the production of purple pigment. But what if there were more? Multiple pair-wise crossings can give geneticists a clue as to how many genes might be controlling a phenotype.

Consider a situation where there are 6 true-breeding lines of white flowering plants. How can we use a complementation test to determine how many genes are affecting flower color? We can perform multiple crosses, crossing every line with every other line, one pairing at a time. The results of such a test are shown in **Table 4**. The boxes of the table list the phenotype of the F1 offspring for each of the indicated crosses. For example, when strain 1 is crossed with strain 2, the offspring are wild type (purple), but when strain 1 is self-crossed the offspring are white. The empty/dark gray boxes in the table are crosses that are represented elsewhere in the table: the cross between strains 1 and 2 is in the first column second row, but the first row second column is empty.

Table 4. Complementation test to determine how many ge

Strain	**1**	**2**	**3**
1	Mutant		
2	Wild-type	Mutant	
3	Wild-type	Wild-type	Mutant
4	Wild-type	Wild-type	Mutant
5	Mutant	Wild-type	Wild-type
6	Mutant	Wild-type	Wild-type

For the example shown in **Table 4**, you'll see that all the self-crosses listed (1×1, 2×2, 3×3, etc) give a mutant phenotype. This is what you would expect from a true-breeding strain! Complementation tests will usually include the self-cross offspring as a control.

If you look at all of the crosses involving strain 1, you'll see that crosses with strains 2, 3, and 4 produce wild-type offspring. This means that strain 1 **complements** strains 2, 3, and 4. Those strains have a different mutation than strain 1.

But strains 5 and 6 **do not complement** strain 1: offspring of those crosses have a mutant phenotype. From this, we say that strains 1, 5, and 6 belong to the same **complementation group**. Members of a complementation group are presumed to have mutations in the same gene.

Can you find a set of two more genes that do not complement? Answer: Strains 3 and 4 also belong to the same complementation group.

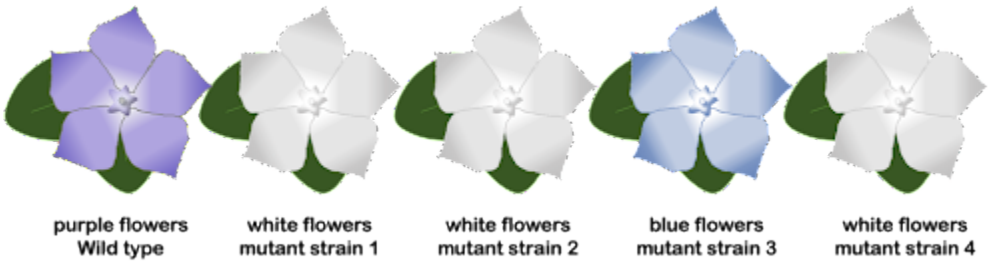
So which strain is left? Strain 2 complements all other strains, so we say that strain 2 is in a "group" by itself. That gives us three groups:

- Group A: Strains 1, 5, 6
- Group B: Strains 3, 4
- Group C: Strain 2

We started with 6 strains, and all strains must be assigned to one group. Because there are three complementation groups, we can conclude that there are likely to be *at least* three genes controlling flower pigmentation. Why the *at least*? If we expanded our sample size to include additional strains, we might find additional strains that belong to each group, or we might discover additional complementation groups.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=372#h5p-65



What about a situation where there are more mutant phenotypes? Well, a complementation test works here, too. Flowers come in more than just two colors, and remember, there may be more than two alleles of a gene in a population. Let's look at a hypothetical situation where we have flowers with multiple mutant phenotypes.

In a complementation test for the four mutant strains, we see the results shown in **Table 5**.

Table 5. Complementation test for the four mutant strains.

Strain	**1**	**2**	**3**	**4**
1	White			
2	Wild type	White		
3	Wild type	Blue	Blue	
4	White	Wild type	Wild type	White

As with the simpler two-phenotype example, here we look for complementation (wild type offspring) and failure to complement (mutant offspring). The cross between strains 2 and 3 fails to complement since the offspring have a mutant phenotype (blue). From this, we can conclude that strains 2 and 3 have mutations in the same gene. Further, because strains 2 and 3 have different phenotypes, we conclude that strains 2 and 3 have two different alleles of the mutant gene. We also can conclude that blue is dominant to white in the allelic series.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=372#h5p-66

This example also reveals something important about complementation groups: although every member of the group harbors a mutation in the same gene, they do not all share the same mutant allele. As an example, the *white* gene in Drosophila has over 300 known mutant alleles. All of these alleles would be part of the same complementation group.

Limits of the complementation test

The complementation test is limited in that it will only work when the mutant phenotype is recessive and if each tested individual has a mutation in one gene. In some cases, similar phenotypes may be caused by a dominant mutation in one gene or a recessive mutation in another. In such cases, the F1 generation of a cross with the dominant mutation will always include mutant individuals, regardless of the genotype of the other parent. This may be easily confused with a failure to complement, so care must be taken in interpreting results of a complementation test. Results also can be complicated if a mutant strain has a mutation in more than one gene.

Variations of the complementation test in human biology

A complementation group refers to a group of alleles of a single gene that contributes to a complex, multigenic

phenotype. The members of a complementation group will always *fail to complement* each other, since they are all missing the ability to perform the same biological function. Many complex human phenotypes are grouped in complementation groups, even though the traditional complementation test described above cannot be used in humans.

But many complex human traits also confer a cellular phenotype that can be measured in the lab. Cellular and molecular genetics techniques can be used to determine if the phenotype of a patient's *cells* can be complemented. An example of this is Fanconi anemia. Patients with Fanconi anemia are usually diagnosed in childhood with bone marrow failure. They are highly susceptible to certain forms of cancer, and they may have characteristic skeletal abnormalities, hyper-pigmented spots on the skin, and/or small stature. There is no cure for the disease, although many patients are ultimately treated with bone marrow transplant.

As of the writing of this text, there are 22 known Fanconi anemia complementation groups[1]. The groups are named for the disease followed by a letter: FancA, FancB, FancC... all the way to FancY. Each group corresponds to a particular gene. The Fanconi genes encode proteins that participate in the repair of a particular form of DNA damage: inter-strand crosslinks. An inter-strand crosslink is a covalent bond that links the two complementary strands of the double helix in a way that prevents them from separating during replication or transcription.

Healthy cells constantly sustain DNA damage, but cells have repair mechanisms built in place to fix the damage before DNA is replicated and/or the cell divides. Different types of damage are repaired with a different set of repair enzymes, so Fanconi cells can repair some damage but not inter-strand crosslinks.

Cells collected from Fanconi anemia patients are highly susceptible to DNA crosslinking agents since they cannot repair the damage. In fact, this is how patients are often diagnosed: blood or skin cells isolated from a patient are treated with a DNA crosslinking drug. If a patient's cells are killed by the DNA crosslinks but not other forms of DNA damage, it is an indication of Fanconi anemia. This is also the basis for a version of a complementation test that, until recently, was the most common way to determine which FA-associated gene was affected in each patient.

Cells from a newly diagnosed Fanconi anemia patient would be fused with a panel of cells from known FA complementation groups. The fused cell (now with four copies of each chromosome) would be treated with DNA crosslinking agents. If the fused cells die, the patient's cells *do not complement* the test cells, indicating the patient's cells have a loss of function pathogenic variant in the *same* gene as the test cell line. If the fused cells survive, this indicates the patient's cells *complement* the test cells, and the patient has a pathogenic allele in *different* gene from the test cell line. This is important in treating the symptoms of Fanconi anemia, since patients belonging to different complementation groups may have slightly different disease phenotypes and react differently to treatment.

If a new complementation group was identified, a related experiment could then be used to identify the

1. Mehta, P. A. & Ebens, C. Fanconi Anemia. in *GeneReviews*® (eds. Adam, M. P. et al.) (University of Washington, Seattle, Seattle (WA), 1993).

genomic locus responsible for the FA phenotype. Genomic DNA from a healthy (non-FA) individual is broken into thousands of fragments and inserted into small, circular DNA molecules called plasmids. The plasmids carrying the genomic segments are introduced into FA cells, one segment per cell. The cells are then subjected to DNA crosslinking agents. Cells that survive have been complemented by the genomic segment they host. The genomic DNA is recovered from the cells for further study, thus identifying the gene linked with FA.

This process revealed that many of the genes represented by the complementation groups were already known by other names, studied by molecular geneticists in other contexts. For example, FANCD2 is the BRCA2 gene. Some mutations in the BRCA2 gene are associated with a predisposition to breast cancer, while other mutations in the BRCA2 gene are associated with Fanconi anemia. (Most people have a healthy version of the gene and have neither Fanconi anemia nor a genetic predisposition to breast cancer.)

Both of these techniques have largely been superseded by newer diagnostic tests, including whole-genome sequencing. But the vocabulary of complementation remains: like many multigenic disorders, Fanconi anemia is still subdivided into complementation groups, and as new Fanconi-associated mutations are identified through whole-genome sequencing, new complementation groups are added to the list.

Media Attributions

WRAP-UP QUESTIONS

1. In addition to the coat coloration genes discussed in this chapter, other genes affect patterns like saddle markings and other black/tan patterns. The "Dominant Black" allele, K, overpowers the effects of several patterning genes, making a dog mostly black in color regardless of the alleles at those other genes. Thinking about the language of genetics, from the Allele Interactions and Multigenic Inheritance modules, explain why the name "Dominant Black" is a misleading way of describing this phenotype.

2. The Cu locus determines whether a dog has a curly or straight coat (Curly is incompletely dominant to straight). The L locus determines whether a dog has a long or short coat (Short is dominant to long). However, a dog with a short coat is likely to have straight fur, regardless of its genotype at the Cu locus.

a. How would you describe the relationship of the Cu and L loci?
b. What is the phenotype of a dihybrid dog?
c. Predict the phenotypic ratio of the F1 offspring of two dihybrid dogs.

3. Chickens have fleshy protuberances called combs on the top of their heads. Combs can come in shapes called "single", "rose", "pea" and "walnut", shown in **Figure 18**. Two chickens with walnut combs are crossed, with offspring counted over multiple years. The F1 offspring are listed below.

- 27 Walnut chicks
- 10 Pea chicks
- 9 Rose chicks
- 3 Single chicks

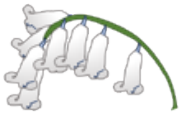
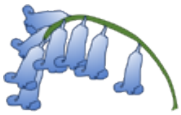
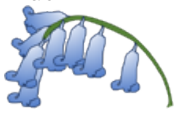
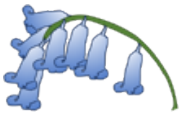
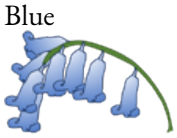
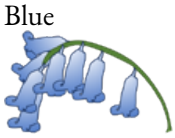
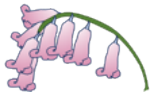
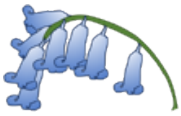
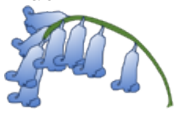
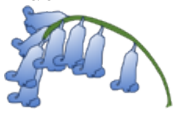
Figure 18. Four roosters with different shaped combs.

Give the genotypes of the parents and the F1 offspring. Use an underscore to indicate where multiple alleles might be possible.

4. You'll recall that early scientists studying heredity hypothesized that the "blending hypothesis" would explain how traits of parents might be transmitted to offspring. Explain how QTLs work to support and contradict the blending hypothesis. For example: why might a very tall parent and a very short parent be expected to have a child of medium height? But how might two medium-height parents have a very tall kid?

5. The table below shows the results of a complementation test in English bluebells. Wildtype bluebells are blue, but some varieties have white or pink flowers. How many complementation groups are listed in this table? Based on the result of this test, how many genes control flower color in bluebells?

Complementation test in English bluebells

Strain	**1** **White**	**2** **White**	**3** **White**	**4** **Pink**	**5** **Pink**
1: White	White 				
2. White	White 	White 			
3. White	Blue 	Blue 	White 		
4. Pink	Blue 	Blue 	Blue 	Pink 	
5. Pink	Blue 	Blue 	Blue 	Pink 	Pink 

6. In this module, we discussed how recessive homozygosity at the "e" locus results in a dog with yellow, cream, or reddish fur. These dogs are all true-breeding: a cross between two yellow labrador retrievers always gives yellow puppies. However, when certain reddish-yellow dogs are interbred with other reddish-yellow dogs known to have the "ee" genotype, the litters have dark-colored puppies. Propose an explanation for this result.

Media Attributions

- Four roosters © Shannan Muskopf is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- White bluebell flower
- Blue bluebell flower
- Pink bluebell flower

PART XIII

GENETICS OF SEX

Objectives

1. Define sex chromosome, autosome, heterogametic, homogametic, and haploinsufficient.
2. Compare and contrast mechanisms of sex determination in different species.
3. Recognize that both sex-linked and autosomal genes play a role in sex development in humans, and explain how disruptions in the sex development pathway can lead to differences of sex development.
4. Distinguish between sex-determination, sex-linked, and sex-influenced genes
5. Recognize that sex-linked traits show different inheritance patterns from autosomal traits
6. Use a modified Punnett square to predict the inheritance of sex-linked traits.
7. Explain what is meant by dosage compensation, and explain how X-inactivation can lead to codominance, incomplete dominance, or variable expressivity of X-linked traits.

Introduction to sex chromosomes: it's not just X and Y

In the chapter on Genome Structure, we first introduced the difference between sex chromosomes and autosomes. Historically they were identified because they were easily visible under the microscope, and, depending on the species of organism being studied, these distinctive chromosomes tracked with sex through generations. Geneticists later found that these chromosomes carry genes associated with the development of sex structures, called **sex-determination genes**.

In humans and other mammals, males typically have one X and one Y chromosome and are said to be **heterogametic.** Females typically have two X chromosomes and are **homogametic**.

Note 1: In this text, you'll see the XX genotype described as "typically" or "most commonly" associated with a female phenotype, and the XY genotype with males. The reason for the "typically" or "most commonly" description is that there are many exceptions to this both in humans and other organisms. For example,

there are individuals with **sex chromosome aneuploidies** (SCAs), with only one X chromosome (described sometimes as XO) or more than two sex chromosomes (XXX, XXY, or XXXX, for example). There are also many individuals for whom their sex chromosome pair (XY or XX) does not align with either the appearance of sex-associated traits or gender.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=381#h5p-67

Note 2: Although humans and many other species (including some plants[1]!) use X and Y sex chromosomes, other organisms use different systems of sex determination. Birds, for example, use Z and W sex chromosomes. In yet other species, sex can be determined by environmental conditions or even the ploidy of an organism. In honeybees, for example, males are typically haploid and females are diploid[2].

The X and Y chromosomes share similar DNA sequence at the ends

During meiosis I, the autosomes are paired with their homolog. For both autosomes and sex chromosomes, the pairing happens due to sequence similarities between the sister homologs. This is discussed in more detail in the chapter on meiosis and mitosis. This is possible, even when that "pairing" is between XY or ZW chromosomes because the very ends of the sex chromosomes contain homologous sequences. These are called pseudoautosomal (PAR) regions because both males and females have two copies of all the genes in those regions. The pseudoautosomal regions of the X and Y chromosomes are shown in **Figure 1**.[3]

During meiosis, chromosome pairs – including XX or XY pairs – are separated into different daughter cells. This results in haploid daughter cells. Each daughter cell ultimately contains one copy of each autosome and one sex chromosome. For mammals, eggs typically carry an X chromosome and sperm carry either an X or

1. Charlesworth, D. Plant sex determination and sex chromosomes. *Heredity* **88**, 94–101 (2002).
2. Sex Determination in Honeybees | Learn Science at Scitable. http://www.nature.com/scitable/topicpage/sex-determination-in-honeybees-2591764.
3. Image Source: Kelkar A, Thakur V, Ramaswamy R, Deobagkar D (2009) Characterisation of Inactivation Domains and Evolutionary Strata in Human X Chromosome through Markov Segmentation. PLoS ONE 4(11): e7885.

Y chromosome. For birds, eggs typically contain either a Z or a W chromosome, while all sperm contain a Z chromosome.

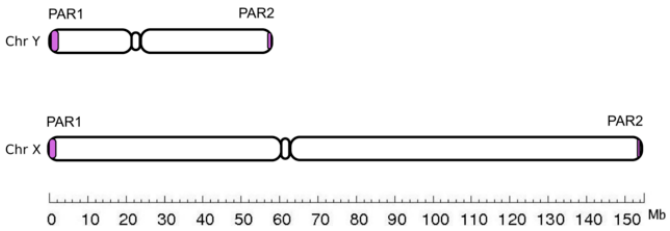


Figure 1. In humans and other mammals, the X and Y chromosomes share regions of homology at their very ends, called pseudoautosomal regions (PAR).

Media Attributions

- Pseudoautosomal-region © Image Source: Kelkar A, Thakur V, Ramaswamy R, Deobagkar D (2009) Characterisation of Inactivation Domains and Evolutionary Strata in Human X Chromosome through Markov Segmentation. PLoS ONE 4(11): e7885. is licensed under a CC0 (Creative Commons Zero) license

CHROMOSOMES AND SEX DEVELOPMENT

Sex is a phenotype, like any other observable trait. **Sex determination genes** are genes that control the development of sex-associated traits.

In mammals, fruit flies, and some flowering plants embryos, females usually (but not always) have an XX genotype while males usually (but not always) have an XY genotype. In birds and some reptiles and amphibians, males typically have a ZZ genotype and females have ZW genotype. Note that females are not always the homogametic sex – they don't always have two of the same chromosome.

These genotypes are associated with either male or female phenotypes due to the presence of **sex determination genes** located on the sex chromosomes. But the mechanism for development of sex-associated phenotypes is different depending on species. For example, although humans and Drosophila both have X and Y sex chromosomes, they have different mechanisms for determining sex.

Table 2 lists chromosomal methods of sex determination in several species. In humans and other mammals, the **SRY gene** on the Y chromosome triggers the development of testes in the early embryo.

Fruit flies and certain other insects also use an XY system of sex determination, but the mechanism of sex determination is different. For fruit flies, the ratio of X chromosomes to autosomes determines sex phenotype, due to the expression of autosomal genes that in turn influences the expression of sex-determination genes on the X chromosome.

In other insects, there's no Y chromosome at all! The number of X chromosomes influences maleness. And in birds and some other species, the DMRT1 gene on the Z chromosome initiates the process of sex development, but it is **haploinsufficient**: one copy of the gene does not produce enough gene product to trigger maleness, so ZW individuals will typically develop female anatomy.

Note: some species do not use sex chromosomes! Sex in some species can be determined by autosomes or even environmental conditions. In honeybees, sex is determined by whether eggs are fertilized: unfertilized eggs develop into males, while fertilized eggs develop into females. In many turtles, sex is determined by environmental temperature: cooler temperatures are associated with male development, and warmer temperatures with female development[1].

1. US Department of Commerce, N. O. and A. A. What causes a sea turtle to be born male or female? https://oceanservice.noaa.gov/facts/temperature-dependent.html.

Table 2 Chromosomal systems of sex determination

Organism	**Human and other mammals**	**Drosophila (fruit fly)**	**Some insects and other organisms**	**Birds, reptiles, some fishes and amphibians**
Chromosomal System	XX-XY	XX-XY (XA system)	XX-XO	ZZ-ZW
Chromosomal method of determination	Presence of Y chromosome triggers maleness during fetal development	Ratio of X:autosomal chromosomes determines sex	Number of X chromosomes determines sex; there is no Y chromosome.	Two Z chromosomes required for maleness
Mechanism	SRY gene on the Y chromosome sets off a series of events leading to development of testes. (Other genes both autosomal and sex-linked are also involved)	Expression of sex-determining genes on the X-chromosome is affected by timing (and level) of expression of autosomal genes	May vary among species: In *C. elegans* is related to the amount of an X-encoded protein which can inhibit maleness	DMRT1 gene on the Z chromosome is haploinsufficient; one copy of the gene can't produce enough protein to trigger male development

In all organisms, the sex chromosomes are associated with sex determination because they house some, but not all, of the genes responsible for determining sex. (Other sex-determination genes are located on the autosomes, as discussed below.). The sex chromosomes also house genes not responsible for sex determination at all. These are called **sex-linked genes**, but they have nothing to do with sex beyond their chromosomal locus. Some examples in humans are a gene linked with color-blindness and a gene that affects the strength of tooth enamel.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=384#h5p-68

Sex differentiation during human development

So, given what we know about genes, chromosomes, and gene expression, why does the presence of a Y chromosome typically trigger maleness during human development? The answer to that involves a network of both sex-linked and autosomal genes.

Sex is a phenotype. Typically, in most sexually dimorphic species, there are multiple characteristics, in addition to sex organs, that distinguish male from female individuals. Those sex-associated traits are phenotypes just like hair color or eye color or wing shape. Those phenotypes can be genetically (or in some cases environmentally) determined.

In humans and other mammals, the Y chromosome carries the SRY gene. The SRY protein encoded by the gene is a transcription factor. In early stages of human development, a human embryo develops a bi-potential genital ridge – that is, tissue that has the potential to become either ovaries or testes. Early embryos also have two systems of ducts, Wolffian and Müllerian, which can develop into the male and female reproductive tracts, respectively.

If a Y chromosome is present, the transcription factor SRY is produced. In turn, SRY activates the expression of Sox9, which is also a transcription factor. Sox9 in turn activates other genes, which eventually lead to the development of testes.

The testes in turn produce testosterone and anti-Müllerian hormone, or AMH. Testosterone (and other hormones, including 5α-dihydrotestosterone) trigger formation of other organs in the male reproductive system from the Wolffian duct tissue, while AMH causes degeneration of the Müllerian duct and suppresses the development of female sex structures.

In the absence of SRY (as in individuals with an XX genotype), an alternative set of molecular signals is typically activated, including WNT4, RSPO1, DHH (Desert Hedgehog), and β-catenin. These lead to development of ovaries. The ovaries then produce estrogen and trigger development of the uterus, oviducts, and cervix from the Müllerian duct.

The sequential activation of these genes is illustrated in **Figure 7**, below.

Thus, in humans and other mammals, the SRY protein product sets into motion this cascade of events that leads to maleness. However, SRY is not the only gene involved in this process – and, in fact, all the other genes shown in Figure 7 are autosomal, not sex-linked! A loss of function in any of the autosomal genes will disrupt the sex determination pathway, leading to differences in sex development, discussed in greater detail below.

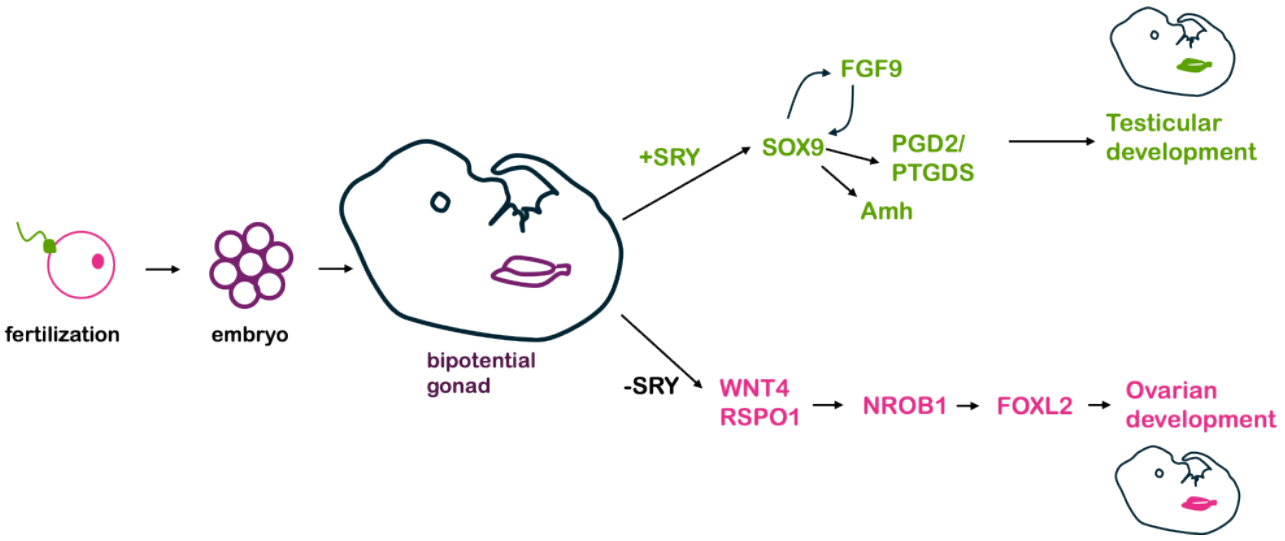


Figure 7 Selected genes required for sex determination and differentiation in humans and other mammals.

Differences of sex development in humans

Although the term "biological sex" is used quite often, this term is an oversimplification of the biology underlying the development of sex characteristics. There are a number of ways males and females differ from one another. As a result, there are many ways "biological sex" might be defined.

In this module, we've so far been focused on **chromosomal sex**, where an XY individual is considered male and an XX individual is considered female. But sex is usually determined or assigned at birth based on the visible presence of a penis or a vulva in a newborn. Sex can be determined by genitalia even before birth, when genitals are clearly visible by ultrasound.

Genitalia define **anatomical sex**. This is also sometimes called phenotypic sex, although this is an oversimplification of term phenotype. **Gonadal sex** refers to the presence of either testes or ovaries. Although individuals show a range of hormonal levels, human males and females tend to have different ranges of androgens (males typically have more, females typically have less) and estrogens (females typically have more, males typically have less). This is **hormonal sex**.

At puberty, both males and females develop secondary sex characteristics. In males, this includes deepening of the voice, growth of facial and body hair, and broadening of the shoulders. In females, secondary sex characteristics include growth of breast tissue, widening of hips, and onset of menses. These secondary sex characteristics typically (though not always) develop according to hormonal sex.

You might expect that these ways of defining biological sex (chromosomal, anatomical, gonadal, and hormonal) all align: XY individuals would have a penis, have testes, have relatively high levels of androgens like testosterone and low levels of estrogens, and develop male secondary sex characteristics at puberty. But this is not always true: it is possible to have an XY genotype but have female genitalia, gonads, and secondary sex characteristics. It is possible to have an XX genotype and have male genitalia, gonads, and secondary sex

characteristics. It is possible to have female external genitalia and internal testes. It is possible to have genitalia with both male and female characteristics, or indeterminate characteristics. In fact, it is possible to have almost any combination of chromosomal, genital, gonadal, and hormonal sex.

Genetically, these sex phenotypes are due to genotypic changes in any one of the networks of genes responsible for sex determination and differentiation – and because there are so many genes involved in the process, there are many Differences of Sex Development (DSDs) in the human population. Another term for DSD is intersex. Some examples of DSDs in humans are listed below, although this is not an exhaustive list.

Sex chromosome aneuploidy. Remember that aneuploidy refers to an atypical number of chromosomes. They often result from nondisjunction during meiosis or mitosis: either homologous pairs or sister chromatids fail to separate during anaphase. Aneuploidies of autosomes are relatively rare in the human population, since additional copies – or too few copies – of most chromosomes would have lethal effects on a developing embryo. (Some exceptions are Trisomy 21 – Down syndrome, Trisomy 13, and Trisomy 18, all of which have significant phenotypic effects and are linked with shortened lifespan.) Aneuploidies of sex chromosomes, however, are by far more common. This is likely because the Y chromosome has very few genes – none required for life – and any extra copies of the X chromosome can be inactivated (discussed more in the next section).

Sex chromosome aneuploidies are the most common DSDs in the human population as well as the most common aneuploidies in the human population. Humans with sex chromosome aneuploidies usually have a normal lifespan but may be infertile, may have neurological and/or cognitive disabilities, or they may have few or no measurable phenotypic differences. It's estimated that 50-75% of people with sex chromosome abnormalities never even know they have a sex chromosome aneuploidy.

Individuals without a Y chromosome typically – but not always – show a female phenotype, regardless of the total number of sex chromosomes. This is because the SRY gene on the Y chromosome is needed to initiate development of male reproductive structures. Individuals with one or more Y chromosomes typically – but not always – show a male phenotype since the presence of SRY triggers development of male reproductive structures during embryogenesis.

A selected list of human sex chromosome abnormalities is listed in **Table 3:** Sex chromosome aneuploidies in humans, below, with associated phenotype.

Table 3.

Genotype	Phenotype
XO (One sex chromosome only)	Turner syndrome. Anatomical females present with symptoms that vary in severity but may include short stature, ovarian failure, cardiac defects, and/or infertility that may be corrected with fertility treatment.
XXY	Klinefelter syndrome. Anatomical males present with signs and symptoms that vary in severity but may include taller than average stature, weak bones, delayed puberty, decreased muscle mass, and low sex drive. Some may go undiagnosed.
XYY	XYY syndrome. Anatomical males with normal fertility and sexual development. May be taller than average, have increased risk of cystic acne, ADHD, and, to a lesser extent, autism spectrum disorder. May go undiagnosed.
XXX	Triple X syndrome. Anatomical females who are generally taller than average, some with subtle physical differences including wide-spaced eyes. Some patients may have learning disabilities and medical problems including infertility due to premature ovarian failure, but others may have mild or no symptoms and go undiagnosed.

Translocation of SRY. In some anatomical males with an XX genotype, the SRY gene has been translocated to one of the X chromosomes or an autosome. This is often the result of an aberrant recombination event – crossing over between the X and Y chromosomes (or the Y and an autosome) during meiosis in the father. Such individuals appear phenotypically male but may have small testes or a urethra opening on the underside of the penis. Affected individuals may be shorter than average, require hormone treatments at puberty to trigger development of secondary sex characteristics, and may be infertile, but are otherwise healthy.

Swiyer syndrome. XY individuals who are phenotypically female, with functional vagina, uterus, and fallopian tubes but lacking ovaries. Because they lack ovaries, women with Swiyer syndrome are usually diagnosed in adolescence when they do not undergo a first period. Secondary sex characteristics do not develop without hormonal supplementation, but women can become pregnant with donated embryos. Women with Swiyer syndrome often have a deletion on the Y chromosome that inactivates the SRY gene, but the syndrome is also associated with mutations in the gene NROB1 on the X chromosome and several autosomal genes as well.

Complete androgen insensitivity (CAIS). 46, XY female phenotype. Mutations in the androgen receptor gene prevent cells of a developing embryo from responding to androgens, including testosterone. Although people with CAIS have normal-appearing female external genitalia and secondary sex characteristics, they have internal, undescended testes, and usually have a shortened vagina and lack a uterus.

Congenital adrenal hyperplasia (CAH). Mutations in the enzyme 21-hydroxylase block one step of cortisol synthesis. This in turn leads to a build-up of testosterone since cortisol and testosterone share parts of their biosynthetic pathway. This causes masculinization of the external genitalia in XX individuals, who may have ambiguous genitalia or even the appearance of normal male external genitalia. Such XX individuals have female gonads and internal genitalia.

5-alpha-reductase deficiency. These XY individuals have a loss of function mutation in the enzyme that

converts testosterone to dihydrotestosterone (DHT). DHT is the hormone required for the development of external male genitalia during fetal development. Children with 5-alpha-reductase deficiency may appear to have female external genitalia or ambiguous external genitalia at birth. However, they have internal testes, and, at puberty, the testes produce much higher levels of testosterone that cause penile and scrotal tissue to grow and male secondary sex characteristics to develop. This particular DSD occurs at high frequency in certain communities in the Dominican Republic, where such people are called "guavedoces". The term "guavedoces" very loosely translates to "penis at age twelve", since children assumed to be female appear to grow a penis.

With this expanded understanding of biological sex, it becomes apparent that while individuals with unambiguously male or unambiguously female characteristics make up the majority of the human population, individuals frequently do not completely align with either category. Thus, although sex is often considered binary – male or female – this is not an accurate picture of human biology, since there are many ways differences in sex development can arise. Humans may be chromosomally male but phenotypically female, or vice versa. They may be chromosomally male, hormonally male, but phenotypically female. Or they may have nearly any other combination of chromosomal, anatomical, gonadal, and hormonal sex.

Estimates of how many individuals in the human population have DSDs range depending on who is counted. If only individuals with differences in external genitalia are counted, such individuals make up about 1/4500-1/2000 of the human population. However, DSDs may result in typical external genitalia but atypical alignment of chromosomal, hormonal, or gonadal sex. They may also include individuals with atypical numbers of sex chromosomes – many of whom may not even know they have such a condition. If individuals with atypical sex chromosome ploidy are included, some estimates are that individuals with DSDs make up about 2% of the human population. For comparison, this is about the same frequency as the red hair phenotype worldwide – not as uncommon as one might think. If you know people with red hair, you likely know someone with a difference of sex development, too, making sex not very binary at all.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=384#h5p-69

Media Attributions

- Selected genes required for sex © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

SEX VS GENDER

It is important to note that, to this point, we have been discussing sex, not gender. Although the terms are sometimes used interchangeably, sex and gender refer to two different ideas. According to the American Psychological Association: "Sex is typically assigned at birth (or before during ultrasound) based on the appearance of external genitalia. When the external genitalia are ambiguous, other indicators (e.g., internal genitalia, chromosomal and hormonal sex) are considered to assign a sex, with the aim of assigning a sex that is most likely to be congruent with the child's gender identity." Gender is "a person's deeply felt, inherent sense of being a girl, woman, or female; a boy, a man, or male; a blend of male or female; or an alternative gender[1]."

Just like there are many individuals for whom chromosomal, hormonal, gonadal, and anatomical sex do not align, so too are there individuals whose innate gender identity or gender expression differs from their phenotypic sex as determined by external genitalia. This is called **transgender.** Individuals for whom their gender identity is the same as their phenotypic sex are **cisgender.**

Historically, the biology of gender has not been as well studied as sex, and, as such, it is poorly understood. However, a number of studies provide strong evidence that gender is innate, with both genetic and other biological causes. For example, transgender individuals may have brain structures that more closely match their gender than their phenotypic sex. A majority of individuals with 5-alpha-reductase deficiency choose to live as male after puberty, despite culturally having been raised as female to that point. XY babies with abnormal genitalia who are surgically assigned a female sex (a practice more common in previous decades) have a much higher incidence of transgender. And twin studies and adoption studies – standard ways to determine whether genetic factors affect a trait – strongly suggest a genetic component to gender. Twin studies and adoption studies are discussed in a later module.

Sex-determination vs sex-linked vs sex-influenced genes

In the previous section, we looked at sex-determination genes: genes that are responsible for the development of sex-associated structures in the developing embryo. SRY, Sox9, and RSPO1 are three examples mentioned previously. Sex-determination genes are not necessarily carried on the sex chromosomes: for example, both

1. American Psychological Association. Guidelines for psychological practice with transgender and gender nonconforming people. *American Psychologist* **70**, 832–864 (2015).

Sox9 and RSPO1 are encoded by autosomes. Other traits may be influenced by sex, but encoded by genes on autosomes. For example, male pattern baldness – also called androgenetic alopecia – appears to be linked with several autosomal loci. All sexes may experience alopecia (baldness), but androgens appear to influence the progression of hair loss, making the trait appear more often in hormonal males than hormonal females. Androgenic alopecia is thus an example of a **sex-influenced** trait. Some traits that are influenced by sex are actually limited by sex: many secondary sex characteristics fall in this category. Ovarian cancer would be an example of a sex-limited trait, since it only affects individuals with ovaries (gonadal females).

But what about genes located on the sex chromosomes? The sex chromosomes, especially the X chromosome, carry many genes that do not affect either sex-determination or sex-influenced traits. These are called **sex-linked** genes. **Sex-linked** traits are those that are encoded by genes on the sex chromosomes. Most of these sex-linked genes have functions that have nothing to do with sex-associated phenotypes.

The Y-chromosome is small and has few genes. True Y-linked traits are very rare, but only chromosomal males would ever display Y-linked traits. There are no dominant or recessive Y-linked traits, since most individuals have only one copy of the Y chromosome.

X-linked traits are more common, since the X chromosome carries about 900 genes[2]. Although chromosomal females are diploid for X-linked genes, males are haploid for X-linked genes. Because chromosomal males and females have different numbers of sex chromosomes, inheritance of **sex-linked** traits does not follow standard rules of Mendelian inheritance. Genetic crosses tracking sex-linked traits show differences in phenotypes among male and female offspring.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=393#h5p-70

X-Linked Gene example: the *white* gene in *Drosophila melanogaster*

The first examples of X-linked traits came from studies in *Drosophila melanogaster,* the common fruit fly. In

2. X chromosome: MedlinePlus Genetics. https://medlineplus.gov/genetics/chromosome/x/.

the early 1900's, Thomas Hunt Morgan followed up on Mendel's work in pea plants, using fruit flies as a model organism. Like peas, fruit flies were easy to grow with a short generation time (about two weeks). Different phenotypes could also be easily identified.

Normally, fruit flies have red eyes, so the red eye phenotype is considered **wild type**. Morgan found a fly with white eyes. The white eye phenotype was heritable. Morgan named the mutated gene "white", and described the wild-type allele as "w+" and the mutant allele "w-". Note the deviation from the capital/lowercase notation that Mendel used: even today, geneticists may describe wild-type alleles as "+" and mutant alleles as "-".

The white phenotype did not appear to follow Mendel's predictions for inheritance. Instead, different offspring ratios were observed for reciprocal crosses: a white female crossed with a wild-type male gave different offspring than a cross between a wild-type female and white male. Morgan was able to conclude that this was because the gene was linked to the X chromosome. More than a century later, we now know that the mutation in the white gene prevents the flies from synthesizing the (many) pigments that contribute to red eye color.

The white phenotype is recessive to wild-type: a heterozygous individual can produce pigment and its eye color is red. But because male flies (XY) only have one copy of the X chromosome, they always express the phenotype determined by that one allele. Female flies can be homozygous for the wild-type allele, heterozygous, or homozygous for the white allele. This is illustrated in Figure 8.

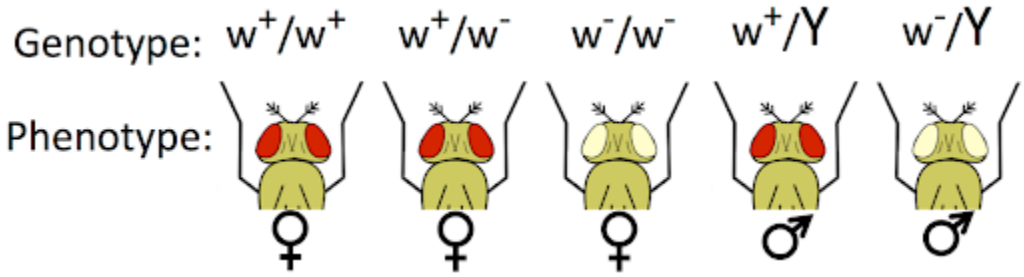


Figure 8. Relationship between genotype and phenotype for the white gene on the X-linked gene in Drosophila melanogaster. The Y chromosome is indicated with a capital Y because it does not have a copy of the white gene.

To better distinguish sex-linked genes from autosomal genes, we can use a modification of standard genetic notation for dominant and recessive alleles. While dominant or recessive autosomal alleles are typically indicated by capital or lowercase letters, respectively, if we want to indicate that a gene is X-linked we would use an X and a superscript to indicate the allele: X^W or X^w; X^{w+} or X^{w-}

We can likewise use a modified Punnett square, which will allow us to visually depict differences between male and female offspring. This is shown in **Figure 9**, below, where the two X chromosomes from the female parent are listed to the left, the X and Y from the male parent is across the top, and we can see the offspring include two dominant females, one dominant male, and one recessive male. This is an offspring ratio of 1:1 for males, but all dominant for females. Differing offspring ratios for males and females is characteristic of a sex-linked trait.

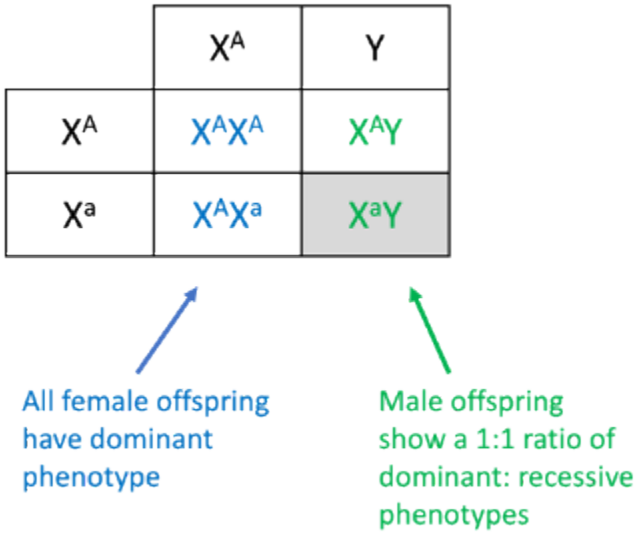


Figure 9. Punnett Square of an X-linked gene. X-linked cross between a dominant male and a heterozygous female. Female offspring are all dominant, and male offspring are present in a 1:1 ratio of dominant: recessive.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=393#h5p-71

Crosses that track sex-linked traits often show different phenotype ratios among male and female offspring. In addition, **reciprocal crosses** of sex-linked traits may show different offspring ratios. Reciprocal crosses test the influence of parental sex toward the offspring phenotypes. For example, a male of phenotype A might be crossed with a female of phenotype B (AxB), and in the reciprocal cross a male of phenotype B might be crossed with a female of phenotype A (BxA). Traits that follow Mendelian patterns of inheritance show no difference in reciprocal crosses, but sex-linked genes do.

Morgan's results with the white flies demonstrate the results of a reciprocal cross (**Figure 10**). In the first cross, a male with the mutant phenotype is crossed with a wild-type female. In the second cross, a female with the mutant phenotype is crossed with a wild-type male. The two crosses showed different offspring phenotypes.

When male and female offspring from a single cross show distinct phenotypes, or if reciprocal crosses give different offspring phenotypes, a sex-linked gene is often the cause.

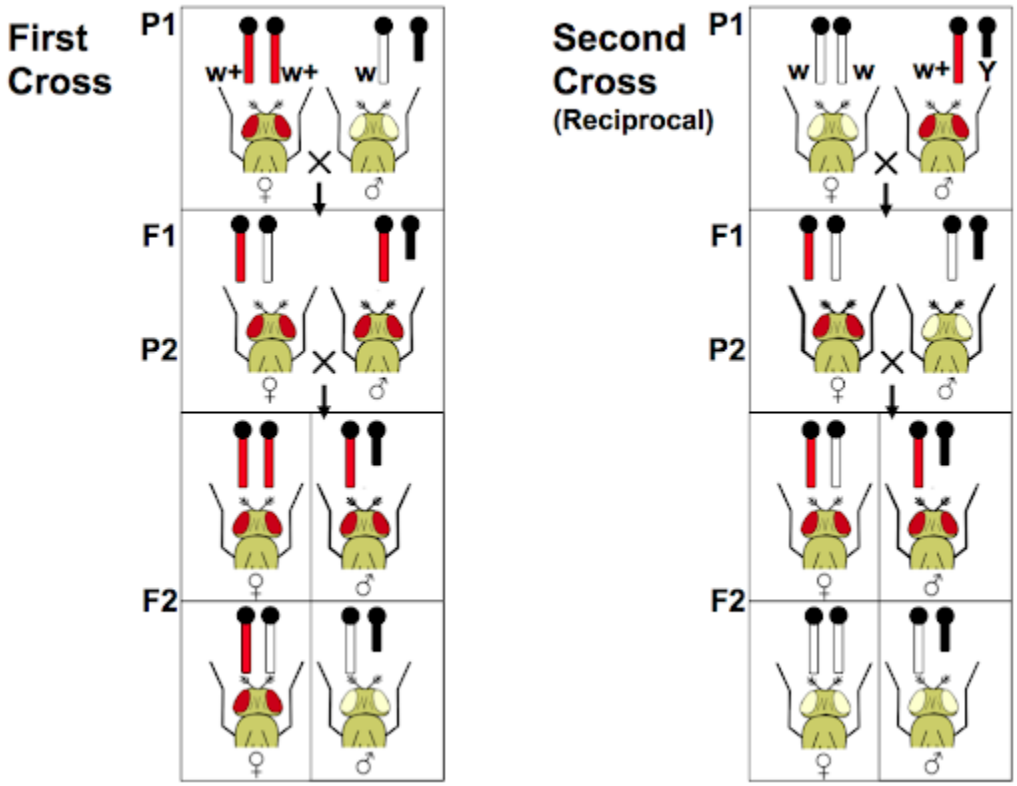


Figure 10. Reciprocal crosses involving an X-linked gene in Drosophila melanogaster. In the first cross (left) all of the offspring have red eyes. In second (reciprocal) cross (right) all of the female offspring have red eyes and the male offspring all have white eyes. If the F1 progeny are crossed (to make the P2), the F2 progeny will be different in each cross. The first cross has all red-eyed females and half red-eyed males. The reciprocal cross has half red-eyed males and females. Thomas Morgan won the Nobel Prize for using these crosses to demonstrate that genes (such as white) were on chromosomes (in this case the X-chromosome).

Dosage Compensation

As for autosomal traits, X-linked traits often show incomplete dominance, codominance, variable expressivity, or incomplete penetrance. For many traits, molecular mechanisms of **dosage compensation** cause these types of phenotypes.

What is meant by dosage compensation? Because chromosomal males have only one copy of the X chromosome, and chromosomal females have two copies of the X chromosome, it would be expected that chromosomal females produce twice as much protein product from X-linked genes as do chromosomal males. This could be problematic, since the amount of protein produced can affect phenotype, and most genes on the sex chromosomes are not sex-related. However, it turns out that males and females do generally produce similar

amounts of protein product from sex-linked genes. This is because most species with sex chromosomes use some form of **dosage compensation**, which normalizes the amount of expression from the sex chromosomes so that there is no difference in protein produced in males and females.

Dosage compensation works differently in different species, as illustrated in Figure 5. For example, in the fruit fly *D. melanogaster,* the X chromosome is hyper-transcribed: it is transcribed twice as often in males as it is in females. This results in similar amounts of RNA produced in males and females, and similar levels of protein. In the round worm *C. elegans*, however, the opposite is true: X chromosomes in XX individuals are hypo-transcribed, or transcribed half as much as in individuals with only one X chromosome. (*C. elegans* has no Y chromosome, and the genotype is written as XO to indicate there is only one sex chromosome.)

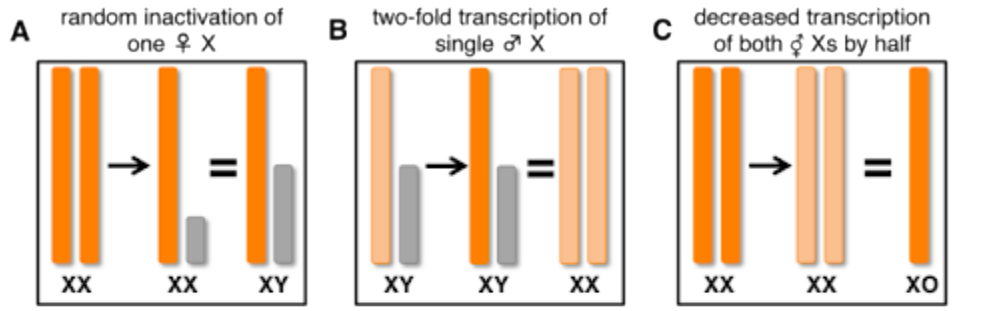


Figure 1. Three main mechanisms of dosage compensation observed in common model eukaryotic organisms.
(A) Mus musculus (as well as humans and many mammalian organisms) follow a pattern of dosage compensation in which females randomly inactive one X chromosome in every cell in their body to match male expression of a single X chromosome.
(B) D. melanogaster follow a mechanism of dosage compensation in which males increase the transcription of their single X chromosome by two-fold in order to match female expression of two X chromosomes.
(C) C. elegans hermaphroditic worms express both X chromosomes by half to match male worms with only one X chromosome.

Figure 11. Methods of Dosage compensation.

Mammals use a process called **X inactivation** to compensate. In each cell with more than one X chromosome, the "extra" X chromosomes are inactivated by packaging the DNA into tightly packed heterochromatin, full of chromatin modifications that make the chromosome inaccessible to transcription machinery. This is a form of **epigenetic** modification, as the chromatin modifications (and the X-inactivation) is inherited by mitotic daughter cells. Epigenetics is discussed in more detail in a separate module.

Interestingly, for individuals with additional copies of the X-chromosome (for example XXY or XXXX) all but one X chromosome will typically be silenced. This silencing may explain why sex chromosome aneuploidies typically cause few phenotypic effects: extra copies of X-linked genes are not expressed. Autosomal aneuploidies, on the other hand, are far less common, because an extra copy of an autosome affects the overall level of expression of nearly every gene on the chromosome.

The tightly packed X chromosome is called a Barr body, and it is clearly visible by light microscopy. This is shown in Figure 6. The alleles on the inactivated X-chromosome are never expressed, leaving only one protein-producing X chromosome per cell.

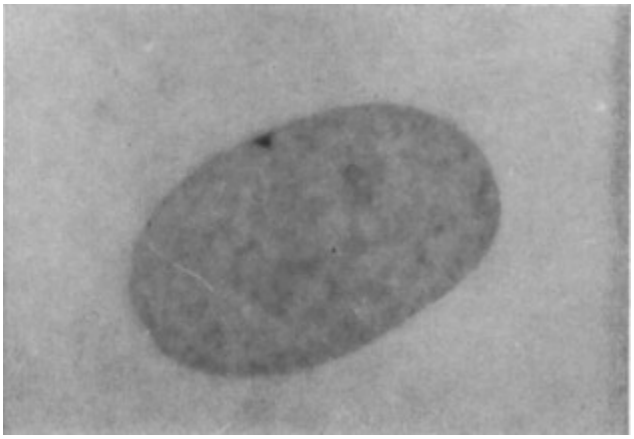


Figure 12. Barr body.

Which X chromosome is inactivated is mostly random, with the maternal X inactivated in some cells of a developing embryo and the paternal X inactivated in others. Inactivation appears to happen early during embryo development. As the cells of the embryo continue to divide, cells with an inactive X chromosome give rise to cells that have the same inactivated X-chromosome. In the mature organism, this results in patches of cells that all have the same X-chromosome expressed (and the same one silenced).

Calico cats

The expression of a fur color gene on the X-chromosome in cats gives a striking visual depiction of X-inactivation. This fur color gene *O* exists in two alleles: one that encodes black pigment, O^B, and one that encodes orange, O^O. Male cats generally only have one X chromosome and one allele. Female cats, with two X chromosomes, can be heterozygous for both orange and black. Due to inactivation of the maternal X in some cells and the paternal X in other cells, such heterozygous XX cats end up with both black and orange fur in a patchwork of orange and black. This patterning is called "calico" or "tortoiseshell" patterning as shown in Figure 11a. A tortoiseshell cat is shown in Figure 11b. (Calico cats have a white background color that is due to the action of a separate gene.)

To tie this back to terminology used in other chapters of this text: the orange and black alleles are **codominant**, since both are visible in the phenotype, and X-inactivation explains the mechanism of co-dominance. Note the unusual notation to describe the alleles: rather than use a capital and a lowercase letter, the alleles are both described by a capital letter with a superscript distinguishing the alleles from each other.

Figure 13.1 Different alleles of the O gene give either an orange (homozygous O^OO^O), patchwork (heterozygous), or black phenotype (homozygous O^BO^B) in females. Males are hemizygous with only one allele. **Figure 13.2** Tia, a tortoiseshell cat. Image courtesy Aline Davis.

Tortoiseshell and calico cats are nearly always female. Notably, fewer than 1/1000 calico cats are male. Those rare male calico cats often have extra copies of their sex chromosomes: for example, an XXY or XXXY karyotype[3].

Examples in humans

The *Orange* gene in cats is a good visual demonstration of how the mammalian dosage compensation system affects gene expression. However, most X-linked genes do not produce such striking mosaic phenotypes in heterozygous females. For most X-linked alleles, provided that some cells are expressing the dominant allele, a heterozygous individual will have a dominant phenotype. Two examples are the F8 gene encoding Factor VIII, a protein important for blood clotting, and the green cone pigment gene (GCP) that participates in color vision.

Hemophilia A is a bleeding disorder associated with certain loss of function alleles of the F8 gene. The F8 gene is normally expressed in liver cells and released into the bloodstream, where it plays an important role in blood clotting. Patients with hemophilia A have impaired blood-clotting ability and are prone to excessive bleeding, both spontaneously and after injury. This can be life-threatening.

Like most X-linked recessive traits, hemophilia A Is much more common in XY individuals than in XX individuals, since XY individuals need only one copy of the disease-associated allele to have the phenotype. XX individuals heterozygous for the recessive hemophilia-associated allele typically do not have hemophilia, even though only half (on average) of their liver cells produce Factor VIII protein. About half of their liver

3. Pedersen, A. S., Berg, L. C., Almstrup, K. & Thomsen, P. D. A tortoiseshell male cat: chromosome analysis and histologic examination of the testis. *Cytogenet Genome Res* **142**, 107–111 (2014).

cells inactivate the healthy allele, and half of the cells inactivate the allele associated with hemophilia. The cells with the active healthy X chromosome release enough Factor VIII into the bloodstream to prevent excessive bleeding.

A second example is impaired color vision (or color blindness). In the eye, photoreceptor cells called "rods" and "cones" detect light. Rod photoreceptors are highly sensitive and function well in conditions of low light, while cone photoreceptors detect different wavelengths of light (different colors) but are less sensitive under low light conditions. Humans typically have three types of cones, each of which senses a different spectrum of light best (short, medium, or long wavelength). The short, medium and long wavelength cones are sometimes called blue cones, green cones, and red cones, respectively, named for the color of light to which they are most sensitive. Color vision is a result of the brain's interpreting a combination of signals from all three types of cones.

The three types of cones sense different wavelengths of light due to the expression of different cone pigment proteins (called "opsins"). For some individuals with impaired color vision, the gene sequence of one of the cone pigments is altered, resulting in two pigments that sense similar wavelengths of light. This makes such people less able to distinguish between colors. Others with impaired color vision may be missing one of the cone pigment genes entirely – they are missing an entire segment of the chromosome![4]

Figure 14 approximates what is seen by individuals with three forms of color blindness, caused by differences in each of the three cone pigments.

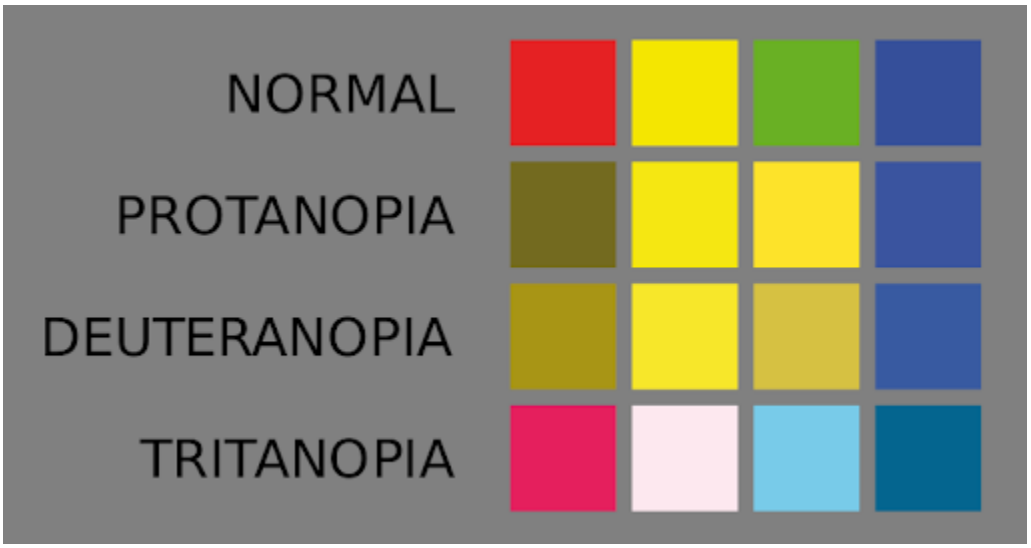


Figure 14. Comparison of the colors seen by people with normal color vision and three types of impaired color vision. Protanopia is caused by a loss of function of the green cone pigment. Deuteranopia is caused by a loss of function of the red cone pigment. Tritanopia is caused by a loss of function of blue cone pigment.

4. Deeb, S. The molecular basis of variation in human color vision. *Clinical Genetics* **67**, 369–377 (2005).

Both the red cone pigment gene and the green cone pigment gene are found on the X chromosome.

