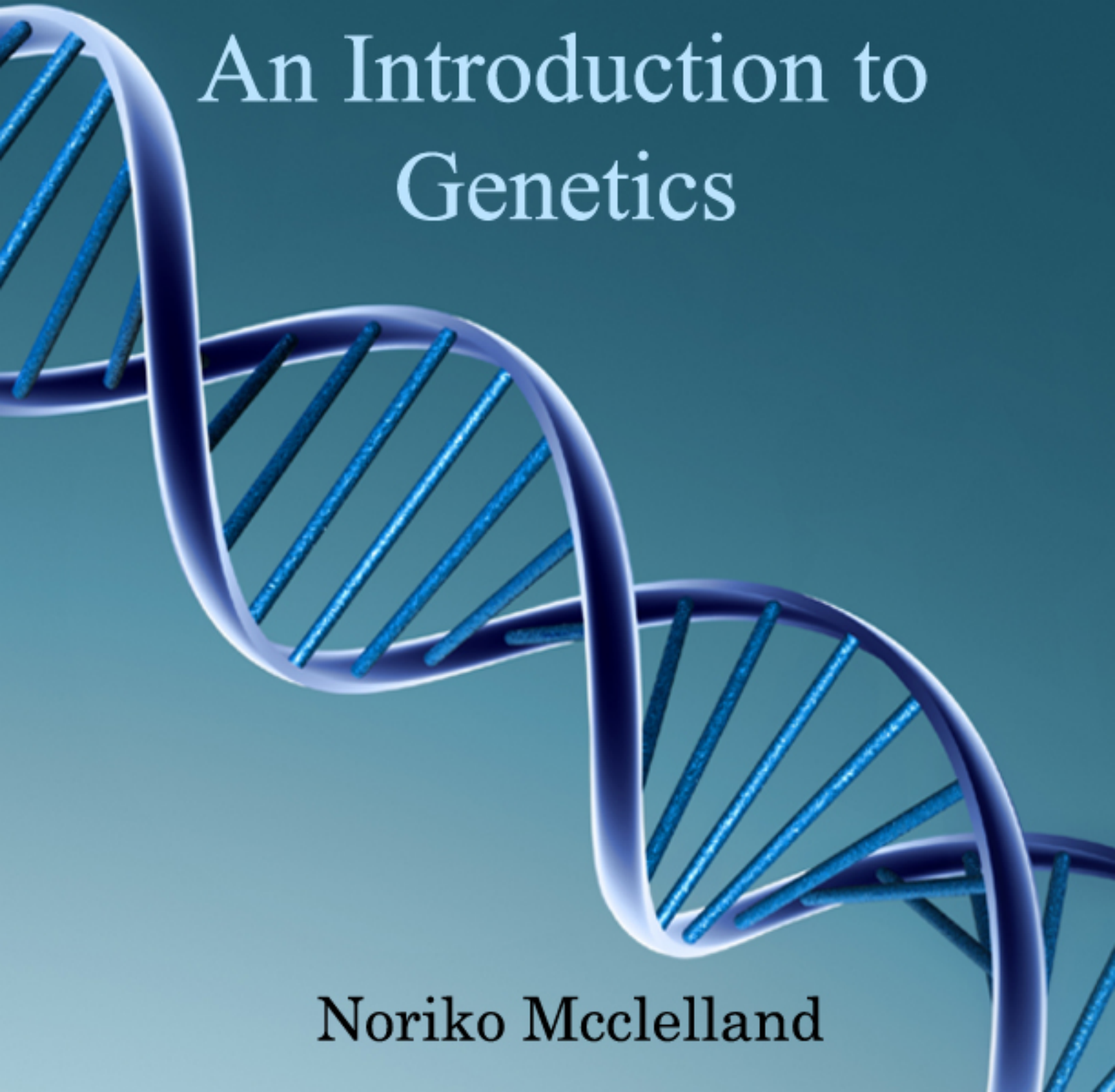
An Introduction to Genetics



Noriko Mcclelland

First Edition, 2012

ISBN 978-81-323-3234-3

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Genetics

Genetics (from Ancient Greek γενετικός *genetikos*, "genitive" and that from γένεσις *genesis*, "origin"), a discipline of biology, is the science of genes, heredity, and variation in living organisms.

Genetics deals with the molecular structure and function of genes, with gene behavior in the context of a cell or organism (e.g. dominance and epigenetics), with patterns of inheritance from parent to offspring, and with gene distribution, variation and change in populations. Given that genes are universal to living organisms, genetics can be applied to the study of any living system from viruses and bacteria, through plants (especially crops) to humans (for example in Medical Genetics)

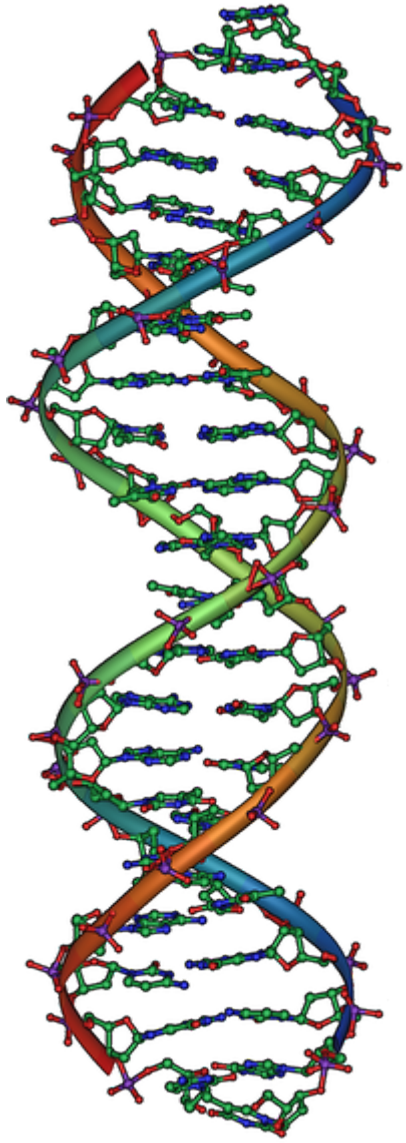
The fact that living things inherit traits from their parents has been used since prehistoric times to improve crop plants and animals through selective breeding. However, the modern science of genetics, which seeks to understand the process of inheritance, only began with the work of Gregor Mendel in the mid-19th century. Although he did not know the physical basis for heredity, Mendel observed that organisms inherit traits via discrete units of inheritance, which are now called genes.

Genes correspond to regions within DNA, a molecule composed of a chain of four different types of nucleotides—the sequence of these nucleotides is the genetic information organisms inherit. DNA naturally occurs in a double stranded form, with nucleotides on each strand complementary to each other. Each strand can act as a template for creating a new partner strand—this is the physical method for making copies of genes that can be inherited.

The sequence of nucleotides in a gene is translated by cells to produce a chain of amino acids, creating proteins—the order of amino acids in a protein corresponds to the order of nucleotides in the gene. This relationship between nucleotide sequence and amino acid sequence is known as the genetic code. The amino acids in a protein determine how it folds into a three-dimensional shape; this structure is, in turn, responsible for the protein's function. Proteins carry out almost all the functions needed for cells to live. A change to the DNA in a gene can change a protein's amino acids, changing its shape and function: this can have a dramatic effect in the cell and on the organism as a whole.

Although genetics plays a large role in the appearance and behavior of organisms, it is the combination of genetics with what an organism experiences that determines the ultimate outcome. For example, while genes play a role in determining an organism's size, the nutrition and other conditions it experiences after inception also have a large effect.

History



DNA, the molecular basis for inheritance. Each strand of DNA is a chain of nucleotides, matching each other in the center to form what look like rungs on a twisted ladder.

Although the science of genetics began with the applied and theoretical work of Gregor Mendel in the mid-19th century, other theories of inheritance preceded Mendel. A popular theory during Mendel's time was the concept of blending inheritance: the idea that individuals inherit a smooth blend of traits from their parents. Mendel's work disproved this, showing that traits are composed of combinations of distinct genes rather

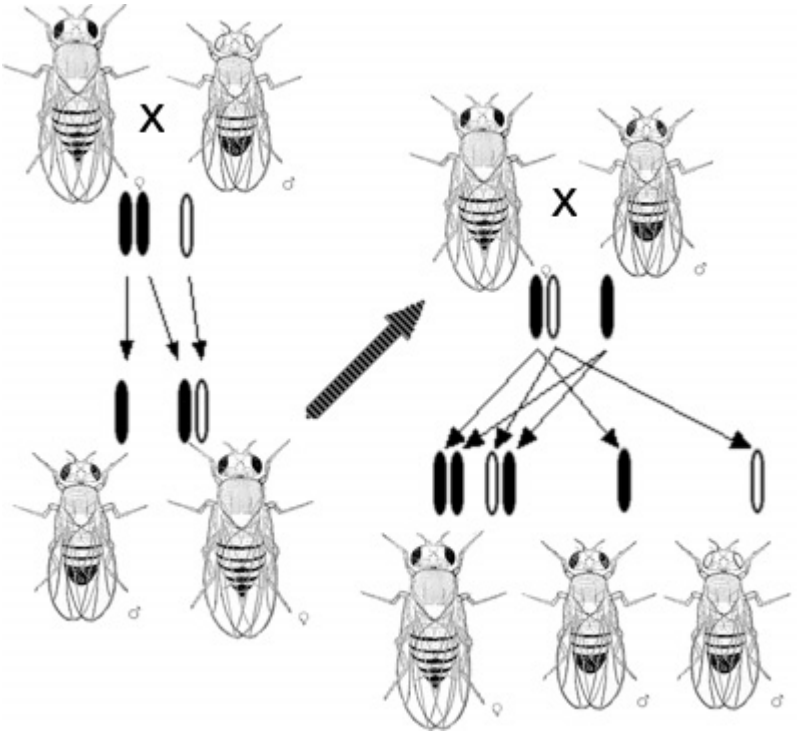
than a continuous blend. Another theory that had some support at that time was the inheritance of acquired characteristics: the belief that individuals inherit traits strengthened by their parents. This theory (commonly associated with Jean-Baptiste Lamarck) is now known to be wrong—the experiences of individuals do not affect the genes they pass to their children. Other theories included the pangenesis of Charles Darwin (which had both acquired and inherited aspects) and Francis Galton's reformulation of pangenesis as both particulate and inherited.

Mendelian and classical genetics

Modern genetics started with Gregor Johann Mendel, a German-Czech Augustinian monk and scientist who studied the nature of inheritance in plants. In his paper "Versuche über Pflanzenhybriden" ("Experiments on Plant Hybridization"), presented in 1865 to the *Naturforschender Verein* (Society for Research in Nature) in Brünn, Mendel traced the inheritance patterns of certain traits in pea plants and described them mathematically. Although this pattern of inheritance could only be observed for a few traits, Mendel's work suggested that heredity was particulate, not acquired, and that the inheritance patterns of many traits could be explained through simple rules and ratios.

The importance of Mendel's work did not gain wide understanding until the 1890s, after his death, when other scientists working on similar problems re-discovered his research. William Bateson, a proponent of Mendel's work, coined the word *genetics* in 1905. (The adjective *genetic*, derived from the Greek word *genesis*—γένεσις, "origin", predates the noun and was first used in a biological sense in 1860.) Bateson popularized the usage of the word *genetics* to describe the study of inheritance in his inaugural address to the Third International Conference on Plant Hybridization in London, England, in 1906.

After the rediscovery of Mendel's work, scientists tried to determine which molecules in the cell were responsible for inheritance. In 1910, Thomas Hunt Morgan argued that genes are on chromosomes, based on observations of a sex-linked white eye mutation in fruit flies. In 1913, his student Alfred Sturtevant used the phenomenon of genetic linkage to show that genes are arranged linearly on the chromosome.



Morgan's observation of sex-linked inheritance of a mutation causing white eyes in *Drosophila* led him to the hypothesis that genes are located upon chromosomes.

Molecular genetics

Although genes were known to exist on chromosomes, chromosomes are composed of both protein and DNA—scientists did not know which of these is responsible for inheritance. In 1928, Frederick Griffith discovered the phenomenon of transformation: dead bacteria could transfer genetic material to "transform" other still-living bacteria. Sixteen years later, in 1944, Oswald Theodore Avery, Colin McLeod and Maclyn McCarty identified the molecule responsible for transformation as DNA. The Hershey-Chase experiment in 1952 also showed that DNA (rather than protein) is the genetic material of the viruses that infect bacteria, providing further evidence that DNA is the molecule responsible for inheritance.

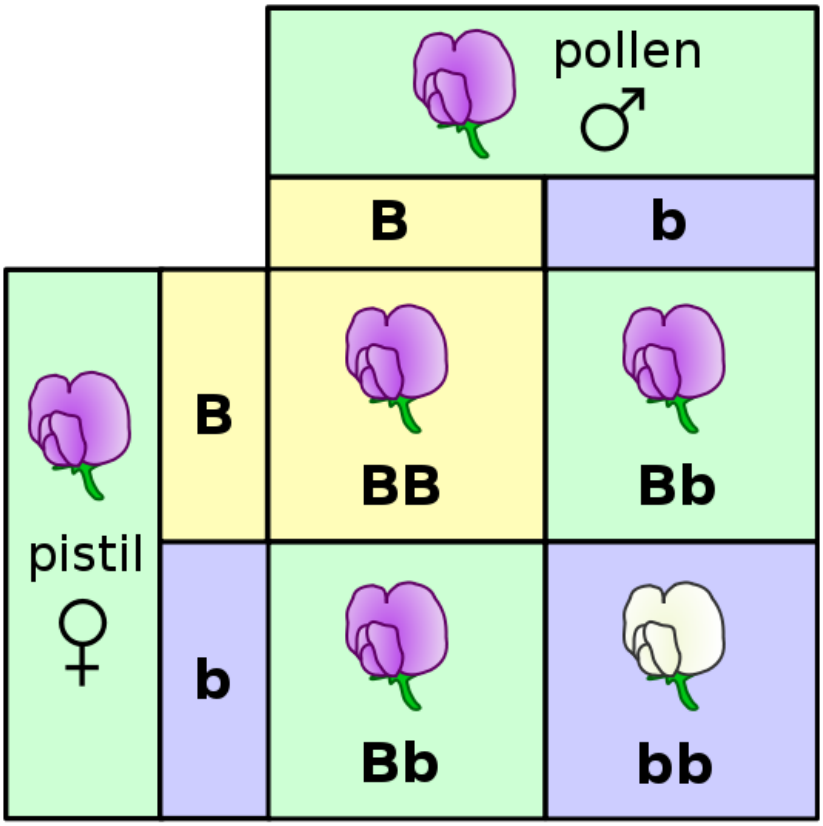
James D. Watson and Francis Crick determined the structure of DNA in 1953, using the X-ray crystallography work of Rosalind Franklin and Maurice Wilkins that indicated DNA had a helical structure (i.e., shaped like a corkscrew). Their double-helix model had two strands of DNA with the nucleotides pointing inward, each matching a complementary nucleotide on the other strand to form what looks like rungs on a twisted ladder. This structure showed that genetic information exists in the sequence of nucleotides on each strand of DNA. The structure also suggested a simple method for duplication: if the strands are separated, new partner strands can be reconstructed for each based on the sequence of the old strand.

Although the structure of DNA showed how inheritance works, it was still not known how DNA influences the behavior of cells. In the following years, scientists tried to understand how DNA controls the process of protein production. It was discovered that the cell uses DNA as a template to create matching messenger RNA (a molecule with nucleotides, very similar to DNA). The nucleotide sequence of a messenger RNA is used to create an amino acid sequence in protein; this translation between nucleotide and amino acid sequences is known as the genetic code.

With this molecular understanding of inheritance, an explosion of research became possible. One important development was chain-termination DNA sequencing in 1977 by Frederick Sanger. This technology allows scientists to read the nucleotide sequence of a DNA molecule. In 1983, Kary Banks Mullis developed the polymerase chain reaction, providing a quick way to isolate and amplify a specific section of a DNA from a mixture. Through the pooled efforts of the Human Genome Project and the parallel private effort by Celera Genomics, these and other techniques culminated in the sequencing of the human genome in 2003.

Features of inheritance

Discrete inheritance and Mendel's laws



A Punnett square depicting a cross between two pea plants heterozygous for purple (B) and white (b) blossoms

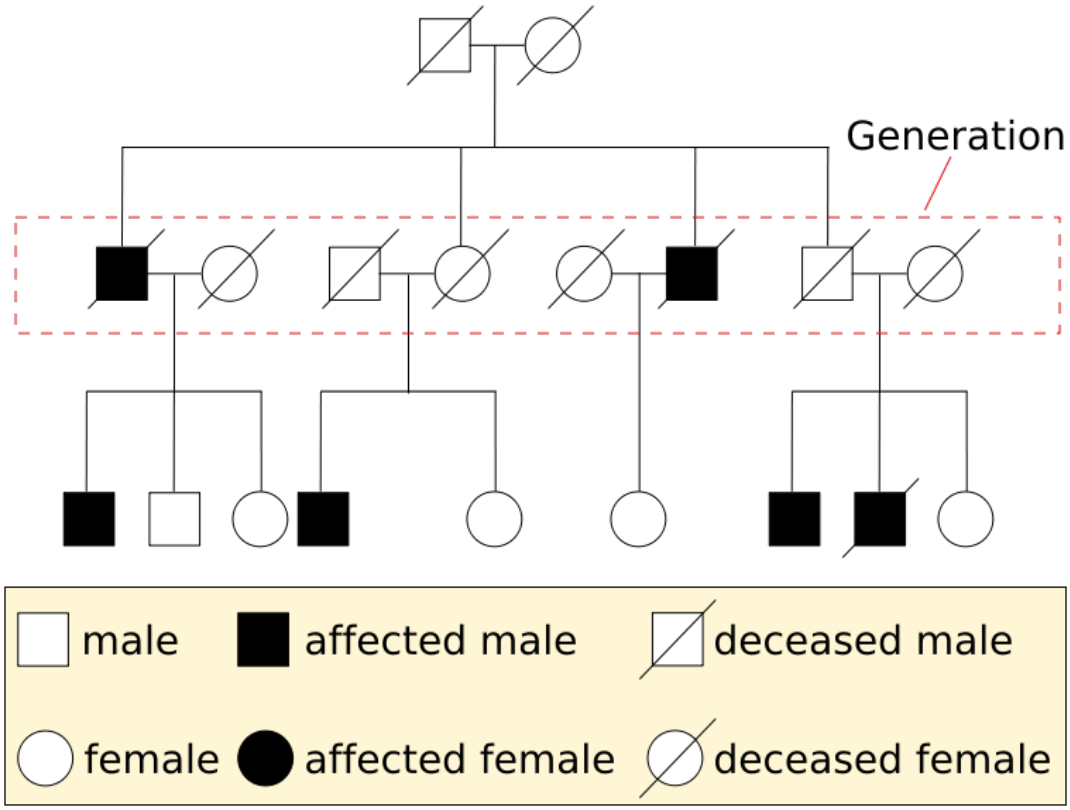
At its most fundamental level, inheritance in organisms occurs by means of discrete traits, called genes. This property was first observed by Gregor Mendel, who studied the segregation of heritable traits in pea plants. In his experiments studying the trait for flower color, Mendel observed that the flowers of each pea plant were either purple or white—but never an intermediate between the two colors. These different, discrete versions of the same gene are called alleles.

In the case of pea, which is a diploid species, each individual plant has two alleles of each gene, one allele inherited from each parent. Many species, including humans, have this pattern of inheritance. Diploid organisms with two copies of the same allele of a given gene are called homozygous at that gene locus, while organisms with two different alleles of a given gene are called heterozygous.

The set of alleles for a given organism is called its genotype, while the observable traits of the organism are called its phenotype. When organisms are heterozygous at a gene, often one allele is called dominant as its qualities dominate the phenotype of the organism, while the other allele is called recessive as its qualities recede and are not observed. Some alleles do not have complete dominance and instead have incomplete dominance by expressing an intermediate phenotype, or codominance by expressing both alleles at once.

When a pair of organisms reproduce sexually, their offspring randomly inherit one of the two alleles from each parent. These observations of discrete inheritance and the segregation of alleles are collectively known as Mendel's first law or the Law of Segregation.

Notation and diagrams



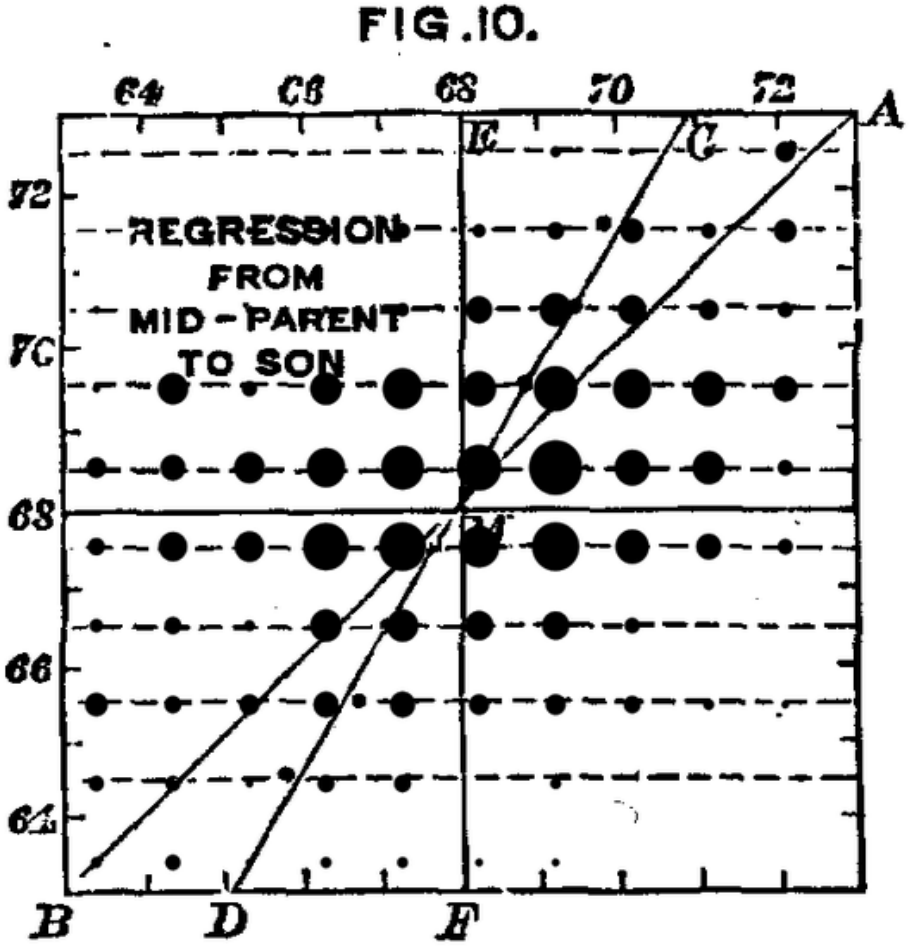
Genetic pedigree charts help track the inheritance patterns of traits.

Geneticists use diagrams and symbols to describe inheritance. A gene is represented by one or a few letters. Often a "+" symbol is used to mark the usual, non-mutant allele for a gene.

In fertilization and breeding experiments (and especially when discussing Mendel's laws) the parents are referred to as the "P" generation and the offspring as the "F1" (first filial) generation. When the F1 offspring mate with each other, the offspring are called the "F2" (second filial) generation. One of the common diagrams used to predict the result of cross-breeding is the Punnett square.

When studying human genetic diseases, geneticists often use pedigree charts to represent the inheritance of traits. These charts map the inheritance of a trait in a family tree.

Interactions of multiple genes



Human height is a trait with complex genetic causes. Francis Galton's data from 1889 shows the relationship between offspring height as a function of mean parent height. While correlated, remaining variation in offspring heights indicates environment is also an important factor in this trait.

Organisms have thousands of genes, and in sexually reproducing organisms these genes generally assort independently of each other. This means that the inheritance of an allele for yellow or green pea color is unrelated to the inheritance of alleles for white or purple flowers. This phenomenon, known as "Mendel's second law" or the "Law of independent assortment", means that the alleles of different genes get shuffled between parents to form offspring with many different combinations.

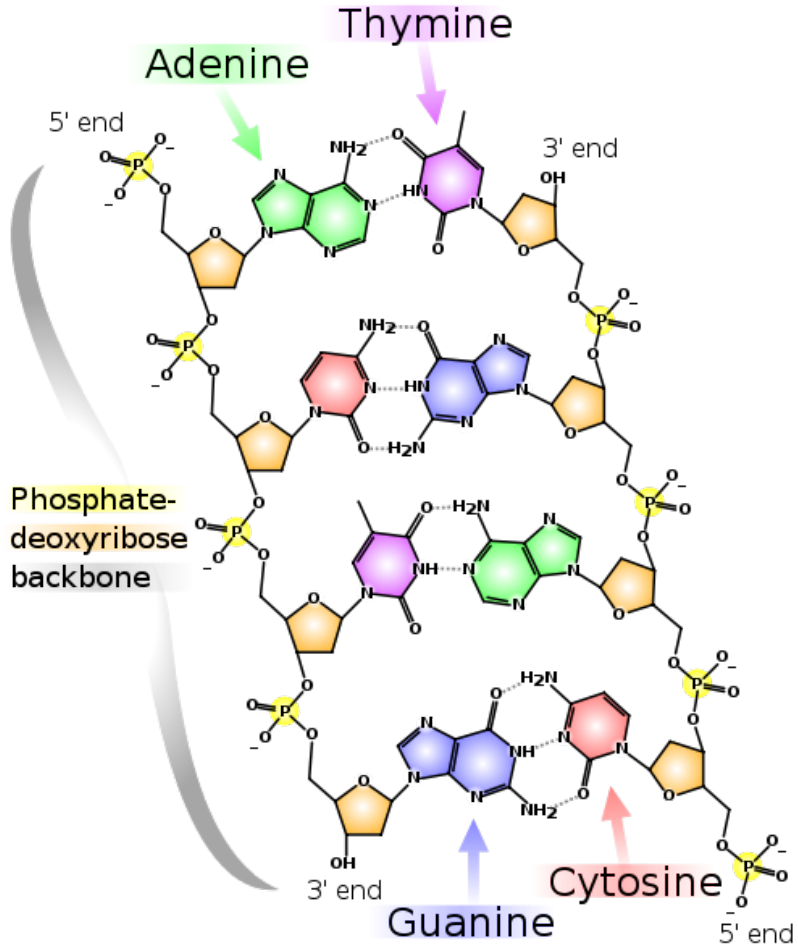
Often different genes can interact in a way that influences the same trait. In the Blue-eyed Mary (*Omphalodes verna*), for example, there exists a gene with alleles that determine the color of flowers: blue or magenta. Another gene, however, controls whether the flowers have color at all or are white. When a plant has two copies of this white allele, its

flowers are white—regardless of whether the first gene has blue or magenta alleles. This interaction between genes is called epistasis, with the second gene epistatic to the first.

Many traits are not discrete features (e.g. purple or white flowers) but are instead continuous features (e.g. human height and skin color). These complex traits are products of many genes. The influence of these genes is mediated, to varying degrees, by the environment an organism has experienced. The degree to which an organism's genes contribute to a complex trait is called heritability. Measurement of the heritability of a trait is relative—in a more variable environment, the environment has a bigger influence on the total variation of the trait. For example, human height is a trait with complex causes. It has a heritability of 89% in the United States. In Nigeria, however, where people experience a more variable access to good nutrition and health care, height has a heritability of only 62%.

Molecular basis for inheritance

DNA and chromosomes



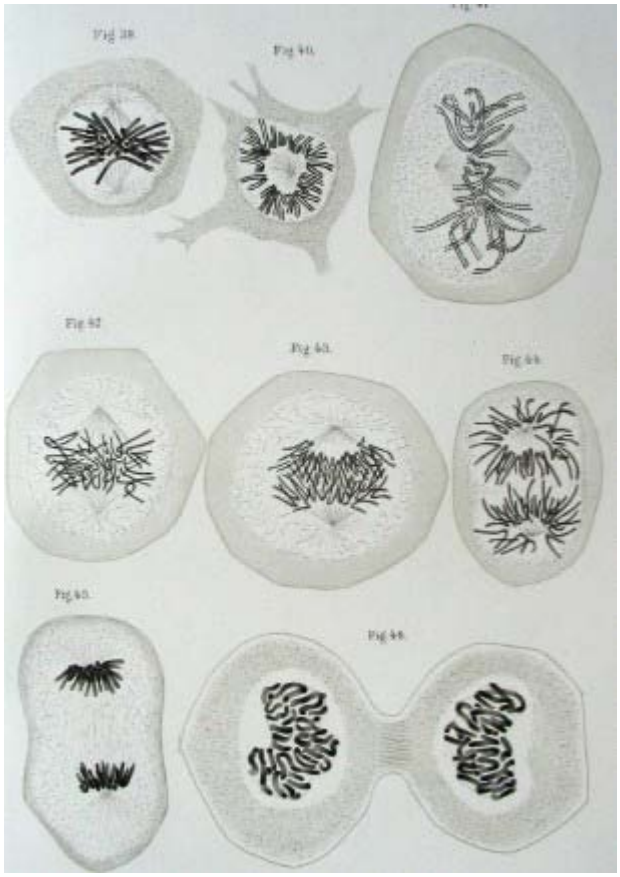
The molecular structure of DNA. Bases pair through the arrangement of hydrogen bonding between the strands.

The molecular basis for genes is deoxyribonucleic acid (DNA). DNA is composed of a chain of nucleotides, of which there are four types: adenine (A), cytosine (C), guanine (G), and thymine (T). Genetic information exists in the sequence of these nucleotides, and genes exist as stretches of sequence along the DNA chain. Viruses are the only exception to this rule—sometimes viruses use the very similar molecule RNA instead of DNA as their genetic material.

DNA normally exists as a double-stranded molecule, coiled into the shape of a double-helix. Each nucleotide in DNA preferentially pairs with its partner nucleotide on the opposite strand: A pairs with T, and C pairs with G. Thus, in its two-stranded form, each strand effectively contains all necessary information, redundant with its partner strand. This structure of DNA is the physical basis for inheritance: DNA replication duplicates the genetic information by splitting the strands and using each strand as a template for synthesis of a new partner strand.

Genes are arranged linearly along long chains of DNA sequence, called chromosomes. In bacteria, each cell usually contains a single circular chromosome, while eukaryotic organisms (including plants and animals) have their DNA arranged in multiple linear chromosomes. These DNA strands are often extremely long; the largest human chromosome, for example, is about 247 million base pairs in length. The DNA of a chromosome is associated with structural proteins that organize, compact, and control access to the DNA, forming a material called chromatin; in eukaryotes, chromatin is usually composed of nucleosomes, segments of DNA wound around cores of histone proteins. The full set of hereditary material in an organism (usually the combined DNA sequences of all chromosomes) is called the genome.

While haploid organisms have only one copy of each chromosome, most animals and many plants are diploid, containing two of each chromosome and thus two copies of every gene. The two alleles for a gene are located on identical loci of sister chromatids, each allele inherited from a different parent.



Walther Flemming's 1882 diagram of eukaryotic cell division. Chromosomes are copied, condensed, and organized. Then, as the cell divides, chromosome copies separate into the daughter cells.

Many species have so called sex chromosomes. They are special in that they determine the sex of the organism. In humans and many other animals, the Y chromosome contains the gene that triggers the development of the specifically male characteristics. In evolution, this chromosome has lost most of its content and also most of its genes, while the X chromosome is similar to the other chromosomes and contains many genes. The X and Y chromosomes form a very heterogeneous pair before cell division.

Reproduction

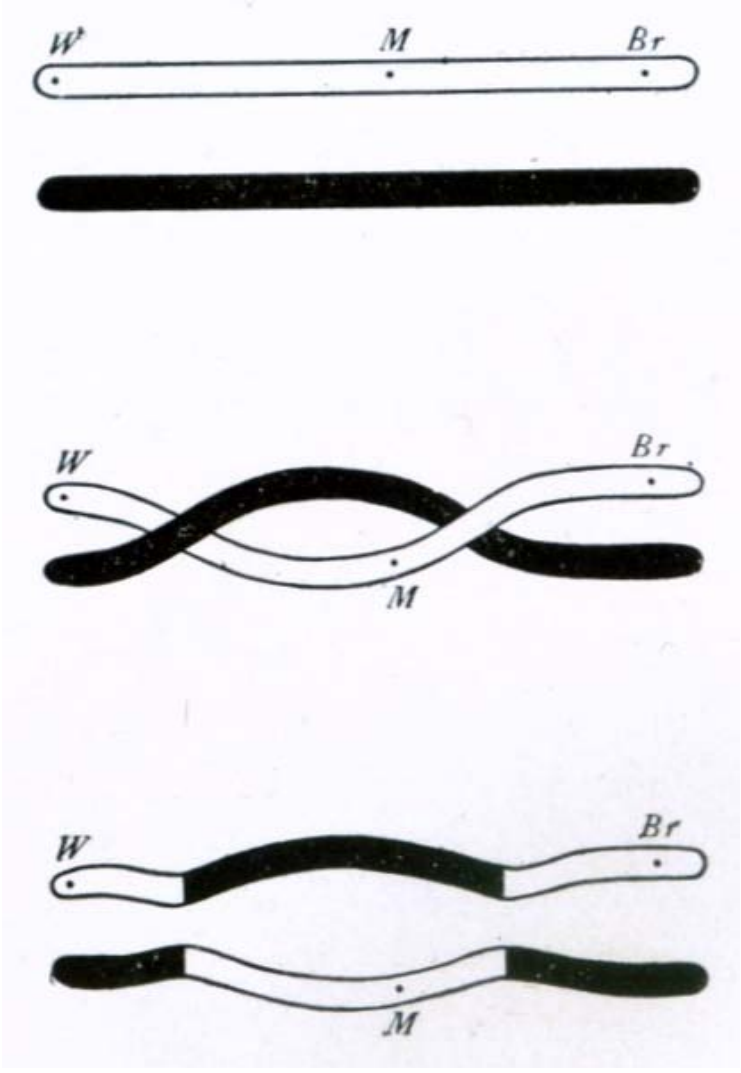
When cells divide, their full genome is copied and each daughter cell inherits one copy. This process, called mitosis, is the simplest form of reproduction and is the basis for asexual reproduction. Asexual reproduction can also occur in multicellular organisms, producing offspring that inherit their genome from a single parent. Offspring that are genetically identical to their parents are called clones.

Eukaryotic organisms often use sexual reproduction to generate offspring that contain a mixture of genetic material inherited from two different parents. The process of sexual reproduction alternates between forms that contain single copies of the genome (haploid)

and double copies (diploid). Haploid cells fuse and combine genetic material to create a diploid cell with paired chromosomes. Diploid organisms form haploids by dividing, without replicating their DNA, to create daughter cells that randomly inherit one of each pair of chromosomes. Most animals and many plants are diploid for most of their lifespan, with the haploid form reduced to single cell gametes such as sperm or eggs.

Although they do not use the haploid/diploid method of sexual reproduction, bacteria have many methods of acquiring new genetic information. Some bacteria can undergo conjugation, transferring a small circular piece of DNA to another bacterium. Bacteria can also take up raw DNA fragments found in the environment and integrate them into their genomes, a phenomenon known as transformation. These processes result in horizontal gene transfer, transmitting fragments of genetic information between organisms that would be otherwise unrelated.

Recombination and linkage



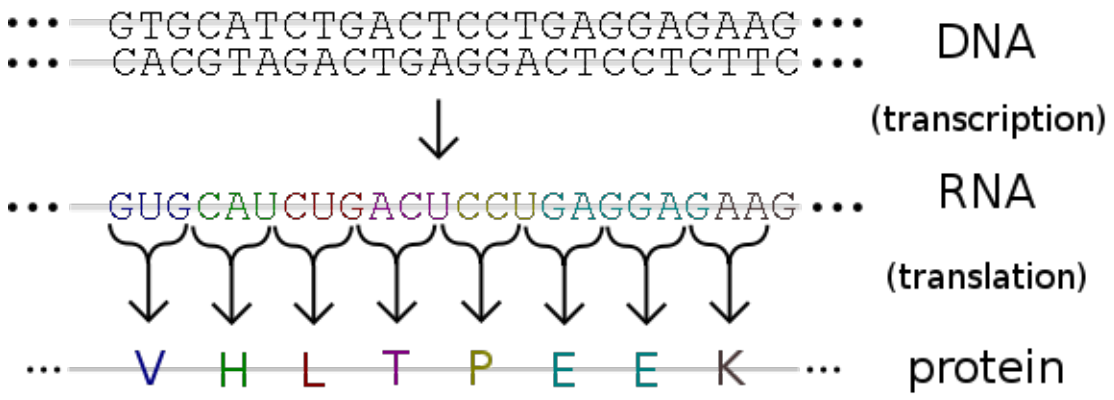
Thomas Hunt Morgan's 1916 illustration of a double crossover between chromosomes

The diploid nature of chromosomes allows for genes on different chromosomes to assort independently during sexual reproduction, recombining to form new combinations of genes. Genes on the same chromosome would theoretically never recombine, however, were it not for the process of chromosomal crossover. During crossover, chromosomes exchange stretches of DNA, effectively shuffling the gene alleles between the chromosomes. This process of chromosomal crossover generally occurs during meiosis, a series of cell divisions that creates haploid cells.

The probability of chromosomal crossover occurring between two given points on the chromosome is related to the distance between the points. For an arbitrarily long distance, the probability of crossover is high enough that the inheritance of the genes is effectively uncorrelated. For genes that are closer together, however, the lower probability of crossover means that the genes demonstrate genetic linkage—alleles for the two genes tend to be inherited together. The amounts of linkage between a series of genes can be combined to form a linear linkage map that roughly describes the arrangement of the genes along the chromosome.

Gene expression

Genetic code

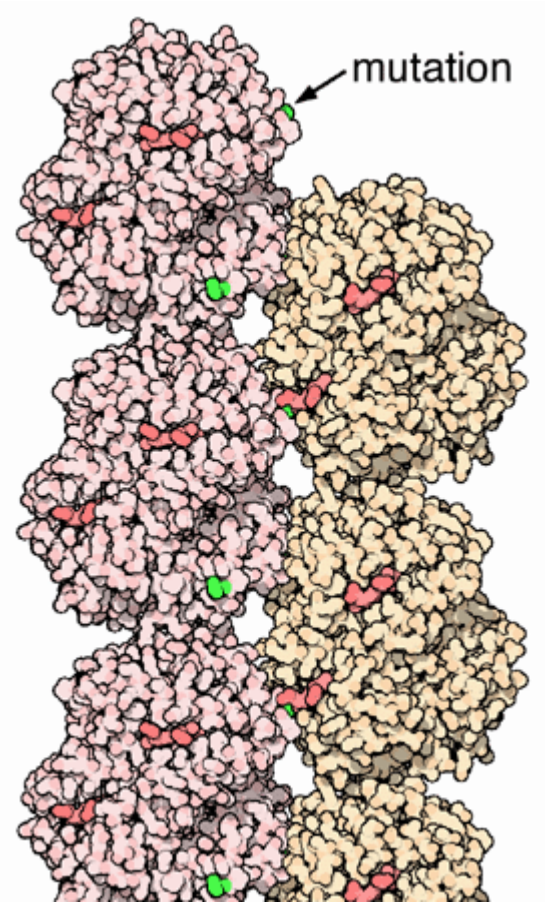


The genetic code: DNA, through a messenger RNA intermediate, codes for protein with a triplet code.

Genes generally express their functional effect through the production of proteins, which are complex molecules responsible for most functions in the cell. Proteins are chains of amino acids, and the DNA sequence of a gene (through an RNA intermediate) is used to produce a specific protein sequence. This process begins with the production of an RNA molecule with a sequence matching the gene's DNA sequence, a process called transcription.