Chromosomal males who inherit one recessive allele of the green cone pigment gene have defective cone photoreceptor cells and cannot distinguish between certain shades of red and green. Chromosomal female heterozygotes usually have normal color vision, although there are some exceptions to this. As long as an XX individual retains some photoreceptor cells that express the healthy cone pigment allele, that person will be able to distinguish colors with a sensitivity close to that of someone with two healthy alleles. However, there are rare examples of heterozygotes with impaired color vision due to X-inactivation that is highly skewed toward the inactivation of the healthy allele[5]. This is similar to a calico cat that is mostly black with only a few orange patches (which can happen too).

Red-green colorblindness is one of the most common examples of X-linked recessive traits in humans. The trait seems to be most common among people of European descent, where up to 8% of males have some difficulty distinguishing between red and green. It is less common among people of other geographic ancestry; for example, some estimate about 5% of males of Chinese and Japanese descent and less than 4% of males of African descent have some form of red-green colorblindness[6].

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=393#h5p-72

Media Attributions

- Genotype and phenotype © Open Online Genetics, Nickle and Barrette-Ng. is licensed under a CC0 (Creative Commons Zero) license
- Punnett Square of an X-linked gene © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Reciprocal crosses © Online Genetics, Nickle and Barrette-Ng. is licensed under a CC0 (Creative

5. Jørgensen, A. L. *et al.* Different patterns of X inactivation in MZ twins discordant for red-green color-vision deficiency. *Am J Hum Genet* **51**, 291–298 (1992).

6. Deeb, S. The molecular basis of variation in human color vision. *Clinical Genetics* **67**, 369–377 (2005).

Commons Zero) license

- Three main dosages © CarmenMariaConroy via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Sex-chromatin body © Barr Body is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Alleles of the O gene © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Comparison of the colors © Japs 88 via. Wikipedia is licensed under a CC0 (Creative Commons Zero) license

WRAP-UP QUESTIONS

1. Explain why autosomal aneuploidies result in far more severe phenotypic consequences than sex chromosome aneuploidies.

2. Haploinsufficient genes are rare on the X-chromosome of humans and other mammals. But there are many haploinsufficient genes on the Drosophila X-chromosome[1]. Explain why, given the differences in dosage compensation in each species. (Hint: remember that haploinsufficient genes require the action of both copies of the gene to produce the phenotypic effect.)

3. In birds, males are the homogametic sex and females are the heterogametic sex. Draw a modified Punnett Square that tracks the inheritance of a Z-linked recessive trait from the cross $Z^{a}W$ x $Z^{A}Z^{A}$.

Science and Society

4. The term Differences of Sex Development (DSD) is used to describe conditions in which a person's sex phenotype is different from what is most common for their genotype. This reflects a recent transition in vocabulary: in the past, DSDs were more frequently described as Disorders of Sex Development. In your opinion, what does the change in terminology indicate about changing views of society?

1. de Clare, M., Pir, P. & Oliver, S. G. Haploinsufficiency and the sex chromosomes from yeasts to humans. *BMC Biology* **9**, 15 (2011).

PART XIV

FAMILY TREES AND PEDIGREES

Objectives

1. Use concordance data from twin studies to determine whether a trait is caused by genetic or non-genetic factors.
2. Interpret a multi-generational pedigree to
    a. Determine mode of inheritance of a trait
    b. Calculate risk of inheritance of a trait

Is a trait caused by genetic factors?

Sometimes, it's not so easy to determine whether a trait is caused by genetic factors. Traits can be variably expressive or incompletely penetrant. Traits may be controlled by multiple genes, as we saw with epistasis and quantitative trait loci. A single trait can also be influenced by both genetic and non-genetic factors. In previous modules, we've looked at expected patterns of inheritance and phenotypic ratios for parents and offspring, in some cases tracking alleles and traits through multiple generations of controlled crosses. Refer back to the module on Sex for a description of how such crosses can be used to show sex-linkage for the white eye and other phenotypes in *Drosophila*.

Controlled crosses are powerful genetic tools in organisms that can be manipulated in the lab. But in humans and other organisms where controlled crosses are inappropriate, medical researchers turn to other tools of genetics to understand the causes of diseases and other phenotypic variation. In this module, we will look at concordance studies and pedigree analysis. While this section will focus primarily on examples from human genetics, these tools can be used to study any trait or any species.

Example: Celiac Disease

Do you know someone who follows a gluten-free diet? This has become pretty common in the last several decades, with food suppliers meeting the demand for gluten-free products like breads, pasta, snacks, and desserts.

Glutens are a class of proteins found in wheat and certain other grains like barley and rye. These proteins give products like breads and cakes their elasticity: during the baking process, crosslinks between protein chains form a mesh-like network. This gives structure to the dough or batter, and air bubbles trapped in the network of protein chains make cakes and breads rise. Different types of flour have different amounts of gluten. Bread flour typically has a higher gluten content, giving a denser network of gluten fibers and a chewier texture to the final product. Pastry flour typically has a lower gluten content, and baked goods made with pastry flour are usually more delicate and tender.

Gluten-free flours are made from plants like rice or tapioca, which do not have gluten. If you've had baked goods made with gluten-free flour, you may have noticed that the texture tends to be more crumbly than would be expected from the same food made with wheat-based flour. This is because, without the network of gluten proteins, there is less structure to the batter or dough.

For some people, consuming gluten triggers unpleasant symptoms like digestive issues (diarrhea or constipation), skin rashes, or bloating. Celiac disease is a specific type of gluten sensitivity caused by an immune response to the gluten proteins. The symptoms of celiac disease have been known for centuries, the connection of symptoms to wheat consumption was not recognized until the 1950's. This was after several decades of treating symptoms with restrictive diets that coincidently excluded wheat, often by promoting single foods like bananas or rice[1].

Understanding why some people but not others develop gluten sensitivity (and immune response) was difficult. Celiac disease does appear to run in families: someone with a relative with celiac disease is more likely to also develop the disease than someone with no close relatives with the disease.

But is this because of shared DNA or a shared environment? Families share DNA – and genetic traits – but they also share foods, environment, culture, and behaviors. Those non-genetic factors can also influence family clusters of disease or other traits. For example, if a family were all exposed to the same contaminated water, they might all develop the same rare disease, even if it were not caused by genetics. So how do we determine the cause?

Today, we recognize celiac disease as a multifactorial immune disease, caused by both genetic and non-genetic factors. To earn this knowledge required decades of work from many medical researchers. Genetic

1. History of Celiac Disease | BeyondCeliac.org. *Beyond Celiac* https://www.beyondceliac.org/celiac-disease/celiac-history/.

tools like pedigree analysis, twin studies, and genetic association studies were used together to find the causes. This module describes pedigree analysis and twin studies.

NOT ALL TRAITS AND DISEASES ARE GENETIC: CONCORDANCE STUDIES

Although many human traits and conditions are caused by genetics, many are not. Environmental factors, diet and lifestyle, and even randomness can play a role. And, for many conditions, there are both genetic and non-genetic factors.

Some diseases are caused by a mutation in a single gene: examples are cystic fibrosis (the CFTR gene), sickle-cell anemia (the HBB gene), and Huntingdon's disease (the HTT gene) are all highly penetrant, caused by genetic mutations in a single gene. Other traits are controlled solely by genetics, but are influenced by multiple genes; for example, eye color is controlled by several genes working together, which makes inheritance patterns less straightforward than we'd expect from single-gene inheritance.

On the other hand, diseases like scarlet fever, COVID-19, influenza, and botulism are caused by exposure to bacteria or viruses. Exposure to heavy metals like lead can cause neurodevelopmental problems in children. These are solely non-genetic causes.

And, to make things more complicated, some traits have both a genetic and a non-genetic component. Examples of this include certain hereditary cancer syndromes (discussed in the Cancer Genetics Module) and phenylketonuria.

People with phenylketonuria have a homozygous loss of function mutation in the gene encoding phenylalanine hydroxylase (PAH). This enzyme is responsible for the metabolism of phenylalanine. Exposure to phenylalanine – an amino acid that is common in many foods – can cause severe neurological and other health problems. A diet low in phenylalanine can largely prevent the most extreme conditions associated with phenylketonuria. Thus, both genetic (PAH mutations) and nongenetic (diet) factors influence the condition.

In some cases, non-genetic influences can affect the penetrance or expressivity of a trait; in the example of phenylketonuria, changes in diet can lessen or even prevent the neurological symptoms of the disease, making the neurological symptoms variably expressive or incompletely penetrant.

Because of shared environment, factors like infectious disease, diet, and exposure to toxins can be shared in families, which might make it appear as if there is a genetic cause even if there isn't. So, when studying a condition, **how can geneticists distinguish between genetic and non-genetic causes?**

Twin Studies

Studies of twins and adoptive families (where families share an environment and culture, but not DNA) can

help geneticists distinguish between traits that have a genetic component (even if multifactorial) and traits that are shared among family members because of a shared environment.

Twin studies typically compare the **concordance** among monozygotic (identical) twins to the concordance among dizygotic (fraternal) twins. **"Concordance"** means agreement – it is the percentage of twin pairs that match each other.

For a fully genetic trait or disease, identical twins match in phenotype 100% of the time. They share all of their DNA, and their concordance would be 100. Dizygotic twins, on the other hand, share no more DNA than any other siblings – about 50%. Concordance in dizygotic twins for a genetic trait is lower than for monozygotic twins.

For a trait or disease that is wholly environmental – symptoms due to exposure to a toxin or a virus, for example – both monozygotic twins and dizygotic twins are exposed to similar environments and sometimes may both be affected, although sometimes one twin might not be affected. Concordance of less than 100 in monozygotic twins thus indicates an environmental component.

For many diseases, though, there might be both a genetic and environmental component. One might inherit a predisposition to disease, for example, that is mitigated by diet (like phenylketonuria). For traits with both an environmental *and* genetic cause, both conditions would be met: monozygotic concordance less than 100, and monozygotic concordance greater than dizygotic concordance. Some examples are given in Table 1 below.

Table 1. Concordance for example conditions

Condition	Monozygotic Concordance	Dizygotic Concordance	Results	Cause of trait
Eye color[1]	100	49	MZ>DZ	Genetic
Colon cancer[2]	4.7	2.6	MZ>DZ and MZ<100	Genetic and Environmental
Scarlet Fever[3]	88	92	MZ<100	Environmental

1. Mackey, D. A. What colour are your eyes? Teaching the genetics of eye colour & colour vision. Edridge Green Lecture RCOphth Annual Congress Glasgow May 2019. *Eye* **36**, 704–715 (2022).

2. Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315**, 68–76 (2016).

3. Gedda, L. *et al.* Heredity and Infectious Diseases: A Twin Study. *Acta Genet. Medicae Gemellol. Twin Res.* **33**, 497–500 (1984).

Note that for many traits with both genetic and non-genetic components, monozygotic concordance may actually be quite low. It's not the monozygotic concordance by itself that shows a genetic component: it is the monozygotic concordance **relative to** dizygotic concordance. We see that for colon cancer in Table 1.

For colon cancer, the low rates of monozygotic concordance are due to the fact that colon cancer is a heterogeneous disease, caused by different factors in different people. Sometimes colon cancer occurs randomly, with no previous family history of the disease. This can be influenced somewhat by diet and lifestyle. In other cases, individuals may inherit an allele that predisposes them to colon cancer. This can be due to heterozygous loss of function mutations in tumor suppressor genes like *MSH2* or *MLH3*. The families of such patients may have a high number of relatives who also develop colon cancer. But whether or not an individual with a mutant allele of *MSH2 or MLH3* actually develops cancer depends on the acquisition of additional somatic mutations over the course of their lifetime. Not all such individuals will acquire the additional mutation, and so not all will develop cancer (it is incompletely penetrant). This is discussed more in the Cancer Genetics module.

Test Your Understanding

Concordance Conditions

Condition	Monozygotic Concordance	Dizygotic Concordance
A	72	74
B	12	3
C	23	24
D	100	75

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=399#h5p-73

Twin studies are limited in how they can be interpreted. Sometimes this is because only incomplete data is available: there are many case studies in the literature where only one or two sets of twins are studied. In these cases, if a set of monozygotic twins are not concordant, it's possible to conclude that there are non-genetic factors at play. Several case studies on celiac disease in twins from the 1970's-1990's fall in this category, with monozygotic twins occasionally showing discordance. But without a large enough sample size to compare

monozygotic twins and dizygotic twins, it's not possible to conclude whether there is also a genetic component to the trait. A large-scale twin study of celiac disease was published in 2001.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=399#h5p-74

Another example of this is relatively recent work on gender, comparing cisgender and transgender twins. You'll recall from the module on Sex the difference between sex and gender. Sex refers to measurable features associated with maleness or femaleness, such as gonads, external genitalia, or chromosomes. Gender identity refers to an innate sense of self. An individual might self-identify as male, female, or another gender. Sometimes sex is referred to as "biological sex", suggesting that there is no biological component to gender. But is that true? Some twin studies have suggested otherwise.

One comparison of concordance[4] in trans- and cisgender twins shows that for male twin pairs, MZ =33 and DZ = 5. For female twin pairs, MZ = 23 and DZ = 0. Because MZ < 100, there does appear to be a nongenetic component to gender. But because MZ > DZ, there is also strong evidence for a genetic component to gender. These studies also thus strongly suggest a biological component to gender as well as to sex.

Some interpretations of this type of work wrongly suggest that because there is a nongenetic component to gender, there cannot also be a genetic component. But that is contrary to our understanding of many traits influenced by both genetic and nongenetic factors. This is similar to the example of colon cancer: As shown in Table 1, monozygotic concordance for colon cancer is relatively low (4.7) but is still higher than dizygotic concordance (2.6). Colon cancer is linked to both genetic factors (mutations in tumor suppressor genes like *MSH2* and *MLH3*) and non-genetic factors.

It is important to recognize the limitations of twin studies. Misinterpretation can happen when the dizygotic control is ignored, or if a small number of twin pairs makes a good comparison difficult. Remember, it's not the actual value of the monozygotic concordance that matters but the value relative to the dizygotic concordance. Sometimes a low monozygotic concordance might be misinterpreted as "no genetic component" if it is not compared to the dizygotic control. Also, keep in mind that identical twins may be treated more similarly than fraternal twins, so the monozygotic/dizygotic comparison is not a perfect control.

4. Diamond, M. Transsexuality Among Twins: Identity Concordance, Transition, Rearing, and Orientation. *Int. J. Transgenderism* **14**, 24–38 (2013).

As a summary: if monozygotic concordance is less than 100 (MZ<100), there is a non-genetic or environmental cause for a trait. Monozygotic concordance greater than dizygotic concordance (MZ>DZ) suggests there is a genetic cause for a trait. If both conditions are met (MZ<100 and MZ>DZ) there is likely both a genetic component and a non-genetic component to the trait.

Adoption studies

Twin studies can identify conditions for which a genetic component exists by comparing two groups of people with shared environment but differing amounts of shared DNA: monozygotic twins (with 100% shared DNA) and dizygotic twins (with about 50% shared DNA). Adoption studies are similar, comparing individuals with either shared DNA (biological relationships) or shared environment (adoptive relationships). If a trait is shared more often between biological family members than adoptive family members, this suggests that there is a genetic component for the trait.

Adoption studies have shown that body mass index (BMI) and obesity are correlated for biological parents and children but not adoptive children[5], suggesting a strong genetic component for obesity. Adoption studies have also suggested a genetic component to alcohol use disorder and drug dependency[6].

5. Stunkard, A. J. *et al.* An Adoption Study of Human Obesity. *N. Engl. J. Med.* **314**, 193–198 (1986).

6. Dinwiddie, S. & Cloninger, R. Family and Adoption Studies in Alcoholism and Drug Addiction. *Psychiatr. Ann.* **21**, 206–214 (1991).

PEDIGREE ANALYSIS

While adoption and twin studies can suggest a genetic or non-genetic cause for a trait, tracking a trait through multiple generations of an extended family can provide additional information.

Pedigrees are pictorial representations of a family tree, used to track phenotypes (or sometimes genotypes) through a family. They are most commonly used to study rare traits in a population.

Pedigrees can be used to determine the **mode of inheritance** for a particular trait. The mode of inheritance refers to how a trait is encoded by the genome: is it a dominant or recessive trait? Is it encoded on an autosome, sex chromosome, or mitochondrial chromosome? We will look at the patterns for six different modes of inheritance: autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive, Y-linked, and mitochondrial. We will also use pedigrees to determine the risk of an individual developing a trait or disease.

By convention, a standard set of symbols are used to represent males and females, parents and children. Squares represent males, and circles represent females. An individual of unknown sex is indicated by a diamond. A diamond is also sometimes used to indicate someone who is intersex or of nonbinary gender, although this is not common usage and is not accepted by all in the nonbinary or genetics communities.

A horizontal line connects a pair of parents, with a vertical line connecting parents to offspring. An individual who has the phenotype tracked in the pedigree will be indicated by a filled-in shape, while an individual who does not show the phenotype will be indicated by an empty shape. An individual with a trait is also said to be **affected** by the trait.

If parents have more than one offspring, the vertical offspring line will be connected with a square fork, with siblings oldest to youngest arranged left to right. Each generation is indicated by a Roman numeral, with "I" being the oldest generation. Within a generation, individuals are numbered with Arabic numerals, starting with "1" and counting left to right.

Two example pedigrees are shown in **Figure 1**. On the left is a family with one child who is affected by the trait. On the right is a pedigree with three children, the oldest being male and the two youngest female.

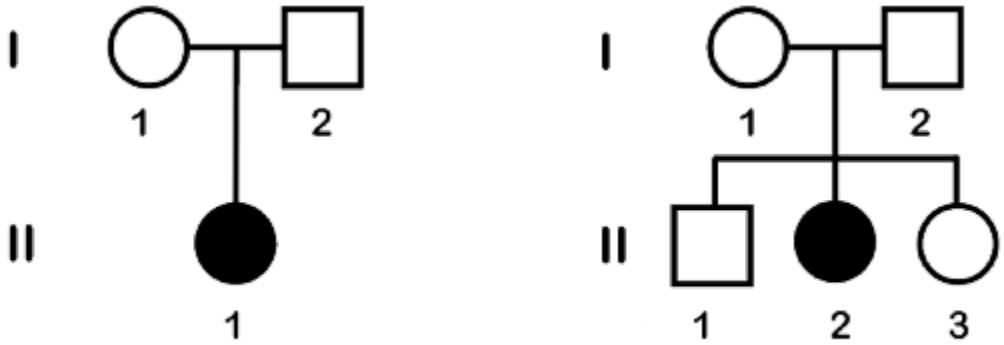


Figure 1. Two examples of simple pedigrees.

In many cases, a pedigree is first constructed in response to an individual who comes to the medical community. This person is called the **propositus** or the **proband**, and they are indicated on the pedigree with an arrow. In **Figure 1**, the proband is individual II-1.

Additional symbols may be also used in genetic pedigrees, including symbols that indicate heterozygous carriers of an allele (who do not show the trait), symbols to represent consanguineous mating (parents who are blood relations), symbols to represent twins, and symbols that include adopted family members. A selection of some pedigree symbols is shown in **Figure 2**.

A note on how sex is depicted in pedigrees

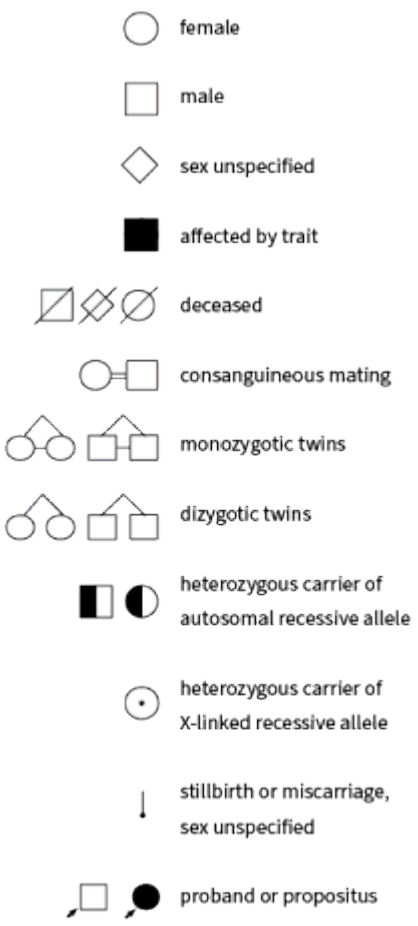


Figure 2. A selection of symbols used in genetic pedigrees.

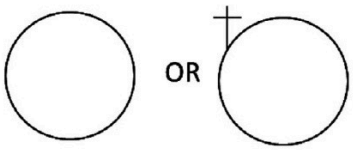
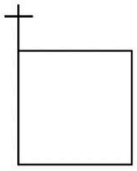
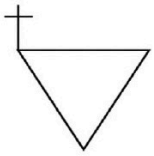
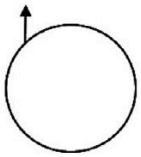
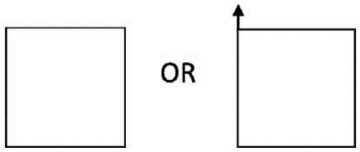
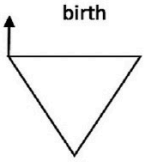
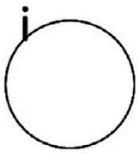
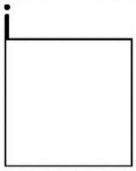
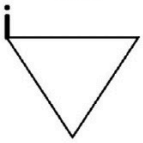
Historically, circles have been used to represent females and squares males in pedigrees. As geneticists gain a better understanding of sex and gender, however, it becomes increasingly apparent that this is not always an accurate representation of either an individual's sex or gender. Remember from the module on sex that there are multiple ways to define "male" and "female". For example, an individual with chromosomes typically associated with males (XY) can have gonads and/or external anatomy that is typically associated with females. Likewise, an individual with chromosomes typically associated with females can have gonads and/or external anatomy that is typically associated with males. Depending on the purpose of the pedigree, different definitions of male and female may be more useful than others.

In this module, we will use pedigrees to determine mode of inheritance, including whether a gene is carried on the X- or Y-chromosome. Therefore, unless otherwise indicated, we will most often use the symbols to represent chromosomal sex, with circles to represent people with the most common female chromosomal genotype (XX) and squares to represent people with the most common male chromosomal genotype (XY). It should be understood that this is not a complete description of any individual's sex (sex assigned at birth, sex determined by external genitalia, or gonadal sex may be different may be different from chromosomal sex) or gender (which may be cisgender or transgender).

Other symbols may offer a more complete or accurate representation of a person's sex or gender. While there is not currently a single standard for representation of transgender and gender diverse individuals or

individuals with differences of sex development, there are a number of proposals from within the genetic counseling community[12]. One example, from Tuite et al (2020), is shown in **Table 2**.

Table 2. Gender inclusive pedigree symbols.

	Identifies as girl/woman	Identifies as boy/man	Identifies as non-binary
Assigned female at birth	Cis girl/woman 	Trans boy/man 	Non-binary 
Assigned male at birth	Trans girl/woman 	Cis boy/man 	Non-binary, assigned male at birth 
Assigned intersex at birth	Girl/woman, assigned intersex at birth 	Boy/man, assigned intersex at birth 	Non-binary, assigned intersex at birth 

Uses and limitations of pedigree analysis

Like twin and adoption studies, pedigree analysis has limitations. A single family is, by definition, only a very small fraction of the population. Small sample sizes can make it seem like patterns are caused by an underlying genetic mechanism when it is really just random variation. As an example, we expect about 50% of the human population as a whole to be male. But due to the randomness inherent in meiosis (and biology in general!) many individual families may have all male offspring or all female offspring. In this way, as we interpret pedigrees, we must distinguish between what is *most likely*, what is *possible* (even though unlikely), and what can be ruled out. It's also important to remember that pedigrees represent the best-available information about a family. It often relies on information recalled from family members. If signs and symptoms of a trait are

1. Sheehan, E., Bennett, R. L., Harris, M. & Chan-Smutko, G. Assessing transgender and gender non-conforming pedigree nomenclature in current genetic counselors' practice: The case for geometric inclusivity. *J. Genet. Couns.* **29**, 1114–1125 (2020).

2. Tuite, A., Dalla Piazza, M., Brandi, K. & Pletcher, B. A. Beyond circles and squares: A commentary on updating pedigree nomenclature to better represent patient diversity. *J. Genet. Couns.* **29**, 435–439 (2020).

variable, or information is incomplete, a pedigree can lead to a mistaken conclusion. Incomplete penetrance of a trait can also make interpretation difficult.

Inferring the Mode of Inheritance

Pedigree analysis can be used to determine the **mode of inheritance** of a particular trait, allowing geneticists to begin to identify a gene (or genes) responsible. The mode of inheritance refers to how a trait is inherited: is it dominant or recessive? Is the gene located on a sex chromosome, an autosome, or the mitochondrial chromosome? We will look at six modes of inheritance in this module: autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive , Y-linked, and mitochondrial inheritance.

We will look for patterns among small nuclear families (parent/child relationships) within the extended family. Certain inheritance patterns allow geneticists to conclusively rule out modes of inheritance, so the remaining modes can be evaluated for likelihood. Each of the six modes is considered below.

Y-linked inheritance

One of the easiest modes of inheritance to identify in pedigrees is Y-linked inheritance. This is because Y-linked traits are present only in chromosomal males (who have a Y chromosome). The traits are passed from father to son, with no exceptions.

An example of a Y-linked pedigree is shown in Figure 3. Assuming complete penetrance, every affected son has an affected father. No females are affected. Y-linked inheritance can be ruled out if chromosomal females have the trait or if fathers and sons do not have the same phenotype.

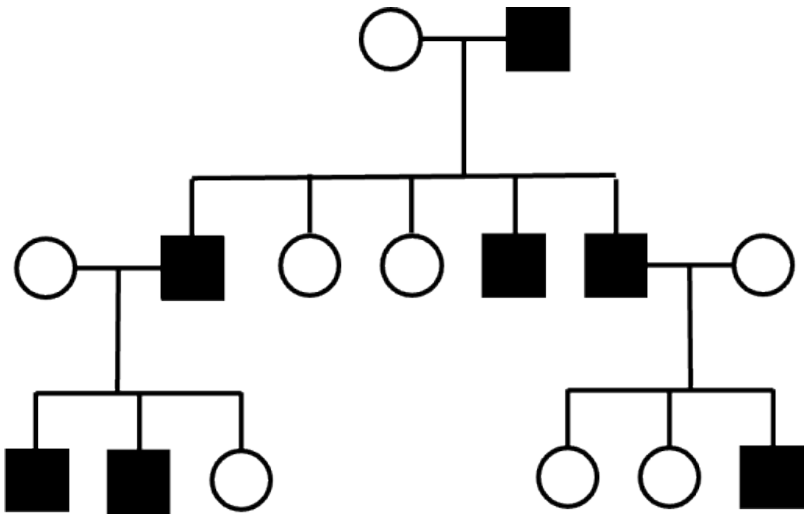


Figure 3. Pedigree consistent with Y-linked inheritance. Traits are passed from father to son, with sons and father always sharing the same phenotype.

Although Y-linked pedigrees are easy to identify, Y-linked pedigrees are quite rare outside of textbooks. The Y chromosome is the smallest of the human nuclear chromosomes, and, by some estimates, the Y chromosome may carry only about 55-70 protein-coding genes. By contrast, the smallest autosome (chromosome 21) likely carries about 200-300 protein-coding genes, and the X chromosome carries about 800 protein-coding genes. The small number of Y-linked genes makes Y-linked traits quite rare. Some regions of the Y chromosome appear to be linked to defects in sperm development, and Y chromosome variants may be associated with coronary artery disease.[3] But evidence linking other traits to the Y chromosome has proven to be insufficient. The most commonly cited example of a rare Y-linked trait is hairy pinnae (the outer rim of the ear). Although pedigree analysis suggested Y-linkage, subsequent analysis seems to show that the trait is an autosomally inherited sex-influenced trait.

Mitochondrial inheritance

Like Y-linked inheritance, mitochondrial inheritance is also quite easy to identify in a pedigree, but rare in the human population. The mitochondrial chromosome contains 37 genes (13 protein-coding). Sperm do not contribute mitochondria to a fertilized zygote: all mitochondria (and thus, mitochondrial DNA) comes from the egg. Because of this, mitochondrial traits are passed from a mother to all of her offspring, both male and female. Assuming complete penetrance, a child with a mitochondrial trait must have a mother with the trait. Fathers can never pass the trait to their sons.

An example of a pedigree consistent with a mitochondrial trait is shown in **Figure 4**. Mitochondrial inheritance can be ruled out if the trait passes from father to child, or if children of an affected mother do not have the trait.

3. Maan, A. A. *et al.* The Y chromosome: a blueprint for men's health? *Eur. J. Hum. Genet.* **25**, 1181–1188 (2017).

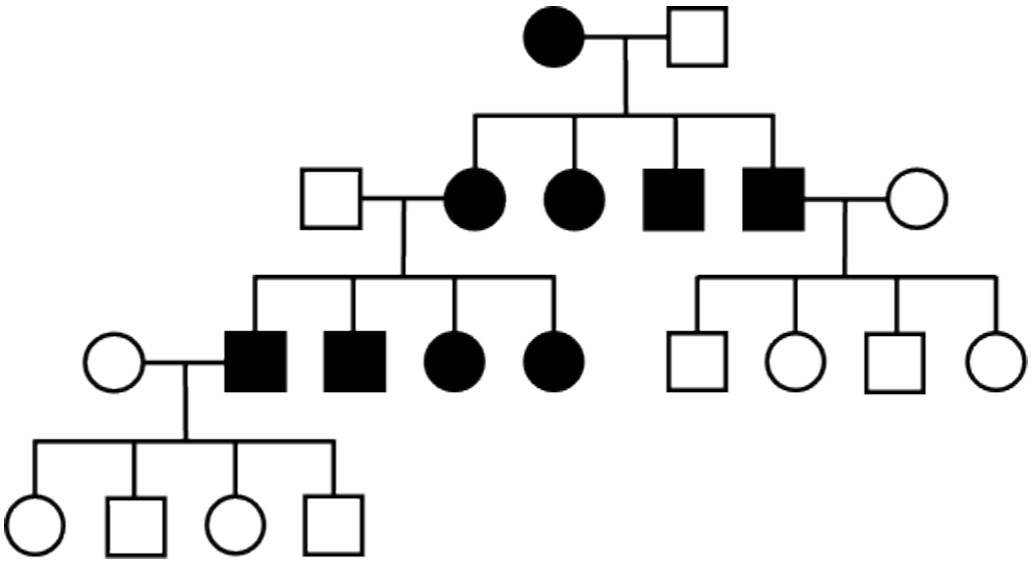


Figure 4. Example of a pedigree showing mitochondrial inheritance. Children always have the same phenotype as their mother.

X-linked dominant inheritance (XD)

As with AD inheritance, all children affected with an XD trait must also have an affected parent. X-linked dominance means that the gene causing the phenotype is located on the X-chromosome. All XY individuals who inherit the allele will show the trait, since they have no second allele to modify the phenotype. All XX individuals who inherit at least one copy of the allele will show the trait as well: a second allele on the second X chromosome will make no difference.

Examples of X-linked dominant inheritance in humans include a form of rickets, in which problems with phosphate absorption lead to bone deformity and soft teeth. A second example is Rett syndrome, which is characterized by a number of neurological challenges including differences in language development and coordination, as well as short stature.

XD traits tend to affect chromosomal females more often than males in a population, since XX females have twice as many X chromosomes as XY males. Although it can be difficult to distinguish between AD and XD pedigrees, often XD pedigrees show more affected females than males. Equal numbers of males and females are not, however, sufficient to rule out XD inheritance. An example of a pedigree that is consistent with X-linked dominant inheritance is shown in **Figure 5**.

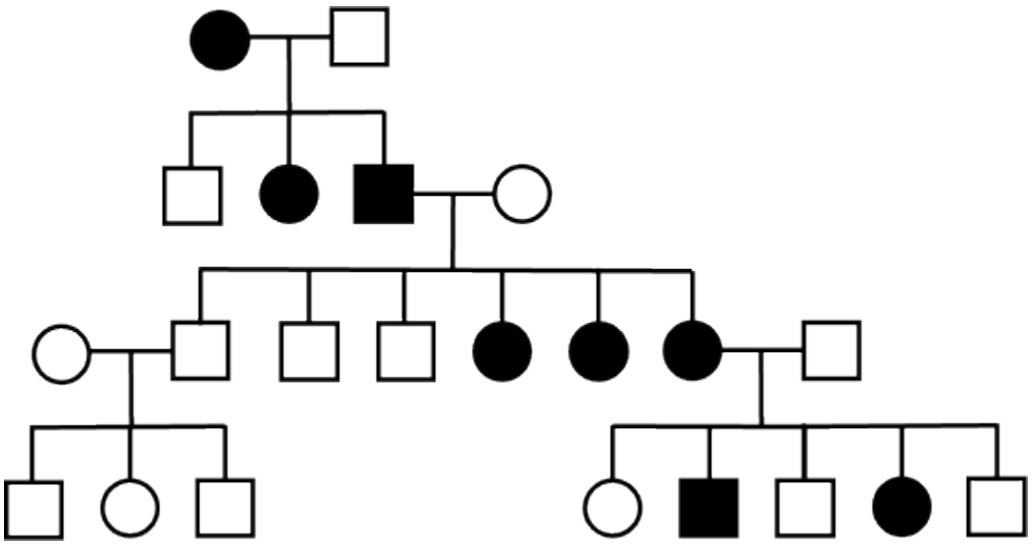


Figure 5. An example of a pedigree suggesting X-linked dominant inheritance. Affected individuals always have an affected parent, and all daughters of affected dads also have the trait.

XD inheritance can be ruled out if two unaffected parents have a child with the trait. Because fathers do not pass X chromosomes to their sons (except in rare cases of a son who inherits an extra sex chromosome), X-linked dominance can be ruled out if a trait is transmitted from father to son. XD inheritance can also be ruled out if an affected father has an unaffected daughter.

Autosomal dominant (AD) inheritance

Assuming complete penetrance, the most characteristic feature of an autosomal dominant pedigree is that every child who has the trait also has an affected parent with the trait. This is because every individual with the allele will display the trait. Any offspring with the trait must get the allele from a parent, so the parent must also show the trait.

Pedigrees tracking autosomal dominant traits show affected individuals in every generation, tracking up through older generations, assuming complete penetrance. However, keep in mind that it is possible that an AD trait vanishes in younger generations: parents who are Aa (affected) and aa (unaffected) can certainly have aa (unaffected) children! An example of a pedigree from an autosomal dominant trait is shown in Figure 6. Note that there are no affected individuals in the fourth (youngest) generation. This kind of pattern is often seen with dominant traits. It means that the affected father, who is heterozygous, did not pass down the affected allele to his children.

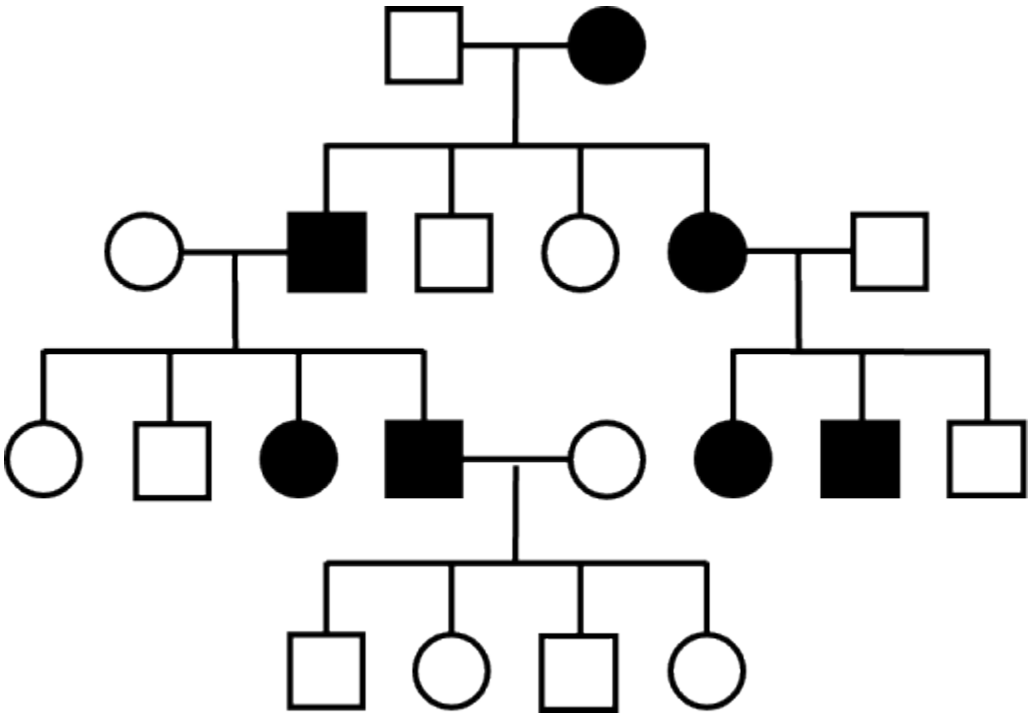


Figure 6. Example of a pedigree suggesting autosomal dominant inheritance. Affected individuals always have an affected parent.

Huntington's disease and polydactyly (extra fingers and toes) are examples of human traits that are inherited in an autosomal dominant fashion. People with either Huntington's disease or polydactyly have a parent affected by the trait as well. But they only have a 50% chance of passing the affect allele to their children, so they may not have children with the trait.

Note: other modes of inheritance can also show a similar pattern with the trait in every generation: for both Y-linked and mitochondria pedigrees, all affected children also have an affected parent.

Autosomal dominant pedigrees usually show roughly equal numbers of males and females affected by the phenotype. But an overabundance of one sex of the other is not sufficient on its own to rule out autosomal dominance. Assuming complete penetrance, autosomal dominant inheritance can be ruled out if an affected individual does not have an affected parent.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=420#h5p-76

X-linked recessive (XR) inheritance

Recall that chromosomal males have only one X chromosome, while chromosomal females have two. Males are **hemizygous** for all X-linked genes. Because of this, if a chromosomal male inherits an X-linked recessive allele, he will show the trait. There is no second copy to dominate. Because of this, X-linked recessive traits often affect more males than females in a population. Examples of XR traits in the human population include red-green colorblindness, some forms of muscular dystrophy, and some forms of immunodeficiency. In all of these example cases, it is more common for boys to be affected by the trait than girls in the general population.

It is very common for a pedigree tracking an X-linked recessive trait to show mostly (or even only) males affected. While it can be tempting to look at such a pedigree and infer a Y-linked mode of inheritance, remember that Y-linked inheritance requires that affected sons have affected fathers. This will not be the case for XR inheritance, since sons inherit the X chromosome from their mothers. A typical XR pedigree is shown in **Figure 7**.

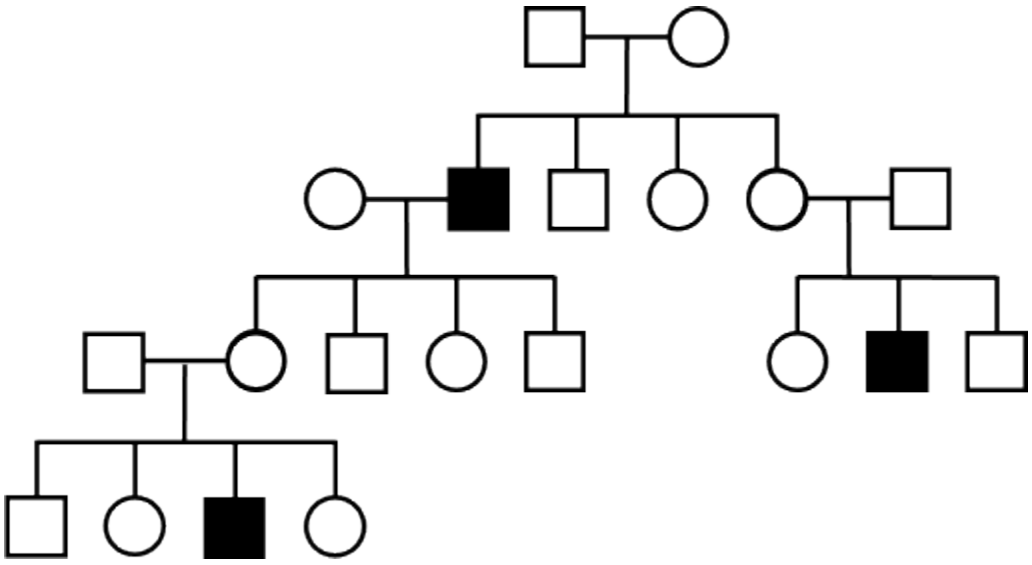


Figure 7. Example of a pedigree that suggests X-linked recessive inheritance.

In contrast, both sexes tend to be affected equally by autosomal recessive and autosomal dominant inherited traits, while females are more frequently affected by XD traits. Distribution of sexes alone is not sufficient to infer a mode of inheritance. Additional information must be extracted from the pedigree.

XR inheritance can be ruled out if we see an affected daughter with an unaffected father. A daughter affected with an X-recessive trait must have genotype X^aX^a, which means her father must have genotype X^aY to pass it down. He must also be affected by the phenotype. X-recessive inheritance can also be ruled out if an affected mother (who must have genotype X^aX^a) has an unaffected son (since it is not possible for such a son to inherit genotype X^AY from his parents).

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=420#h5p-77

Autosomal recessive (AR)

For diploid organisms, autosomal recessive traits are only expressed if an individual receives recessive alleles from both parents. An **unaffected carrier** of a trait (sometimes just called a carrier) is a person who has the recessive allele but does not show the trait. Unaffected carriers are heterozygous, but they can pass on the recessive allele to their offspring.

A pedigree of a rare AR trait often shows fewer affected individuals, and the trait may "skip" a generation, meaning that an affected child does not have an affected parent.

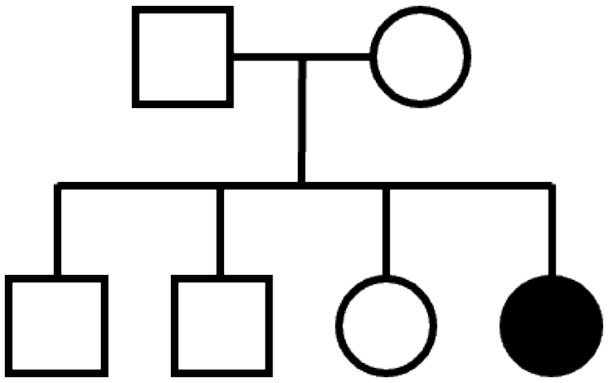


Figure 8. An example of an autosomal recessive pedigree.

An example of an AR pedigree is shown in Figure 8. As is typical of a rare AR trait, most affected individuals do not have an affected parent. However, this is not a requirement for AR inheritance: note that it's possible for affected children to also have an affected parent.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=420#h5p-78

It can be difficult to rule out AR as a mode of inheritance. For example, even if a trait appears in every generation of a pedigree – all affected children having an affected parent – AR cannot be ruled out. It just would require multiple individuals, through several generations, all carrying the same allele associated with the trait. This is illustrated in **Figure 9**. For a rare allele in the general population, this is unlikely, but not impossible.

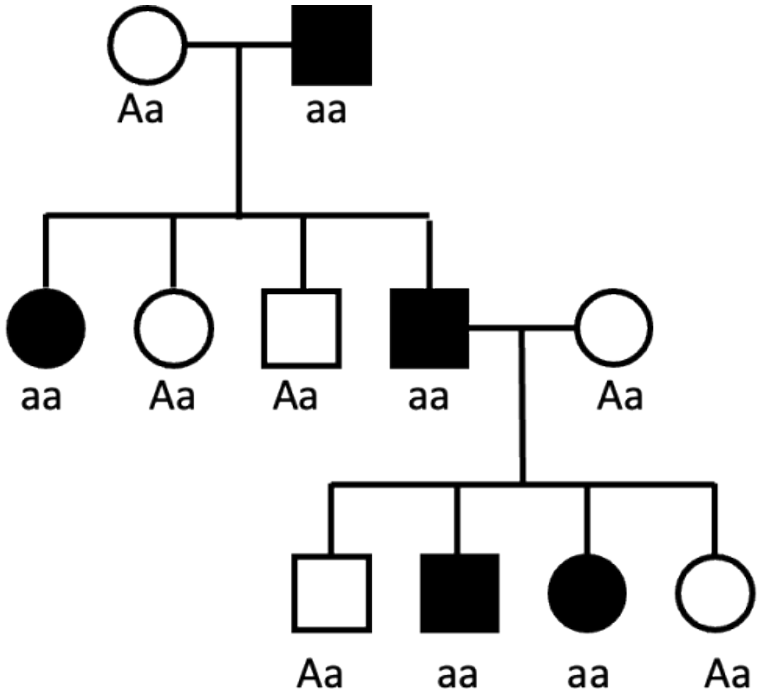


Figure 9. Example of a pedigree where an autosomal recessive trait appears in every generation.

However, if two individuals with the trait (genotype aa) have children together, it is expected that all of their children will have the trait as well (because only genotype aa is possible in their offspring). If two affected individuals have an unaffected child, this rules out AR as a possible mode of inheritance.

Possible vs most likely

The examples given above allow you to rule out modes of inheritance. In many cases, though, the small amount of data provided in a pedigree will leave you with multiple possible modes of inheritance. In those cases, it's necessary to determine which is the *most likely* mode of the remaining options.

After ruling out all impossible modes, if fathers always match sons and no females are affected, it is likely to be Y-linked. It would be very unusual to see such a pattern with other modes of inheritance: autosomal traits would be expected to show even numbers of male and female. Thus, if fathers always match sons, and no females are affected, Y-linked is the most likely mode of inheritance.

After ruling out all impossible modes, if there is a great discrepancy between the number of affected females and males, it is most likely that the mode is sex-linked rather than autosomal. This is because we would expect a roughly even distribution of affected males vs affected females for a sex-linked trait.

After ruling out all impossible modes, if all affected children have an affected parent, it is most likely that the mode is dominant rather than recessive. This is because it would be relatively uncommon for multiple unrelated individuals with the same rare recessive allele to join a family, which is necessary for the parents of every affected individual. The exception is if the pedigree involves a consanguineous mating, or a mating between individuals who are blood relations.

Rare recessive disorders are common in families with consanguineous matings. This is due to the fact that almost everyone carries one or more rare recessive alleles. But the phenotype does not manifest in offspring unless a mate also carries the same rare recessive alleles. For the most part, it is very uncommon for partners to share the same rare allele, so the phenotypes remain rare in the general population. However, consanguineous partners are far more likely to share the same rare alleles – they've inherited them from the same ancestors. An example of this is seen in **Figure 10**.

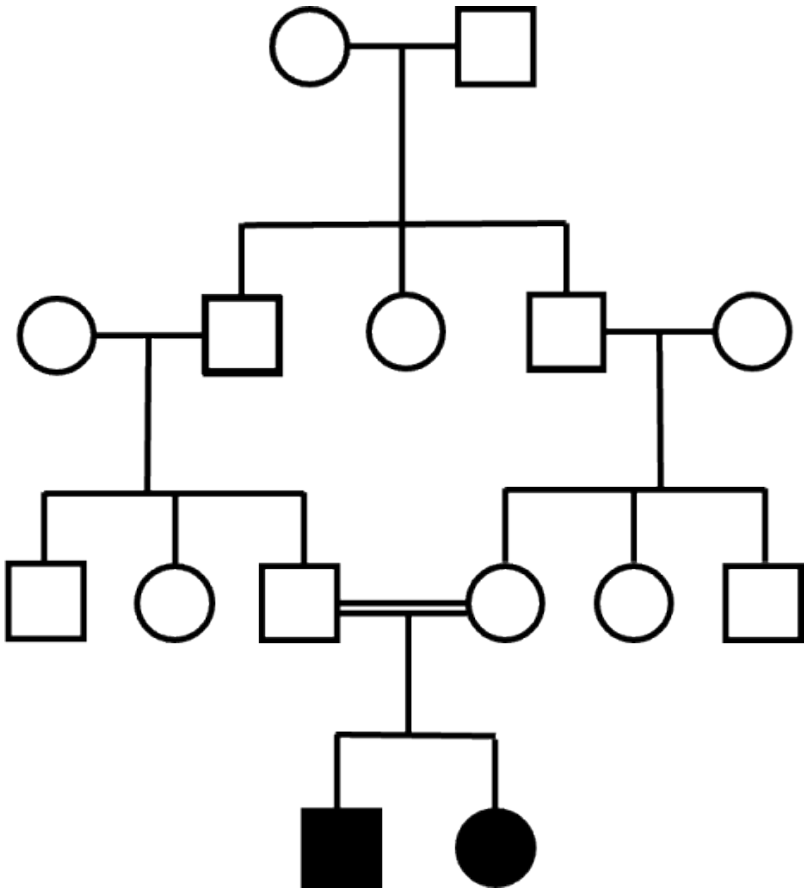


Figure 10. Consanguineous families often have a high incidence of autosomal recessive traits.

In human populations (and in other species), distinctive genetic traits may be observed in populations where consanguineous relationships are common. A classic example, cited in many genetics textbooks, is the Habsburg jaw, a distinctive protrusion of the lower jaw. It was so-named from its frequency among the Habsburg royal family in Europe, and likely persisted among family members due to interbreeding over many generations.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=420#h5p-79

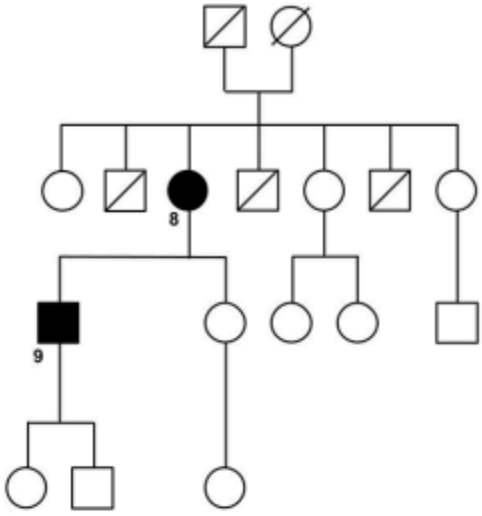
Real pedigrees are usually more complicated

In this module, straightforward examples of pedigrees are given, assuming complete penetrance of a trait caused by a single gene. In "real-life" practice, however, determining the mode of inheritance is complicated by traits that are incompletely penetrant, variably expressive, or multifactorial. Incomplete penetrance can make a dominant trait appear to skip a generation, variable expressivity can complicate identification of a phenotype, and multifactorial traits can be attenuated by other genes or non-genetic factors. If members of a family are deceased, it can also be difficult to identify whether they had been affected by a trait, given that the identification may rely on the memory of family members.

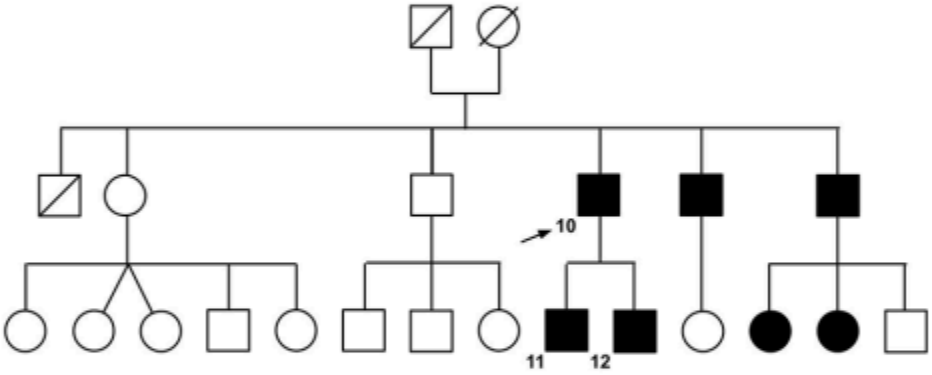
Figure 11 shows examples of pedigrees adapted from an early paper tracking celiac disease in families[4]. The authors of the paper, David and Ajdukiewicz, tracked 13 families total: 4 pedigrees are reproduced here. This paper draws pedigrees using an alternative convention where unrelated partners are not shown in the pedigree. We can presume that they do not show signs of celiac disease.

4. David, T. J. & Ajdukiewicz, A. B. A family study of coeliac disease. *J. Med. Genet.* **12**, 79–82 (1975).

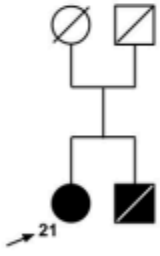
Celiac C



Celiac D



Celiac I



Celiac J

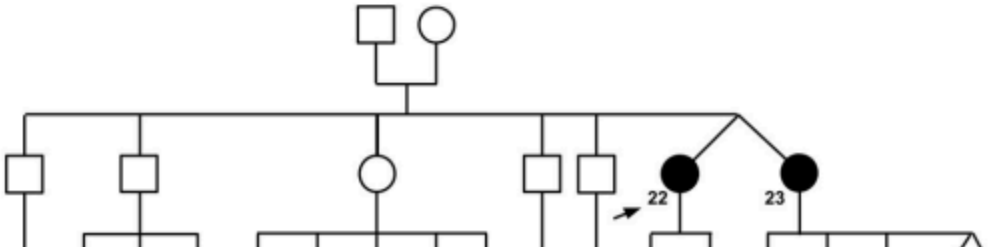


Figure 11 Pedigrees adapted from David and Ajdukiewicz, 1975, tracking celiac disease in several unrelated families.

Figure 11 Pedigrees from David and Ajdukiewicz, 1975, tracking celiac disease in several unrelated families.

In the full paper, of the 36 children born to a parent with celiac disease, 14 also have the disease (~39%). For a dominant trait, we would expect about 50% of offspring to also show the trait. However, none of the oldest generation have the trait in any of these four pedigrees. Because a dominant trait would normally require affected children to have an affected parent, one possible conclusion is that this is an incompletely penetrant trait. Notably, in pedigree J, the only affected individuals are twins (presumed in the paper to be identical twins).

Because of the number of pedigrees that show similar patterns – only one parent has a family history of the relatively rare disease – David and Ajdukiewicz ruled out a simple autosomal recessive pattern of inheritance, since it would require so many unrelated individuals to carry a rare allele associated with celiac disease.

The authors of the paper stated that the pedigrees matched what would be expected from an incompletely penetrant, dominant condition, but described their contemporary (1975) understanding of celiac disease as heterogeneous. This still holds true at present, with an imperfect understanding of the inheritance of a multifactorial disease that's influenced by both genetic and non-genetic factors.

Calculating Probabilities from Pedigrees

Pedigree analysis can also be used to calculate risk. If the mode of inheritance of a trait is known, we can use information about others in the family to calculate the likelihood that another individual will develop the trait. This is useful in situations like genetic counseling: if a couple comes to a genetic counselor due to a family history of a genetic disorder, what is the risk that their child will also suffer from the disorder?

If the mode of inheritance is known, it's often possible to assign probable genotypes to some individuals in the pedigree, based on their phenotypes and relationships to others in the pedigree. For example, all individuals affected by an autosomal recessive trait have genotype "aa", and any of their offspring who are unaffected by the trait must have genotype "Aa". With this information, it's then possible to calculate the probability of other individuals either having the allele (being unaffected carriers) or having kids with the trait. This makes pedigrees an important tool in genetic counseling if, for example, parents with a family history of a genetic disease would like to know the likelihood of passing the disease to their child.

The rules of probability – and the laws of Mendelian inheritance – make these calculations possible. Remember from earlier modules, we used two rules of probability: the **multiplication rule of probability** and the **addition rule of probability**. The multiplication rule of probability states that the probability of two independent events occurring is the product of the probability of each event occurring independently. For

example, in a cross between parents of genotypes Aa and Aa, the probability of having a child with phenotype A *and* a child with phenotype a is ¾ * ¼ = 3/8.

The addition rule of probability states that the probability of one event *or* another is the sum of their individual probabilities. In a cross between parents of genotypes Aa and Aa, the probability of having a child of genotype AA *or* Aa is ¼ + ½ = ¾.

In many of these complex family pedigrees, in order for a child to have a particular trait, the allele for the trait must be passed down from multiple individuals, often through several generations. But because all of these inheritance events must happen, we use the multiplication rule to calculate the combined probability.

When calculating the probabilities for a rare trait in the general population, unless there is evidence otherwise in the pedigree, we usually assume that unrelated individuals who are joining the family do not carry the same rare allele that "blood" relatives do, since this would be a pretty unlikely occurrence.

An example of this is shown using the pedigree shown in Figure 12. What is the probability that a child of III-1 and III-2 will be affected by the autosomal recessive trait?