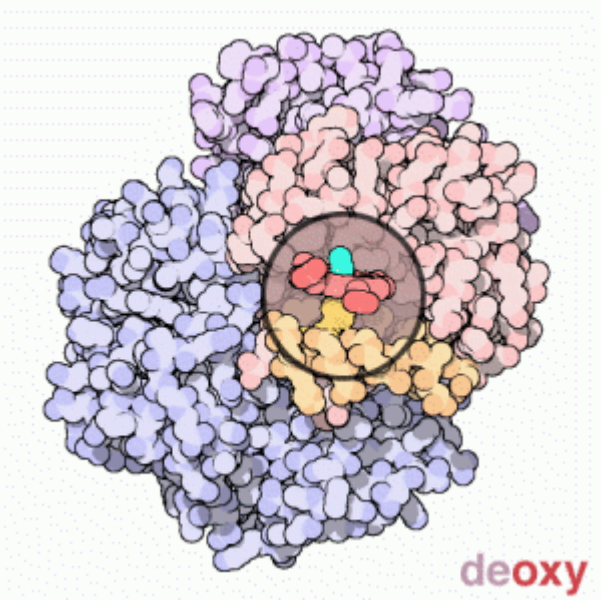
This messenger RNA molecule is then used to produce a corresponding amino acid sequence through a process called translation. Each group of three nucleotides in the sequence, called a codon, corresponds either to one of the twenty possible amino acids in a protein or an instruction to end the amino acic sequence; this correspondence is called

the genetic code. The flow of information is unidirectional: information is transferred from nucleotide sequences into the amino acid sequence of proteins, but it never transfers from protein back into the sequence of DNA—a phenomenon Francis Crick called the central dogma of molecular biology.



A single amino acid change causes hemoglobin to form fibers

The specific sequence of amino acids results in a unique three-dimensional structure for that protein, and the three-dimensional structures of proteins are related to their functions. Some are simple structural molecules, like the fibers formed by the protein collagen. Proteins can bind to other proteins and simple molecules, sometimes acting as enzymes by facilitating chemical reactions within the bound molecules (without changing the structure of the protein itself). Protein structure is dynamic; the protein hemoglobin bends into slightly different forms as it facilitates the capture, transport, and release of oxygen molecules within mammalian blood.



The dynamic structure of hemoglobin is responsible for its ability to transport oxygen within mammalian blood.

A single nucleotide difference within DNA can cause a change in the amino acid sequence of a protein. Because protein structures are the result of their amino acid sequences, some changes can dramatically change the properties of a protein by destabilizing the structure or changing the surface of the protein in a way that changes its interaction with other proteins and molecules. For example, sickle-cell anemia is a human genetic disease that results from a single base difference within the coding region for the β-globin section of hemoglobin, causing a single amino acid change that changes hemoglobin's physical properties. Sickle-cell versions of hemoglobin stick to themselves, stacking to form fibers that distort the shape of red blood cells carrying the protein. These sickle-shaped cells no longer flow smoothly through blood vessels, having a tendency to clog or degrade, causing the medical problems associated with this disease.

Some genes are transcribed into RNA but are not translated into protein products—such RNA molecules are called non-coding RNA. In some cases, these products fold into structures which are involved in critical cell functions (e.g. ribosomal RNA and transfer RNA). RNA can also have regulatory effect through hybridization interactions with other RNA molecules (e.g. microRNA).

Nature versus nurture

Siamese cats have a temperature-sensitive mutation in pigment production

Although genes contain all the information an organism uses to function, the environment plays an important role in determining the ultimate phenotype—a phenomenon often referred to as "nature vs. nurture". The phenotype of an organism depends on the interaction of genetics with the environment. One example of this is the case of temperature-sensitive mutations. Often, a single amino acid change within the sequence of a protein does not change its behavior and interactions with other molecules, but it does destabilize the structure. In a high temperature environment, where molecules are moving more quickly and hitting each other, this results in the protein losing its structure and failing to function. In a low temperature environment, however, the protein's structure is stable and it functions normally. This type of mutation is visible in the coat

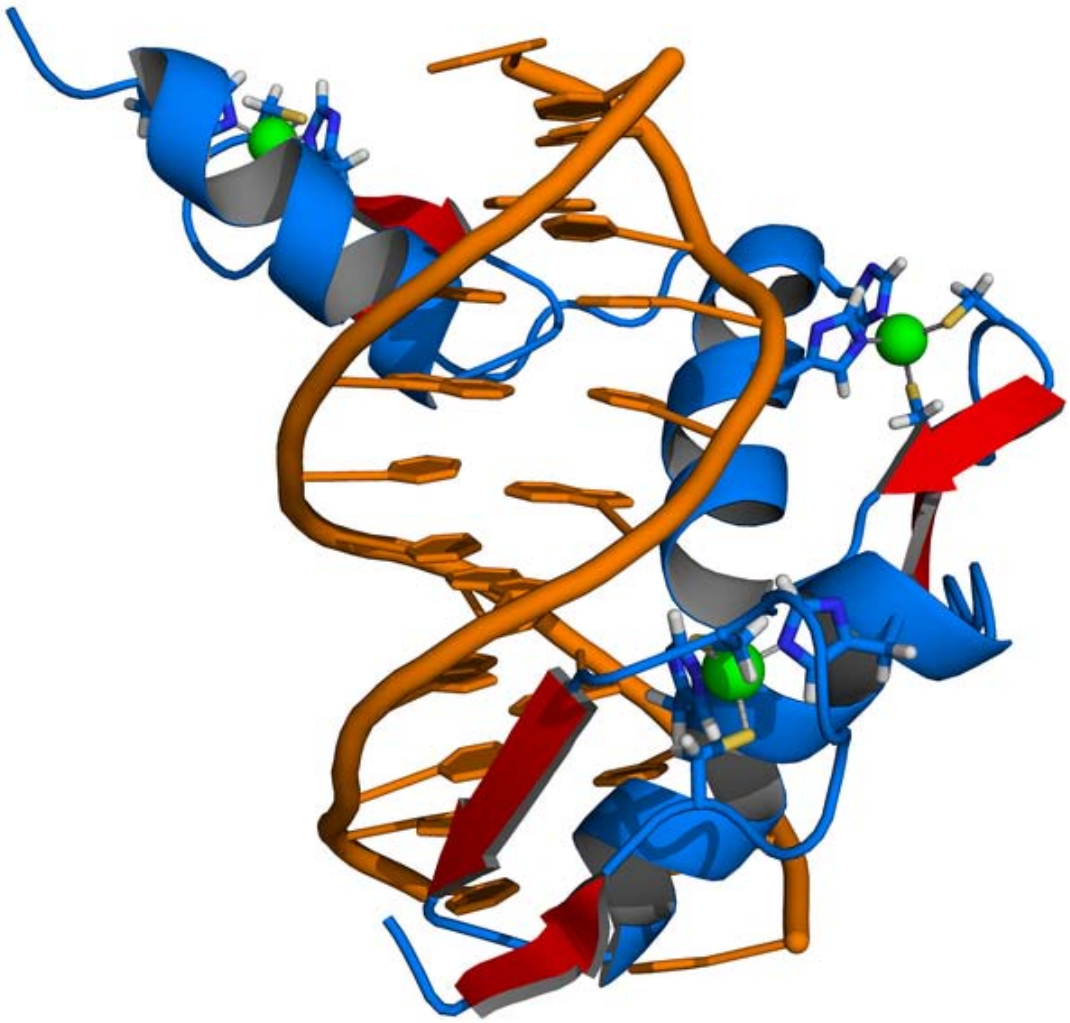
coloration of Siamese cats, where a mutation in an enzyme responsible for pigment production causes it to destabilize and lose function at high temperatures. The protein remains functional in areas of skin that are colder—legs, ears, tail, and face—and so the cat has dark fur at its extremities.

Environment also plays a dramatic role in effects of the human genetic disease phenylketonuria. The mutation that causes phenylketonuria disrupts the ability of the body to break down the amino acid phenylalanine, causing a toxic build-up of an intermediate molecule that, in turn, causes severe symptoms of progressive mental retardation and seizures. If someone with the phenylketonuria mutation follows a strict diet that avoids this amino acid, however, they remain normal and healthy.

A popular method to determine how much role nature and nurture play is to study identical and fraternal twins or siblings of multiple birth. Because identical siblings come from the same zygote they are genetically the same. Fraternal siblings however are as different genetically from one another as normal siblings. By comparing how often the twin of a set has the same disorder between fraternal and identical twins, scientists can see whether there is more of a nature or nurture effect. One famous example of a multiple birth study includes the Genain quadruplets, who were identical quadruplets all diagnosed with schizophrenia.

Gene regulation

The genome of a given organism contains thousands of genes, but not all these genes need to be active at any given moment. A gene is expressed when it is being transcribed into mRNA (and translated into protein), and there exist many cellular methods of controlling the expression of genes such that proteins are produced only when needed by the cell. Transcription factors are regulatory proteins that bind to the start of genes, either promoting or inhibiting the transcription of the gene. Within the genome of *Escherichia coli* bacteria, for example, there exists a series of genes necessary for the synthesis of the amino acid tryptophan. However, when tryptophan is already available to the cell, these genes for tryptophan synthesis are no longer needed. The presence of tryptophan directly affects the activity of the genes—tryptophan molecules bind to the tryptophan repressor (a transcription factor), changing the repressor's structure such that the repressor binds to the genes. The tryptophan repressor blocks the transcription and expression of the genes, thereby creating negative feedback regulation of the tryptophan synthesis process.



Transcription factors bind to DNA, influencing the transcription of associated genes.

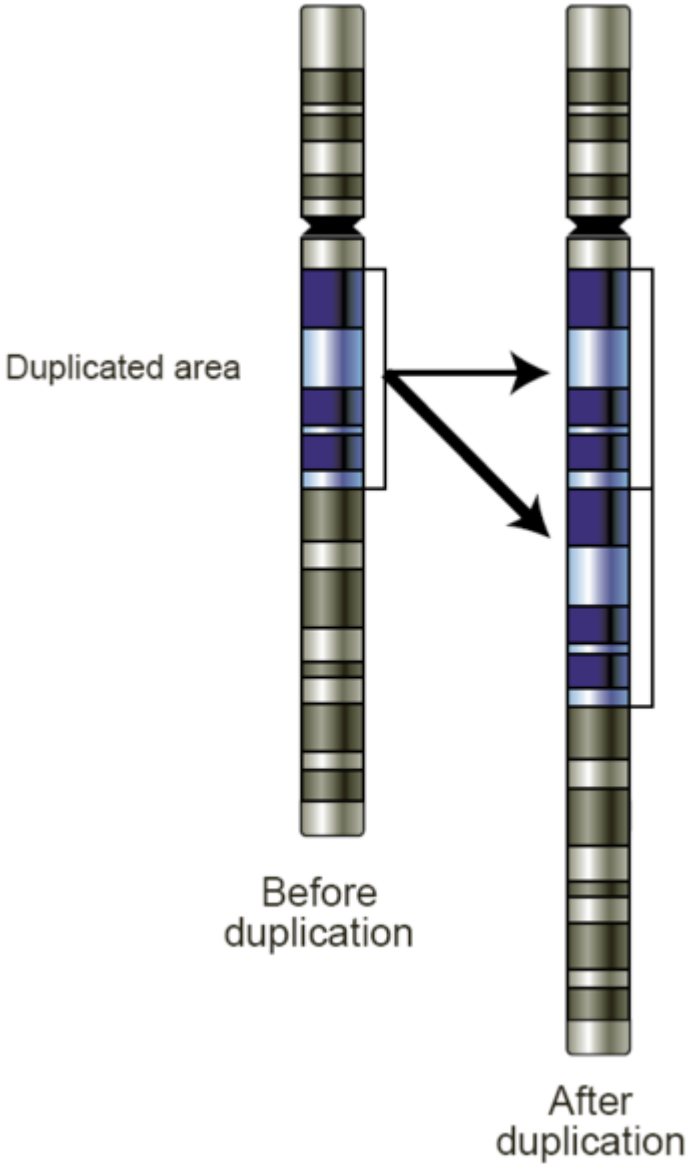
Differences in gene expression are especially clear within multicellular organisms, where cells all contain the same genome but have very different structures and behaviors due to the expression of different sets of genes. All the cells in a multicellular organism derive from a single cell, differentiating into variant cell types in response to external and intercellular signals and gradually establishing different patterns of gene expression to create different behaviors. As no single gene is responsible for the development of structures within multicellular organisms, these patterns arise from the complex interactions between many cells.

Within eukaryotes there exist structural features of chromatin that influence the transcription of genes, often in the form of modifications to DNA and chromatin that are stably inherited by daughter cells. These features are called "epigenetic" because they exist "on top" of the DNA sequence and retain inheritance from one cell generation to the next. Because of epigenetic features, different cell types grown within the same medium

can retain very different properties. Although epigenetic features are generally dynamic over the course of development, some, like the phenomenon of paramutation, have multigenerational inheritance and exist as rare exceptions to the general rule of DNA as the basis for inheritance.

Genetic change

Mutations



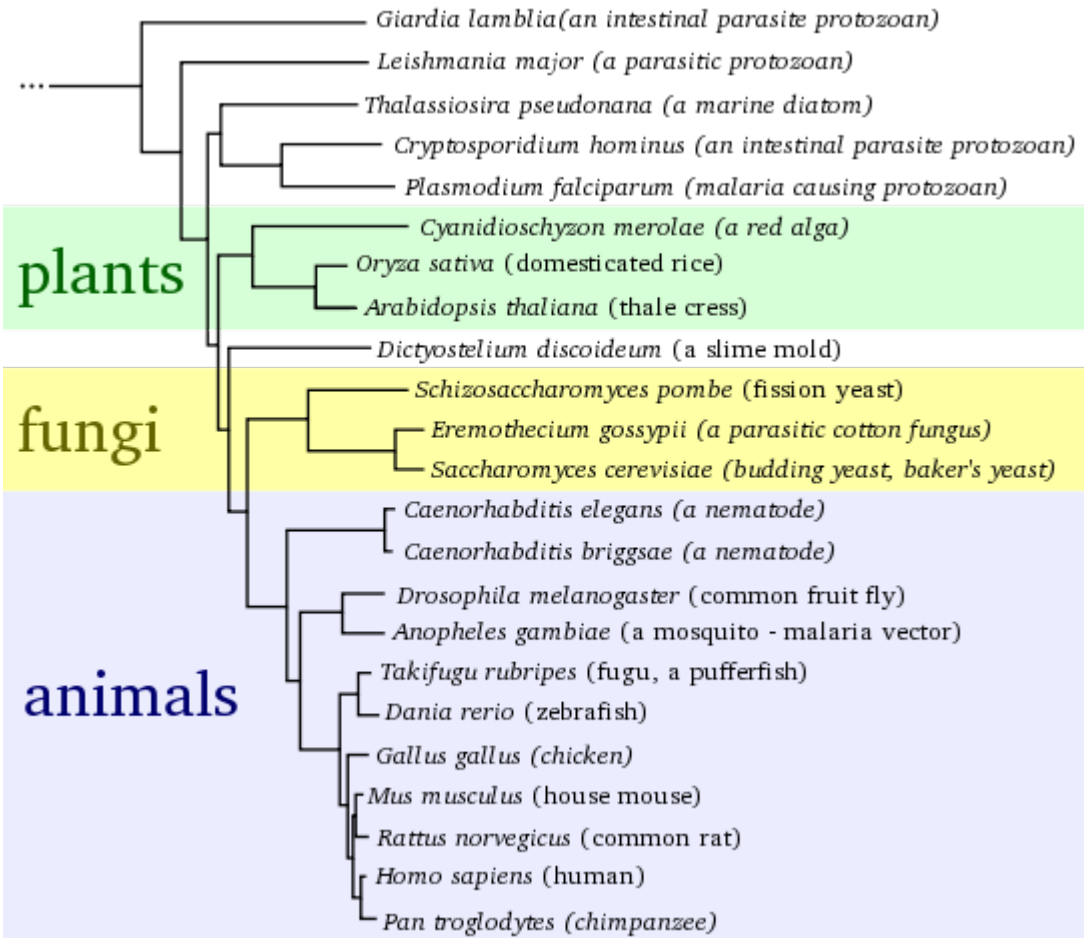
Gene duplication allows diversification by providing redundancy: one gene can mutate and lose its original function without harming the organism.

During the process of DNA replication, errors occasionally occur in the polymerization of the second strand. These errors, called mutations, can have an impact on the phenotype of an organism, especially if they occur within the protein coding sequence of a gene. Error rates are usually very low—1 error in every 10–100 million bases—due to the "proofreading" ability of DNA polymerases. (Without proofreading error rates are a thousandfold higher; because many viruses rely on DNA and RNA polymerases that lack proofreading ability, they experience higher mutation rates.) Processes that increase the rate of changes in DNA are called mutagenic: mutagenic chemicals promote errors in DNA replication, often by interfering with the structure of base-pairing, while UV radiation induces mutations by causing damage to the DNA structure. Chemical damage to DNA occurs naturally as well, and cells use DNA repair mechanisms to repair mismatches and breaks in DNA—nevertheless, the repair sometimes fails to return the DNA to its original sequence.

In organisms that use chromosomal crossover to exchange DNA and recombine genes, errors in alignment during meiosis can also cause mutations. Errors in crossover are especially likely when similar sequences cause partner chromosomes to adopt a mistaken alignment; this makes some regions in genomes more prone to mutating in this way. These errors create large structural changes in DNA sequence—duplications, inversions or deletions of entire regions, or the accidental exchanging of whole parts between different chromosomes (called translocation).

Natural selection and evolution

Mutations alter an organisms genotype and occasionally this causes different phenotypes to appear. Most mutations have little effect on an organism's phenotype, health, or reproductive fitness. Mutations that do have an effect are usually deleterious, but occasionally some can be beneficial. Studies in the fly *Drosophila melanogaster* suggest that if a mutation changes a protein produced by a gene, about 70 percent of these mutations will be harmful with the remainder being either neutral or weakly beneficial.



An evolutionary tree of eukaryotic organisms, constructed by comparison of several orthologous gene sequences

Population genetics studies the distribution of genetic differences within populations and how these distributions change over time. Changes in the frequency of an allele in a population are mainly influenced by natural selection, where a given allele provides a selective or reproductive advantage to the organism, as well as other factors such as genetic drift, artificial selection and migration.

Over many generations, the genomes of organisms can change significantly, resulting in the phenomenon of evolution. Selection for beneficial mutations can cause a species to evolve into forms better able to survive in their environment, a process called adaptation. New species are formed through the process of speciation, often caused by geographical separations that prevent populations from exchanging genes with each other. The application of genetic principles to the study of population biology and evolution is referred to as the modern synthesis.

By comparing the homology between different species' genomes it is possible to calculate the evolutionary distance between them and when they may have diverged (called a molecular clock). Genetic comparisons are generally considered a more accurate method

of characterizing the relatedness between species than the comparison of phenotypic characteristics. The evolutionary distances between species can be used to form evolutionary trees; these trees represent the common descent and divergence of species over time, although they do not show the transfer of genetic material between unrelated species (known as horizontal gene transfer and most common in bacteria).

Research and technology

Model organisms and genetics

The common fruit fly (*Drosophila melanogaster*) is a popular model organism in genetics research.

Although geneticists originally studied inheritance in a wide range of organisms, researchers began to specialize in studying the genetics of a particular subset of organisms. The fact that significant research already existed for a given organism would encourage new researchers to choose it for further study, and so eventually a few model organisms became the basis for most genetics research. Common research topics in model organism genetics include the study of gene regulation and the involvement of genes in development and cancer.

Organisms were chosen, in part, for convenience—short generation times and easy genetic manipulation made some organisms popular genetics research tools. Widely used

model organisms include the gut bacterium *Escherichia coli*, the plant *Arabidopsis thaliana*, baker's yeast (*Saccharomyces cerevisiae*), the nematode *Caenorhabditis elegans*, the common fruit fly (*Drosophila melanogaster*), and the common house mouse (*Mus musculus*).

Medical genetics research

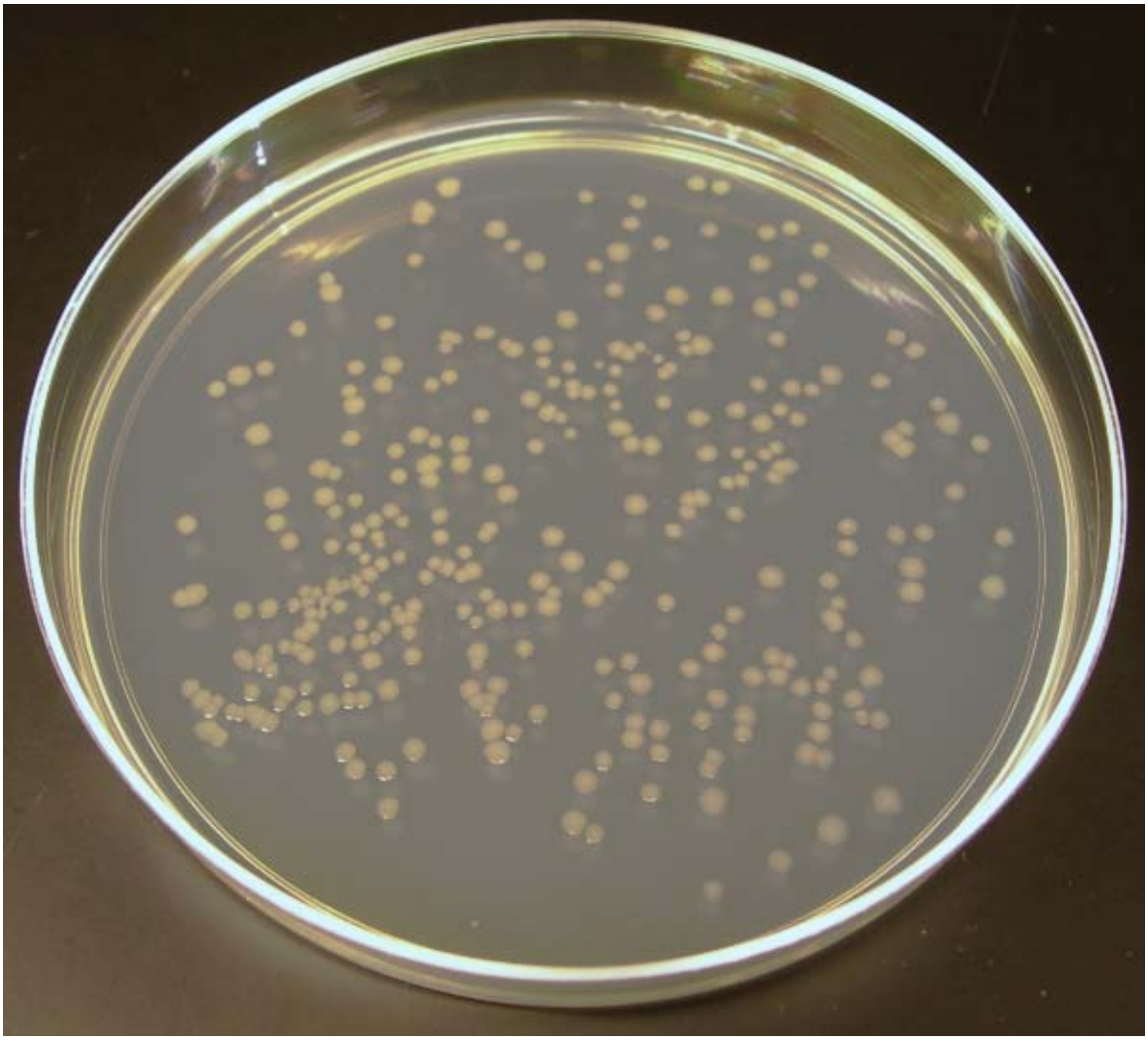
Medical genetics seeks to understand how genetic variation relates to human health and disease. When searching for an unknown gene that may be involved in a disease, researchers commonly use genetic linkage and genetic pedigree charts to find the location on the genome associated with the disease. At the population level, researchers take advantage of Mendelian randomization to look for locations in the genome that are associated with diseases, a technique especially useful for multigenic traits not clearly defined by a single gene. Once a candidate gene is found, further research is often done on the corresponding gene (called an orthologous gene) in model organisms. In addition to studying genetic diseases, the increased availability of genotyping techniques has led to the field of pharmacogenetics—studying how genotype can affect drug responses.

Individuals differ in their inherited tendency to develop cancer, and cancer is a genetic disease. The process of cancer development in the body is a combination of events. Mutations occasionally occur within cells in the body as they divide. Although these mutations will not be inherited by any offspring, they can affect the behavior of cells, sometimes causing them to grow and divide more frequently. There are biological mechanisms that attempt to stop this process; signals are given to inappropriately dividing cells that should trigger cell death, but sometimes additional mutations occur that cause cells to ignore these messages. An internal process of natural selection occurs within the body and eventually mutations accumulate within cells to promote their own growth, creating a cancerous tumor that grows and invades various tissues of the body.

Research techniques

DNA can be manipulated in the laboratory. Restriction enzymes are commonly used enzymes that cut DNA at specific sequences, producing predictable fragments of DNA. DNA fragments can be visualized through use of gel electrophoresis, which separates fragments according to their length.

The use of ligation enzymes allows DNA fragments to be connected, and by ligating fragments of DNA together from different sources, researchers can create recombinant DNA. Often associated with genetically modified organisms, recombinant DNA is commonly used in the context of plasmids—short circular DNA fragments with a few genes on them. By inserting plasmids into bacteria and growing those bacteria on plates of agar (to isolate clones of bacteria cells), researchers can clonally amplify the inserted fragment of DNA (a process known as molecular cloning). (Cloning can also refer to the creation of clonal organisms, through various techniques.)



Colonies of *E. coli* on a plate of agar, an example of cellular cloning and often used in molecular cloning.

DNA can also be amplified using a procedure called the polymerase chain reaction (PCR). By using specific short sequences of DNA, PCR can isolate and exponentially amplify a targeted region of DNA. Because it can amplify from extremely small amounts of DNA, PCR is also often used to detect the presence of specific DNA sequences.

DNA sequencing and genomics

One of the most fundamental technologies developed to study genetics, DNA sequencing allows researchers to determine the sequence of nucleotides in DNA fragments. Developed in 1977 by Frederick Sanger and coworkers, chain-termination sequencing is now routinely used to sequence DNA fragments. With this technology, researchers have been able to study the molecular sequences associated with many human diseases.

As sequencing has become less expensive, researchers have sequenced the genomes of many organisms, using computational tools to stitch together the sequences of many different fragments (a process called genome assembly). These technologies were used to sequence the human genome, leading to the completion of the Human Genome Project in 2003. New high-throughput sequencing technologies are dramatically lowering the cost of DNA sequencing, with many researchers hoping to bring the cost of resequencing a human genome down to a thousand dollars.

The large amount of sequence data available has created the field of genomics, research that uses computational tools to search for and analyze patterns in the full genomes of organisms. Genomics can also be considered a subfield of bioinformatics, which uses computational approaches to analyze large sets of biological data.

Chapter- 2

History of Genetics

The **history of genetics** started with the work of the Augustinian friar Gregor Johann Mendel. His work on pea plants, published in 1866, described what came to be known as Mendelian Inheritance. In the centuries before—and for several decades after—Mendel's work, a wide variety of theories of heredity proliferated.

1900 marked the "rediscovery of Mendel" by Hugo de Vries, Carl Correns and Erich von Tschermak, and by 1915 the basic principles of Mendelian genetics had been applied to a wide variety of organisms—most notably the fruit fly *Drosophila melanogaster*. Led by Thomas Hunt Morgan and his fellow "drosophilists", geneticists developed the Mendelian, which was widely accepted by 1925. Alongside experimental work, mathematicians developed the statistical framework of population genetics, bring genetical explanations into the study of evolution.

With the basic patterns of genetic inheritance established, many biologists turned to investigations of the physical nature of the gene. In the 1940s and early 1950s, experiments pointed to DNA as the portion of chromosomes (and perhaps other nucleoproteins) that held genes. A focus on new model organisms such as viruses and bacteria, along with the discovery of the double helical structure of DNA in 1953, marked the transition to the era of molecular genetics.

In the following years, chemists developed techniques for sequencing both nucleic acids and proteins, while others worked out the relationship between the two forms of biological molecules: the genetic code. The regulation of gene expression became a central issue in the 1960s; by the 1970s gene expression could be controlled and manipulated through genetic engineering. In the last decades of the 20th century, many biologists focused on large-scale genetics projects, sequencing entire genomes.

Pre-Mendelian ideas on heredity

Ancient theories

The most influential early theories of heredity were that of Hippocrates and Aristotle. Hippocrates' theory (possibly based on the teachings of Anaxagoras) was similar to

Darwin's later ideas on pangenesis, involving heredity material that collects from throughout the body. Aristotle suggested instead that the (nonphysical) form-giving principle of an organism was transmitted through semen (which he considered to be a purified form of blood) and the mother's menstrual blood, which interacted in the womb to direct an organism's early development. For both Hippocrates and Aristotle—and nearly all Western scholars through to the late 19th century—the inheritance of acquired characters was a supposedly well-established fact that any adequate theory of heredity had to explain. At the same time, individual species were taken to have a fixed essence; such inherited changes were merely superficial.

In the 9th century CE, the Afro-Arab writer Al-Jahiz considered the effects of the environment on the likelihood of an animal to survive, and first described the struggle for existence. His ideas on the struggle for existence in the *Book of Animals* have been summarized as follows:

"Animals engage in a struggle for existence; for resources, to avoid being eaten and to breed. Environmental factors influence organisms to develop new characteristics to ensure survival, thus transforming into new species. Animals that survive to breed can pass on their successful characteristics to offspring."

In 1000 CE, the Arab physician, Abu al-Qasim al-Zahrawi (known as Albucasis in the West), wrote the first clear description of haemophilia, a hereditary genetic disorder, in his *Al-Tasrif*. In this work, he wrote of an Andalusian family whose males died of bleeding after minor injuries.

Plant systematics and hybridization

In the 18th century, with increased knowledge of plant and animal diversity and the accompanying increased focus on taxonomy, new ideas about heredity began to appear. Linnaeus and others (among them Joseph Gottlieb Kölreuter, Carl Friedrich von Gärtner, and Charles Naudin) conducted extensive experiments with hybridization, especially species hybrids. Species hybridizers described a wide variety of inheritance phenomena, include hybrid sterility and the high variability of back-crosses.

Plant breeders were also developing an array of stable varieties in many important plant species. In the early 19th century, Augustin Sageret established the concept of dominance, recognizing that when some plant varieties are crossed, certain characters (present in one parent) usually appear in the offspring; he also found that some ancestral characters found in neither parent may appear in offspring. However, plant breeders made little attempt to establish a theoretical foundation for their work or to share their knowledge with current work of physiology, although Gartons Agricultural Plant Breeders in England explained their system in their seed catalogue of 1901.

Mendel

In breeding experiments between 1856 and 1865, Gregor Mendel first traced inheritance patterns of certain traits in pea plants and showed that they obeyed simple statistical rules. Although not all features show these patterns of Mendelian inheritance, his work acted as a proof that application of statistics to inheritance could be highly useful. Since that time many more complex forms of inheritance have been demonstrated.

From his statistical analysis Mendel defined a concept that he described as an *allele*, which was the fundamental unit of heredity. The term *allele* as Mendel used it is nearly synonymous with the term *gene*, and now means a specific variant of a particular gene.

Mendel's work was published in 1866 as *"Versuche über Pflanzen-Hybriden" (Experiments on Plant Hybridization)* in the *Verhandlungen des Naturforschenden Vereins zu Brünn (Proceedings of the Natural History Society of Brünn)*, following two lectures he gave on the work in early 1865.

Post-Mendel, pre-re-discovery

Mendel's work was published in a relatively obscure scientific journal, and it was not given any attention in the scientific community. Instead, discussions about modes of heredity were galvanized by Darwin's theory of evolution by natural selection, in which mechanisms of non-Lamarckian heredity seemed to be required. Darwin's own theory of heredity, pangenesis, did not meet with any large degree of acceptance. A more mathematical version of pangenesis, one which dropped much of Darwin's Lamarckian holdovers, was developed as the "biometrical" school of heredity by Darwin's cousin, Francis Galton. Under Galton and his successor Karl Pearson, the biometrical school attempted to build statistical models for heredity and evolution, with some limited but real success, though the exact methods of heredity were unknown and largely unquestioned.

Classical genetics

The significance of Mendel's work was not understood until early in the twentieth century, after his death, when his research was re-discovered by other scientists working on similar problems. Hugo de Vries, Carl Correns and Erich von Tschermak

There was then a feud between Bateson and Pearson over the hereditary mechanism. Fisher solved this in The Correlation Between Relatives on the Supposition of Mendelian Inheritance

1865 Gregor Mendel's paper, *Experiments on Plant Hybridization*
1869 Friedrich Miescher discovers a weak acid in the nuclei of white blood cells that today we call DNA
1880-1890 Walther Flemming, Eduard Strasburger, and Edouard van Beneden elucidate chromosome distribution during cell division

1889 Hugo de Vries postulates that "inheritance of specific traits in organisms comes in particles", naming such particles "(pan)genes"
1903 Walter Sutton hypothesizes that chromosomes, which segregate in a Mendelian fashion, are hereditary units
1905 William Bateson coins the term "genetics" in a letter to Adam Sedgwick and at a meeting in 1906
1908 Hardy-Weinberg law derived.
1910 Thomas Hunt Morgan shows that genes reside on chromosomes
1913 Alfred Sturtevant makes the first genetic map of a chromosome
1913 Gene maps show chromosomes containing linear arranged genes
1918 Ronald Fisher publishes "The Correlation Between Relatives on the Supposition of Mendelian Inheritance" the modern synthesis of genetics and evolutionary biology starts.
1928 Frederick Griffith discovers that hereditary material from dead bacteria can be incorporated into live bacteria
1931 Crossing over is identified as the cause of recombination
1933 Jean Brachet is able to show that DNA is found in chromosomes and that RNA is present in the cytoplasm of all cells.
1941 Edward Lawrie Tatum and George Wells Beadle show that genes code for proteins

The DNA era

1944 The Avery–MacLeod–McCarty experiment isolates DNA as the genetic material (at that time called transforming principle)
1950 Erwin Chargaff shows that the four nucleotides are not present in nucleic acids in stable proportions, but that some general rules appear to hold (e.g., that the amount of adenine, A, tends to be equal to that of thymine, T).
Barbara McClintock discovers transposons in maize
1952 The Hershey-Chase experiment proves the genetic information of phages (and all other organisms) to be DNA
1953 DNA structure is resolved to be a double helix by James D. Watson and Francis Crick
1956 Joe Hin Tjio and Albert Levan established the correct chromosome number in humans to be 46
1958 The Meselson-Stahl experiment demonstrates that DNA is semiconservatively replicated
1961-1967 Combined efforts of scientists "crack" the genetic code, including Marshall Nirenberg, Har Gobind Khorana, Sydney Brenner & Francis Crick
1964 Howard Temin showed using RNA viruses that the direction of DNA to RNA transcription can be reversed
1970 Restriction enzymes were discovered in studies of a bacterium, *Haemophilus influenzae*, enabling scientists to cut and paste DNA

The genomics era

1972, Walter Fiers and his team at the Laboratory of Molecular Biology of the University of Ghent (Ghent, Belgium) were the first to determine the sequence of a gene: the gene for bacteriophage MS2 coat protein.
1976, Walter Fiers and his team determine the complete nucleotide-sequence of bacteriophage MS2-RNA
1977 DNA is sequenced for the first time by Fred Sanger, Walter Gilbert, and Allan Maxam working independently. Sanger's lab sequence the entire genome of bacteriophage Φ-X174.
1983 Kary Banks Mullis discovers the polymerase chain reaction enabling the easy amplification of DNA
1989 The human gene that encodes the CFTR protein was sequenced by Francis Collins and Lap-Chee Tsui. Defects in this gene cause cystic fibrosis.
1995 The genome of *Haemophilus influenzae* is the first genome of a free living organism to be sequenced
1996 *Saccharomyces cerevisiae* is the first eukaryote genome sequence to be released
1998 The first genome sequence for a multicellular eukaryote, *Caenorhabditis elegans*, is released
2001 First draft sequences of the human genome are released simultaneously by the Human Genome Project and Celera Genomics.
2003 (14 April) Successful completion of Human Genome Project with 99% of the genome sequenced to a 99.99% accuracy

Chapter- 3

Mendelian Inheritance

Mendelian inheritance (or **Mendelian genetics** or **Mendelism**) is a set of primary tenets relating to the transmission of hereditary characteristics from parent organisms to their offspring; it underlies much of genetics. They were initially derived from the work of Gregor Johann Mendel published in 1865 and 1866 which was "re-discovered" in 1900, and were initially very controversial. When they were integrated with the chromosome theory of inheritance by Thomas Hunt Morgan in 1915, they became the core of classical genetics.

History

The laws of inheritance were derived by Gregor Johann Mendel, a 19th century Austrian Priest/monk conducting hybridization experiments in garden peas (*Pisum sativum*). Between 1856 and 1863, he cultivated and tested some 29,000 pea plants. From these experiments he deduced two generalizations which later became known as *Mendel's Principles of Heredity* or *Mendelian inheritance*. He described these principles in a two part paper, *Experiments on Plant Hybridization* that he read to the Natural History Society of Brno on February 8 and March 8, 1865, and which was published in 1866.

Mendel's conclusions were largely ignored. Although they were not completely unknown to biologists of the time, they were not seen as generally applicable, even by Mendel himself, who thought they only applied to certain categories of species or traits. A major block to understanding their significance was the importance attached by 19th century biologists to the apparent blending of inherited traits in the overall appearance of the progeny, now known to be due to multigene interactions, in contrast to the organ-specific binary characters studied by Mendel. In 1900, however, his work was "re-discovered" by three European scientists, Hugo de Vries, Carl Correns, and Erich von Tschermak. The exact nature of the "re-discovery" has been somewhat debated: De Vries published first on the subject, mentioning Mendel in a footnote, while Correns pointed out Mendel's priority after having read De Vries's paper and realizing that he himself did not have priority. De Vries may not have acknowledged truthfully how much of his knowledge of the laws came from his own work, or came only after reading Mendel's paper. Later scholars have accused Von Tschermak of not truly understanding the results at all.

Regardless, the "re-discovery" made Mendelism an important but controversial theory. Its most vigorous promoter in Europe was William Bateson, who coined the term "genetics", "gene", and "allele" to describe many of its tenets. The model of heredity was highly contested by other biologists because it implied that heredity was discontinuous, in opposition to the apparently continuous variation observable for many traits. Many biologists also dismissed the theory because they were not sure it would apply to all species, and there seemed to be very few true Mendelian characters in nature. However later work by biologists and statisticians such as R.A. Fisher showed that if multiple Mendelian factors were involved in the expression of an individual trait, they could produce the diverse results observed. Thomas Hunt Morgan and his assistants later integrated the theoretical model of Mendel with the chromosome theory of inheritance, in which the chromosomes of cells were thought to hold the actual hereditary material, and create what is now known as classical genetics, which was extremely successful and cemented Mendel's place in history.

Mendel's findings allowed other scientists to predict the expression of traits on the basis of mathematical probabilities. A large contribution to Mendel's success can be traced to his decision to start his crosses only with plants he demonstrated were true-breeding. He also only measured absolute (binary) characteristics, such as color, shape, and position of the offspring, rather than quantitative characteristics. He expressed his results numerically and subjected them to statistical analysis. His method of data analysis and his large sample size gave credibility to his data. He also had the foresight to follow several successive generations (f2, f3) of his pea plants and record their variations. Finally, he performed "test crosses" (back-crossing descendants of the initial hybridization to the initial true-breeding lines) to reveal the presence and proportion of recessive characters. Without his hard work and careful attention to procedure and detail, Mendel's work could not have had the impact it made on the world of genetics.