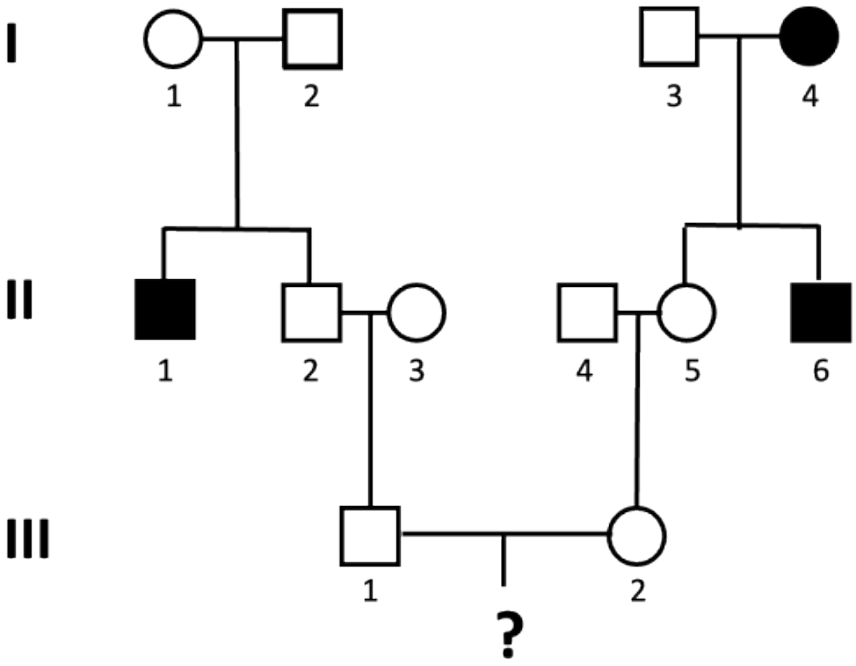
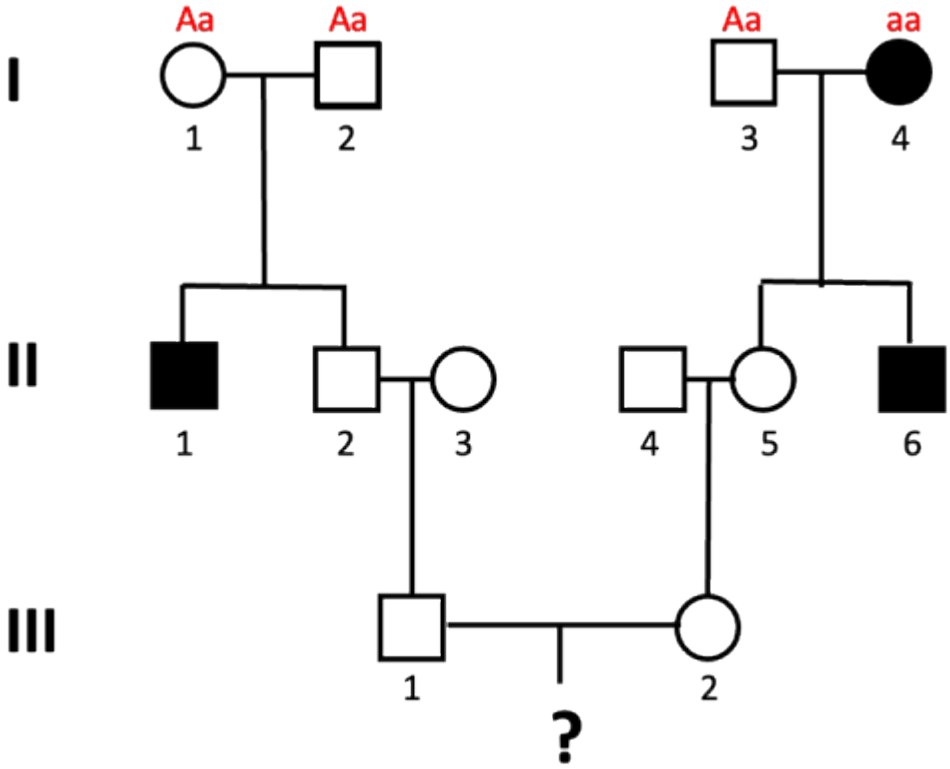


Figure 12. Pedigree tracking an autosomal recessive trait

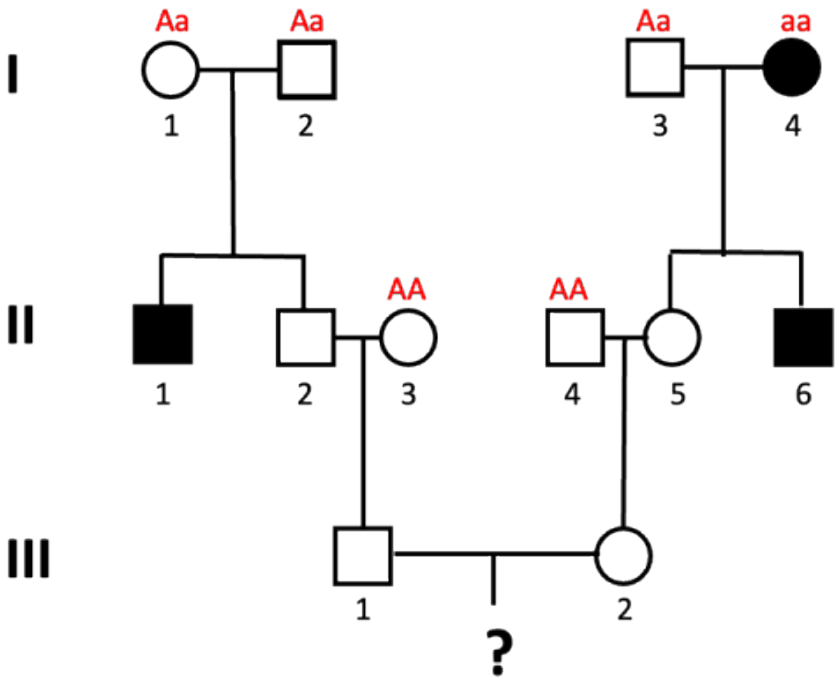
We solve this type of problem by determining the genotypes of the direct ancestors to the individual for whom we will calculate the probability. The next figures walk through this process, step by step. Let's assume the allele associated with the recessive trait is "a".

1. We can assign genotypes to all of the individuals in generation I.
   Individuals I-1 and I-2 must both have genotype Aa, since they have an affected child (II-1) who is

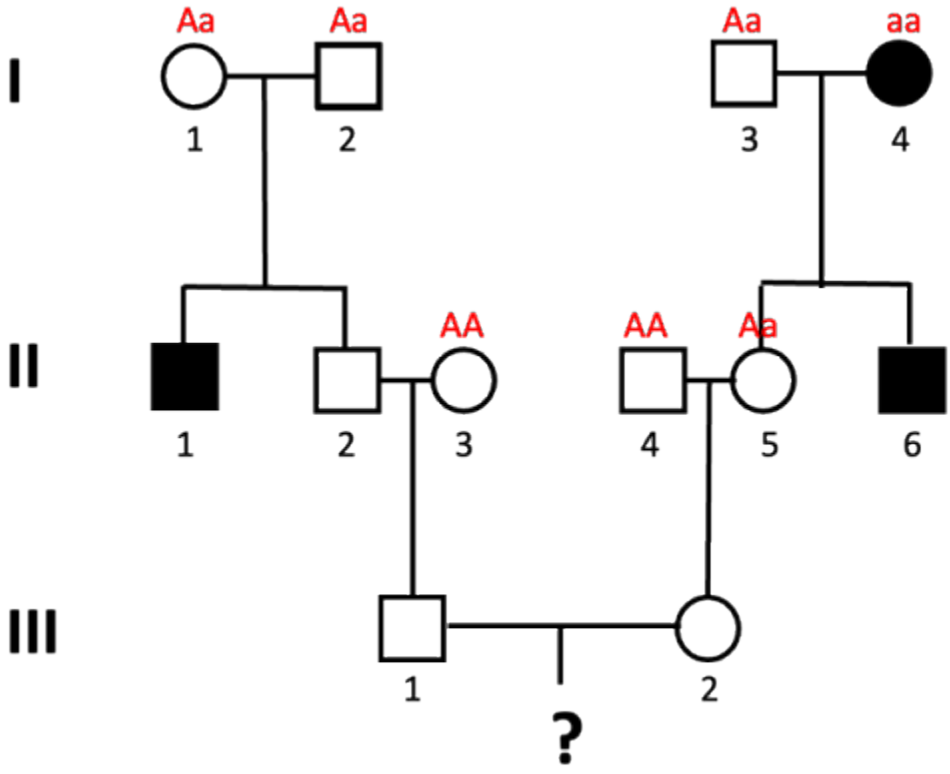
presumed to have genotype aa. This is also true for individual I-3, who has a child with the trait as well. I-4 has the genotype aa, since they have the trait.



2. Since this is a rare trait in the population, in Generation II we assume that individuals II-3 and II-4 do not carry the allele (they have genotype AA).

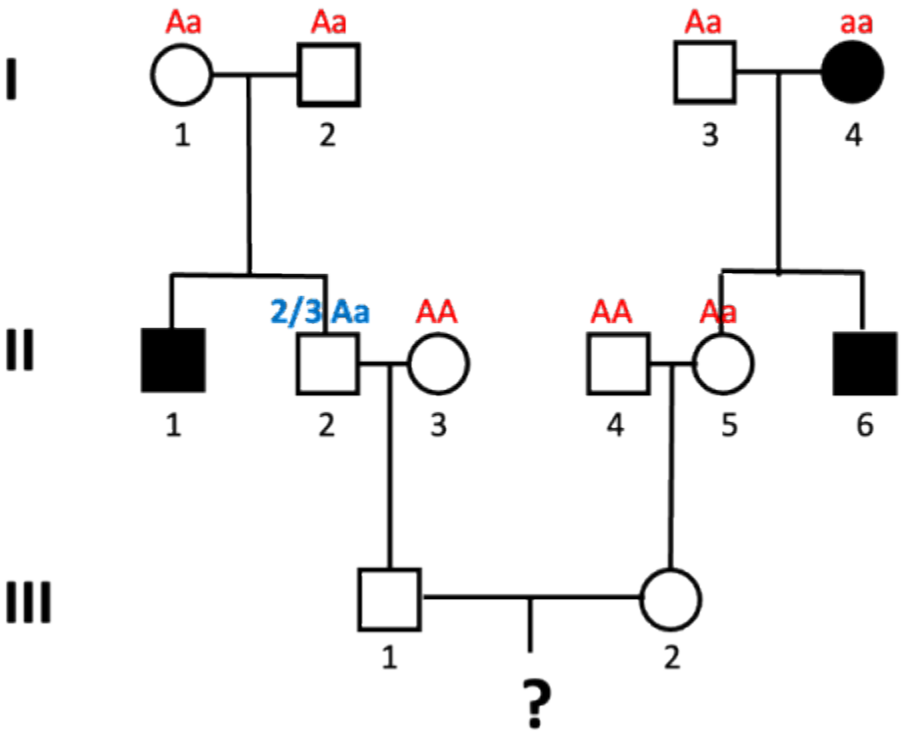


3. Individual II-5 must have genotype Aa, since she inherits a dominant (unaffected) allele from her dad, but must inherit the recessive allele from mom.



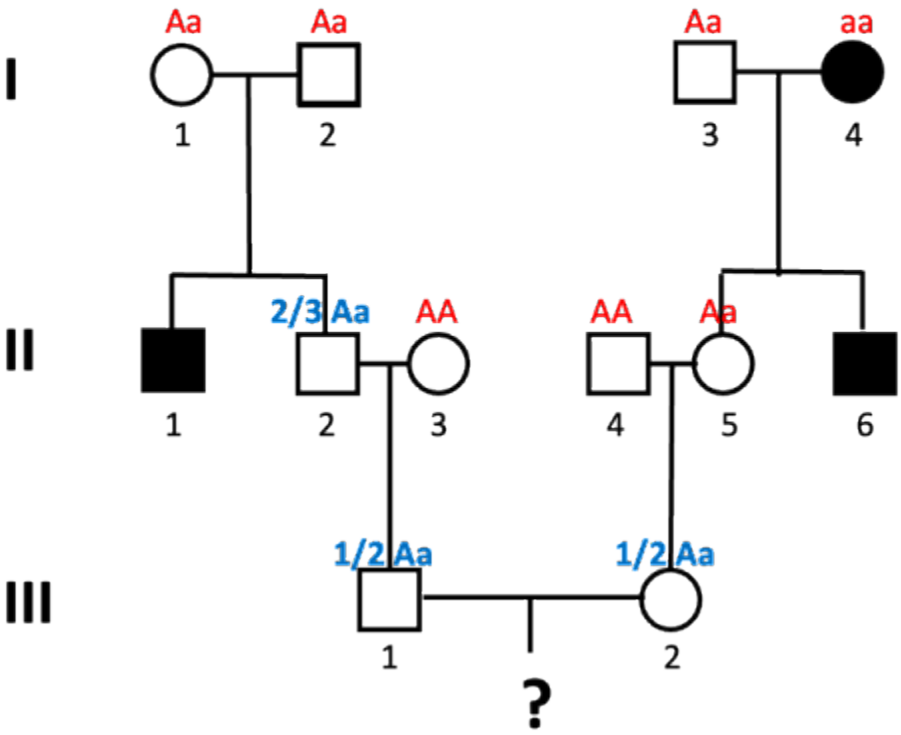
4. What about individual II-2? In order for our final offspring to show the trait, individual II-2 must be a carrier of the "a" allele.
   What's the probability that they inherited the "a" allele? Well, let's look at the Punnett Square expected from Aa x Aa parents (which is what individual II-2 has). This requires a bit of tricky reasoning: it might be tempting to say that ½ are Aa, but that is incorrect. We know that II-2 does not have an aa genotype because II-2 does not show the trait. Of the remaining possibilities, 2/3 are Aa.

	A	a
A	AA	Aa
a	Aa	aa

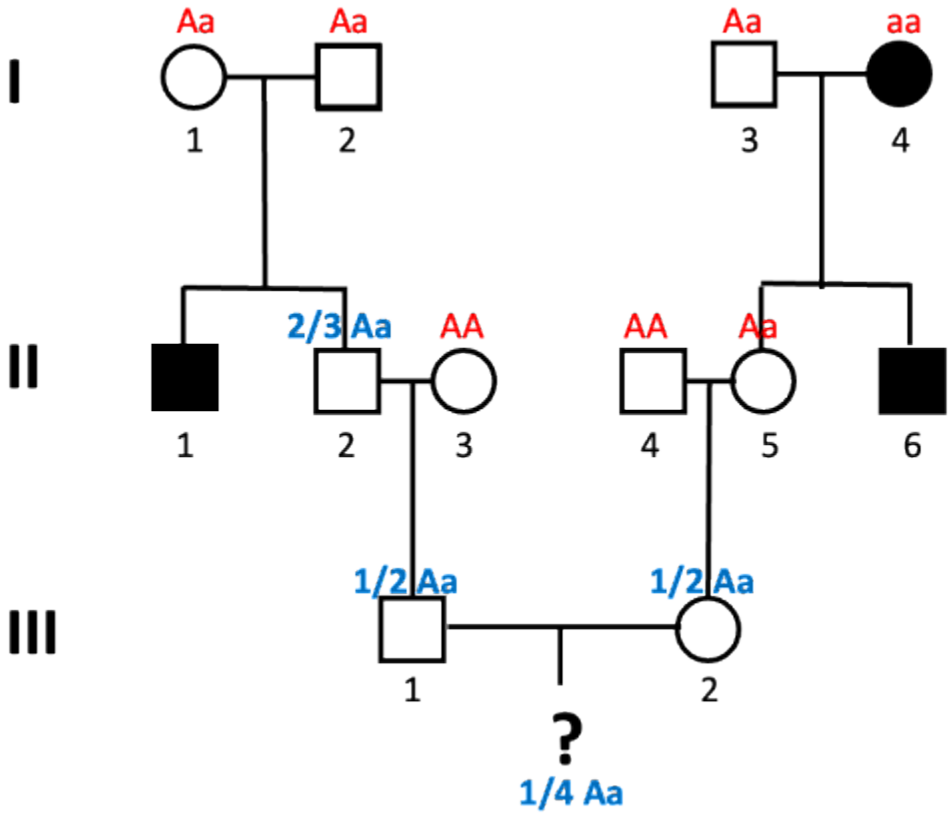


5. In generation III, we likewise need to determine the probability of III-1 and III-2 carrying the allele. We do this as we did for II-2: we draw a Punnett square to illustrate the cross, and determine which fraction of offspring carry the allele. For both III-1 and III-2, there is a ½ probability that they will carry the allele (and have the Aa genotype).

	A	A
A	AA	AA
a	Aa	Aa



6. Lastly, we can't forget the child! A Punnett square of Aa x Aa shows that there is a ¼ chance of aa offspring.



In order for the child of III-1 and III-2 to show the trait, all these four things shown in blue in the images must be true: II-2 must be Aa, III-1 must be Aa, III-2 must be Aa, *and* the child must be aa. We use the multiplication rule to determine the overall probability: 2/3 * 1/2 *1/2 * 1/4 = 2/48.

Some things to keep an eye out for, as you solve these problems: it's very common to forget the last step: the probability of the final offspring showing the trait! Don't forget to do this, even if there isn't a symbol representing the unborn child. And watch out for those Aa x Aa crosses, where you can rule out aa as a possible genotype. 2/3 of the possible unaffected offspring are Aa, not ½.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=420#h5p-80

Media Attributions

- Pedigree examples © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Pedigree symbols © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

PART XV
LINKAGE AND MAPPING

Objectives

After completing this module, students will be able to:

1. Define: linked vs unlinked genes, parental vs. recombinant classes of offspring, locus (plural loci), morph, polymorphism, haplotype
2. Calculate recombination frequencies from offspring ratios and construct a genetic map using a two- or three-point testcross
3. Use a pedigree to infer linkage
4. Explain how tools of molecular genetics can be used in mapping
5. Explain how genome association studies can be used to identify loci associated with a phenotype

Source Material

- Selected images and text modified from Open Genetics Lectures[1]. Chapter 37: DNA variation studied with microarrays
- Selected images and text remixed from Online Open Genetics (Nickle and Barrette-Ng), available through Biology LibreTexts[2]. Chapter 7: Linkage and mapping and Chapter 10: Molecular Markers and Quantitative Traits

1. Locke, J. 'Open Genetics Lectures' textbook for an Introduction to Molecular Genetics and Heredity (BIOL207). Borealis https://doi.org/10.7939/DVN/XMUPO6 (2017).

Introduction

Figure 1. A person performing the ice bucket challenge

In the summer of 2014, the Ice Bucket Challenge (**Figure 1**) swept across the United States before spreading worldwide. Participants posted short videos of themselves pouring a bucket of ice water over their heads, challenging others to either do the same or donate to ALS research[3].

ALS stands for amyotrophic lateral sclerosis. It is also known as motor neuron disease and Lou Gehrig's Disease, after the famous baseball player who developed the disease in 1939. ALS is a progressive neurodegenerative disease characterized by the loss of motor neurons. Motor neurons control the function of muscles, so ALS patients gradually lose muscle function. As the disease progresses, patients become paralyzed,

2. Nickle and Barrette-Ng. Open Online Genetics. in *Open Online Genetics* (2016).

3. Ice bucket challenge: What's happened since? *BBC News* (2015).

without the ability to move, speak, or eat. In the final stages of the disease, patients lose their ability to breathe. Respiratory failure is the most common cause of death for ALS patients.

There is no cure for ALS: treatment can slow but not stop progression of the disease. Death usually comes within 2-5 years after initial diagnosis, although about 10% of ALS patients live longer than 10 years after diagnosis. The theoretical physicist Stephen Hawking (shown with his daughter, Lucy Hawking, in **Figure 2)** lived for more than 5 decades with a slow-progressing form of ALS.

Figure 2. Theoretical physicist Stephen Hawking and his daughter Lucy.

Most cases of ALS are sporadic, meaning that a newly diagnosed patient most likely has no family history or risk factors for the disease. But about 10% of ALS cases are **familial**: more than one member of the family can be affected by the disease. This indicates a genetic component to the disease. One important step in ALS research is therefore identifying causative genes: understanding the normal gene function and how variants cause disease can eventually lead to targeted treatments.

The Ice Bucket Challenge resulted in a windfall of about $115 million dollars to the ALS Association in the United States in the first six weeks after the challenge went viral. Similar organizations worldwide also benefitted. The ALS Association used much of these funds to support patients and their families and fund research. In 2015, the ALS Association awarded nearly four times the amount of research funds as in 2014. Within 18 months, funded researchers had identified a gene linked with susceptibility to ALS[4]. By 2019, five

4. Kenna, K. P. *et al.* NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1037–1042 (2016).

genes had been identified by researchers funded at least in part with Ice Bucket Challenge dollars, with at least one of these genes a target for a therapy in development.

So how are genes associated with a trait identified? Mendelian inheritance treats genes as an abstract concept – and, in fact, Mendel did not know that the patterns of heredity he described were a result of inheritance of physical chromosomes. However, today we understand genes to be information housed on molecules of DNA. The equal segregation and independent assortment that Mendel observed is due to the behavior of the physical chromosomes during meiosis.

This means that phenotypes that we study using genetic crosses and pedigrees can be **mapped** to specific parts of a chromosome, where we can analyze differences in DNA sequence to predict changes in function. We use the word "mapping" to describe the identification of parts of a chromosome that are associated with specific traits, and the word **locus** to describe the specific location on a chromosome.

We also use the work **linkage** to describe the relationship of a trait to a chromosome and the relationship of genes near to one another on a chromosome. You've seen this term used when we describe sex-linked genes – a sex-linked gene has been mapped to a sex chromosome. Linked genes are located close together on a chromosome.

Mapping a trait to a chromosome historically involved classical genetic approaches, including controlled crosses and tracking traits through a pedigree. The beginning of this chapter describes the classical approach to gene mapping from a historical perspective although those techniques have largely been supplanted by tools of molecular genetics.

Since the early 2000s, when genomic sequence data became widely available, genomic comparisons called genetic association studies have more commonly been used to map phenotypes to particular parts of the genome. Genetic association studies can also be used to identify regions of the genome that influence a multifactorial trait.

This module begins with a discussion of landmark studies that allowed researchers to link phenotypes with specific chromosomes using two- and three-point test crosses. It continues with examples of how tools of molecular genetics are used in mapping and ends with modern examples of mapping and a brief discussion of genome-wide association studies.

Media Attributions

NOT ALL GENES INDEPENDENTLY ASSORT

In the module on Basic Heredity, we discussed Mendel's Second Law: Independent Assortment. You'll recall that Mendel observed that the heredity of separate traits (seed shape vs seed color, for example) was independent of one another.

But Mendel got a bit lucky: the seven traits he chose are on separate chromosomes. Some genes, though, located close together on the same chromosome, do not independently assort. These genes are said to be **linked**, and linked alleles tend to be inherited together during meiosis. This phenomenon gives geneticists a tool to map genes to particular **loci** on a chromosome (the singular of loci is locus, meaning location).

Remember: during meiosis 1, homologous chromosomes pair along the metaphase plate of the cell. The pairing of maternal vs paternal chromosomes is random, contributing to the independent assortment of genes on different chromosomes (**Figure 3**). In addition, crossing over occurs between homologous chromosomes, resulting in **recombinant** chromosomes, or chromosomes that are a combination of both maternal and paternal sequences. This recombination allows the independent assortment of genes that are on the same chromosome, but very far apart.

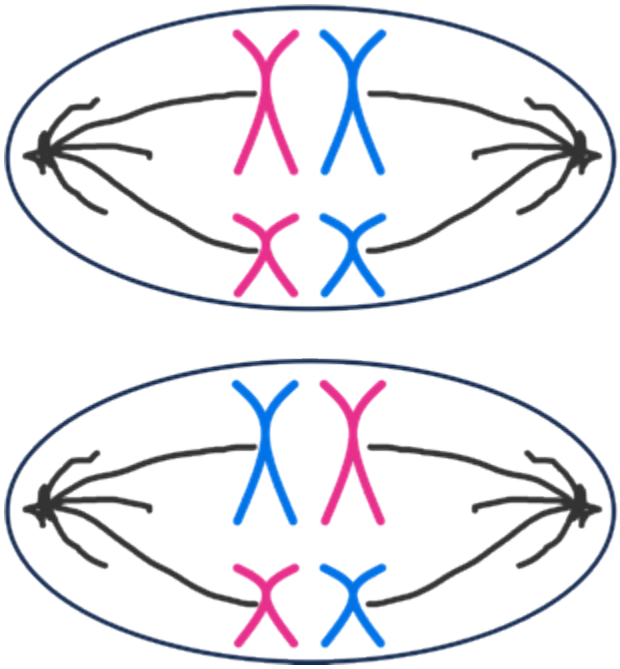


Figure 3. During metaphase I, the random arrangement of the homologous pairs results in the independent assortment of genes on separate chromosomes. In the top meiotic cell, both pink chromosomes will be inherited together, as will the blue chromosomes. On the bottom, each daughter cell will inherit one blue and one pink chromosome.

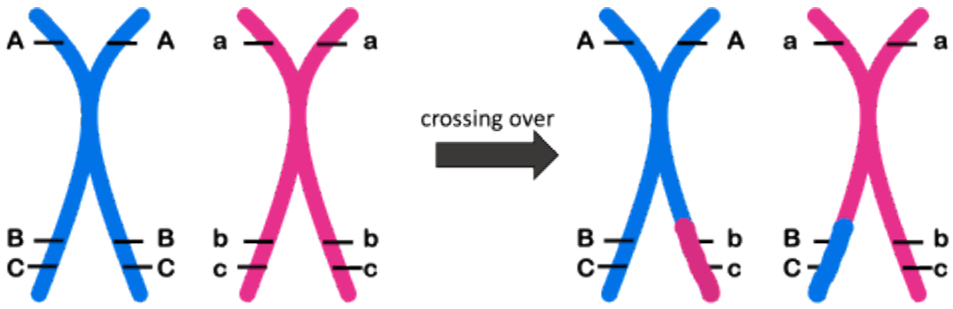


Figure 4. Crossing over between homologous pairs contributes to the independent assortment of genes far apart on the same chromosome, like Genes A and B in this image. Genes B and C are very close together, and so a crossing over event will happen only infrequently between genes B and C. Genes B and C are linked and do not independently assort.

Linked genes are readily distinguished from independently assorting genes in a controlled cross. In the unit on Basic Heredity, we observed that a dihybrid testcross (AaBb x aabb) should always give a 1:1:1:1 ratio of offspring with phenotypes AB, ab, Ab, and aB. The gametes contributing to that ratio are shown in **Figure**

5. This is assuming independent assortment. In **Figure 5**, the genotypes Ab and aB are said to be parental, because these are the exact chromosomes the parental cells contributed to the dihybrid. The genotypes AB and ab are recombinant, since the original parental cells did not have that combination of alleles.

But if the genes are linked, two classes of offspring will be overrepresented, and two classes of offspring will be underrepresented. An extreme example of that is shown in **Figure 6**. In **Figure 6**, there is no recombination between genes A and B. Thus, the **parental** chromosomes Ab and aB are inherited intact. AB and ab genotypes are recombinant chromosomes, but recombinant chromosomes can occur only rarely if two genes are very close together (and not at all in **Figure 6**)

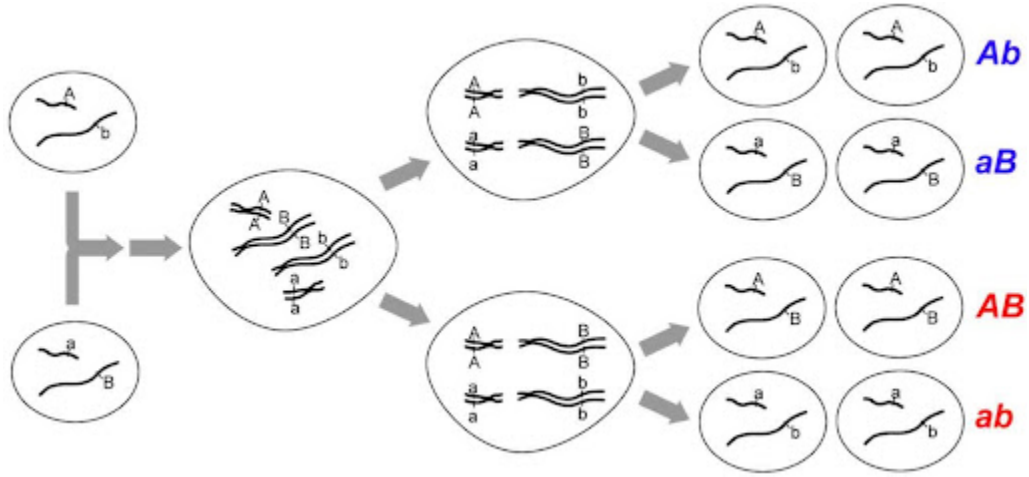


Figure 5. In this image, a dihybrid inherited alleles A and b from one parent and a and B from another. Those alleles are randomly redistributed when the dihybrid produces gametes, as shown on the right of the image.

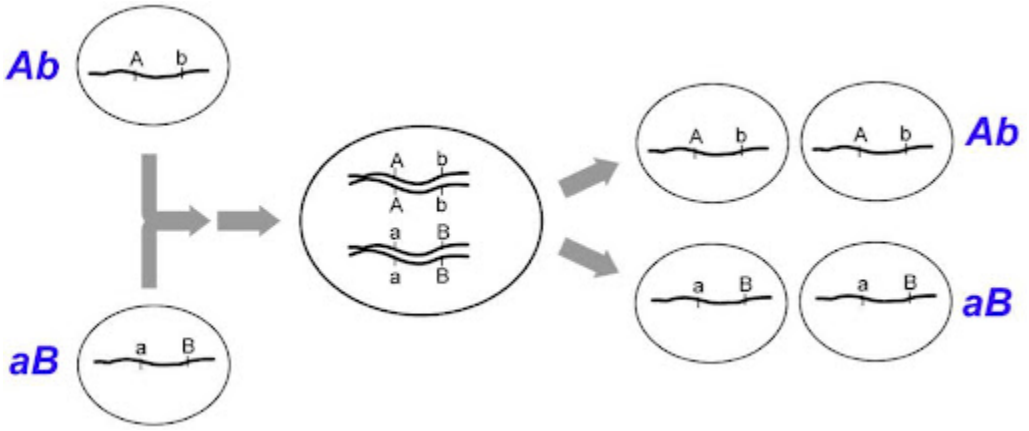


Figure 6 Illustration showing no independent assortment of genes on the same chromosomes during the production of gametes. Alleles of genes A and B are inherited together because they are linked on the same chromosome.

Because the rate of recombination depends on how far apart two genes are, the **recombination frequency** was historically used to estimate map distances on a chromosome. Let's take a look at some examples.

Media Attributions

- Metaphase I © Amanda Simons is licensed under a CC0 (Creative Commons Zero) license
- Crossing over between homologous pairs © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Dihybrid inherited alleles © Nickle and Barrette-Ng via. Online Open Genetics is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Genes on the same chromosomes © Nickle and Barrette-Ng via. Online Open Genetics is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license

THOMAS HUNT MORGAN AND ALFRED STURTEVANT: CHROMOSOME MAPPING VIA LINKAGE ANALYSIS

In the early 1900's, Mendel's work in pea plants was rediscovered, and several researchers began testing the laws of Equal Segregation and Independent Assortment in other species. Very quickly, exceptions to the rule of Independent Assortment were found in multiple species. William Bateson and Reginald Punnett (yes, *that* Punnett) observed that in the sweet pea plant, a plant heterozygous for red/blue flower color and round/long pollen shape would produce red, blue, round, and long offspring. But the blues mostly had long pollen, and the reds mostly had round pollen. This matched the combination of traits seen in the parents of the heterozygote[1]. Bateson and Punnett described the coinciding traits as "coupling" and the traits that hardly ever coincided as "in repulsion".

More examples came from the work of Thomas Hunt Morgan, who observed that sex phenotypes (maleness and femaleness) were linked with other traits not associated with sex (in the fruit fly *Drosophila melanogaster,* white eyes, yellow body, and miniature wings were all associated with sex)[2]. Because those traits all coupled with each other and with the sex factor for femaleness, and because the sex chromosome was visible microscopically and known to couple with sex, Morgan hypothesized that all three genetic factors were coupled on the sex chromosome and that the rare examples of uncoupling were a result of crossing over, which was also visible microscopically. His drawings of this[3] are illustrated in **Figure 7a**.

1. Bateson, William, Punnett, Reginald C. On the inter-relations of genetic factors. *Proc. R. Soc. B* **84**, (1911).

2. Morgan, T. H., Sturtevant, A. H., Muller, H. J. & Bridges, C. B. *The Mechanism of Mendelian Heredity*. (H. Holt and Company, 1923).

3. ibid

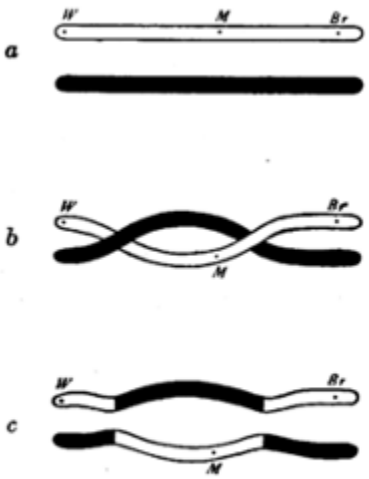


FIG. 25.—Diagram to illustrate double crossing over. The white and the black rods (*a*) twist and cross at two points. Where they cross they are represented as uniting (shown in *c*). That an interchange of pieces has taken place between *W* and *Br* is demonstrated by the factor *M* having gone over to the other chromosome.

Figure 7a. An image of a double crossover from TH Morgan's The Mechanism of Mendelian Heredity

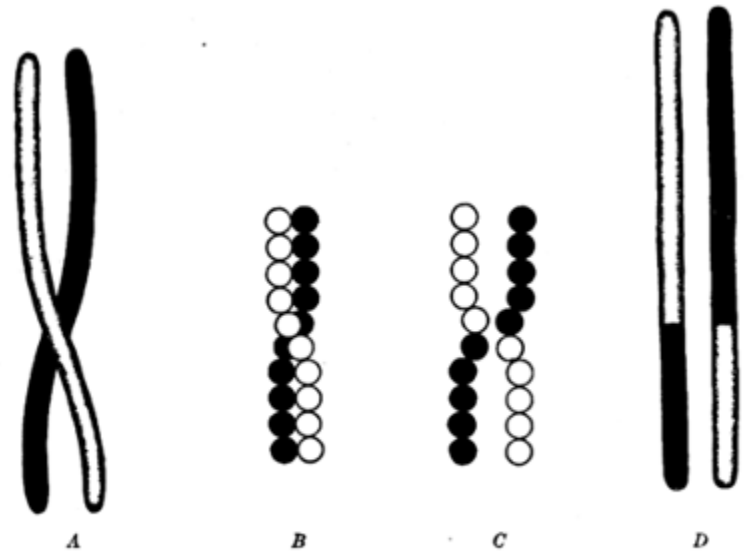


FIG. 24.—Diagram to represent crossing over. At the level where the black and the white rod cross in *A*, they fuse and unite as shown in *D*. The details of the crossing over are shown in *B* and *C*.

Figure 7b. Morgan's 1923 diagrams of crossing over, and an explanation of how multiple crossover events could result in an internal patch of exchanged genetic information.

Morgan further hypothesized that the distance between genes could be estimated by the frequency of uncoupled traits since crossing over was more likely to happen if the genes were farther apart. Morgan's

undergraduate student, Alfred Sturtevant, set out to use the frequency of crossovers (what we'd now call **recombination** frequency) to draw a map of the *Drosophila* X chromosome[4]. He accomplished this through a series of genetic test crosses like the one illustrated in **Figure 8**. Morgan and Sturtevant later went on to map additional chromosomes as well.

Media Attributions

- Internal patch of exchanged © The Mechanism of Mendelian Heredity, by TH Morgan, via Google Books is licensed under a Public Domain license
- Diagrams of crossing over © The Mechanism of Mendelian Heredity, by TH Morgan, via Google Books is licensed under a Public Domain license

4. Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J. Exp. Zool.* (1913).

CALCULATING MAP DISTANCES WITH A DIHYBRID TESTCROSS

Morgan's mapping experiments typically began with a cross between true-breeding individuals with two different mutant traits, like that shown in **Figure 8**.

In **Figure 8**, an *AAbb* parent meiocyte produces Ab gametes, an *aaBB* parent produces *aB* gametes, and the F1 offspring is dihybrid. The dihybrid can potentially produce 4 different gametes: *Ab, aB, AB* and *ab*. But if genes A and B are on the same chromosome as shown in the figure, the **recombinant** AB and ab gametes shown in red will only be produced if crossing over occurs between the two genes. [Note: Figure 8 shows a cross that begins with AAbb x aaBB, but the same principles apply with a cross between AABB x aabb to generate a dihybrid F1. In that case, Ab and aB would be recombinant.]

If the F1 offspring are subsequently testcrossed to a homozygous recessive individual with genotype aabb, the phenotype of the F2 offspring will be determined by the gametes of the dihybrid. So in this way, the phenotype of the F2 offspring can be used to infer the frequency of crossover events, or the frequency of **recombination.**

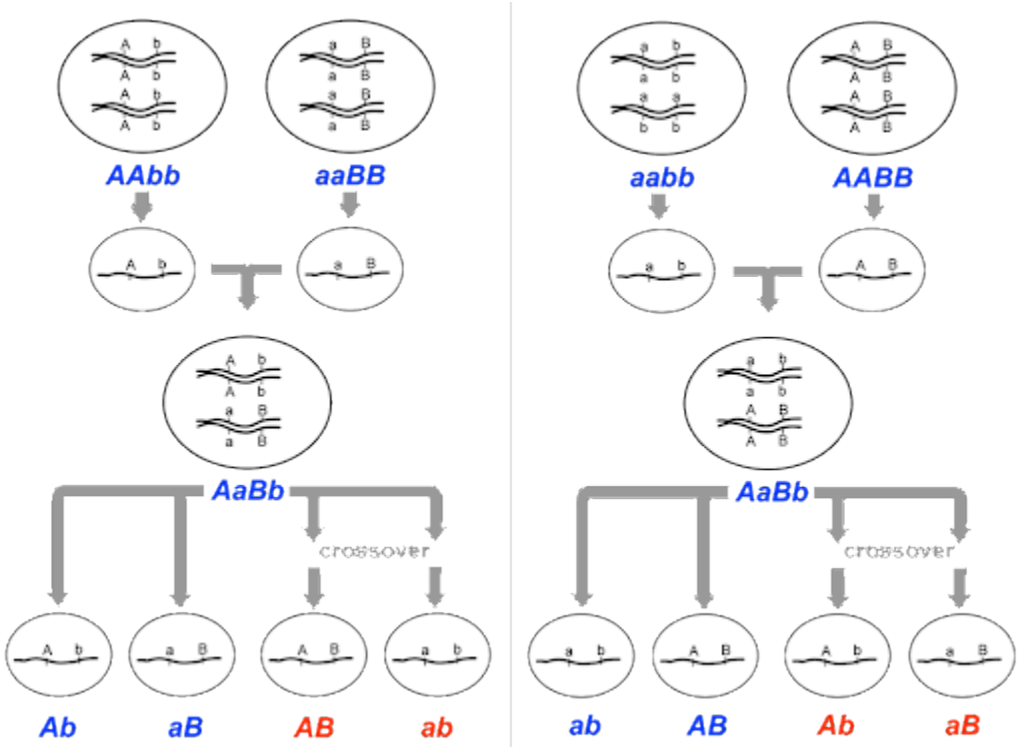


Figure 8. Crossing over between homologous pairs allows recombinant chromosomes to be produced. Two parental meiocytes (AAbb and aaBB) produce gametes Ab and aB, which come together to make a dihybrid F1 with genotype AaBb. The dihybrid can make two parental gametes (Ab and aB) and two recombinant gametes (AB and ab). The recombinant gametes are only produced if crossing over occurs. This leads to recombinant gametes (shown in red).

For unlinked genes, all four potential phenotypic classes – Ab, aB, AB, and ab – will be equally represented among the F2 offspring. This is the expected 1:1:1:1 ratio for a dihybrid testcross. But if the genes are linked, the parental classes Ab and aB will be over-represented, and the recombinant classes (which only result from crossing over) will be under-represented.

The percentage of recombinant F2 offspring indicates the distance between the two genes. The unit of distance is called a "map unit" or a "**centiMorgan**", abbreviated m.u. or cM.

An example of this from Sturtevant's 1913 paper is shown in **Figure 9.** Wild-type flies have long wings and red eyes, but some mutant flies have miniature wings and/or white eyes. Using flies from true-breeding lines, females with miniature wings and white eyes were crossed with males with red eyes and long wings. All the F1 females were long red and presumably dihybrid. These dihybrid females were crossed to miniature white males, and the F2 offspring were counted.

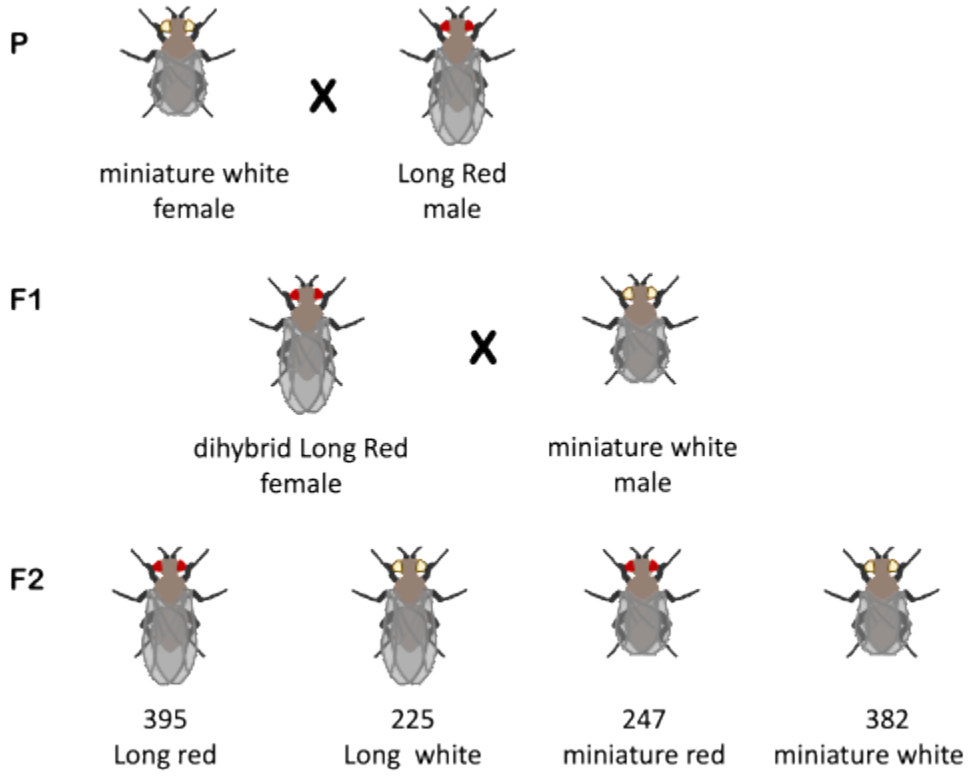


Figure 9. Example of a cross Sturtevant used to calculate recombination frequency. Miniature white female flies crossed with Long Red males give F1 dihybrid Long Red females. The F1 dihybrids testcrossed with miniature white males give F2 offspring with both parental and recombinant phenotypes. The phenotypes of the F2 offspring can be used to calculate recombination frequency as the number of recombinant offspring divided by the total number. Here, the recombination frequency is 38%. Note that Sturtevant combined this cross with others to have a more robust data set: the actual recombination frequency between white and miniature is closer to 34.6%.

Among the F2 offspring, Sturtevant observed 395 Long Red, 225 Long white, 247 miniature red, and 382 miniature white. Long Red and miniature white are parental classes: those are the phenotypes of the P generation. The long white and miniature red phenotypic classes are recombinant, and they are underrepresented compared to the parental classes.

To calculate the recombination frequency, the equation below is used:

$$\textit{recombination frequency (\%)} = 100 * \frac{no.\ recombinants}{total\ no.\ of\ offspring}$$

For the cross shown in **Figure 9**, the recombinant frequency is calculated as 38%.

$$100 * \frac{225 + 247}{395 + 225 + 247 + 382}$$

We would therefore conclude from these data that the white and miniature are 38 map units apart. Note however that Sturtevant combined this cross with others, so he ended with 1643 recombinants out of 4749 total flies, for a recombination frequency of 34.6%.

From recombination frequencies, a **chromosome map** can be drawn, indicating the distance between the two genes. Maps usually depict chromosomes as a straight line, with "tick marks" indicating the location of genes. Maps for the miniature and white genes are drawn in **Figure 10**.

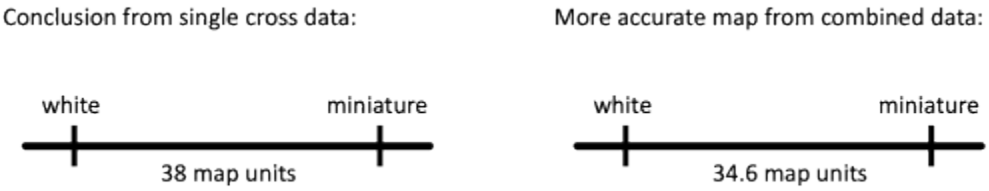


Figure 10. Two chromosome maps estimating distance between white and miniature.

On the left in **Figure 10** is the map that would be drawn from the single cross shown in Figure 9. On the right is the map that would be drawn from Sturtevant's combined data. The greater the number of offspring counted, the more reliable a map becomes. *Drosophila* are a powerful model organism for genetics, since many offspring can be generated in a short period of time – each female can produce hundreds of eggs, and a new generation reaches reproductive maturity in about two weeks.

Test Your Understanding

Table 1

Phenotype	# of F2 offspring
AB	302
Ab	706
aB	695
ab	297
total	2000

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=439#h5p-81

Table 2

Phenotype	# of F2 offspring
EF	305
eF	195
Ef	202
ef	298
total	1000

Media Attributions

- Figure 8 © Online Open Genetics, Chapter 7. Original-Deyholos is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- White female flies and Red males flies © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Distance between white and miniature © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

50% RECOMBINATION FREQUENCY IS THE MAX, BUT CHROMOSOMES ARE LONGER THAN 50 MAP UNITS

Mapping with a two-point testcross is most accurate for genes that are close together. The further apart they are, the less reliable the distance. This is for at least two reasons: first, the maximum recombination frequency is 50%, and second, more than one crossover event can occur between two genes.

Why is 50% recombination frequency the max? Remember, for independently assorting genes, testcrossing a dihybrid gives a 1:1:1:1 ratio of phenotypes among the offspring. Two of those classes are always parental and two are always recombinant, so a 1:1:1:1 ratio gives a 50% recombination frequency.

But why? It goes back to the mechanism of crossing over during meiosis. When crossovers form, they do not always generate recombinant products. The crossovers may be resolved, or cut apart, by separating the chromosomes back into their original conformation. Or, they can be cut apart generating recombinant products. This is shown in **Figure 11**. This happens randomly, so about 50% of crossovers are resolved non-recombinantly, restoring the parental chromosome.

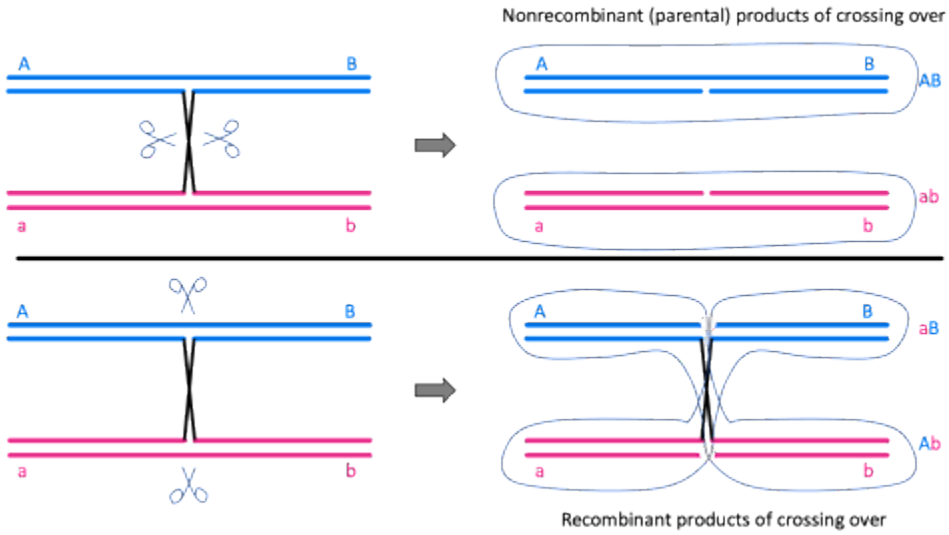


Figure 11. During meiotic recombination, crossovers may be resolved by cutting apart the chromatids, returning the chromosomes to their original conformation. Or, they may be cut apart to generate recombinant products.

Although the recombination frequency maxes out at 50%, genes can be much farther apart than 50 map units. To get around this experimental limitation, maps are typically constructed by compiling evidence from

multiple pair-wise recombination frequencies. For example, for three genes Q, R, and S, we might see the following recombination frequencies:

Table 3. Recombination frequencies for hypothetical genes Q, R, S

Gene pair	Recombination frequency
Q and R	5%
R and S	10%
S and Q	15%

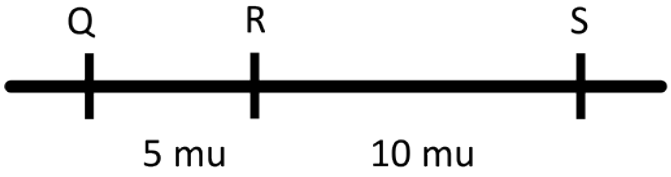


Figure 12. QRS chromosome map using the data from Table 3.

This indicates that Q and R are 5 map units apart, R and S are 10, and S and T are 15 map units apart. We can draw the map in Figure 12.

One of the tricky things for this, though, is that it might appear that genes are unlinked, with a 50% recombination frequency. But that doesn't mean that the map distance is 50 map units! The example in **Table 4** and **Figure 13** illustrate this.

Table 4. Recombination frequencies for hypothetical genes T, U, V

Gene pair	Recombination frequency
T and U	50%
U and V	30%
V and T	40%

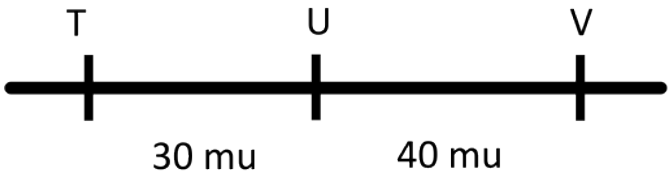


Figure 13. TUV chromosome map using the data from Table 4.

In **Table 4** and **Figure 13**, it might be tempting to say that T and U are 50 map units apart. But that is not correct – a 50% recombination frequency means the genes are unlinked, either on two different chromosomes or far apart on the same chromosome. And the data for U and V, and V and T, tells us that T and U are far apart on the same chromosome (about 70 map units, since 30 + 40 = 70). Always use the shortest "steps" to construct a map.

Test Your Understanding

Draw a chromosome map for genes L, M, and N.

Gene pair	Recombination frequency
L and M	50%
M and N	35%
N and L	25%

While the specific examples given in this section describe two-point test crosses in fruit flies, the same principles were historically used to map genes in other organisms as well.

Media Attributions

- Cutting the chromatids © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 12 © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 13 © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

MULTIPLE CROSSOVERS: THE THREE-POINT TESTCROSS

The recombination frequency underestimates long map distances as discussed in the previous section. In addition, if multiple crossover events occur between two genes, those crossover events will not be detected using a two-point test cross. This is illustrated in Morgan's 1923 text, reprinted in **Figure 7b** from earlier in this chapter. Although two crossovers have occurred between genes W and Br, they would not be counted in a two-point test cross between W and Br because the parental alleles would still appear to be linked. Multiple crossover events are increasingly likely to happen the further apart the genes.

To partially solve this, a three-point test cross can be used. Having a third point intermediate to the other two allows detection of double-crossover events, since the middle gene would be swapped out in the case of a double recombinant. This is also called a trihybrid testcross – you might recall that a trihybrid would be heterozygous for three genes, for example AaBbCc.

An example of a hypothetical three point testcross is shown in **Figure 14**, reprinted from *Online Open Genetics* (Nickle and Barrette-Ng)[1]. Here, three traits are tracked in mice: tail length (Long or short), fur color (Brown or white), and whisker length (Long or short).

While a dihybrid testcross gives 4 potential classes of offspring (AB, Ab, aB, ab), a trihybrid testcross potentially gives 8 classes of offspring (ABC, abc, Abc, aBC, aBc, AbC, ABc, abC). The 8 possible classes are shown for the F2 generation in **Figure 14.**

1. Nickle and Barrette-Ng. Open Online Genetics. in *Open Online Genetics* (2016).

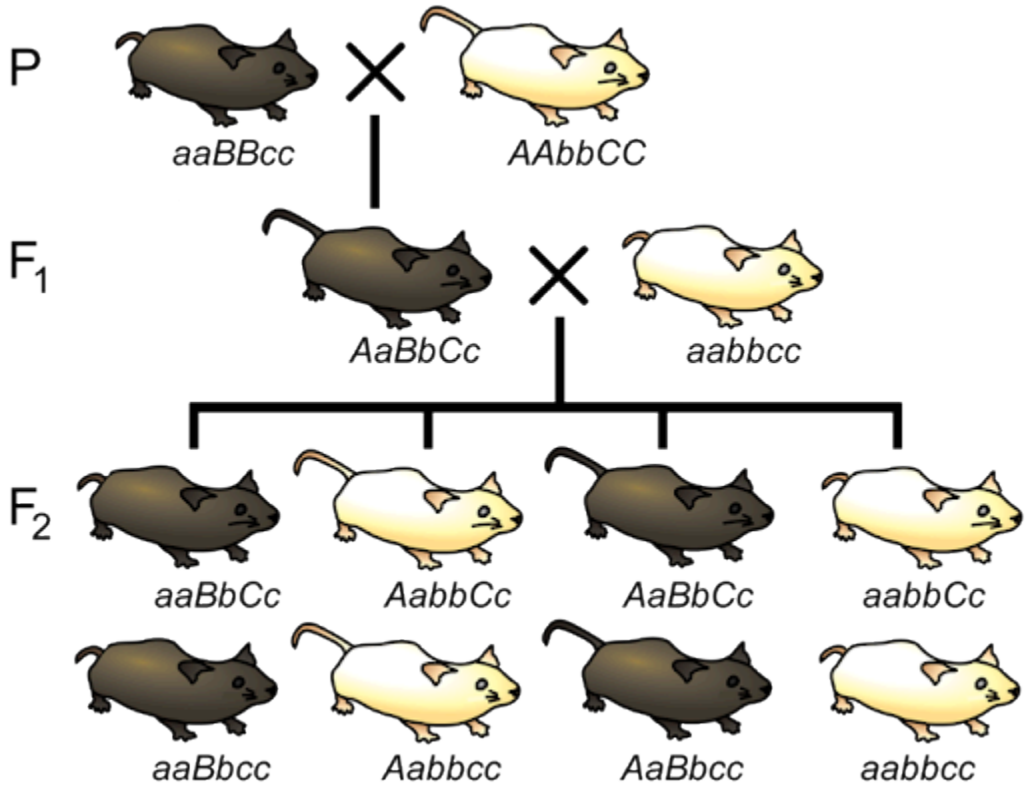


Figure 14. A three-point testcross tracking fur color, whisker length, and tail length in mice.

Table 5 gives sample data for the testcross in **Figure 14**. The parental phenotypes are highlighted. *For any testcross with linkage, the parental phenotypes will be the most numerous class.*

Table 5 Three-point testcross results

Class number	tail phenotype (A)	fur phenotype (B)	whisker phenotype (C)	gamete from trihybrid	genotype of F2 from test cross	number of progeny
1	short	Brown	Long	*aBC*	*aaBbCc*	5
2	Long	white	Long	*AbC*	*AabbCc*	38
3	short	white	Long	*abC*	*aabbCc*	1
4	Long	Brown	Long	*ABC*	*AaBbCc*	16
5	short	Brown	short	*aBc*	*aaBbcc*	42
6	Long	white	short	*Abc*	*Aabbcc*	5
7	short	white	short	*abc*	*aabbcc*	12
8	Long	Brown	short	*ABc*	*AaBbcc*	1
Total						120

The rest of the data are analyzed by looking at two genes at a time to calculate recombination frequency. So, for example, we can look at just tail and fur phenotypes, as if we'd never tracked whisker phenotype at all.

The parents for this cross had short tail, Brown fur and Long tail, white fur. Those are highlighted in yellow in **Table 6.** That means the recombinants are short, white and Long, brown. Those are highlighted in green. There are 30 total recombinant offspring out of 120, or 25%. This means the tail gene (A) and fur gene (B) are 25 map units apart.

We can then repeat this for the remaining pair-wise combinations. **Table 7** shows the count for recombination between fur (B) and whisker (C). **Table 8** shows the count for recombination between whisker (C) and tail (A).

Table 6. Calculating recombination frequency for tail and fur, The parental pheontypes are Long, white and short, Brown. Long, Brown and short, white are recombinant. Adding the progeny from all recombinant classes gives 30/120, or 25%.

Tail (A)	Fur (B)	whisker (C)	# of progeny	Recombinant for tail and fur?	Recombinant for fur and whisker?	Recombinant for tail/whisker?
short	Brown	Long	5	P		
Long	white	Long	38	P		
short	white	Long	1	R = 1		
Long	Brown	Long	16	R = 16		
short	Brown	short	42	P		
Long	white	short	5	P		
short	white	short	12	R =12		
Long	Brown	short	1	R = 1		
total			120	30		
% recombinant				30/120 = 25%		

Table 7. Calculating recombination frequency for fur and whiskers. The parental phenotypes are white, Long and Brown, s
Adding the progeny from all recombinant classes gives 38/120, or 32%

Tail (A)	Fur (B)	whisker (C)	# of progeny	Recombinant for tail and fur?	Recombinant for fur and
short	Brown	Long	5	P	R = 5
Long	white	Long	38	P	P
short	white	Long	1	R = 1	P
Long	Brown	Long	16	R = 16	R =16
short	Brown	short	42	P	P
Long	white	short	5	P	R = 5
short	white	short	12	R =12	R = 12
Long	Brown	short	1	R = 1	P
total			120	30	38
% recombinant				30/120 = 25%	38/120 = 32%

Table 8. Calculating recombination frequency for tail and whiskers. The parental phenotypes are Long, Long and short, Adding the progeny from all recombinant classes gives 12/120, or 10%

Tail (A)	Fur (B)	whisker (C)	# of progeny	Recombinant for tail/fur?	Recombinant for fur/wh
short	Brown	Long	5	P	R = 5
Long	white	Long	38	P	P
short	white	Long	1	R = 1	P
Long	Brown	Long	16	R = 16	R =16
short	Brown	short	42	P	P
Long	white	Short	5	P	R = 5
short	white	short	12	R =12	R = 12
Long	Brown	short	1	R = 1	
total			120	30	38
% recombinant				30/120 = 25%	38/120 = 32%

Finally, we can use those recombination frequencies to draw a map of the chromosomes. Fur (B) and Whisker (C) are farthest apart, so Tail (A) is in the middle of the three. Remember, the smallest map distances are the most reliable! So our map would look like the one seen in **Figure 15**, with C separated from A by 10 map units, and B another 25 map units from A.

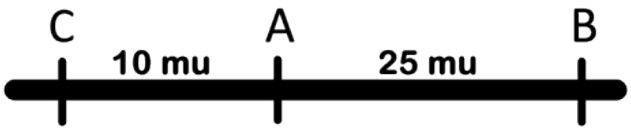


Figure 15. Map constructed from the recombination frequencies in Table 8.

Remember that we started this section by saying that a three point testcross can pick up double recombinants and give us more accurate map distances? Well, in each three point testcross two classes have the lowest number of progeny. *The two least-common phenotypes are always the reciprocal products of a double crossover.*

Comparing to the parental AbC and aBc, we see that gene A is the one that has been switched in the double recombinants abC and ABc. Just from this, we can conclude that gene A is in the middle in relation to the other two, as shown in **Figure 16**.

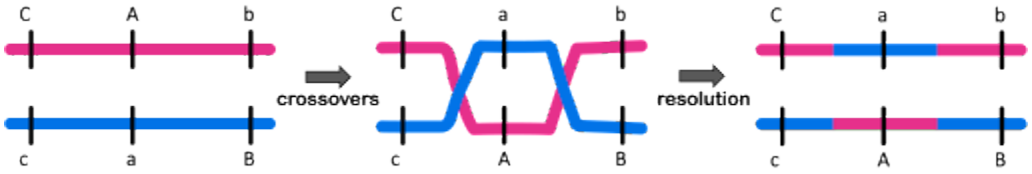


Figure 16. Double crossovers switch the alleles in the middle, while the outermost alleles remain apparently linked.

But it also gives us a way to count the double crossovers. Although we originally classified those as parental phenotypes for fur/whisker and didn't count them in our total recombinants, in actuality, there were *two* crossover events between fur/whisker! So we can edit our calculations: our new total number of recombinants between fur (B) and whisker (C) is 5+1+1+16+5+12+1+1 = 35%. This now matches the number we would expect if we added the two shorter map distances, C-A and A-B, together!

Table 9. Revised recombination frequency for fur/whisker, taking into account double recombinants. Although the short, must be two crossovers between the fur and whisker loci. Therefore, the short, white, Long, and Long Brown short proge frequency – and map distance – better match what is expected fro

Tail (A)	Fur (B)	whisker (C)	# of progeny	Recombinant for tail/fur?
short	Brown	Long	5	P
Long	white	Long	38	P
short	white	Long	1	R = 1
Long	Brown	Long	16	R = 16
short	Brown	short	42	P
Long	white	Short	5	P
short	white	short	12	R =12
Long	Brown	short	1	R = 1
total			120	30
% recombinant				30/120 = 25%

In this way, a three-point test cross gives an advantage over the two-point test cross: it allows the determination of the linear arrangement of multiple genes and it allows the calculation of double recombinants. Because of this, a three-point test cross typically provides a better estimate of map distances than multiple two-point test crosses.

However, keep in mind that it still isn't perfect. There are still additional double recombinants that are not detected – in this case, between A-C and A-B. In addition, these calculations assume that crossovers happen

independently of one another. But they don't. A crossover in one location can prevent the formation of a second crossover nearby, in a phenomenon called **interference.**

Although variations of test crosses are still sometimes used today, particularly in plants, mapping a phenotype to a particular part of a chromosome typically involves much more powerful methods of molecular genetics.

Media Attributions

- Three-point testcross © Original-Modified Deyholos-CC: AN, via Online Open Genetics (Nickle and Barrette-Ng). is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Figure 15 © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Switch the alleles in the middle © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

TRACKING LINKED TRAITS THROUGH PEDIGREE ANALYSIS

The results of multiple test crosses were historically used to build a chromosome (and genome) map. Hundreds of genes could be mapped in relation to each other in this fashion – even some of Morgan's initial chromosome maps included nearly one hundred genes.[1]

If a new phenotype was identified, test crosses with known genes could help map the trait to the causative region of a chromosome. This was especially true in organisms where large numbers of offspring could be generated by a single cross. However, test crosses were not useful in organisms where controlled crosses cannot be performed (including humans, where such a thing would be deeply unethical) or for organisms where few progeny might be generated at a time. Test crosses were also less useful for complex multifactorial traits.

In humans, there are many obvious clinical benefits to identifying genes associated with particular phenotypes. By the mid-1950s, geneticists were interested in building linkage groups in humans. Since controlled crosses could not be used, linked traits were tracked through pedigree analysis in large families. They began by looking for phenotypes that could be easily observed and tended to coincide in members of a family.

This sort of pedigree analysis was most useful for traits where both alleles were known for all individuals in a pedigree. So early pedigree analysis often looked for linkage with traits like the ABO blood group: With three different blood type alleles common in human populations, homozygosity and heterozygosity could often be determined from pedigree analysis. One of the earliest linkages found was between Nail Patella Syndrome and the ABO blood group.

1. Bridges, C. B. & Morgan, T. H. *The Third-Chromosome Group of Mutant Characters of Drosophila Melanogaster*. (Carnegie Institution of Washington, 1923).

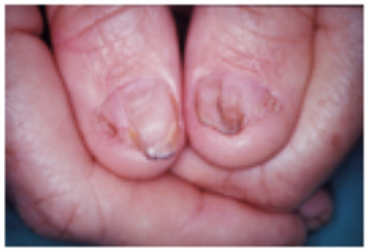


Figure 17. Nail Patella Syndrome is characterized by atypical growth of fingernails (shown here) and absent or underdeveloped patellae (kneecaps), as well as other developmental differences in the elbow, kidney, and pelvis.

Nail Patella Syndrome is characterized by atypical fingernail growth, missing or misshapen kneecaps, and other developmental differences in the elbow, kidney, and pelvis[2]. An image of the fingernail phenotype is shown in **Figure 17**. Nail Patella Syndrome is inherited in autosomal dominant fashion. Most examples of Nail Patella Syndrome are familial, but there have been documented examples of NPS patients with *de novo* mutations that were not inherited from parents.

An example of how NPS and ABO blood type might track in a family is shown in **Figure 18**.

2. Nail-patella syndrome: MedlinePlus Genetics. https://medlineplus.gov/genetics/condition/nail-patella-syndrome/.

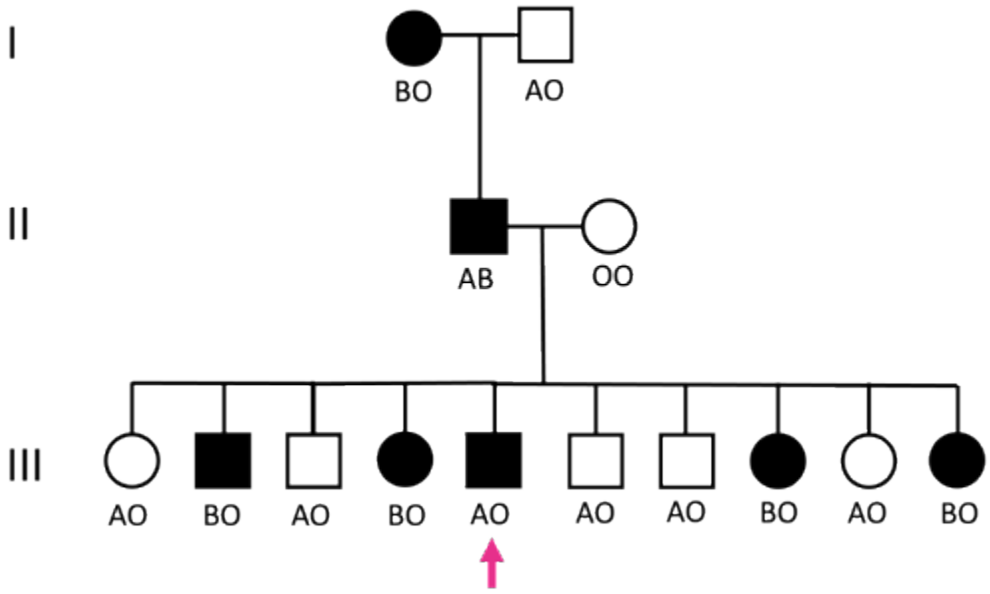


Figure 18. Hypothetical pedigree of a family with Nail Patella Syndrome. Note that almost all family members with NPS also have blood type B, indicated by the BO genotype under each individual in the pedigree. The exception is the individual in generation III marked with a pink arrow.

The ABO blood group locus determines whether you have blood type A, B, AB, or O. There are three alleles: A, B, and O, where A and B are codominant, and O is recessive. In the pedigree in **Figure 18**, Nail Patella Syndrome tracks with the B allele in all individuals except the individual in generation III marked with a pink arrow. The individual marked with a pink arrow has a recombinant phenotype, arising from crossing over between the NPS locus and the ABO locus in the NPS-affected parent in generation II.

This type of pattern is typical of linked traits in a pedigree. With only a small number of people, it is difficult to have much confidence in recombination frequency. But just as Morgan and Sturtevant pooled data from multiple crosses, so can human geneticists pool data from multiple families. When this was done for Nail Patella Syndrome, a recombination frequency of about 10% between the NPS locus and the ABO locus was found, suggesting that the genes are about 10 map units apart. Although not discussed in this chapter, geneticists use statistical tests to determine confidence in their estimates of map distance. If you're interested in learning more about this, you can find an explanatory video from GnomX on YouTube.

Also like Morgan and Sturtevant, geneticists were able to combine additional data to build a larger linkage map; for example, NPS was later found to also be tightly linked to the adenylate kinase locus[3].

3. McIntosh, I., Dunston, J. A., Liu, L., Hoover-Fong, J. E. & Sweeney, E. Nail Patella Syndrome Revisited: 50 Years After Linkage. *Ann. Hum. Genet.* **69**, 349–363 (2005).

As the tools of molecular genetics were developed, testcross mapping experiments and pedigree analysis might serve as a first step in **cloning** a gene, meaning to isolate it for further study.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=453#h5p-82

Media Attributions

- Nps © Wikipedia is licensed under a CC0 (Creative Commons Zero) license
- Hypothetical pedigree © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

DNA VARIANTS CAN BE MOLECULAR MARKERS FOR LINKAGE

Linkage maps have limitations.

First, as it turns out, crossing-over does not occur uniformly across a chromosome. So-called "hotspots" for recombination exist, so a centiMorgan in one part of the genome may correspond to many more base pairs than a centiMorgan in another part of the genome. And recombination may occur at different rates in different organisms, so a centiMorgan in one species might include many more base pairs than in another. (In humans, a cM equals about 1 million base pairs[1] depending on genomic location.) Therefore, a linkage map is good for relative orientation of genes, but it does not directly translate to a physical map of DNA sequence.

Second, test crosses were most useful in organisms like fruit flies and plants, where thousands of offspring with variable traits could be measured in one controlled cross. But test crosses are not possible in all organisms or under all conditions, and not all organisms have such easily identifiable single-gene phenotypes. So, instead, in humans and other organisms, tracking the linkage of phenotypes with molecular markers was often more useful.

In the chapter on Mutation, we discuss different types of mutations and polymorphisms. Many neutral polymorphisms exist in the population of any species. Most are not within gene coding sequences or regulatory regions, although there are some exceptions. These polymorphisms are nevertheless quite useful as **molecular markers**, differences in DNA that can be detected molecularly and treated as codominant alleles.

Vocabulary: Just as alleles are versions of a gene, **morphs** are versions of a locus.

These polymorphisms, like all DNA variants, arise in any population through mutation. But because most do not affect reproductive fitness, there is no selective pressure for them to be maintained in or lost from the gene pool. As a result, there is a lot of variability among individuals in a population in certain regions of the genome. [In contrast, a mutation in part of the genome that is very important to an organism's function is likely to negatively affect reproductive fitness and be lost from a population. Such regions of the genome are said to be **highly conserved**, meaning that little variation is observed.]

Some examples of variants used as molecular markers are shown in **Figure 19**.

1. Centimorgan (cM). https://www.genome.gov/genetics-glossary/Centimorgan.

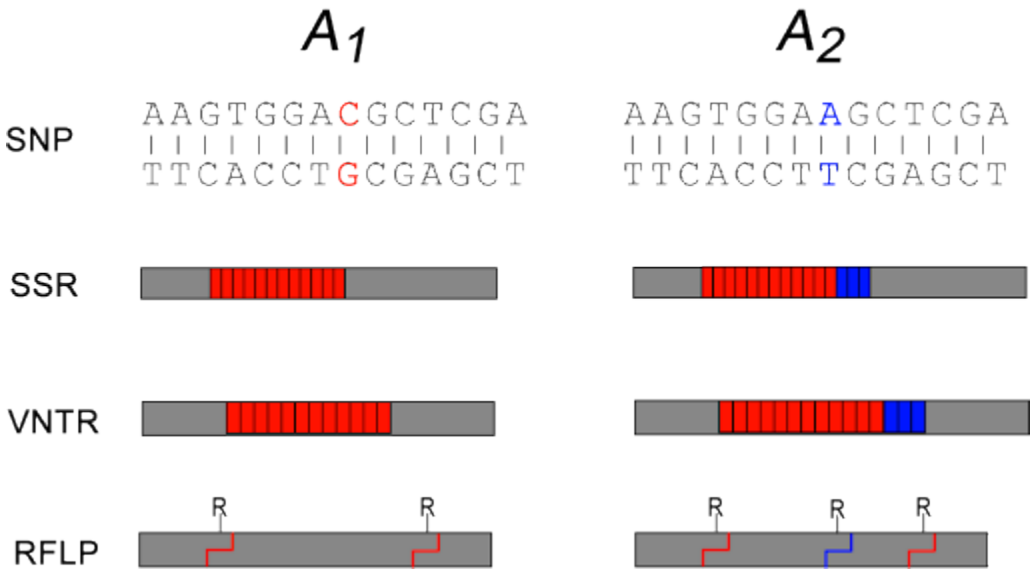


Figure 19. Reprinted from Online Open Genetics (Nickle and Barrette-Ng). Some examples of DNA polymorphisms. The variant region is marked in blue, and each variant sequence is arbitrarily assigned one of two allele labels (here, A1 and A2). Abbreviations: SNP (Single Nucleotide Polymorphism) = a single base change; SSR (Simple Sequence Repeat) = short sequences of 2-6 bases repeated variable numbers of times. Also called microsatellites, VNTR (Variable Number of Tandem Repeats) = longer sequences repeated variable numbers of times; RFLP (Restriction Fragment Length Polymorphisms = a change in sequence that creates or destroys a restriction site, so if the DNA is cut by a restriction enzyme different size fragments result. VNTRs and SSRs differ in the size of the repeat unit; VNTRs are larger than SSRs. (Original-Deyholos-CC:AN)

Some SNPs may change **restriction sites** in the genome. Restriction sites are sequences that are recognized by a **restriction endonuclease** which cut DNA in a sequence-specific manner. The change in sequence either creates or destroys a restriction site, affecting the length of the DNA fragment that is generated by the cut. This special subset of SNP is therefore called a **restriction fragment length polymorphism,** or RFLP.

SSRs are **simple sequence repeats** – repeating sequences like CAGCAGCAG or CCGCCGCCG. They are also called **short tandem repeats (STRs) or microsatellites**. **VNTR** stands for **variable number tandem repeat.** These are longer repeated sequences, where the repeated element is longer than just a few nucleotides. These sequences are usually detected by PCR of the region around the repeat, with longer repeats yielding a longer PCR product.

Detection of SNPs via PCR, RFLP analysis, and microsatellite analysis via PCR all generate different size DNA fragments that can be distinguished via gel electrophoresis. So gel electrophoresis analysis of morphs can substitute for analysis of phenotypes, with the added advantage that both alleles can usually be detected simultaneously (instead of just a dominant allele). An example of this is shown in **Figure 20**, where 3/15 F2 offspring are recombinant, making the recombination frequency 20%.

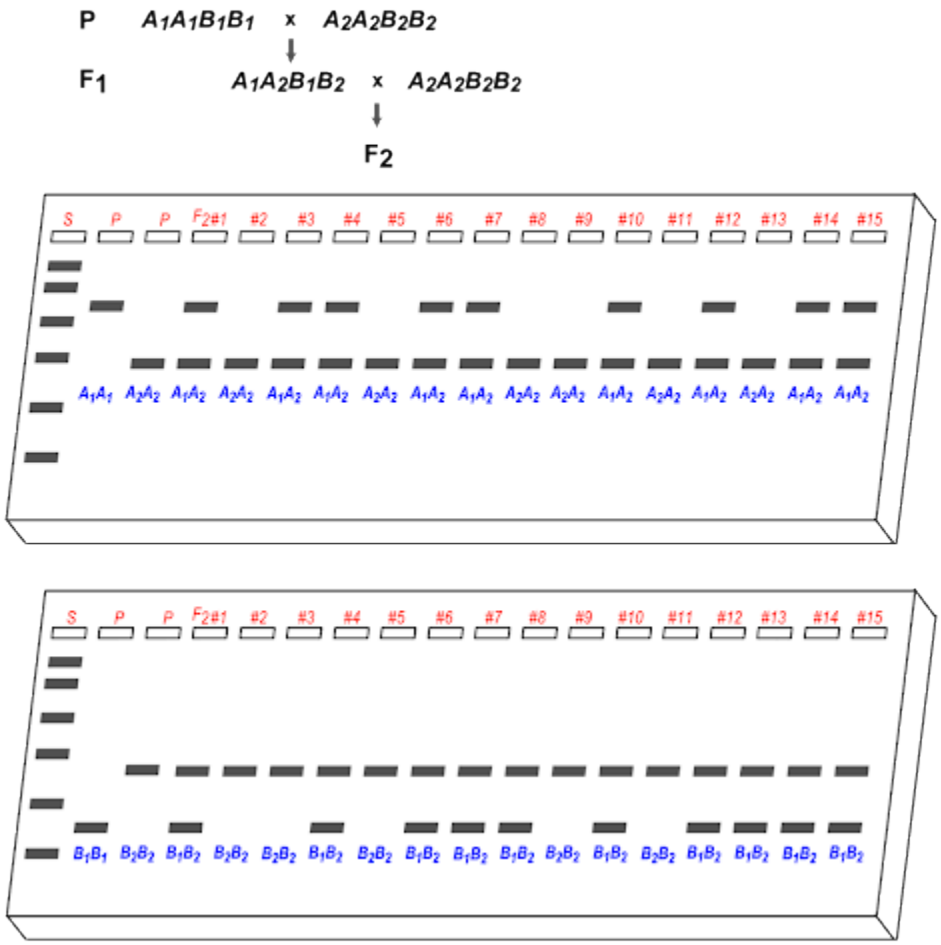


Figure 20. Measuring recombination frequency between two molecular marker loci, A and B. Two sets of PCR reactions are performed, one for locus A (top gel) and one for locus B (bottom gel), with the interpretation of the genotype indicated in blue below the bands. Results from parents (P) and 15 of the F2 offspring from the cross are shown. Recombinant progeny will have the genotype A1A2B2B2 or A2A2B1B2. Individuals #3, #8, and #13 are recombinant, so the recombination frequency is 3/15=20%. Image source: Reprinted from Online Open Genetics, Chapter 10. (Original-Deyholos-CC:AN).

Using gel electrophoresis, a band on a gel can be treated as a "phenotype".

RFLP mapping was used in 1983 to identify the first disease-associated gene: the causative gene for Huntington's disease (HD). HD is an autosomal dominant, progressive neurological disorder with an age-of-onset in middle age. Disease symptoms include physical tremors, dyscoordination, and cognitive decline. Thoughts and mood can be affected, and many patients have hallucinations. There is no effective treatment to slow the progression of the disease, and it is ultimately fatal.

HD is caused by a triplet-repeat expansion in the *HTT* gene, which encodes the protein huntingtin. This is discussed further in the chapter on Mutation. To identify the gene that causes HD, several very large families

with multiple members with Huntington's disease were tested for linkage with RFLPs, which identified an RFLP on chromosome 4 that was often co-inherited with the HD phenotype[2].

Media Attributions

- DNA polymorphisms © Nickle and Barrette-Ng via. Online Open Genetics is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Measure recombination frequency © Nickle and Barrette-Ng via. Online Open Genetics is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license

2. Gusella, J.F., Wexler, N.S., Conneally, P.M, Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., Young, A.B., Shoulson, I., Bonilla, E., Martin, J.B. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).

HAPLOTYPES

It's important to recognize that linkage studies like the HD RFLP analysis do not identify a causative gene, so the coinheritance of the chromosome 14 RFLP did not mean that polymorphism was within the *HTT* gene. Rather, the polymorphism was close enough to the gene that recombination did not occur in these families. Much more work was done to pinpoint the *HTT* more accurately within chromosome 4.

Like any DNA variant, polymorphisms arise due to a new (or *de novo*) mutation in a single individual. As that individual reproduces, the new variant may spread through the population. But remember that recombination is relatively rare between sites that are close together. So any new polymorphism tends to be inherited along with the sequence information that surrounds this. The new polymorphism would only be separated from surrounding polymorphisms due to recombination.

As a result, when we compare sequence information within a population of individuals, we tend to see blocks of sequence that are inherited together. For example, there are three polymorphic sites in the *TAS2R38* gene, which encodes a bitter taste receptor. These SNPs affect the protein-coding region of the gene: position 145 of the gene specifies either a proline or alanine in the protein depending on the variant, position 785 specifies alanine or valine, and 886 valine or isoleucine. The alleles of the gene are named for the amino acid specified by each SNP. For example, PAV has a proline-alanine-valine combination, and AVI has an alanine-valine-isoleucine combination.

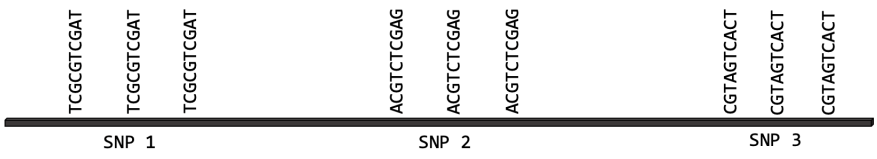
One might expect that, due to recombination, we would see all possible combinations of SNPs: PAV, PAI, PVI, etc. But, in fact, two are by far the most common, with about 95% of the human population having either PAV or AVI. People with the AVI variant usually cannot taste certain bitter substances, while people with the PAV variant can.

When we see certain morphs co-occurring more often than we'd expect, like the PAV or AVI combinations, we say that the loci are in **linkage disequilibrium**. The morphs that co-occur are said to be part of a **haplotype.** A haplotype is a group of SNPs (or other molecular markers) that tends to be inherited together in a block, as shown in **Figure 21.**

Figure 21. Haplotypes are a group of genomic variants that tend to be inherited together. In this image, a segment of DNA including three SNPs is shown. Two haplotypes exist: ATT and GAA.

Haplotypes often arise when a relatively recent mutation event occurs and the variant becomes fixed in the population. Although recombination can separate that new mutation from surrounding SNPs, that may take many, many generations for that to occur. We tend to find that certain haplotypes occur more often in some populations than others, so haplotypes can be used to infer ancestry.

SNPs can be measured via a variety of molecular methods, including specialized versions of PCR and microarray analysis. There are about 5 million single nucleotide variants in the human genome. A SNP microarray or "SNP chip" may be able to measure an individual's genotype at up to 2,000,000 SNPs with one test. Below is an animated GIF showing SNP microarrays.



SNP microarrays are constructed by immobilizing short pieces of DNA (called oligonucleotides) in clusters onto a support. For one type of SNP microarray manufactured by Illumina, these oligonucleotides correspond to part of the genome near a SNP, stopping one nucleotide short of the variable site. Each cluster of oligos corresponds to a different SNP. A SNP microarray can have hundreds of thousands of such clusters, or spots.

Animated GIF showing SNP microarrays.

SNP arrays are used for genome-wide association studies (discussed more in the next section) as well as other purposes, including personalized medicine. Most of the commercially available direct-to-consumer DNA test kits use SNP microarrays to test for ancestry and health-associated genetic variants. But it's not just individual SNPs that are used to determine ancestry. Very few SNPs belong to a single ancestral group, and no single SNP can be used to infer ancestry. Instead, these tests look for haplotypes and combinations of SNPs that are most common in people of certain ancestry.

Media Attributions

GENOME-WIDE ASSOCIATION STUDIES (GWAS)

All the linkage examples above, including two- and three-point test crosses and pedigree analysis, rely on pairwise analysis of phenotypes and/or molecular markers. But this is slow.

In the case of the Huntington's Disease mapping, it took decades of work and the construction of extensive pedigrees in a large Venezuelan family with Huntington Disease to identify the *HTT* gene. The project was propelled by the geneticist Nancy Wexler **(Figure 22)**, who began the project because she watched her mother, grandfather, and three uncles suffer from the disease. It also greatly depended on the willing participation of the Venezuelan family, many of whom understandably were initially suspicious of Dr. Wexler's motivations for the project.

Figure 22. Nancy Wexler.

Over three decades, Dr. Wexler and her colleagues tracked phenotypes in 10 generations and 18,000 individuals, processing 4,000 blood samples to look for molecular markers. (Dr. Wexler's foundation also founded a nursing home for HD patients in this region of Venezuela, sending tens of millions of dollars in clinical support to the affected community.) If you're interested in learning more about Dr. Wexler and her work, you can read interviews with her from the Lasker Foundation and the New York Times.

The identification of the *HTT* gene was truly an accomplishment in the days before the Human Genome Project. But with the advent of genome sequencing in the early 2000s, identifying regions of a chromosome associated with a phenotype shifted to large-scale data analysis. Rather than looking for a link between two genes at a time, we can now study the association of a trait with millions of sequence variants in one experiment.

Genome-wide association studies (GWAS) compare sequence information for the entire genome at once. The most common version of a GWAS uses SNP microarrays that test hundreds of thousands of SNPs (or even millions!) at once.

GWAS do not need to test every SNP in the genome to have a good understanding of genomic sequence as a whole, since several adjacent SNPs may be linked within a haplotype.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=464#h5p-83

To perform GWAS, sequence variants from two groups of people are compared: those with a certain phenotype (cases) and those without the phenotype (control). The bigger the cohorts, the more likely a genetic association can be identified. Powerful statistical software is used to compare the frequency of each individual variant (up to 2,000,000 at a time) in each group (case or control). The statistical significance of any difference is often presented in a Manhattan plot like that seen in **Figure 23**, which looks at GWAS for pigmentation in skin, hair, and eyes.

A Manhattan plot arranges all SNPs along the X-axis according to position, and the statistical significance of an association between SNP and phenotype on the Y-axis, reported as -log10(p-value). Each dot on the plot therefore represents one SNP. The threshold for statistical significance is indicated by the red line in **Figure 23**, so a dot that falls at or above the line represents a SNP that is a candidate for association with the measured phenotype. (The threshold for significance is corrected for the number of SNPs, so the threshold will be different for different studies.)

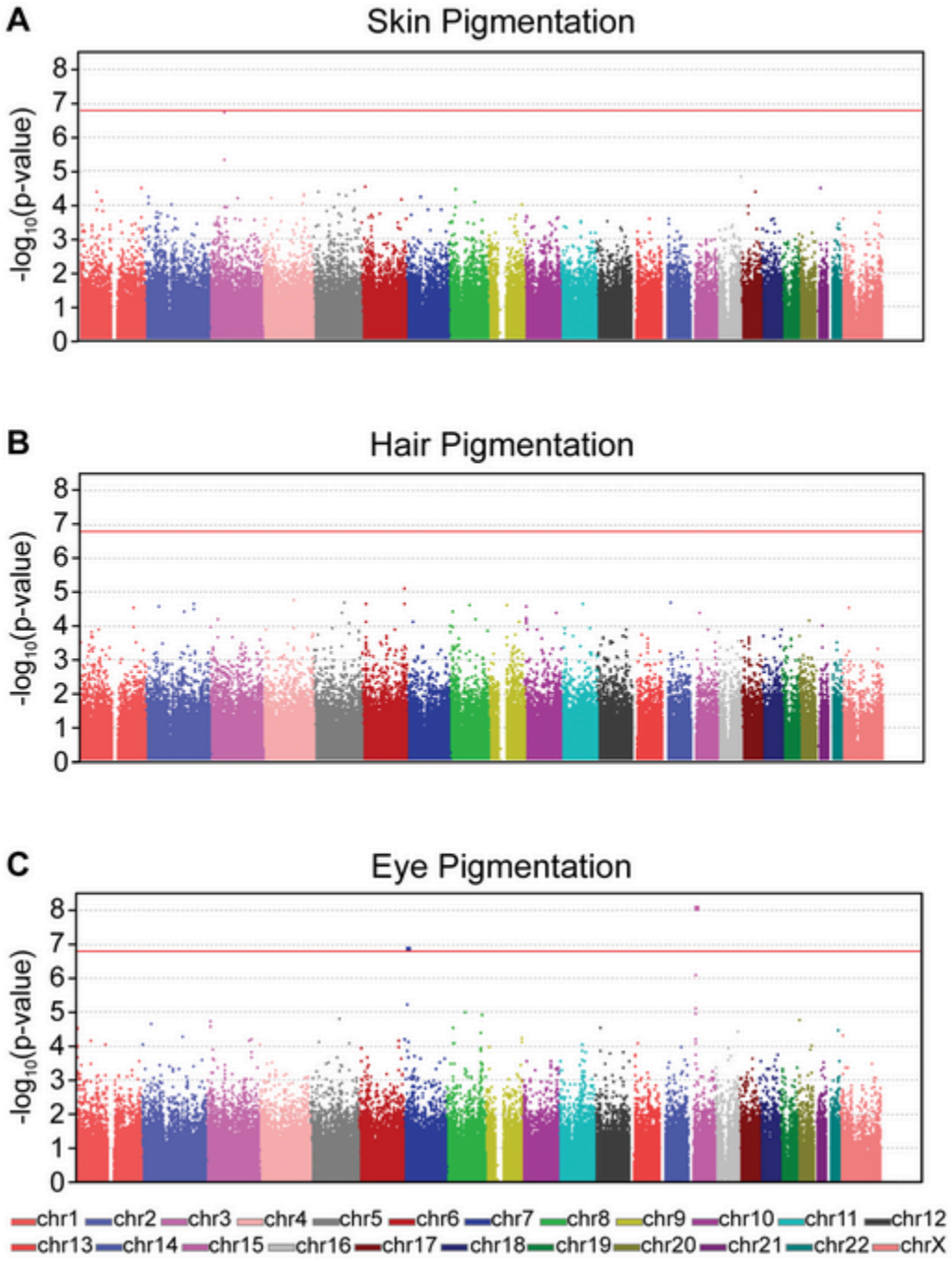


Figure 23. Manhattan plots for the GWAS results for the skin (A), hair (B), and eye (C) pigmentation. The threshold for genome-wide significance is indicated as a red line. For skin pigmentation, one SNP on chromosome 3 in the FLNB gene almost reaches genome-wide significance (p-value = 1.8×10^{-7}). No SNP achieves genome-wide significance in the hair pigmentation GWAS. For eye pigmentation, two SNPs, one near the pigmentation gene OCA2 on chromosome 15 and one in the SCIN gene on chromosome 7 achieve genome-wide significance.

GWAS can be complicated by the fact that many phenotypes are multifactorial, with multiple genes associated with the phenotype and environmental factors increasing or decreasing susceptibility. However, because they

compare such large data sets, GWAS can identify associated variants that contribute in part to a complex phenotype.

The example presented in **Figure 23** uses SNP microarray analysis with 317,503 SNPs. So, as with the molecular linkage experiments described earlier, the assay only identifies markers that are associated with the trait, finding potential haplotypes within the study participants. These SNPs represent only 0.01% of the genome, but since they are spread throughout the genome they give good coverage to identify linked genes. Further analysis is still needed afterward to identify candidate genes near the SNP.

Genome-wide association studies can also be performed with exome sequencing and whole-genome sequencing. These give much more coverage and can identify specific sequences associated with a trait. But they are also more expensive to perform.

Exome analysis provides the sequence for the expressed regions of the genome – those regions transcribed into RNA. The idea is that the protein-coding regions of the genome are the most likely to affect phenotype, but only make up about 1% of the whole genome. So exome analysis can identify many sequence variants with only slight sacrifice to impactful genomic coverage. But exome sequencing cannot pick up variants in regulatory sequence. Since promoter regions, enhancer regions, and even some intron sequences can have a profound effect on the amount of protein produced, exome sequencing can miss some associations.

Family exome analysis compares the exome of members of a family and can be particularly helpful in identifying *de novo* mutations. Because families share genomic sequence (50% for parent and child, 25% for siblings), sequence that is different among family members with different phenotypes is a candidate for a mutation associated with the phenotype. Remember, though, every child will have a handful of mutations that are not present in the parent. Thus, even if a mutation is identified through family exome analysis, it does not mean that the mutation causes the phenotype in question.

Whole genome sequencing, in contrast, offers full coverage of the genome. But as it is much more expensive to sequence a whole genome than it is to sequence 1% of the genome, whole genome sequencing is not yet the most common method for GWAS. As costs for sequencing continue to decrease, it's likely that whole genome sequencing will become more readily used.

We started this module with a description of ALS. Ice Bucket Challenge funding has allowed the identification of genes involved in familial ALS. Much of this work has used GWAS, family exome sequencing, and various applications of SNP microarrays.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it

online here:
https://rotel.pressbooks.pub/genetics/?p=464#h5p-84

Media Attributions

- Nancy Wexler © Christopher Michel via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Manhattan plots for the GWAS © Sophie I. Candille , Devin M. Absher, Sandra Beleza, Marc Bauchet, Brian McEvoy, Nanibaa' A. Garrison, Jun Z. Li, Richard M. Myers, Gregory S. Barsh, Hua Tang , Mark D. Shriver - Candille SI, Absher DM, Beleza S, Bauchet M, McEvoy B, Garrison NA, et al. via. Wikipedia is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license

SUMMARY

In this module, we looked at landmark historical experiments that began the task of mapping phenotypes to specific loci in the genome. This included linkage mapping experiments in *Drosophila melanogaster*, early linkage mapping in humans using pedigrees, mapping via molecular markers, and, most recently, genome-wide association studies.

Although the test-cross strategy used by T.H.Morgan and other early geneticists is no longer commonly used to map phenotypes to chromosomes, it's nonetheless useful to understand the concept of linkage. Modern mapping experiments via SNP association still are based on the principle that little recombination occurs between loci that are close together on the same chromosome, so when SNPs are co-inherited with a phenotype it suggests they may be tightly linked.

WRAP-UP QUESTIONS

Questions 1-11 are reprinted from *Online Open Genetics* (Nickle and Barrette-Ng), Chapter 7.

1. If you knew that a locus that affected earlobe shape was tightly linked to a locus that affected susceptibility to cardiovascular disease human, under what circumstances would this information be clinically useful?

2. In a previous chapter, we said a 9:3:3:1 phenotypic ratio was expected among the progeny of a dihybrid cross, in absence of gene interaction.

a. What does this ratio assume about the linkage between the two loci in the dihybrid cross?
b. What ratio would be expected if the loci were completely linked? Be sure to consider every possible configuration of alleles in the dihybrids.

3. In corn (i.e. maize, a diploid species), imagine that alleles for resistance to a particular pathogen are recessive and are linked to a locus that affects tassel length (short tassels are recessive to long tassels). Design a series of crosses to determine the map distance between these two loci. You can start with any genotypes you want, but be sure to specify the phenotypes of individuals at each stage of the process. Outline the crosses similar to what is shown in Figure 9, and specify which progeny will be considered recombinant. You do not need to calculate recombination frequency.

4. In a mutant screen in *Drosophila*, you identified a gene related to memory, as evidenced by the inability of recessive homozygotes to learn to associate a particular scent with the availability of food. Given another line of flies with an autosomal mutation that produces orange eyes, design a series of crosses to determine the map distance between these two loci. Outline the crosses and specify which progeny will be considered recombinant. You do not need to calculate recombination frequency.

5. Image that methionine heterotrophy, chlorosis (loss of chlorophyll), and absence of leaf hairs (trichomes) are each caused by recessive mutations at three different loci in Arabidopsis. Given a triple mutant, and assuming the loci are on the same chromosome, explain how you would determine the order of the loci relative to each other.

6. If the progeny of the cross *aaBB* x *AAbb* is testcrossed, and the following genotypes are observed among the progeny of the testcross, what is the frequency of recombination between these loci?

- *AaBb,* 135
- *Aabb,* 430
- *aaBb,* 390
- *aabb,* 120

7. Three loci are linked in the order B-C-A. If the A-B map distance is 1cM, and the B-C map distance is 0.6cM, given the lines *AaBbCc* and *aabbcc*, what will be the frequency of *Aabb* genotypes among their progeny if one of the parents of the dihybrid had the genotypes *AABBCC*?

8. Genes for body color (B black dominant to b yellow) and wing shape (C straight dominant to c curved) are located on the same chromosome in flies. If single mutants for each of these traits are crossed (i.e. a yellow fly crossed to a curved-wing fly), and their progeny is testcrossed, the following phenotypic ratios are observed among their progeny.

Body Color, Wing Shape	**Phenotypic Ratios**
Black, straight	17
yellow, curved	12
black, curved	337
yellow, straight	364

a. Calculate the map distance between B and C.
b. Why are the frequencies of the two smallest classes not exactly the same?

9. Given the map distance you calculated between B-C in question 12, if you crossed a double mutant (i.e. yellow body and curved wing) with a wild-type fly, and testcrossed the progeny, what phenotypes in what proportions would you expect to observe among the F_2 generation?

10. In a three-point cross, individuals *AAbbcc* and *aaBBCC* are crossed, and their F^1 progeny is testcrossed. Answer the following questions based on these F^2 frequency data.

aaBbCc	480
AaBbcc	15
AaBbCc	10
aaBbcc	1
aabbCc	13
Aabbcc	472
AabbCc	1
aabbcc	8

a. Without calculating recombination frequencies, determine the relative order of these genes.
b. Calculate pair-wise recombination frequencies (without considering double cross overs) and produce a genetic map.
c. Recalculate recombination frequencies accounting for double recombinants.

11. Wild-type mice have brown fur and short tails. Loss of function of a particular gene produces white fur, while loss of function of another gene produces long tails, and loss of function at a third locus produces agitated behavior. Each of these loss of function alleles is recessive. If a wild-type mouse is crossed with a triple mutant, and their F1 progeny is test-crossed, the following recombination frequencies are observed among their progeny. Produce a genetic map for these loci.

Fur	**tail**	**behavior**	**Number of offspring**
white	short	normal	16
brown	short	agitated	0
brown	short	normal	955
white	short	agitated	36
white	long	normal	0
brown	long	agitated	14
brown	long	normal	46
white	long	agitated	933

12. This module spends a great deal of time talking about building linkage maps of a chromosome. These are classical genetics experiments, but they are no longer performed very often (or at least not in this manner). Nonetheless, this topic is still included in most introductory Genetics textbooks.

Make a case for why linkage should (or should not) still be covered in introductory genetics textbooks. How is linkage connected to more contemporary methods for mapping genes to chromosomes?

13. Every child will have *de novo* mutations that make their genome different slightly different from that of their parents. Most are not associated with any change in phenotype, although occasionally some result in a change in phenotype. For example, would a SNP microarray, exome sequencing, or whole genome sequencing be most suitable for identifying *de novo*

14. Discuss the importance of GWAS in understanding human genetic diversity. How do these studies contribute to our understanding of the genetic basis of complex traits and diseases?

Science and Society

15. The Ice Bucket Challenge drew in $115 million in funding for ALS research, in just a few weeks after gaining attention on social media. The privately-funded ALS Association used these funds to sponsor ALS research. But much of the medical research in the United States is actually government funded. Interestingly, the increased attention on ALS may have subsequently driven government funding for ALS as well. The National Institutes of Health allocated $60 million to ALS research in 2014, but almost twice that in 2017[1]. Between 2020-2023, the estimated NIH funding for ALS research nearly doubled again, from $107-$206 million[2].

The additional attention and private funding almost certainly influenced NIH funding, either directly or indirectly. However, there are known gender-based and race-based disparities in research funding[3].

By what criteria should the NIH allocate funds? Some factors to consider might be the overall number of people affected by a disease, the severity of disease, who is affected by the disease, the likelihood of developing treatment quickly, the attention a disease receives in media (including awareness campaigns), etc.

---

1. EVALUATION OF THE ALS ASSOCIATION GRANT PROGRAMS Executive Summary Report. (2019).

2. Focus on Amyotrophic Lateral Sclerosis | National Institute of Neurological Disorders and Stroke. https://www.ninds.nih.gov/current-research/focus-disorders/focus-amyotrophic-lateral-sclerosis.

3. Mirin, A. A. Gender Disparity in the Funding of Diseases by the U.S. National Institutes of Health. *J. Womens Health* **30**, 956–963 (2021).

16. GWAS compare the genomes of hundreds, thousands, or even millions of individuals[4], looking for variants associated with particular traits. Careful consideration must be given to ensure that the populations compared are appropriate. For example, certain genetic disorders are more common in people of specific geographic ancestry. An example is cystic fibrosis, which is most common among people of European ancestry. A GWAS comparing cystic fibrosis patients with a control group of varying ancestry might flag other SNPs common in people of European ancestry rather than SNPs associated with cystic fibrosis.

    The **GWAS** in **Figure 23** are from a study looking at skin, hair, and eye pigmentation in European populations: specifically people from Ireland, Poland, Italy, and Portugal. What are the benefits in looking for this type of variation in these populations? Do the results of this study reflect the variation seen in the human population as a whole? Explain your reasoning.

17. Reflect on the ethical implications of GWAS. Consider issues such as privacy concerns, potential misuse of genetic information, and disparities in genetic research representation.

4. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).

PART XVI
EPIGENETICS

Objectives

1. Recognize that gene expression can change in different cell types over time and in response to changing conditions.
2. Define epigenetics.
3. Explain how modification of histone proteins affects gene expression by remodeling chromatin.
4. Explain how DNA methylation affects gene expression.
5. Describe examples of how epigenetics influences human phenotypes and behavior.

Introduction

In the history of science, Jean Baptiste Lamarck, an early evolutionary biologist, is often treated as an aside or a footnote. Although he, like Charles Darwin, proposed that the diversity of species on earth arose because of evolution, Lamarck proposed a very different mechanism from Darwin's natural selection. You may recall that Darwin proposed that species evolved because, in any heterogenous population, certain individuals were more "fit" and thus more likely to reproduce and pass their genetic information on to the next generation. Over time, less advantageous variations were outnumbered in the population in a manner called "natural selection". Thus future generations evolved to resemble the most fit individuals.

Figure 1. Lamarck proposed that populations evolve as characteristics acquired throughout an individual's lifetime are passed to the next generation. An example would be that a giraffe's neck is long because a short-necked ancestor, reaching to reach high leaves on a tree, stretched its neck through use. That giraffe's offspring then stretched their necks even farther, passing longer necks to their offspring. We now understand that this is not how evolution occurs.

The underlying principle of Darwinian evolution depends on the variation inherent in any population. Today, we know that variation arises in a population due to random mutations, some of which might be advantageous (and be selected for) and some of which might not (and be selected against).

Lamarck, in contrast, proposed that individuals within a population vary because of traits acquired during their lifetime. The classic example of the Lamarckian hypothesis is the evolution of long-necked giraffes: because a giraffe might have to reach its neck to eat leaves on a high tree, its neck would become stretched. And that giraffe would pass its longer, stretched neck to its offspring. (**Figure 1**) To use a more modern example: this would be like saying that your kids would be born understanding Genetics since you, their parent, had studied it.

We know now that this is not how evolution works. But Lamarck's work did get one thing right: in some cases, our children do inherit traits that we acquire during our lifetime. Our behavior and environment can have long-term effects on gene expression, linked with changes in chromatin structure. Those changes can influence traits over the course of a lifetime and even, in some cases, be inherited by offspring.

The mechanism by which we pass acquired traits to our children is not due to typical Mendelian inheritance. Instead, this inheritance is dependent on mechanisms outside of or on top of the rules of genetics we've discussed so far. The mechanism has thus been named **epigenetic** inheritance. A closer look reveals that epigenetic inheritance results from the regulation of gene expression through changes in chromatin structure that are transmitted to daughter cells during cell division.

In this module, we discuss **epigenetic** changes to chromatin that affect gene expression. In this module, we look at three specific molecular modifications, or **epigenetic marks**: histone acetylation, histone methylation, and DNA methylation.

Those epigenetic modifications are heritable through mitosis, so individual somatic cells can share epigenetic marks with other cells of the same lineage – meaning, other cells that came from the same parent cell during development. This type of epigenetic inheritance influences cell fate during the early development of multicellular organisms as well as certain complex genetic phenotypes, including obesity and cancer. In some cases, epigenetic modifications are also maintained through meiosis (so epigenetic marks can be shared from parent to offspring). Inter-generational epigenetic inheritance is linked with cardiovascular disease and stress response in humans and other mammals and other complex phenotypes.

Collectively, the combination of epigenetic marks found in a cell is called the **epigenome**.

Media Attributions

- Giraffes

CHROMATIN AND CHROMATIN DYNAMICS

Typically, we think of our genome as static, with genetic information passed along mostly intact from generation to generation. However, the overall structure of chromatin is not static. It changes by the stage of the cell cycle, with chromosomes condensing in preparation for mitosis or meiosis. It also changes in response to other stimuli affecting gene regulation. Different parts of the genome are packaged differently within chromatin. The Genome Structure module described how areas of euchromatin and heterochromatin – loosely packed and tightly condensed, respectively – can be observed microscopically.

Recall from the module on Genome Structure that chromatin consists of DNA wrapped around packaging proteins, which allow meters of DNA to be condensed into a nucleus that is measured in micrometers. Even the least condensed DNA in the cell is wrapped around histone proteins, forming a nucleosome core, as shown in **Figure 2**. The nucleosome core consists of about 147 base pairs of DNA wrapped around eight histone proteins: two copies each of histones H2A, H2B, H3, and H4. Because of this, the core histones are called a **histone octamer**. As illustrated in the bottom two panels, the least-condensed chromatin has the appearance of beads on a string when viewed using an electron microscope, but a chromosome is further condensed by packing nucleosomes tightly together. Not shown in this image is the linker histone H1, which helps pack nucleosomes more tightly, and other non-histone proteins that affect higher-order packing.

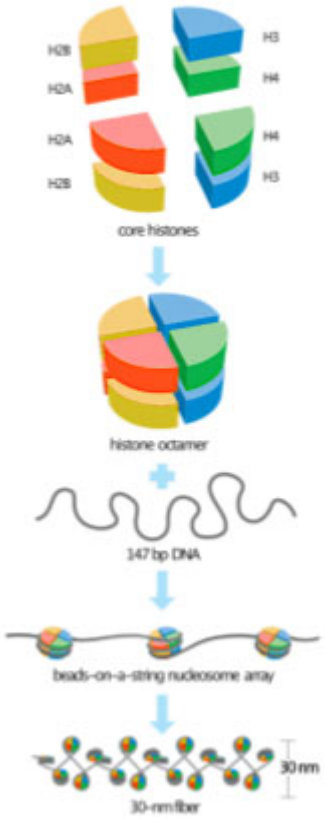


Figure 2. Nucleosome and chromatin structure. A nucleosome consists of about 147 basepairs wrapped around an octomer consisting of two copies each of histones H2A, H2B, H3, and H4.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=476#h5p-85

Three views of the molecular structure of a nucleosome are shown in **Figure 3**. The left panel shows a space-filling model of the nucleosome, with DNA shown in grey. The middle panel shows a ribbon diagram of the proteins, with DNA shown in purple. The right panel shows the same structure as in the middle but rotated 90 degrees to view the barrel-shaped structure from "above." In all three panels, histone H2A is shown in red, H2B in yellow, H3 in blue, and H4 in green.

Also recall: The modules on transcription and gene expression discussed how eukaryotic genes are regulated via promoter elements and enhancers. Protein factors bind to those DNA elements to regulate expression. But the proteins of chromatin can interfere with or block that binding. Usually, only short stretches of unbound DNA are accessible to proteins involved in cell functions like transcription. The rest wraps around histones, and those histones must be circumvented for the cell to access the DNA for gene regulation. The DNA is unwrapped from the histones, or the histones are moved before the DNA double helix is melted for replication. A class of proteins called **chromatin remodelers** helps make this happen.

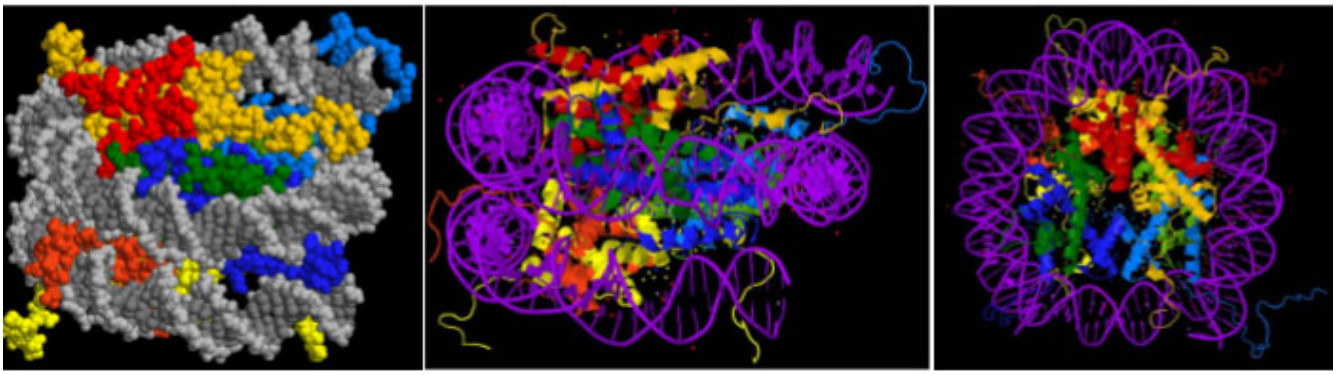


Figure 3. Three views of nucleosome structure. DNA wraps around a core of eight histone proteins: two copies each of H2A, H2B, H3, and H4, shown in yellow, red, blue, and green, respectively. Left: a space-filling diagram shows DNA in grey. Middle: a ribbon diagram shows DNA in purple. Right: a ribbon diagram of the middle structure, shown from above.

In the nucleosome structure shown in **Figure 3**, you can see that the bulk of each protein fits in the center of the coiled DNA. But the ends of the histone polypeptides – called the **histone tails** – stick out from the core. These are the colorful strings seen around the outside of the core, most readily visible in the ribbon diagrams in the two right panels. These are the N-terminal ends of all four histone polypeptides and, to a lesser extent, the C-termini of H2A and H2B. Positively charged amino acids (lysine and arginine) are over-represented in the tails, and this positive charge facilitates the interaction between negatively charged DNA and the histone octamer.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=476#h5p-86

The histone tails can be modified through the covalent linkage of functional groups like phosphate, acetyl, and methyl groups to the side chains of particular amino acids. Most commonly, the positively charged lysines and arginines are targets for modification. **Figure 4** illustrates the amino acids on each histone that can potentially be modified. This image uses the one-letter code for each amino acid, so lysine is indicated with a K and arginine with an R.[1] The modifications are sometimes called the **histone code**, signaling to the cell which parts of the genome should be transcribed. These changes in histone structure impact how tightly histones associate with the DNA, how densely chromatin is packaged within the nucleus, and the accessibility of the DNA for transcription and gene expression.

Figure 4. Histone modifications.The single-letter code for the amino acids in each histone tail is listed. Numbers below the sequence represent the position within the polypeptide, numbered starting with 1 at the N-terminus.

Differences in chromatin packaging can be seen within the nucleus using both light and electron microscopes. Densely packed chromatin is called **heterochromatin**. Epigenetically silenced regions of the genome are typically packed in heterochromatin, and heterochromatin has few actively transcribed genes. **Euchromatin** is more loosely packed, and most actively transcribed genes are found in euchromatic regions of the nucleus.

1. Rodriguez-Paredes, M. & Esteller, M. Cancer epigenetics reaches mainstream oncology. *Nat. Med.* **17**, 330 (2011).

Media Attributions

- Nucleosome and chromatin structure © David O Morgan via. Wikimedia is licensed under a CC0 (Creative Commons Zero) license
- Three views of nucleosome © Darekk2 via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Histone modifications © Modified from Mariuswalker, Based on Rodriguez-Paredes and Esteller, Nature, 2011 via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

THE HISTONE CODE

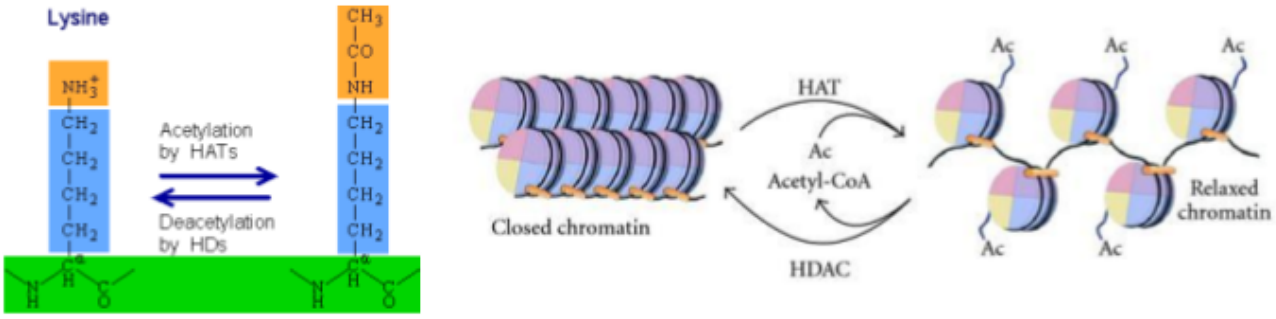


Figure 5. Histone acetylation. Left panel: Histone acetyl transferases (HATs) add acetyl groups to the side chains of lysine residues within histone tails, making the tails less positive. Acetyl groups can be removed by histone decacetylases (HDACs or HDs). Right panel: Acetylation relaxes chromatin packaging, making DNA more accessible to transcription factors.