Mendel's Laws

Mendel discovered that when crossing white flower and purple flower plants, the result is not a blend. Rather than being a mix of the two, the offspring was purple flowered. He then conceived the idea of heredity units, which he called "factors", one of which is a recessive characteristic and the other dominant. Mendel said that factors, later called genes, normally occur in pairs in ordinary body cells, yet segregate during the formation of sex cells. Each member of the pair becomes part of the separate sex cell. The dominant gene, such as the purple flower in Mendel's plants, will hide the recessive gene, the white flower. After Mendel self-fertilized the F1 generation and obtained the 3:1 ratio, he correctly theorized that genes can be paired in three different ways for each trait: AA, aa, and Aa. The capital "A" represents the dominant factor and lowercase "a" represents the recessive. (The last combination listed above, Aa, will occur roughly twice as often as each of the other two, as it can be made in two different ways, Aa or aA.)

Mendel stated that each individual has two factors for each trait, one from each parent. The two factors may or may not contain the same information. If the two factors are identical, the individual is called homozygous for the trait. If the two factors have

different information, the individual is called heterozygous. The alternative forms of a factor are called alleles. The genotype of an individual is made up of the many alleles it possesses. An individual's physical appearance, or phenotype, is determined by its alleles as well as by its environment. An individual possesses two alleles for each trait; one allele is given by the female parent and the other by the male parent. They are passed on when an individual matures and produces gametes: egg and sperm. When gametes form, the paired alleles separate randomly so that each gamete receives a copy of one of the two alleles. The presence of an allele doesn't promise that the trait will be expressed in the individual that possesses it. In heterozygous individuals the only allele that is expressed is the dominant. The recessive allele is present but its expression is hidden.

Mendel summarized his findings in two laws; the **Law of Segregation** and the **Law of Independent Assortment**.

Law of Segregation (The "First Law")

The Law of Segregation states that when any individual produces gametes, the copies of a gene separate so that each gamete receives only one copy. A gamete will receive one allele or the other. The direct proof of this was later found following the observation of meiosis by two independent scientists, the German botanist, Oscar Hertwig in 1876, and the Belgian zoologist, Edouard Van Beneden in 1883. In meiosis the paternal and maternal chromosomes get separated and the alleles with the traits of a character are segregated into two different gametes.

OR

The two coexisting alleles of an individual for each trait segregate (separate) during gamete formation so that each gamete gets only one of the two alleles. Alleles again unite at random fertilization of gametes.

Law of Independent Assortment (The "Second Law")

The Law of Independent Assortment, also known as "Inheritance Law" states that alleles of different genes assort independently of one another during gamete formation. While Mendel's experiments with mixing one trait always resulted in a 3:1 ratio (Fig. 1) between dominant and recessive phenotypes, his experiments with mixing two traits (dihybrid cross) showed 9:3:3:1 ratios (Fig. 2). But the 9:3:3:1 table shows that each of the two genes are independently inherited with a 3:1 phenotypic ratio. Mendel concluded that different traits are inherited independently of each other, so that there is no relation, for example, between a cat's color and tail length. This is actually only true for genes that are not linked to each other.

Independent assortment occurs during meiosis I in eukaryotic organisms, specifically metaphase I of *meiosis*, to produce a gamete with a mixture of the organism's maternal and paternal chromosomes. Along with chromosomal crossover, this process aids in increasing genetic diversity by producing novel genetic combinations.

Of the 46 chromosomes in a normal diploid human cell, half are maternally-derived (from the mother's egg) and half are paternally-derived (from the father's sperm). This occurs as sexual reproduction involves the fusion of two haploid gametes (the egg and sperm) to produce a new organism having the full complement of chromosomes. During gametogenesis—the production of new gametes by an adult—the normal complement of 46 chromosomes needs to be halved to 23 to ensure that the resulting haploid gamete can join with another gamete to produce a diploid organism. An error in the number of chromosomes, such as those caused by a diploid gamete joining with a haploid gamete, is termed aneuploidy.

In independent assortment the chromosomes that end up in a newly-formed gamete are randomly sorted from all possible combinations of maternal and paternal chromosomes. Because gametes end up with a random mix instead of a pre-defined "set" from either parent, gametes are therefore considered assorted independently. As such, the gamete can end up with any combination of paternal or maternal chromosomes. Any of the possible combinations of gametes formed from maternal and paternal chromosomes will occur with equal frequency. For human gametes, with 23 pairs of chromosomes, the number of possibilities is 2^{23} or 8,388,608 possible combinations. The gametes will normally end up with 23 chromosomes, but the origin of any particular one will be randomly selected from paternal or maternal chromosomes. This contributes to the genetic variability of progeny.

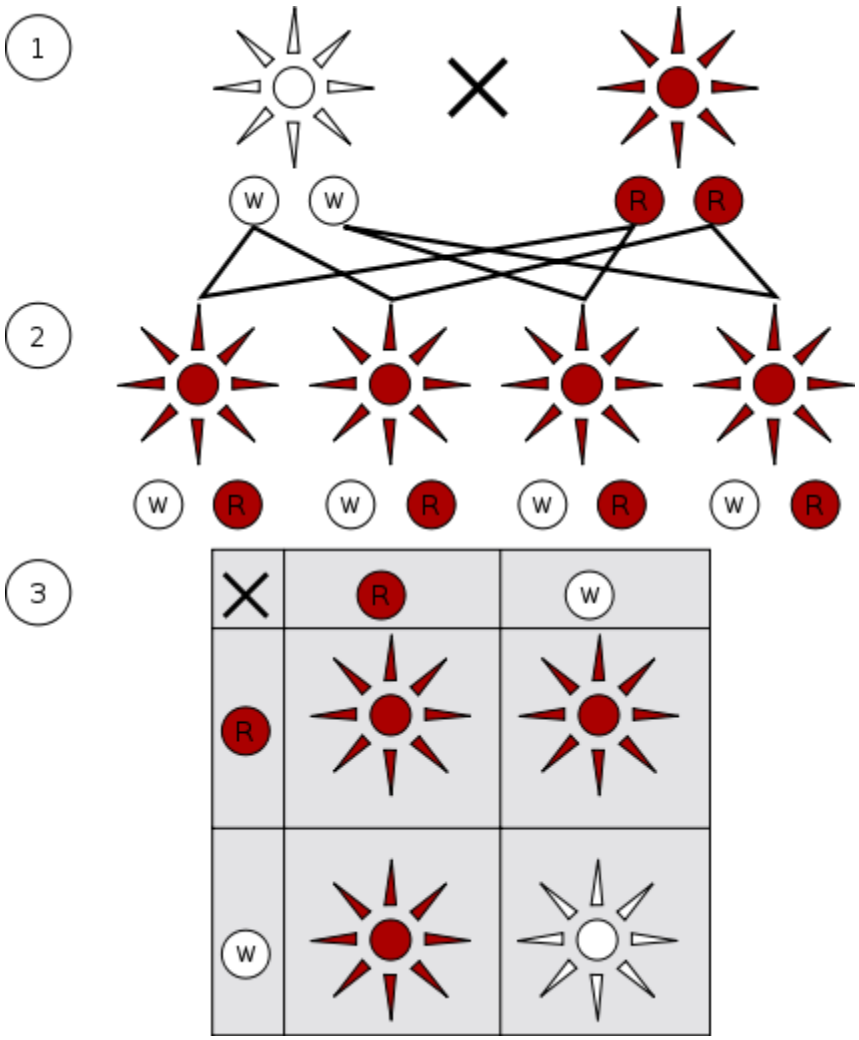


Figure 1: Dominant and recessive phenotypes.
(1) Parental generation. (2) F_1 generation. (3) F_2 generation. Dominant (red) and recessive (white) phenotype look alike in the F_1 (first) generation and show a 3:1 ratio in the F_2 (second) generation

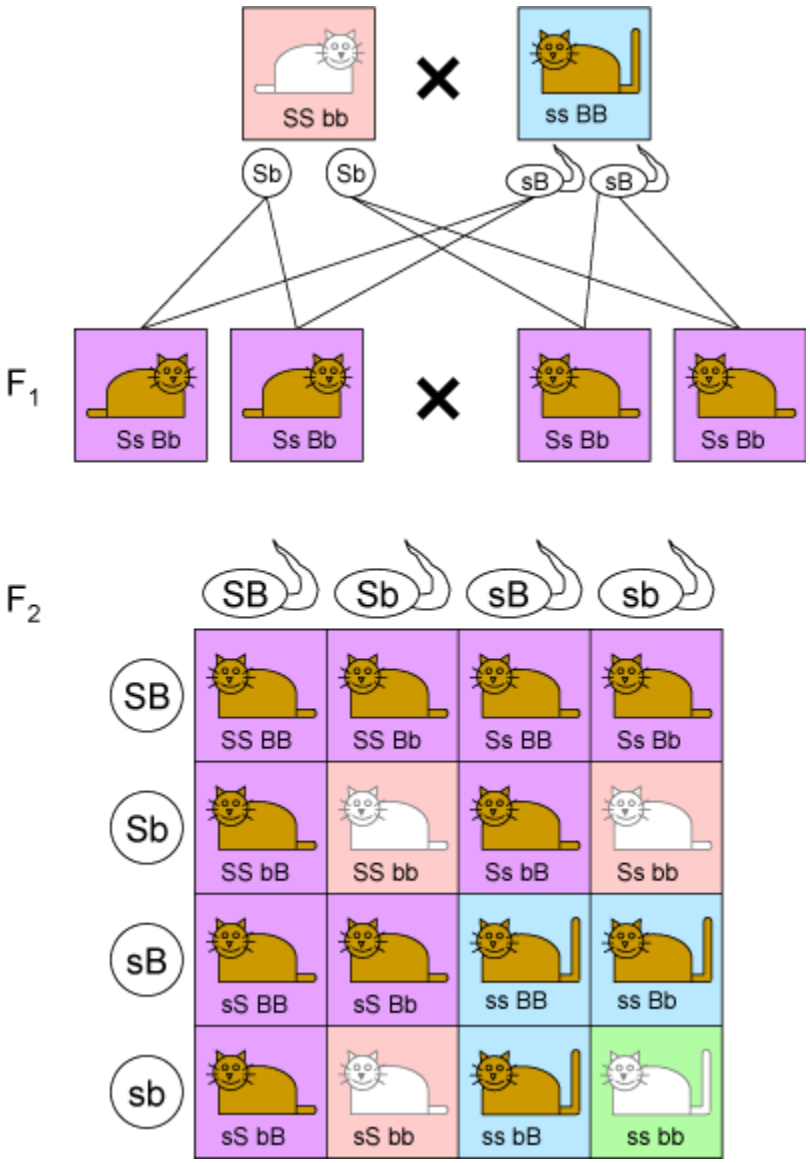


Figure 2: The phenotypes of two independent traits show a 9:3:3:1 ratio in the F_2 generation. In this example, coat color is indicated by **B** (brown, dominant) or **b** (white) while tail length is indicated by **S** (short, dominant) or **s** (long). When parents are homozygous for each trait ('*SSbb* ***and*** *ssBB*), their children in the F_1 generation are heterozygous at both loci and only show the dominant phenotypes. If the children mate with each other, in the F_2 generation all combination of coat color and tail length occur: 9 are brown/short (purple boxes), 3 are white/short (pink boxes), 3 are brown/long (blue boxes) and 1 is white/long (green box).

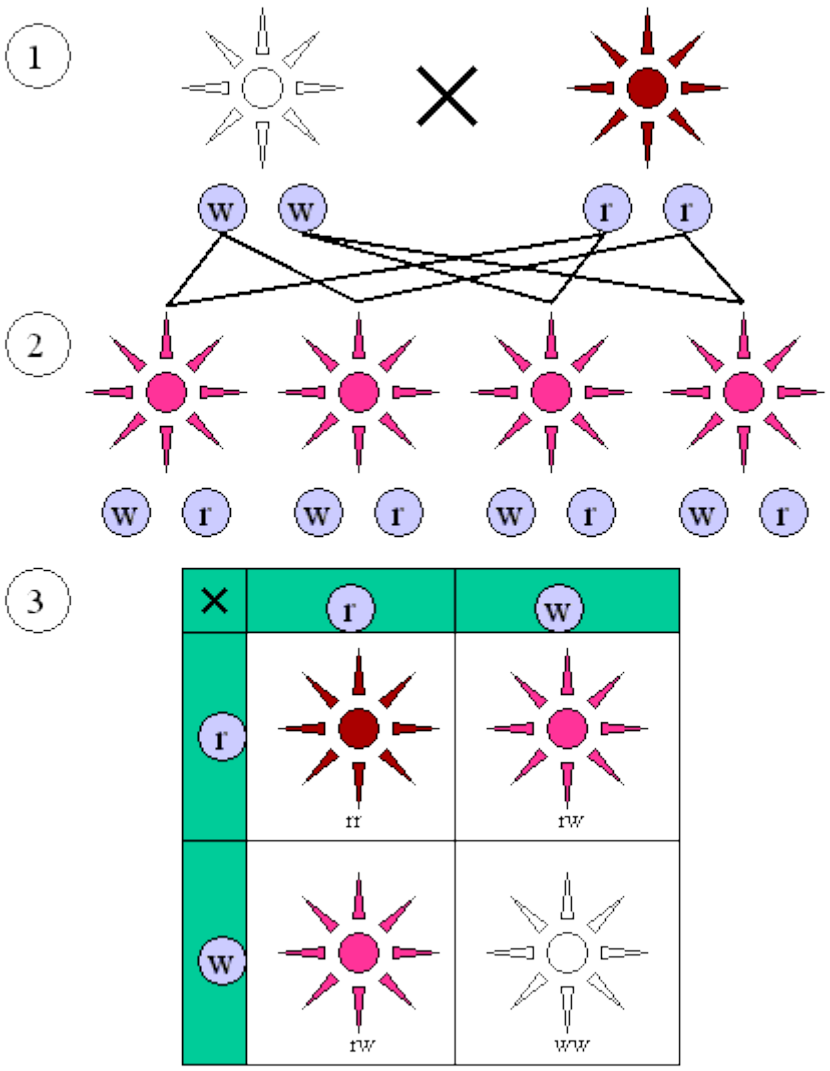


Figure 3: The color alleles of *Mirabilis jalapa* are not dominant or recessive.
(1) Parental generation. (2) F_1 generation. (3) F_2 generation. The "red" and "white" allele together make a "pink" phenotype, resulting in a 1:2:1 ratio of red:pink:white in the F_2 generation.

Background

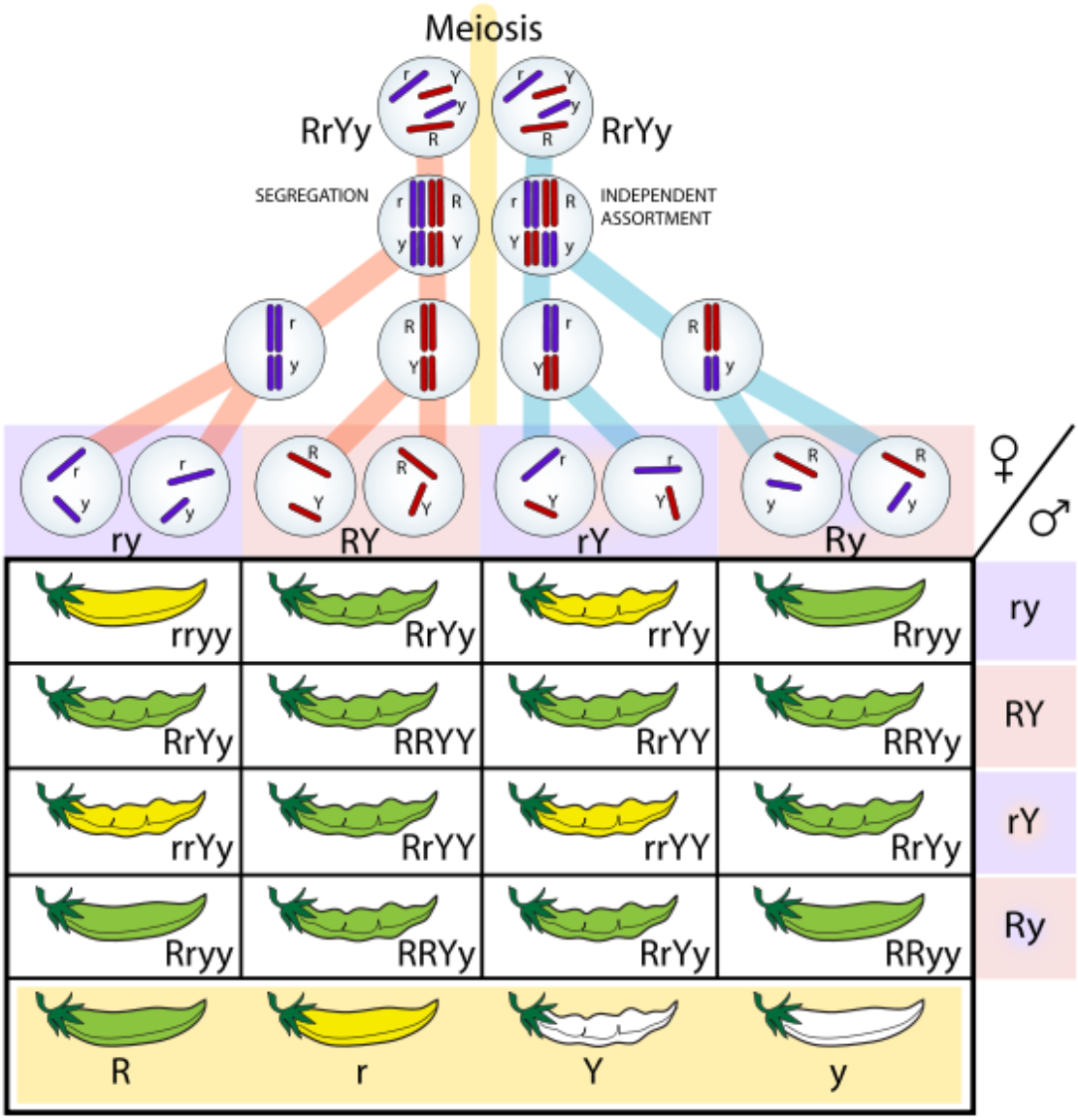


Table showing how the genes exchange according to segregation or independent assortment during meiosis and how this translates into Mendel's laws

Mendel's law: Forget-me-not

The reason for these laws is found in the nature of the cell nucleus. It is made up of several chromosomes carrying the genetic traits. In a normal cell, each of these chromosomes has two parts, the chromatids. A reproductive cell, which is created in a process called meiosis, usually contains only one of those chromatids of each chromosome. By merging two of these cells (usually one male and one female), the full set is restored and the genes are mixed. The resulting cell becomes a new embryo. The fact that this new life has half the genes of each parent (23 from mother, 23 from father for total of 46 in the case of humans) is one reason for the Mendelian laws. The second most important reason is the varying dominance of different genes, causing some traits to appear unevenly instead of averaging out (whereby dominant doesn't mean more likely to reproduce—recessive genes can become the most common, too).

There are several advantages of this method (sexual reproduction) over reproduction without genetic exchange:

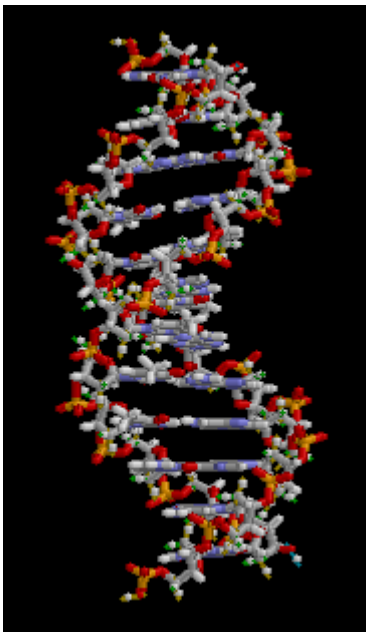
1. Instead of nearly identical copies of an organism, a broad range of offspring develops, allowing more different abilities and evolutionary strategies.
2. There are usually some errors in every cell nucleus. Copying the genes usually adds more of them. By distributing them randomly over different chromosomes and mixing the genes, such errors will be distributed unevenly over the different children. Some of them will therefore have only very few such problems. This helps reduce problems with copying errors somewhat.
3. Genes can spread faster from one part of a population to another. This is for instance useful if there's a temporary isolation of two groups. New genes developing in each of the populations don't get reduced to half when one side replaces the other, they mix and form a population with the advantages of both sides.
4. Sometimes, a mutation can have positive side effects. For example, sickle cell anemia is a mutation that can causethe benefit of malaria resistance. The mechanism behind the Mendelian laws can make it possible for some offspring to carry the advantages without the disadvantages until further mutations solve the problems.

Mendelian trait

A **Mendelian trait** is one that is controlled by a single locus and shows a simple Mendelian inheritance pattern. In such cases, a mutation in a single gene can cause a disease that is inherited according to Mendel's laws. Examples include sickle-cell anemia, Tay-Sachs disease, cystic fibrosis and xeroderma pigmentosa. A disease controlled by a single gene contrasts with a multi-factorial disease, like arthritis, which is affected by several loci (and the environment) as well as those diseases inherited in a non-Mendelian fashion. The Mendelian Inheritance in Man database is a catalog of, among other things, genes in which Mendelian traits cause disease.

Chapter- 4

DNA



The structure of part of a DNA double helix

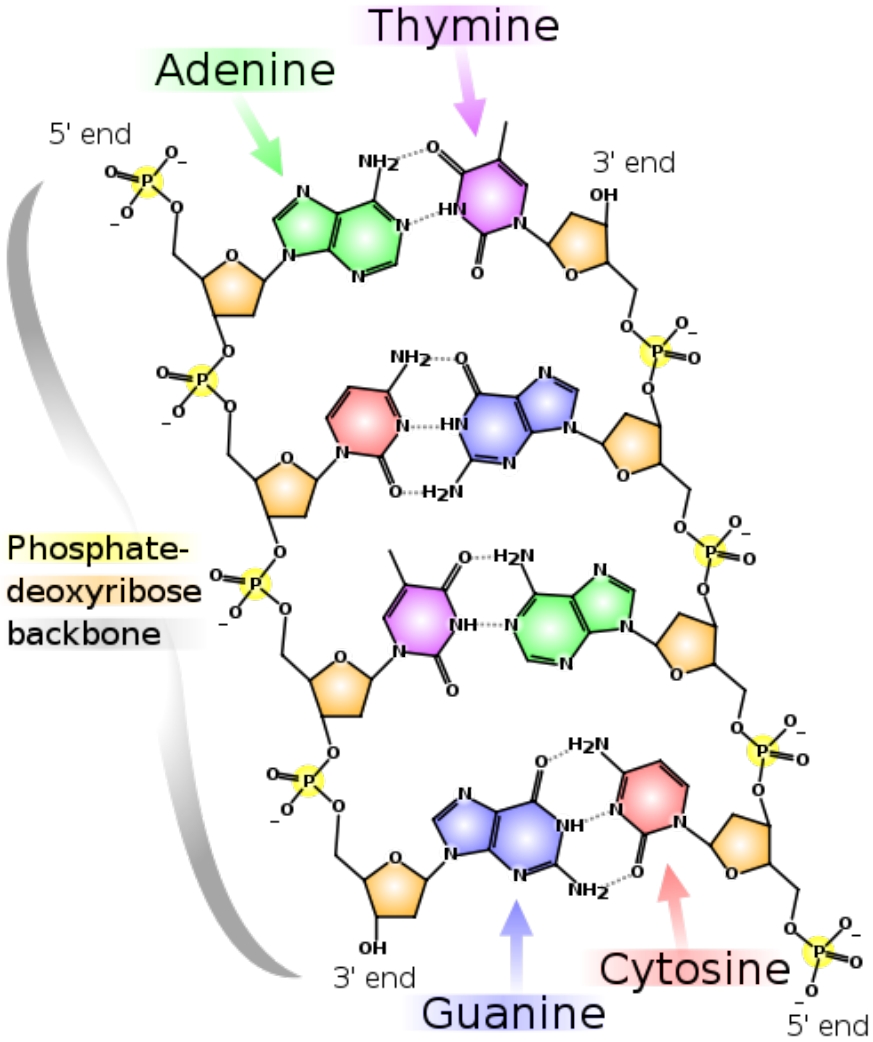
Deoxyribonucleic acid, or **DNA**, is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms (with the exception of RNA viruses). The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints, like a recipe or a code, since it contains the instructions needed to construct other components of cells, such as proteins and RNA molecules. The DNA segments that carry this genetic information are called genes, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information.

DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are therefore anti-parallel. Attached to each sugar is one of four types of molecules called bases. It is the sequence of these four bases along the backbone that encodes information. This information is read using the genetic code, which specifies the sequence of the amino acids within proteins. The code is read by

copying stretches of DNA into the related nucleic acid RNA, in a process called transcription.

Within cells, DNA is organized into long structures called chromosomes. These chromosomes are duplicated before cells divide, in a process called DNA replication. Eukaryotic organisms (animals, plants, fungi, and protists) store most of their DNA inside the cell nucleus and some of their DNA in organelles, such as mitochondria or chloroplasts. In contrast, prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm. Within the chromosomes, chromatin proteins such as histones compact and organize DNA. These compact structures guide the interactions between DNA and other proteins, helping control which parts of the DNA are transcribed.

Properties

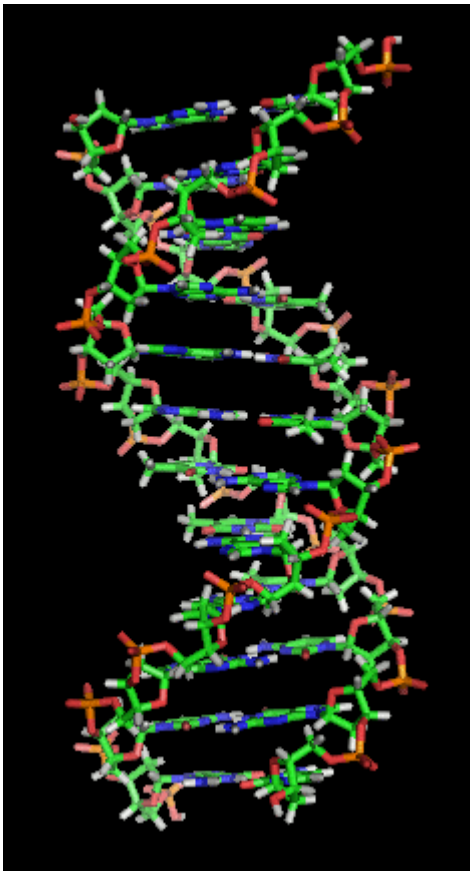


Chemical structure of DNA. Hydrogen bonds shown as dotted lines

DNA is a long polymer made from repeating units called nucleotides. As first discovered by James D. Watson and Francis Crick, the structure of DNA of all species comprises two helical chains each coiled round the same axis, and each with a pitch of 34 Ångströms (3.4 nanometres) and a radius of 10 Ångströms (1.0 nanometres). According to another study, when measured in a particular solution, the DNA chain measured 22 to 26 Ångströms wide (2.2 to 2.6 nanometres), and one nucleotide unit measured 3.3 Å (0.33 nm) long. Although each individual repeating unit is very small, DNA polymers can be very large molecules containing millions of nucleotides. For instance, the largest human chromosome, chromosome number 1, is approximately 220 million base pairs long.

In living organisms, DNA does not usually exist as a single molecule, but instead as a pair of molecules that are held tightly together. These two long strands entwine like vines, in the shape of a double helix. The nucleotide repeats contain both the segment of the backbone of the molecule, which holds the chain together, and a base, which interacts with the other DNA strand in the helix. A base linked to a sugar is called a nucleoside and a base linked to a sugar and one or more phosphate groups is called a nucleotide. If multiple nucleotides are linked together, as in DNA, this polymer is called a polynucleotide.

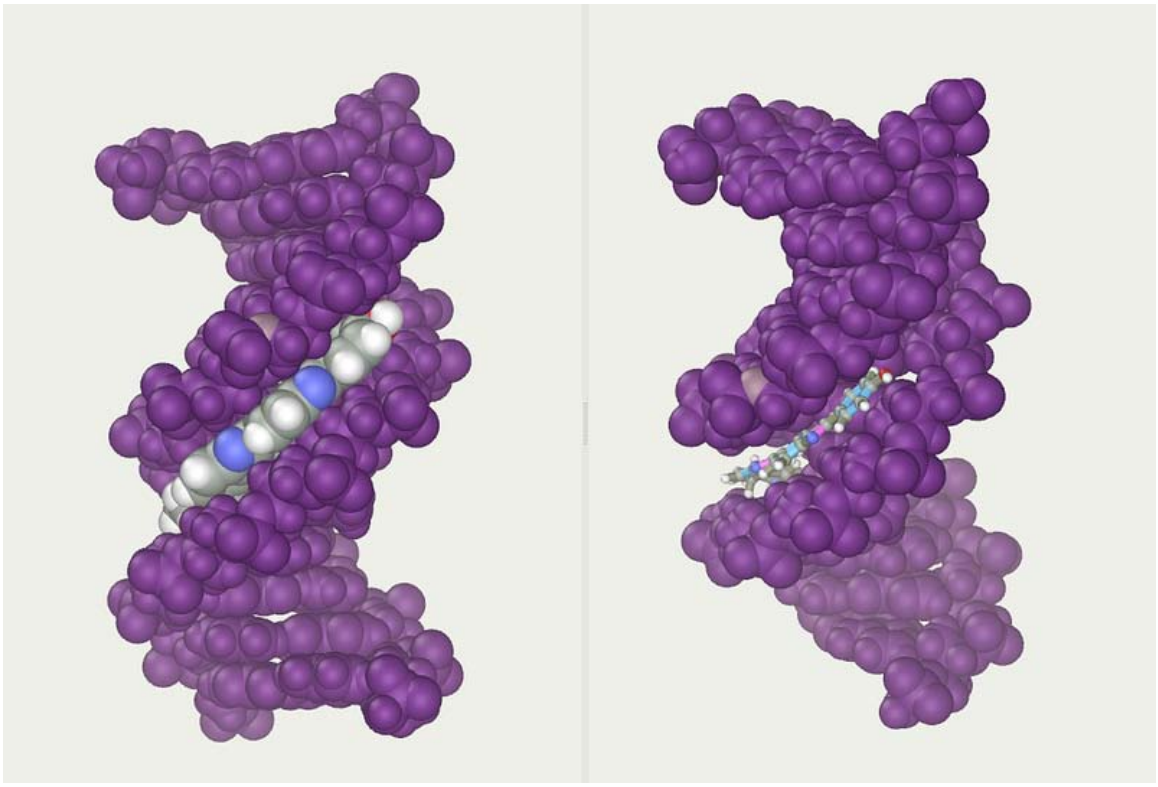
The backbone of the DNA strand is made from alternating phosphate and sugar residues. The sugar in DNA is 2-deoxyribose, which is a pentose (five-carbon) sugar. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These asymmetric bonds mean a strand of DNA has a direction. In a double helix the direction of the nucleotides in one strand is opposite to their direction in the other strand: the strands are *antiparallel*. The asymmetric ends of DNA strands are called the 5′ (*five prime*) and 3′ (*three prime*) ends, with the 5' end having a terminal phosphate group and the 3' end a terminal hydroxyl group. One major difference between DNA and RNA is the sugar, with the 2-deoxyribose in DNA being replaced by the alternative pentose sugar ribose in RNA.



A section of DNA. The bases lie horizontally between the two spiraling strands.

The DNA double helix is stabilized by hydrogen bonds between the bases attached to the two strands. The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). These four bases are attached to the sugar/phosphate to form the complete nucleotide, as shown for adenosine monophosphate.

These bases are classified into two types; adenine and guanine are fused five- and six-membered heterocyclic compounds called purines, while cytosine and thymine are six-membered rings called pyrimidines. A fifth pyrimidine base, called uracil (U), usually takes the place of thymine in RNA and differs from thymine by lacking a methyl group on its ring. Uracil is not usually found in DNA, occurring only as a breakdown product of cytosine. In addition to RNA and DNA, a large number of artificial nucleic acid analogues have also been created to study the proprieties of nucleic acids, or for use in biotechnology.



Major and minor grooves of DNA. Minor groove is a binding site for the dye Hoechst 33258.

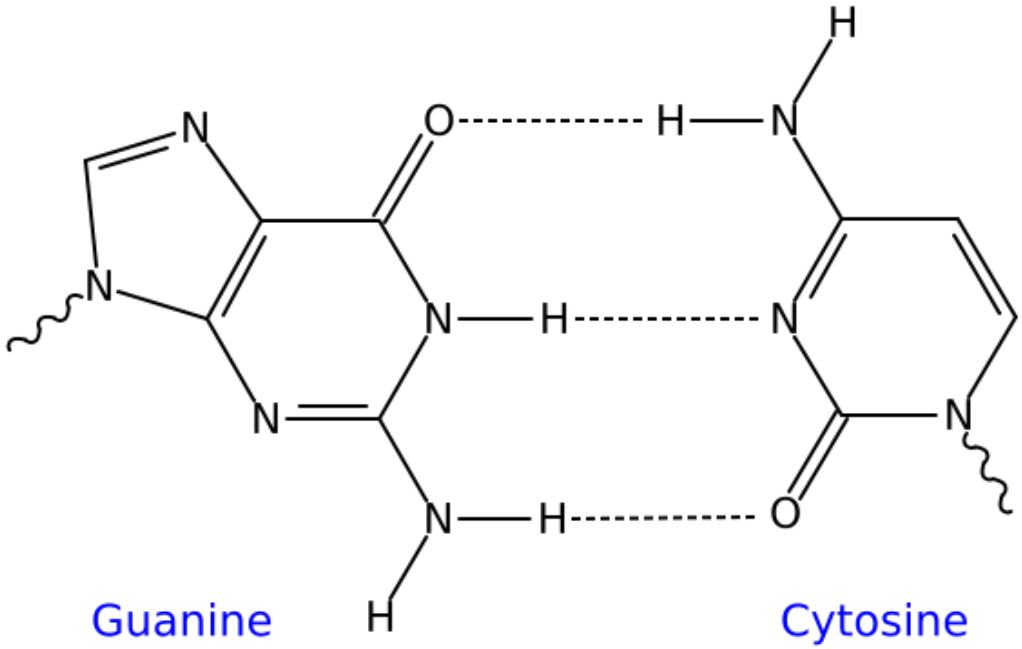
Grooves

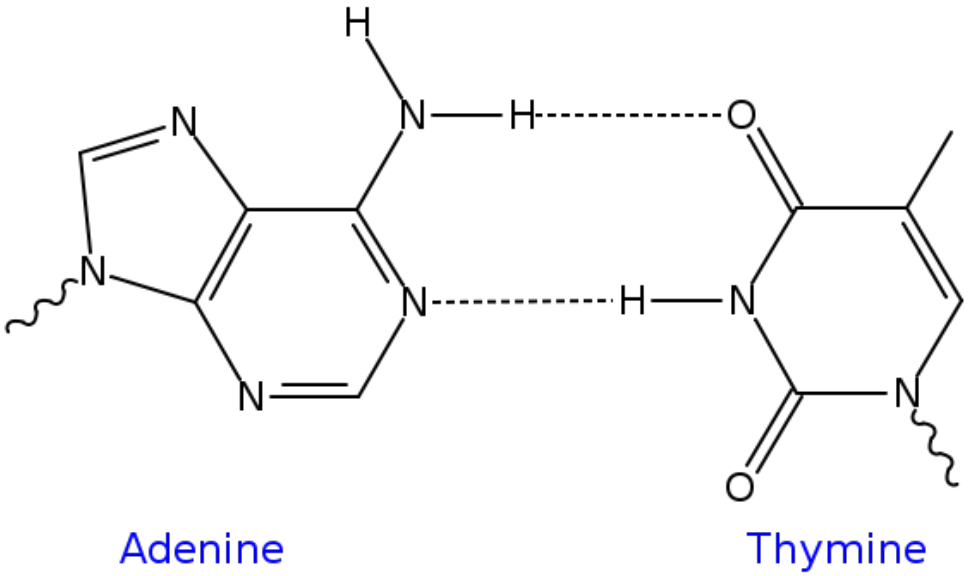
Twin helical strands form the DNA backbone. Another double helix may be found by tracing the spaces, or grooves, between the strands. These voids are adjacent to the base pairs and may provide a binding site. As the strands are not directly opposite each other, the grooves are unequally sized. One groove, the major groove, is 22 Å wide and the other, the minor groove, is 12 Å wide. The narrowness of the minor groove means that the edges of the bases are more accessible in the major groove. As a result, proteins like transcription factors that can bind to specific sequences in double-stranded DNA usually make contacts to the sides of the bases exposed in the major groove. This situation varies in unusual conformations of DNA within the cell *(see below)*, but the major and minor grooves are always named to reflect the differences in size that would be seen if the DNA is twisted back into the ordinary B form.

Base pairing

Each type of base on one strand forms a bond with just one type of base on the other strand. This is called complementary base pairing. Here, purines form hydrogen bonds to pyrimidines, with A bonding only to T, and C bonding only to G. This arrangement of two nucleotides binding together across the double helix is called a base pair. As hydrogen bonds are not covalent, they can be broken and rejoined relatively easily. The

two strands of DNA in a double helix can therefore be pulled apart like a zipper, either by a mechanical force or high temperature. As a result of this complementarity, all the information in the double-stranded sequence of a DNA helix is duplicated on each strand, which is vital in DNA replication. Indeed, this reversible and specific interaction between complementary base pairs is critical for all the functions of DNA in living organisms.





Top, a **GC** base pair with three hydrogen bonds. Bottom, an **AT** base pair with two hydrogen bonds. Non-covalent hydrogen bonds between the pairs are shown as dashed lines.

The two types of base pairs form different numbers of hydrogen bonds, AT forming two hydrogen bonds, and GC forming three hydrogen bonds (see figures, left). DNA with high GC-content is more stable than DNA with low GC-content, but contrary to popular belief, this is not due to the extra hydrogen bond of a GC base pair but rather the contribution of stacking interactions (hydrogen bonding merely provides specificity of the pairing, not stability).

As a result, it is both the percentage of GC base pairs and the overall length of a DNA double helix that determine the strength of the association between the two strands of DNA. Long DNA helices with a high GC content have stronger-interacting strands, while short helices with high AT content have weaker-interacting strands. In biology, parts of the DNA double helix that need to separate easily, such as the TATAAT Pribnow box in some promoters, tend to have a high AT content, making the strands easier to pull apart.

In the laboratory, the strength of this interaction can be measured by finding the temperature required to break the hydrogen bonds, their melting temperature (also called T_m value). When all the base pairs in a DNA double helix melt, the strands separate and exist in solution as two entirely independent molecules. These single-stranded DNA molecules (*ssDNA*) have no single common shape, but some conformations are more stable than others.

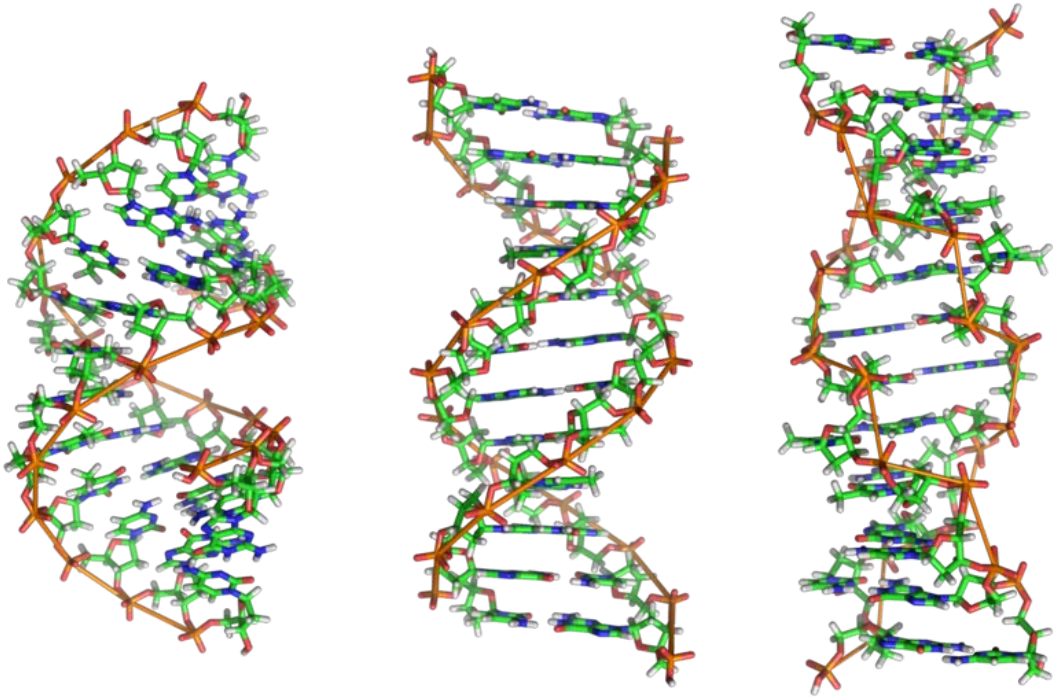
Sense and antisense

A DNA sequence is called "sense" if its sequence is the same as that of a messenger RNA copy that is translated into protein. The sequence on the opposite strand is called the "antisense" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands contain both sense and antisense sequences). In both prokaryotes and eukaryotes, antisense RNA sequences are produced, but the functions of these RNAs are not entirely clear. One proposal is that antisense RNAs are involved in regulating gene expression through RNA-RNA base pairing.

A few DNA sequences in prokaryotes and eukaryotes, and more in plasmids and viruses, blur the distinction between sense and antisense strands by having overlapping genes. In these cases, some DNA sequences do double duty, encoding one protein when read along one strand, and a second protein when read in the opposite direction along the other strand. In bacteria, this overlap may be involved in the regulation of gene transcription, while in viruses, overlapping genes increase the amount of information that can be encoded within the small viral genome.

Supercoiling

DNA can be twisted like a rope in a process called DNA supercoiling. With DNA in its "relaxed" state, a strand usually circles the axis of the double helix once every 10.4 base pairs, but if the DNA is twisted the strands become more tightly or more loosely wound. If the DNA is twisted in the direction of the helix, this is positive supercoiling, and the bases are held more tightly together. If they are twisted in the opposite direction, this is negative supercoiling, and the bases come apart more easily. In nature, most DNA has slight negative supercoiling that is introduced by enzymes called topoisomerases. These enzymes are also needed to relieve the twisting stresses introduced into DNA strands during processes such as transcription and DNA replication.



From left to right, the structures of A, B and Z DNA

Alternate DNA structures

DNA exists in many possible conformations that include A-DNA, B-DNA, and Z-DNA forms, although, only B-DNA and Z-DNA have been directly observed in functional organisms. The conformation that DNA adopts depends on the hydration level, DNA sequence, the amount and direction of supercoiling, chemical modifications of the bases, the type and concentration of metal ions, as well as the presence of polyamines in solution.

The first published reports of A-DNA X-ray diffraction patterns— and also B-DNA used analyses based on Patterson transforms that provided only a limited amount of structural information for oriented fibers of DNA. An alternate analysis was then proposed by

Wilkins *et al.*, in 1953, for the *in vivo* B-DNA X-ray diffraction/scattering patterns of highly hydrated DNA fibers in terms of squares of Bessel functions. In the same journal, James D. Watson and Francis Crick presented their molecular modeling analysis of the DNA X-ray diffraction patterns to suggest that the structure was a double-helix.

Although the `B-DNA form' is most common under the conditions found in cells, it is not a well-defined conformation but a family of related DNA conformations that occur at the high hydration levels present in living cells. Their corresponding X-ray diffraction and scattering patterns are characteristic of molecular paracrystals with a significant degree of disorder.

Compared to B-DNA, the A-DNA form is a wider right-handed spiral, with a shallow, wide minor groove and a narrower, deeper major groove. The A form occurs under non-physiological conditions in partially dehydrated samples of DNA, while in the cell it may be produced in hybrid pairings of DNA and RNA strands, as well as in enzyme-DNA complexes. Segments of DNA where the bases have been chemically modified by methylation may undergo a larger change in conformation and adopt the Z form. Here, the strands turn about the helical axis in a left-handed spiral, the opposite of the more common B form. These unusual structures can be recognized by specific Z-DNA binding proteins and may be involved in the regulation of transcription.

Alternate DNA chemistry

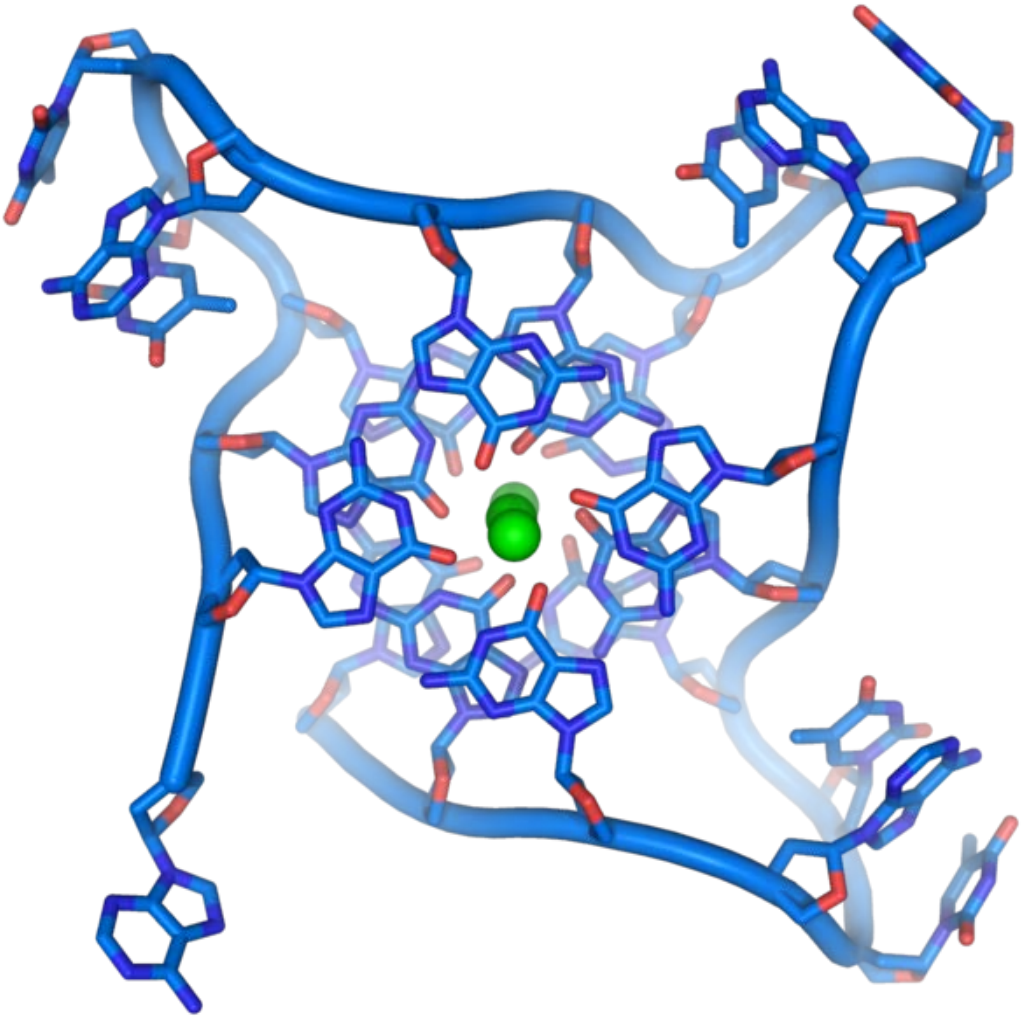
For a number of years exobiologists have proposed the existence of a shadow biosphere, a postulated microbial biosphere of Earth that uses radically different biochemical and molecular processes than currently known life. One of the proposals was the existence of lifeforms that use arsenic instead of phosphorus in DNA.

A December 2010 NASA press conference revealed that the bacterium GFAJ-1, which has evolved in an arsenic-rich environment, is the first terrestrial lifeform found which may have this ability. The bacterium was found in Mono Lake, east of Yosemite National Park. GFAJ-1 is a rod-shaped extremophile bacterium in the family Halomonadaceae that, when starved of phosphorus, may be capable of incorporating the usually poisonous element arsenic in its DNA. This discovery lends weight to the long-standing idea that extraterrestrial life could have a different chemical makeup from life on Earth. The research was carried out by a team led by Felisa Wolfe-Simon, a geomicrobiologist and geobiochemist, a Postdoctoral Fellow of the NASA Astrobiology Institute with Arizona State University.

Quadruplex structures

At the ends of the linear chromosomes are specialized regions of DNA called telomeres. The main function of these regions is to allow the cell to replicate chromosome ends using the enzyme telomerase, as the enzymes that normally replicate DNA cannot copy the extreme 3′ ends of chromosomes. These specialized chromosome caps also help protect the DNA ends, and stop the DNA repair systems in the cell from treating them as

damage to be corrected. In human cells, telomeres are usually lengths of single-stranded DNA containing several thousand repeats of a simple TTAGGG sequence.

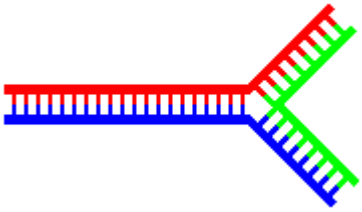


DNA quadruplex formed by telomere repeats. The looped conformation of the DNA backbone is very different from the typical DNA helix.

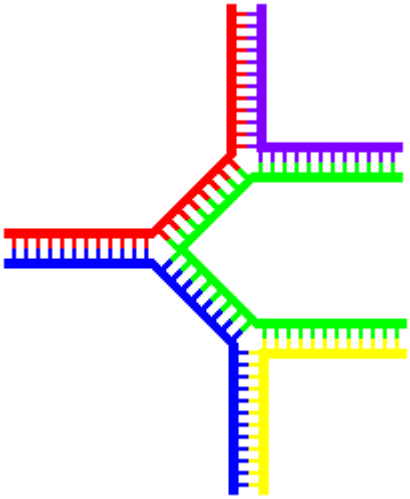
These guanine-rich sequences may stabilize chromosome ends by forming structures of stacked sets of four-base units, rather than the usual base pairs found in other DNA molecules. Here, four guanine bases form a flat plate and these flat four-base units then stack on top of each other, to form a stable *G-quadruplex* structure. These structures are stabilized by hydrogen bonding between the edges of the bases and chelation of a metal ion in the centre of each four-base unit. Other structures can also be formed, with the central set of four bases coming from either a single strand folded around the bases, or several different parallel strands, each contributing one base to the central structure.

In addition to these stacked structures, telomeres also form large loop structures called telomere loops, or T-loops. Here, the single-stranded DNA curls around in a long circle

stabilized by telomere-binding proteins. At the very end of the T-loop, the single-stranded telomere DNA is held onto a region of double-stranded DNA by the telomere strand disrupting the double-helical DNA and base pairing to one of the two strands. This triple-stranded structure is called a displacement loop or D-loop.



Single branch



Multiple branches

Branched DNA can form networks containing multiple branches.

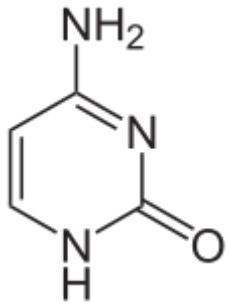
Branched DNA

In DNA fraying occurs when non-complementary regions exist at the end of an otherwise complementary double-strand of DNA. However, branched DNA can occur if a third strand of DNA is introduced and contains adjoining regions able to hybridize with the frayed regions of the pre-existing double-strand. Although the simplest example of branched DNA involves only three strands of DNA, complexes involving additional strands and multiple branches are also possible. Branched DNA can be used in nanotechnology to construct geometric shapes, see the section on uses in technology below.

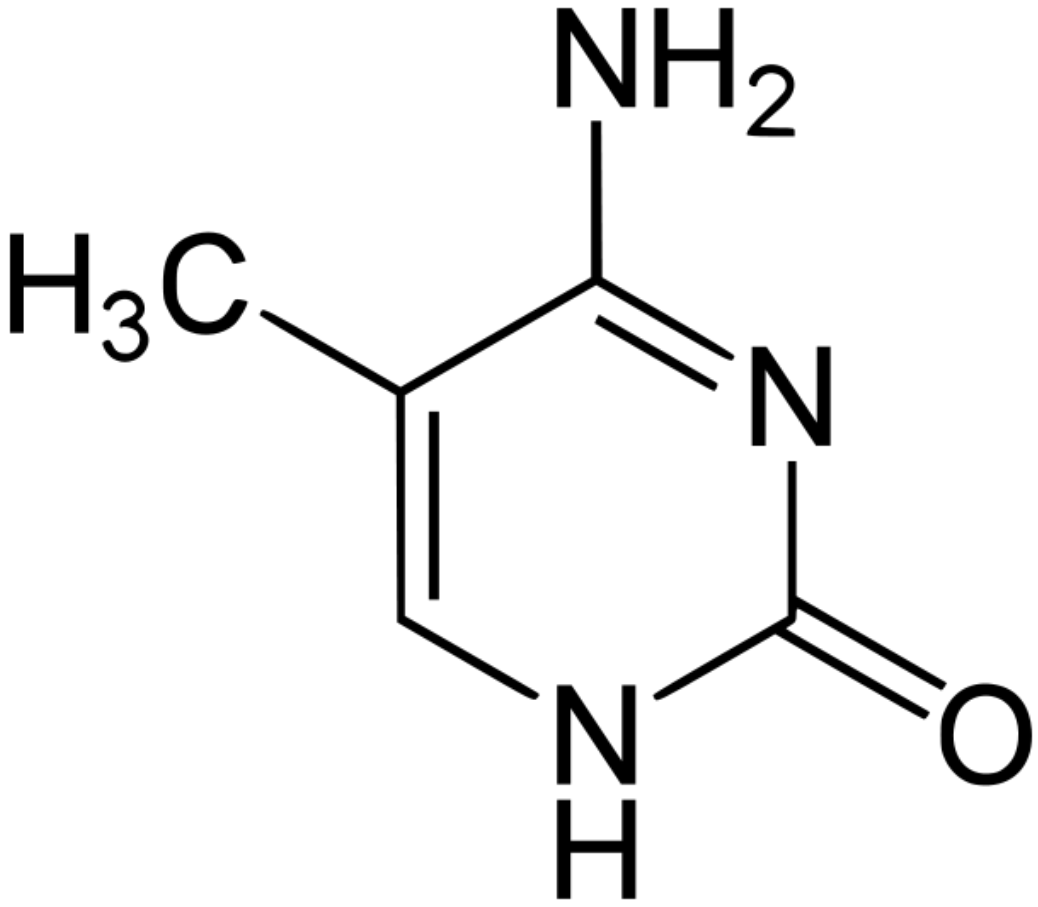
Vibration

DNA may carry out low-frequency collective motion as observed by the Raman spectroscopy and analyzed with a quasi-continuum model.

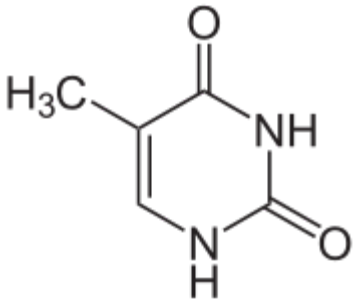
Chemical modifications



cytosine



5-methylcytosine



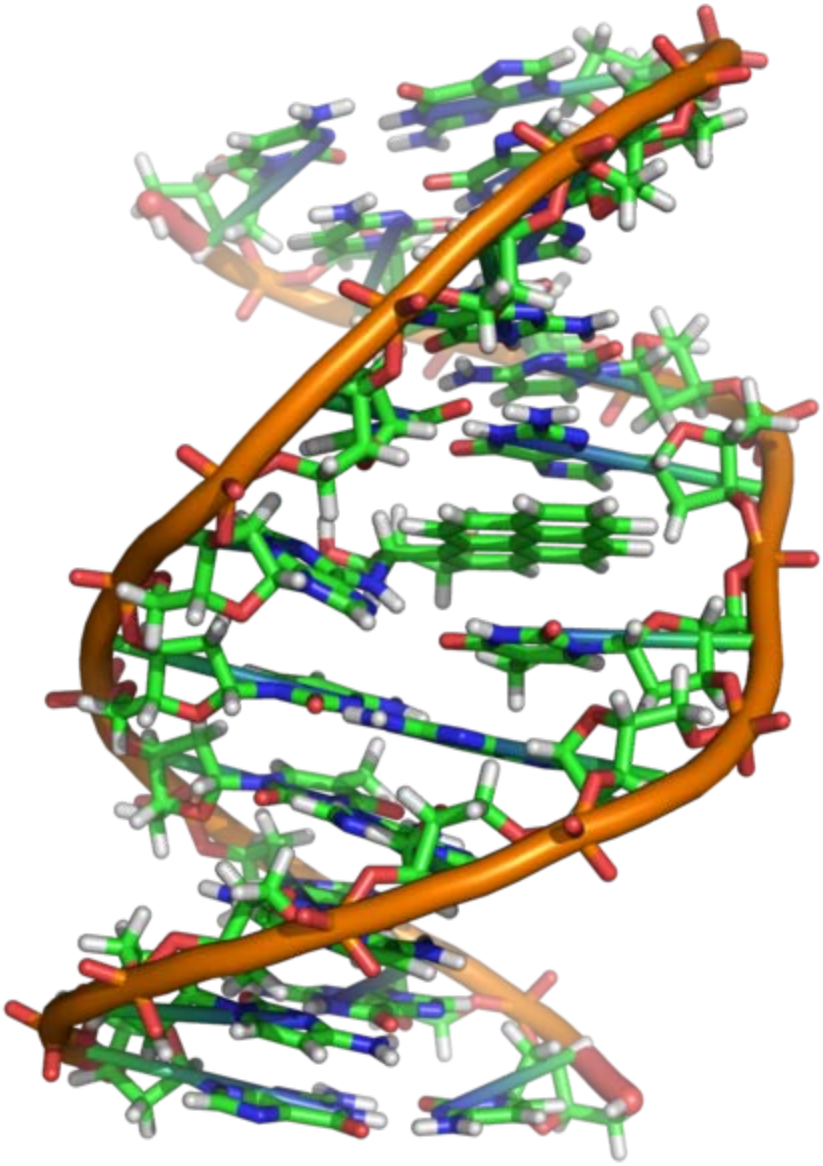
thymine

Structure of cytosine with and without the 5-methyl group. Deamination converts 5-methylcytosine into thymine.

Base modifications

The expression of genes is influenced by how the DNA is packaged in chromosomes, in a structure called chromatin. Base modifications can be involved in packaging, with regions that have low or no gene expression usually containing high levels of methylation of cytosine bases. For example, cytosine methylation, produces 5-methylcytosine, which is important for X-chromosome inactivation. The average level of methylation varies between organisms - the worm *Caenorhabditis elegans* lacks cytosine methylation, while vertebrates have higher levels, with up to 1% of their DNA containing 5-methylcytosine. Despite the importance of 5-methylcytosine, it can deaminate to leave a thymine base, so methylated cytosines are particularly prone to mutations. Other base modifications include adenine methylation in bacteria, the presence of 5-hydroxymethylcytosine in the brain, and the glycosylation of uracil to produce the "J-base" in kinetoplastids.

Damage



A covalent adduct between a metabolically activated form of benzo[*a*]pyrene, the major mutagen in tobacco smoke, and DNA

DNA can be damaged by many sorts of mutagens, which change the DNA sequence. Mutagens include oxidizing agents, alkylating agents and also high-energy electromagnetic radiation such as ultraviolet light and X-rays. The type of DNA damage produced depends on the type of mutagen. For example, UV light can damage DNA by producing thymine dimers, which are cross-links between pyrimidine bases. On the other hand, oxidants such as free radicals or hydrogen peroxide produce multiple forms of damage, including base modifications, particularly of guanosine, and double-strand breaks. A typical human cell contains about 150,000 bases that have suffered oxidative damage. Of these oxidative lesions, the most dangerous are double-strand breaks, as these

are difficult to repair and can produce point mutations, insertions and deletions from the DNA sequence, as well as chromosomal translocations.

Many mutagens fit into the space between two adjacent base pairs, this is called *intercalation*. Most intercalators are aromatic and planar molecules; examples include ethidium bromide, daunomycin, and doxorubicin. In order for an intercalator to fit between base pairs, the bases must separate, distorting the DNA strands by unwinding of the double helix. This inhibits both transcription and DNA replication, causing toxicity and mutations. As a result, DNA intercalators are often carcinogens, and benzo[*a*]pyrene diol epoxide, acridines, aflatoxin and ethidium bromide are well-known examples. Nevertheless, due to their ability to inhibit DNA transcription and replication, other similar toxins are also used in chemotherapy to inhibit rapidly growing cancer cells.

Biological functions

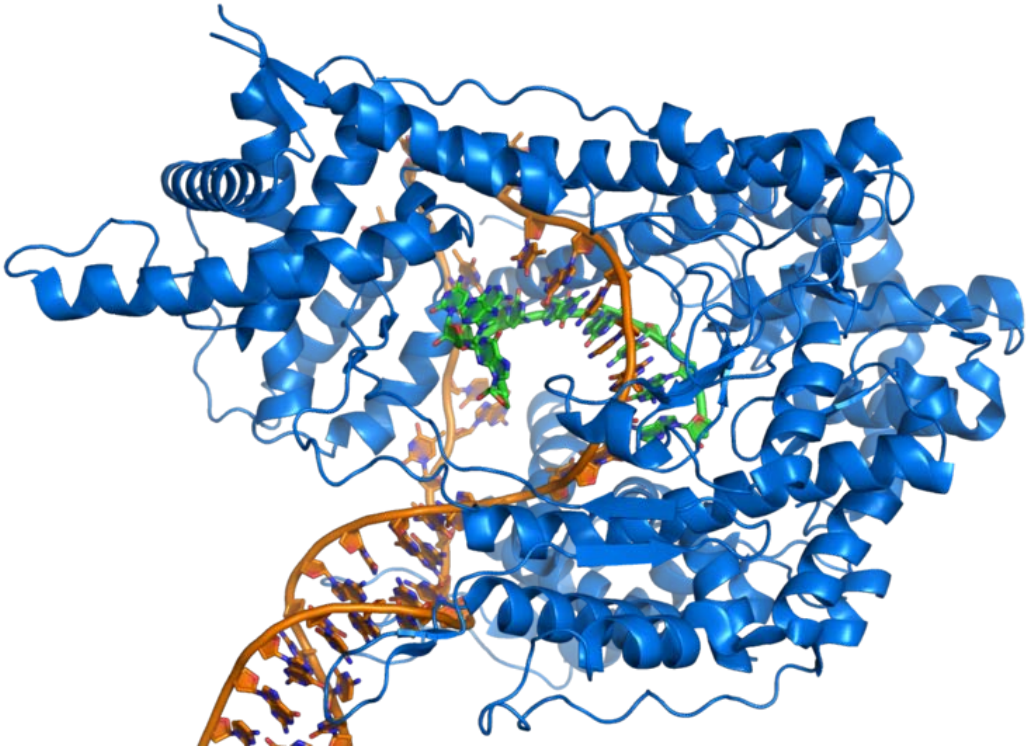
DNA usually occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The set of chromosomes in a cell makes up its genome; the human genome has approximately 3 billion base pairs of DNA arranged into 46 chromosomes. The information carried by DNA is held in the sequence of pieces of DNA called genes. Transmission of genetic information in genes is achieved via complementary base pairing. For example, in transcription, when a cell uses the information in a gene, the DNA sequence is copied into a complementary RNA sequence through the attraction between the DNA and the correct RNA nucleotides. Usually, this RNA copy is then used to make a matching protein sequence in a process called translation, which depends on the same interaction between RNA nucleotides. In alternative fashion, a cell may simply copy its genetic information in a process called DNA replication. The details of these functions are covered in other articles; here we focus on the interactions between DNA and other molecules that mediate the function of the genome.

Genes and genomes

Genomic DNA is tightly and orderly packed in the process called DNA condensation to fit the small available volumes of the cell. In eukaryotes, DNA is located in the cell nucleus, as well as small amounts in mitochondria and chloroplasts. In prokaryotes, the DNA is held within an irregularly shaped body in the cytoplasm called the nucleoid. The genetic information in a genome is held within genes, and the complete set of this information in an organism is called its genotype. A gene is a unit of heredity and is a region of DNA that influences a particular characteristic in an organism. Genes contain an open reading frame that can be transcribed, as well as regulatory sequences such as promoters and enhancers, which control the transcription of the open reading frame.

In many species, only a small fraction of the total sequence of the genome encodes protein. For example, only about 1.5% of the human genome consists of protein-coding exons, with over 50% of human DNA consisting of non-coding repetitive sequences. The reasons for the presence of so much noncoding DNA in eukaryotic genomes and the extraordinary differences in genome size, or *C-value*, among species represent a long-

standing puzzle known as the "C-value enigma". However, DNA sequences that do not code protein may still encode functional non-coding RNA molecules, which are involved in the regulation of gene expression.



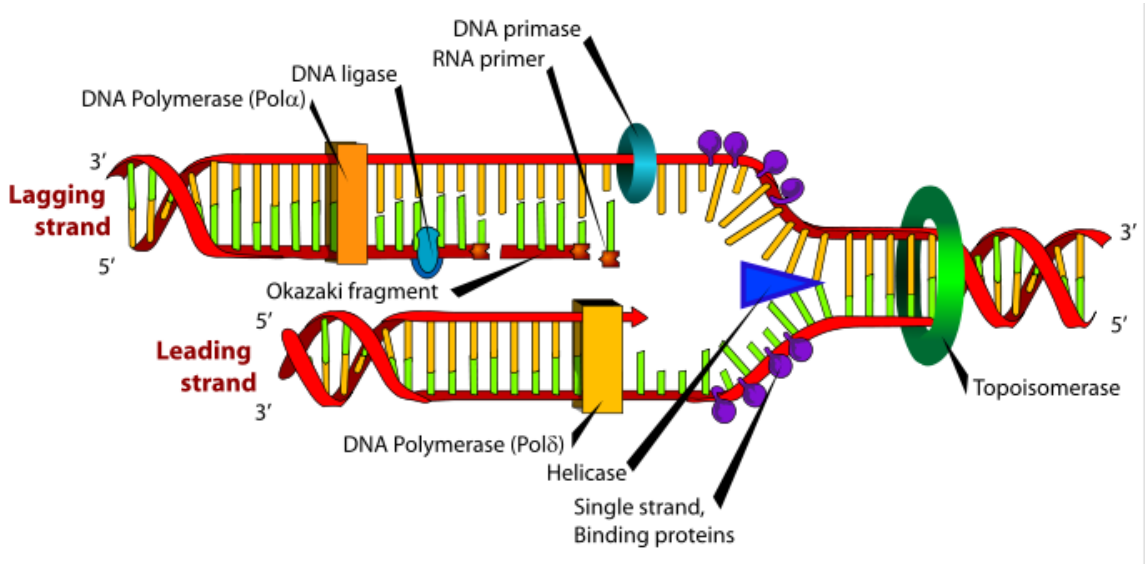
T7 RNA polymerase (blue) producing a mRNA (green) from a DNA template (orange).

Some noncoding DNA sequences play structural roles in chromosomes. Telomeres and centromeres typically contain few genes, but are important for the function and stability of chromosomes. An abundant form of noncoding DNA in humans are pseudogenes, which are copies of genes that have been disabled by mutation. These sequences are usually just molecular fossils, although they can occasionally serve as raw genetic material for the creation of new genes through the process of gene duplication and divergence.

Transcription and translation

A gene is a sequence of DNA that contains genetic information and can influence the phenotype of an organism. Within a gene, the sequence of bases along a DNA strand defines a messenger RNA sequence, which then defines one or more protein sequences. The relationship between the nucleotide sequences of genes and the amino-acid sequences of proteins is determined by the rules of translation, known collectively as the genetic code. The genetic code consists of three-letter 'words' called *codons* formed from a sequence of three nucleotides (e.g. ACT, CAG, TTT).

In transcription, the codons of a gene are copied into messenger RNA by RNA polymerase. This RNA copy is then decoded by a ribosome that reads the RNA sequence by base-pairing the messenger RNA to transfer RNA, which carries amino acids. Since there are 4 bases in 3-letter combinations, there are 64 possible codons (4^3 combinations). These encode the twenty standard amino acids, giving most amino acids more than one possible codon. There are also three 'stop' or 'nonsense' codons signifying the end of the coding region; these are the TAA, TGA and TAG codons.



DNA replication. The double helix is unwound by a helicase and topoisomerase. Next, one DNA polymerase produces the leading strand copy. Another DNA polymerase binds to the lagging strand. This enzyme makes discontinuous segments (called Okazaki fragments) before DNA ligase joins them together.

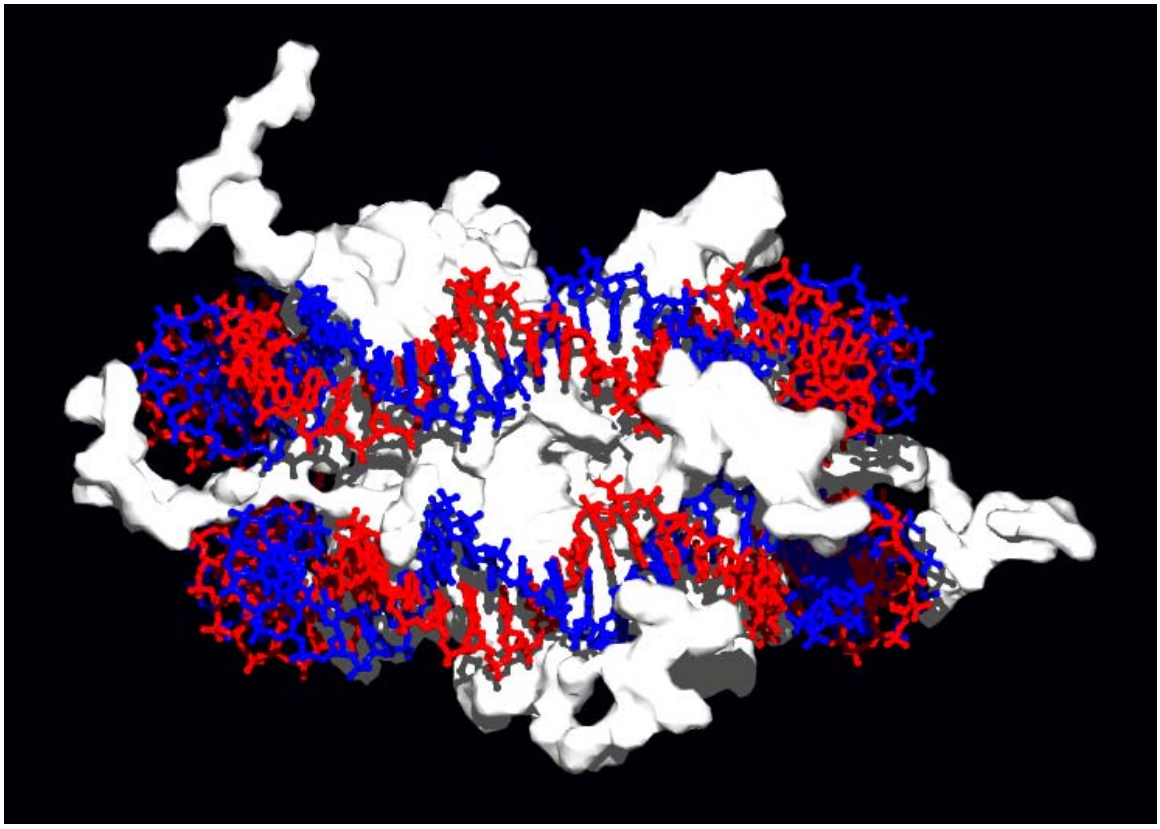
Replication

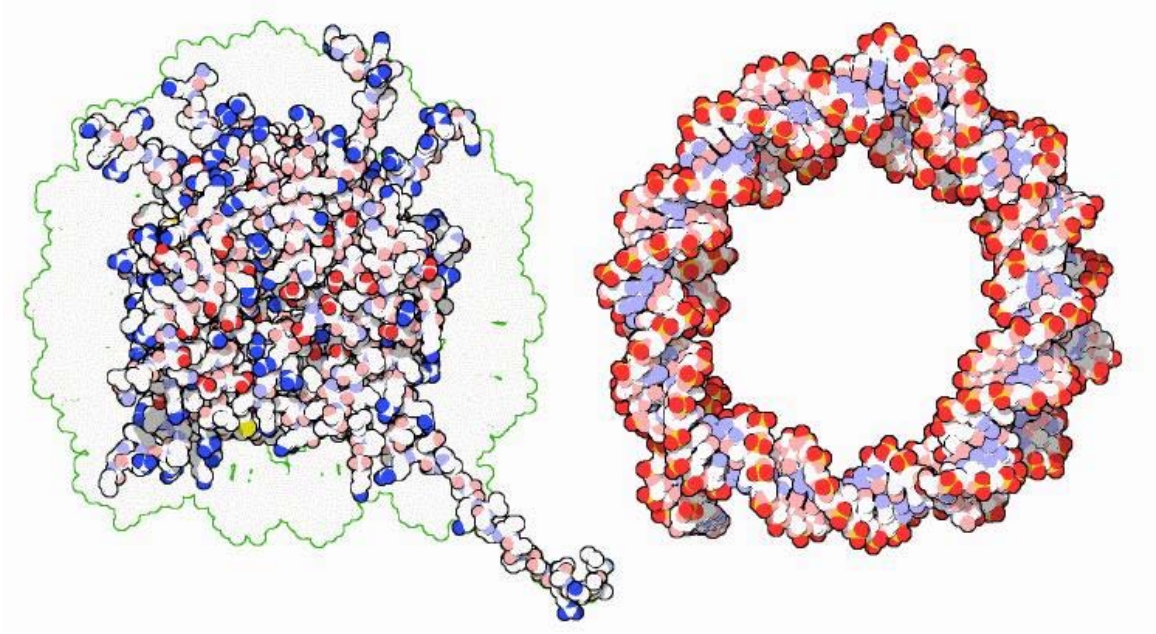
Cell division is essential for an organism to grow, but, when a cell divides, it must replicate the DNA in its genome so that the two daughter cells have the same genetic information as their parent. The double-stranded structure of DNA provides a simple mechanism for DNA replication. Here, the two strands are separated and then each strand's complementary DNA sequence is recreated by an enzyme called DNA polymerase. This enzyme makes the complementary strand by finding the correct base through complementary base pairing, and bonding it onto the original strand. As DNA polymerases can only extend a DNA strand in a 5′ to 3′ direction, different mechanisms are used to copy the antiparallel strands of the double helix. In this way, the base on the old strand dictates which base appears on the new strand, and the cell ends up with a perfect copy of its DNA.

Interactions with proteins

All the functions of DNA depend on interactions with proteins. These protein interactions can be non-specific, or the protein can bind specifically to a single DNA sequence. Enzymes can also bind to DNA and of these, the polymerases that copy the DNA base sequence in transcription and DNA replication are particularly important.

DNA-binding proteins

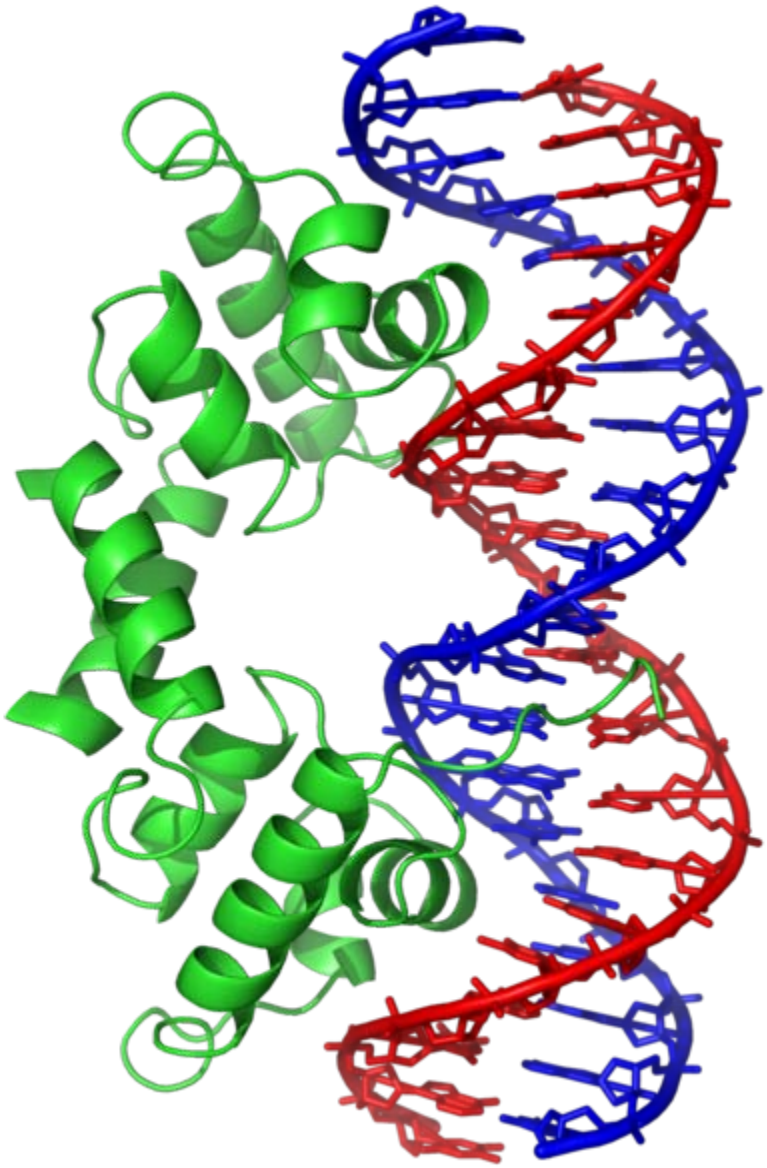




Interaction of DNA with histones (shown in white, top). These proteins' basic amino acids (below left, blue) bind to the acidic phosphate groups on DNA (below right, red).

Structural proteins that bind DNA are well-understood examples of non-specific DNA-protein interactions. Within chromosomes, DNA is held in complexes with structural proteins. These proteins organize the DNA into a compact structure called chromatin. In eukaryotes this structure involves DNA binding to a complex of small basic proteins called histones, while in prokaryotes multiple types of proteins are involved. The histones form a disk-shaped complex called a nucleosome, which contains two complete turns of double-stranded DNA wrapped around its surface. These non-specific interactions are formed through basic residues in the histones making ionic bonds to the acidic sugar-phosphate backbone of the DNA, and are therefore largely independent of the base sequence. Chemical modifications of these basic amino acid residues include methylation, phosphorylation and acetylation. These chemical changes alter the strength of the interaction between the DNA and the histones, making the DNA more or less accessible to transcription factors and changing the rate of transcription. Other non-specific DNA-binding proteins in chromatin include the high-mobility group proteins, which bind to bent or distorted DNA. These proteins are important in bending arrays of nucleosomes and arranging them into the larger structures that make up chromosomes.