Acetylation of histones is associated with increased gene expression

Acetylation of lysine covalently links an acetyl group ($-COCH_3$) to the amine group ($-NH_3^+$) in the lysine sidechain as shown in **Figure 5a**. This makes the side chain being less positive, which in turn loosens the interaction between DNA and histones and relaxes chromatin structure (**Figure 5b**).

Acetylation of lysines is catalyzed by a class of enzymes called **histone acetyl transferases**, or **HATs** for short. Acetylation is a dynamic process, and acetyl groups can be removed by enzymes called **histone deacetylases**, or **HDACs** for short. In general, acetylation is associated with loose chromatin structure in actively transcribed regions of the genome. Parts of the genome that are not actively transcribed are generally associated with lower levels of acetylation. Heterochromatin has little acetylation.

Methylation of histones has varied effects

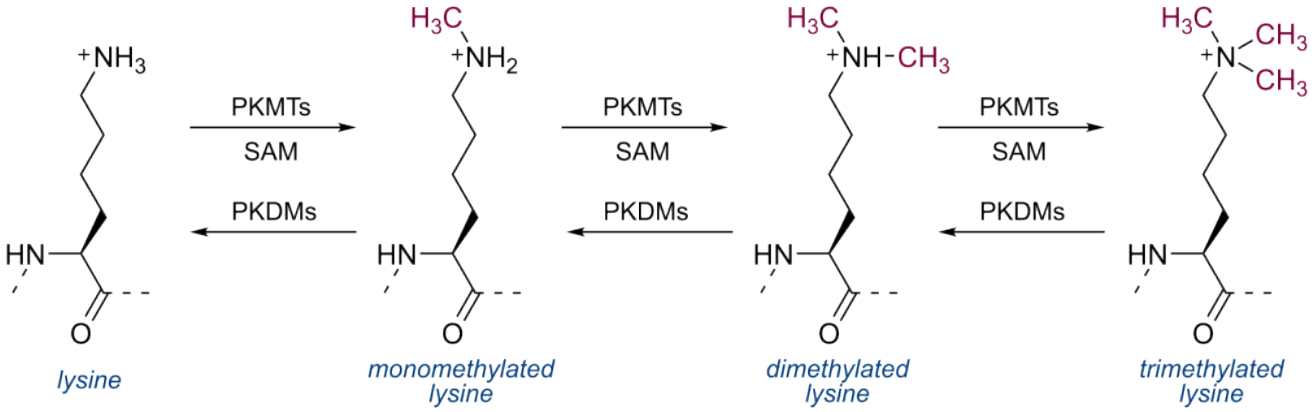


Figure 6. Mono- di- and trimethylation of lysine. Enzymes called histone methyltransferases (abbreviated MTs in this image) add methyl groups to the side chain of lysine using S-adenosylmethionine (SAM) as the methyl donor. Enzymes called histone demethylases (abbreviated DMs in this image) can remove methyl groups.

Histones can also be methylated. Enzymes called **histone methyltransferases** (often abbreviated as HMTs) catalyze the transfer of one or more -CH3 functional groups to selected lysine or arginine side chains in the histone tails. The structure of mono, di, and trimethylated lysine is shown in **Figure 6**. The cofactor S-adenosylmethionine (abbreviated SAM or SAMe) serves as the source or donor of the methyl group. **Histone demethylases** can remove methyl groups.

In contrast to acetylation, which is almost always associated with transcriptional activation, methylation of histones has a more varied effect. For example, methylation of histone H3 at position K4 is generally associated with increased transcription, but methylation of H3 at position K9 is associated with decreased transcription.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=480#h5p-87

Histone modifications are maintained after replication

Histone modifications are maintained in daughter DNA after replication and mitosis. During replication, nucleosomes must be disassembled from the parent DNA to unwind the double helix. Components of the replication machinery help reassemble nucleosomes on daughter strands of DNA, but twice as many nucleosomes are needed post-replication because there is twice as much DNA. Nucleosomes are reassembled on daughter DNA from a mixture of parental histones (which have been modified) and new, largely unmodified histones[1]. The histone code of the parental histones may then be read and re-written to the new histones, although the mechanism by which this occurs is not well understood[2].

Media Attributions

- Histone acetylation © Annabelle L. Rodd, Katherine Ververis, and Tom C. Karagiannis via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Lysine methylation © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

1. Zhang, W., Feng, J. & Li, Q. The replisome guides nucleosome assembly during DNA replication. *Cell Biosci.* **10**, 37 (2020).

2. Escobar, T. M., Loyola, A. & Reinberg, D. Parental nucleosome segregation and the inheritance of cellular identity. *Nat. Rev. Genet.* **22**, 379–392 (2021).

METHYLATION OF DNA IS OFTEN ASSOCIATED WITH DECREASED TRANSCRIPTION

In addition to modification of histones, DNA itself can be epigenetically modified via methylation of cytosine. Cytosines in DNA can be methylated at position 5 of the base. The methylation does not change the underlying sequence of the DNA: Methylated cytosines still base pair with guanine, because the base-pairing parts of the molecule are unchanged. This is illustrated in **Figure 7**. But the methylation changes the part of the base that is facing the major groove, so methylation affects how transcription factors can interact with the major groove of DNA.

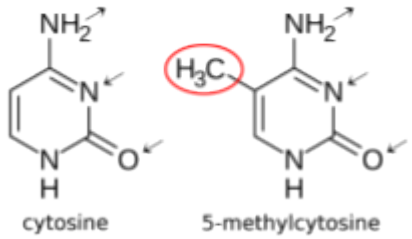


Figure 7. Comparison of cytosine and 5-methylcytosine. The arrows indicate the H-bond donors and acceptors that participate in base pairing with guanine.

Methylation of cytosines typically occurs at specific positions within the genome. Potential methylation sites usually have a cytosine immediately followed by a guanine. These are called **CpG** sites, where the p represents the phosphate between the two sugars. This is a palindromic sequence, with the complementary strand also having the sequence CpG. Usually both strands are methylated, if a site is methylated at all, as shown in **Figure 8**. Regions of the genome with many clustered CpG sites known to function in gene regulation are sometimes called **CpG islands**.

Methylation of DNA itself is generally associated with decreased transcription regulation, including the silencing of highly methylated regions of the genome. The converse statement is also true: transcriptionally silenced regions of the genome tend to be highly methylated.

As with histone modification, DNA methylation is maintained after replication. Immediately after replication, the daughter double helix is methylated only on one strand. But DNA methyltransferases can recognize this **hemimethylated** DNA and modify the second strand to match. This is shown in **Figure 8**.

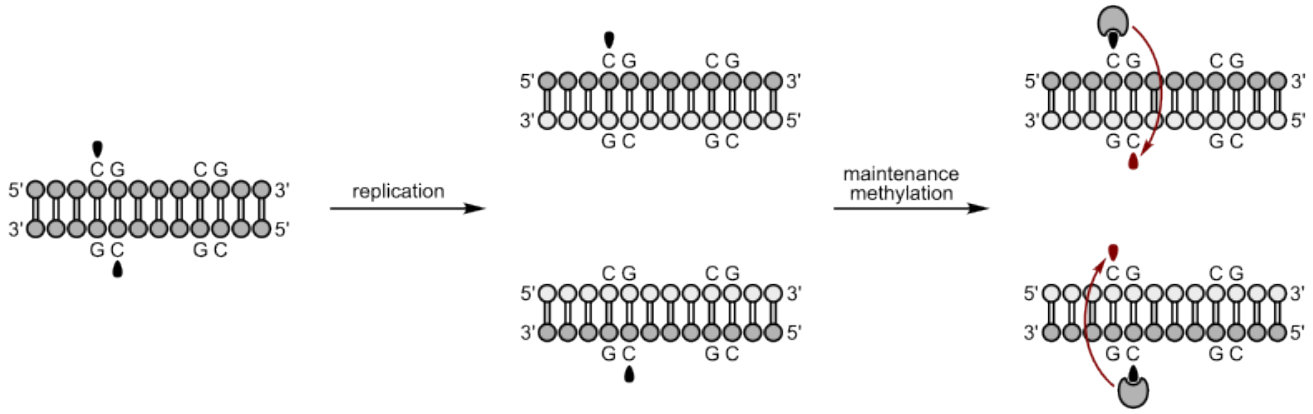


Figure 8. Methylation of DNA is maintained after replication. Newly replicated DNA is hemimethylated, methylated only on one strand of the CpG site. DNA methyltransferases add a methyl group to the second strand, matching the parental methylation pattern.

Enzymes called **DNA methyl transferases** catalyze the transfer of methyl groups from the methyl donor SAM — the same substrate that donates methyl groups to histones. DNA methylation is more stable than histone methylation: demethylation seems to be relatively rare. Methyl groups can be passively eliminated from DNA if daughter strands are not methylated following DNA replication. Methyl groups can also be eliminated from DNA via a series of enzymatic reactions that ultimately remove the methylated cytosine and replace it with an unmethylated cytosine.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=484#h5p-88

Media Attributions

- Cytosine and 5-methylcytosin © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Maintenance methylation © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

EPIGENETICS IN ACTION

The examples of epigenetic modification described above – DNA methylation and histone acetylation and methylation – are all heritable epigenetic changes that do not affect the DNA sequence. The epigenetic marks influence gene expression in a manner that is stable through cell division. The gene expression status, in turn, affects the phenotype of a cell (and the organism). The epigenetic status is maintained through mitosis, affecting the fate of somatic cells that share a parent cell. In some cases, the epigenetic status is maintained through meiosis, affecting the phenotype of offspring.

Although not discussed in this text, there are additional mechanisms of epigenetic inheritance. These include other forms of histone modification (including phosphorylation, ubiquitination, and citrullination) that contribute to the histone code. Other molecules in the cell – including non-coding RNAs – play an important role in regulating epigenetic silencing.

Epigenetics during development

Most, but not all, epigenetic marks are removed in the zygote. So as the embryo begins to develop, it starts with almost a clean slate, epigenetically. This likely contributes to the **totipotency** of the early embryonic cells. Totipotency describes the ability of a cell to divide and give rise to any other cell type. The cells are called totipotent **stem cells**. But as the fertilized egg divides, daughter cells become committed to a particular cell fate, beginning 4-5 days post-fertilization with the differentiation of trophoblast cells and embryonic stem cells of the inner cell mass (shown in yellow and red, respectively, in **Figure 9**). This commitment is correlated with a loss of some potential: the trophoblast cells develop into much of the placenta but cannot develop into cells of the embryo proper. Likewise, the embryonic stem cells of the inner cell mass cannot give rise to the trophoblastic structures. These embryonic cells are thus called **pluripotent** rather than totipotent.

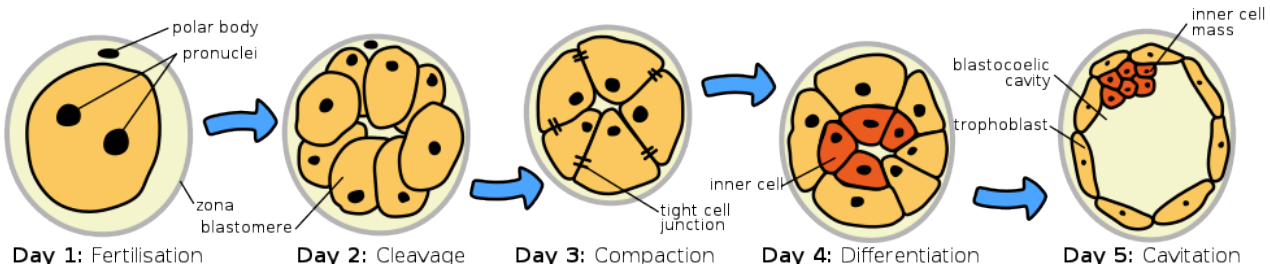


Figure 9. Early stages of mammalian embryogenesis, beginning with the zygote on Day 1. The yellow cells in Day 4 and Day 5 (labeled "trophoblast" will eventually give rise to structures that will support the growing embryo, including part of the placenta. The red cells labeled "inner cell mass" will continue to divide and develop into the embryo proper.

The loss of potential is associated with epigenetic silencing of certain parts of the genome[1], as are subsequent stages of differentiation. For example, genes called Nanog and Oct-4 play a role in "stemness." Histones around these genes are acetylated in embryonic stem cells, but in trophoblast cells, the DNA around the genes is hypermethylated[2]. As cells of the inner cell mass divide and commit to a particular cell fate, additional epigenetic modifications occur.

We've already seen an example of epigenetics in the module on Sex: X-chromosome inactivation is a form of epigenetic inheritance. In individuals with more than one X chromosome (most commonly XX genotypes, but also XXX, XXY, or other less common genotypes), only one X chromosome is typically active in each cell. The additional X chromosome(s) are highly condensed, and most (but not all) of genes are transcriptionally inactive. The condensed X chromosome is visible under the microscope as a structure called a **Barr body**, which is a densely staining region within the nucleus. An example of a Barr body stained is shown in **Figure 10**.

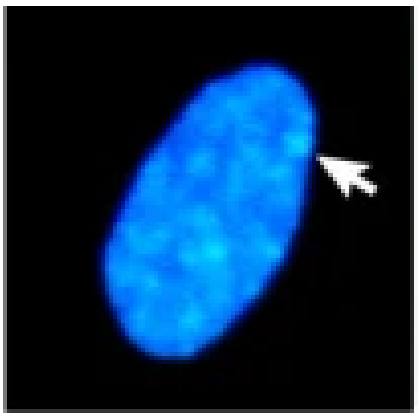


Figure 10. The nucleus of a cell stained with DAPI, with the Barr body indicated by an arrow. DAPI is used to stain DNA in fluorescence microscopy.

Which X chromosome is inactivated occurs randomly in individual cells early in embryogenesis. However, the inactivation is inherited, with the same X chromosome consistently inactive in daughter cells after cell division. X-inactivation depends on the action of a long non-coding RNA molecule called Xist. Xist molecules coat the to-be-inactivated X chromosome, recruiting additional protein factors and ultimately leading to the remodeling of the X chromatin. The inactivated X chromosome shows high levels of DNA methylation around the promoters of silenced genes as well as low levels of histone acetylation[3].

The phenotype of a calico cat is a visually striking example of how silencing of part of the genome is inherited through replication and mitosis: the patches of color result from epigenetic silencing of one or the other of the X chromosomes in a cell with XX genotype. Patches include cells that were derived from the same parent cell.

On average, because the X chromosomes are inactivated randomly, about 50% of the cells in the body of an XX individual will be inactivated for each X chromosome. But in some individuals, X inactivation is skewed, with more than 75% of cells inactivating the chromosome inherited from one parent. Interestingly, this appears to play a role in the variable expressivity of some X-linked genetic traits, including red-green colorblindness

1. Xu, R., Li, C., Liu, X. & Gao, S. Insights into epigenetic patterns in mammalian early embryos. *Protein Cell* **12**, 7–28 (2021).

2. Hattori, N. *et al.* Epigenetic regulation of Nanog gene in embryonic stem and trophoblast stem cells. *Genes Cells* **12**, 387–396 (2007).

3. Fang, H., Disteche, C. M. & Berletch, J. B. X Inactivation and Escape: Epigenetic and Structural Features. *Front. Cell Dev. Biol.* **7**, (2019).

and diseases like Hemophilia B, Duchenne muscular dystrophy, and Fabry syndrome. The skewed inactivation of a healthy allele increases the severity of the disease, and the skewed inactivation of a disease-associated allele lessens the severity and reduces symptoms[45].

Imprinting

Imprinted genes are differently expressed depending on the parent-of-origin for each allele. For some imprinted genes, only (or almost only) the maternal copy of a gene is expressed, and for others, only (or almost only) the paternal copy is expressed. The other allele was silenced with DNA methylation and histone modification.

For most genes, epigenetic marks are cleared in the zygote. But for imprinted genes, epigenetic marks are reprogrammed in the egg or the sperm, and the egg- or sperm-specific epigenetic marks are maintained in the zygote beyond the stage of development when most other epigenetic marks are removed. Imprinting is maintained throughout the lifespan and is only reset during gametogenesis when new egg- or sperm-specific epigenetic marks are reprogrammed for the next generation. **Figure 11** shows this for a maternally imprinted gene. Paternal imprinting is similar but with the paternal allele silenced.

The phenomenon of imprinting appears to have evolved independently in flowering plants and mammals; about 1% of human genes are imprinted.

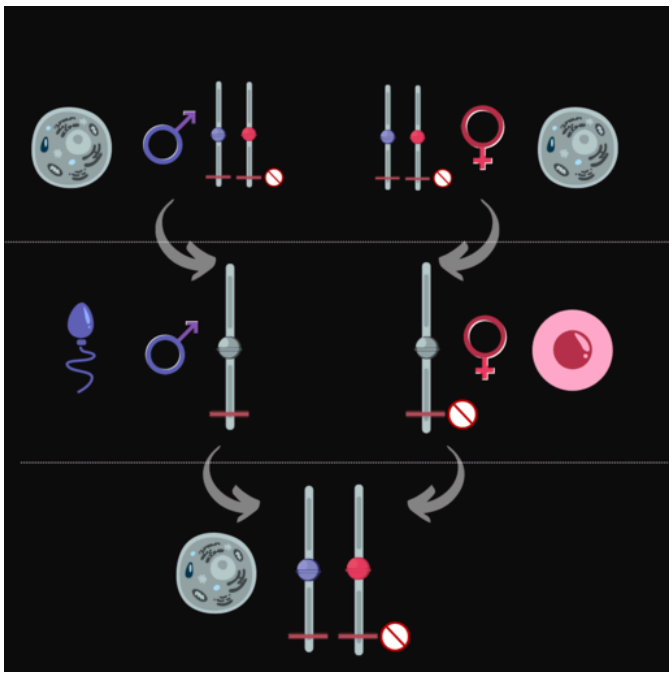


Figure 11. Image depicting the process of maternal imprinting. In the top panel, chromosome pairs from male and female parent are shown with the location of an imprinted gene indicated by a red horizontal hash mark. In the middle panel, each parent has donated one chromosome to the egg or sperm. All epigenetic marks are removed in the sperm, but the egg imprints its chromosome. In the bottom panel, this results in the zygote only expressing the imprinted gene from the paternal chromosome.

4. Sun, Z., Fan, J. & Wang, Y. X-Chromosome Inactivation and Related Diseases. *Genet. Res.* **2022**, 1391807 (2022).

5. Jørgensen, A. L. *et al.* Different patterns of X inactivation in MZ twins discordant for red-green color vision deficiency. *Am. J. Hum. Genet.* **51**, 291–298 (1992).

Many imprinted genes appear to be involved in growth rate, metabolism, brain function, and behavior, with imprinted genes linked to phenotypes like body size and behaviors like risk-taking and impulsive behavior[6].

In humans, several genetic diseases are caused by a disruption in imprinting. For example, Beckwith-Wiedemann syndrome is caused by a disruption in imprinted genes on Chromosome 11. These genes include the growth factor IGF2, which is normally only expressed from the paternal chromosome. Altered DNA methylation, duplication of the paternal chromosome, and other changes result in two copies of the gene being expressed. Among other signs associated with the syndrome, children with Beckwith-Wiedemann syndrome tend to be larger in size than their same-age peers[7].

Angelman syndrome is linked with a loss of function of the gene UBE3A on Chromosome 15. UBE3A is imprinted and in cells of the central nervous system, only the maternal copy of the gene is expressed. Children with Angelman syndrome show intellectual disability, small head size, and, often, a happy and excitable personality. These traits are linked with a deletion or loss of function in the maternal chromosome, which leaves patients without a functional copy of the gene in cells of the central nervous system. Mutations in the paternal allele will not cause Angelman syndrome[8].

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=492#h5p-89

Epigenetic changes are influenced by behavior and environment

Over the course of a lifetime, diet, behavior, and environment contribute to the accumulation of epigenetic marks. For example, some nutrients in the diet are a source of methyl groups that are eventually part of S-adenosyl methionine (SAM/SAMe), which in turn donates methyl groups to methylate cytosine or methylate

6. Tucci, V. *et al.* Genomic Imprinting and Physiological Processes in Mammals. *Cell* **176**, 952–965 (2019).

7. Beckwith-Wiedemann syndrome: MedlinePlus Genetics. https://medlineplus.gov/genetics/condition/beckwith-wiedemann-syndrome/.

8. Angelman syndrome: MedlinePlus Genetics. https://medlineplus.gov/genetics/condition/angelman-syndrome/.

histones[9]. Nutrients like folate and choline are examples, and SAMe itself can be purchased as a dietary supplement. Other nutrients may promote or inhibit the action of DNA methyltransferase or histone modifiers. Examples include molecules like resveratrol (found in red wine, activates HDACs), curcumin (found in turmeric, inhibits HATs), and diallyl disulfide (found in garlic and onions, inhibits HDACs)[10]. Stress, drug use, trauma, and lifestyle choices have also been shown to impact the epigenome.

Environmental effects can add up over time. Identical twins, who share 100% of their DNA, are phenotypically similar but not actually indistinguishable from one another. Identical twins tend to have very similar epigenomes early in life. However, as identical twins age, their genomes stay the same, but their epigenomes diverge. This likely contributes to identical twins becoming more dissimilar as they age[11].

These environmental effects on the epigenome likely influence the incomplete penetrance of some genetic traits. For example, the methylation of a gene called REELIN appears to be linked with the development of schizophrenia. Schizophrenia is a psychiatric disorder that manifests in early adulthood. The phenotype shows a strong genetic component, but twin studies reveal that less than 50% of identical twins are concordant for schizophrenia. Differences in methylation status between twins may account for at least some of this difference[12].

Epigenetic changes are also linked with cancer. Loss-of-function mutation of tumor suppressor genes contributes to oncogenesis. However, in some cases, the DNA sequence of tumor suppressor genes is unchanged, but the gene is highly methylated, resulting in loss of function without underlying mutation. Acetylation of genes that drive the cell cycle can play a role in the uncontrolled growth associated with cancer. A number of HDAC inhibitors have been approved for the treatment of cancer[13].

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it

9. Tiffon, C. The Impact of Nutrition and Environmental Epigenetics on Human Health and Disease. *Int. J. Mol. Sci.* **19**, 3425 (2018).

10. Davis, C. D. & Ross, S. A. Dietary Components Impact Histone Modifications and Cancer Risk. *Nutr. Rev.* **65**, 88–94 (2008).

11. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci.* **102**, 10604–10609 (2005).

12. Imamura, A. *et al.* Genetic and environmental factors of schizophrenia and autism spectrum disorder: insights from twin studies. *J. Neural Transm.* **127**, 1501–1515 (2020).

13. Yoon, S. & Eom, G. H. HDAC and HDAC Inhibitor: From Cancer to Cardiovascular Diseases. *Chonnam Med. J.* **52**, 1–11 (2016).

online here:
https://rotel.pressbooks.pub/genetics/?p=492#h5p-90

Transgenerational epigenetics

The examples discussed above all involve mitotic inheritance of epigenetic marks, with maintenance of the epigenome in somatic cells of a single individual. As was discussed earlier, most epigenetic marks are erased during gametogenesis or in the fertilized egg (zygote). However, some epigenetic marks escape this zygotic reprogramming. These allow epigenetic information to be passed through meiotic cell division from parent to offspring – a mechanism for Lamarckian inheritance of acquired traits.

There are many known examples of transgenerational epigenetic inheritance in non-mammalian species, where the epigenome can be manipulated and tracked for many generations. For example, in the plant Arabidopsis thaliana, a mustard weed used as a model organism in the genetics lab, DNA methylation induced in one generation can be measured for at least 8 generations of offspring[14].

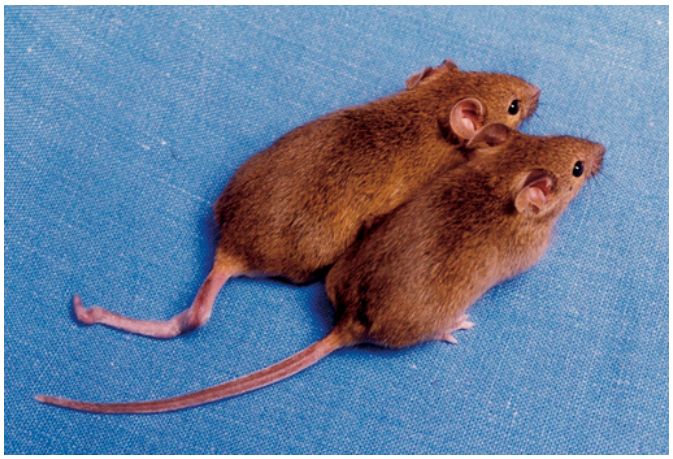


Figure 12. The kinked tail phenotype of some mice is linked with DNA methylation of the AxinFu gene. Both mice in this image have the same genotype.

In mice, a "tail kink" phenotype (**Figure 11**) is associated with a mutation of a gene called $Axin^{Fu}$, but whether the tail kink phenotype is penetrant depends on the methylation status of the gene. The phenotype can be inherited across generations[15].

In contrast, in humans, many of the traits suspected to be affected by epigenetics are likely multifactorial, and causative genes may or may not yet be linked with the phenotype. Many studies are retrospective analyses of human populations, tracking traits associated with environmental factors over the course of several generations.

Some of the most commonly cited of these retrospective studies track the descendants of individuals who

14. Fitz-James, M. & Cavalli, G. Molecular mechanisms of transgenerational epigenetic inheritance. *Nat. Rev. Genet.* **23**, 325–341 (2022).

15. ibid

were exposed to famine *in utero.* These studies include the Dutch famine winter of 1944-1945, food scarcity in the Överkalix municipality in Sweden, the Ukrainian famine of 1932-1933, and the Chinese famine of 1959-1961.

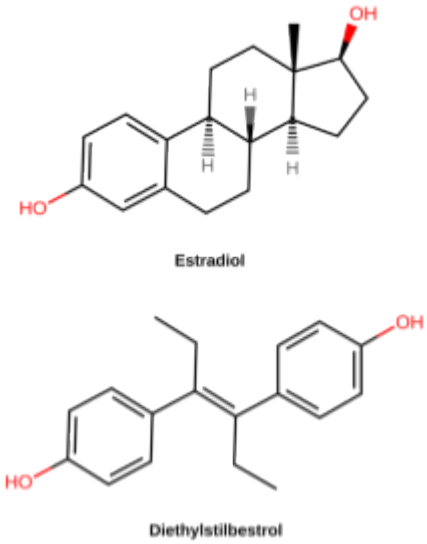


Figure 13. Diethylstilbestrol (DES) is chemically similar to estradiol, the main form of estrogen used in the human body. DES was historically used to prevent miscarriage in pregnant women. It was also used as a drug to treat prostate cancer, symptoms of menopause, and other estrogen-responsive medical conditions. However, DES use is linked with certain kinds of cancers and reproductive tract abnormalities.

Exposure to famine as a fetus increased the risk for obesity and other metabolic disorders, Type 2 diabetes, schizophrenia, and depression. It also altered the risk of disease in their children and grandchildren[16].

Other large-scale retrospective studies track the descendants of women treated with diethylstilbestrol (DES) during pregnancy. DES is a synthetic estrogen that was used to prevent miscarriage in pregnant women in the 1940's-1960's (**Figure 12**). By the 1970s, however, studies had shown that DES was not effective in preventing miscarriage and, even worse, was causing vaginal and breast cancers and reproductive anomalies in "DES daughters" – children born to women treated with DES. Follow-up studies have shown that DES grandchildren are also affected: DES grandchildren born to DES daughters are more likely to be born premature, and DES grandsons born to DES daughters appear to have an increased risk of certain anomalies of the reproductive tract. It is not clear whether the grandchildren born to DES sons are similarly affected[17].

16. Vaiserman, A. & Lushchak, O. Prenatal famine exposure and adult health outcomes: an epigenetic link. *Environ. Epigenetics* **7**, dvab013 (2021).

17. Zamora-León, P. Are the Effects of DES Over? A Tragic Lesson from the Past. *Int. J. Environ. Res. Public. Health* **18**, 10309 (2021).

Multigenerational retrospective studies like the famine and DES studies can be very powerful in linking phenotypes to environmental stressors. And, with modern genome sequencing tools, both mutations and changes to DNA methylation can be compared across generations. However, it can be difficult to distinguish between true transgenerational epigenetic inheritance and the direct effects of environmental exposure.

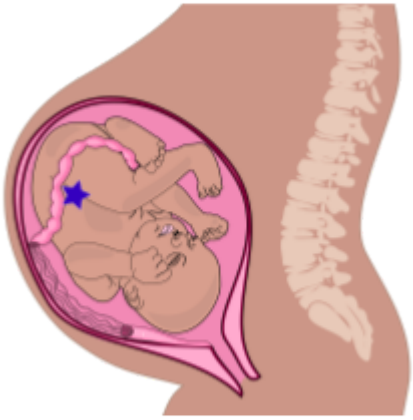


Figure 14. Three generations are shown in this image: Parent, child (fetus), and grandchildren (with the star representing the eggs of the fetus)

You may have heard that when a baby is born, ovaries already contain all the eggs the baby will have in their lifetime (this is not true for sperm, which are produced lifelong after puberty). Those future egg cells form very early during fetal development: as early as 8 weeks post-fertilization. So as soon as those primitive ovaries form during development, exposure during pregnancy potentially exposes three generations to a stressor: the pregnant parent (P or F0 generation), the fetus (F1 generation), and, potentially, the future eggs of the fetus (F2 generation). This is illustrated in **Figure 13**, which shows three generations in the same image: parent, fetus, and the eggs of the fetus, illustrated by the blue star.

Because of this, to show true transgenerational epigenetic inheritance – with the maintenance of epigenetic marks through meiosis – retrospective studies of environmental stress must track great-grandchildren from pregnant women and grandchildren from non-pregnant women and/or men.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=492#h5p-91

Media Attributions

- HumanEmbryogenesis © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- DAPI © BMC Biology via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Imprinting maternal © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Cloned mice © Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license
- Diethylstilbestrol © Medgirl131 via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike)

license

- Placenta system © Modified from Wilfredor, via. Wikipedia is licensed under a CC BY-SA (Attribution ShareAlike) license

SUMMARY

Epigenetic inheritance refers to heritable characteristics that are controlled by changed in gene expression rather than differences in underlying DNA sequence. Epigenetic inheritance often depends on changes in chromatin state that can be transmitted through mitosis to daughter cells. Molecularly, this may depend on modifications of histone tails – acetylation and methylation were discussed in this chapter, although there are other covalent modifications seen in chromatin. Acetylation of histones is associated with active transcription of nearby genes, while methylation has more varied effects. Methylation of cytosines at CpG sequences in the genome is also associated with decreased transcription of nearby genes.

These epigenetic marks may be inherited through mitosis, and epigenetic silencing of genes plays a role in the development and aging of somatic cells. Accumulation of epigenetic marks happens over the course of a lifetime and is influenced by environment, diet, behavior, and lifestyle. These epigenetic marks are usually, but not always, cleared and "reset" in gametes and/or the fertilized embryo, and new marks are acquired as the cells divide and age. However, some marks persist through meiosis and fertilization. These persistent epigenetic marks include but are not limited to imprinted genes, which are genes that are only, or mostly, expressed from either the paternal or maternal chromosome rather than both chromosomes equally. Transgenerational epigenetic inheritance may influence phenotypes like growth and metabolism, schizophrenia, and depression.

WRAP-UP QUESTIONS

1. Early evolutionary biologists like Jean-Baptiste Lamarck hypothesized that traits acquired over the course of a lifetime could be inherited by offspring, thus driving evolution. Explain how Lamarckian evolution differs from the theory of natural selection proposed by Charles Darwin.

2. Using online resources, find two examples of how epigenetic changes during a person's lifetime affect human health. Are your examples of histone modification, DNA methylation, or a mechanism not discussed in this text?

3. Using online resources, find two examples of how epigenetics influences gene expression in specific cell types during embryonic or fetal development. Are your examples of histone modification, DNA methylation, or a mechanism not discussed in this text?

4. Describe how epigenetic changes might explain the incomplete penetrance or variable expressivity of certain genetic traits.

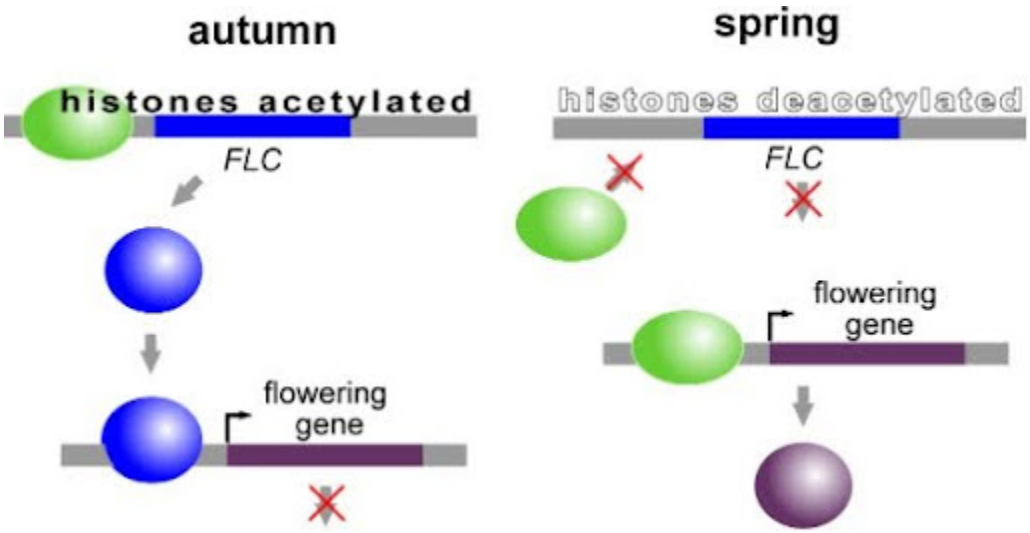


Figure 15. In the autumn, histones associated with the gene FLC are acetylated, allowing this repressor of flowering genes to be expressed. During winter, enzymes progressive deacetylate FLC, preventing it from being expressed, and therefore allowing flowering genes to respond to other signals that induce flowering.

5. Adapted from *Online Open Genetics* (Nickle and Barrett-Ng)[1]: Epigenetics plays a role in the vernalization

1. Nickle and Barrette-Ng. Open Online Genetics. in *Open Online Genetics* (2016).

of flowering plants – the process by which certain winter annual plants only begin flowering after a prolonged exposure to cold temperatures. This process is controlled by expression of a gene called FLC, which encodes a repressor of genes required for flowering. In the fall and early winter, the histones around the FLC gene are highly acetylated. With prolonged exposure to the cold, histones are gradually deacetylated, reducing transcription of the FLC gene. Without expression of FLC, the plant can then respond to pro-flowering hormones and signals, inducing flowering at the proper time in the spring after cold temperatures have passed. This is illustrated in **Figure 15**.

What do you think might be the phenotype of a winter annual plant that lacked HDAC function? Explain your reasoning

6. Oncogenesis, or the development of cancer, results when a single cell sustains mutations in tumor suppressors and proto-oncogenes. Tumor suppressor proteins are responsible for protecting somatic cells against becoming cancerous, and most cancer cells have loss-of-function mutations in one or more tumor suppressors.

Loss-of-function mutations in the BRCA1 tumor suppressor gene are linked with breast cancer: up to 70% of women with one loss-of-function allele in either gene will eventually get breast cancer, and their tumors typically show that the second allele has sustained a somatic mutation. This makes such tumor cells homozygous for BRCA1 mutations.

But less than 10% of breast cancer patients have inherited such mutations. Instead, in about 30% of breast cancer patients, tumors show hypermethylation in the region around the BRCA1 gene. What is the most likely consequence of that hypermethylation on BRCA1 expression, and what impact did it likely have on oncogenesis? Explain your reasoning.

7. Hemophilia B, a disorder characterized by excessive bleeding, is caused by a loss of function in factor IX clotting activity. Factor IX is encoded by a gene on the X chromosome. Chromosomal males (XY) who inherit one loss of function allele or chromosomal females (XX) who inherit two loss of function alleles have impaired clotting ability. Heterozygous females typically have lower levels of circulating Factor IX and mild hemophilia but are largely unaffected by the health problems caused by excessive bleeding. How would you predict the phenotype of heterozygous females to be affected by skewed X-inactivation? Explain your reasoning.

Science and Society

8. An increasing body of evidence supports the role of both genetics and epigenetics in human phenotypes like sexual orientation and gender identity. Studies have shown a link between epigenetic changes in gene expression established early in embryonic development and sexually dimorphic brain structures[2].

2. Swaab, D. F., Wolff, S. E. C. & Bao, A.-M. Chapter 31 – Sexual differentiation of the human hypothalamus: Relationship to gender identity and

These studies overwhelmingly validated the experiences of gay and transgender people who understand that their gender identity or sexual orientation is innate and not a matter of social conditioning. However, genetic studies looking for differences between groups of people (often racial or ethnic groups) have also historically been misused to mark groups of people as different or socially inferior in some way, for example connecting brain size to intelligence, despite no evidence to show such a connection.

Explain why epigenetics is rooted in biology as much as genetics. In your opinion, what are the advantages and risks of such studies on sex and gender? How might such information be used for or against the affected community? Use outside resources as appropriate to answer this question and support your argument.

9. The cloning of adult organisms is done through **somatic cell nuclear transfer (SCNT).** Somatic cell nuclear transfer is the process by which the nucleus from an adult somatic cell is transferred to an egg that has had the nucleus removed. The re-nucleated egg is then induced to divide and form an embryo, which can be implanted into a surrogate mother for gestation. Given what you learned about epigenetics and development, how might the epigenetic markings of the donor nucleus impact the fetus resulting from SCNT? What ethical issues should be considered before cloning of mammals should be performed? Use outside resources as appropriate to answer this question and support your argument.

---

sexual orientation. in *Handbook of Clinical Neurology* (eds. Swaab, D. F., Buijs, R. M., Lucassen, P. J., Salehi, A. & Kreier, F.) vol. 181 427–443 (Elsevier, 2021).

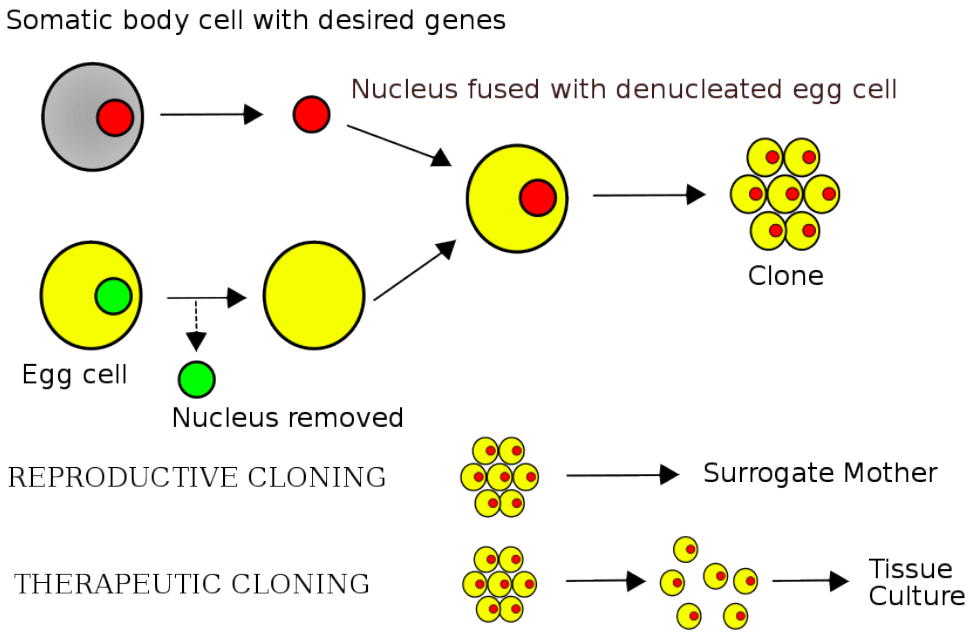


Figure 16. Reproductive cloning via somatic cell nuclear transfer.

Media Attributions

- Autumn vs spring © Online Open Genetics, Nickle and Barrett-Ng. Origianl-Deyholos is licensed under a CC BY-SA (Attribution ShareAlike) license
- Cloning diagram

PART XVII
EVOLUTION AND ANCESTRY

Learning Objectives

Objectives

1. Recognize that evolution provides a framework for how geneticists view biology.
2. Explain the role mutation and variation play in evolution
3. Describe the processes by which evolution occurs, including genetic drift, gene flow, assortative mating, and natural selection.
4. Explain how molecular genetics is used to deduce evolutionary relationships, recognizing that DNA sequence, protein sequence, and chromosome structure can all offer evolutionary clues
5. Define: mutation, molecular clock, bottleneck, founder effect, assortative mating, haplotype, ortholog, paralog

Introduction

The evolutionary biologist Theodosius Dobzhansky said in 1973, "Nothing in biology makes sense except in light of evolution."[1]This oft-quoted statement describes biologists' understanding of the relationship among organisms on Earth. Evolution gives geneticists a lens for studying the mechanisms of inheritance. We see this in many of the fundamental concepts in this text:

- In the chapters on DNA structure and molecular genetics, we see that all organisms use the same

1. Dobzhansky, T. Nothing in Biology Makes Sense except in the Light of Evolution. *Am. Biol. Teach.* **35**, 125–129 (1973).

biomolecules and the same genetic code. The mechanisms of molecular genetics – replication, transcription, translation, and control of gene expression – are likewise shared among all organisms. With the understanding that these processes are shared among all organisms because all organisms share a common ancestor, biologists can study these processes in one model organism and gain an understanding of how these properties work in all organisms. See Figure 1.

- In the chapter on Linkage, we discussed **linkage disequilibrium** and how parts of the genome are inherited as a unit, because there has not been enough evolutionary time to separate SNPs within a haplotype or because the area of the genome is under selection.
- In the chapter on Gene Expression in Eukaryotes, we looked at commonalities among the enhancers of genes important during development. Differences in an enhancer in snakes result in absent limbs, suggesting a mutation occurred in a snake ancestor after snakes diverged evolutionarily from other vertebrates. This is an example of how genetic variation can aid in our understanding of speciation, or the creation of new species over time.
- In the chapter on DNA Repair and Cancer, we see natural selection and microevolution on a cellular scale. Mutations in cancer cells give them a growth advantage over their healthy surrounding tissue, so the cancer cells overrun the tissue or organ. Duplication of the p53 tumor suppressor gene allowed the evolution of increased body size in elephants without a concomitant increase in the incidence of cancer.

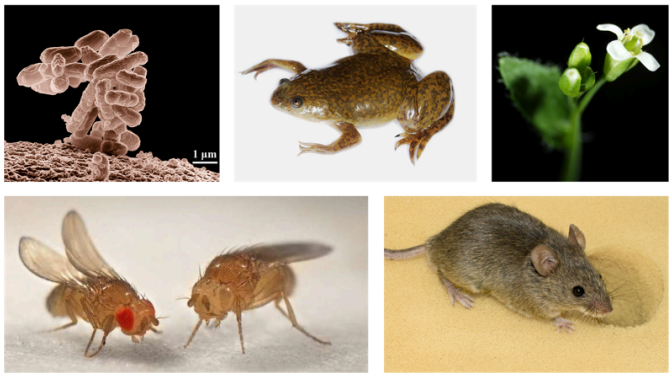


Figure 1. Model organisms. Clockwise from top left: Escherichia coli, Xenopus laevis, Arabidopsis thaliana, Mus musculus, Drosophila melanogaster. Many of the central dogma mechanisms were first studied in E.coli. Xenopus eggs have been used to study cell cycle regulation. Arabidopsis has been used as a model plant. Drosophila were used in early genetic studies on linkage and later work on gene expression during development. Mice share about 85% of their genetic sequence with humans and are a model mammal used in medical research.

Microevolution refers to measurable changes in population over time. These changes don't create new species, but they can change the appearance or behavior of a population. We can often observe microevolution in action, either in the field or in the lab. We can also observe microevolution on a cellular level. The growth of a tumor, and its shrinkage following exposure to chemotherapy, are examples of microevolution on a cellular level.

Macroevolution refers to the accumulation of changes over a longer period of time. This includes significant enough changes to result in the creation of new species. Because macroevolution requires change over a long time, we typically cannot observe or measure macroevolution directly. Historically, macroevolution has been studied through comparison of homologous and analogous structures in different species, with dating dependent on the fossil record. In the case of human evolution, archaeological remains can offer clues as well.

But in the last 20-30 years, the advent of molecular genetics and the genomic age have given us a way to study evolution on a molecular level.

In this chapter, we'll look at the evidence for and mechanisms of evolution from a genetic perspective. We'll look at examples of both microevolution and macroevolution, using humans and other organisms as examples.

Media Attributions:

Ecoli: By Photo byfkfkrErbe, digital colorization by Christopher Pooley, both of USDA, ARS, EMU. – This image was released by the Agricultural Research Service, the research agency of the United States Department of Agriculture, with the ID K11077-1 (next)., Public Domain, https://commons.wikimedia.org/w/index.php?curid=958857 Xenopus: By Brian Gratwicke – Flickr: Xenopus laevis, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=23752908 Arabidopsis: By I, Suisetz, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2448732 Mouse: By George Shuklin (talk) – Own work, CC BY-SA 1.0, https://commons.wikimedia.org/w/index.php?curid=5521043 Drosophila: By Fruit fly (Drosophila melanogaster), CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=112257273

Media Attributions

- Model organisms is licensed under a CC BY (Attribution) license

EVOLUTION IS CHANGE IN A POPULATION OVER TIME

In casual conversation, we might use the word evolution to describe change. I might say that my opinions have evolved over time, for example. But for a biologist, evolution refers to change not in an individual but in a population.

An example from humans: people who live in areas close to sea level may experience altitude sickness if they travel to regions at high elevation. Symptoms of altitude sickness include headache, nausea, and dizziness, among others. These symptoms are largely a result of difficulty in absorbing oxygen: although the percentage of oxygen in the air is similar at low and high elevations, the overall air pressure is lower, resulting in a **hypoxic** environment. (**Hypoxia** refers to conditions of low oxygenation.)

Initially, a person's body may compensate for high altitude with increased breathing and heart rates. But in the first few days at high altitude, a person's body begins to acclimate: more red blood cells are produced, increasing hemoglobin concentration, and the lungs may increase in capacity. Because these same oxygen-carrying adaptations confer a benefit to athletes, many training centers for elite athletes are located at high elevations.

These physiological changes in individual experiences are acclimatization but not evolution. These changes gradually revert if the person heads back to lower elevations and there is no underlying genetic change.

In contrast, if we look at populations of people whose ancestors have lived at high elevations, we do see examples of evolution in action. The Sherpa are a group of people who live in the Khumbu region of Nepal, 3,000+ meters (11,000 feet) above sea level, near Mount Everest. The Sherpa are renowned for mountaineering skill and often serve as guides for other Mount Everest climbers. Tenzing Norgay was one of the first two climbers to have been recorded in summitting Mount Everest (with Edmund Hilary). And as of 2024, of the 25 climbers who have each summitted Everest more than 10 times, all but 5 are from Nepal and of Sherpa ancestry. (If you're curious, you can find an unofficial list of these mountaineers at Wikipedia.)

Figure 2. Map of Nepal. Image source: By Globe-trotter – Own work based on Perry-Castañeda Map Collection, University of Texas Libraries (public domain), CC BY-SA 4.0,

The Sherpa share recent ancestry with ethnic Tibetans whose ancestors have lived at altitude for 25,000-40,000 years. The Sherpa and Tibetans share genetic variants in genes that participate in the Hypoxia-Induced Factor (HIF) pathway. These variants affect hemoglobin concentrations in ways presumed to give a physiological advantage at high elevation[1]. Their frequency in Sherpa and Tibetan populations are thought to be a result of **selective pressure** caused by the hypoxic environment. In other words, those ancestors with these particular genetic variants were the most successful in reproducing in this environment.

Test your understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=692#h5p-96

1. Bhandari, S. & Cavalleri, G. L. Population History and Altitude-Related Adaptation in the Sherpa. Front. Physiol. 10, (2019).

Media Attributions

- Fig 2 Map of Nepal © Perry-Castañeda Map Collection, University of Texas Libraries (public domain) adapted by Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

MUTATION AND VARIATION ARE REQUIRED FOR EVOLUTION

Evolution is dependent on genetic variation in a population. With evolution, the overall frequency of variants in a population changes over time. Without genetic variation, there's no change.

Where does genetic variation come from? As discussed in the chapters on Mutation and DNA repair, genetic variation arises from mutation. Mutation arises naturally through unrepaired DNA damage. The rate at which this occurs in an individual depends on species (for example, viruses tend to have very high rates of mutation due to error-prone polymerases) and environmental influences (like exposure to high levels of radiation). Remember that mutations can be advantageous ("good"), disadvantageous ("bad"), or neutral. Variants can include single nucleotide polymorphisms (SNPs), larger insertions or deletions, copy-number variants, and structural variants that rearrange larger parts of the genome.

Most new variants disappear quickly from a large population, regardless of whether they are advantageous, disadvantageous, or neutral. This is for several reasons. A new mutation will typically only be present in one copy of a diploid genome, so only one-half of offspring from a mutant individual will harbor the mutation; there's no guarantee that any one individual will produce offspring, and most reproducing individuals do not harbor the mutation.

But sometimes, a new allele may begin to accumulate in a population. Eventually, the new allele may become fixed in a population. In the section on population genetics, we discuss Hardy-Weinberg equilibrium. At equilibrium, allele frequency does not change from one generation to the next (so no evolution). But for allele frequency to be consistent from generation to generation, some conditions must be met:

1. The alleles are not accumulating mutations (converting one allele to the other, for example)
2. There must be a large population
3. No migration in or out of the population
4. No selective pressure
5. Individuals choose mates randomly

When those conditions are not met, mutation, genetic drift, gene flow, natural selection, and assortative mating can influence changes in allele frequencies in the population. Each of these is a factor in the evolution of a population. They are discussed in the next section.

Test your understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=597#h5p-109

MECHANISMS OF EVOLUTION

In addition to mutation, genetic drift, gene flow, natural selection, and assortative mating are factors that influence the evolution of a population.

Natural selection

Natural selection is probably what many people think of as driving evolution. This is Charles Darwin's proposal for "survival of the fittest", where "fitness" is determined by the likelihood of reproducing. Selection results when one variant confers some sort of advantage over another. An individual with the advantageous variant is more likely to reproduce, contributing the advantageous allele to the gene pool of the next generation (**Figure 3**). An individual with the disadvantageous variant is less likely to reproduce, so the disadvantageous allele is likely to make up a smaller percentage of the gene pool.

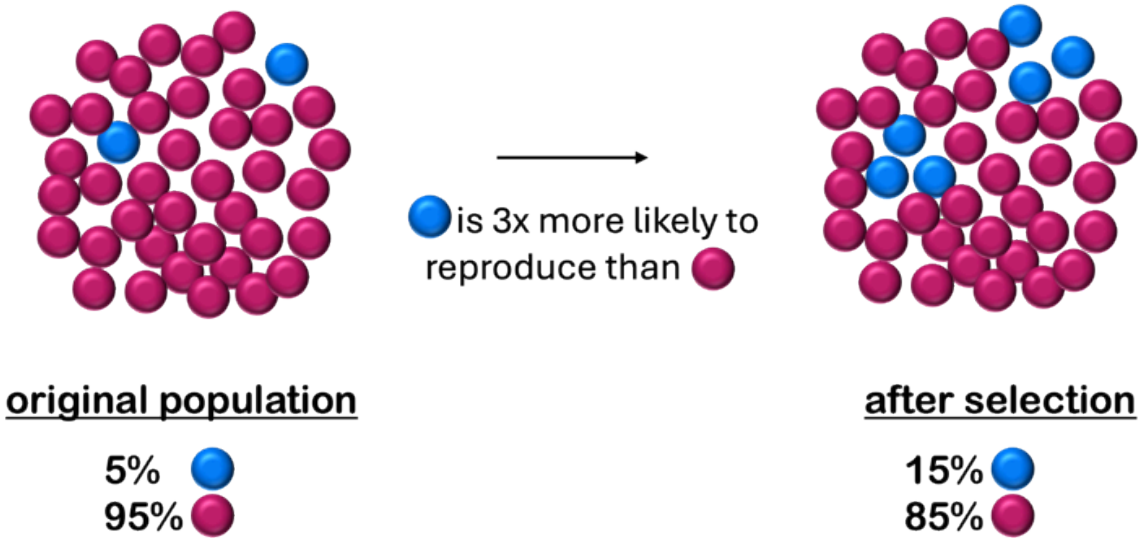


Figure 3. Natural Selection

On a molecular level, this is the reason why certain sequences in the genome are more conserved than others. For example, the function of many of the consensus sequences we described in the molecular genetics modules depend on DNA: protein interactions. Any changes to those sequences would reduce function and presumably negatively impact the fitness of the organism. So, any individual who acquires a mutation in a consensus sequence may be less likely to reproduce. Such conserved regions of the genome are typically under great selective pressure against mutation.

Note that in populations of diploid organisms, it is not always a homozygote that will have an advantage! This, called the heterozygote advantage, is discussed further in the chapter on Population Genetics. Heterozygote advantage can result in the maintenance of otherwise deleterious alleles in a population.

Note also that if environmental conditions change, the selective pressure on a population can change, too. For example, the introduction of a new pathogen might selectively affect individuals with some variants rather than others. Or an increase in environmental temperature might favor some variants vs others. Because of this, genetically diverse populations tend to adapt better to changing conditions than populations that are genetically homogenous.

Genetic drift

While natural selection is a directional change due to some sort of pressure, **genetic drift** refers to the random variations that happen from generation to generation. In any population, not all individuals will reproduce at equal levels. Which individuals reproduce can affect the frequency of alleles in a population. Over many generations, genetic drift can change a population, illustrated in Figure 4.

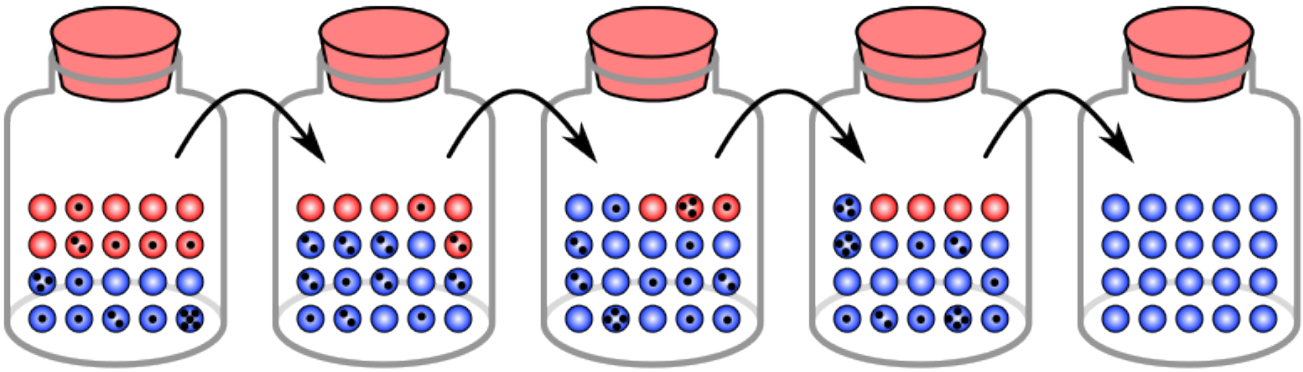


Figure 4. Simulation of a common example used describing the effect random sampling has in genetic drift. Dots indicate samples from each generation that are transferred to the next generation; multiple dots indicate that an individual contributed multiple times. In this population of 20, there is a shift from an allele frequency of 50% for the blue allele to 100% for the blue allele in just 5 generations.

Genetic drift tends to be amplified in small populations where small random fluctuations can have a big effect. To use an analogy: you'd expect a coin flip to generate heads about 50% of the time. But if you flip a coin 4 times, you won't always get 2 heads and 2 tails. Sometimes, you might get 3 heads/1 tail (75% heads). And sometimes you might even get all heads (100%)! On the other hand, if you were to flip a coin 1,000,000 times, you might not have *exactly* 50% heads, but you would not expect 1,000,000 heads in a row.

Two extreme examples of genetic drift magnified by a change in population size should be mentioned: the **founder effect** and the **bottleneck effect**.

The founder effect happens when a small fraction of a population founds, or begins, a new population, as shown in **Figure 5**. The smaller population can have a different genetic makeup than the original population.

For example, if a rare allele is present among the founders, that allele may be more common in the new population. If the populations remain apart for generations, additional drift can further separate the populations.

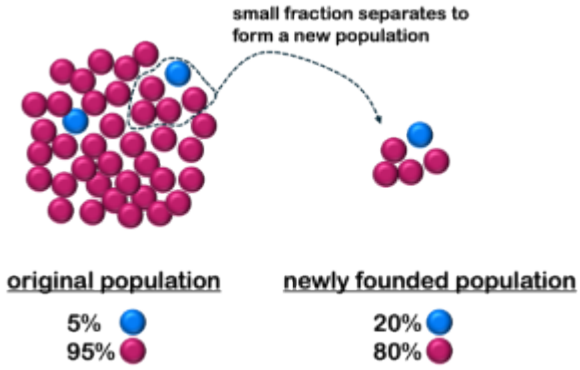


Figure 5. Genetic Drift, Founder effect.

Figure 6. Black wolf in Yellowstone National Park, USA.

An example of the founder effect is seen in the wolves of Yellowstone National Park (**Figure 6**). Hundreds of years ago, grey wolves were found across the northern and western United States. But in the past 150 years, wolves vanished across most of the United States, mostly due to hunting. This had profound consequences on the ecosystem. In 1995, biologists reintroduced a small population of wolves to locations in Yellowstone National Park, in the US states of Wyoming and Montana. The wolves of Yellowstone have been monitored ever since.

Even though it is called a grey wolf, the animal naturally varies in coat color. In most North American wolf populations, less than 5% of wolves are black. The Yellowstone population is about 50% black, reflecting the gene pool of the small number of wolves reintroduced to the park[1].

The **bottleneck effect** occurs when a natural disaster decimates a large percentage of a population. The resulting population may have different allele frequencies from the parental population, as shown in **Figure 7**. In contrast to an event of natural selection, in which one variant has an advantage over another, a bottleneck event kills off individuals randomly. Bottleneck events typically reduce the genetic diversity of a population. Many endangered species have reduced genetic diversity (due to bottlenecks caused by overhunting or overfishing).

1. Park, M. A. P. B. 168 Y. N. & Us, W. 82190-0168 P. 307-344-7381 C. Gray Wolf – Yellowstone National Park (U.S. National Park Service). https://www.nps.gov/yell/learn/nature/wolves.htm.

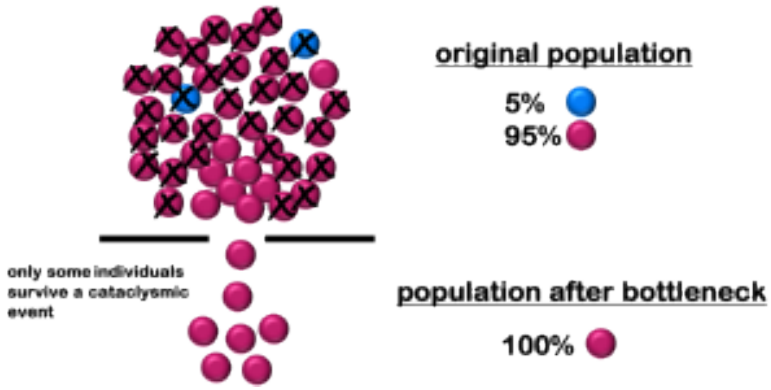


Figure 7. Bottleneck effect

Gene flow

Gene flow results from the migration of individuals (and genes) from one previously isolated population to another. It is also called **admixture**. Admixture can keep somewhat isolated populations of a species from diverging much from one another. This can happen between populations of the same species or different species. Interspecies gene flow is called **horizontal gene transfer.**

Analysis of the human genome reveals that many people have Neanderthal DNA in their genomes, offering evidence that ancient *Homo sapiens* interbred with *Homo neanderthalensis.* This is an example of gene flow between two populations of hominins.

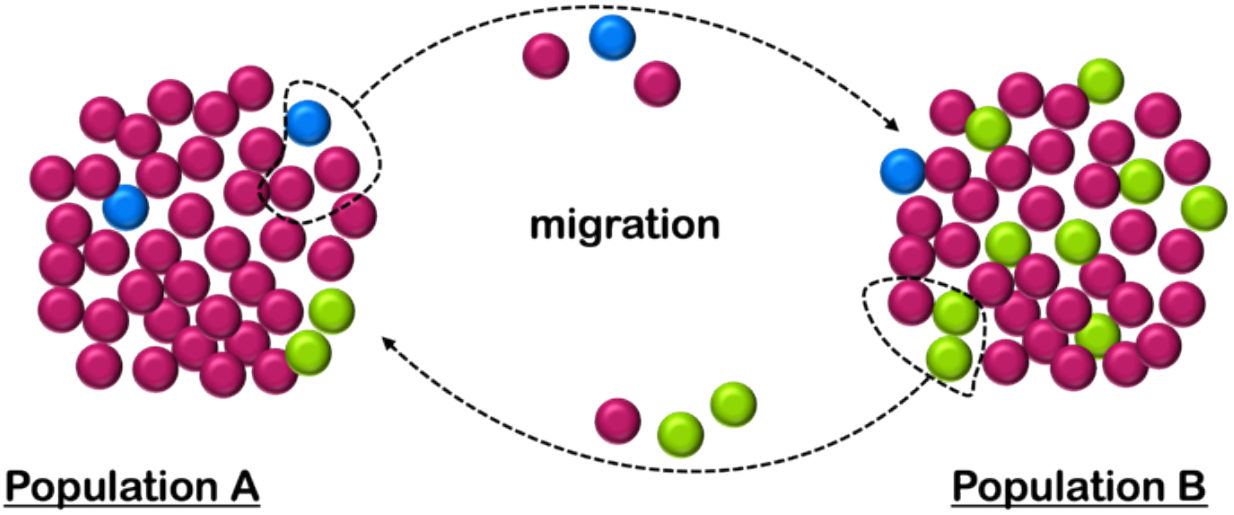


Figure 8. Gene Flow

Assortative mating

Assortative mating can also influence evolution in a population. Most studies of populations assume that mating occurs randomly within a geographical population, but that's not always the case. In human populations, for example, choice of partner greatly depends on social and cultural influences, as well as geography. And even in animal populations, selection of a mate can be nonrandom. For example, several studies show that butterflies have a preference to choose a mate with similar color patterns[2].

If individuals preferentially select mates like themselves, **inbreeding** can further amplify rare traits in a population. Discussed briefly with consanguineous pairings in the Pedigree chapter, inbreeding means that individuals with shared genetics produce offspring. Every individual in a population is likely to have a few rare, recessive variants. Most of the time, the rare recessive traits will not manifest in offspring since it would be unlikely to have two parents *both* share the same rare variant. But with inbreeding that becomes more likely.

The selective breeding that gave rise to many dog breeds is an example of this. Dog breeds tend to be genetically homogenous, specifically bred for desired characteristics. As a result, many dog breeds have an increased incidence of genetic disease: Dalmatians are prone to deafness, German Shepherds are prone to hip dysplasia, and Doberman Pinchers are prone to hypothyroidism, among others.

H5P Test your understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=600#h5p-110

Media Attributions

- Figure 3 Natural Selection © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 4 Genetic Drift © Based on File Random_sampling_genetic_drift.gif is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 5 Founder Effect © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike)

2. Butterflies are genetically wired to choose a mate that looks just like them. ScienceDaily https://www.sciencedaily.com/releases/2019/02/190207142148.htm.

SPECIATION

We can observe instances of mutation, genetic drift, natural selection, and gene flow by monitoring populations over time. But speciation – and the relationship of one species to another – must be inferred through comparison of the anatomy of extant and fossilized organisms. In the last few decades, however, genomic comparisons have become much more common in the analysis of evolutionary relationships.

For speciation to occur, two (or more) populations must be reproductively isolated from one another. For reproductive isolation, gene flow must be minimized between the populations, with each population separately subjected to mutation, natural selection, and genetic drift. While speciation generally happens on much too long a timescale to observe in real-time, we can see instances of microevolution that are likely steps in speciation.

One example is seen in the three-spine stickleback fish, *Gasterosteus aculeatus*, shown in **Figure 9**. These fish can live in both deep water and shallow water, but there are phenotypic differences between deep water and shallow water populations. Deep-water fish have an extra, spiny pelvic fin protruding from their ventral side, while shallow-water fish lack this ventral fin. It is thought that the pelvic fin is disadvantageous to the shallow-water fish since it might drag in shallow-water sediment and serve as an attachment point for parasitic insects like dragonfly larvae. The spiny pelvic fin is advantageous to the deep water fish, though, because it protects against predators.

A comparison of the genomes reveals the underlying genetics of these phenotypes: the pelvic fin seems to depend on an enhancer upstream of the *Pitx* (paired-like homeobox 1) gene. *Pitx* is expressed in certain clusters of cells during development, including the cells that eventually develop into the pelvic fin. The shallow-water fish carry a loss-of-function mutation in the enhancer. Although the *Pitx* is fully functional in the shallow-water fish and is expressed in other parts of the embryo, it is not expressed in the cells that would otherwise give rise to the pelvic spine.

These populations are reproductively isolated, and natural selection has favored fish either with or without the spiny fin, depending on the ecosystem. This has resulted in the divergence of these two populations. Some scientists think that the stickleback is in the early stages of speciation.

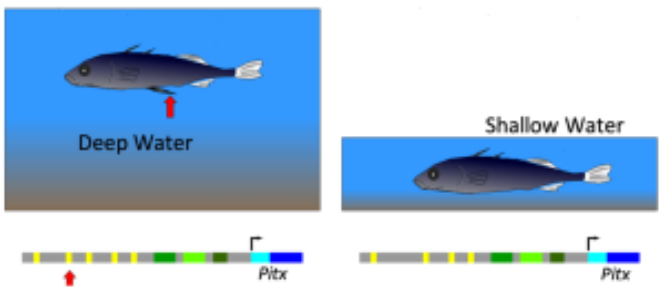


Figure 9. Deep water sticklebacks are reproductively isolated and are genotypically and phenotypically distinct.

Test your understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=616#h5p-111

Media Attributions

- Figure 9. (right) Speciation © JaySo83 is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 9. (left) Speciation © Open Genetics Lectures, Original-Deyholos is licensed under a CC BY-NC (Attribution NonCommercial) license

MEASURING EVOLUTION: MOLECULAR CLOCKS

Mutations occur naturally and accumulate in a population over time. For this reason, the oldest populations – the ones that have been sustained through the most generations – tend to have the most variation. Young, newly established populations often have very little variation. The number of variants *within* a single population and *between two* populations can thus be used as a sort of **molecular clock** if we assume that mutations accumulate at a relatively consistent rate.

Figure 10 shows an example of a hypothetical molecular clock. In this example, two lineages diverged 50 million years ago about share 4 differences . If a third lineage was found to have twice as many differences, one might hypothesize that this third lineage last shared a common ancestor with these two 100 million years ago.

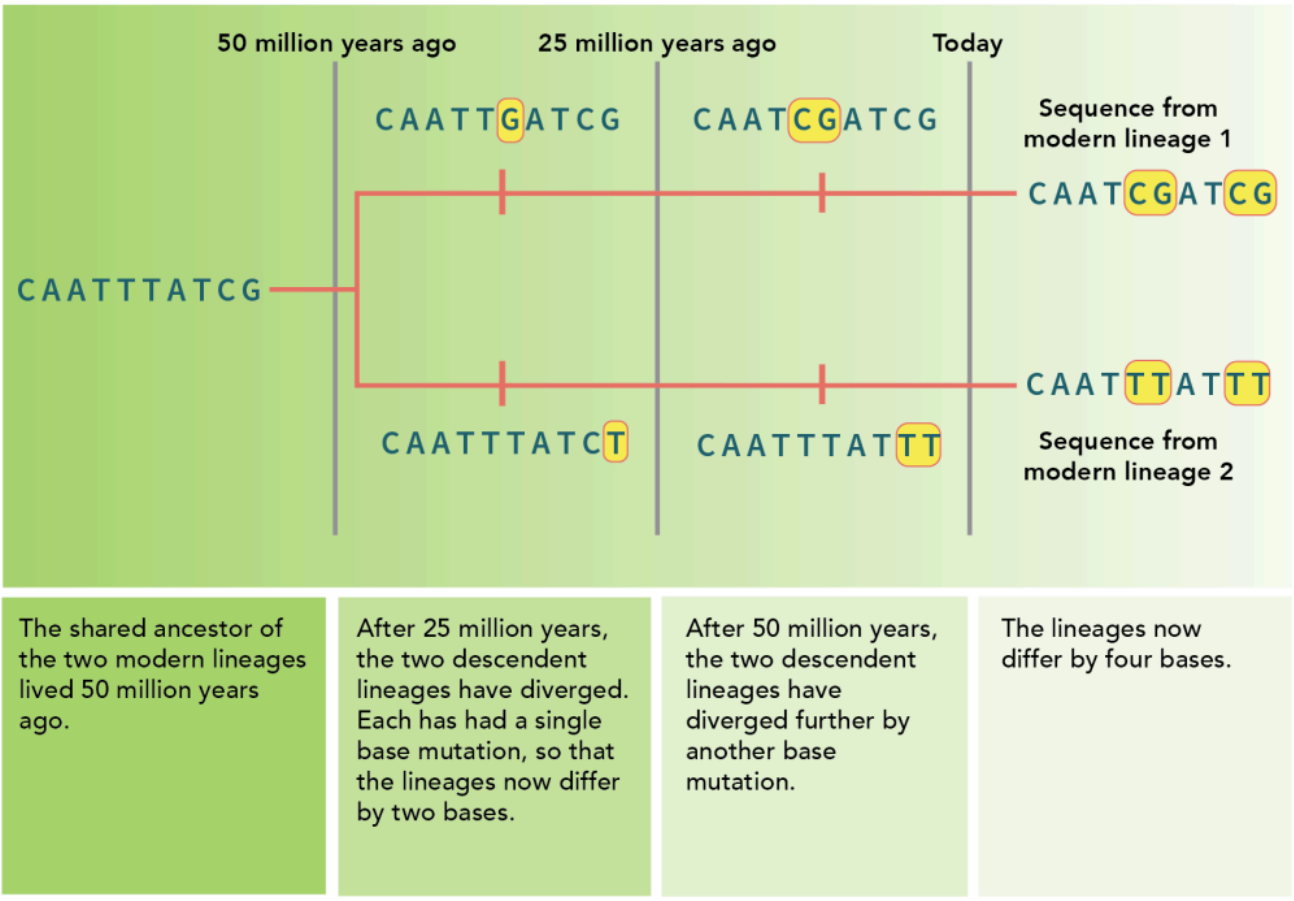


Figure 10. Molecular clocks. The accumulation of mutations can be used to estimate evolutionary time. Assuming that mutations

The rate of mutation can be different depending on species, environment, and part of the genome. Therefore, the number of mutations can be very useful in determining relative timeframes for the age of a population, but they need to be "calibrated" by comparing them to fossil or archaeology records for a more precise estimate of age.

As an example from humans, humans as a species (*Homo sapiens*) are thought to have originated in Africa. Ancient ancestral humans then migrated out of Africa, gradually populating other parts of the globe. Figure 11 shows a map that diagrams the routes early humans took as they migrated out of Africa. These migration patterns were originally hypothesized by dating archaeological remains.

The oldest human populations are thus found in Africa, and the youngest in South America and the Pacific Islands. The molecular clock supports this hypothesis: by far the most diverse indigenous populations are found in Africa. In fact, most of the genetic variation in the human population worldwide is found in African populations! The younger indigenous populations in the Americas and Pacific Islands are more genetically homogenous.

Note that the amount of diversity alone, however, does not allow determination of precise dates. Instead, a comparison of the molecular variation with the archaeological data allows the molecular clock to be calibrated. Then, age can be estimated for populations where minimal archaeological data is available.

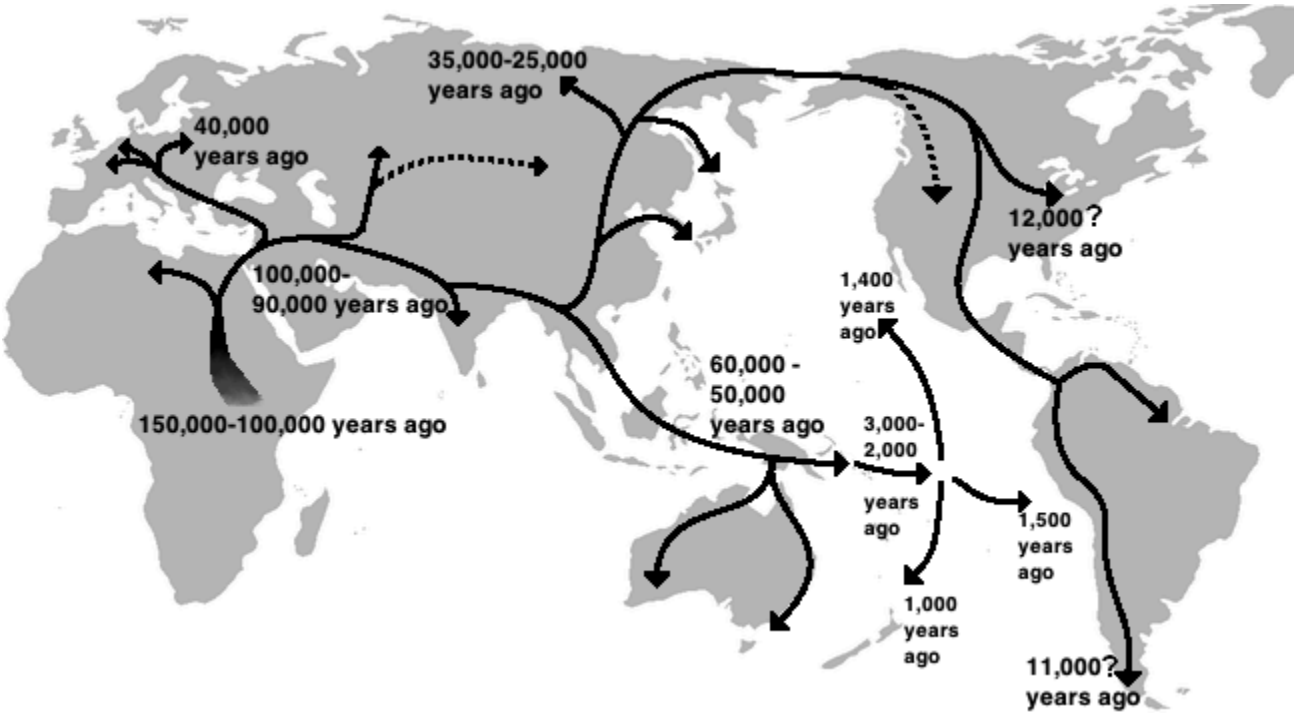


Figure 11. Humans are hypothesized to have originated in Africa and migrated to populate other parts of the world. Note: the indicated dates are one estimate only.

The accumulation of mutations over time can also be used to determine relationships among species (and

individuals within a species). Because some sequences are conserved more than others, the choice of a comparative sequence can affect the analysis.

For example, some parts of the genome don't vary much because they are under strong selection: almost any change to the sequence would decrease the reproductive fitness of an individual. These conserved regions of the genome are less useful in establishing relationships within a species because most of the sequence is the same regardless of which individual is tested. But other parts of the genome are highly variable: certain intragenic regions of the genome away from important regulatory sequences can show lots of variation, as are certain regions of mitochondrial DNA.

On the other hand, these highly variable regions of the genome are less useful in comparing relationships among species: if they are too different, no comparison is possible. For example, the so-called D-loop region of mtDNA, shown in **Figure 12**, is highly variable and useful in comparing relationships among humans. mtDNA overall is estimated to undergo mutations at twice the rate of nuclear DNA, with several hypervariable regions in the noncoding D-loop of the chromosome[1]. But these regions are so different in humans compared to other primates that it is less useful in constructing a phylogenetic tree even of primates.[2] In comparing relationships among species, more conserved regions of the genome may be better suited.

1. Nicholls, T. J. & Minczuk, M. In D-loop: 40 years of mitochondrial 7S DNA. Exp. Gerontol. 56, 175–181 (2014).

2. Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. & Takahata, N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc. Natl. Acad. Sci. U. S. A. 92, 532–536 (1995).

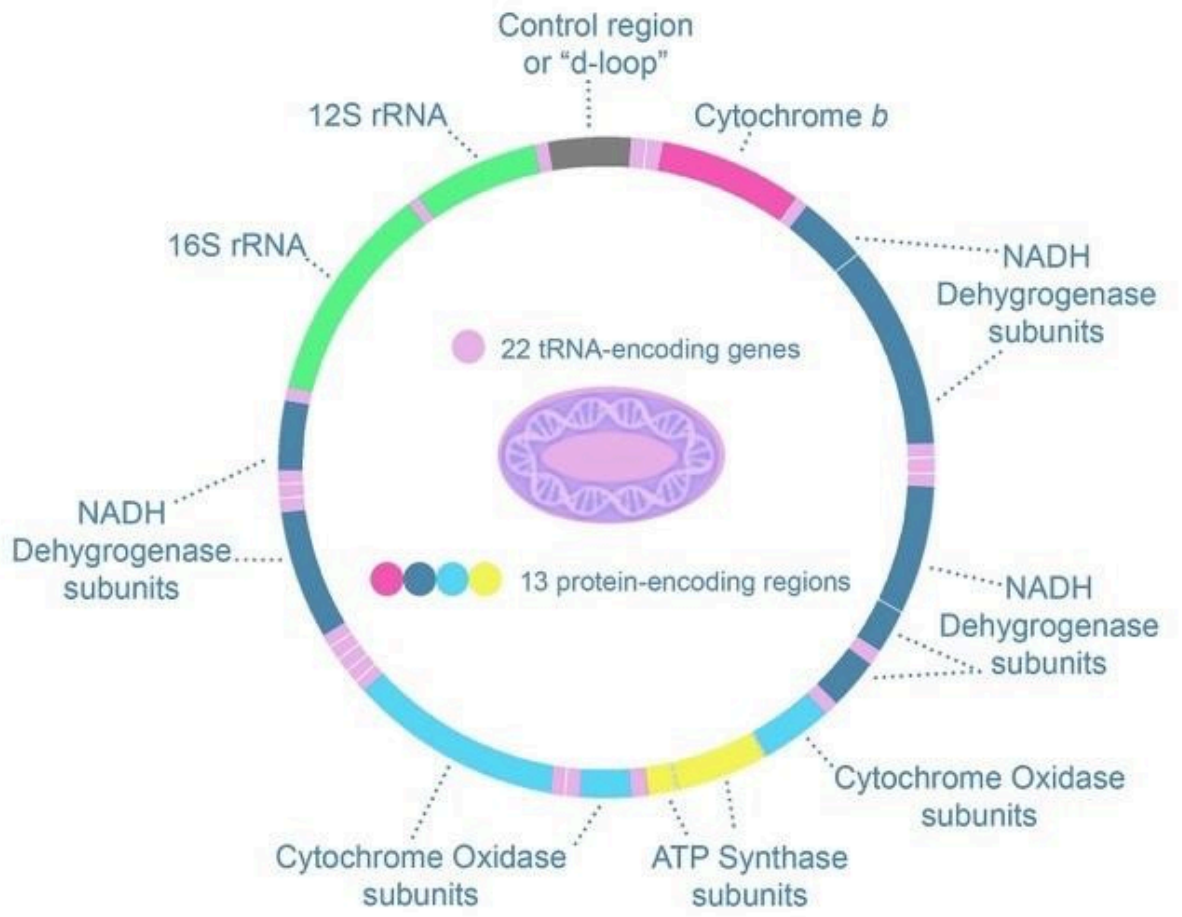


Figure 12. Map of the mtDNA chromosome.

Constructing an Evolutionary Family Tree

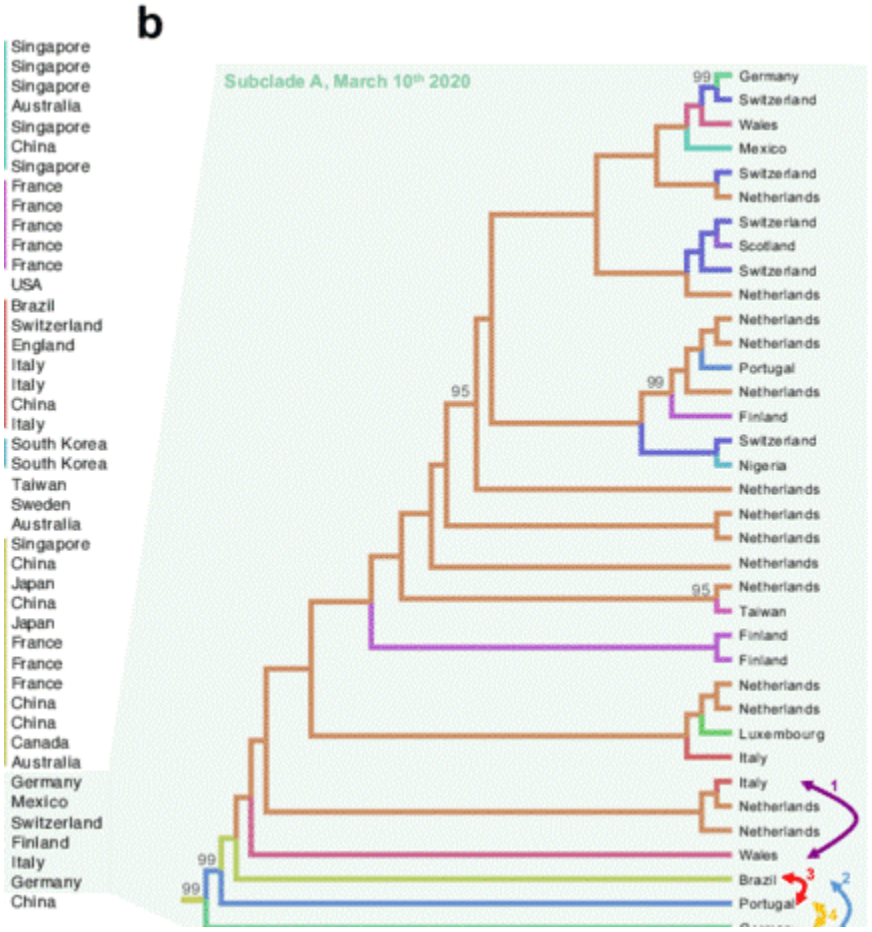
The number of genomic differences can then be compared among individuals in a population or between species. There are a few different ways to mathematically weight the differences, but the general principle is that the fewer the differences, the more closely the two subjects are related. Geneticists construct a **cladogram** to illustrate these relationships.

Note: The terms phylogenetic tree, cladogram, and dendrogram all describe similar images. While there are some differences between the terms, many biologists use these terms interchangeably, and the differences will not be addressed here. A **clade** is a branch on the family tree.

Figure 13 gives an example of a cladogram and shows how sequence analysis was used in virus tracking during the early days of the COVID-19 pandemic. In the tree shown on the left, the sequences of individual patient samples were compared, and labeled with their country of origin. Shorter branches represent patient

samples with a more recent common ancestor, which gives a clue to how the virus spread geographically. The tree shown on the right tracks COVID variants as they arose among patient samples in 2020-2021.

Note that some variants disappeared from the population early on and were out-competed by Delta variants by the end of 2021. Although these data stop at the end of 2021, the beginning of the rise of Omicron variants can be seen in the red clade.



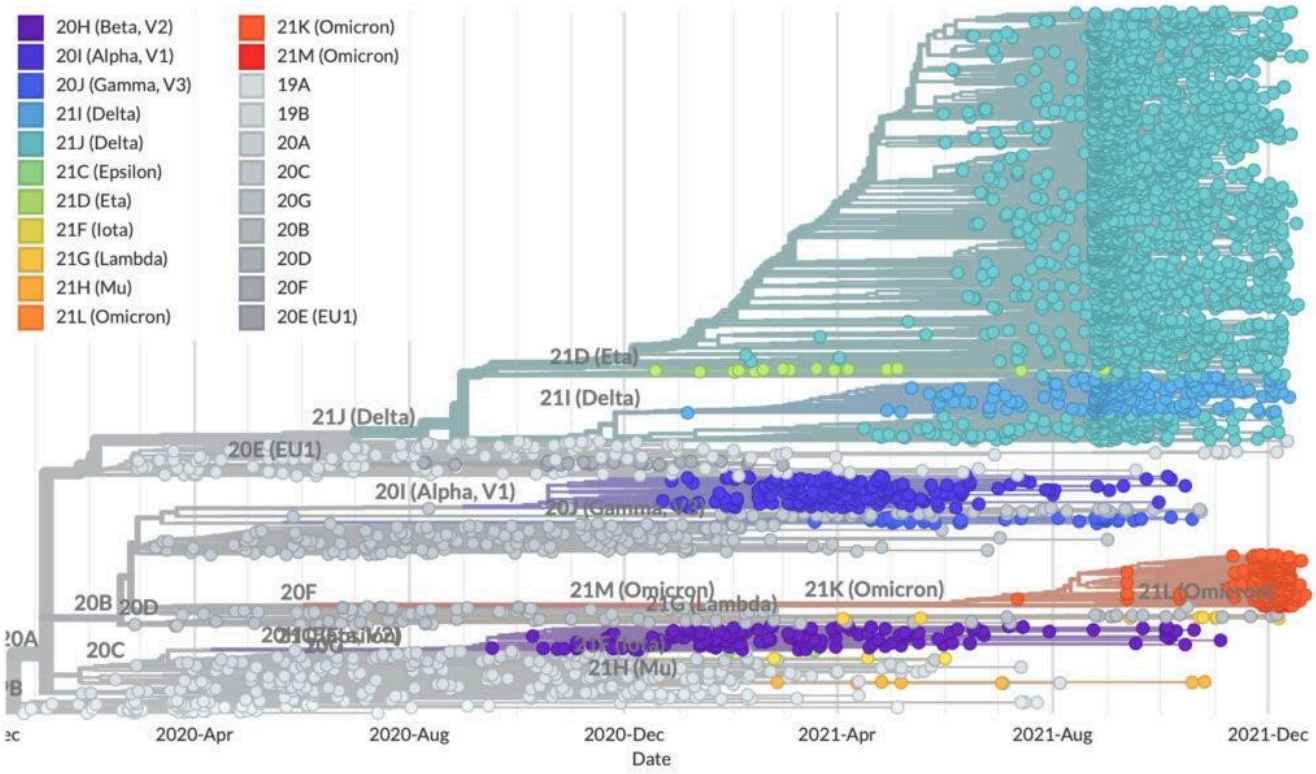


Figure 13. In the early days of the COVID-19 pandemic, sequence analysis of viral samples from patients was used to track the source of the virus. On the top is a cladogram showing the relationship of SARS-Cov2 samples isolated from patients in the early days of the COVID-19 pandemic. Such genetic comparisons were used in tracking the spread of the virus worldwide. For example, in the tree on the top right (b), the only Taiwanese sample is most closely related to samples from the Netherlands, suggesting that the Taiwanese case spread from the Netherlands. On the bottom is a more complex cladogram tracking the rise of SARS-Cov2 variants. The earliest Delta variants appeared around August of 2020, while Omicron variants began spreading in August of 2021. These variants arise from new mutations of the virus that are spread through human infection. Note that viruses accumulate mutations at a much faster rate than most organisms.

Media Attributions

- Figure 10 Molecular Clocks is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license
- Figure 11 Human Global Migration Patterns © Ephert is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 12 Map of the mtDNA chromosome © BQUB17-JHolguera is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 13 left Sequence Analysis © Researchgate.net is licensed under a CC BY-NC-ND (Attribution

SEQUENCE, CHROMOSOME STRUCTURE, STRUCTURAL VARIANTS: EVOLUTION OF GENOMES

While a lot of information can be gathered from a comparison of DNA sequence, chromosome structure, and structural variants also offer a wealth of molecular information about evolution. Genomes evolve, too, and new genes can arise through large-scale structural variants.

Over the course of evolution, genomes don't just accumulate single nucleotide polymorphisms. Larger structural variants also accumulate. Like all mutations, structural variants can be advantageous, disadvantageous, or neutral, but most will only be maintained in a population if they confer some sort of reproductive advantage. Some large-scale chromosomal rearrangements may affect far more than just one gene at a time, and they likely play a correspondingly larger role in speciation.

Structural variants include chromosomal rearrangements, exon shuffling, gene deletions, and duplications, and even whole-genome duplications in some cases! This can result in multiple copies of a gene. Exon shuffling is a chromosomal rearrangement that rearranges exons. This may involve moving, duplicating, or deleting exons of a single gene, or it can involve a recombination event that links exons of two different genes.

Not all duplications result in functional genes: some may copy only part of a gene or reassemble exons in a nonproductive manner, resulting in an incomplete, nonfunctional gene. Gene duplication and exon shuffling events are followed by further mutations, so the duplicated sequences also diverge from each other over time. These events can generate new genes with new functions. These structurally related genes are **paralogs** of one another. Sometimes, additional loss-of-function mutations may accumulate, causing the inactivation of one or more of the paralogs. These nonfunctional **pseudogenes** persist in the genome but are usually not translated into protein. (**Figure 14**)

Pseudogenes actually make up quite a bit of most eukaryotic genomes. In humans, for example, there appear to be more pseudogenes than protein-coding genes![1]

1. Karro, J. E. et al. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res. 35, D55–D60 (2007).

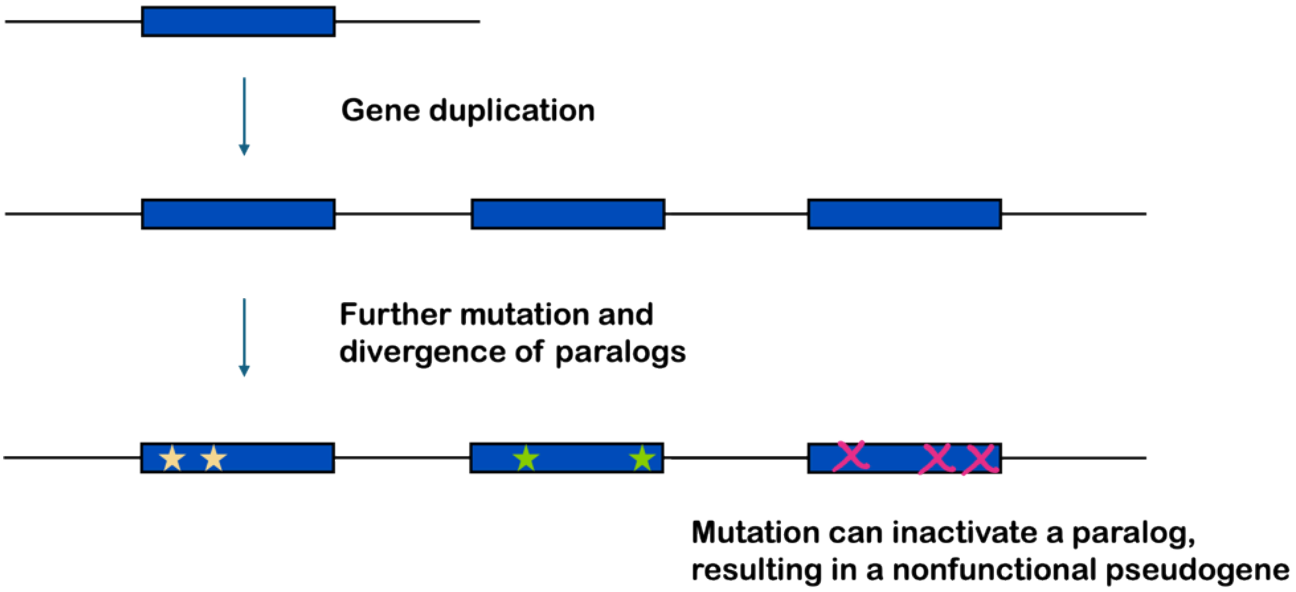


Figure 14. Paralogs result from gene duplication. Duplicated genes can undergo further mutation over time, resulting in the divergence of the paralogs. Mutations can inactivate a paralog, resulting in a nonfunctional pseudogene. Eukaryotic genomes have many pseudogenes.

We've seen examples of large-scale rearrangements and paralogs in this text already: the chapter on gene expression during eukaryotic development looked at the *hedgehog* gene in *Drosophila* and its three corresponding **orthologs** *sonic hedgehog, desert hedgehog, and Indian hedgehog* in vertebrates. (Remember that the word **ortholog** refers to a related gene found in a different species, and the word **paralog** refers to duplicated genes within the genome of an organism.)

The chapter on cancer biology looked at paralogs of p53 in proboscideans (elephants), which likely evolved concomitantly with an increase in body size.

Large-scale structural variants also result in genome structures that can be quite varied from species to species, so a comparison of chromosome structure among species can often give additional clues to evolutionary relationships. For example, mouse and human protein-coding genes are about 85% identical, depending on what sequences are counted. (This varies from gene to gene: some genes are >85% identical, and some are <85% identical.) But if we compare how those sequences are arranged within the genome, genome structure looks quite different, as shown in Figure 15. In Figure 15, genome **synteny** refers to how sequences align within two or more genomes

Figure 15. Synteny between the human and mouse genome. The 23 autosomes plus the X and Y of the human genome are arranged in the inner circle, each chromosome represented by a different colored arc. The Mouse genome has 19 autosomes, plus X and Y. The mouse chromosomes are color-coded to match where a similar sequence is found.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=639#h5p-112

Media Attributions

- Figure 14 Pseudogenes © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 15 Synteny © generated with JAX Synteny Browser

HUMAN ANCESTRY TESTS

We started this chapter with an example of genetic variation and adaptation to high altitude in humans. Let's return to human ancestry, now with an understanding of the underlying mechanisms that affect evolution.

Humans, like all living things, can be better understood within the context of evolution.

The most ancient human populations likely arose in Africa, with several successive migrations from Africa populating other parts of the world. As these migrations occurred, there is evidence that the dispersing *Homo sapiens* interbred with Neanderthals and Denisovans , now-extinct hominin species that coexisted with early *Homo sapiens.*

The migrating populations continued to acquire genetic variants. Selective pressures in different geographic regions likely influenced which variants were maintained in different populations. For example, darker skin is usually found among people whose ancestors lived near the equator, and lighter skin is usually found among people whose ancestors lived in areas nearer to the poles. These adaptations confer protection against UV sun damage among darker-skinned populations near the equator, while lighter skin may be an adaptation that makes it easier for Northern populations to synthesize vitamin D.

The adaptation to high altitude described at the beginning of this chapter is another example. Genetic drift played a role in the accumulation of variants as well, with phenotypically neutral SNPs accumulating in some populations and not others due to the randomness of time and reproduction.

The migration of early humans was not unidirectionally "out" of Africa: there is evidence of "reverse migration" as well. Populations of early humans thus did not exist as discrete, reproductively isolated groups. Because of this, humans show considerable admixture among populations of different geographic ancestry. Rather than discrete ethnic groups, human variation appears to exist on a geographic continuum. Any one person is more likely to be similar to someone with ancestry from a nearby geographic location and less similar to someone of ancestry separated by more distance but rarely is there evidence of true reproductive isolation.

Human ancestry tests marketed by companies like 23andme, AncestryDNA, and others typically take advantage of the geographic association of certain variants; that is, the frequency of SNPs varies with geographic ancestry. These companies compare SNPs across the genome, looking for patterns that are similar to other groups of people of known ancestry.

However, single SNPs cannot predict ancestry. Instead, these genome analyses typically look at **haplotypes**, longer stretches of adjacent SNPs that tend to remain linked in multiple generations. (Haplotypes were first discussed in the chapter on linkage.) These ancestry tests also look at maternal lineage through mitochondrial DNA and paternal lineage through Y chromosome DNA. Because there is no recombination in mtDNA and the Y chromosome, these chromosomes are passed relatively intact from generation to generation. Closely

related mtDNA and Y chromosome DNA arise from shared ancestry and are called **haplogroups**. A simple video overview of haplotypes and haplogroups can be found at Learn.Genetics.

Ancestry-testing companies look for the frequency of those haplotypes and haplogroups in reference populations of known ancestry. This frequency can predict the likelihood that someone of that geographic ancestry will have those combinations of SNPs. The larger the reference pool, the more reliable the test becomes. Because of this, such ancestry tests can pinpoint smaller and smaller geographic regions every year, but they are still limited by the people who have contributed to the reference pool. Historically, such companies have had more European reference data than for other continents, although this is gradually changing.

By looking at short stretches of haplotypes along a chromosome, these tests can also detect the admixture that results from ancestors of varying geographic ancestry, illustrated in **Figure 16** (top)**.**

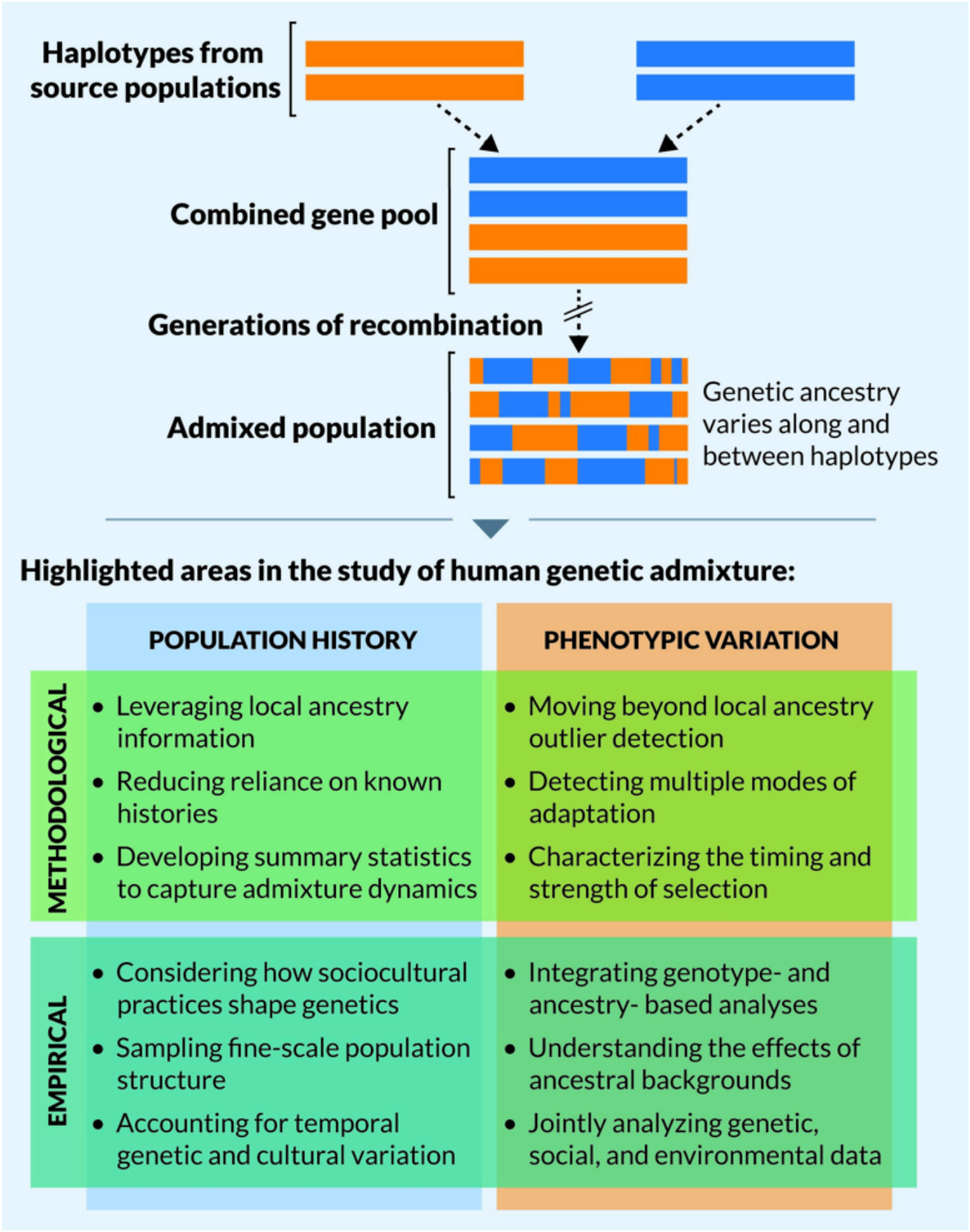


Figure 16. From Korunes and Goldberg, 2023. Large-scale movements of individuals allow haplotypes from

previously isolated populations to come together in a combined gene pool. Generations of recombination between these haplotypes lead to an admixed population with genetic ancestry that varies between individuals and along haplotypes. The distribution of this variation is governed by the demographic and selection history of the admixed population and its sources. (Bottom) To leverage these patterns of admixed ancestry to better understand human history and phenotypic variation, we highlight key areas of recent progress and possible future directions in the study of human genetic admixture.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://rotel.pressbooks.pub/genetics/?p=658#h5p-113

Media Attributions

- Figure 16 Haplotypes © Korunes and Goldberg, 2023

A NOTE ON RACE AND HUMAN ANCESTRY

Humans like to categorize things, including other humans. And the concept of race permeates nearly every part of life in the United States in one way or another.

Early geneticists assumed that phenotypic variation among races arose through the dispersal of human populations and parallel divergence from a common ancestor. This assumption often formed the basis for egregious social practices designed to maintain power for one racial or ethnic group over another.

But there's a lot of data to show that this assumption just isn't true.

First of all, even self-identification of race is very subjective and differs from country to country across the world. For example, people who self-identify as white in Brazil often have more African ancestry than people who self-identify as Black in the United States.[1]

In addition, there is more genetic variation within racial groupings than there is between people of different races. In other words, in a genomic comparison of your DNA with someone of the same race and someone of a different race, you're just as likely to be a closer genetic match to the person of a different race9. This is especially true for the oldest African populations, which tend to be far more genetically diverse than populations founded later in human expansion.

And finally, genomic data reveals so much admixture that there are no reproductive separations among different racial groups. Instead, people tend to be most closely related to others with proximal geographic ancestry, but there does not seem to be any discrete separation. Human variation exists along a continuous variation, mostly along the exact migration paths described back in**Figure 11.** The haplotypes used to predict geographic ancestry are hyperlocal: they do not fall along the borders one might expect for racial separation, and different haplotypes have overlapping geographies.

These data collectively form the scientific basis for the statement that there is no genetic separation of racial groups.

Despite this, race continues to be used as a proxy for genetics in medicine, sometimes with good intentions but often to patients' detriment. Historically, clinical trials were overwhelmingly performed using white male patients. This led to treatments that worked well in white male patients but sometimes less well for other populations. In recent years, efforts have been made to increase diversity in clinical trials, with efforts to specifically recruit BIPOC patients as well as women and gender-diverse patients historically underrepresented

1. Templeton, A. R. Biological Races in Humans. Stud. Hist. Philos. Biol. Biomed. Sci. 44, 262–271 (2013).

in clinical trials.[2] But identifying patients solely by race does not necessarily lead to improved outcomes either, since race is not a good proxy for genetics.

With that being said, a lack of genetic basis for race does not mean that race is not real. Race and cultural influences can have a profound effect on human health: for just two examples, Black women have a three-fold higher chance of dying in childbirth compared with white women, and there is strong evidence for disparities in how pain is treated and managed in BIPOC populations compared with white populations[3] Salmond, S. & Dorsen, C. Time to Reflect and Take Action on Health Disparities and Health Inequities. Orthop. Nurs. 41, 64–85 (2022)

An increased understanding of both genetic and social/cultural influences on human health is necessary to improve outcomes. The bottom panel of **Figure 16** offers some suggestions for the use of genetic and genomic data.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=663#h5p-114

2. Igwe, J. et al. Opportunities to Increase Science of Diversity and Inclusion in Clinical Trials: Equity and a Lack of a Control. J. Am. Heart Assoc. Cardiovasc. Cerebrovasc. Dis. 12, e030042 (2023).

3. Njoku, A., Evans, M., Nimo-Sefah, L. & Bailey, J. Listen to the Whispers before They Become Screams: Addressing Black Maternal Morbidity and Mortality in the United States. Healthcare 11, 438 (2023).

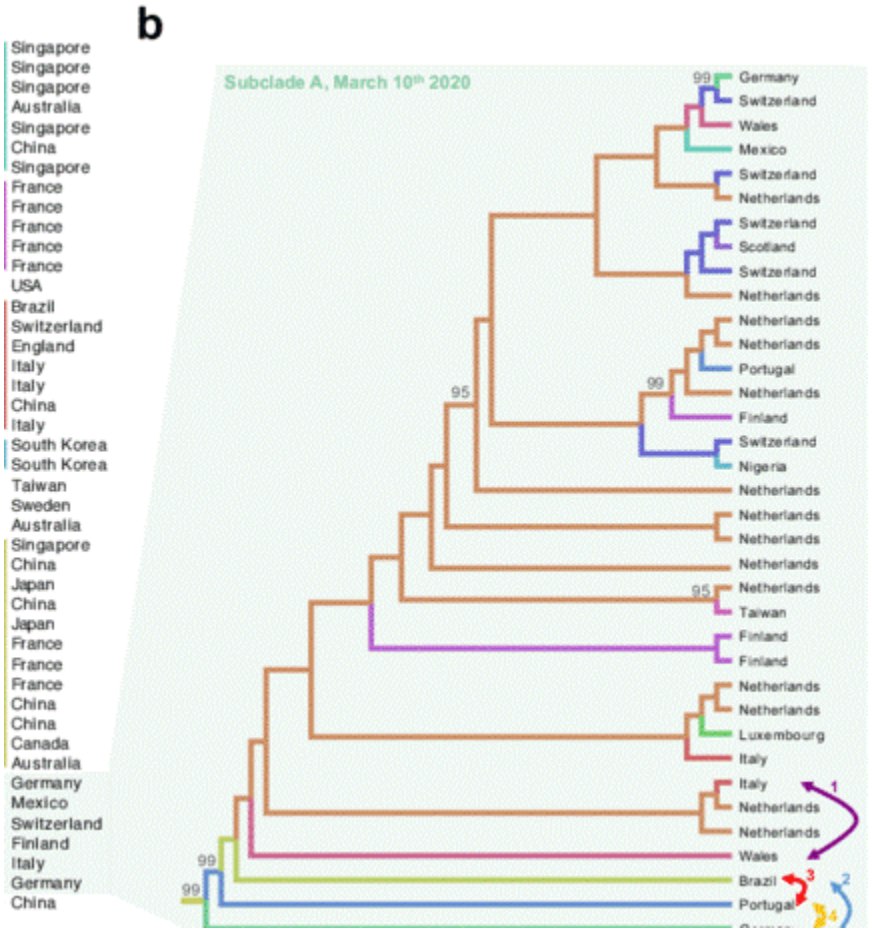
WRAP-UP QUESTIONS

1. Adaptation and evolution depend on variation in a population. Other chapters mentioned the genetic homogeneity of bananas: virtually all bananas commercially available today are genetically identical Cavendish bananas. The Cavendish banana supplanted the Gros Michel banana, which was the dominant market banana until the 1950s when the banana disease Fulsarium wilt decimated the crop worldwide. Explain why genetically homogenous populations are particularly susceptible to disease and why a genetically varied population might be more resistant.

2. Would you choose a highly conserved or poorly conserved region of the genome to determine the relationship among individuals of the same species? Explain your reasoning.

3. Describe how a molecular clock is used to determine how long ago two species shared a common ancestor and explain why additional information like the fossil record is needed to calibrate the clock.

4. **Figure 13** from this chapter, reprinted here, shows a cladogram of the relationship of SARS-Cov2 samples isolated from patients in the early days of the COVID-19 pandemic. From where did the Luxemburg patient most likely acquire their COVID-19 infection?



5. Explain how **Figure 19** below, reprinted from the Cancer chapter, illustrates natural selection.

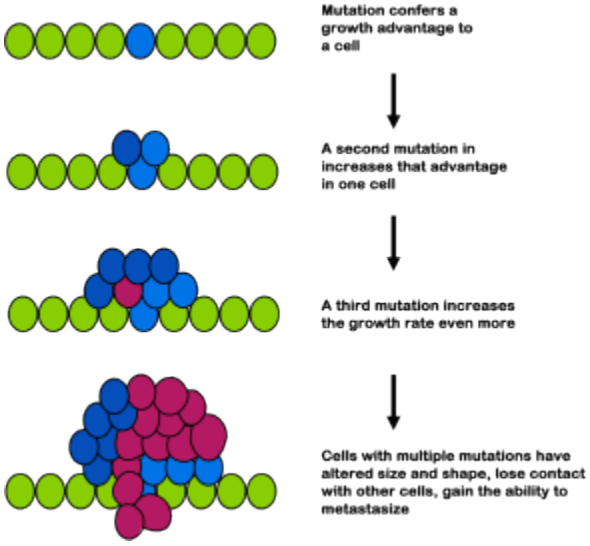


Figure 19. Tumors are genetically heterogeneous

Science and Society

6. Race is based on culture and not genetics, but in human medicine, patients are often identified by race. Racial categories can be used to influence treatment plans despite race being a poor proxy for genetics. Given that both biology and society play a role in human health, how should race be used in data collection and medical recommendations? Use outside resources as needed to support your argument.

Media Attributions

PART XVIII

DNA REPAIR AND CANCER

Learning Objectives

1. Define cancer, tumor, tumor-suppressor, proto-oncogene, oncogene.
2. Compare and contrast mismatch repair, nucleotide excision repair, and base excision repair.
3. Compare and contrast non-homologous end-joining and homologous recombination.
4. Explain why cancer is described as a genetic disease.
5. Describe the process by which a cell becomes cancerous, including the types of mutations in proto-oncogenes and tumor suppressors.
6. Recognize the relationship between age and incidence of cancer.
7. Explain why elephants and naked mole rates rarely get cancer, using the hallmarks of cancer to justify the explanation.

Source material:

Wong, EV. Cells – Molecules and Mechanisms1. Chapter 7.6 DNA Repair and Chapter 16: Viruses, Cancer, and the Immune System

https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Cells_-

_Molecules_and_Mechanisms_(Wong)/16%3A_Viruses_Cancer_and_the_Immune_System/16.03%3A_Cancer [1]

Introduction

Before we get started, stop and think about what you know about cancer. You may think of words like "tumor." You may have heard of people saying they "have the breast cancer gene." You may have seen images of people who lose their hair as they undergo cancer treatments. In this chapter, we will look at the genetic mechanisms by which cancer arises.

The process by which cancer forms is called **oncogenesis**. The prefix "onco-" refers to cancer. You'll see this used in words like "oncologist," which is a doctor who specializes in cancer, or "oncogene," which, as we will see, is a gene associated with oncogenesis.

A **tumor** is a mass of cells that grows much faster than surrounding healthy tissue. All multicellular organisms can grow tumors – even plants, as shown in **Figure 1**. A tumor typically arises from one cell that begins to proliferate faster than its surrounding cells.

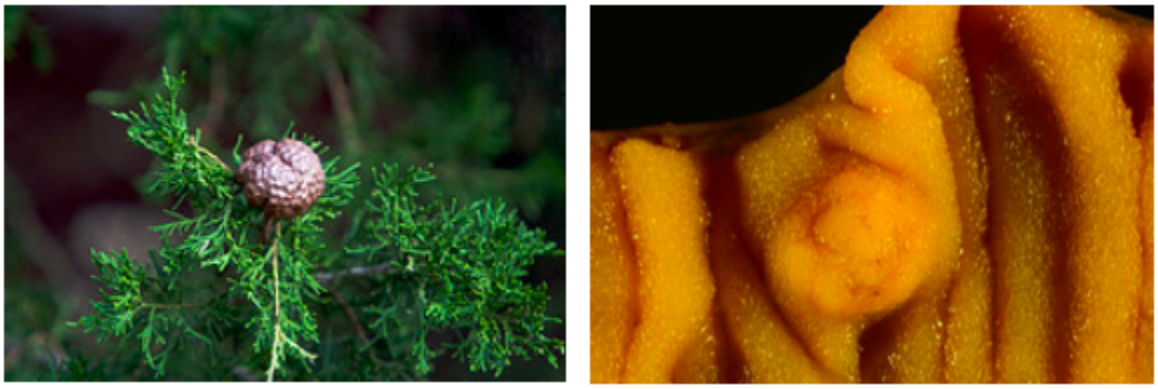


Figure 1. (left) A tumor on a cypress branch. (right) A tumor of the small intestine. Cypress tumor photo by W. Calder, cc licensed 2009. Small bowel tumor by E. Uthman, public domain 1999.