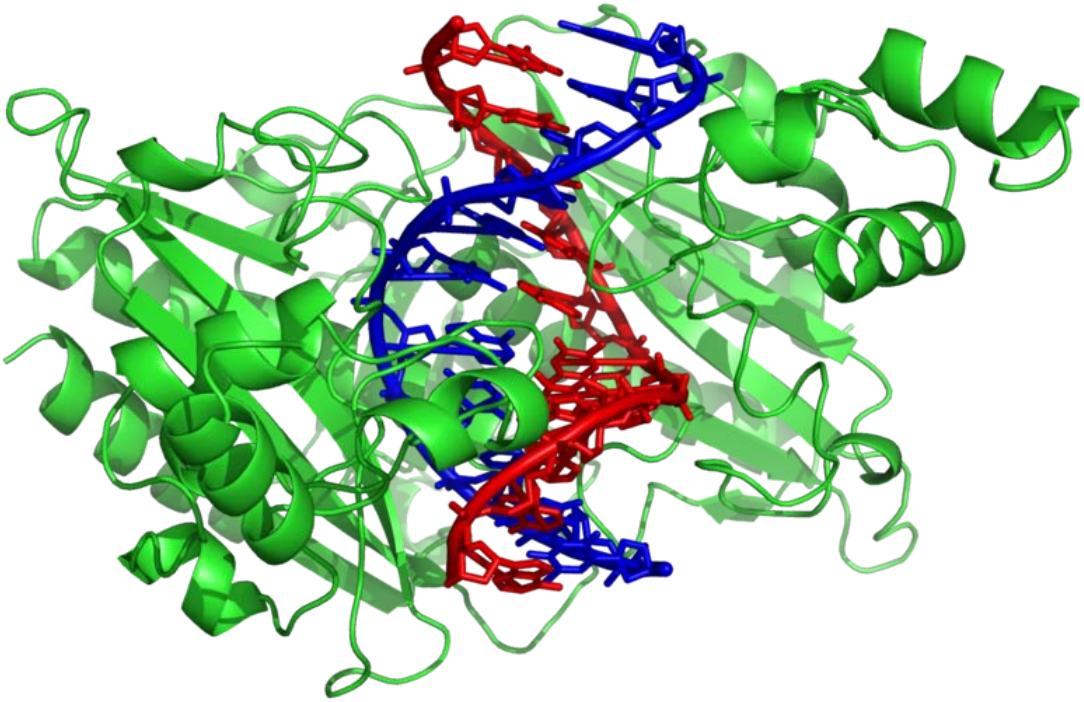
A distinct group of DNA-binding proteins are the DNA-binding proteins that specifically bind single-stranded DNA. In humans, replication protein A is the best-understood member of this family and is used in processes where the double helix is separated, including DNA replication, recombination and DNA repair. These binding proteins seem to stabilize single-stranded DNA and protect it from forming stem-loops or being degraded by nucleases.



The lambda repressor helix-turn-helix transcription factor bound to its DNA target

In contrast, other proteins have evolved to bind to particular DNA sequences. The most intensively studied of these are the various transcription factors, which are proteins that regulate transcription. Each transcription factor binds to one particular set of DNA sequences and activates or inhibits the transcription of genes that have these sequences close to their promoters. The transcription factors do this in two ways. Firstly, they can bind the RNA polymerase responsible for transcription, either directly or through other mediator proteins; this locates the polymerase at the promoter and allows it to begin transcription. Alternatively, transcription factors can bind enzymes that modify the histones at the promoter; this will change the accessibility of the DNA template to the polymerase.

As these DNA targets can occur throughout an organism's genome, changes in the activity of one type of transcription factor can affect thousands of genes. Consequently, these proteins are often the targets of the signal transduction processes that control responses to environmental changes or cellular differentiation and development. The specificity of these transcription factors' interactions with DNA come from the proteins making multiple contacts to the edges of the DNA bases, allowing them to "read" the DNA sequence. Most of these base-interactions are made in the major groove, where the bases are most accessible.



The restriction enzyme EcoRV (green) in a complex with its substrate DNA

DNA-modifying enzymes

Nucleases and ligases

Nucleases are enzymes that cut DNA strands by catalyzing the hydrolysis of the phosphodiester bonds. Nucleases that hydrolyse nucleotides from the ends of DNA strands are called exonucleases, while endonucleases cut within strands. The most frequently used nucleases in molecular biology are the restriction endonucleases, which cut DNA at specific sequences. For instance, the EcoRV enzyme shown to the left recognizes the 6-base sequence 5′-GAT|ATC-3′ and makes a cut at the vertical line. In nature, these enzymes protect bacteria against phage infection by digesting the phage DNA when it enters the bacterial cell, acting as part of the restriction modification

system. In technology, these sequence-specific nucleases are used in molecular cloning and DNA fingerprinting.

Enzymes called DNA ligases can rejoin cut or broken DNA strands. Ligases are particularly important in lagging strand DNA replication, as they join together the short segments of DNA produced at the replication fork into a complete copy of the DNA template. They are also used in DNA repair and genetic recombination.

Topoisomerases and helicases

Topoisomerases are enzymes with both nuclease and ligase activity. These proteins change the amount of supercoiling in DNA. Some of these enzymes work by cutting the DNA helix and allowing one section to rotate, thereby reducing its level of supercoiling; the enzyme then seals the DNA break. Other types of these enzymes are capable of cutting one DNA helix and then passing a second strand of DNA through this break, before rejoining the helix. Topoisomerases are required for many processes involving DNA, such as DNA replication and transcription.

Helicases are proteins that are a type of molecular motor. They use the chemical energy in nucleoside triphosphates, predominantly ATP, to break hydrogen bonds between bases and unwind the DNA double helix into single strands. These enzymes are essential for most processes where enzymes need to access the DNA bases.

Polymerases

Polymerases are enzymes that synthesize polynucleotide chains from nucleoside triphosphates. The sequence of their products are copies of existing polynucleotide chains - which are called *templates*. These enzymes function by adding nucleotides onto the 3′ hydroxyl group of the previous nucleotide in a DNA strand. As a consequence, all polymerases work in a 5′ to 3′ direction. In the active site of these enzymes, the incoming nucleoside triphosphate base-pairs to the template: this allows polymerases to accurately synthesize the complementary strand of their template. Polymerases are classified according to the type of template that they use.

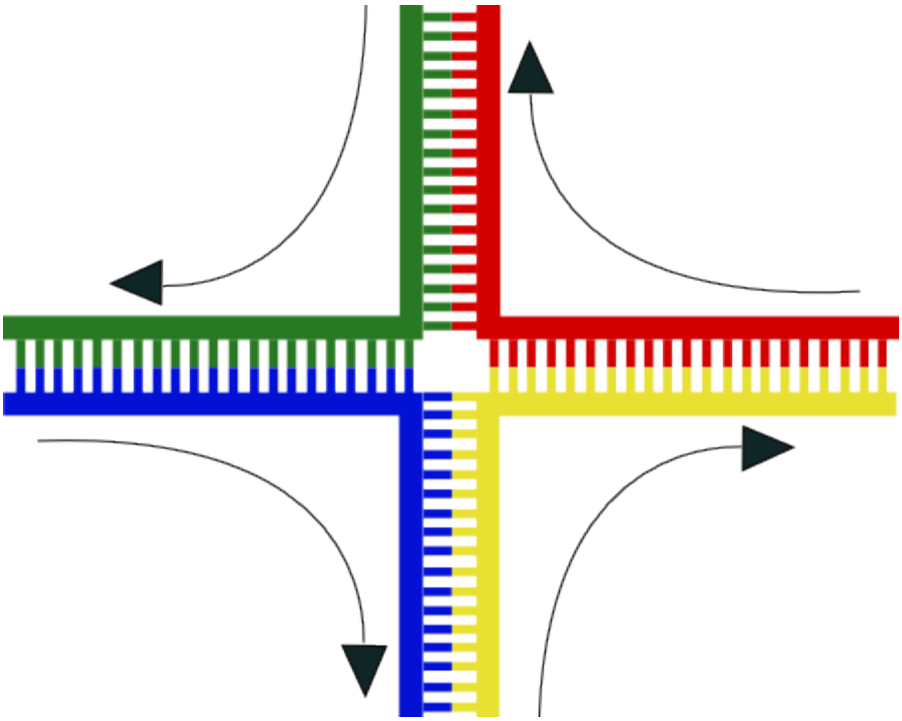
In DNA replication, a DNA-dependent DNA polymerase makes a copy of a DNA sequence. Accuracy is vital in this process, so many of these polymerases have a proofreading activity. Here, the polymerase recognizes the occasional mistakes in the synthesis reaction by the lack of base pairing between the mismatched nucleotides. If a mismatch is detected, a 3′ to 5′ exonuclease activity is activated and the incorrect base removed. In most organisms, DNA polymerases function in a large complex called the replisome that contains multiple accessory subunits, such as the DNA clamp or helicases.

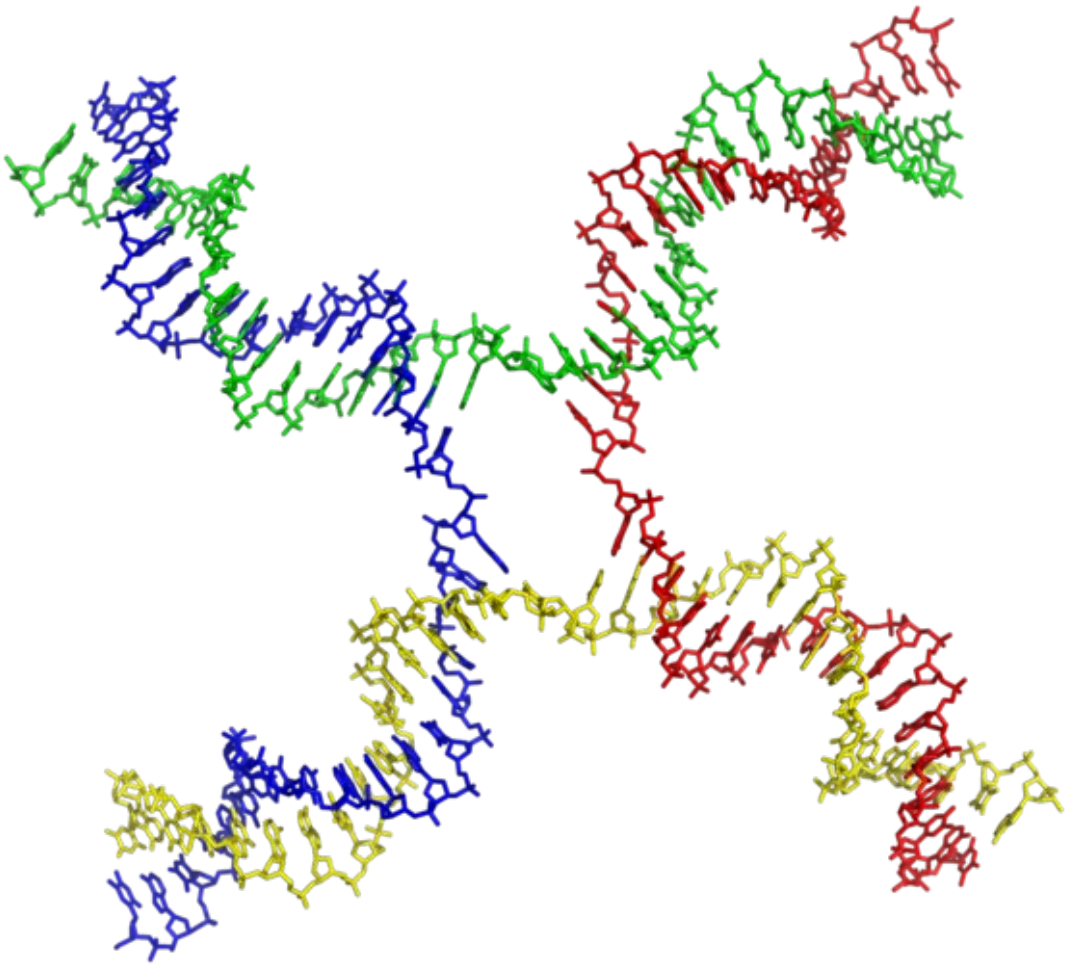
RNA-dependent DNA polymerases are a specialized class of polymerases that copy the sequence of an RNA strand into DNA. They include reverse transcriptase, which is a viral enzyme involved in the infection of cells by retroviruses, and telomerase, which is

required for the replication of telomeres. Telomerase is an unusual polymerase because it contains its own RNA template as part of its structure.

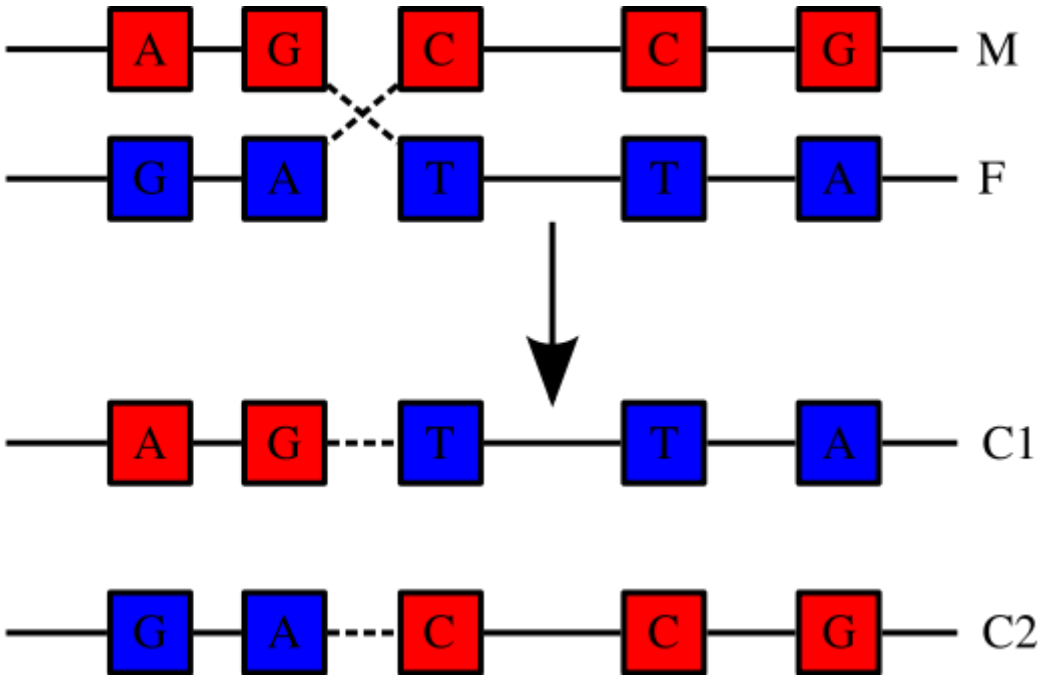
Transcription is carried out by a DNA-dependent RNA polymerase that copies the sequence of a DNA strand into RNA. To begin transcribing a gene, the RNA polymerase binds to a sequence of DNA called a promoter and separates the DNA strands. It then copies the gene sequence into a messenger RNA transcript until it reaches a region of DNA called the terminator, where it halts and detaches from the DNA. As with human DNA-dependent DNA polymerases, RNA polymerase II, the enzyme that transcribes most of the genes in the human genome, operates as part of a large protein complex with multiple regulatory and accessory subunits.

Genetic recombination





Structure of the Holliday junction intermediate in genetic recombination. The four separate DNA strands are coloured red, blue, green and yellow.



Recombination involves the breakage and rejoining of two chromosomes (M and F) to produce two re-arranged chromosomes (C1 and C2).

A DNA helix usually does not interact with other segments of DNA, and in human cells the different chromosomes even occupy separate areas in the nucleus called "chromosome territories". This physical separation of different chromosomes is important for the ability of DNA to function as a stable repository for information, as one of the few times chromosomes interact is during chromosomal crossover when they recombine. Chromosomal crossover is when two DNA helices break, swap a section and then rejoin.

Recombination allows chromosomes to exchange genetic information and produces new combinations of genes, which increases the efficiency of natural selection and can be important in the rapid evolution of new proteins. Genetic recombination can also be involved in DNA repair, particularly in the cell's response to double-strand breaks.

The most common form of chromosomal crossover is homologous recombination, where the two chromosomes involved share very similar sequences. Non-homologous recombination can be damaging to cells, as it can produce chromosomal translocations and genetic abnormalities. The recombination reaction is catalyzed by enzymes known as recombinases, such as RAD51. The first step in recombination is a double-stranded break either caused by an endonuclease or damage to the DNA. A series of steps catalyzed in part by the recombinase then leads to joining of the two helices by at least one Holliday junction, in which a segment of a single strand in each helix is annealed to the complementary strand in the other helix. The Holliday junction is a tetrahedral junction structure that can be moved along the pair of chromosomes, swapping one strand for another. The recombination reaction is then halted by cleavage of the junction and re-ligation of the released DNA.

Evolution

DNA contains the genetic information that allows all modern living things to function, grow and reproduce. However, it is unclear how long in the 4-billion-year history of life DNA has performed this function, as it has been proposed that the earliest forms of life may have used RNA as their genetic material. RNA may have acted as the central part of early cell metabolism as it can both transmit genetic information and carry out catalysis as part of ribozymes. This ancient RNA world where nucleic acid would have been used for both catalysis and genetics may have influenced the evolution of the current genetic code based on four nucleotide bases. This would occur, since the number of different bases in such an organism is a trade-off between a small number of bases increasing replication accuracy and a large number of bases increasing the catalytic efficiency of ribozymes.

However, there is no direct evidence of ancient genetic systems, as recovery of DNA from most fossils is impossible. This is because DNA will survive in the environment for less than one million years and slowly degrades into short fragments in solution. Claims for older DNA have been made, most notably a report of the isolation of a viable bacterium from a salt crystal 250 million years old, but these claims are controversial.

Uses in technology

Genetic engineering

Methods have been developed to purify DNA from organisms, such as phenol-chloroform extraction, and to manipulate it in the laboratory, such as restriction digests and the polymerase chain reaction. Modern biology and biochemistry make intensive use of these techniques in recombinant DNA technology. Recombinant DNA is a man-made DNA sequence that has been assembled from other DNA sequences. They can be transformed into organisms in the form of plasmids or in the appropriate format, by using a viral vector. The genetically modified organisms produced can be used to produce products such as recombinant proteins, used in medical research, or be grown in agriculture.

Forensics

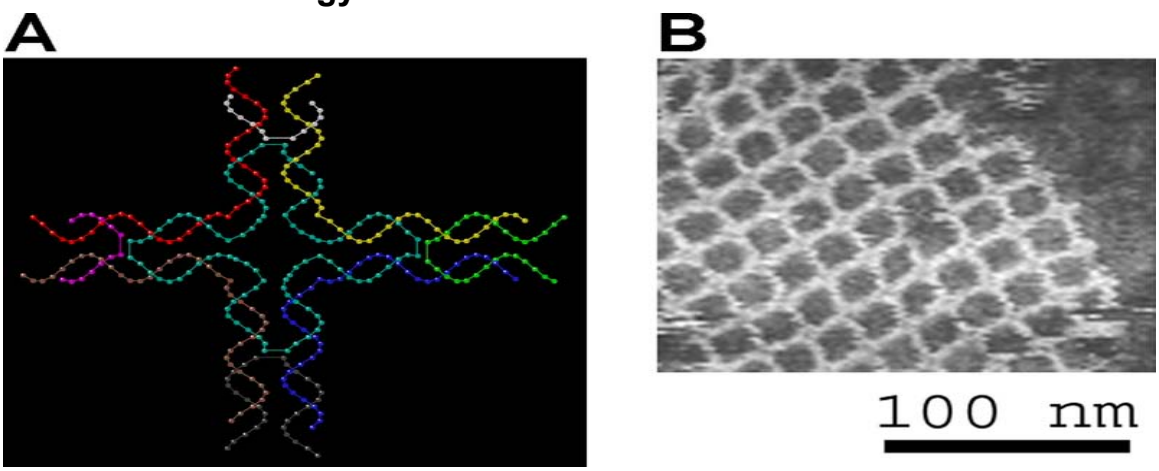
Forensic scientists can use DNA in blood, semen, skin, saliva or hair found at a crime scene to identify a matching DNA of an individual, such as a perpetrator. This process is formally termed DNA profiling, but may also be called "genetic fingerprinting". In DNA profiling, the lengths of variable sections of repetitive DNA, such as short tandem repeats and minisatellites, are compared between people. This method is usually an extremely reliable technique for identifying a matching DNA. However, identification can be complicated if the scene is contaminated with DNA from several people. DNA profiling was developed in 1984 by British geneticist Sir Alec Jeffreys, and first used in forensic science to convict Colin Pitchfork in the 1988 Enderby murders case.

People convicted of certain types of crimes may be required to provide a sample of DNA for a database. This has helped investigators solve old cases where only a DNA sample was obtained from the scene. DNA profiling can also be used to identify victims of mass casualty incidents. On the other hand, many convicted people have been released from prison on the basis of DNA techniques, which were not available when a crime had originally been committed.

Bioinformatics

Bioinformatics involves the manipulation, searching, and data mining of biological data, and this includes DNA sequence data. The development of techniques to store and search DNA sequences have led to widely applied advances in computer science, especially string searching algorithms, machine learning and database theory. String searching or matching algorithms, which find an occurrence of a sequence of letters inside a larger sequence of letters, were developed to search for specific sequences of nucleotides. The DNA sequenced may be aligned with other DNA sequences to identify homologous sequences and locate the specific mutations that make them distinct. These techniques, especially multiple sequence alignment, are used in studying phylogenetic relationships and protein function. Data sets representing entire genomes' worth of DNA sequences, such as those produced by the Human Genome Project, are difficult to use without the annotations that identify the locations of genes and regulatory elements on each chromosome. Regions of DNA sequence that have the characteristic patterns associated with protein- or RNA-coding genes can be identified by gene finding algorithms, which allow researchers to predict the presence of particular gene products and their possible functions in an organism even before they have been isolated experimentally. Entire genomes may also be compared which can shed light on the evolutionary history of particular organism and permit the examination of complex evolutionary events.

DNA nanotechnology



The DNA structure at left (schematic shown) will self-assemble into the structure visualized by atomic force microscopy at right. DNA nanotechnology is the field that seeks to design nanoscale structures using the molecular recognition properties of DNA molecules.

DNA nanotechnology uses the unique molecular recognition properties of DNA and other nucleic acids to create self-assembling branched DNA complexes with useful properties. DNA is thus used as a structural material rather than as a carrier of biological information. This has led to the creation of two-dimensional periodic lattices (both tile-based as well as using the "DNA origami" method) as well as three-dimensional structures in the shapes of polyhedra. Nanomechanical devices and algorithmic self-assembly have also been demonstrated, and these DNA structures have been used to template the arrangement of other molecules such as gold nanoparticles and streptavidin proteins.

History and anthropology

Because DNA collects mutations over time, which are then inherited, it contains historical information, and, by comparing DNA sequences, geneticists can infer the evolutionary history of organisms, their phylogeny. This field of phylogenetics is a powerful tool in evolutionary biology. If DNA sequences within a species are compared, population geneticists can learn the history of particular populations. This can be used in studies ranging from ecological genetics to anthropology; For example, DNA evidence is being used to try to identify the Ten Lost Tribes of Israel.

DNA has also been used to look at modern family relationships, such as establishing family relationships between the descendants of Sally Hemings and Thomas Jefferson. This usage is closely related to the use of DNA in criminal investigations detailed above. Indeed, some criminal investigations have been solved when DNA from crime scenes has matched relatives of the guilty individual.

History of DNA research

James D. Watson and Francis Crick (right), co-originators of the double-helix model, with Maclyn McCarty (left).

DNA was first isolated by the Swiss physician Friedrich Miescher who, in 1869, discovered a microscopic substance in the pus of discarded surgical bandages. As it resided in the nuclei of cells, he called it "nuclein". In 1919, Phoebus Levene identified the base, sugar and phosphate nucleotide unit. Levene suggested that DNA consisted of a string of nucleotide units linked together through the phosphate groups. However, Levene thought the chain was short and the bases repeated in a fixed order. In 1937 William Astbury produced the first X-ray diffraction patterns that showed that DNA had a regular structure.

Raymond Gosling, co-creator of the single X-ray diffraction image

In 1928, Frederick Griffith discovered that traits of the "smooth" form of the *Pneumococcus* could be transferred to the "rough" form of the same bacteria by mixing killed "smooth" bacteria with the live "rough" form. This system provided the first clear suggestion that DNA carries genetic information—the Avery–MacLeod–McCarty experiment—when Oswald Avery, along with coworkers Colin MacLeod and Maclyn McCarty, identified DNA as the transforming principle in 1943. DNA's role in heredity

was confirmed in 1952, when Alfred Hershey and Martha Chase in the Hershey–Chase experiment showed that DNA is the genetic material of the T2 phage.

In 1953, James D. Watson and Francis Crick suggested what is now accepted as the first correct double-helix model of DNA structure in the journal *Nature*. Their double-helix, molecular model of DNA was then based on a single X-ray diffraction image (labeled as "Photo 51") taken by Rosalind Franklin and Raymond Gosling in May 1952, as well as the information that the DNA bases are paired — also obtained through private communications from Erwin Chargaff in the previous years. Chargaff's rules played a very important role in establishing double-helix configurations for B-DNA as well as A-DNA.

Experimental evidence supporting the Watson and Crick model were published in a series of five articles in the same issue of *Nature*. Of these, Franklin and Gosling's paper was the first publication of their own X-ray diffraction data and original analysis method that partially supported the Watson and Crick model; this issue also contained an article on DNA structure by Maurice Wilkins and two of his colleagues, whose analysis and *in vivo* B-DNA X-ray patterns also supported the presence *in vivo* of the double-helical DNA configurations as proposed by Crick and Watson for their double-helix molecular model of DNA in the previous two pages of *Nature*. In 1962, after Franklin's death, Watson, Crick, and Wilkins jointly received the Nobel Prize in Physiology or Medicine. However, Nobel rules of the time allowed only living recipients, but a vigorous debate continues on who should receive credit for the discovery.

In an influential presentation in 1957, Crick laid out the central dogma of molecular biology, which foretold the relationship between DNA, RNA, and proteins, and articulated the "adaptor hypothesis". Final confirmation of the replication mechanism that was implied by the double-helical structure followed in 1958 through the Meselson–Stahl experiment. Further work by Crick and coworkers showed that the genetic code was based on non-overlapping triplets of bases, called codons, allowing Har Gobind Khorana, Robert W. Holley and Marshall Warren Nirenberg to decipher the genetic code. These findings represent the birth of molecular biology.

Chapter- 5

Chromosome

A **chromosome** is an organized structure of DNA and protein that is found in cells. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences. Chromosomes also contain DNA-bound proteins, which serve to package the DNA and control its functions. The word *chromosome* comes from the Greek χρῶμα (*chroma*, colour) and σῶμα (*soma*, body) due to their property of being very strongly stained by particular dyes.

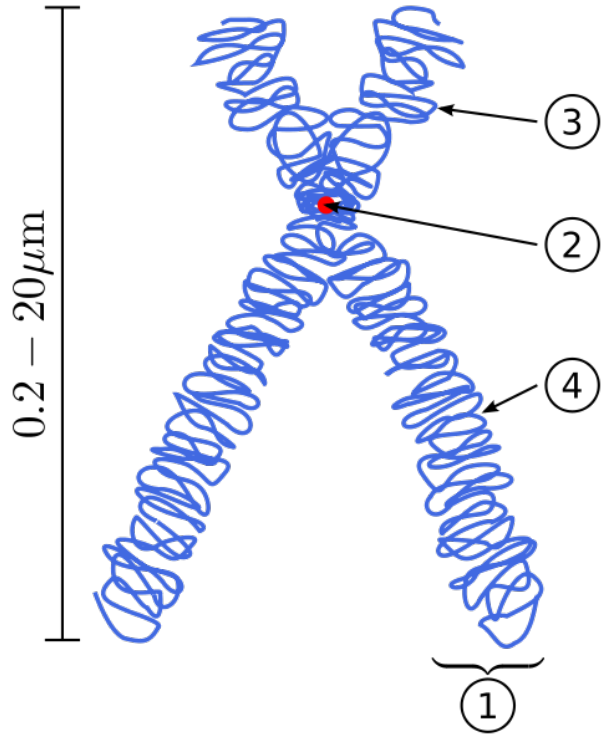


Diagram of a replicated and condensed metaphase eukaryotic chromosome. (1) Chromatid – one of the two identical parts of the chromosome after S phase. (2) Centromere – the point where the two chromatids touch, and where the microtubules attach. (3) Short arm. (4) Long arm.

Chromosomes vary widely between different organisms. The DNA molecule may be circular or linear, and can be composed of 10,000 to 1,000,000,000 nucleotides in a long chain. Typically eukaryotic cells (cells with nuclei) have large linear chromosomes and prokaryotic cells (cells without defined nuclei) have smaller circular chromosomes, although there are many exceptions to this rule. Furthermore, cells may contain more than one type of chromosome; for example, mitochondria in most eukaryotes and chloroplasts in plants have their own small chromosomes.

In eukaryotes, nuclear chromosomes are packaged by proteins into a condensed structure called chromatin. This allows the very long DNA molecules to fit into the cell nucleus. The structure of chromosomes and chromatin varies through the cell cycle. Chromosomes are the essential unit for cellular division and must be replicated, divided, and passed successfully to their daughter cells so as to ensure the genetic diversity and survival of their progeny. Chromosomes may exist as either duplicated or unduplicated—unduplicated chromosomes are single linear strands, whereas duplicated chromosomes (copied during synthesis phase) contain two copies joined by a centromere.

Compaction of the duplicated chromosomes during mitosis and meiosis results in the classic four-arm structure (pictured to the right). Chromosomal recombination plays a vital role in genetic diversity. If these structures are manipulated incorrectly, through processes known as chromosomal instability and translocation, the cell may undergo mitotic catastrophe and die, or it may unexpectedly evade apoptosis leading to the progression of cancer.

In practice "chromosome" is a rather loosely defined term. In prokaryotes and viruses, the term genophore is more appropriate when no chromatin is present. However, a large body of work uses the term chromosome regardless of chromatin content. In prokaryotes DNA is usually arranged as a circle, which is tightly coiled in on itself, sometimes accompanied by one or more smaller, circular DNA molecules called plasmids. These small circular genomes are also found in mitochondria and chloroplasts, reflecting their bacterial origins. The simplest genophores are found in viruses: these DNA or RNA molecules are short linear or circular genophores that often lack structural proteins.

History

Chromosomes as vectors of heredity

In a series of experiments, Theodor Boveri gave the definitive demonstration that chromosomes are the vectors of heredity. His two principles were based upon the *continuity* of chromosomes and the *individuality* of chromosomes. It is the second of these principles that was so original. Boveri was able to test the proposal put forward by Wilhelm Roux, that each chromosome carries a different genetic load, and showed that Roux was right. Upon the rediscovery of Mendel, Boveri was able to point out the connection between the rules of inheritance and the behaviour of the chromosomes. It is interesting to see that Boveri influenced two generations of American cytologists:

Edmund Beecher Wilson, Walter Sutton and Theophilus Painter were all influenced by Boveri (Wilson and Painter actually worked with him).

In his famous textbook *The Cell*, Wilson linked Boveri and Sutton together by the Boveri-Sutton theory. Mayr remarks that the theory was hotly contested by some famous geneticists: William Bateson, Wilhelm Johannsen, Richard Goldschmidt and T.H. Morgan, all of a rather dogmatic turn-of-mind. Eventually complete proof came from chromosome maps in Morgan's own lab.

Chromosomes in eukaryotes

Eukaryotes (cells with nuclei such as those found in plants, yeast, and animals) possess multiple large linear chromosomes contained in the cell's nucleus. Each chromosome has one centromere, with one or two arms projecting from the centromere, although, under most circumstances, these arms are not visible as such. In addition, most eukaryotes have a small circular mitochondrial genome, and some eukaryotes may have additional small circular or linear cytoplasmic chromosomes.

In the nuclear chromosomes of eukaryotes, the uncondensed DNA exists in a semi-ordered structure, where it is wrapped around histones (structural proteins), forming a composite material called chromatin.

Chromatin

Chromatin is the complex of DNA and protein found in the eukaryotic nucleus, which packages chromosomes. The structure of chromatin varies significantly between different stages of the cell cycle, according to the requirements of the DNA.

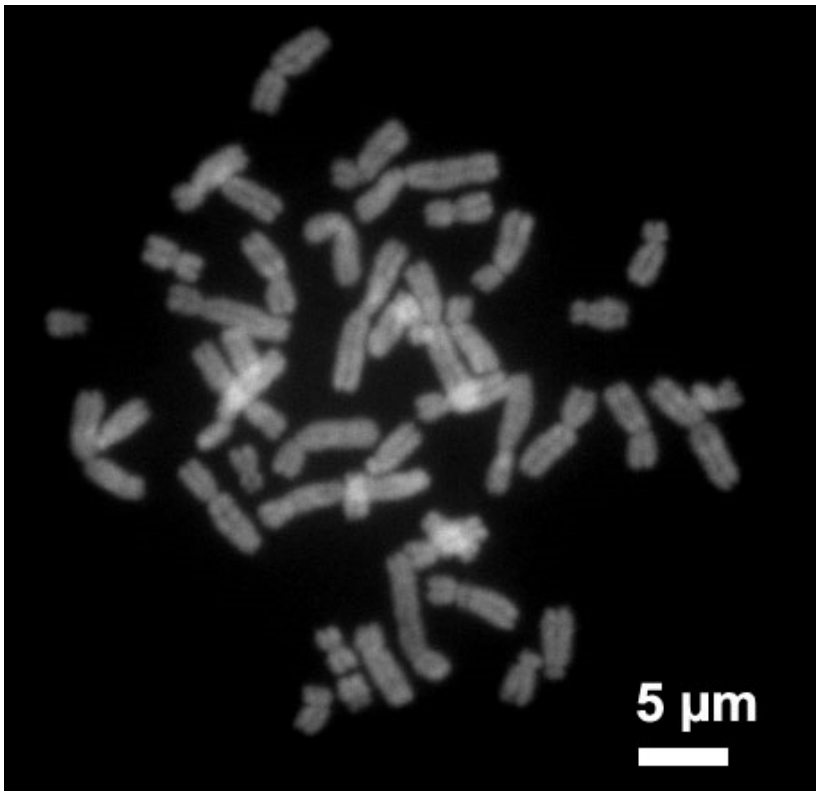
Interphase chromatin

During interphase (the period of the cell cycle where the cell is not dividing), two types of chromatin can be distinguished:

- Euchromatin, which consists of DNA that is active, e.g., being expressed as protein.
- Heterochromatin, which consists of mostly inactive DNA. It seems to serve structural purposes during the chromosomal stages. Heterochromatin can be further distinguished into two types:
    - *Constitutive heterochromatin*, which is never expressed. It is located around the centromere and usually contains repetitive sequences.
    - *Facultative heterochromatin*, which is sometimes expressed.

Individual chromosomes cannot be distinguished at this stage – they appear in the nucleus as a homogeneous tangled mix of DNA and protein.

Metaphase chromatin and division



Human chromosomes during metaphase

In the early stages of mitosis or meiosis (cell division), the chromatin strands become more and more condensed. They cease to function as accessible genetic material (transcription stops) and become a compact transportable form. This compact form makes the individual chromosomes visible, and they form the classic four arm structure, a pair of sister chromatids attached to each other at the centromere. The shorter arms are called *p arms* (from the French *petit*, small) and the longer arms are called *q arms* (*q* follows *p* in the Latin alphabet). This is the only natural context in which individual chromosomes are visible with an optical microscope.

During divisions, long microtubules attach to the centromere and the two opposite ends of the cell. The microtubules then pull the chromatids apart, so that each daughter cell inherits one set of chromatids. Once the cells have divided, the chromatids are uncoiled and can function again as chromatin. In spite of their appearance, chromosomes are structurally highly condensed, which enables these giant DNA structures to be contained within a cell nucleus (Fig. 2).

The self-assembled microtubules form the spindle, which attaches to chromosomes at specialized structures called kinetochores, one of which is present on each sister chromatid. A special DNA base sequence in the region of the kinetochores provides, along with special proteins, longer-lasting attachment in this region.

Chromosomes in prokaryotes

The prokaryotes – bacteria and archaea – typically have a single circular chromosome, but many variations do exist. Most bacteria have a single circular chromosome that can range in size from only 160,000 base pairs in the endosymbiotic bacterium *Candidatus Carsonella ruddii*, to 12,200,000 base pairs in the soil-dwelling bacterium *Sorangium cellulosum*. Spirochaetes of the genus *Borrelia* are a notable exception to this arrangement, with bacteria such as *Borrelia burgdorferi*, the cause of Lyme disease, containing a single linear chromosome.

Structure in sequences

Prokaryotic chromosomes have less sequence-based structure than eukaryotes. Bacteria typically have a single point (the origin of replication) from which replication starts, whereas some archaea contain multiple replication origins. The genes in prokaryotes are often organized in operons, and do not usually contain introns, unlike eukaryotes.

DNA packaging

Prokaryotes do not possess nuclei. Instead, their DNA is organized into a structure called the nucleoid. The nucleoid is a distinct structure and occupies a defined region of the bacterial cell. This structure is, however, dynamic and is maintained and remodeled by the actions of a range of histone-like proteins, which associate with the bacterial chromosome. In archaea, the DNA in chromosomes is even more organized, with the DNA packaged within structures similar to eukaryotic nucleosomes.

Bacterial chromosomes tend to be tethered to the plasma membrane of the bacteria. In molecular biology application, this allows for its isolation from plasmid DNA by centrifugation of lysed bacteria and pelleting of the membranes (and the attached DNA).

Prokaryotic chromosomes and plasmids are, like eukaryotic DNA, generally supercoiled. The DNA must first be released into its relaxed state for access for transcription, regulation, and replication.

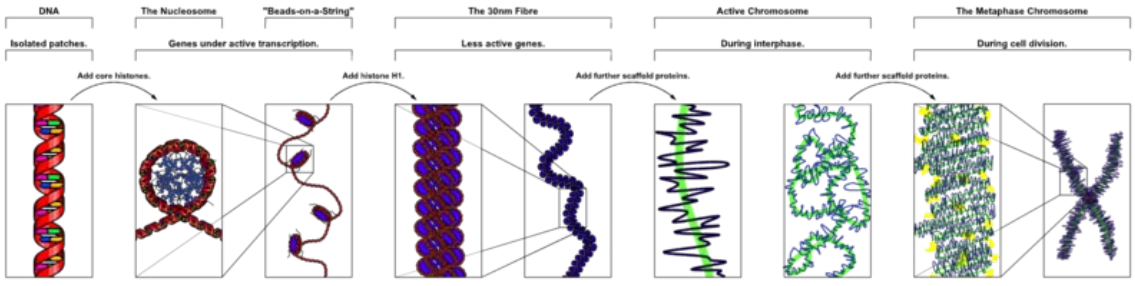


Fig. 2: The major structures in DNA compaction; DNA, the nucleosome, the 10nm "beads-on-a-string" fibre, the 30nm fibre and the metaphase chromosome.

Number of chromosomes in various organisms

Eukaryotes

These tables give the total number of chromosomes (including sex chromosomes) in a cell nucleus. For example, human cells are diploid and have 22 different types of autosome, each present as two copies, and two sex chromosomes. This gives 46 chromosomes in total. Other organisms have more than two copies of their chromosomes, such as bread wheat, which is *hexaploid* and has six copies of seven different chromosomes – 42 chromosomes in total.