Tumors can be a solid mass of cells like those seen in **Figure 1**, or they can be "liquid tumors" – individual

1. Cells – Molecules and Mechanisms (Wong). *Biology LibreTexts* https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Cells_-_Molecules_and_Mechanisms_(Wong) (2018).

cells that are dividing uncontrollably but are not all clumped together. Blood cancers like leukemias are liquid tumors because there is no solid mass of cells, but the cancer cells end up overrunning the bloodstream. Leukemia is shown in Figure 2. On the left is a normal blood smear, where you can see many pink-red blood cells and far fewer purple-staining white blood cells. On the right is a blood smear from a leukemia patient. Note the abundance of purple-stained leukemia cells among red blood cells in the leukemia smear.

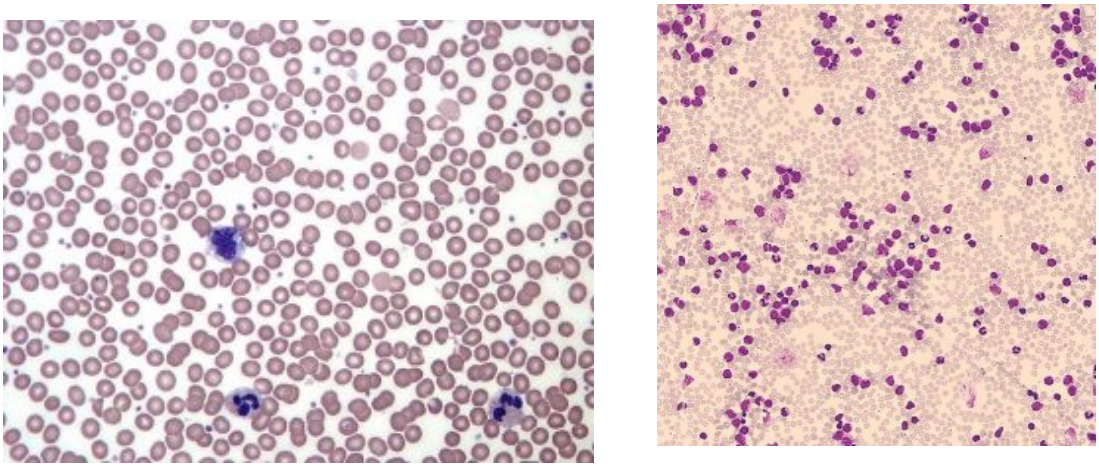


Figure 2. Healthy peripheral blood smear (left) vs leukemia (right).

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=720#h5p-96

Tumors begin to cause health problems when they impede the normal function of the tissue in which they are found. For example, lung cancers will impede the function of the lungs and eventually block the ability to breathe. Tumors that arise in nonessential parts of the body – like breast tissue, for example – cause health problems not because they block the function of their tissue of origin but because they can **metastasize**. Metastasis means that cells escape from the original (or primary) tumor, travel through the bloodstream, and form a secondary tumor at a distant site. The process of metastasis is described in more detail in **Figure 3**.

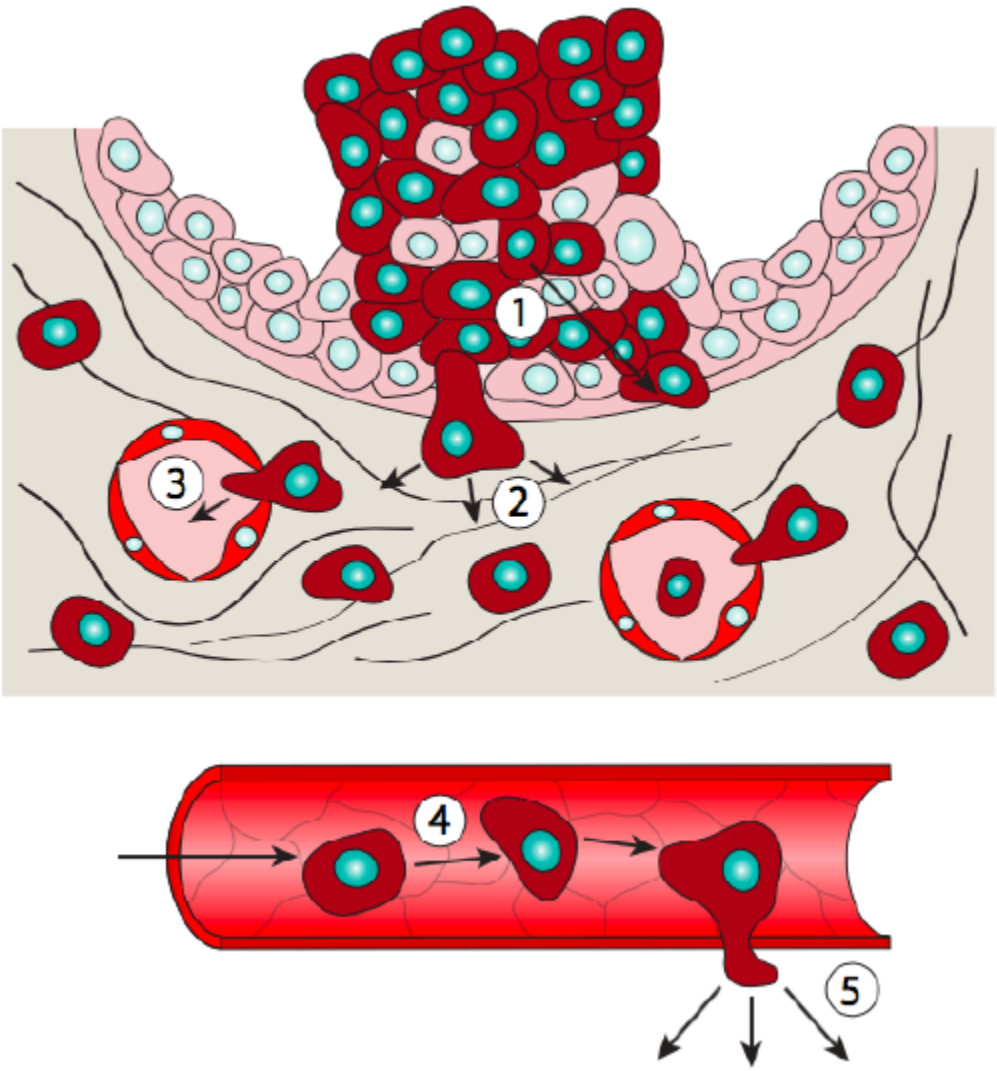


Figure 3. Metastasis starts with the loss of adhesion between tumor cells (1). Non-metastatic tumors are surrounded by a capsule of extracellular matrix that contains the tumor in its location. To escape this capsule, the metastasizing cell must secrete proteases (usually metalloproteases) that can break down the extracellular matrix proteins (2). Once out into the looser connective tissue, the metastatic cell increases its locomotive activity and heads for a blood vessel. Intravasation (3) into a small, low-flow blood vessel allows the cell to be carried to nearly any destination in the body by the circulatory system (4). At some point, the metastatic cell will attach to the interior wall of a blood vessel and exit the circulation (5). The molecules and situations that determine the point of exit are not clear yet although there are clearly preferred sites of metastasis for some types of tumors. Image source: Wong, Cells-Molecules and Mechanisms.

A tumor is not necessarily cancerous. A benign tumor stays localized to its original location in the body, surrounded by an extracellular matrix that encapsulates the tumor. A benign tumor can still cause health problems by putting pressure on surrounding healthy tissue and impeding normal function. However, a

cancerous or **malignant** tumor has the potential to spread beyond its original boundaries and can metastasize to new tissue, impeding the function of both the primary tissue and the secondary tissue.

Cancers are usually defined by their tissue of origin. So, in humans, breast cancer arises because of the uncontrolled growth of breast cells, lung cancers are an uncontrolled growth of cells in the lung, and leukemias (a blood cancer) are the uncontrolled growth of leukocytes (a type of white blood cell). Even after metastasis, such cancers are still describable by their primary tissue.

The growth advantage that tumor cells have over their surroundings comes from the **somatic mutation** of genes involved in cell growth and maintenance of the genome. For this reason, cancer is said to be a genetic disease. You'll recall from our discussion of mutation that somatic cells are not reproductive cells, so the cancer-associated somatic mutations will not be inherited by offspring.

In the chapter on mutations, we discussed some of ways mutations can arise. But DNA is damaged far more often than it is mutated because the vast majority of damage is repaired. Both prokaryotic and eukaryotic cells have specific mechanistic pathways to repair damage. These pathways are overlapping in function and some proteins are shared between pathways. When these repair pathways fail, mutations can accumulate. A mutation in exactly the wrong place of the genome can cause the cell proliferation that leads to the growth of a tumor.

This chapter will begin with a look at how damage arises and how different types of damage are repaired. We'll then transition to the types of mutations that can lead to cancer, with some examples of specific genes involved in oncogenesis. Finally, we'll talk briefly about cancer treatments, including a few examples of targeted therapies that exploit what we know about cancer genetics.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=720#h5p-97

Media Attributions

- Figure 1 DNA Repair is licensed under a CC0 (Creative Commons Zero) license
- Figure 2-left DNA Repair © By Keith Chambers - Scooter Project is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 2-right DNA Repair © By El*Falaf is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 3 DNA Repair © Metastasis is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts. is licensed under a CC0 (Creative Commons Zero) license

REPAIR OF DNA DAMAGE

The chapter on Mutation discussed some of the ways that DNA could be damaged and the types of mutations that could result from each type of lesion. But as mentioned at the end of the chapter on mutation, not all DNA lesions result in mutation. Cells have numerous repair mechanisms in place to fix damage as it occurs, and the (im)balance between damage and repair is what results in accumulation of mutations.

DNA damage can occur as a result of normal cellular processes: metabolic byproducts damage DNA, as can the act of replication. But DNA damage can also be a result of exposure to exogenous **mutagens,** environmental agents like chemicals or radiation that damage DNA. Anything that increases the rate of DNA damage tends to also increase the possibility of cancer, since it shifts that balance toward damage rather than repair.

Most DNA repair processes detect DNA lesions due to their effect on the overall shape of the DNA double helix: lesions disrupt the nice, even helical structure of the molecule. The goal is to repair the DNA before it is replicated or distributed to daughter cells during mitosis. If a lesion is used as a template for replication, the resulting daughter may have the wrong complementary base installed. And if *that* strand is used as a template for replication, the result is a perfect double helix that can't be recognized by DNA repair proteins. Now, the cell usually cannot tell what the original base should have been and can no longer deploy repair mechanisms. It is at this point a lesion becomes a mutation.

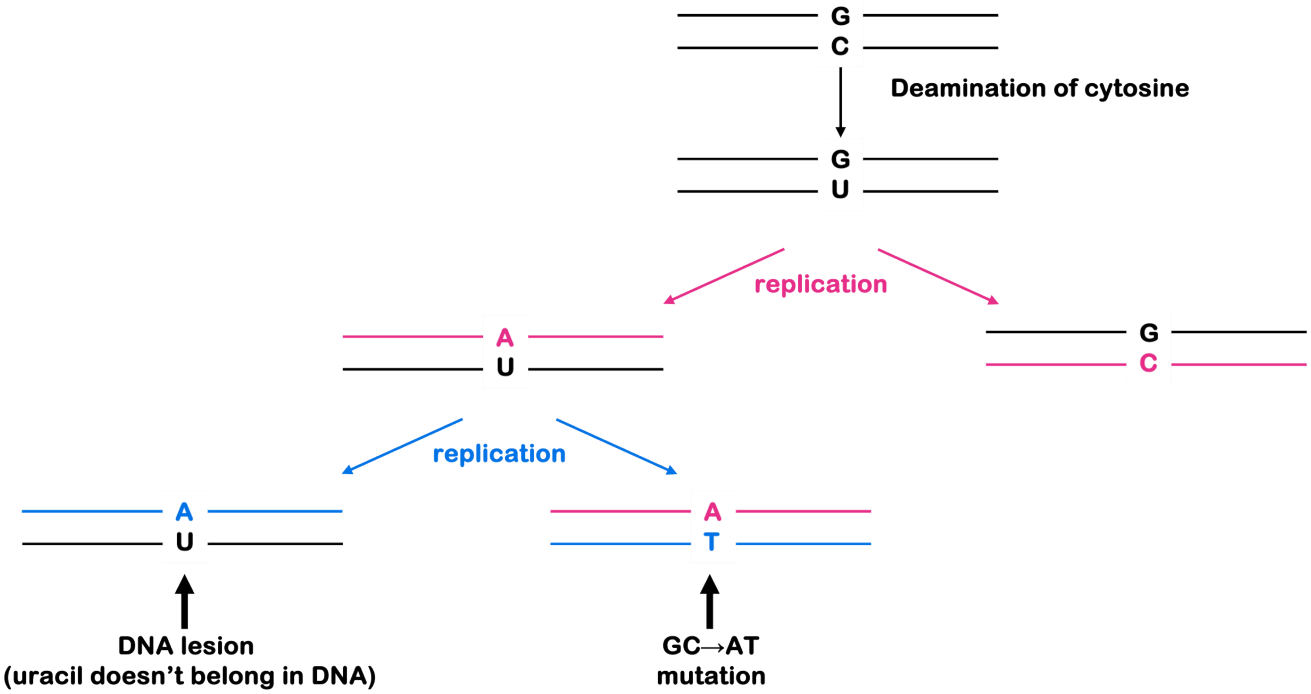


Figure 4. If DNA with a lesion is replicated, the replicated DNA may have a mutation.

There are a few methods by which cells can directly reverse DNA damage. One is dealkylation, which removes alkyl groups like —CH3 or —C2H5 from alkylated bases. The alkyl group is directly transferred to the protein **O6-alkylguanyl-DNA alkyltransferase**. A second example of direct reversal are the **photolyases**, which can undo the linkage between **pyrimidine dimers**. Pyrimidine dimers are caused by exposure to UV light, and these enzymes use the energy from light to catalyze the repair reaction. This is called photoreactivation. These lesions are shown in **Figure 5**. Photolyases are used by many organisms from bacteria to eukaryotes, but they are not found in humans or other mammals.

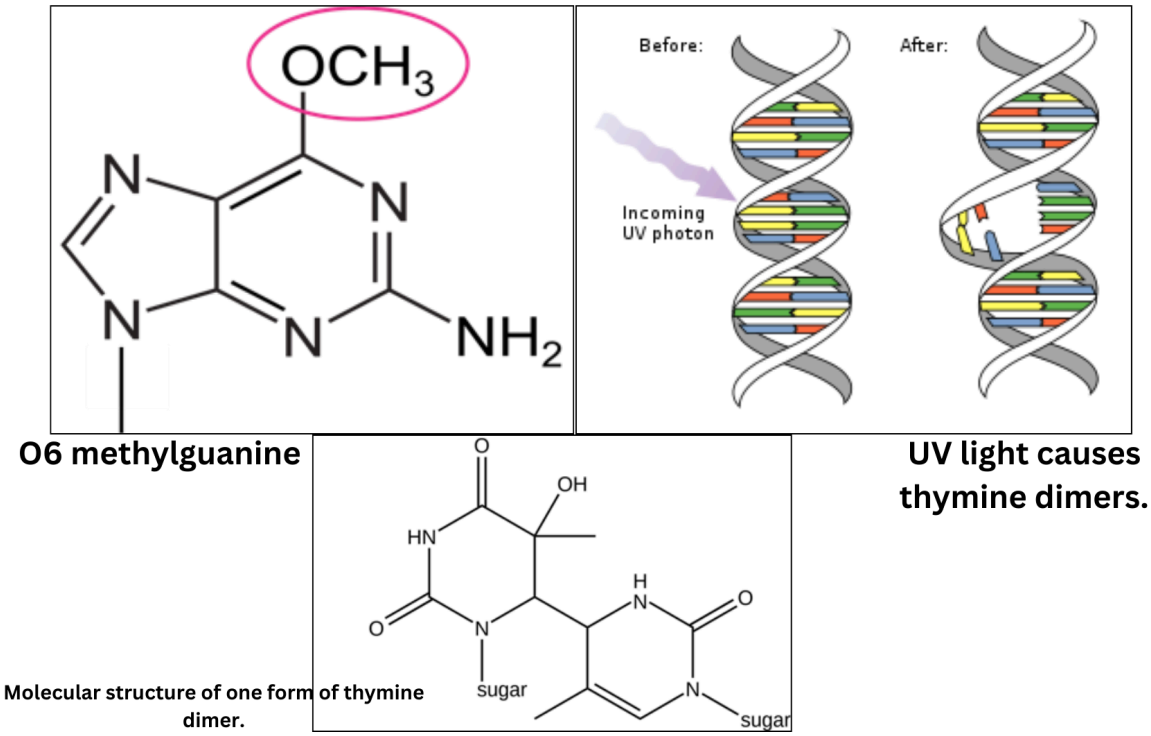


Figure 5. DNA lesions repaired through direct reversal. Left: O6 methylguanine. Right: UV light causes thymine dimers. Bottom Center: Molecular structure of one form of thymine dimer.

But most other forms of DNA repair do not directly reverse the chemical damage to DNA. Instead, they fix lesions indirectly by performing a similar series of steps:

1. Detection of the DNA lesion
2. Excision (removal) of the damaged part of the chromosome
3. Use of an undamaged strand as a template to synthesize replacement DNA
4. Ligation of old with new DNA

The enzymes involved in this process depend on the type of lesion.

- Damaged bases are repaired via base excision repair (BER)
- Larger DNA lesions affecting multiple nucleotides are repaired via nucleotide excision repair (NER)
- Mismatches in DNA caused by replication errors are repaired by the mismatch repair pathway
- Double-strand breaks are repaired via homologous recombination or non-homologous end joining

Each pathway includes a family of proteins that work together to perform these same basic steps of the repair process.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=733#h5p-98

Base Excision repair

Base excision repair recognizes and removes small lesions to single bases. This repair can include deaminated bases (like the conversion of cytosine to uracil), oxidative damage to bases (which happens as a result of normal cell metabolism and exposure to oxidative chemicals), or other damaged bases.

Base excision repair uses a family of enzymes called **glycosylases** to remove the damaged base. Glycosylases typically recognize a specific base: for example, there are uracil glycosylases that would repair the lesions shown in **Figure 4** (uracil never belongs in DNA). The removal of the damaged base leaves behind an **abasic site** (missing a base). The abasic nucleotide is cut out of the DNA with an endonuclease, a DNA polymerase fills in the gap, and the nick is sealed with ligase. This process is shown in **Figure 6**.

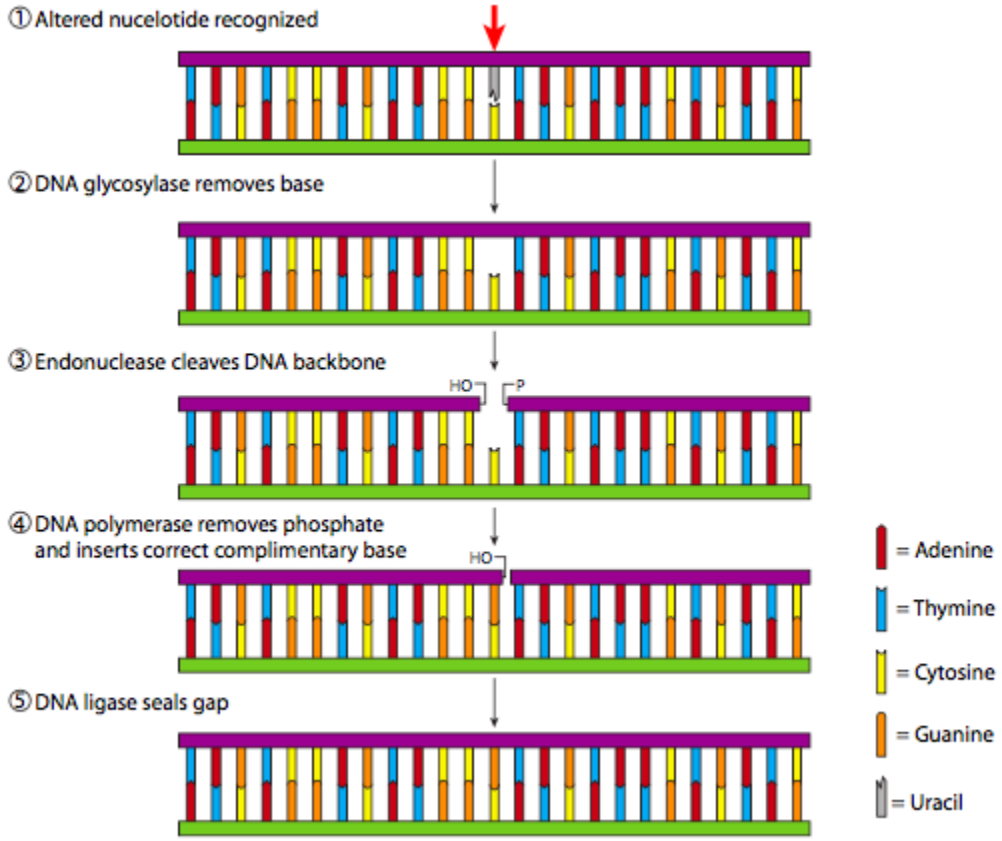


Figure 6. Base excision repair uses glycosylases to recognize and remove damaged bases. The nucleotide with the missing base is then excised by endonuclease, DNA polymerase fills in the gap, and ligase seals the nick in the backbone.

Nucleotide Excision Repair

If the DNA lesion is bulky or involves more than one base, nucleotide excision repair is used instead. In the absence of photolyases in mammals, nucleotide excision repair is used for pyrimidine dimers and other larger lesions to the bases.

In eukaryotes, many of the proteins involved in this process belong to the XP family of proteins, shown in **Figure 7**.

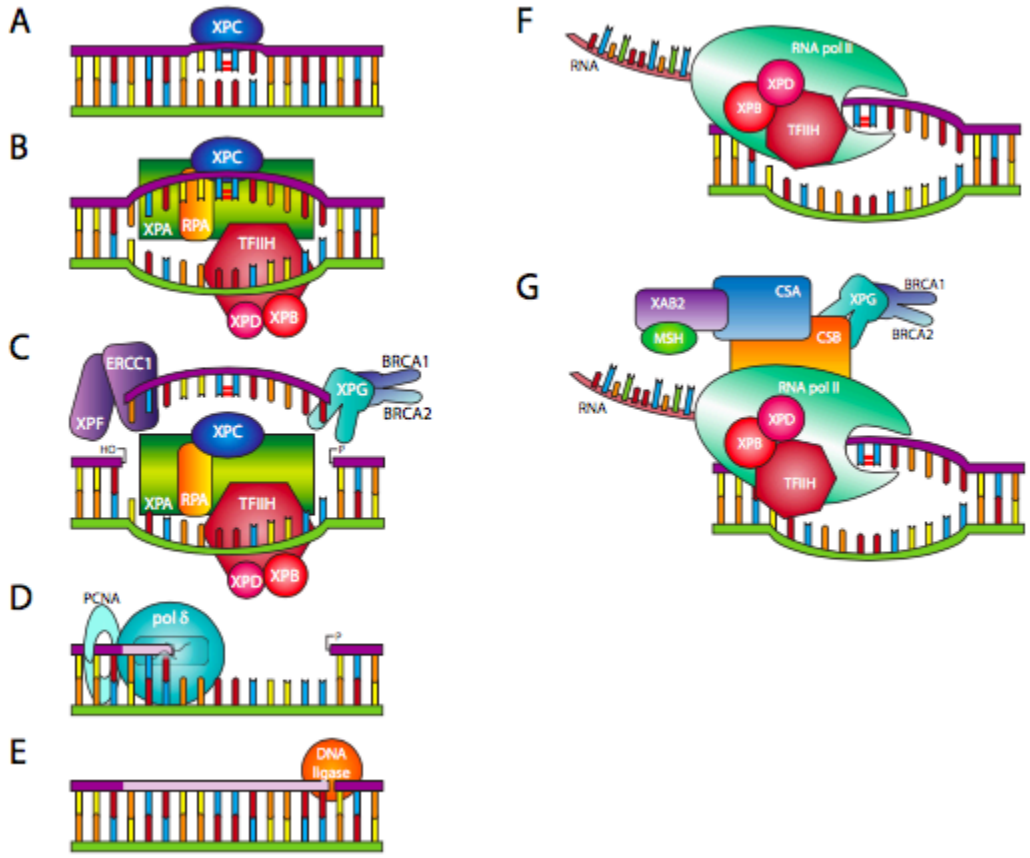


Figure 7. Nucleotide excision repair. A bulky lesion like a pyrimidine dimer is recognized by XP proteins. An entire segment of DNA is removed from around the lesion. DNA polymerase synthesizes new DNA across the gap, and DNA ligase seals the nick in the DNA backbone.

The XP family of proteins is named for the genetic disease Xeroderma Pigmentosum (or XP for short). XP is an autosomal recessive disorder that causes extreme sensitivity to UV light – even just a few minutes of exposure to sunlight results in severe sunburns. People with XP are up to 10,000 times more likely to develop skin cancer than the general population (that's right, 4 zeros). This likelihood of skin cancer comes about because they have a loss of function variants in the XP genes and they cannot repair the damage caused by UV light.

This explanation gives you an idea of how frequently healthy XP proteinsworkg to repair DN.–Mostt people can withstand more than a few minutes of sunlight because their cells can keep up with the repair.

Mismatch Repair

Mismatch repair recognizes mismatched bases in DNA and other replication errors. Mismatches occur during replication due to the misincorporation of bases. In the chapter on Mutation, we discussed how this can happen due to the tautomeric shifts of the bases, which can affect base pairing properties.

In prokaryotes, enzymes called MutS and MutL recognize the "kink" in the DNA caused by a mismatch. Additional proteins bind to the protein complex and remove part of the daughter strand around the mismatch. A DNA polymerase then synthesizes new DNA across the gap, and ligase seals the nick in the backbone. This process is shown in **Figure 8**.

How does the cell know which is parent and which is a daughter? In prokaryotes, DNA is methylated at particular sites in the genome, usually on both strands of the double helix. But immediately after replication, the new strand still needs to be methylated. You can see this in **Figure 8**, where the green oval representing a methyl group is only seen on the top strand. The daughter strand is the one without methyl groups. Mismatch repair doesn't work for more than a few minutes after replication, because after that both strands of the double helix are methylated.

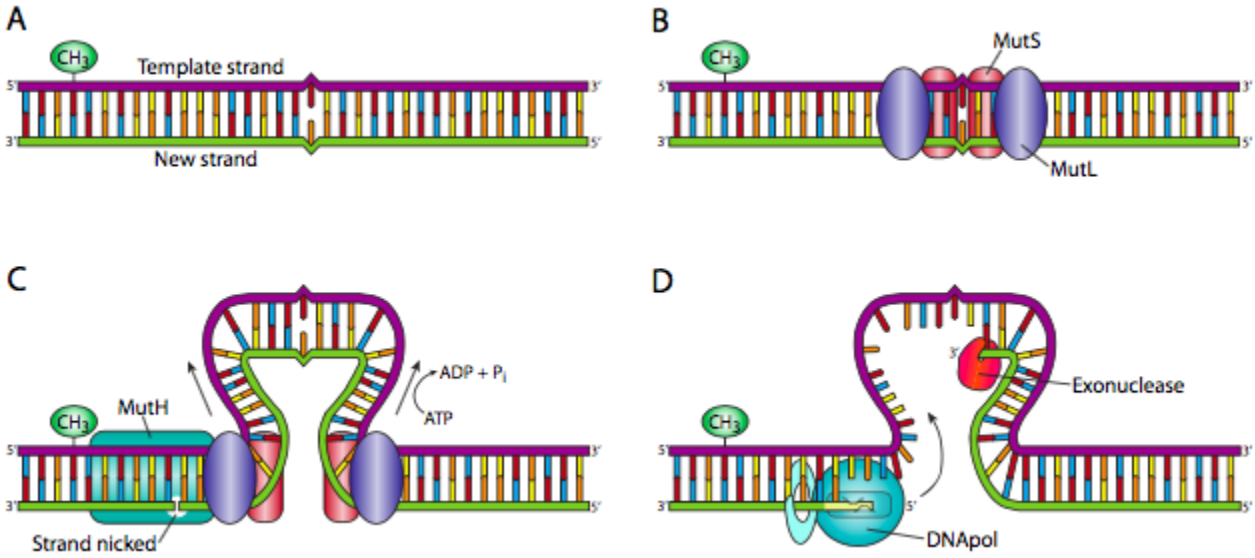


Figure 8. Mismatch repair. MutS and MutL proteins recognize a mismatch shortly after replication. MutH nicks the DNA and unwinds the region around the lesion. DNApol synthesizes new DNA across the gap, and ligase seals the nick.

There is no methylation to mark the parent strand in eukaryotes. Instead, eukaryotes use a similar mechanism but detect the daughter strand by the presence of Okazaki fragments. Again, the process must be completed within minutes of replication since the lagging strands are ligated together very quickly after replication.

In eukaryotes, there are families of homologous proteins to MutS and MutL called MSH (for MutS Homolog) and MLH (for MutL Homolog).

In humans, the MSH and MLH family of proteins are associated with Lynch syndrome, a form of hereditary colorectal cancer. MSH and MLH mutations are also associated with microsatellite instability, which is the expansion or shrinkage of short repeated elements in the genome. These are also replicative errors.

Double-strand break repair

Some of the most damaging lesions to the cell are double-strand breaks, which are breaks across both strands of the DNA backbone. DNA breaks can be caused by exposure to ionizing radiation (like X-rays) as well as certain kinds of chemicals. Other types of DNA damage can lead to double-strand breaks if they cause the replication fork to get "stuck" during replication.

Unlike the previous lesions in this chapter, double-strand breaks can affect many genes all at once: any part of the chromosome separated from the centromere will not be properly sorted during mitosis or meiosis. Thus, failure of the double-strand break repair pathway can lead to aneuploidies with missing parts of a chromosome. Improper repair of breaks also contributes to translocations and inversions.

A double-strand break, shown at the top of **Figure 9**, can be repaired via one of two mechanisms: Non-homologous end-joining (NHEJ) or homology-directed repair (HR).

Non-homologous end-joining (shown on the left) is pretty much exactly what it sounds like. NHEJ proteins recognize the break, recruit additional proteins to trim up the end, and a ligase joins the ends back together. But this is an error-prone process. Nucleotides are always lost around the break. So why might a cell choose to repair in a way that is certain to introduce mutations? Remember that the majority of the eukaryotic genome is not a coding sequence. Hence, a loss of a few nucleotides is preferable to the loss of a whole chunk of chromosome.

Homology-directed Repair, on the right in **Figure 9**, uses many of the same proteins used during homologous recombination (crossing over) in meiosis. In homology-directed repair, the double-strand break is recognized and processed to expose a long single-stranded region of DNA. The single-stranded region is used to search for homology, preferably in a sister chromatid, which should share identical sequence to the broken DNA. When a match is found, the single-stranded region invades the intact double-helix, pushing the second strand out of the way and base pairing with its complement. The intact strands are then used as a template for DNA polymerase to synthesize new DNA across the gap. DNA ligase seals nicks in the backbone, and then the crossovers are resolved, or cut apart. In bacteria, RecA is the major recombinase involved in strand invasion, and RecBCD participates in later steps. In eukaryotes, protein complexes like MRN, RPA, Rad51, and other genes are linked with cancer phenotypes .

At first glance, this seems like homology-directed repair would be the best choice to repair double-strand breaks. But it's not without downsides, either. A sister chromatid is only available in the second half of the cell cycle – after DNA synthesis. While a homologous chromosome can be used for homology-directed repair in the G1 phase, this runs the risk of **gene conversion** if the paternal and maternal chromosomes have different alleles. Gene conversion occurs when a cell or organism starts as heterozygous but ends with two of the same allele. It happens because one allele is used to patch the other. While gene conversion is often not so bad, on rare occasions, gene conversion can lead to tumorigenesis, as discussed

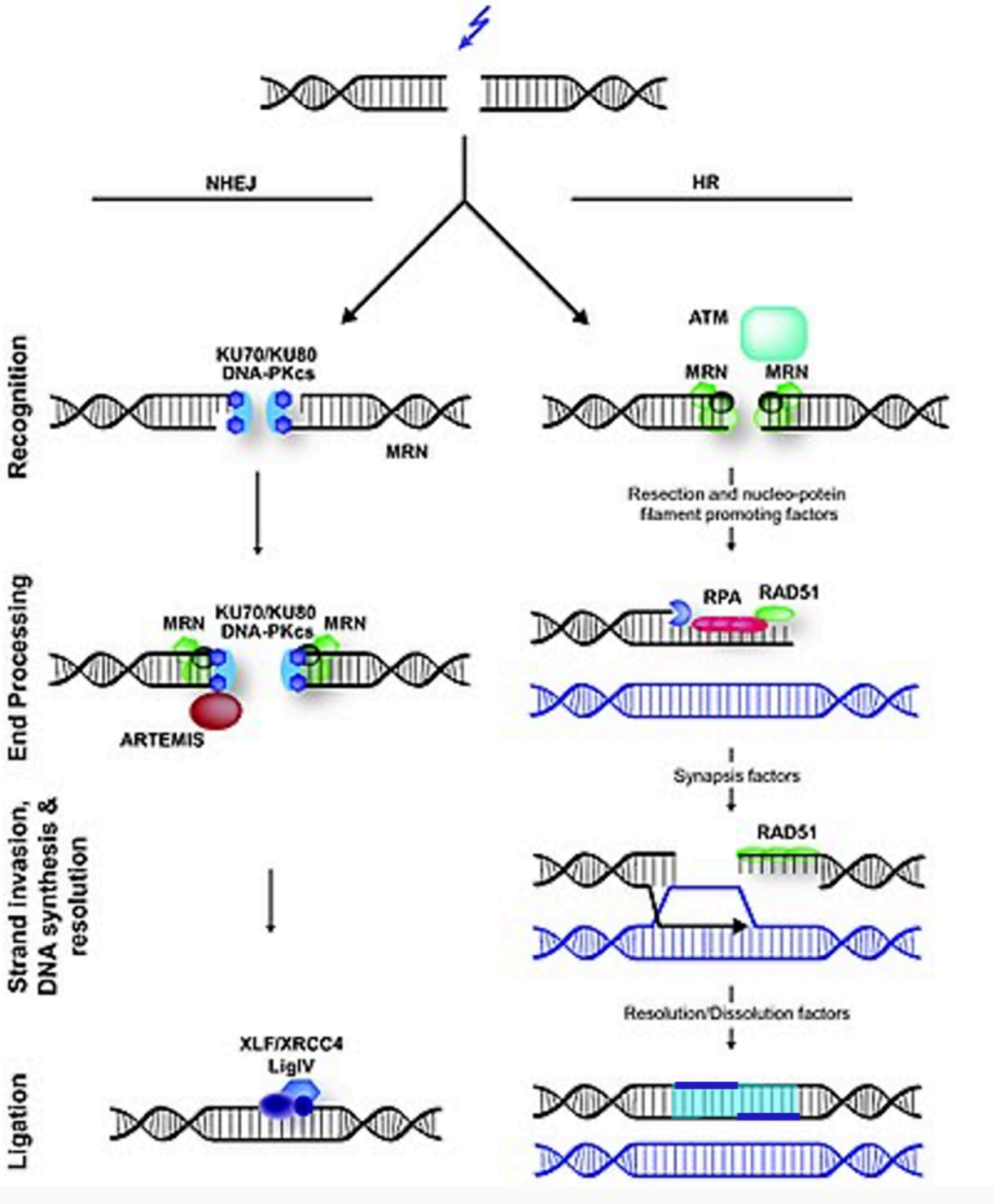


Figure 9. Left: Overview of non-homologous end-joining. Right: Overview of homology-directed repair.

Summary of repair mechanisms

Whoa, that's a lot! Here's a summary of the types of DNA damage and their repair in eukaryotes.

Type of lesion	Repair Mechanism	Selected Proteins Involved
Damaged base like deamination, oxidative damage	Base excision repair	Glycosylases
Large, bulky DNA lesions	Nucleotide excision repair	XP family of proteins
Replicative errors	Mismatch repair	MSH, MLH family of proteins
Double-strand breaks	Non-homologous end joining	MRN, Ku70/80, DNA-PKcs, XRCC4, Ligase IV
Double-strand breaks	Homology-directed repair	MRN, Rad51, BRCA1, BRCA2 and others

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=733#h5p-99

Media Attributions

- Figure 4 DNA Repair © Amanda Simons is licensed under a CC BY (Attribution) license
- O6 methylguaninet © Amanda Simons is licensed under a CC BY (Attribution) license
- Figure 6 DNA Repair © Molecules and Mechanisms (Wong) is licensed under a CC0 (Creative Commons Zero) license
- Figure 7 DNA Repair © Molecules and Mechanisms (Wong) is licensed under a CC0 (Creative Commons Zero) license
- Figure 8 DNA Repair © Molecules and Mechanisms (Wong) is licensed under a CC0 (Creative Commons Zero) license
- Figure 9 DNA Repair © Modified from: Ingo Schubert, Veit Schubert, Jörg Fuchs is licensed under a CC BY (Attribution) license

CANCER IS CAUSED BY DYSREGULATION OF THE CELL CYCLE

Not all mutations have the potential to lead to cancer. Remember: many mutations have no phenotypic effect on either a cellular or organismal level. Other mutations may cause cell death or a cellular phenotype not associated with proliferation. For example, a somatic mutation in a gene important for cell metabolism would probably just lead to the death of the cell – it would not trigger uncontrollable growth. However, certain mutations in two types of genes do lead to inappropriate cell growth: **proto-oncogenes** and **tumor suppressors.**

Both proto-oncogenes and tumor suppressors play important roles in regulating the cell cycle.

Cell proliferation is a part of the normal physiology of a healthy organism as new cells are needed to repair injured or worn-out cells. Different tissues undergo cell division at different rates. For example, the cells lining the small intestine are normally replaced every 2-4 days, but cells of the central nervous system divide rarely, if at all, over the course of a lifetime[1]. (This is why spinal cord injuries do not heal.) However, the process of cell proliferation is tightly controlled to ensure that cells only divide when appropriate.

You'll recall from the chapter on Meiosis and Mitosis that there are four stages to the cell cycle: G1, S, G2, and M (**Figure 10**). Transition from one stage to the next is tightly regulated in healthy tissue: cell division only happens under appropriate conditions. There is both positive and negative regulation of the cell cycle.

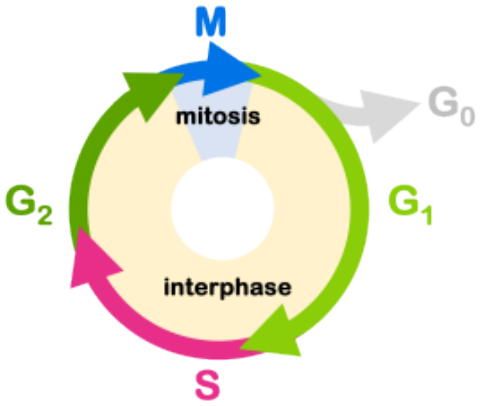


Figure 10. Cell Cycle

1. Philips, R. M. & R. » How quickly do different cells in the body replace themselves? http://book.bionumbers.org/how-quickly-do-different-cells-in-the-body-replace-themselves/.

We can use the analogy of a car to describe this: the gas pedal positively regulates the car's movement, and the brake pedal negatively regulates the car's movement. For the car to go, the brake must be off and the gas must be on. For the cell cycle to proceed, the cell must release the negative regulation and activate the positive regulators.

Proto-oncogenes

Proto-oncogenes are positive regulators of the cell cycle. These are proteins that signal for the cell to divide – but only when conditions are right. Many of these positive regulators participate in signaling cascades that respond to growth factors. They are called "cascades" because the growth factor sets off a series of events that lead to cell division, much like a chain of dominos triggered by one push. One example of such a signaling cascade is shown in **Figure 11**.

In **Figure 11** an extracellular signaling protein called a cytokine binds to a receptor on the cell surface, initiating a series of events inside the cell that leads to the transcription of genes involved in the immune system. This includes the proliferation of immune cells, which is necessary for an immune response. If any of the genes involved sustain a gain of function mutation, this can result in cell proliferation even in the absence of the extracellular signal.

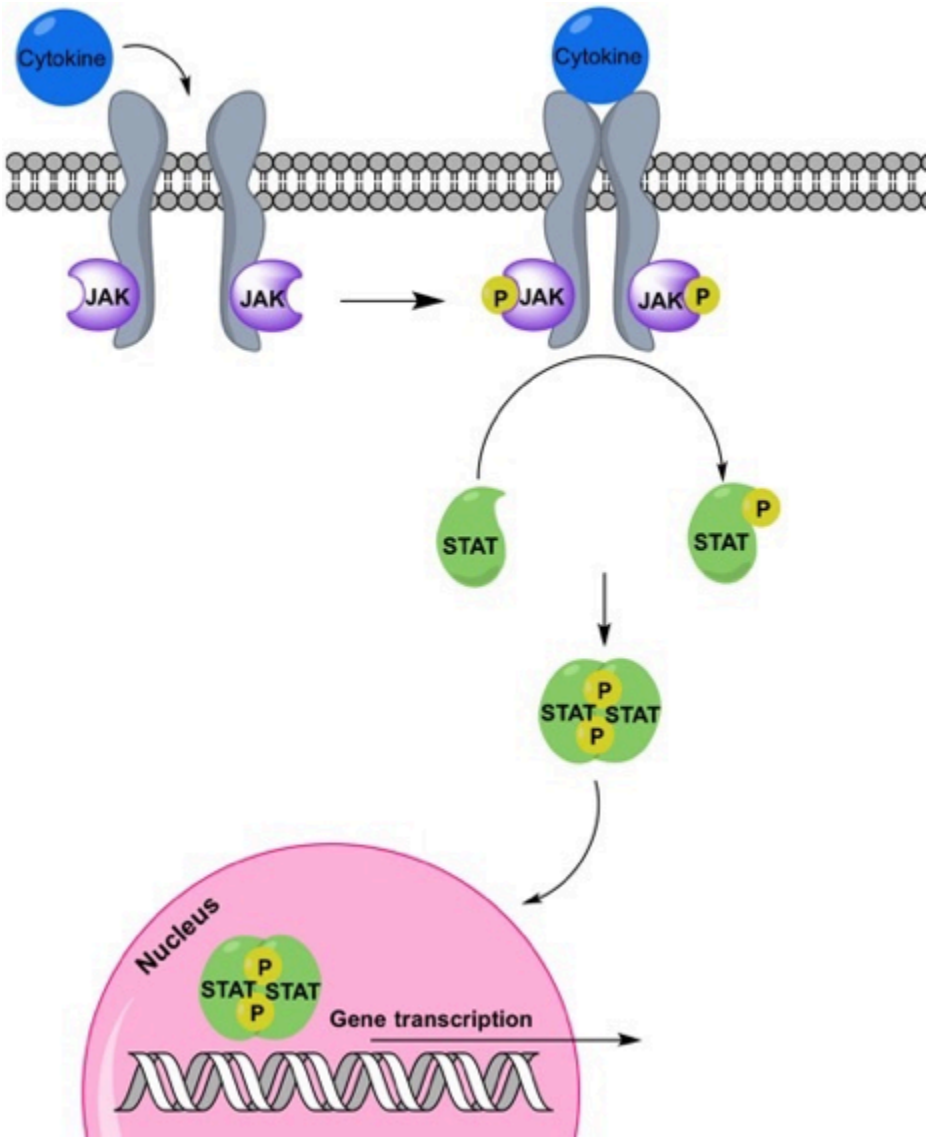


Figure 11. JAK-STAT signaling cascade activates genes involved in cell proliferation. An extracellular signal called a cytokine binds to a receptor on the cell surface (shown in gray). Cytokines are small proteins that are involved in regulating the immune response, including the proliferation of immune cells in response to an infection. Cytokine binding leads to the phosphorylation of JAK (purple). The activated JAK protein in turn, phosphorylates STAT (green). The phosphorylated STAT is transported to the nucleus where it acts as a transcription factor, activating genes involved in cell proliferation. Both JAK and STAT are representative of a family of proteins. In mammals there are four JAK proteins and seven STAT proteins. Mutations in JAKs, STATs, and transcriptional targets are associated with oncogenesis.

Gain of function mutations in proto-oncogenes convert the genes to **oncogenes**. Oncogenes cause the cell to divide regardless of whether conditions are appropriate. They've lost the regulatory part of their job. The

family of proteins shown in the signaling cascade in **Figure 11** – JAK, STAT, and many of their transcriptional targets – are mutated in many different kinds of cancers[2].

Although some oncogenic mutations can be point mutations, other larger chromosomal rearrangements are also common in the conversion of proto-oncogene to oncogene. Some of these are shown in **Figure 12.** For example, gene duplication can result in additional copies of a gene, which leads to extra protein production (and gain of function). A translocation can bring together a proto-oncogene with the regulatory region of another gene, resulting in the misexpression of the oncogene. A translocation can also fuse the coding sequences of two genes together, resulting in a fusion protein that is improperly regulated and overactive.

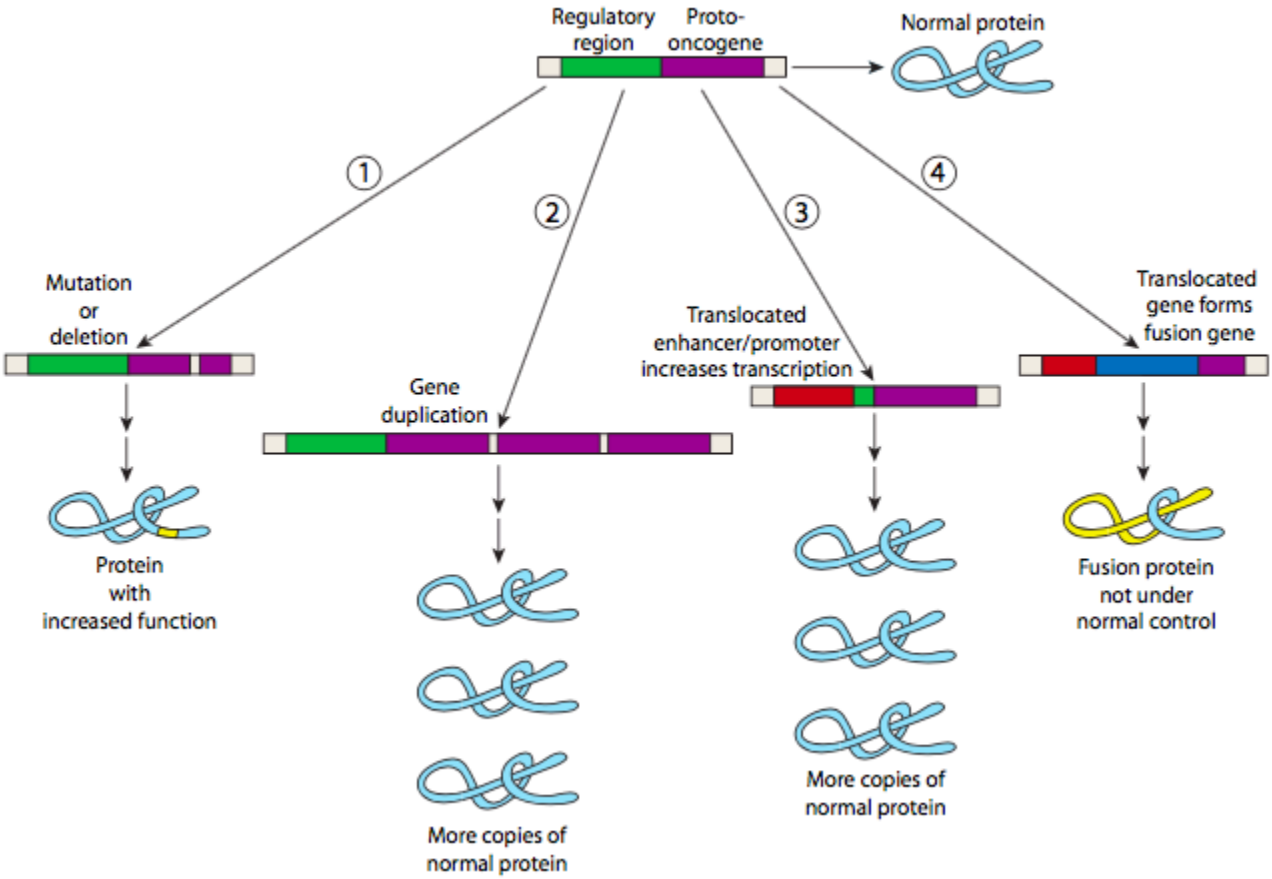


Figure 12. Types of mutation that result in gain of function.

To go back to the analogy of a car, these oncogene mutations are akin to a gas pedal that is stuck to the floor.

Because these are gain-of-function mutations, they are typically dominant on a cellular level. Only one allele of a proto-oncogene needs be mutated to cause overactive growth of a cell.

2. Bromberg, J. Stat proteins and oncogenesis. J. Clin. Invest. 109, 1139–1142 (2002).

The somatic gain of function mutations in proto-oncogenes is the only type of mutation in these genes that would lead to cancer. Other types of mutations in proto-oncogenes do not lead to cancer. For example, a somatic *loss of function* mutation in a proto-oncogene would not lead to cancer because loss of function would prevent a mutant cell from proceeding through the cell cycle. And *germline* mutations in proto-oncogenes are typically embryonic lethal since proper regulation of the cell cycle is necessary for a zygote to grow into an adult multicellular organism.

Tumor suppressors

If proto-oncogenes are the gas pedal of the cell cycle, the brakes are tumor suppressors. Tumor suppressor genes are involved in cell cycle checkpoints, putting the brakes on until conditions are right and ensuring the integrity of the genome.

The cell cycle is negatively regulated by checkpoint proteins that monitor conditions inside and outside the cell to be sure it is appropriate for the cell to divide. Four checkpoints are shown in Figure 13. The G1 checkpoint blocks cell cycle progression from G1 to S, monitoring for adequate cell size and nutrients as well as DNA damage. The G1 checkpoint proteins will block the transition to S phase until conditions are appropriate and any DNA damage is repaired.

The intra-S checkpoint ensures that replication is completed appropriately, since incompletely replicated DNA cannot be properly divided among daughter cells. These checkpoint proteins block cell cycle progression until replication is complete. They also monitor for DNA damage.

The G2/M checkpoint monitors for DNA damage and ensures replication is complete before the transition to M phase. And the M checkpoint monitors for spindle assembly, blocking progression if chromosomes are not properly attached to the spindle.

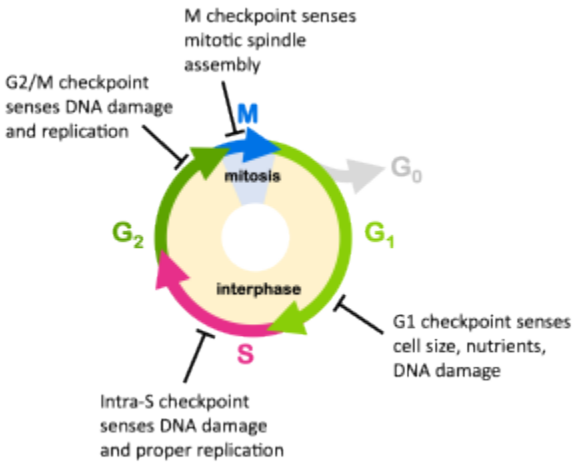


Figure 13. cell cycle checkpoints. Image source: AS CC BY SA

You'll notice from **Figure 13** that virtually all of the checkpoints are meant to prevent mutations: the G1, S, and G2/M checkpoints sense for DNA damage which would lead to mutations if not repaired, while the M checkpoint senses for chromosomal attachment to the spindle, which could lead to the gain or loss of entire chromosomes after cytokinesis.

Some tumor suppressors directly block the cell cycle, allowing other tumor suppressors to act in DNA repair. You can think of these two tumor suppressor functions as gatekeeper proteins and caretaker proteins. Gatekeepers block the cell from proceeding through the cell cycle, while caretaker proteins repair DNA and ensure genomic integrity.

Loss of function mutations in either type of tumor suppressor allows the cell cycle to proceed, even if DNA damage is present. These mutations are usually recessive on a cellular level; if just one healthy copy is present, this is typically enough to prevent a cell from becoming cancerous. This is shown in the cartoon in **Figure 14.**

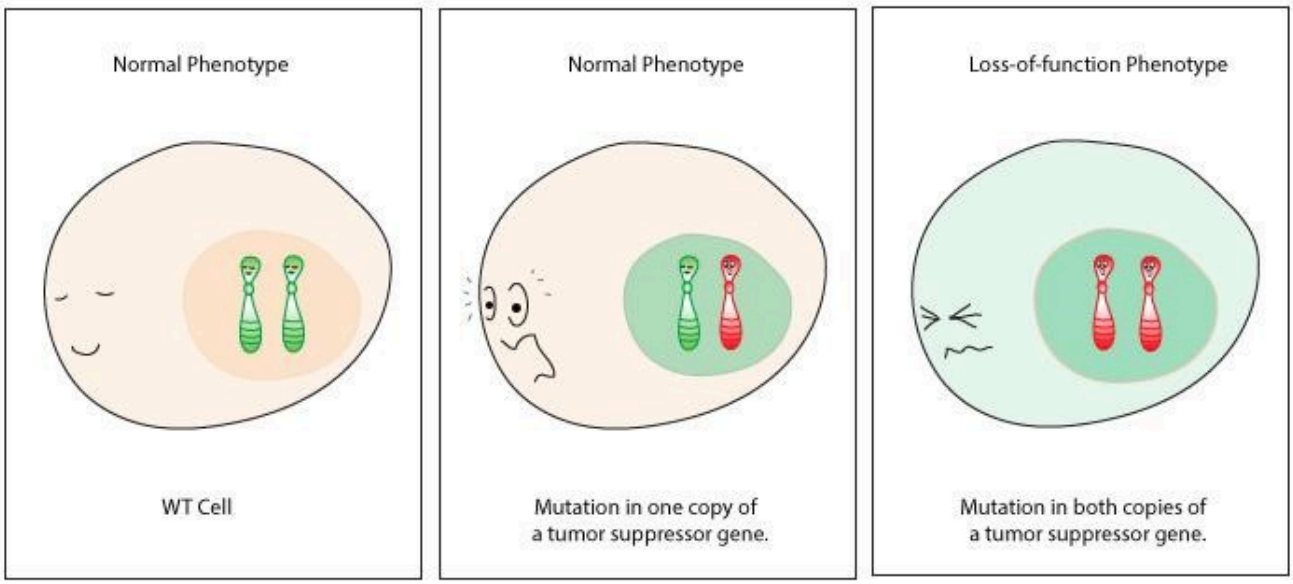


Figure 14. A loss of function mutation in both copies of a tumor suppressor gene is necessary to alter cellular phenotype.

As with the proto-oncogenes, only a subset of mutations in tumor suppressor genes can lead to cancer. It typically requires loss of function mutations in *both* copies of a tumor suppressor for a somatic cell to become cancerous. A somatic gain of function would not lead to oncogenesis – in fact, evolutionarily, there is evidence that gain of function mutations in tumor suppressors may protect against cancer. And a *germline*, homozygous loss of function in a tumor suppressor is often embryonic lethal: without that function, an early embryo likely accumulates too many mutations to be compatible with life.

Media Attributions

- Figure 10 DNA Repair © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 11 DNA Repair is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 12 DNA Repair © Molecules and Mechanisms(Wong) is licensed under a CC0 (Creative Commons Zero) license
- Figure 13 DNA Repair © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license
- Figure 14 DNA Repair © WassermanLab is licensed under a CC BY-SA (Attribution ShareAlike) license

EXAMPLES OF TUMOR SUPPRESSOR PROTEINS

Most of the DNA repair proteins mentioned in the first half of this module are tumor suppressor proteins that act as caretakers. If DNA repair pathways fail, this results in a rapid accumulation of mutations. The rapid accumulation of mutations makes it more likely for a proto-oncogene to be mutated.

Two examples of gatekeeper tumor suppressors are Rb and p53. Rb was the first identified tumor suppressor, with mutations in Rb linked with the rare childhood cancer retinoblastoma. Rb mutations have since been found in other cancers as well[1]. Mutations in p53 are found in 50% of all human cancers. p53 plays a central role in multiple mechanisms of tumor suppression.

Rb

Retinoblastoma is a tumor of the retina. It is most diagnosed in children under the age of five. Historically, early stages of retinoblastoma were visible as white- or yellow-reflecting eyes in flash photography, shown on the left in **Figure 15**. While a healthy retina often produces a typical "red-eye" reflection in flash photography, the reflection of the flash against the tumor would produce a white or yellow reflection. (Note that with modern cell phone cameras the red-eye reduction features can produce yellow or white reflections too!) On the right in **Figure 15** is the imaging of a patient's retina, with the tumor visible as a lighter colored mass.

1. Engeland, K. Cell cycle regulation: p53-p21-RB signaling. Cell Death Differ. 29, 946–960 (2022).

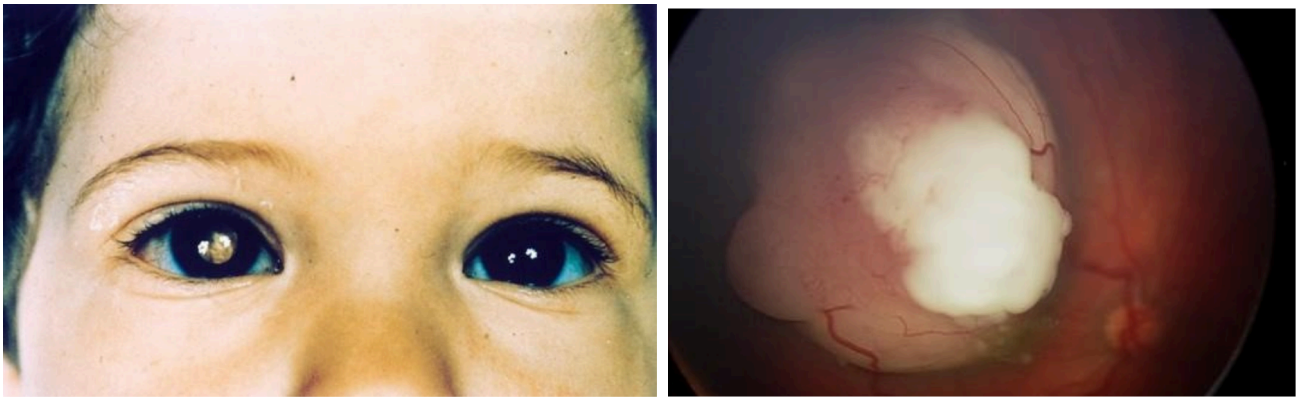


Figure 15. Retinoblastoma is a cancer of the retina. It most commonly occurs in children under five. The tumor can be seen in certain flash photographs as a white reflection in the eye as shown on the left. On the right is an image of the retina in a patient with retinoblastoma. The tumor is the light-colored mass within the red retinal tissue.

Figure 15. *Image sources:*
(left) http://visualsonline.cancer.gov/details.cfm?imageid=2418 , Public Domain; (right) Cieslik et al[2] https://www.researchgate.net/figure/Retinoblastoma-before-treatment-Visible-perforation-of-the-internal-limiting-membrane-in_fig1_371224804 CC BY 3.0

It most commonly occurs in children under five. The tumor can be seen in certain flash photographs as a white reflection in the eye as shown on the left. On the right is an image of the retina in a patient with retinoblastoma. The tumor is the light-colored mass within the red retinal tissue.

Retinoblastoma is caused by loss of function mutations in Rb. As shown in **Figure 16**, in healthy cells Rb blocks the transcription factor E2F. Cell cycle progression leads to the phosphorylation of Rb, which releases E2F. E2F transcribes genes required for the transition to S-phase of the cell cycle. A loss of function in Rb means E2F (and its downstream genes) are constantly active, leading to inappropriate cell cycling and proliferation. This is what leads to retinoblastoma tumors.

With modern medicine, most children in the United States with retinoblastoma are successfully treated through a combination of surgery and drug therapies.

2. Cieslik, K., Rogowska, A., Danowska, M. & Hautz, W. Efficacy of intravitreal injections of melphalan in the treatment of retinoblastoma vitreous seeding. Adv. Clin. Exp. Med. Off. Organ Wroclaw Med. Univ. 33, (2023).

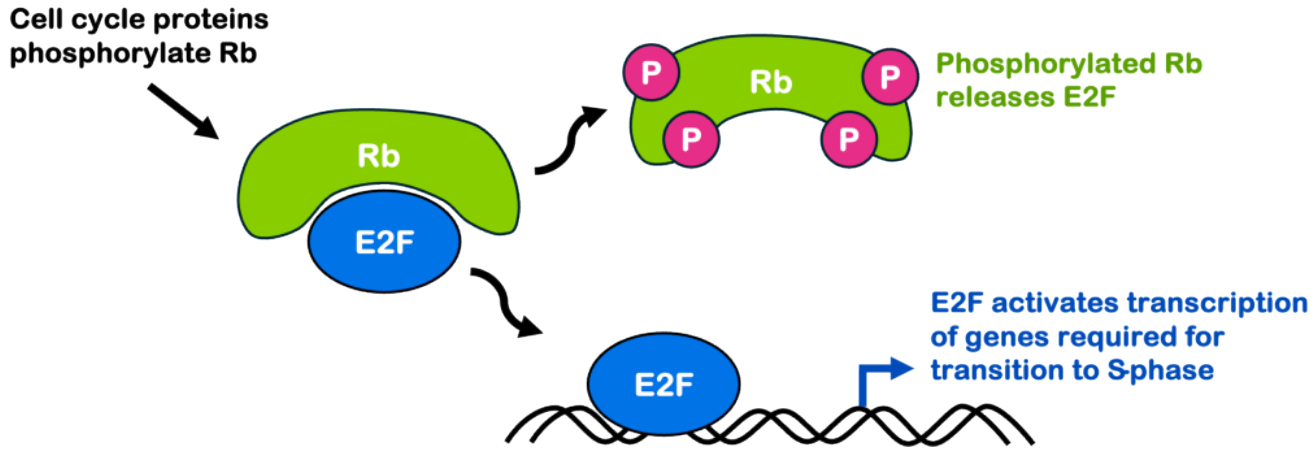


Figure 16. Rb is a tumor suppressor. Rb inactivates the E2F transcription factor. Cell cycle signaling proteins phosphorylate Rb, causing it to release E2F, which then activates the transcription of genes required for the transition to S phase.

p53

p53 has been called the "guardian of the genome", with good reason. It can act as a sensor for DNA damage, relaying the signal to downstream targets and allowing the cell to respond in multiple ways. The cartoon in **Figure 17**, reprinted from a recent review article[3], is a bit complicated but gives a good overview of how many different ways p53 can act.

p53 is normally present at only low levels in the cell. DNA damage (shown at the top of the image) can be sensed by a variety of proteins in the cell. These are labeled "sensors" in this image. Those sensors relay a signal to **transducers,** which in turn stabilize p53 in the cell and allow it to accumulate at higher amounts. p53 is a transcription factor that, in turn, activates transcription of genes involved in DNA repair and cell cycle arrest.

p53 also can promote **apoptosis** and **senescence**. Apoptosis is a controlled form of cell death. A cell undertakes a series of molecular steps that will eventually trigger its own death. It might be counter-intuitive to think of cell death as a protective measure, but for a multicellular organism, the loss of one cell is not that big of a deal. It would be far worse for a damaged cell to become cancerous since that can potentially kill the whole organism. Senescence is when a cell permanently ceases cycling.

3. Vadivel Gnanasundram, S., Bonczek, O., Wang, L., Chen, S. & Fahraeus, R. p53 mRNA Metabolism Links with the DNA Damage Response. Genes 12, 1446 (2021).

When a cell sustains low levels of DNA damage, p53 likely promotes cell cycle arrest and DNA repair, followed by re-entry into the cell cycle. If a cell sustains too much damage, though, p53 promotes apoptosis.[4]

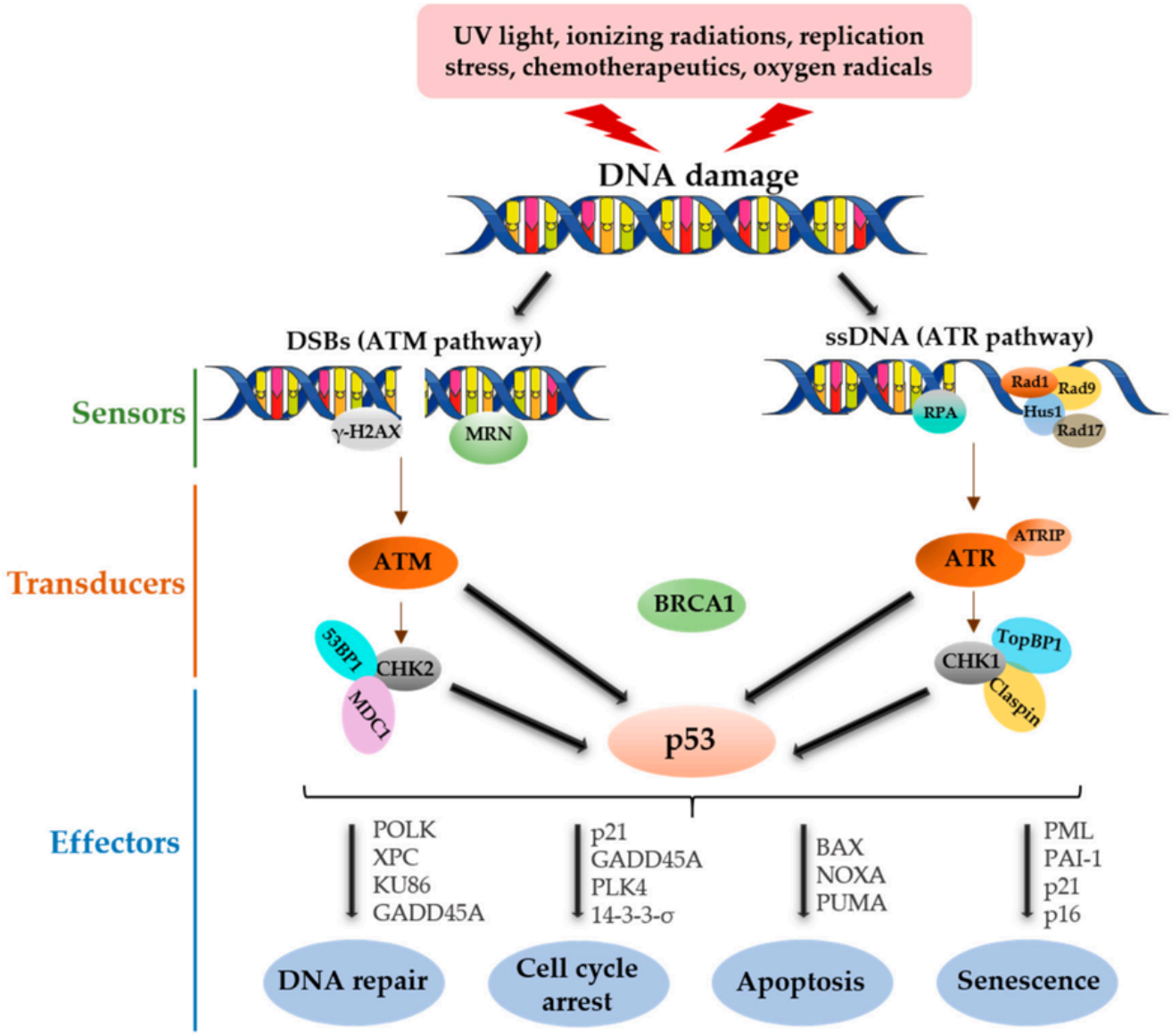


Figure 17. p53 is central to multiple mechanisms of tumor suppression.

Media Attributions

- Figure 15 DNA Repair combined © Cieslik et al is licensed under a CC BY-SA (Attribution ShareAlike)

4. Chen, J. The Cell-Cycle Arrest and Apoptotic Functions of p53 in Tumor Initiation and Progression. Cold Spring Harb. Perspect. Med. 6, a026104 (2016).

CANCER REQUIRES MULTIPLE MUTATIONS

The mechanism by which a cell gains a growth advantage can vary widely from cancer to cancer. There are likely over a thousand human tumor suppressors and proto-oncogenes that can contribute to oncogenesis. This makes cancers genetically distinct from patient to patient and tumor to tumor. Even cancers of the same tissue can have a different combination of mutations.

In healthy cells, proto-oncogenes promote cell division, but only when conditions are right. When mutated, oncogenes cause cell division all the time (even when conditions are inappropriate). Rapidly dividing cells are more likely to accumulate additional mutations, since each round of replication is a chance to make a mistake. This rapid division of cells is described as **genomic instability.** If tumor suppressors are mutated, the mutation leads to a loss of cell cycle inhibition and/or a rapid accumulation of mutations, including additional oncogenic mutations that lead to overactive cell division.

In this respect, mutations in either a proto-oncogene or tumor suppressor can make it more likely to accumulate even more mutations.

Typically, mutations in both proto-oncogenes and tumor suppressors are needed for a cell to become cancerous – and usually mutations in multiple proto-oncogenes and multiple tumor suppressors. One estimate is that there are, on average, 5-6 mutations driving most cancers although there are notable exceptions with fewer mutations or far more.[1] Because of this, oncogenesis typically happens in stages, with each acquired mutation giving a cell (and its offspring) a slight growth advantage over surrounding healthy tissue.

An example of colon cancer is shown in **Figure 18**. Typically, each stage is associated with a slight growth advantage. So, in stage 1, the loss of APC tumor suppression confers a slight growth advantage to a mutant cell, leading to the growth of a polyp in the colon. In subsequent stages, the rapid growth increases the likelihood of acquiring more mutations. More mutations overall mean an increased chance an oncogene will be mutated in one cell of the polyp. This, in turn, makes the cell divide more often, making the cell (and its descendants) more likely to acquire additional mutations. The cell growth and genomic instability are therefore interconnected.

1. Knudson, A. G. Two genetic hits (more or less) to cancer. Nat. Rev. Cancer 1, 157–162 (2001).

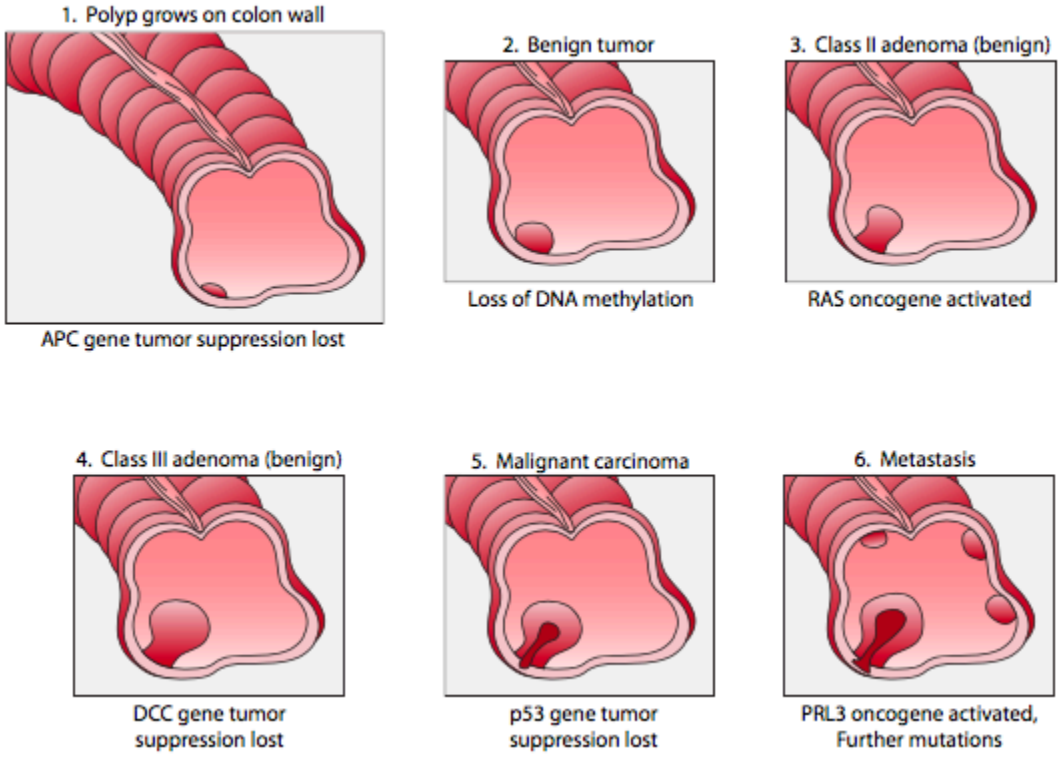


Figure 18. Development of colon cancer takes time and multiple mutations. An individual cancer-causing mutation generally creates a problem that can be corrected by some other cellular mechanism. Therefore, the development of cancer comes about through the accumulation of multiple mutations and not the acquisition of just one. One example of this gradual development of cancer is colon cancer. There is a fairly characteristic progression of mutations in the genes APC, RAS, DCC, TP53, and PRL3. Note that the progression depicted here is not inevitable: the presence of polyps does not lead invariably to colon cancer. RAS and PRL3 are oncogenes, while APC, TP53, and DCC are tumor suppressor genes.Image source: reprinted from Wong, Cells – Molecules and Mechanisms.

Because oncogenesis occurs in stages, the cells of a single tumor can be quite heterogeneous genetically, as illustrated in **Figure 19**. This can make it challenging for treatment, as not all cells of a tumor may respond equally to a drug regimen. As a cancer is treated, cells will continue to divide and acquire mutations, even as other cells are killed by a regiment. This can ultimately lead to a drug-resistant tumor, in much the same way as antibiotic misuse can lead to antibiotic resistance.

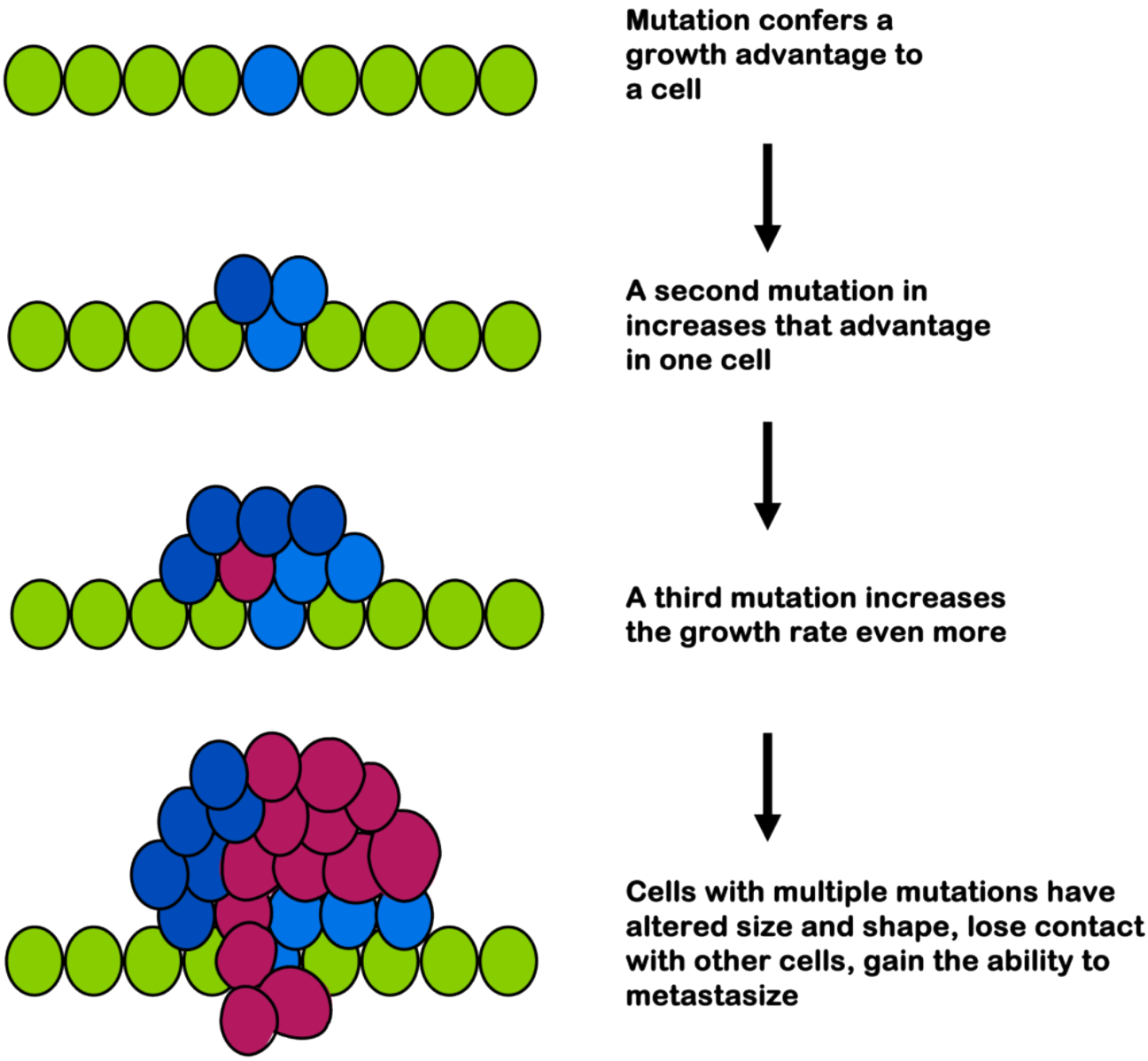


Figure 19. Tumors are genetically heterogeneous

In later stages, cells may appear de-differentiated, as if they've "forgotten" the type of cell they were. They have altered size and shape compared to their surroundings. And they no longer stop growth on contact with surrounding cells: they can invade surrounding tissues.

As the tumor grows and additional mutations accumulate, most tumors acquire certain characteristics that distinguish them from their healthy surroundings. These are called the Hallmarks of Cancer, first proposed by Douglas Hanahan and Robert Weinberg in 2000, with updates in 2011 and 2022 as the physiology of cancer became better understood.

Among the original Hallmarks of Cancer:

1. Cell division without growth signals
2. Cell division in the presence of inhibitory signals
3. Evading cell death
4. Cell immortality (activation of telomerase, for example)
5. Accessing or inducing vasculature (blood vessels)
6. Metastasis

To these hallmarks were later added: avoiding immune detection, inflammation, genome instability, and deregulating cell metabolism. And most recently: epigenetic changes, microbiome influences, senescent cells, and reshaping cell fate (including de-differentiation or assuming a different cellular phenotype).

Because these hallmarks distinguish cancer cells from healthy tissue, they are attractive targets for cancer therapeutics. Cancer therapeutics are designed to kill cells with these characteristics.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/genetics/?p=802#h5p-100

Media Attributions

- Figure 18 DNA Repair © Cells – Molecules and Mechanisms (Wong) is licensed under a CC BY (Attribution) license
- Figure 19 DNA Repair © Amanda Simons is licensed under a CC BY-SA (Attribution ShareAlike) license

FAMILIAL VS SPORADIC CANCERS

Although cancer is a genetic disease, it is not inherited. A person who has cancer will not pass that tumor to their offspring because it arose from somatic mutations. But there are examples where a family may have a higher-than-expected incidence of cancer. These are called **familial** cancers, in contrast to **sporadic** cancers that occur in people with no family history of the disease.

Familial incidence of cancer occurs because family members share a disease-associated allele in a tumor suppressor gene. An important distinction is that the cancer itself is not inherited from a parent; rather, the *susceptibility* to disease is inherited.

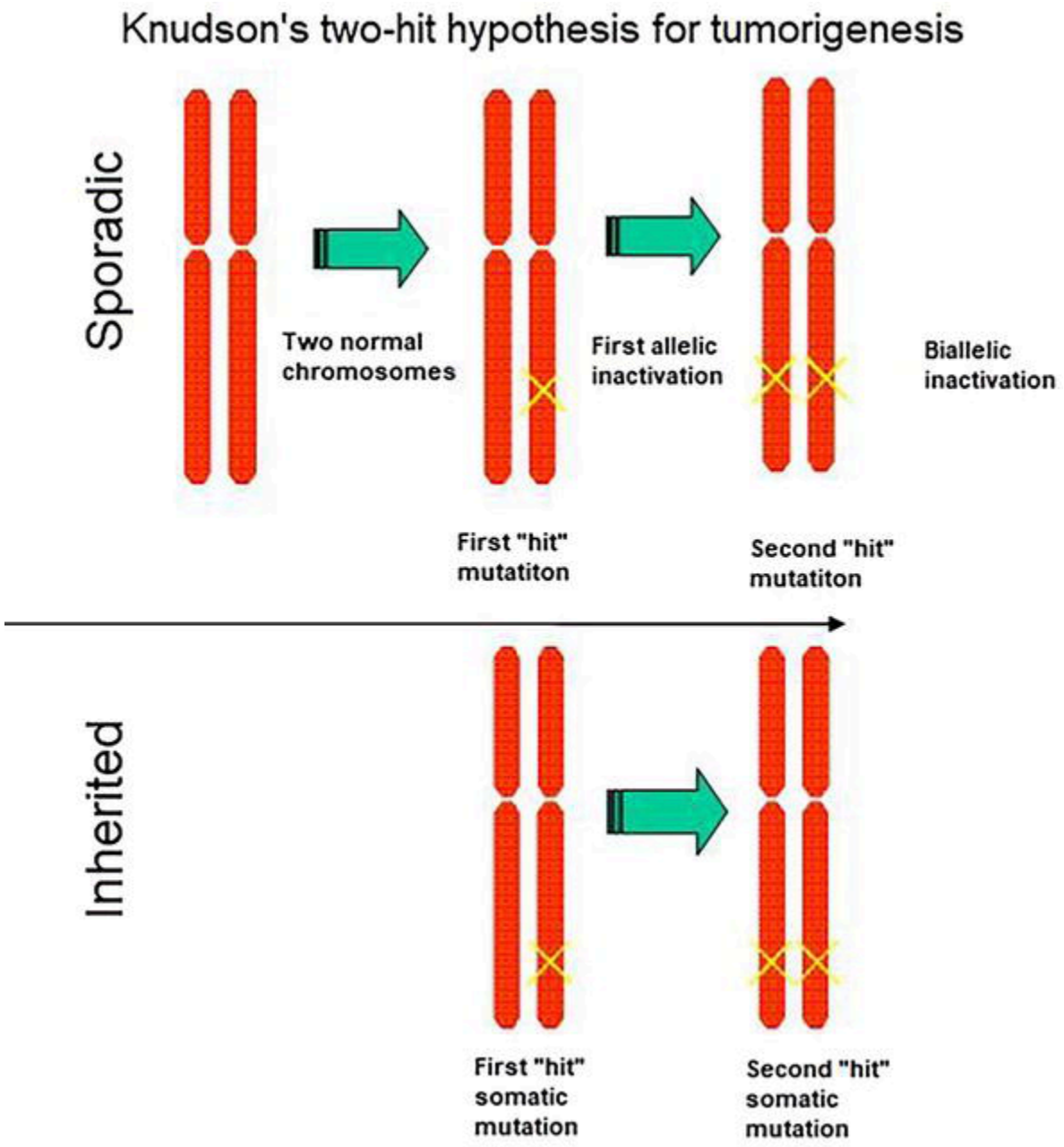


Figure 20. Sporadic cancer requires two successive mutations, one in each allele of a tumor suppressor. Inherited cancers begin with a germline mutation in one copy of a tumor suppressor, so only one additional mutation is needed.

For most people to develop cancer, they need two mutations, or "hits," to occur in a single cell: one in each allele. This is extremely rare, although the longer one lives, the more likely it is to happen just because all organisms accumulate mutations throughout their lifespan. For those who inherit a disease-associated variant, though, every cell in their body already has a head start. If only one cell acquires a mutation in the second allele, that cell will become cancerous. This is called the "two-hit hypothesis" for how cancer develops. (In practice, for most cancers, it's more like 5-6 hits, as discussed in the previous section!)

Tumor suppressor Rb provides a good example of a two-hit cancer. Most incidences of retinoblastoma are sporadic, and patients only have unilateral tumors (meaning in one eye). But there are examples of familial retinoblastoma. Patients with a family history of the disease typically showed signs of the disease at a younger age, and they often developed tumors in both eyes. This is because the retinal cells of those patients only needed to acquire one more mutation to begin oncogenesis. Children who inherit a germline disease-associated variant of Rb have a 90% chance of developing retinoblastoma in one or both eyes. They also have an increased chance of developing other cancers later in life. The incomplete penetrance is due to the random nature of that second mutation.

Other familial cancers are similar. People with one germline disease-associated variant of p53 have Li-Fraumeni syndrome. People with Li-Fraumeni syndrome have nearly a 90% chance of developing at least one cancer throughout their lifetime.

Cisgender females who inherit one disease-associated allele of the tumor suppressors BRCA1 or BRCA2 have a 70% chance of developing breast cancer. Or, in other words, they have a 70% chance of losing BRCA1/2 function in the second allele. Note that although certain variants in BRCA1 and BRCA2 are associated with increased breast cancer risk in cisgender men, there is not currently enough data to calculate the risk for transgender individuals using gender-affirming hormones who have BRCA1/2 variants.[12]

The gene conversion mentioned in the double-strand break repair section sometimes plays a role in whether an individual gets cancer. The healthy cells of a patient who has inherited a cancer-associated allele will be heterozygous. But often, we see a loss of heterozygosity in tumor cells: the tumor cells are homozygous for the mutant allele. This is not just the acquisition of a new mutation since both alleles are the same. Instead, it is an example of improper homology-directed repair.

There are no known familial cancers associated with proto-oncogenes. Inheritance of even one germline mutation in a proto-oncogene is presumably lethal at early embryonic stages since the embryo cannot properly regulate cell division.

Media Attributions

- Figure 20 DNA Repair

1. Piñar-Gutiérrez, A. et al. Difficulties of gender-affirming treatment in trans women with BRCA1+ mutation: A case report. Endocrinol. Diabetes Nutr. 71, 144–148 (2024).

2. Valentini, V. et al. Gender-Specific Genetic Predisposition to Breast Cancer: BRCA Genes and Beyond. Cancers 16, 579 (2024).

EPIGENETICS AND CANCER

Although this chapter focuses primarily on how mutations cause a cancer phenotype, it's important to remember that the changes in gene expression and protein function are what are more directly relevant. Epigenetic modification of genes can have as much effect on cellular phenotype as a mutation. For example, about half of all familial breast cancers are linked with mutations in BRCA1, but only about 10% of breast cancers are familial – the remainder are sporadic. Only some of those sporadic cancers have mutations in BRCA111. However, other sporadic cancers show methylation of BRCA1, suggesting that epigenetic silencing may play a role in cancer progression.[1]

Recent work even suggests that epigenetic changes alone are enough to trigger oncogenesis! [2]

1. Ruscito, I. et al. The Clinical and Pathological Profile of BRCA1 Gene Methylated Breast Cancer Women: A Meta-Analysis. Cancers 13, 1391 (2021).

2. Parreno, V. et al. Transient loss of Polycomb components induces an epigenetic cancer fate. Nature 1–9 (2024) doi:10.1038/s41586-024-07328-w.

VIRUSES AND CANCER

Although cancer is not "contagious" in its strictest sense, there are some viruses that are associated with cancer. In humans, one notable example is the human papilloma virus (HPV). Some strains of HPV cause warts, which are a form of benign tumor. Other strains of HPV can cause cervical cancer.

There are vaccines available for HPV – you may have received one as a tween. They are typically administered around age 11-12.

During infection, the HPV genome can integrate into the genome of a host cell. The HPV genome encodes two proteins, E6 and E7, that affect cells' normal tumor suppression. When expressed by the integrated viral genome, E6 triggers the degradation of the p53 protein. The loss of p53 makes cells unable to respond to DNA damage or begin p53-mediated apoptosis. E7 binds to Rb, preventing Rb from inhibiting E2F and ultimately destroying E2F as well. Together, these two proteins dysregulate cell division and block the dysregulated cell from apoptosis. This process is illustrated in **Figure 21**. These actions together lead to infected cells becoming cancerous.

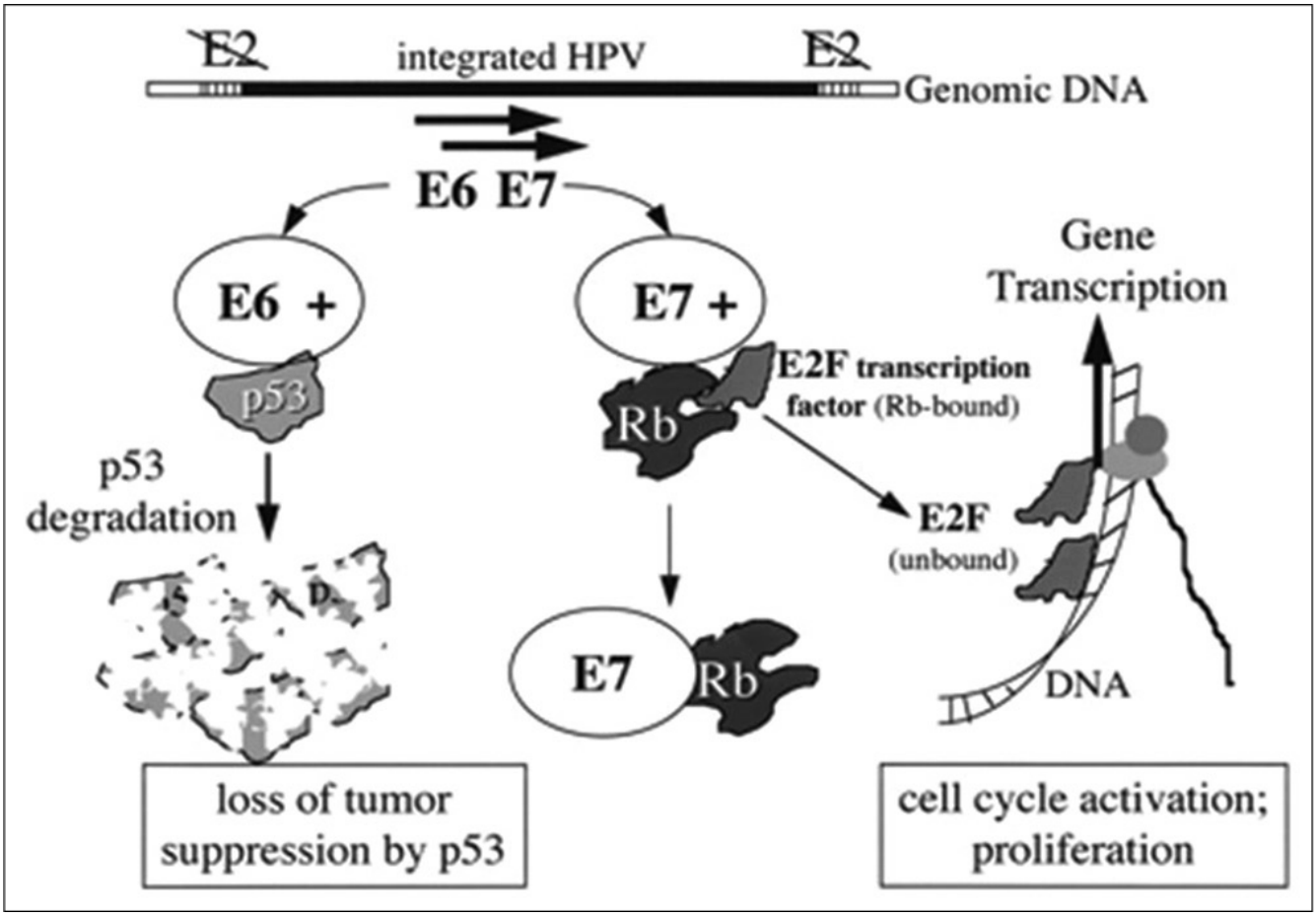


Figure 21. *The HPV E6 and E7 oncogenes degrade p53 and Rb, leading to a loss of tumor suppression by p53 and an activation of E2F-responsive genes.*[1]

Other cancer-associated viruses act differently. Integration of a viral genome into the host genome can bring a viral promoter close to the coding sequence of a proto-oncogene. This can cause the misexpression of the proto-oncogene. One example of this is seen in birds: avian leukosis virus can insert near the c-myc proto-oncogene, resulting in over-expression of c-myc and the formation of tumors. In humans, the human T-lymphotropic virus (HTLV) infects T-cells and is associated with T-cell leukemia and lymphoma.

Media Attributions

- Figure 21 DNA Repair © Dharitri Bhat14 is licensed under a CC BY-NC-SA (Attribution NonCommercial ShareAlike) license

1. Bhat, D. The 'Why and How' of Cervical Cancers and Genital HPV Infection. Cytojournal 19, (2022).

TREATING CANCER: TARGETING THE HALLMARKS OF CANCER

If a person is diagnosed with a solid tumor, typically the first goal will be surgical removal of the tumor. This often includes treatment to kill off any remaining cancer cells, including any cells that may have metastasized. Once metastasis has occurred, it can be very difficult, if not impossible, to surgically remove all cancer cells.

It is a challenge to use nonsurgical means to kill off cancer cells. Unlike taking an antibiotic to kill a bacterial infection, cancer cells are part of a patient's own body. Any drug or treatment that can kill cancer cells will likely also damage healthy tissue. So the goal is to find a drug or treatment that kills cancer cells but leaves healthy tissue alone as much as possible. Sometimes, this therapeutic window is quite small.

Traditionally, oncology treatments have targeted two main characteristics of cancer that make them different from healthy cells: uncontrolled growth and genetic instability.

Most traditional chemotherapies work by exploiting the uncontrolled growth of cancer cells, using various drugs to kill the most rapidly dividing cells in the body. Unfortunately, cancer cells are not the only rapidly dividing cells: hair follicles and the lining of the digestive tract are two examples of other rapidly dividing cell types. Thus, patients being treated with chemotherapies often lose their hair and have digestive challenges as side effects of the drug treatment.

Other chemotherapies as well as radiation exploit the DNA repair defects of cancer cells. Although radiation induces DNA damage and therefore can lead to cancer, tumor cells that lack functional repair mechanisms are especially susceptible to radiation. Essentially, low levels of damage can cause cancer, but high amounts of damage will just kill the cell. Certain chemotherapies also work by causing DNA damage. But again, these therapies also damage healthy cells, just hopefully not as much as the cancer cells.

But in the last twenty years, cancer research has been focused on finding targeted strategies for fighting cancer: ones that are more specific to killing cancer cells but will have little effect on healthy ones.

The most successful example to date has been the drug imatinib, which targets an oncogene called BCR-ABL. The BCR-ABL oncogene does not exist in healthy cells. It is a fusion protein created by a translocation common to chronic myelogenous leukemia (CML) and other cancers, as shown in **Figure 22**. The drug inhibits the BCR-ABL mutation that drives the cancer and blocks cancer cell proliferation, but it has little effect on healthy cells because they do not have the oncogene. With imatinib treatment, life expectancy of CML patients is similar to that of noncancer patients.

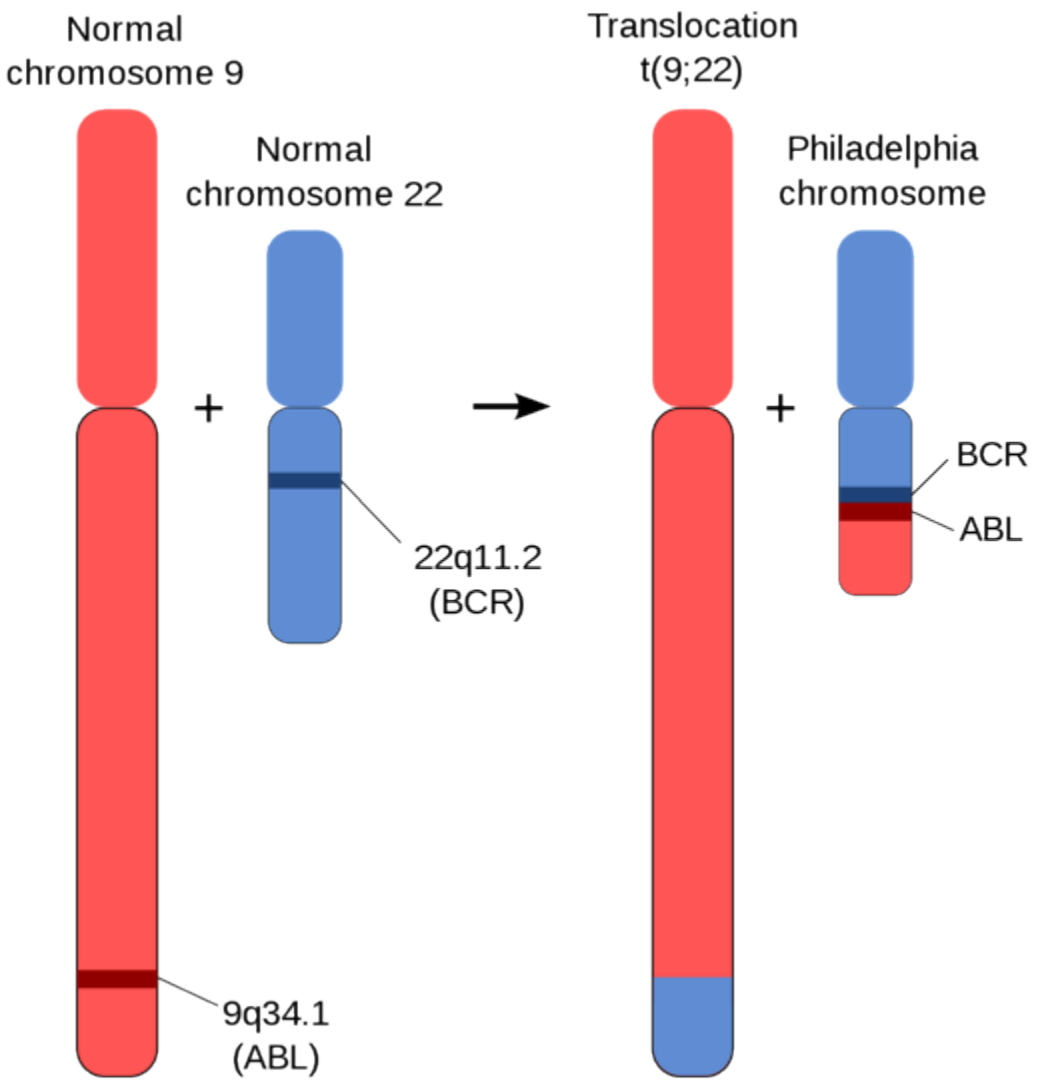


Figure 22. The "Philadelphia chromosome" is found in certain types of cancers. It results from a swap of genetic material between chromosomes 9 and 22. The fusion site brings together two genes into one: BCR from chromosome 22 and ABL from chromosome 9. This forms the oncogene BCR-ABL, which drives tumorigenesis cells with translocation.

Other examples target growth receptors found on the cell surface of cancer cells or exploit the hallmarks of cancer. This includes new therapies that program the immune system to recognize and fight specific types of cancer.

Despite these remarkable advances, we still remain a long way from "a cure" for cancer. Because of the varied mutations that contribute to cancer, each of these innovations only helps a very select population of patients.

Media Attributions

- Figure 22 DNA Repair © By Aryn89 is licensed under a CC BY-SA (Attribution ShareAlike) license

FACTORS THAT INFLUENCE CANCER INCIDENCE: LESSONS FROM EVOLUTION

As mentioned at the beginning of this chapter, humans are not the only organisms to get cancer. It turns out that we can learn a lot from other organisms.

The effect of DNA damage is cumulative over time as lesions escape repair and become ensconced in the genome as mutations. And the more cells (and cell divisions) there are, one might expect, the more likely that one will be damaged in exactly the right way.

Within a species, this holds true: older organisms are far more likely to get most cancers than young ones, simply because it takes time to accumulate the mutations. The exceptions – childhood cancers like Rb or cancers that show up in young adulthood – appear to require fewer overall "hits". Retinoblastoma, for example, is a two-hit cancer (one in each copy of the Rb gene), and CML appears to be a one-hit cancer (driven by the BCR-ABL oncogene).

But when we compare cancer across species, this doesn't hold true. We might expect that larger and long-lived organisms, having more cells and presumably equally prone to DNA damage, would get cancer more frequently, and small shorter-lived organisms would get cancer more infrequently.

But that's not the case. For example, elephants are large, long-lived organisms that rarely get cancer. So what's going on?

The answer is that body size in elephants likely co-evolved with additional anti-cancer defense mechanisms. Elephant genomes have multiple copies of the genes similar to the tumor suppressor protein p53, called TP53RTGs. Essentially, the p53 gene was duplicated multiple times over the modern elephant's evolutionary history. More tumor suppressor genes = more tumor suppressor proteins = more tumor suppression!

In **Figure 23**, you can see a comparison of the number of copies of the p53 and related genes in a number of living and extinct proboscidean (elephant) species, arranged by evolutionary relationship. The African savannah elephant, the largest of the bunch, has the highest number of copies – 2 copies each of 10 related genes! Other proboscideans have fewer, but still more than the 2 copies of a single p53 gene that humans have.

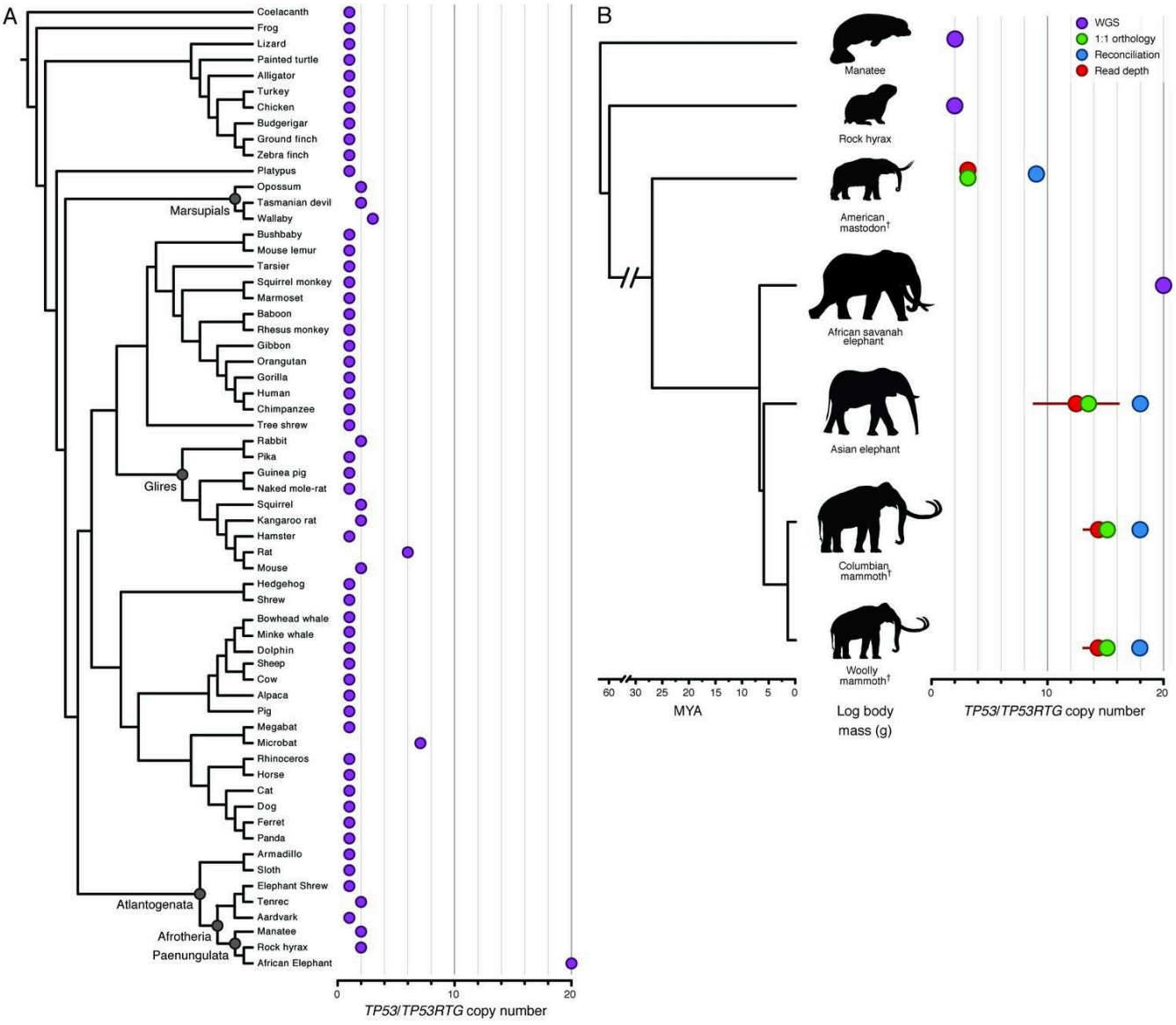


Figure 23. Comparison of TP53 copy number and related genes among proboscidean species. The colored dots represent an approximation of copy numbers using different methods of genome analysis.

Other long-lived or large-bodied organisms appear to fight cancer differently. For example, the naked mole rat (**Figure 24**) is extraordinarily long-lived for its body size, living up to three decades. The naked mole rat's physiology blocks one of the hallmarks of cancer: metastasis.

In most organisms, healthy cells are controlled by contact inhibition: that is, when they come into contact with other cells, they stop dividing. Metastasizing cancer cells lose their contact inhibition. Even when grown in the lab, healthy cells stop dividing when they cover the surface of their dish, but cancer cells will keep dividing, piling on top of one another.

The cells of the naked mole rat, on the other hand, seem to be *hyper*sensitive to contact inhibition.[1] They are also resistant to the epigenetic reprogramming that accompanies de-differentiation during cancer progression.[2]

Figure 24. Naked mole rats live a really, really long time and don't get cancer.

Research on other large-bodied or long-lived species is likely to yield more insights into cancer physiology.

Test Your Understanding

An interactive H5P element has been excluded from this version of the text. You can view it

1. Seluanov, A. et al. Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. Proc. Natl. Acad. Sci. 106, 19352–19357 (2009).

2. Tan, L. et al. Naked Mole Rat Cells Have a Stable Epigenome that Resists iPSC Reprogramming: Stem Cell Rep. 9, 1721–1734 (2017).

online here:
https://rotel.pressbooks.pub/genetics/?p=815#h5p-101

Media Attributions

- Figure 23 DNA Repair © Sulak et al, 2016 is licensed under a CC BY (Attribution) license
- Figure 24 DNA Repair © By Roman Klementschitz, Wien is licensed under a CC BY-SA (Attribution ShareAlike) license

SUMMARY

Cancer is a genetic disease characterized by abnormal cell proliferation. It is caused by somatic mutations in tumor suppressor genes and proto-oncogenes.

Many of these mutations accumulate through failure of DNA repair pathways, and many DNA repair proteins are tumor suppressors, involved in genomic caretaker roles. Tumor suppressors like Rb and p53, on the other hand, serve as gatekeepers to the cell cycle, blocking the cell cycle unless conditions are appropriate for cell division. A loss of function in tumor suppressors leads to abnormal proliferation and/or an increase in mutation rate. Cancer-associated mutations in tumor suppressor genes are recessive on a cellular level.

In contrast, mutations in proto-oncogenes are dominant on a cellular level. Proto-oncogenes normally promote cell cycle progression, but only when conditions are right. Gain of function mutations in proto-oncogenes convert them to oncogenes, driving the cell cycle forward regardless of whether the timing is appropriate.

Most cells require an accumulation of 5-6 distinct mutations in both proto-oncogenes and tumor suppressors, although there are examples of so-called “one-hit” and “two-hit” cancers. With each subsequent mutation, cells acquire a growth advantage over their neighboring cells. This advantage results in a genetically heterogeneous tumor. Familial cancers essentially have a head start on this process, beginning with a germline tumor suppressor mutation.

The accumulation of mutations leads to characteristic hallmarks of cancer, many of which have been targeted for cancer therapeutics in recent years.

Although we might expect that large-bodied or long-lived organisms would have a greater incidence of cancer than small short-lived organisms, this is not the case. Our understanding of cancer and its physiology has been enhanced by studying cancer in other organisms, where the rate of cancer is higher or lower than would be expected by body size or lifespan.

AFTERWARD

In the spring of 2020, I was watching many of my students struggle with access to education, complicated by the pandemic. Many of them were juggling either job loss or extra hours at work, taking care of family members, online classes, and other pandemic stresses. The United States was reeling from the aftermath of the murder of George Floyd, just the latest in a series of civil rights violations across the country. At the same time, I was participating in the first year of the HHMI-sponsored STEM Racial Equity Institute at Framingham State University, a 5-week intensive workshop designed to take a close look at racial equity in STEM.

This project grew out of that experience. A big part of equity, of course, is financial. STEM textbooks are ridiculously expensive to purchase. The project was initially prompted so I could move away from expensive textbooks in my 200-level Genetics course. But it quickly moved beyond that.

This textbook was developed with assistance from a grant from the ROTEL project, a multi-institutional collaboration among six Massachusetts colleges and universities. The mission of ROTEL project, Remixing Open Textbooks through an Equity Lens, is to:

> "*[promote] textbook affordability, student success, and inclusion and equity to benefit all students, particularly minoritized students at our six institutions. The project will provide stipends for faculty to remix and/or develop accessible, intentionally inclusive open textbooks and other OER that reflect students' local and lived experiences.*"

This textbook was written with an eye toward that mission, with three goals in mind:

1. Consider how social and cultural pressures influenced what we know about genetics and who gets to participate in the field today.
2. Remember that people are not model organisms. When using examples from humans, be respectful that real people are affected by genetic differences we are discussing. We learn a lot about biology from looking at rare differences, but people are more than just a phenotype. Recognize that social significance is often misapplied to biological differences.
3. Make education more accessible. Create an accessible online textbook that is compatible with screen readers and other adaptive technology, incorporating media that isn't possible with a printed textbook, and which students would retain access to beyond their semester-long course.

This is a work in progress: please share comments and feedback where these three goals are not yet met.

While reading this text, I hope you will consider how scientists' background, culture, and society influenced what we know about Genetics. I hope you will also consider how your background as a scientist is a strength: what do you bring to the field that others do not? What research questions will you ask, and what skills will

you use to answer those questions? What connections can you make – with scientists and nonscientists – that will further influence how your work will be used by others?

Like all science, which builds on previous work, this project was influenced by other Open Educational Resources in genetics. Images and text are remixed from the following OER sources, and attributed within the modules where they are used:

- *Open Genetics Lectures: An Open Source Molecular Genetics Textbook*, from the University of Alberta, Canada, Fall 2017 and *Online Open Genetics (Nickle and Isabelle Barrette-Ng)* 2009, both derivatives of work originally created for Biology 207 at the University of Alberta by Michael Deyholos, Mike Harrington, John Locke, and Mark Wollansky
- *OpenStax Biology 2e*, by Mary Ann Clark, Matthew Douglas, and Jung Choi
- *OpenStax Microbiology*, by Nina Parker, Mark Schneegurt, Anh-Hue Thi Tu, Philip Lister, Brian M. Forster

I'd like to thank many people who helped with this project, directly or indirectly:

- Millie Gonzales and the ROTEL publishing team, in particular Richard Lizotte, Minh Lee, and Vicky Gavin
- Michelle Cromwell, for inspiring conversations about equity, practical suggestions, and encouragement
- Catherine Dignam and the leaders of the HHMI STEM Racial Equity Project, for providing a space to focus on equity in STEM and brainstorm ways to improve
- Chelsea Hudson, for beta-testing and review
- And, of course, my family, for stepping up and filling in at home while I was working on this project.

Amanda Simons
Framingham State University

GRANT INFORMATION

The U.S. Department of Education, the granting agency for the ROTEL project, requires information about the grant be included in the back matter. The text for this section is provided below.

The ROTEL Project is 100% funded by the U.S. Department of Education's grant from the Fund for the Improvement of Postsecondary Education, (FIPSE). The contents of the OERs do not necessarily represent the policy of the Department of Education, and an endorsement by the Federal Government should not be assumed.

For more information about the ROTEL Project, please visit our project website.

VERSION HISTORY

Below is the version history for ***Chromosomes, Genes, and Traits***:

Version	Publication Date	Changes
First Edition: link to first edition	May 18, 2024	—
Revised Edition: link to revised edition	January 24, 2025	**New Chapters:** IX. Eukaryotic Gene Regulation in Action: Examples from Develop XVII. Evolution and Ancestry XVIII. DNA repair and cancer**Stylistic improvements** throughou Afterward Version History