Chromosome numbers in some plants

Plant Species	**#**
Arabidopsis thaliana (diploid)	10
Rye (diploid)	14
Maize (diploid or palaeotetraploid)	20
Einkorn wheat (diploid)	14
Durum wheat (tetraploid)	28
Bread wheat (hexaploid)	42
Cultivated tobacco (tetraploid)	48
Adder's Tongue Fern (diploid)	}

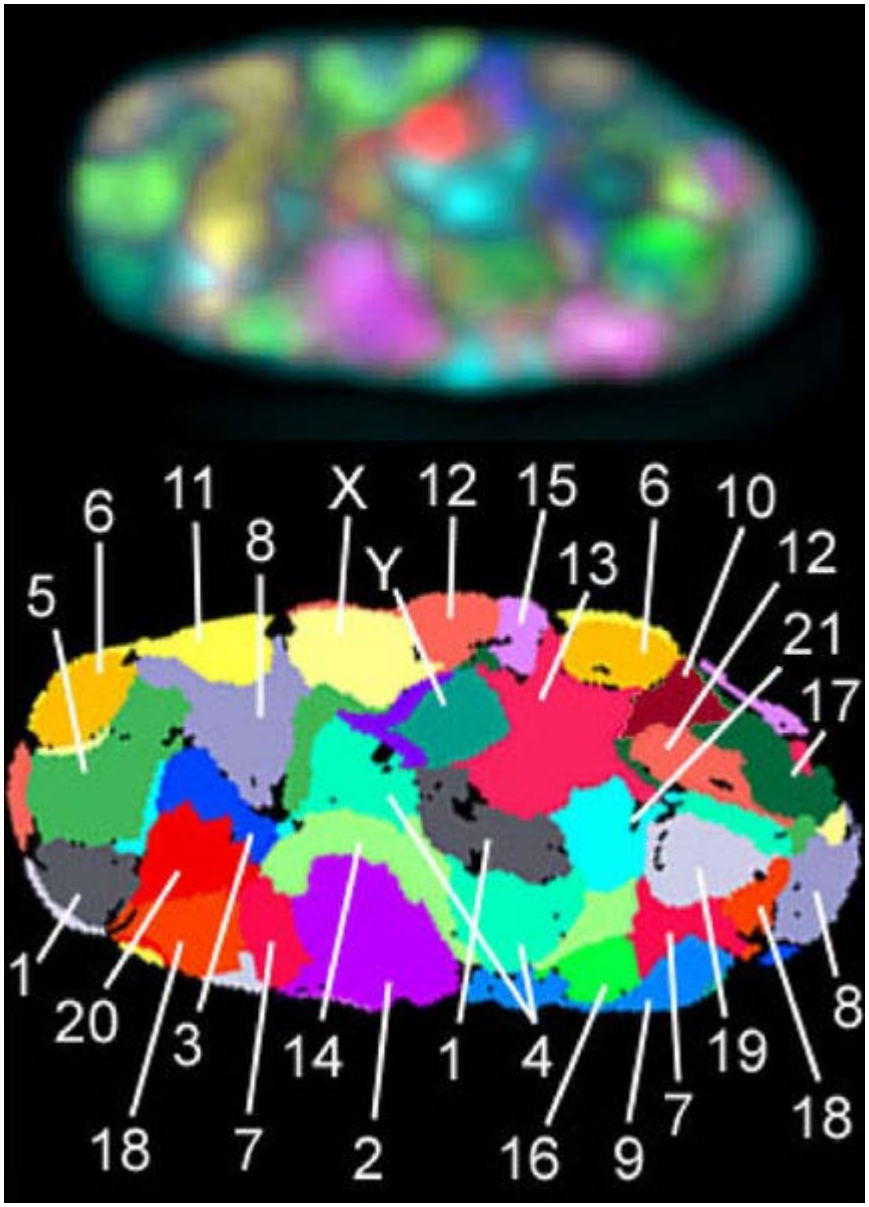
Chromosome numbers (2n) in some animals

Species	#	Species	#
Common fruit fly	8	Guinea Pig	64
Guppy (*poecilia reticulata*)	46	Garden snail	54
Earthworm (*Octodrilus complanatus*)	36	Tibetan fox	36
Domestic cat	38	Domestic pig	38
Laboratory mouse	40	Laboratory rat	42
Rabbit (*Oryctolagus cuniculus*)	44	Syrian hamster	44
Hares	48	Human	46
Gorillas, Chimpanzees	48	Domestic sheep	54
Elephants	56	Cow	60
Donkey	62	Horse	64
Dog	78	Kingfisher	132
Goldfish	100-104	Silkworm	56

Chromosome numbers in other organisms

Species	Large Chromosomes	Intermediate Chromosomes	Microchromosomes
Trypanosoma brucei	11	6	~100
Domestic Pigeon (*Columba livia domestics*)	18	-	59-63
Chicken	8	2 sex chromosomes	60

Normal members of a particular eukaryotic species all have the same number of nuclear chromosomes (see the table). Other eukaryotic chromosomes, i.e., mitochondrial and plasmid-like small chromosomes, are much more variable in number, and there may be thousands of copies per cell.



The 23 human chromosome territories during prometaphase in fibroblast cells

Asexually reproducing species have one set of chromosomes, which are the same in all body cells. However, asexual species can be either haploid or diploid.

Sexually reproducing species have somatic cells (body cells), which are diploid [2n] having two sets of chromosomes, one from the mother and one from the father. Gametes, reproductive cells, are haploid [n]: They have one set of chromosomes. Gametes are produced by meiosis of a diploid germ line cell. During meiosis, the matching chromosomes of father and mother can exchange small parts of themselves (crossover), and thus create new chromosomes that are not inherited solely from either parent. When a male and a female gamete merge (fertilization), a new diploid organism is formed.

Some animal and plant species are polyploid [Xn]: They have more than two sets of homologous chromosomes. Plants important in agriculture such as tobacco or wheat are often polyploid, compared to their ancestral species. Wheat has a haploid number of seven chromosomes, still seen in some cultivars as well as the wild progenitors. The more-common pasta and bread wheats are polyploid, having 28 (tetraploid) and 42 (hexaploid) chromosomes, compared to the 14 (diploid) chromosomes in the wild wheat.

Prokaryotes

Prokaryote species generally have one copy of each major chromosome, but most cells can easily survive with multiple copies. For example, *Buchnera*, a symbiont of aphids has multiple copies of its chromosome, ranging from 10–400 copies per cell. However, in some large bacteria, such as *Epulopiscium fishelsoni* up to 100,000 copies of the chromosome can be present. Plasmids and plasmid-like small chromosomes are, as in eukaryotes, very variable in copy number. The number of plasmids in the cell is almost entirely determined by the rate of division of the plasmid – fast division causes high copy number, and vice versa.

Karyotype

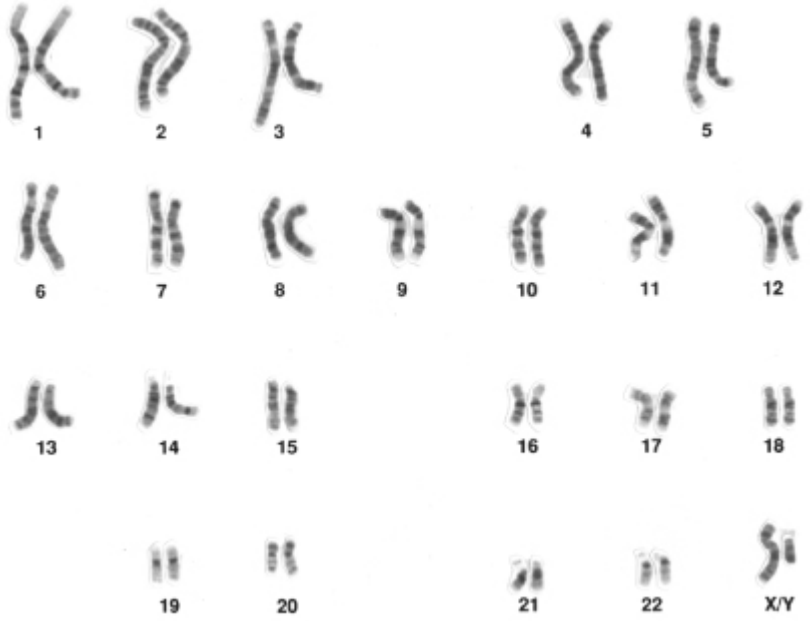


Figure 3: Karyogram of a human male

In general, the **karyotype** is the characteristic chromosome complement of a eukaryote species. The preparation and study of karyotypes is part of cytogenetics.

Although the replication and transcription of DNA is highly standardized in eukaryotes, *the same cannot be said for their karyotypes*, which are often highly variable. There may

be variation between species in chromosome number and in detailed organization. In some cases, there is significant variation within species. Often there is:

1. variation between the two sexes
2. variation between the germ-line and soma (between gametes and the rest of the body)
3. variation between members of a population, due to balanced genetic polymorphism
4. geographical variation between races
5. mosaics or otherwise abnormal individuals.

Also, variation in karyotype may occur during development from the fertilised egg.

The technique of determining the karyotype is usually called *karyotyping*. Cells can be locked part-way through division (in metaphase) in vitro (in a reaction vial) with colchicine. These cells are then stained, photographed, and arranged into a *karyogram*, with the set of chromosomes arranged, autosomes in order of length, and sex chromosomes (here X/Y) at the end: Fig. 3.

Like many sexually reproducing species, humans have special gonosomes (sex chromosomes, in contrast to autosomes). These are XX in females and XY in males.

Historical note

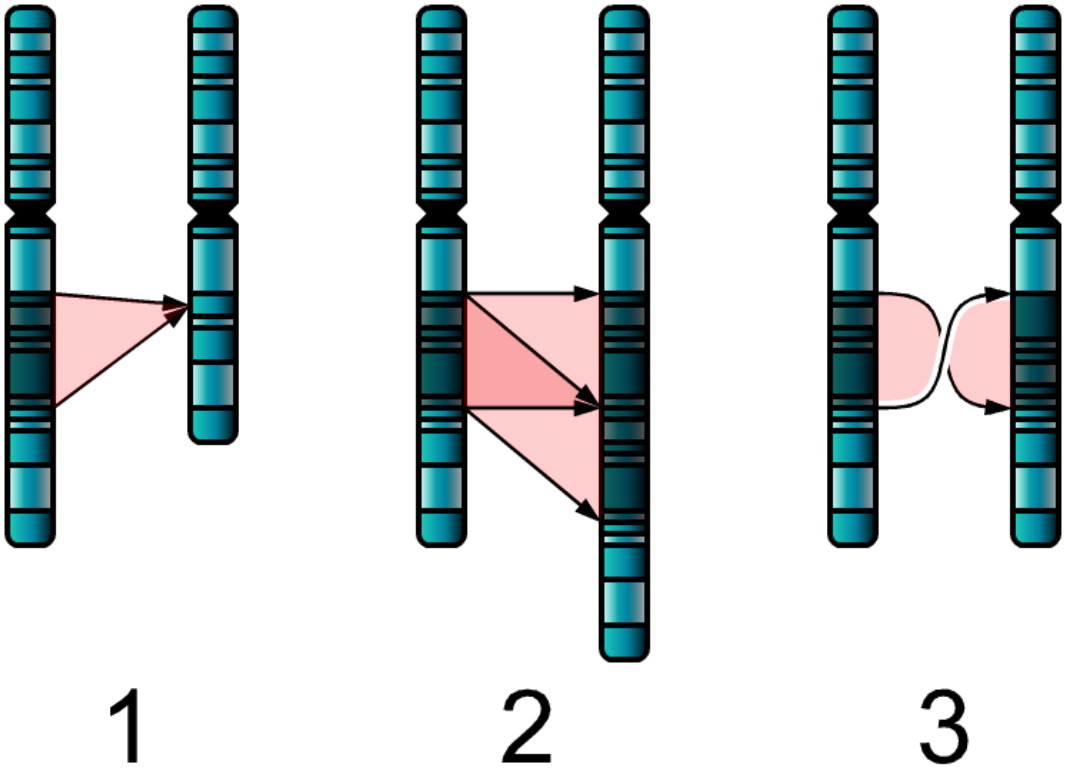
Investigation into the human karyotype took many years to settle the most basic question. How many chromosomes does a normal diploid human cell contain? In 1912, Hans von Winiwarter reported 47 chromosomes in spermatogonia and 48 in oogonia, concluding an XX/XO sex determination mechanism. Painter in 1922 was not certain whether the diploid number of man is 46 or 48, at first favouring 46. He revised his opinion later from 46 to 48, and he correctly insisted on humans having an XX/XY system.

New techniques were needed to definitively solve the problem:

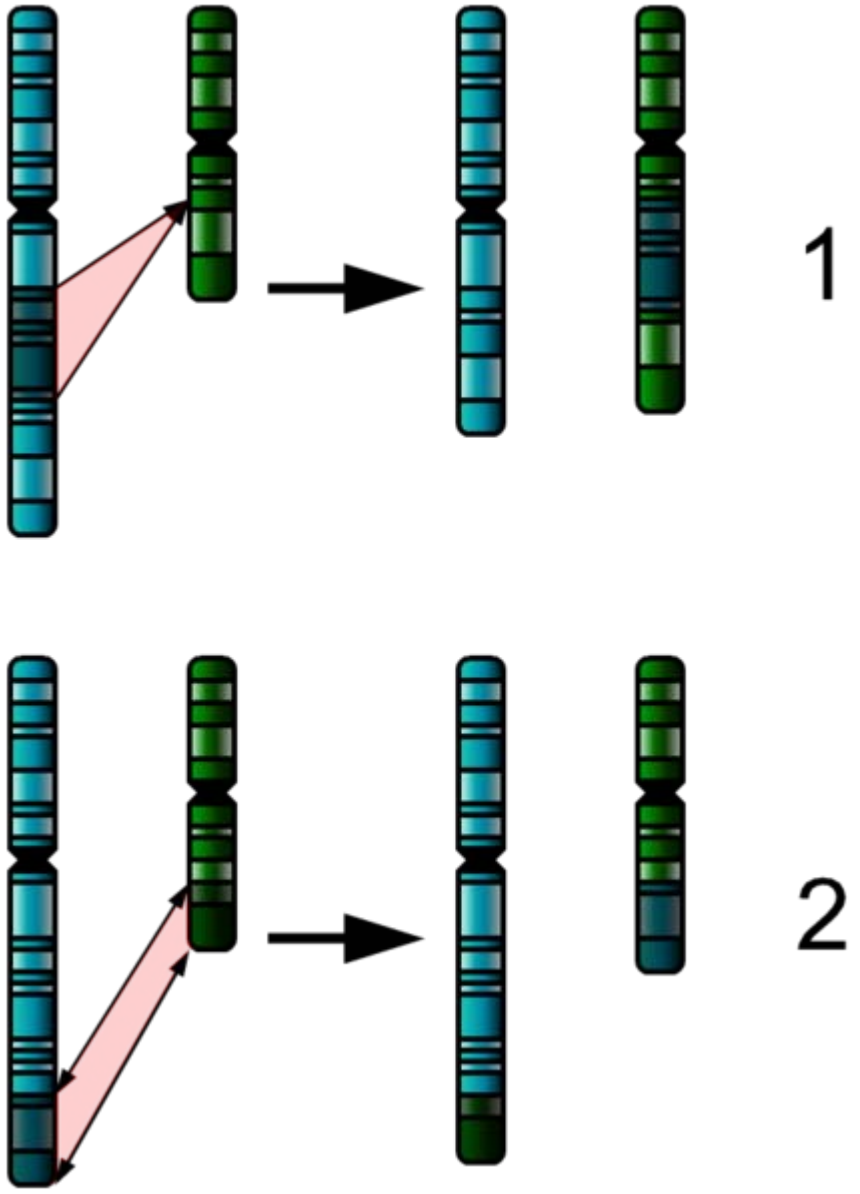
1. Using cells in culture
2. Pretreating cells in a hypotonic solution, which swells them and spreads the chromosomes
3. Arresting mitosis in metaphase by a solution of colchicine
4. Squashing the preparation on the slide forcing the chromosomes into a single plane
5. Cutting up a photomicrograph and arranging the result into an indisputable karyogram.

It took until the mid-1950s for it to become generally accepted that the human karyotype include only 46 chromosomes. Considering the techniques of Winiwarter and Painter, their results were quite remarkable. Chimpanzees (the closest living relatives to modern humans) have 48 chromosomes.

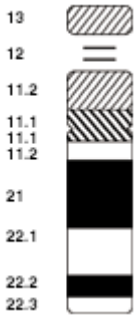
Chromosomal aberrations



The three major single chromosome mutations; deletion (1), duplication (2) and inversion (3).



The two major two-chromosome mutations; insertion (1) and translocation (2).



In Down syndrome, there are three copies of chromosome 21

Chromosomal aberrations are disruptions in the normal chromosomal content of a cell and are a major cause of genetic conditions in humans, such as Down syndrome. Some chromosome abnormalities do not cause disease in carriers, such as translocations, or chromosomal inversions, although they may lead to a higher chance of birthing a child with a chromosome disorder. Abnormal numbers of chromosomes or chromosome sets, aneuploidy, may be lethal or give rise to genetic disorders. Genetic counseling is offered for families that may carry a chromosome rearrangement.

The gain or loss of DNA from chromosomes can lead to a variety of genetic disorders. Human examples include:

- Cri du chat, which is caused by the deletion of part of the short arm of chromosome 5. "Cri du chat" means "cry of the cat" in French, and the condition was so-named because affected babies make high-pitched cries that sound like those of a cat. Affected individuals have wide-set eyes, a small head and jaw, moderate to severe mental health issues, and are very short.
- Down syndrome, usually is caused by an extra copy of chromosome 21 (trisomy 21). Characteristics include decreased muscle tone, stockier build, asymmetrical skull, slanting eyes and mild to moderate developmental disability.
- Edwards syndrome, which is the second-most-common trisomy; Down syndrome is the most common. It is a trisomy of chromosome 18. Symptoms include motor retardation, developmental disability and numerous congenital anomalies causing serious health problems. Ninety percent die in infancy; however, those that live past their first birthday usually are quite healthy thereafter. They have a characteristic clenched hands and overlapping fingers.
- Idic15, abbreviation for Isodicentric 15 on chromosome 15; also called the following names due to various researches, but they all mean the same; IDIC(15), Inverted duplication 15, extra Marker, Inv dup 15, partial tetrasomy 15
- Jacobsen syndrome, also called the terminal 11q deletion disorder. This is a very rare disorder. Those affected have normal intelligence or mild developmental disability, with poor expressive language skills. Most have a bleeding disorder called Paris-Trousseau syndrome.
- Klinefelter's syndrome (XXY). Men with Klinefelter syndrome are usually sterile, and tend to have longer arms and legs and to be taller than their peers. Boys with the syndrome are often shy and quiet, and have a higher incidence of speech delay and dyslexia. During puberty, without testosterone treatment, some of them may develop gynecomastia.
- Patau Syndrome, also called D-Syndrome or trisomy-13. Symptoms are somewhat similar to those of trisomy-18, but they do not have the characteristic hand shape.
- Small supernumerary marker chromosome. This means there is an extra, abnormal chromosome. Features depend on the origin of the extra genetic material. Cat-eye syndrome and isodicentric chromosome 15 syndrome (or Idic15) are both caused by a supernumerary marker chromosome, as is Pallister-Killian syndrome.

- Triple-X syndrome (XXX). XXX girls tend to be tall and thin. They have a higher incidence of dyslexia.
- Turner syndrome (X instead of XX or XY). In Turner syndrome, female sexual characteristics are present but underdeveloped. People with Turner syndrome often have a short stature, low hairline, abnormal eye features and bone development and a "caved-in" appearance to the chest.
- XYY syndrome. XYY boys are usually taller than their siblings. Like XXY boys and XXX girls, they are somewhat more likely to have learning difficulties.
- Wolf-Hirschhorn syndrome, which is caused by partial deletion of the short arm of chromosome 4. It is characterized by severe growth retardation and severe to profound mental health issues.

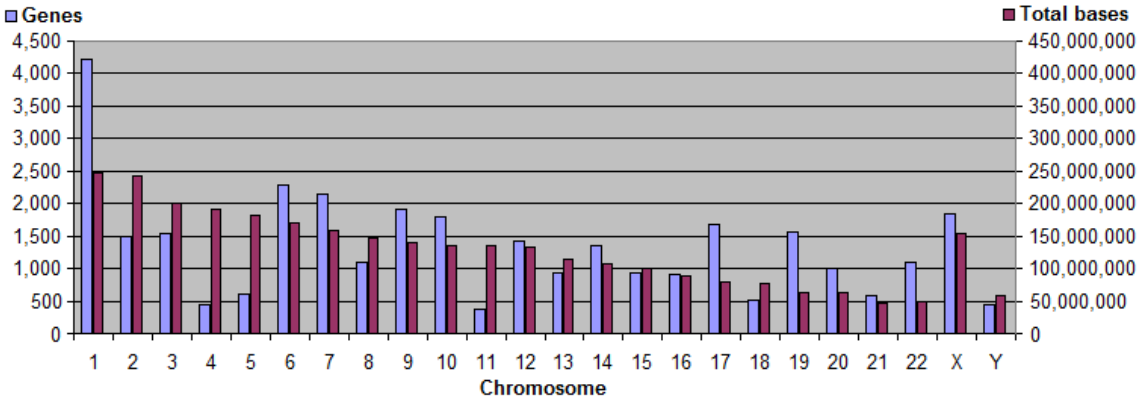
Chromosomal mutations produce changes in whole chromosomes (more than one gene) or in the number of chromosomes present.

- Deletion – loss of part of a chromosome
- Duplication – extra copies of a part of a chromosome
- Inversion – reverse the direction of a part of a chromosome
- Translocation – part of a chromosome breaks off and attaches to another chromosome

Most mutations are neutral – have little or no effect. Chromosomal aberrations are the changes in the structure of chromosomes. It has a great role in evolution. A detailed graphical display of all human chromosomes and the diseases annotated at the correct spot may be found at the Oak Ridge National Laboratory.

Human chromosomes

Chromosomes can be divided into two types—autosomes, and sex chromosomes. Certain genetic traits are linked to your sex, and are passed on through the sex chromosomes. The autosomes contain the rest of the genetic hereditary information. All act in the same way during cell division. Human cells have 23 pairs of large linear nuclear chromosomes, (22 pairs of autosomes and one pair of sex chromosomes) giving a total of 46 per cell. In addition to these, human cells have many hundreds of copies of the mitochondrial genome. Sequencing of the human genome has provided a great deal of information about each of the chromosomes. Below is a table compiling statistics for the chromosomes, based on the Sanger Institute's human genome information in the Vertebrate Genome Annotation (VEGA) database. Number of genes is an estimate as it is in part based on gene predictions. Total chromosome length is an estimate as well, based on the estimated size of unsequenced heterochromatin regions.



Chromosome	Genes	Total bases	Sequenced bases
1	4,220	247,199,719	224,999,719
2	1,491	242,751,149	237,712,649
3	1,550	199,446,827	194,704,827
4	446	191,263,063	187,297,063
5	609	180,837,866	177,702,766
6	2,281	170,896,993	167,273,993
7	2,135	158,821,424	154,952,424
8	1,106	146,274,826	142,612,826
9	1,920	140,442,298	120,312,298
10	1,793	135,374,737	131,624,737
11	379	134,452,384	131,130,853
12	1,430	132,289,534	130,303,534
13	924	114,127,980	95,559,980
14	1,347	106,360,585	88,290,585
15	921	100,338,915	81,341,915
16	909	88,822,254	78,884,754
17	1,672	78,654,742	77,800,220
18	519	76,117,153	74,656,155
19	1,555	63,806,651	55,785,651
20	1,008	62,435,965	59,505,254
21	578	46,944,323	34,171,998
22	1,092	49,528,953	34,893,953
X (sex chromosome)	1,846	154,913,754	151,058,754
Y (sex chromosome)	454	57,741,652	25,121,652
Total	**32,185**	**3,079,843,747**	**2,857,698,560**

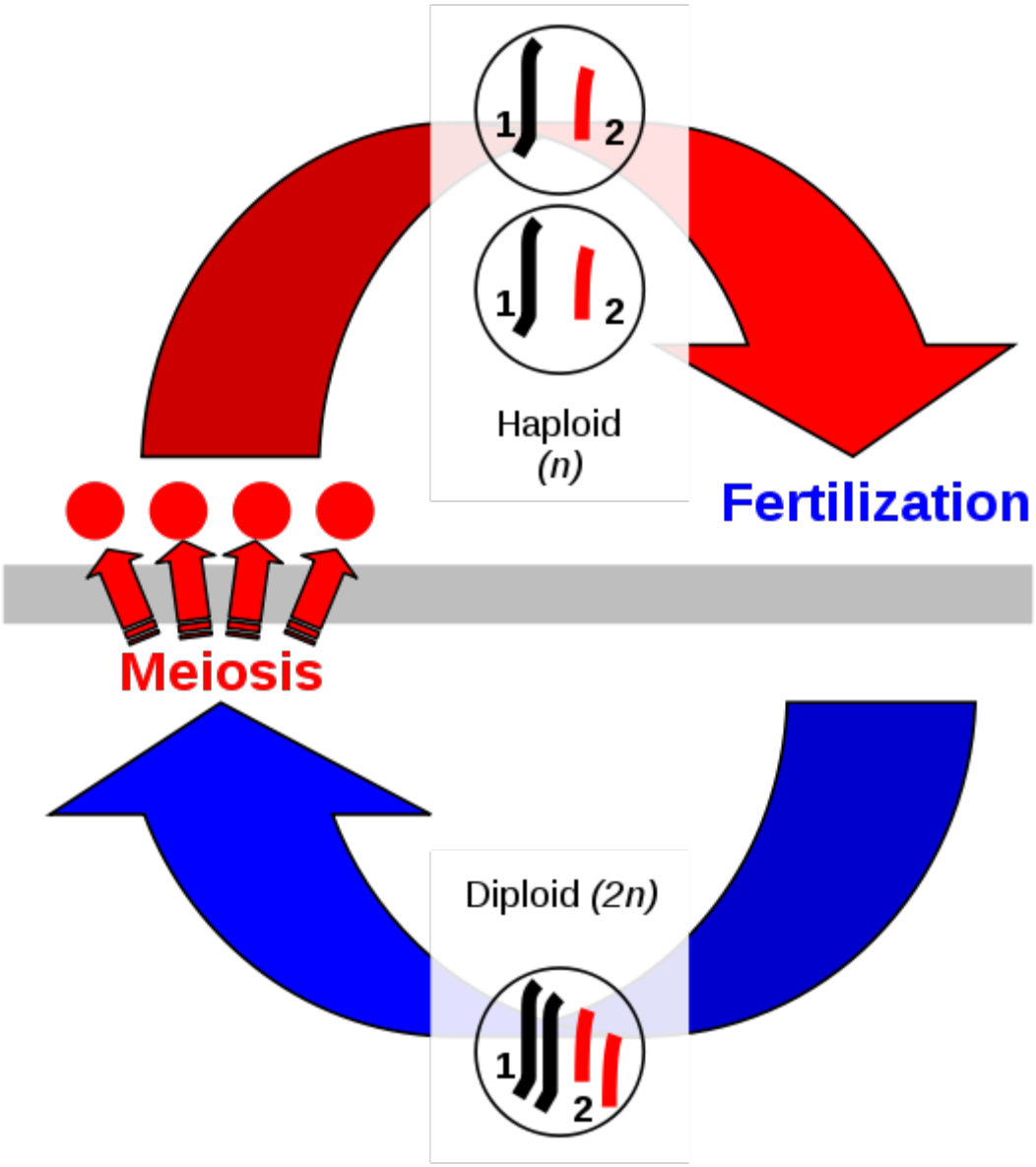
Chapter- 6

Sexual Reproduction

Sexual reproduction is the creation of a new organism by combining the genetic material of two organisms. The two main processes are: meiosis, involving the halving of the number of chromosomes; and fertilization, involving the fusion of two gametes and the restoration of the original number of chromosomes. During meiosis, the chromosomes of each pair usually cross over to achieve homologous recombination.

The evolution of sexual reproduction is a major puzzle. The first fossilized evidence of sexually reproducing organisms is from eukaryotes of the Stenian period, about 1 to 1.2 billion years ago. Sexual reproduction is the primary method of reproduction for the vast majority of macroscopic organisms, including almost all animals and plants. Bacterial conjugation, the transfer of DNA between two bacteria, is often mistakenly confused with sexual reproduction, because the mechanics are similar.

A major question is why sexual reproduction persists when parthenogenesis appears in some ways to be a superior form of reproduction. Contemporary evolutionary thought proposes some explanations. It may be due to selection pressure on the clade itself—the ability for a population to radiate more rapidly in response to a changing environment through sexual recombination than parthenogenesis allows. Alternatively, sexual reproduction may allow for the "ratcheting" of evolutionary speed as one clade competes with another for a limited resource.



In the first stage of sexual reproduction, "meiosis," the number of chromosomes is reduced from a diploid number (2n) to a haploid number (n). During "fertilization," haploid gametes come together to form a diploid zygote and the original number of chromosomes (2n) is restored.

Plants

Animals typically produce male gametes called sperm, and female gametes called eggs and ova, following immediately after meiosis, with the gametes produced directly by meiosis. Plants on the other hand have mitosis occurring in spores, which are produced by meiosis. The spores germinate into the gametophyte phase. The gametophytes of different groups of plants vary in size; angiosperms have as few as three cells in pollen, and mosses and other so called primitive plants may have several million cells. Plants

have an alternation of generations where the sporophyte phase is succeeded by the gametophyte phase. The sporophyte phase produces spores within the sporangium by meiosis.

Flowering plants

Flowers are the sexual organs of flowering plants.

Flowering plants are the dominant plant form on land and they reproduce by sexual and asexual means. Often their most distinguishing feature is their reproductive organs, commonly called flowers. The anther produces male gametophytes, the sperm is produced in pollen grains, which attach to the stigma on top of a carpel, in which the female gametophytes (inside ovules) are located. After the pollen tube grows through the carpel's style, the sex cell nuclei from the pollen grain migrate into the ovule to fertilize

the egg cell and endosperm nuclei within the female gametophyte in a process termed double fertilization. The resulting zygote develops into an embryo, while the triploid endosperm (one sperm cell plus two female cells) and female tissues of the ovule give rise to the surrounding tissues in the developing seed. The ovary, which produced the female gametophyte(s), then grows into a fruit, which surrounds the seed(s). Plants may either self-pollinate or cross-pollinate. Nonflowering plants like ferns, moss and liverworts use other means of sexual reproduction.

Ferns

Ferns mostly produce large diploid sporophytes with rhizomes, roots and leaves; and on fertile leaves called sporangium, spores are produced. The spores are released and germinate to produce short, thin gametophytes that are typically heart shaped, small and green in color. The gametophytes or thallus, produce both motile sperm in the antheridia and egg cells in separate archegonia. After rains or when dew deposits a film of water, the motile sperm are splashed away from the antheridia, which are normally produced on the top side of the thallus, and swim in the film of water to the antheridia where they fertilize the egg. To promote out crossing or cross fertilization the sperm are released before the eggs are receptive of the sperm, making it more likely that the sperm will fertilize the eggs of different thallus. A zygote is formed after fertilization, which grows into a new sporophytic plant. The condition of having separate sporephyte and gametophyte plants is called alternation of generations. Other plants with similar reproductive means include the *Psilotum*, *Lycopodium*, *Selaginella* and *Equisetum*.

Bryophytes

The bryophytes, which include liverworts, hornworts and mosses, reproduce both sexually and vegetatively. They are small plants found growing in moist locations and like ferns, have motile sperm with flagella and need water to facilitate sexual reproduction. These plants start as a haploid spore that grows into the dominate form, which is a multicellular haploid body with leaf-like structures that photosynthesize. Haploid gametes are produced in antherida and archegonia by mitosis. The sperm released from the antherida respond to chemicals released by ripe archegonia and swim to them in a film of water and fertilize the egg cells thus producing a zygote. The zygote divides by mitotic division and grows into a sporophyte that is diploid. The multicellular diploid sporophyte produces structures called spore capsules, which are connected by seta to the archegonia. The spore capsules produce spores by meiosis, when ripe the capsules burst open and the spores are released. Bryophytes show considerable variation in their breeding structures and the above is a basic outline. Also in some species each plant is one sex while other species produce both sexes on the same plant.

Fungi

Fungi are classified by the methods of sexual reproduction they employ. The outcome of sexual reproduction most often is the production of resting spores that are used to survive

inclement times and to spread. There are typically three phases in the sexual reproduction of fungi: plasmogamy, karyogamy and meiosis.

Insects

Insects mating on a *liatris* flower head.

Insect species make up more than two-thirds of all extant animal species, and most insect species use sex for reproduction, though some species are facultatively parthenogenetic. Many species have sexual dimorphism, while in others the sexes look nearly identical. Typically they have two sexes with males producing spermatozoa and females ova. The ova develop into eggs that have a covering called the chorion, which forms before internal fertilization. Insects have very diverse mating and reproductive strategies most often resulting in the male depositing spermatophore within the female, which stores the sperm until she is ready for egg fertilization. After fertilization, and the formation of a zygote, and varying degrees of development; the eggs are deposited outside the female in

many species, or in some, they develop further within the female and live born offspring are produced.

Mammals

There are three extant kinds of mammals: Monotremes, Placentals and Marsupials, all with internal fertilisation. In placental mammals, offspring are born as juveniles: complete animals with the sex organs present although not reproductively functional. After several months or years, the sex organs develop further to maturity and the animal becomes sexually mature. Most female mammals are only fertile during certain periods during their estrous cycle, at which point they are ready to mate. Individual male and female mammals meet and carry out copulation. For most mammals, males and females exchange sexual partners throughout their adult lives.

Male

The male reproductive system contains two main divisions: the penis, and the testicles, the latter of which is where sperm are produced. In humans, both of these organs are outside the abdominal cavity, but they can be primarily housed within the abdomen in other animals (for instance, in dogs, the penis is internal except when mating). Having the testicles outside the abdomen best facilitates temperature regulation of the sperm, which require specific temperatures to survive. Sperm are the smaller of the two gametes and are generally very short-lived, requiring males to produce them continuously from the time of sexual maturity until death. Prior to ejaculation the produced sperm are stored in the epididymis. The sperm cells are motile and they swim using tail-like flagella to propel themselves towards the ovum. The sperm follows temperature gradients (thermotaxis) and chemical gradients (chemotaxis) to locate the ovum.

Female

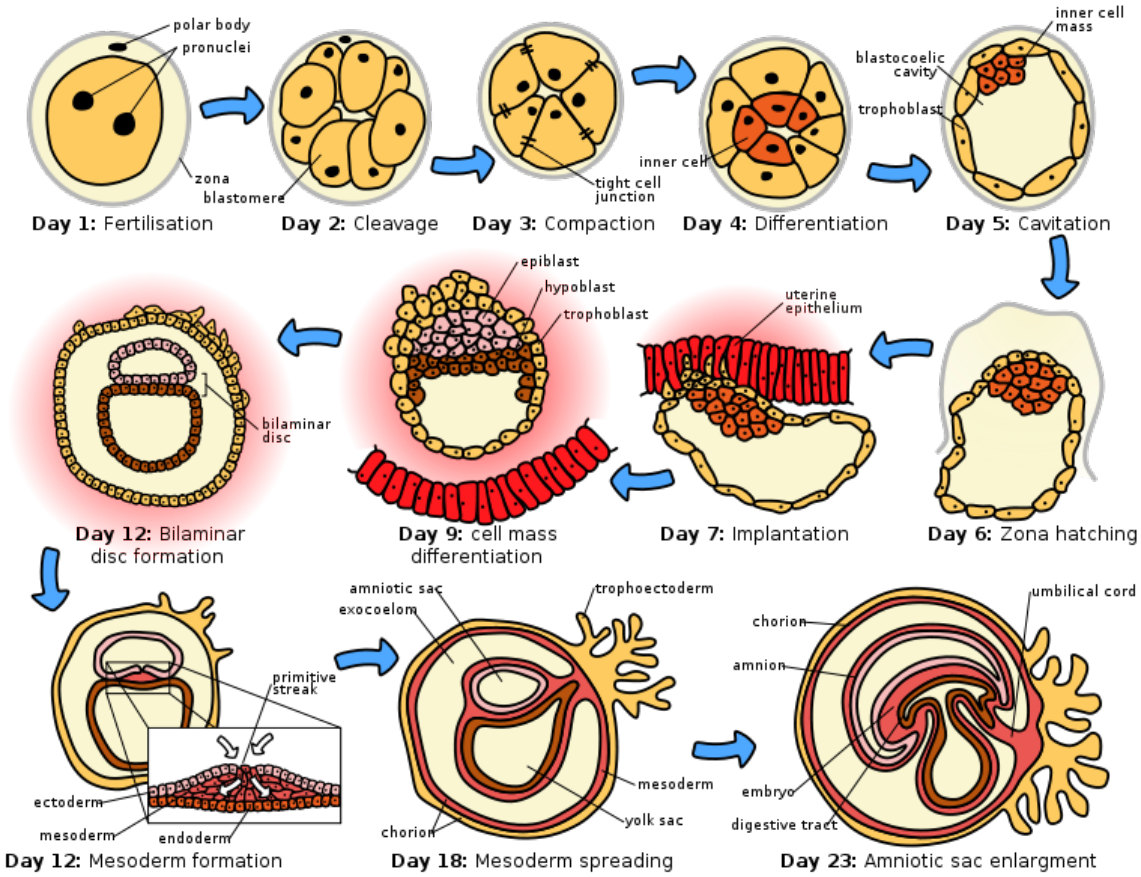
The female reproductive system likewise contains two main divisions: the vagina and uterus, which act as the receptacle for the sperm, and the ovaries, which produce the female's ova. All of these parts are always internal. The vagina is attached to the uterus through the cervix, while the uterus is attached to the ovaries via the Fallopian tubes. At certain intervals, the ovaries release an ovum, which passes through the fallopian tube into the uterus.

If, in this transit, it meets with sperm, the egg selects sperm with which to merge; this is termed fertilization. The fertilization usually occurs in the oviducts, but can happen in the uterus itself. The zygote then implants itself in the wall of the uterus, where it begins the processes of embryogenesis and morphogenesis. When developed enough to survive outside the womb, the cervix dilates and contractions of the uterus propel the fetus through the birth canal, which is the vagina.

The ova, which are the female sex cells, are much larger than the sperm and are normally formed within the ovaries of the fetus before its birth. They are mostly fixed in location

within the ovary until their transit to the uterus, and contain nutrients for the later zygote and embryo. Over a regular interval, in response to hormonal signals, a process of oogenesis matures one ovum which is released and sent down the Fallopian tube. If not fertilized, this egg is released through menstruation in humans and other great apes, and reabsorbed in other mammals in the estrus cycle.

Gestation



The initial stages of human embryogenesis

Gestation, called *pregnancy* in humans, is the period of time during which the fetus develops, dividing via mitosis inside the female. During this time, the fetus receives all of its nutrition and oxygenated blood from the female, filtered through the placenta, which is attached to the fetus' abdomen via an umbilical cord. This drain of nutrients can be quite taxing on the female, who is required to ingest slightly higher levels of calories. In addition, certain vitamins and other nutrients are required in greater quantities than normal, often creating abnormal eating habits. The length of gestation, called the gestation period, varies greatly from species to species; it is 40 weeks in humans, 56–60 in giraffes and 16 days in hamsters.

Birth

Once the fetus is sufficiently developed, chemical signals start the process of birth, which begins with contractions of the uterus and the dilation of the cervix. The fetus then descends to the cervix, where it is pushed out into the vagina, and eventually out of the female. The newborn, which is called an infant in humans, should typically begin respiration on its own shortly after birth. Not long after, the placenta is passed as well. Most mammals eat this, as it is a good source of protein and other vital nutrients needed for caring for the young. The end of the umbilical cord attached to the young's abdomen eventually falls off on its own.

Monotremes

Monotremes, only five species of which exist, all from Australia and New Guinea, lay eggs. They have one opening for excretion and reproduction called the cloaca. They hold the eggs internally for several weeks, providing nutrients, and then lay them and cover them like birds. After less than two weeks the young hatches and crawls into its mother's pouch, much like marsupials, where it nurses for several weeks as it grows.

Marsupials

Marsupials reproductive systems differ markedly from those of placental mammals. Females have two vaginas, both of which open externally through one orifice but lead to different compartments within the uterus. Males generally have a two-pronged penis, which corresponds to the females' two vaginae. The penis is used only for discharging semen into females, and is separate from the urinary tract. Both sexes possess a cloaca, which is connected to a urogenital sac used to store waste before expulsion.

The female develops a kind of yolk sack in her womb which delivers nutrients to the embryo. Embryos of bandicoots, koalas and wombats additionally form placenta-like organs that connect them to the uterine wall, although the placenta-like organs are smaller than in placental mammals and it is not certain that they transfer nutrients from the mother to the embryo.

Pregnancy is very short, typically 4 to 5 weeks. The embryo is born at a very young stage of development, and is usually less than 5 cm (2.0 in) long at birth. It has been suggested that the short pregnancy is necessary to reduce the risk that the mother's immune system will attack the embryo.

The newborn marsupial uses its forelimbs (with relatively strong hands) to climb to a nipple, which is usually in a pouch on the mother's belly. The mother feeds the baby by contracting muscles over her mammary glands, as the baby is too weak to suckle. The newborn marsupial's need to use its forelimbs for climbing to the nipple has prevented the forelimbs from evolving into paddles or wings and has therefore prevented the appearance of aquatic or truly flying marsupials (although there are several marsupial gliders).

Fish

The vast majority of fish species lay eggs that are then fertilized by the male, some species lay their eggs on a substrate like a rock or on plants, while others scatter their eggs and the eggs are fertilized as they drift or sink in the water column. Some fish species use internal fertilization and then disperse the developing eggs or give birth to live offspring. Fish that have live-bearing offspring include the Guppy and Mollies or *Poecilia*. Fishes that give birth to live young can be ovoviviparous, where the eggs are fertilized within the female and the eggs simply hatch within the female body, or in seahorses, the male carries the developing young within a pouch, and gives birth to live young. Fishes can also be viviparous, where the female supplies nourishment to the internally growing offspring. Some fish are hermaphrodites, where a single fish is both male and female and can produce eggs and sperm. In hermaphroditic fish, some are male and female at the same time while in other fish they are serially hermaphroditic; starting as one sex and changing to the other. In at least one hermaphroditic species, self-fertilization occurs when the eggs and sperm are released together. Internal self-fertilization may occur in some other species. One fish species does not reproduce by sexual reproduction but uses sex to produce offspring; *Poecilia formosa* is a unisex species that uses a form of parthenogenesis called gynogenesis, where unfertilized eggs develop into embryos that produce female offspring. *Poecilia formosa* mate with males of other fish species that use internal fertilization, the sperm does not fertilize the eggs but stimulates the growth of the eggs which develops into embryos.

Chapter- 7

Genetic Linkage

Genetic linkage is the tendency of certain loci or alleles to be inherited together. Genetic loci that are physically close to one another on the same chromosome tend to stay together during meiosis, and are thus genetically *linked*.

Background

At the beginning of normal meiosis, a chromosome pair (made up of a chromosome from the mother and a chromosome from the father) intertwine and exchange sections or fragments of chromosome. The pair then breaks apart to form two chromosomes with a new combination of genes that differs from the combination supplied by the parents. Through this process of recombining genes, organisms can produce offspring with new combinations of maternal and paternal traits that may contribute to or enhance survival.

This recombination of genes, called the crossing over of DNA, can cause alleles previously on the same chromosome to be separated and end up in different daughter cells. The further the two alleles are apart, the greater the chance that a cross-over event may occur between them, and the greater the chance that the alleles are separated.

The relative distance between two genes can be calculated by taking the offspring of an organism showing two linked genetic traits, and finding the percentage of the offspring where the two traits do not run together. The higher the percentage of descendants that does not show both traits, the farther apart on the chromosome the two genes are. Genes for which this percentage is lower than 50% are typically thought to be linked.

Genetic linkage can also be understood by looking at the relationships among phenotypes. Among individuals of an experimental population or species, some phenotypes or traits can occur randomly with respect to one another, or with some correlation with respect to one another.

The former is known as independent assortment. Today, scientists understand that independent assortment occurs when the genes affecting the phenotypes are found on

different chromosomes or separated by a great enough distance on the same chromosome that recombination occurs at least half of the time.

The latter is known as genetic linkage. This occurs as an exception to independent assortment, and develops when genes appear near one another on the same chromosome. This phenomenon causes the genes to usually be inherited as a single unit. Genes inherited in this way are said to be linked, and are referred to as "linkage groups". For example, in fruit flies, the genes affecting eye color and wing length are inherited together because they appear on the same chromosome.

Discovery

Genetic linkage was first discovered by the British geneticists William Bateson and Reginald Punnett shortly after Mendel's laws were rediscovered. The understanding of genetic linkage was expanded by the work of Thomas Hunt Morgan. Morgan's observation that the amount of crossing over between linked genes differs led to the idea that crossover frequency might indicate the distance separating genes on the chromosome.

Alfred Sturtevant, a student of Morgan's, first developed genetic maps, also known as linkage maps. Sturtevant proposed that the greater the distance between linked genes, the greater the chance that non-sister chromatids would cross over in the region between the genes. By working out the number of recombinants it is possible to obtain a measure for the distance between the genes. This distance is called a **genetic map unit (m.u.)**, or a **centimorgan** and is defined as the distance between genes for which one product of meiosis in 100 is recombinant. A **recombinant frequency** (RF) of 1 % is equivalent to 1 m.u. But this equivalence is only a good approximate for small percentages; the largest percentage of recombinants cannot exceed 50%, which would be the situation where the two genes are at the extreme opposite ends of the same chromosomes. In this situation, any crossover events would result in an exchange of genes, but only an odd number of crossover events (a 50-50 chance between even and odd number of crossover events) would result in a recombinant product of meiotic crossover. A statistical interpretation of this is through the Haldane mapping function or the Kosambi mapping function, among others. A linkage map is created by finding the map distances between a number of traits that are present on the same chromosome, ideally avoiding having significant gaps between traits to avoid the inaccuracies that will occur due to the possibility of multiple recombination events.

Linkage map

A linkage map is a genetic map of a species or experimental population that shows the position of its known genes or genetic markers relative to each other in terms of recombination frequency, rather than as specific physical distance along each chromosome. Linkage mapping is critical for identifying the location of genes that cause genetic diseases.

A genetic map is a map based on the frequencies of recombination between markers during crossover of homologous chromosomes. The greater the frequency of recombination (segregation) between two genetic markers, the farther apart they are assumed to be. Conversely, the lower the frequency of recombination between the markers, the smaller the physical distance between them. Historically, the markers originally used were detectable phenotypes (enzyme production, eye color) derived from coding DNA sequences; eventually, confirmed or assumed noncoding DNA sequences such as microsatellites or those generating restriction fragment length polymorphisms (RFLPs) have been used.

Genetic maps help researchers to locate other markers, such as other genes by testing for genetic linkage of the already known markers.

A genetic map is **not** a physical map (such as a radiation reduced hybrid map) or gene map.

LOD score method for estimating recombination frequency

The **LOD score** (logarithm (base 10) of odds), developed by Newton E. Morton, is a statistical test often used for linkage analysis in human, animal, and plant populations. The LOD score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance. Positive LOD scores favor the presence of linkage, whereas negative LOD scores indicate that linkage is less likely. Computerized LOD score analysis is a simple way to analyze complex family pedigrees in order to determine the linkage between Mendelian traits (or between a trait and a marker, or two markers).

The method is described in greater detail by Strachan and Read. Briefly, it works as follows:

1. Establish a pedigree
2. Make a number of estimates of recombination frequency
3. Calculate a LOD score for each estimate
4. The estimate with the highest LOD score will be considered the best estimate

The LOD score is calculated as follows:

$$\begin{aligned} LOD = Z &= \log_{10} \frac{\text{probability of birth sequence with a given linkage value}}{\text{probability of birth sequence with no linkage}} \\ &= \log_{10} \frac{(1-\theta)^{NR} \times \theta^{R}}{0.5^{(NR+R)}} \end{aligned}$$

NR denotes the number of non-recombinant offspring, and R denotes the number of recombinant offspring. The reason 0.5 is used in the denominator is that any alleles that are completely unlinked (e.g. alleles on separate chromosomes) have a 50% chance of recombination, due to independent assortment.

Theta is the recombinant fraction, it is equal to R / (NR + R)

In practice, LOD scores are looked up in a table which lists LOD scores for various standard pedigrees and various values of recombination frequency.

By convention, a LOD score greater than 3.0 is considered evidence for linkage. A LOD score of +3 indicates 1000 to 1 odds that the linkage being observed did not occur by chance. On the other hand, a LOD score less than -2.0 is considered evidence to exclude linkage. Although it is very unlikely that a LOD score of 3 would be obtained from a single pedigree, the mathematical properties of the test allow data from a number of pedigrees to be combined by summing the LOD scores. It is important to keep in mind that this traditional cutoff of LOD>+3 is an arbitrary one and that the difference between certain types of linkage studies, particularly analyses of complex genetic traits with hundreds of markers, these criteria should probably be modified to a somewhat higher cutoff.

Recombination frequency

Recombination frequency (θ) is the frequency that a single chromosomal crossover will take place between two genes during meiosis. Double crossovers would turn into no recombination. In this case we cannot tell if crossovers took place. If the loci we're analysing are very close (less than 7 cM) a double crossover is very unlikely. When distances become higher, the likelihood of a double crossover increases. Recombination frequency is a measure of genetic linkage and is used in the creation of a genetic linkage map. A centimorgan (cM) is a unit that describes a recombination frequency of 1%. In this way we can measure the genetic distance between two loci, based upon their recombination frequency. This is a good estimate of the real distance. As the likelihood of a double crossover increases we systematically underestimate the genetic distance between two loci.

During meiosis, chromosomes assort randomly into gametes, such that the segregation of alleles of one gene is independent of alleles of another gene. This is stated in Mendel's Second Law and is known as **the law of independent assortment**. The law of independent assortment always holds true for genes that are located on different chromosomes, but for genes that are on the same chromosome, it does not always hold true.

As an example of independent assortment, consider the crossing of the pure-bred homozygote parental strain with genotype *AABB* with a different pure-bred strain with genotype *aabb*. A and a and B and b represent the alleles of genes A and B. Crossing these homozygous parental strains will result in F1 generation offspring with genotype AaBb. The F1 offspring AaBb produces gametes that are *AB*, *Ab*, *aB*, and *ab* with equal frequencies (25%) because the alleles of gene A assort independently of the alleles for gene B during meiosis. Note that 2 of the 4 gametes (50 %)—*Ab* and *aB*—were not present in the parental generation. These gametes represent **recombinant gametes**. Recombinant gametes are those gametes that differ from both of the haploid gametes that

made up the diploid cell. In this example, the recombination frequency is 50% since 2 of the 4 gametes were recombinant gametes.

The recombination frequency will be 50% when two genes are located on different chromosomes or when they are widely separated on the same chromosome. This is a consequence of independent assortment.

When two genes are close together on the same chromosome, they do not assort independently and are said to be linked. Whereas genes located on different chromosomes assort independently and have a recombination frequency of 50%, linked genes have a recombination frequency that is less than 50%.

As an example of linkage, consider the classic experiment by William Bateson and Reginald Punnett. They were interested in trait inheritance in the sweet pea and were studying two genes—the gene for flower colour (*P*, purple, and *p*, red) and the gene affecting the shape of pollen grains (*L*, long, and *l*, round). They crossed the pure lines *PPLL* and *ppll* and then self-crossed the resulting *PpLl* lines. According to Mendelian genetics, the expected phenotypes would occur in a 9:3:3:1 ratio of PL:Pl:pL:pl. To their surprise, they observed an increased frequency of PL and pl and a decreased frequency of Pl and pL (see table below).

Bateson and Punnett experiment

Phenotype and genotype	**Observed**	**Expected from 9:3:3:1 ratio**
Purple, long (*PpLl*)	284	216
Purple, round (*Ppll*)	21	72
Red, long (*ppLl*)	21	72
Red, round (*ppll*)	55	24

Their experiment revealed **linkage** between the *P* and *L* alleles and the *p* and *l* alleles. The frequency of *P* occurring together with *L* and with *p* occurring together with *l* is greater than that of the recombinant *Pl* and *pL*. The recombination frequency cannot be computed directly from this experiment, but intuitively it is less than 50%.

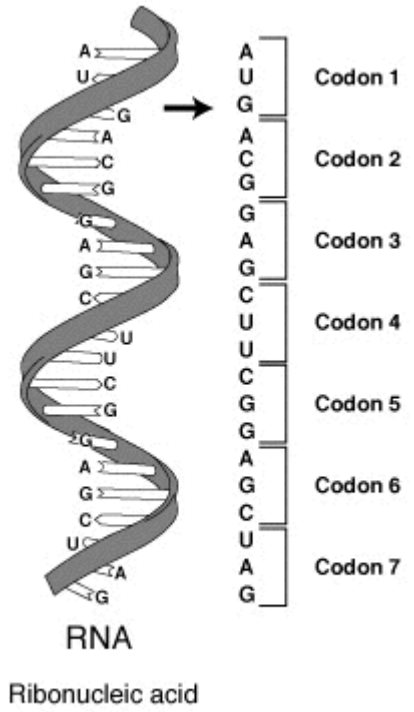
The progeny in this case received two dominant alleles linked on one chromosome (referred to as **coupling** or **cis arrangement**). However, after crossover, some progeny could have received one parental chromosome with a dominant allele for one trait (eg Purple) linked to a recessive allele for a second trait (eg round) with the opposite being true for the other parental chromosome (eg red and Long). This is referred to as **repulsion** or a **trans arrangement**. The phenotype here would still be purple and long but a test cross of this individual with the recessive parent would produce progeny with much greater proportion of the two crossover phenotypes. While such a problem may not seem likely from this example, unfavorable repulsion linkages do appear when breeding for disease resistance in some crops.

When two genes are located on the same chromosome, the chance of a crossover producing recombination between the genes is related to the distance between the two genes. Thus, the use of recombination frequencies has been used to develop **linkage maps** or **genetic maps**.

However, it is important to note that recombination frequency tends to underestimate the distance between two linked genes. This is because as the two genes are located further apart, the chance of double or even number of crossovers between them also increases. Double or even number of crossovers between the two genes results in them being cosegregated to the same gamete, yielding a parental progeny instead of the expected recombinant progeny.

Chapter- 8

Genetic Code



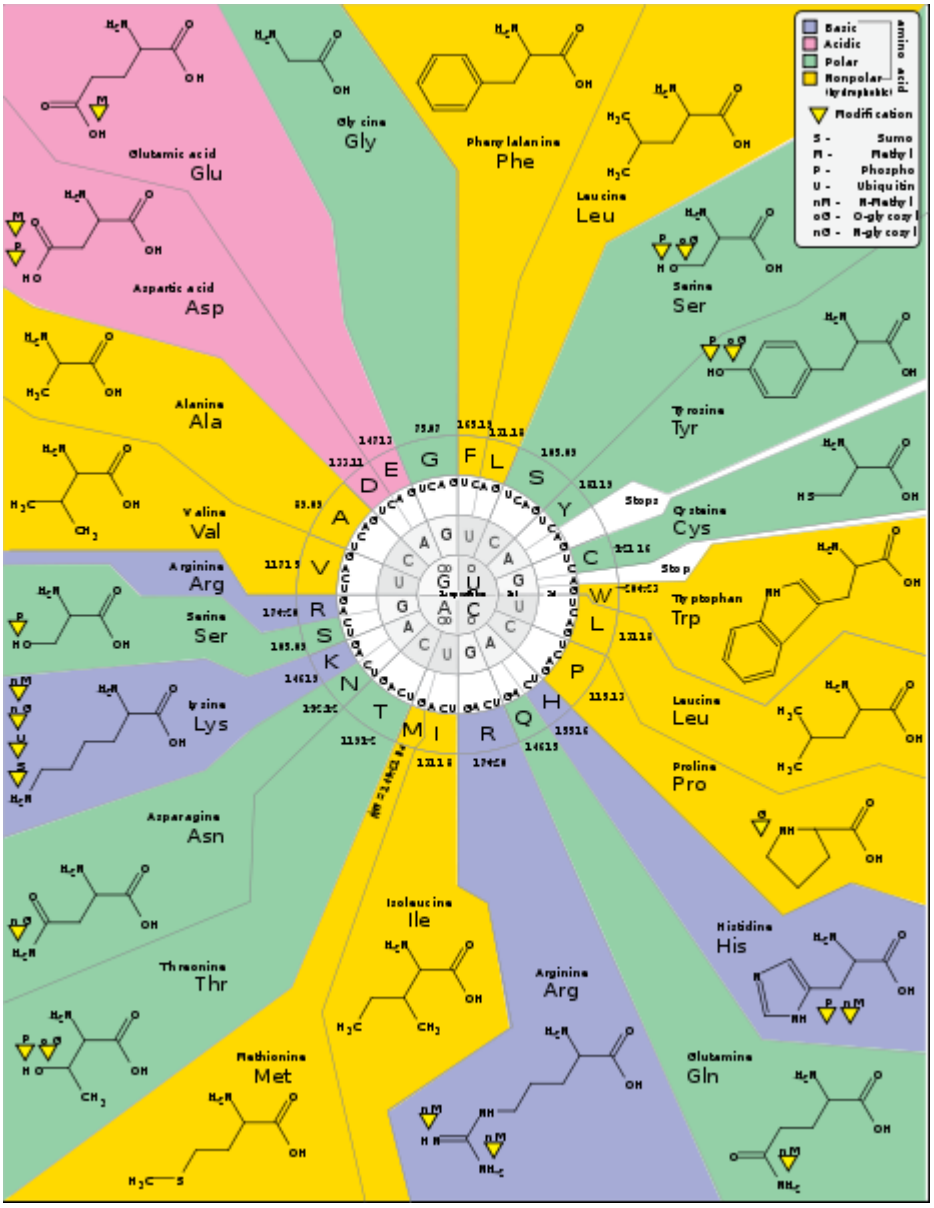
A series of codons in part of a mRNA molecule. Each codon consists of three nucleotides, usually representing a single amino acid.

The **genetic code** is the set of rules by which information encoded in genetic material (DNA or mRNA sequences) is translated into proteins (amino acid sequences) by living cells. The code defines a mapping between tri-nucleotide sequences, called codons, and amino acids. With some exceptions, a triplet codon in a nucleic acid sequence specifies a single amino acid. Because the vast majority of genes are encoded with exactly the same code, this particular code is often referred to as the canonical or standard genetic code, or simply *the* genetic code, though in fact there are many variant codes. For example,

protein synthesis in human mitochondria relies on a genetic code that differs from the standard genetic code.

Not all genetic information is stored using the genetic code. All organisms' DNA contains regulatory sequences, intergenic segments, and chromosomal structural areas that can contribute greatly to phenotype. Those elements operate under sets of rules that are distinct from the codon-to-amino acid paradigm underlying the genetic code.

Discovery



The genetic code

After the structure of DNA was discovered by James Watson and Francis Crick, who used the experimental evidence of Maurice Wilkins and Rosalind Franklin (among others), serious efforts to understand the nature of the encoding of proteins began. George Gamow postulated that a three-letter code must be employed to encode the 20 standard amino acids used by living cells to encode proteins, because 3 is the smallest integer n such that 4^n is at least 20.

The fact that codons consist of three DNA bases was first demonstrated in the Crick, Brenner et al. experiment. The first elucidation of a codon was done by Marshall Nirenberg and Heinrich J. Matthaei in 1961 at the National Institutes of Health. They used a cell-free system to translate a poly-uracil RNA sequence (i.e., UUUUU...) and discovered that the polypeptide that they had synthesized consisted of only the amino acid phenylalanine. They thereby deduced that the codon UUU specified the amino acid phenylalanine. This was followed by experiments in the laboratory of Severo Ochoa demonstrating that the poly-adenine RNA sequence (AAAAA...) coded for the polypeptide, poly-lysine. and the poly-cytosine RNA sequence (CCCCC...) coded for the polypeptide, poly-proline. Therefore the codon AAA specified the amino acid lysine, and the codon CCC specified the amino acid proline. Using different copolymers most of the remaining codons were then determined. Extending this work, Nirenberg and Philip Leder revealed the triplet nature of the genetic code and allowed the codons of the standard genetic code to be deciphered. In these experiments various combinations of mRNA were passed through a filter which contained ribosomes, the components of cells that translate RNA into protein. Unique triplets promoted the binding of specific tRNAs to the ribosome. Leder and Nirenberg were able to determine the sequences of 54 out of 64 codons in their experiments.

Subsequent work by Har Gobind Khorana identified the rest of the genetic code. Shortly thereafter, Robert W. Holley determined the structure of transfer RNA (tRNA), the adapter molecule that facilitates the process of translating RNA into protein. This work was based upon earlier studies by Severo Ochoa, who received the Nobel prize in 1959 for his work on the enzymology of RNA synthesis. In 1968, Khorana, Holley and Nirenberg received the Nobel Prize in Physiology or Medicine for their work.

Transfer of information via the genetic code

The genome of an organism is inscribed in DNA, or in the case of some viruses, RNA. The portion of the genome that codes for a protein or an RNA is called a gene. Those genes that code for proteins are composed of tri-nucleotide units called **codons**, each coding for a single amino acid. Each nucleotide sub-unit consists of a phosphate, deoxyribose sugar and one of the 4 nitrogenous nucleobases. The purine bases adenine (A) and guanine (G) are larger and consist of two aromatic rings. The pyrimidine bases cytosine (C) and thymine (T) are smaller and consist of only one aromatic ring. In the double-helix configuration, two strands of DNA are joined to each other by hydrogen bonds in an arrangement known as base pairing. These bonds almost always form between an adenine base on one strand and a thymine on the other strand and between a cytosine base on one strand and a guanine base on the other. This means that the number

of A and T residues will be the same in a given double helix, as will the number of G and C residues. In RNA, thymine (T) is replaced by uracil (U), and the deoxyribose is substituted by ribose.

Each protein-coding gene is transcribed into a template molecule of the related polymer RNA, known as messenger RNA or mRNA. This, in turn, is translated on the ribosome into an amino acid chain or polypeptide. The process of translation requires transfer RNAs specific for individual amino acids with the amino acids covalently attached to them, guanosine triphosphate as an energy source, and a number of translation factors. tRNAs have anticodons complementary to the codons in mRNA and can be "charged" covalently with amino acids at their 3' terminal CCA ends. Individual tRNAs are charged with specific amino acids by enzymes known as aminoacyl tRNA synthetases, which have high specificity for both their cognate amino acids and tRNAs. The high specificity of these enzymes is a major reason why the fidelity of protein translation is maintained.

There are $4^3 = 64$ different codon combinations possible with a triplet codon of three nucleotides; all 64 codons are assigned for either amino acids or stop signals during translation. If, for example, an RNA sequence, UUUAAACCC is considered and the reading frame starts with the first U (by convention, 5' to 3'), there are three codons, namely, UUU, AAA and CCC, each of which specifies one amino acid. This RNA sequence will be translated into an amino acid sequence, three amino acids long. A given amino acid may be encoded by between one and six different codon sequences. A comparison may be made with computer science, where the codon is similar to a word, which is the standard "chunk" for handling data (like one amino acid of a protein), and a nucleotide is similar to a bit, in that it is the smallest unit.

The standard genetic code is shown in the following tables. Table 1 shows what amino acid each of the 64 codons specifies. Table 2 shows what codons specify each of the 20 standard amino acids involved in translation. These are called forward and reverse codon tables, respectively. For example, the codon AAU represents the amino acid asparagine, and UGU and UGC represent cysteine (standard three-letter designations, Asn and Cys, respectively).

RNA codon table

nonpolar polar basic acidic (stop codon)

		2nd base							
		U		**C**		**A**		**G**	
1st base	**U**	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
		UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
		UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Ochre (*Stop*)	UGA	Opal (*Stop*)

	UUG	(Leu/L) Leucine	UCG	(Ser/S) Serine	UAG	Amber (*Stop*)	UGG	(Trp/W) Tryptophan
	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine
	CUC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
C	CUA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
	CUG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine
	AUC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
A	AUA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
	AUG[A]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine
	GUC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
G	GUA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
	GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

[A] The codon AUG both codes for methionine and serves as an initiation site: the first AUG in an mRNA's coding region is where translation into protein begins.

Inverse table

Ala/A	GCU, GCC, GCA, GCG	**Leu/L**	UUA, UUG, CUU, CUC, CUA, CUG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	**Lys/K**	AAA, AAG
Asn/N	AAU, AAC	**Met/M**	AUG
Asp/D	GAU, GAC	**Phe/F**	UUU, UUC
Cys/C	UGU, UGC	**Pro/P**	CCU, CCC, CCA, CCG
Gln/Q	CAA, CAG	**Ser/S**	UCU, UCC, UCA, UCG, AGU, AGC
Glu/E	GAA, GAG	**Thr/T**	ACU, ACC, ACA, ACG

Gly/G	GGU, GGC, GGA, GGG	**Trp/W**	UGG
His/H	CAU, CAC	**Tyr/Y**	UAU, UAC
Ile/I	AUU, AUC, AUA	**Val/V**	GUU, GUC, GUA, GUG
START	AUG	**STOP**	UAA, UGA, UAG

DNA codon table

The DNA codon table is essentially identical to that for RNA, but with U replaced by T.

Salient features

Sequence reading frame

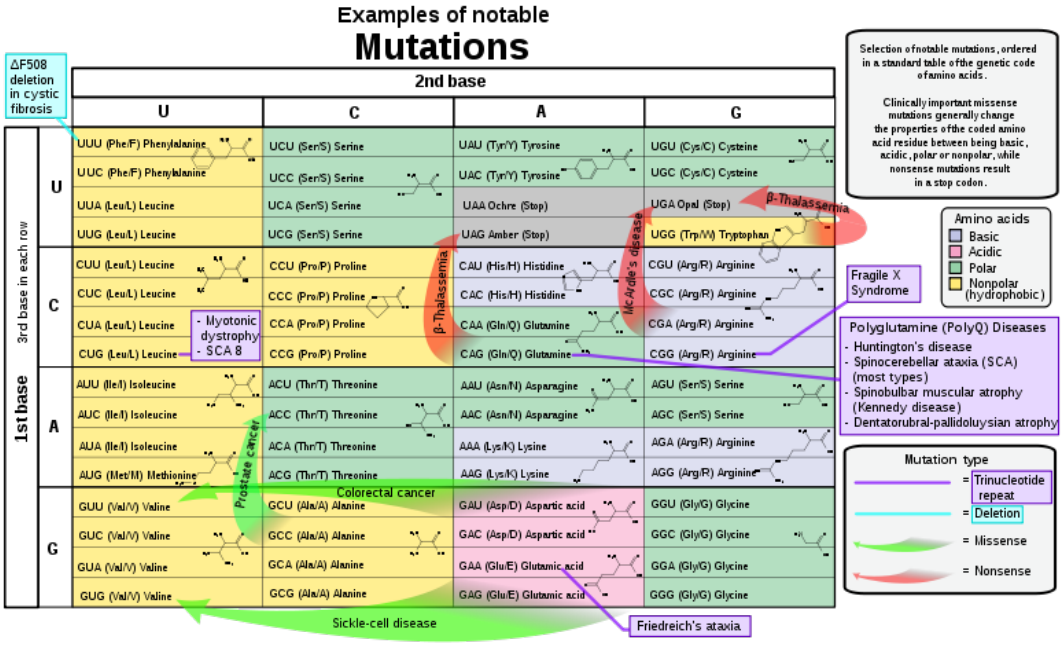
A codon is defined by the initial nucleotide from which translation starts. For example, the string GGGAAACCC, if read from the first position, contains the codons GGG, AAA and CCC; and, if read from the second position, it contains the codons GGA and AAC; if read starting from the third position, GAA and ACC. Every sequence can thus be read in three reading frames, each of which will produce a different amino acid sequence (in the given example, Gly-Lys-Pro, Gly-Asn, or Glu-Thr, respectively). With double-stranded DNA there are six possible reading frames, three in the forward orientation on one strand and three reverse on the opposite strand. The actual frame in which a protein sequence is translated is defined by a start codon, usually the first AUG codon in the mRNA sequence.

Start/stop codons

Translation starts with a chain initiation codon (start codon). Unlike stop codons, the codon alone is not sufficient to begin the process. Nearby sequences (such as the Shine-Dalgarno sequence in *E. coli*) and initiation factors are also required to start translation. The most common start codon is AUG which is read as methionine or, in bacteria, as formylmethionine. Alternative start codons (depending on the organism), include "GUG" or "UUG", which normally code for valine or leucine, respectively. However, when used as a start codon, these alternative start codons are translated as methionine or formylmethionine.

The three stop codons have been given names: UAG is *amber*, UGA is *opal* (sometimes also called *umber*), and UAA is *ochre*. "Amber" was named by discoverers Richard Epstein and Charles Steinberg after their friend Harris Bernstein, whose last name means "amber" in German. The other two stop codons were named "ochre" and "opal" in order to keep the "color names" theme. Stop codons are also called "termination" or "nonsense" codons and they signal release of the nascent polypeptide from the ribosome due to binding of release factors in the absence of cognate tRNAs with anticodons complementary to these stop signals.

Effect of mutations



Examples of notable mutations that can occur in humans

During the process of DNA replication, errors occasionally occur in the polymerization of the second strand. These errors, called mutations, can have an impact on the phenotype of an organism, especially if they occur within the protein coding sequence of a gene. Error rates are usually very low—1 error in every 10–100 million bases—due to the "proofreading" ability of DNA polymerases.

Missense mutations and nonsense mutations are examples of point mutations, which can cause genetic diseases such as sickle-cell disease and thalassemia respectively. Clinically important missense mutations generally change the properties of the coded amino acid residue between being basic, acidic polar or non-polar, whereas nonsense mutations result in a stop codon.

Mutations that disrupt the reading frame sequence by indels (insertions or deletions) of a non-multiple of 3 nucleotide bases are known as frameshift mutations. These mutations usually result in a completely different translation from the original, and are also very likely to cause a stop codon to be read, which truncates the creation of the protein. These mutations may impair the function of the resulting protein, and are thus rare in *in vivo* protein-coding sequences. One reason inheritance of frameshift mutations is rare is that if the protein being translated is essential for growth under the selective pressures the organism faces, absence of a functional protein may cause death before the organism is viable. Frameshift mutations may result in severe genetic diseases such as Tay-Sachs disease.

Although most mutations that change protein sequences are harmful or neutral, some mutations have a positive effect on an organism. These mutations may enable the mutant organism to withstand particular environmental stresses better than wild-type organisms, or reproduce more quickly. In these cases a mutation will tend to become more common in a population through natural selection. Viruses that use RNA as their genetic material have rapid mutation rates, which can be an advantage since these viruses will evolve constantly and rapidly, and thus evade the defensive responses of e.g. the human immune system. In large populations of asexually reproducing organisms, for example, E. coli, multiple beneficial mutations may co-occur, causing competition among them, this phenomenon is called clonal interference.

Degeneracy of the genetic code

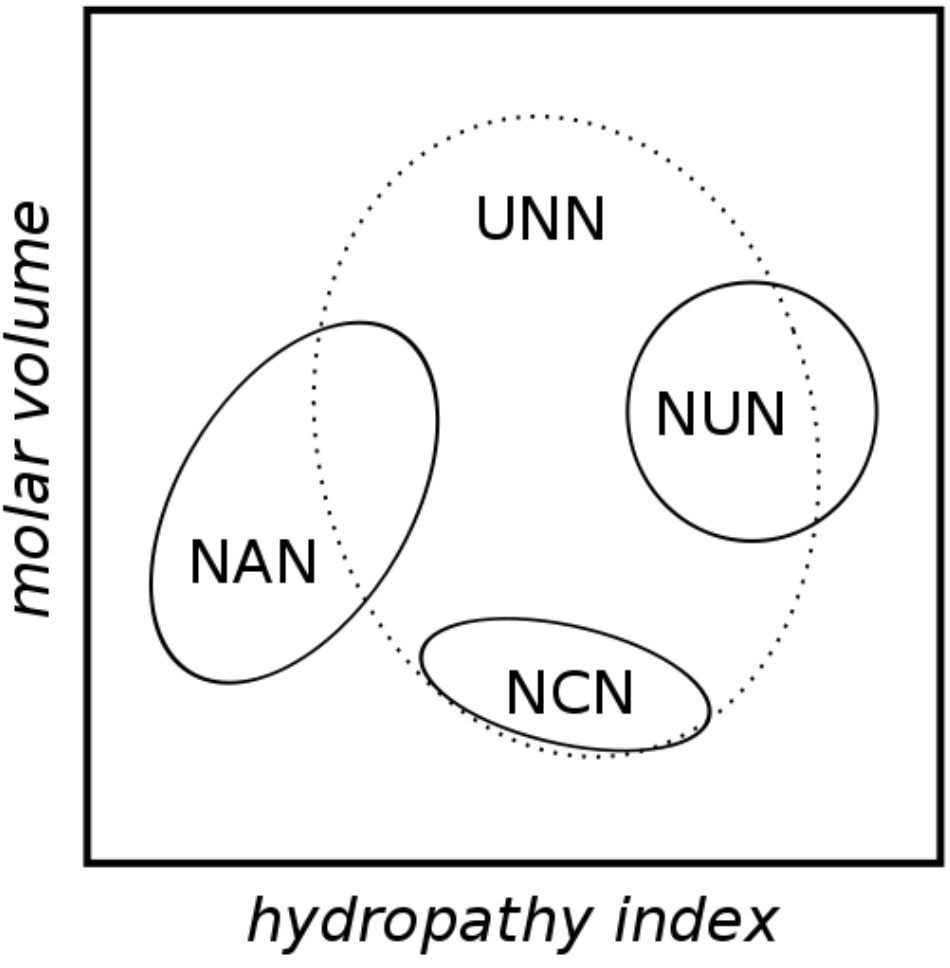
The genetic code has redundancy but no ambiguity. For example, although codons GAA and GAG both specify glutamic acid (redundancy), neither of them specifies any other amino acid (no ambiguity). The codons encoding one amino acid may differ in any of their three positions. For example the amino acid glutamic acid is specified by GAA and GAG codons (difference in the third position), the amino acid leucine is specified by UUA, UUG, CUU, CUC, CUA, CUG codons (difference in the first or third position), while the amino acid serine is specified by UCA, UCG, UCC, UCU, AGU, AGC (difference in the first, second or third position).

A position of a codon is said to be a fourfold degenerate site if any nucleotide at this position specifies the same amino acid. For example, the third position of the glycine codons (GGA, GGG, GGC, GGU) is a fourfold degenerate site, because all nucleotide substitutions at this site are synonymous; i.e., they do not change the amino acid. Only the third positions of some codons may be fourfold degenerate. A position of a codon is said to be a twofold degenerate site if only two of four possible nucleotides at this position specify the same amino acid. For example, the third position of the glutamic acid codons (GAA, GAG) is a twofold degenerate site. In twofold degenerate sites, the equivalent nucleotides are always either two purines (A/G) or two pyrimidines (C/U), so only transversional substitutions (purine to pyrimidine or pyrimidine to purine) in twofold degenerate sites are nonsynonymous. A position of a codon is said to be a non-degenerate site if any mutation at this position results in amino acid substitution. There is only one threefold degenerate site where changing to three of the four nucleotides may have no effect on the amino acid (depending on what it is changed to), while changing to the fourth possible nucleotide always results in an amino acid substitution. This is the third position of an isoleucine codon: AUU, AUC, or AUA all encode isoleucine, but AUG encodes methionine. In computation this position is often treated as a twofold degenerate site.

There are three amino acids encoded by six different codons: serine, leucine, and arginine. Only two amino acids are specified by a single codon. One of these is the amino-acid methionine, specified by the codon AUG, which also specifies the start of translation; the other is tryptophan, specified by the codon UGG. The degeneracy of the genetic code is what accounts for the existence of synonymous mutations.

Degeneracy results because there are more codons than encodable amino acids. For example, if there were two bases per codon, then only 16 amino acids could be coded for ($4^2=16$). Because at least 21 codes are required (20 amino acids plus stop), and the next largest number of bases is three, then 4^3 gives 64 possible codons, meaning that some degeneracy must exist.

These properties of the genetic code make it more fault-tolerant for point mutations. For example, in theory, fourfold degenerate codons can tolerate any point mutation at the third position, although codon usage bias restricts this in practice in many organisms; twofold degenerate codons can tolerate one out of the three possible point mutations at the third position. Since transition mutations (purine to purine or pyrimidine to pyrimidine mutations) are more likely than transversion (purine to pyrimidine or vice-versa) mutations, the equivalence of purines or that of pyrimidines at twofold degenerate sites adds a further fault-tolerance.



Grouping of codons by amino acid residue molar volume and hydropathy.

A practical consequence of redundancy is that some errors in the genetic code only cause a silent mutation or an error that would not affect the protein because the hydrophilicity

or hydrophobicity is maintained by equivalent substitution of amino acids; for example, a codon of NUN (where N = any nucleotide) tends to code for hydrophobic amino acids. NCN yields amino acid residues that are small in size and moderate in hydropathy; NAN encodes average size hydrophilic residues. These tendencies may result from the shared ancestry of the aminoacyl tRNA synthetases related to these codons.

Even so, single point mutations can still cause dysfunctional proteins. For example, a mutated hemoglobin gene causes sickle-cell disease. In the mutant hemoglobin a hydrophilic glutamate (Glu) is substituted by the hydrophobic valine (Val), that is, GAA or GAG becomes GUA or GUG. The substitution of glutamate by valine reduces the solubility of β-globin which causes hemoglobin to form linear polymers linked by the hydrophobic interaction between the valine groups causing sickle-cell deformation of erythrocytes. Sickle-cell disease is generally not caused by a *de novo* mutation. Rather it is selected for in malarial regions (in a way similar to thalassemia), as heterozygous people have some resistance to the malarial *Plasmodium* parasite (heterozygote advantage).

These variable codes for amino acids are allowed because of modified bases in the first base of the anticodon of the tRNA, and the base-pair formed is called a wobble base pair. The modified bases include inosine and the Non-Watson-Crick U-G basepair.

Variations to the standard genetic code

While slight variations on the standard code had been predicted earlier, none were discovered until 1979, when researchers studying human mitochondrial genes discovered they used an alternative code. Many slight variants have been discovered since, including various alternative mitochondrial codes, as well as small variants such as *Mycoplasma* translating the codon UGA as tryptophan and *Candida* species translating CUG as a serine rather than a leucine. In bacteria and archaea, GUG and UUG are common start codons. However, in rare cases, certain specific proteins may use alternative initiation (start) codons not normally used by that species.

In certain proteins, non-standard amino acids are substituted for standard stop codons, depending upon associated signal sequences in the messenger RNA: UGA can code for selenocysteine and UAG can code for pyrrolysine as discussed in the relevant articles. Selenocysteine is now viewed as the 21st amino acid, and pyrrolysine is viewed as the 22nd.

Notwithstanding these differences, all known codes have strong similarities to each other, and the coding mechanism is the same for all organisms: three-base codons, tRNA, ribosomes, reading the code in the same direction and translating the code three letters at a time into sequences of amino acids.

Expanded genetic code

Since 2001, 40 non-natural amino acids have been added into protein by creating a unique codon (recoding) and a corresponding transfer-RNA:aminoacyl – tRNA-synthetase pair to encode it with diverse physicochemical and biological properties in order to be used as a tool to exploring protein structure and function or to create novel or enhanced proteins.

Theories on the origin of the genetic code

Despite the minor variations that exist, the genetic code used by all known forms of life is nearly universal. However, there are a huge number of possible genetic codes. If amino acids are randomly associated with triplet codons, there will be 1.5×10^{84} possible genetic codes.

Phylogenetic analysis of transfer RNA suggests that tRNA molecules evolved before the present set of aminoacyl-tRNA synthetases.

Theoretically the genetic code could be completely random (a "frozen accident"), completely non-random (optimal) or a combination of random and nonrandom. There are sufficient data to refute the first possibility. For a start, a quick view on the table of the genetic code already shows a clustering of amino acid assignments. Furthermore, amino acids that share the same biosynthetic pathway tend to have the same first base in their codons, and amino acids with similar physical properties tend to have similar codons.

There are four themes running through the many theories that seek to explain the evolution of the genetic code (and hence the origin of these patterns):

- **Chemical principles** govern specific RNA interaction with amino acids. Aptamer experiments showed that some amino acids have a selective chemical affinity for the base triplets that code for them. Recent experiments show that of the 8 amino acids tested, 6 show some RNA triplet-amino acid association. This has been called the stereochemical code. The stereochemical code could have created an ancient core of assignments. The current complex translation mechanism involving tRNA and associated enzymes may be a later development, and that originally, protein sequences were directly templated on base sequences.
- **Biosynthetic expansion**. The standard modern genetic code grew from a simpler earlier code through a process of "biosynthetic expansion". Here the idea is that primordial life "discovered" new amino acids (e.g., as by-products of metabolism) and later back-incorporated some of these into the machinery of genetic coding. Although much circumstantial evidence has been found to suggest that fewer different amino acids were used in the past than today, precise and detailed hypotheses about exactly which amino acids entered the code in exactly what order have proved far more controversial.
- **Natural selection** has led to codon assignments of the genetic code that minimize the effects of mutations. A recent hypothesis suggests that the triplet code was

derived from codes that used longer than triplet codons. Longer than triplet decoding has higher degree of codon redundancy and is more error resistant than the triplet decoding. This feature could allow accurate decoding in the absence of highly complex translational machinery such as the ribosome.

- **Information channels:** Information-theoretic approaches see the genetic code as an error-prone information channel. The inherent noise (i.e. errors) in the channel poses the organism with a fundamental question: how to construct a genetic code that can withstand the impact of noise while accurately and efficiently translating information? These "rate-distortion" models suggest that the genetic code originated as a result of the interplay of the three conflicting evolutionary forces: the needs for diverse amino-acids, for error-tolerance and for minimal cost of resources. The code emerges at a coding transition when the mapping of codons to amino-acids becomes nonrandom. The emergence of the code is governed by the topology defined by the probable errors and is related to the map coloring problem.

Chapter- 9

Regulation of Gene Expression

Regulation of gene expression (or **gene regulation**) includes the processes that cells and viruses use to regulate the way that the information in genes is turned into gene products. Although a functional gene product may be an RNA or a protein, the majority of known mechanisms regulate protein coding genes. Any step of the gene's expression may be modulated, from DNA-RNA transcription to the post-translational modification of a protein.

Gene regulation is essential for viruses, prokaryotes and eukaryotes as it increases the versatility and adaptability of an organism by allowing the cell to express protein when needed. The first discovered example of a gene regulation system was the lac operon, discovered by Jacques Monod, in which protein involved in lactose metabolism are expressed by *E. coli* only in the presence of lactose and absence of glucose.

Furthermore, gene regulation drives the processes of cellular differentiation and morphogenesis, leading to the creation of different cell types in multicellular organisms where the different types of cells may possess different gene expression profiles though they all possess the same genome sequence.

Regulated stages of gene expression

Any step of gene expression may be modulated, from the DNA-RNA transcription step to post-translational modification of a protein. The following is a list of stages where gene expression is regulated, the most extensively utilised point is Transcription Initiation:

- Chromatin domains
- Transcription
- Post-transcriptional modification
- RNA transport
- Translation
- mRNA degradation

Modification of DNA

In eukaryotes, the accessibility of large regions of DNA can depend on its chromatin structure, which can be altered as a result of histone modifications directed by DNA methylation, ncRNA, or DNA-binding protein.

Chemical

Methylation of DNA is a common method of gene silencing. DNA is typically methylated by methyltransferase enzymes on cytosine nucleotides in a CpG dinucleotide sequence (also called "CpG islands" when densely clustered). Analysis of the pattern of methylation in a given region of DNA (which can be a promoter) can be achieved through a method called bisulfite mapping. Methylated cytosine residues are unchanged by the treatment, whereas unmethylated ones are changed to uracil. The differences are analyzed by DNA sequencing or by methods developed to quantify SNPs, such as Pyrosequencing (Biotage) or MassArray (Sequenom), measuring the relative amounts of C/T at the CG dinucleotide. Abnormal methylation patterns are thought to be involved in oncogenesis.

Structural

Transcription of DNA is dictated by its structure. In general, the density of its packing is indicative of the frequency of transcription. Octameric protein complexes called nucleosomes are responsible for the amount of supercoiling of DNA, and these complexes can be temporarily modified by processes such as phosphorylation or more permanently modified by processes such as methylation. Such modifications are considered to be responsible for more or less permanent changes in gene expression levels.

Histone acetylation is also an important process in transcription. Histone acetyltransferase enzymes (HATs) such as CREB-binding protein also dissociate the DNA from the histone complex, allowing transcription to proceed. Often, DNA methylation and histone deacetylation work together in gene silencing. The combination of the two seems to be a signal for DNA to be packed more densely, lowering gene expression.

Regulation of transcription

Regulation of transcription controls when transcription occurs and how much RNA is created. Transcription of a gene by RNA polymerase can be regulated by at least five mechanisms:

- **Specificity factors** alter the specificity of RNA polymerase for a given promoter or set of promoters, making it more or less likely to bind to them (i.e., sigma factors used in prokaryotic transcription).

- **Repressors** bind to non-coding sequences on the DNA strand that are close to or overlapping the promoter region, impeding RNA polymerase's progress along the strand, thus impeding the expression of the gene.
- **General transcription factors** position RNA polymerase at the start of a protein-coding sequence and then release the polymerase to transcribe the mRNA.
- **Activators** enhance the interaction between RNA polymerase and a particular promoter, encouraging the expression of the gene. Activators do this by increasing the attraction of RNA polymerase for the promoter, through interactions with subunits of the RNA polymerase or indirectly by changing the structure of the DNA.
- **Enhancers** are sites on the DNA helix that are bound to by activators in order to loop the DNA bringing a specific promoter to the initiation complex. Enhancers are much more common in eukaryote than prokaryotes, where only a few examples exist (to date).

Post-transcriptional regulation

After the DNA is transcribed and mRNA is formed, there must be some sort of regulation on how much the mRNA is translated into proteins. Cells do this by modulating the capping, splicing, addition of a Poly(A) Tail, the sequence-specific nuclear export rates, and, in several contexts, sequestration of the RNA transcript. These processes occur in eukaryotes but not in prokaryotes. This modulation is a result of a protein or transcript that, in turn, is regulated and may have an affinity for certain sequences.

Regulation of translation

The translation of mRNA can also be controlled by a number of mechanisms, mostly at the level of initiation. Recruitment of the small ribosomal subunit can indeed be modulated by mRNA secondary structure, antisense RNA binding, or protein binding. In both prokaryotes and eukaryotes, a large number of RNA binding proteins exist, which often are directed to their target sequence by the secondary structure of the transcript, which may change depending on certain conditions, such as temperature or presence of a ligand (aptamer). Some transcripts act as ribozymes and self-regulate their expression.

Examples of gene regulation

- Enzyme induction is a process in which a molecule (e.g., a drug) induces (i.e., initiates or enhances) the expression of an enzyme.
- The induction of heat shock proteins in the fruit fly *Drosophila melanogaster*.
- The Lac operon is an interesting example of how gene expression can be regulated.
- Viruses, despite having only a few genes, possess mechanisms to regulate their gene expression, typically into an early and late phase, using collinear systems regulated by anti-terminators (lambda phage) or splicing modulators (HIV).

Developmental biology

A large number of studied regulatory systems come from developmental biology. Examples include:

- The colinearity of the Hox gene cluster with their nested antero-posterior patterning
- It has been speculated that pattern generation of the hand (digits - interdigits) The gradient of Sonic hedgehog (secreted inducing factor) from the zone of polarizing activity in the limb, which creates a gradient of active Gli3, which activates Gremlin, which inhibits BMPs also secreted in the limb, resulting in the formation of an alternating pattern of activity as a result of this reaction-diffusion system.
- Somitogenesis is the creation of segments (somites) from a uniform tissue (Pre-somitic Mesoderm, PSM). They are formed sequentially from anterior to posterior. This is achieved in amniotes possibly by means of two opposing gradients, Retinoic acid in the anterior (wavefront) and Wnt and Fgf in the posterior, coupled to an oscillating pattern (segmentation clock) composed of FGF + Notch and Wnt in antiphase.
- Sex determination in the soma of a Drosophila requires the sensing of the ratio of autosomal genes to sex chromosome-encoded genes, which results in the production of sexless splicing factor in females, resulting in the female isoform of doublesex.

Circuitry

Up-regulation and down-regulation

Up-regulation is a process that occurs within a cell triggered by a signal (originating internal or external to the cell), which results in increased expression of one or more genes and as a result the protein(s) encoded by those genes. On the converse, **down-regulation** is a process resulting in decreased gene and corresponding protein expression.

- Up-regulation occurs, for example, when a cell is deficient in some kind of receptor. In this case, more receptor protein is synthesized and transported to the membrane of the cell and, thus, the sensitivity of the cell is brought back to normal, reestablishing homeostasis.

- Down-regulation occurs, for example, when a cell is overstimulated by a neurotransmitter, hormone, or drug for a prolonged period of time, and the expression of the receptor protein is decreased in order to protect the cell.

Inducible vs. repressible systems

Gene Regulation can be summarized as how they respond:

- Inducible systems - An inducible system is off unless there is the presence of some molecule (called an inducer) that allows for gene expression. The molecule is said to "induce expression". The manner by which this happens is dependent on the control mechanisms as well as differences between prokaryotic and eukaryotic cells.

- Repressible systems - A repressible system is on except in the presence of some molecule (called a corepressor) that suppresses gene expression. The molecule is said to "repress expression". The manner by which this happens is dependent on the control mechanisms as well as differences between prokaryotic and eukaryotic cells.

Theoretical circuits

- Repressor/Inducer: an activation of a sensor results in the change of expression of a gene
- negative feedback: the gene product downregulates its own production directly or indirectly, which can result in
    - keeping transcript levels constant/proportional to a factor
    - inhibition of run-away reactions when coupled with a positive feedback loop
    - creating an oscillator by taking advantage in the time delay of transcription and translation, given that the mRNA and protein half-life is shorter
- positive feedback: the gene product upregulates its own production directly or indirectly, which can result in
    - signal amplification
    - bistable switches when two genes inhibit each other and have both positive feedback
    - pattern generation

Methods

In general, most experiments investigating differential expression used whole cell extracts of RNA, called steady-state levels, to determine which genes changed and by how much they did. These are, however, not informative of where the regulation has occurred and may actually mask conflicting regulatory processess, but it is still the most commonly analysed (QPCR and DNA microarray).

When studying gene expression, there are several methods to look at the various stages. In eukaryotes these include:

- The chromatin conformation of the region can be determined by ChIP-chip analysis by pulling down RNA Polymerase II, Histone 3 modifications, Trithorax-group protein, Polycomb-group protein, or any other DNA-binding element to which a good antibody is available.

- Epistatic interactions can be investigated by synthetic genetic array analysis
- Due to post-transcriptional regulation, transcription rates and total RNA levels differ significantly. To measure the transcription rates nuclear run-on assays can be done and newer high-throughput methods are being developed, using thiol labelling instead of radioactivity.
- Only 5% of the RNA polymerised in the nucleus actually exists, and not only introns, abortive products, and non-sense transcripts are degradated. Therefore, the differences in nuclear and cytoplasmic levels can be see by separating the two fractions by gentle lysis.
- Alternative splicing can be analysed with a splicing array or with a tiling array.
- All in vivo RNA is complexed as RNPs. The quantity of transcripts bound to specific protein can be also analysed by RIP-Chip. For example, DCP2 will give an indication of sequestered protein; ribosome-bound gives and indication of transcripts active in transcription (although it should be noted that a more dated method, called polysome fractionation, is still popular in some labs)
- Protein levels can be analysed by Mass spectrometry, which can be compared only to QPCR data, as microarray data is relative and not absolute.
- RNA and protein degradation rates are measured by means of transcription inhibitors (actinomycin D or α-amanitin) or translation inhibitors (Cycloheximide), respectively.

Chapter- 10

Mutation

In molecular biology and genetics, **mutations** are changes in a genomic sequence: the DNA sequence of a cell's genome or the DNA or RNA sequence of a virus. Mutations are caused by radiation, viruses, transposons and mutagenic chemicals, as well as errors that occur during meiosis or DNA replication. They can also be induced by the organism itself, by cellular processes such as hypermutation.

Mutation can result in several different types of change in DNA sequences; these can either have no effect, alter the product of a gene, or prevent the gene from functioning properly or completely. Studies in the fly *Drosophila melanogaster* suggest that if a mutation changes a protein produced by a gene, this will probably be harmful, with about 70 percent of these mutations having damaging effects, and the remainder being either neutral or weakly beneficial. Due to the damaging effects that mutations can have on genes, organisms have mechanisms such as DNA repair to remove mutations.

Therefore, the optimal mutation rate for a species is a trade-off between costs of a high mutation rate, such as deleterious mutations, and the metabolic costs of maintaining systems to reduce the mutation rate, such as DNA repair enzymes. Viruses that use RNA as their genetic material have rapid mutation rates, which can be an advantage since these viruses will evolve constantly and rapidly, and thus evade the defensive responses of e.g. the human immune system.

Description

Mutations can involve large sections of DNA becoming duplicated, usually through genetic recombination. These duplications are a major source of raw material for evolving new genes, with tens to hundreds of genes duplicated in animal genomes every million years. Most genes belong to larger families of genes of shared ancestry. Novel genes are produced by several methods, commonly through the duplication and mutation of an ancestral gene, or by recombining parts of different genes to form new combinations with new functions.

Here, domains act as modules, each with a particular and independent function, that can be mixed together to produce genes encoding new proteins with novel properties. For example, the human eye uses four genes to make structures that sense light: three for color vision and one for night vision; all four arose from a single ancestral gene. Another advantage of duplicating a gene (or even an entire genome) is that this increases redundancy; this allows one gene in the pair to acquire a new function while the other copy performs the original function. Other types of mutation occasionally create new genes from previously noncoding DNA.

Changes in chromosome number may involve even larger mutations, where segments of the DNA within chromosomes break and then rearrange. For example, two chromosomes in the *Homo* genus fused to produce human chromosome 2; this fusion did not occur in the lineage of the other apes, and they retain these separate chromosomes. In evolution, the most important role of such chromosomal rearrangements may be to accelerate the divergence of a population into new species by making populations less likely to interbreed, and thereby preserving genetic differences between these populations.

Sequences of DNA that can move about the genome, such as transposons, make up a major fraction of the genetic material of plants and animals, and may have been important in the evolution of genomes. For example, more than a million copies of the Alu sequence are present in the human genome, and these sequences have now been recruited to perform functions such as regulating gene expression. Another effect of these mobile DNA sequences is that when they move within a genome, they can mutate or delete existing genes and thereby produce genetic diversity.

A mutation has caused this garden moss rose to produce flowers of different colors. This is a somatic mutation that may also be passed on in the germ line.

In multicellular organisms with dedicated reproductive cells, mutations can be subdivided into germ line mutations, which can be passed on to descendants through their reproductive cells, and somatic mutations (also called acquired mutations), which involve cells outside the dedicated reproductive group and which are not usually transmitted to descendants. If the organism can reproduce asexually through mechanisms such as cuttings or budding the distinction can become blurred.

For example, plants can sometimes transmit somatic mutations to their descendants asexually or sexually where flower buds develop in somatically mutated parts of plants. A new mutation that was not inherited from either parent is called a *de novo* mutation.

The source of the mutation is unrelated to the consequence, although the consequences are related to which cells were mutated.

Nonlethal mutations accumulate within the gene pool and increase the amount of genetic variation. The abundance of some genetic changes within the gene pool can be reduced by natural selection, while other "more favorable" mutations may accumulate and result in adaptive changes.

For example, a butterfly may produce offspring with new mutations. The majority of these mutations will have no effect; but one might change the color of one of the butterfly's offspring, making it harder (or easier) for predators to see. If this color change is advantageous, the chance of this butterfly surviving and producing its own offspring are a little better, and over time the number of butterflies with this mutation may form a larger percentage of the population.

Neutral mutations are defined as mutations whose effects do not influence the fitness of an individual. These can accumulate over time due to genetic drift. It is believed that the overwhelming majority of mutations have no significant effect on an organism's fitness. Also, DNA repair mechanisms are able to mend most changes before they become permanent mutations, and many organisms have mechanisms for eliminating otherwise permanently mutated somatic cells.

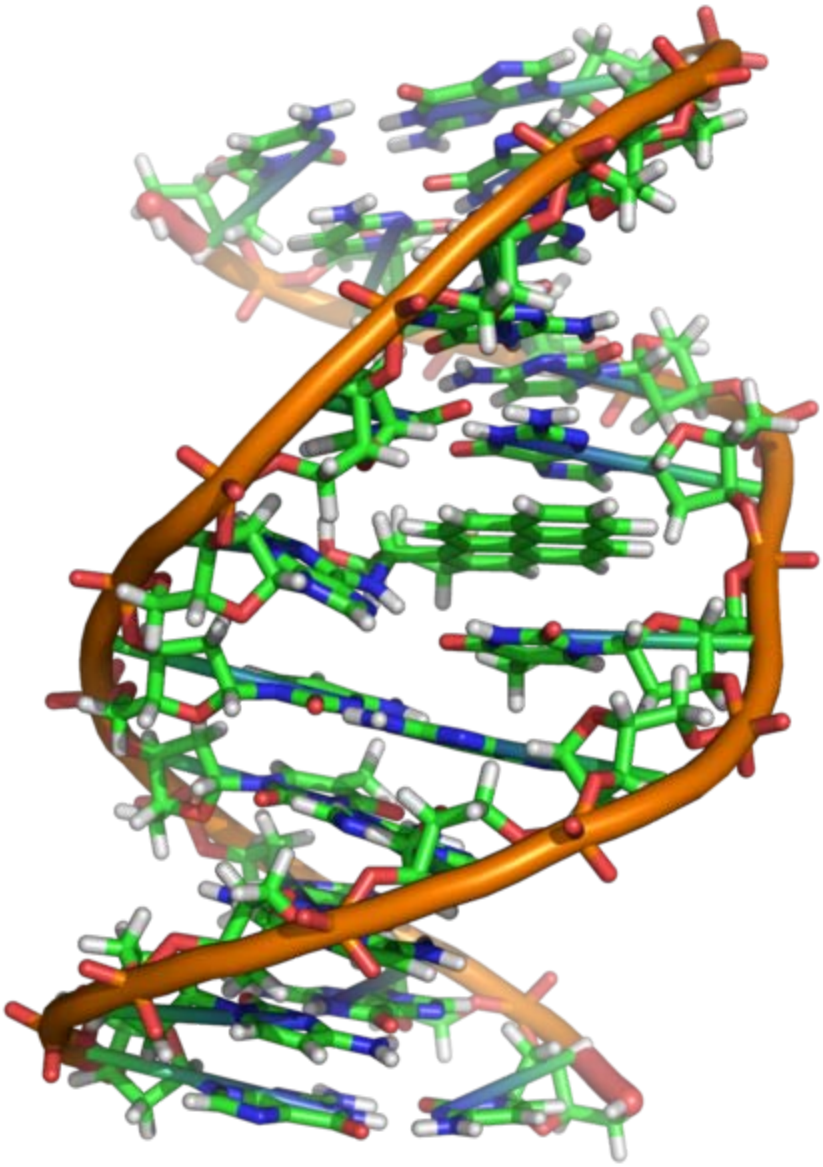
Mutation is generally accepted by biologists as the mechanism by which natural selection acts, generating advantageous new traits that survive and multiply in offspring as well as disadvantageous traits, in less fit offspring, that tend to die out.

Causes

Two classes of mutations are spontaneous mutations (molecular decay) and induced mutations caused by mutagens.

Spontaneous mutations on the molecular level can be caused by:

- Tautomerism – A base is changed by the repositioning of a hydrogen atom, altering the hydrogen bonding pattern of that base resulting in incorrect base pairing during replication.
- Depurination – Loss of a purine base (A or G) to form an apurinic site (AP site).
- Deamination – Hydrolysis changes a normal base to an atypical base containing a keto group in place of the original amine group. Examples include C → U and A → HX (hypoxanthine), which can be corrected by DNA repair mechanisms; and 5MeC (5-methylcytosine) → T, which is less likely to be detected as a mutation because thymine is a normal DNA base.
- Slipped strand mispairing - Denaturation of the new strand from the template during replication, followed by renaturation in a different spot ("slipping"). This can lead to insertions or deletions.



A covalent adduct between benzo[*a*]pyrene, the major mutagen in tobacco smoke, and DNA

Induced mutations on the molecular level can be caused by:

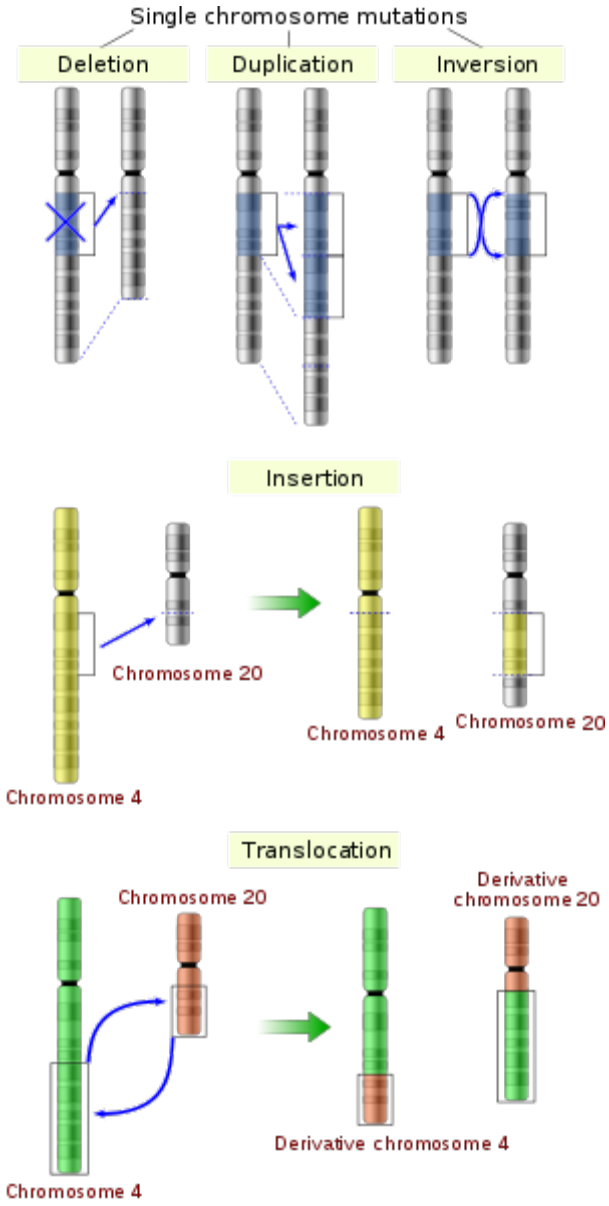
- Chemicals
  - Hydroxylamine NH_2OH
  - Base analogs (e.g. BrdU)
  - Alkylating agents (e.g. *N*-ethyl-*N*-nitrosourea) These agents can mutate both replicating and non-replicating DNA. In contrast, a base analog can only mutate the DNA when the analog is incorporated in replicating the DNA. Each of these classes of chemical mutagens has certain effects that then lead to transitions, transversions, or deletions.

    - Agents that form DNA adducts (e.g. ochratoxin A metabolites)
    - DNA intercalating agents (e.g. ethidium bromide)
    - DNA crosslinkers
    - Oxidative damage
    - Nitrous acid converts amine groups on A and C to diazo groups, altering their hydrogen bonding patterns which leads to incorrect base pairing during replication.
- Radiation
    - Ultraviolet radiation (nonionizing radiation). Two nucleotide bases in DNA – cytosine and thymine – are most vulnerable to radiation that can change their properties. UV light can induce adjacent pyrimidine bases in a DNA strand to become covalently joined as a pyrimidine dimer. UV radiation, particularly longer-wave UVA, can also cause oxidative damage to DNA.
    - Ionizing radiation
    - Radioactive decay, such as ^{14}C in DNA
- Viral infections

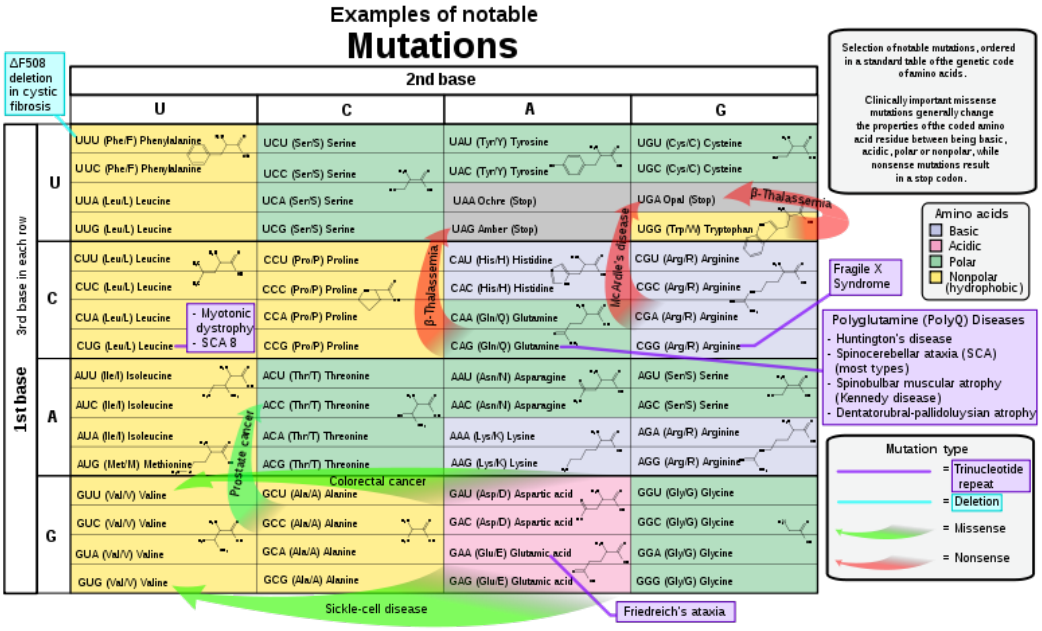
DNA has so-called hotspots, where mutations occur up to 100 times more frequently than the normal mutation rate. A hotspot can be at an unusual base, e.g., 5-methylcytosine.

Mutation rates also vary across species. Evolutionary biologists have theorized that higher mutation rates are beneficial in some situations, because they allow organisms to evolve and therefore adapt more quickly to their environments. For example, repeated exposure of bacteria to antibiotics, and selection of resistant mutants, can result in the selection of bacteria that have a much higher mutation rate than the original population (mutator strains).

Classification of mutation types



Illustrations of five types of chromosomal mutations



Selection of disease-causing mutations, in a standard table of the genetic code of amino acids.

By effect on structure

The sequence of a gene can be altered in a number of ways. Gene mutations have varying effects on health depending on where they occur and whether they alter the function of essential proteins. Mutations in the structure of genes can be classified as:

- Small-scale mutations, such as those affecting a small gene in one or a few nucleotides, including:
  - **Point mutations**, often caused by chemicals or malfunction of DNA replication, exchange a single nucleotide for another. These changes are classified as transitions or transversions. Most common is the transition that exchanges a purine for a purine (A ↔ G) or a pyrimidine for a pyrimidine, (C ↔ T). A transition can be caused by nitrous acid, base mis-pairing, or mutagenic base analogs such as 5-bromo-2-deoxyuridine (BrdU). Less common is a transversion, which exchanges a purine for a pyrimidine or a pyrimidine for a purine (C/T ↔ A/G). An example of a transversion is adenine (A) being converted into a cytosine (C). A point mutation can be reversed by another point mutation, in which the nucleotide is changed back to its original state (true reversion) or by second-site reversion (a complementary mutation elsewhere that results in regained gene functionality). Point mutations that occur within the protein coding region of a gene may be classified into three kinds, depending upon what the erroneous codon codes for:
    - Silent mutations: which code for the same amino acid.
    - Missense mutations: which code for a different amino acid.

        - Nonsense mutations: which code for a stop and can truncate the protein.
    - **Insertions** add one or more extra nucleotides into the DNA. They are usually caused by transposable elements, or errors during replication of repeating elements (e.g. AT repeats). Insertions in the coding region of a gene may alter splicing of the mRNA (splice site mutation), or cause a shift in the reading frame (frameshift), both of which can significantly alter the gene product. Insertions can be reverted by excision of the transposable element.
    - **Deletions** remove one or more nucleotides from the DNA. Like insertions, these mutations can alter the reading frame of the gene. They are generally irreversible: though exactly the same sequence might theoretically be restored by an insertion, transposable elements able to revert a very short deletion (say 1–2 bases) in *any* location are either highly unlikely to exist or do not exist at all. Note that a deletion is not the exact opposite of an insertion: the former is quite random while the latter consists of a specific sequence inserting at locations that are not entirely random or even quite narrowly defined.
- Large-scale mutations in chromosomal structure, including:
    - **Amplifications** (or gene duplications) leading to multiple copies of all chromosomal regions, increasing the dosage of the genes located within them.
    - **Deletions** of large chromosomal regions, leading to loss of the genes within those regions.
    - Mutations whose effect is to juxtapose previously separate pieces of DNA, potentially bringing together separate genes to form functionally distinct fusion genes (e.g. bcr-abl). These include:
        - **Chromosomal translocations**: interchange of genetic parts from nonhomologous chromosomes.
        - **Interstitial deletions**: an intra-chromosomal deletion that removes a segment of DNA from a single chromosome, thereby apposing previously distant genes. For example, cells isolated from a human astrocytoma, a type of brain tumor, were found to have a chromosomal deletion removing sequences between the "fused in glioblastoma" (fig) gene and the receptor tyrosine kinase "ros", producing a fusion protein (FIG-ROS). The abnormal FIG-ROS fusion protein has constitutively active kinase activity that causes oncogenic transformation (a transformation from normal cells to cancer cells).
        - **Chromosomal inversions**: reversing the orientation of a chromosomal segment.
    - **Loss of heterozygosity**: loss of one allele, either by a deletion or recombination event, in an organism that previously had two different alleles.

By effect on function

- **Loss-of-function mutations** are the result of gene product having less or no function. When the allele has a complete loss of function (null allele) it is often called an **amorphic mutation**. Phenotypes associated with such mutations are most often recessive. Exceptions are when the organism is haploid, or when the reduced dosage of a normal gene product is not enough for a normal phenotype (this is called haploinsufficiency).
- **Gain-of-function mutations** change the gene product such that it gains a new and abnormal function. These mutations usually have dominant phenotypes. Often called a neomorphic mutation.
- **Dominant negative mutations** (also called **antimorphic mutations**) have an altered gene product that acts antagonistically to the wild-type allele. These mutations usually result in an altered molecular function (often inactive) and are characterised by a dominant or semi-dominant phenotype. In humans, Marfan syndrome is an example of a dominant negative mutation occurring in an autosomal dominant disease. In this condition, the defective glycoprotein product of the fibrillin gene (FBN1) antagonizes the product of the normal allele.
- **Lethal mutations** are mutations that lead to the death of the organisms which carry the mutations.
- A **back mutation** or **reversion** is a point mutation that restores the original sequence and hence the original phenotype.

By effect on fitness

In applied genetics it is usual to speak of mutations as either harmful or beneficial.

- A **harmful mutation** is a mutation that decreases the fitness of the organism.
- A **beneficial mutation** is a mutation that increases fitness of the organism, or which promotes traits that are desirable.

In theoretical population genetics, it is more usual to speak of such mutations as deleterious or advantageous. In the neutral theory of molecular evolution, genetic drift is the basis for most variation at the molecular level.

- A **neutral mutation** has no harmful or beneficial effect on the organism. Such mutations occur at a steady rate, forming the basis for the molecular clock.
- A **deleterious mutation** has a negative effect on the phenotype, and thus decreases the fitness of the organism.
- An **advantageous mutation** has a positive effect on the phenotype, and thus increases the fitness of the organism.
- A **nearly neutral mutation** is a mutation that may be slightly deleterious or advantageous, although most nearly neutral mutations are slightly deleterious.

In reality, viewing the fitness effects of mutations in these discrete categories is an oversimplification. Attempts have been made to infer the distribution of fitness effects

using mutagenesis experiments or theoretical models applied to molecular sequence data. However, the current distribution is still uncertain, and some aspects of the distribution likely vary between species.

By inheritance

- inheritable generic in pro-generic tissue or cells on path to be changed to gametes.
- non inheritable **somatic** (e.g., carcinogenic mutation)
- non inheritable post mortem aDNA mutation in decaying remains.

By pattern of inheritance The human genome contains two copies of each gene – a paternal and a maternal allele.

- A **heterozygous mutation** is a mutation of only one allele.
- A **homozygous mutation** is an identical mutation of both the paternal and maternal alleles.
- **Compound heterozygous** mutations or a **genetic compound** comprises two different mutations in the paternal and maternal alleles.
- A **wildtype** or **homozygous non-mutated** organism is one in which neither allele is mutated. (Just not a mutation)

By impact on protein sequence

- A **frameshift mutation** is a mutation caused by insertion or deletion of a number of nucleotides that is not evenly divisible by three from a DNA sequence. Due to the triplet nature of gene expression by codons, the insertion or deletion can disrupt the reading frame, or the grouping of the codons, resulting in a completely different translation from the original. The earlier in the sequence the deletion or insertion occurs, the more altered the protein produced is.

- A **nonsense mutation** is a point mutation in a sequence of DNA that results in a premature stop codon, or a *nonsense codon* in the transcribed mRNA, and possibly a truncated, and often nonfunctional protein product.

- **Missense mutations** or *nonsynonymous mutations* are types of point mutations where a single nucleotide is changed to cause substitution of a different amino acid. This in turn can render the resulting protein nonfunctional. Such mutations are responsible for diseases such as Epidermolysis bullosa, sickle-cell disease, and SOD1 mediated ALS (Boillée 2006, p. 39).

- A **neutral mutation** is a mutation that occurs in an amino acid codon which results in the use of a different, but chemically similar, amino acid. The similarity between the two is enough that little or no change is often rendered in the protein. For example, a change from AAA to AGA will encode arginine, a chemically similar molecule to the intended lysine. Neutral mutations occur because of the degenerate nature of the genetic code.

- **Silent mutations** are mutations that do not result in a change to the amino acid sequence of a protein. They may occur in a region that does not code for a protein, or they may occur within a codon in a manner that does not alter the final amino acid sequence. The phrase *silent mutation* is often used interchangeably with the phrase *synonymous mutation*; however, synonymous mutations are a subcategory of the former, occurring only within exons. The name silent could be a misnomer. For example, a silent mutation in the exon/intron border may lead to alternative splicing by changing the splice site, thereby leading to a changed protein.

Special classes

- **Conditional mutation** is a mutation that has wild-type (or less severe) phenotype under certain "permissive" environmental conditions and a mutant phenotype under certain "restrictive" conditions. For example, a temperature-sensitive mutation can cause cell death at high temperature (restrictive condition), but might have no deleterious consequences at a lower temperature (permissive condition).

Nomenclature

A committee of the Human Genome Variation Society (HGVS) has developed the standard human sequence variant nomenclature, which should be used by researchers and DNA diagnostic centers to generate unambiguous mutation descriptions. In principle, this nomenclature can also be used to describe mutations in other organisms. The nomenclature specifies the type of mutation and base or amino acid changes.

- Nucleotide substitution (e.g. 76A>T) - The number is the position of the nucleotide from the 5' end, the first letter represents the wild type nucleotide, and the second letter represents the nucleotide which replaced the wild type. In the given example, the adenine at the 76th position was replaced by a thymine.
    - If it becomes necessary to differentiate between mutations in genomic DNA, mitochondrial DNA, and RNA, a simple convention is used. For example, if the 100th base of a nucleotide sequence mutated from G to C, then it would be written as g.100G>C if the mutation occurred in genomic DNA, m.100G>C if the mutation occurred in mitochondrial DNA, or r.100g>c if the mutation occurred in RNA. Note that for mutations in RNA, the nucleotide code is written in lower case.
- Amino acid substitution (e.g. D111E) – The first letter is the one letter code of the wild type amino acid, the number is the position of the amino acid from the N terminus, and the second letter is the one letter code of the amino acid present in the mutation. Nonsense mutations are represented with an X for the second amino acid (e.g. D111X).
- Amino acid deletion (e.g. ΔF508) – The Greek letter Δ (delta) indicates a deletion. The letter refers to the amino acid present in the wild type and the number is the position from the N terminus of the amino acid were it to be present as in the wild type.

The complete set of rules and more examples of mutation descriptions can be found at the HGVS sequence variant nomenclature website. Since the nomenclature has to cover all sequence variants, descriptions can become very complex. To prevent mistakes and facilitate correct use of this nomenclature, the journal Human Mutation recommends the use of Mutalyzer, which can apply the HGVS human nomenclature guidelines to check and, if necessary, correct sequence variant descriptions.

Harmful mutations

Changes in DNA caused by mutation can cause errors in protein sequence, creating partially or completely non-functional proteins. To function correctly, each cell depends on thousands of proteins to function in the right places at the right times. When a mutation alters a protein that plays a critical role in the body, a medical condition can result. A condition caused by mutations in one or more genes is called a genetic disorder. Some mutations alter a gene's DNA base sequence but do not change the function of the protein made by the gene. Studies of the fly *Drosophila melanogaster* suggest that if a mutation does change a protein, this will probably be harmful, with about 70 percent of these mutations having damaging effects, and the remainder being either neutral or weakly beneficial. However, studies in yeast have shown that only 7% of mutations that are not in genes are harmful.

If a mutation is present in a germ cell, it can give rise to offspring that carries the mutation in all of its cells. This is the case in hereditary diseases. On the other hand, a mutation may occur in a somatic cell of an organism. Such mutations will be present in all descendants of this cell within the same organism, and certain mutations can cause the cell to become malignant, and thus cause cancer.

Often, gene mutations that could cause a genetic disorder are repaired by the DNA repair system of the cell. Each cell has a number of pathways through which enzymes recognize and repair mistakes in DNA. Because DNA can be damaged or mutated in many ways, the process of DNA repair is an important way in which the body protects itself from disease.

Beneficial mutations

Although most mutations that change protein sequences are neutral or harmful, some mutations have a positive effect on an organism. In this case, the mutation may enable the mutant organism to withstand particular environmental stresses better than wild-type organisms, or reproduce more quickly. In these cases a mutation will tend to become more common in a population through natural selection.

For example, a specific 32 base pair deletion in human CCR5 (CCR5-Δ32) confers HIV resistance to homozygotes and delays AIDS onset in heterozygotes. The CCR5 mutation is more common in those of European descent. One possible explanation of the etiology of the relatively high frequency of CCR5-Δ32 in the European population is that it conferred resistance to the bubonic plague in mid-14th century Europe. People with this

mutation were more likely to survive infection; thus its frequency in the population increased. This theory could explain why this mutation is not found in southern Africa, where the bubonic plague never reached. A newer theory suggests that the selective pressure on the CCR5 Delta 32 mutation was caused by smallpox instead of the bubonic plague.

Another example, is Sickle cell disease which is a blood disorder in which the body produces an abnormal type of the oxygen-carrying substance hemoglobin in the red blood cells. One-third of all indigenous inhabitants of Sub-Saharan Africa carry the gene, because in areas where malaria is common, there is a survival value in carrying only a single sickle-cell gene (sickle cell trait). Those with only one of the two alleles of the sickle-cell disease are more resistant to malaria, since the infestation of the malaria plasmodium is halted by the sickling of the cells which it infests.

Prion mutation

Prions are proteins and do not contain genetic material. However, prion replication has been shown to be subject to mutation and natural selection just like other forms of replication.

Chapter- 11

Medical Genetics

Medical genetics is the specialty of medicine that involves the diagnosis and management of hereditary disorders. Medical genetics differs from Human genetics in that human genetics is a field of scientific research that may or may not apply to medicine, but medical genetics refers to the application of genetics to medical care. For example, research on the causes and inheritance of genetic disorders would be considered within both human genetics and medical genetics, while the diagnosis, management, and counseling of individuals with genetic disorders would be considered part of medical genetics.

In contrast, the study of typically non-medical phenotypes such as the genetics of eye color would be considered part of human genetics, but not necessarily relevant to medical genetics (except in situations such as albinism). *Genetic medicine* is a newer term for medical genetics and incorporates areas such as gene therapy, personalized medicine, and the rapidly emerging new medical specialty, predictive medicine.

Scope

Medical genetics encompasses many different areas, including clinical practice of physicians, genetic counselors, and nutritionists, clinical diagnostic laboratory activities, and research into the causes and inheritance of genetic disorders. Examples of conditions that fall within the scope of medical genetics include birth defects and dysmorphology, mental retardation, autism, metabolic and mitochondrial disorders, skeletal dysplasia, connective tissue disorders, cancer genetics, teratogens, and prenatal diagnosis. Medical genetics is increasingly becoming relevant to many common diseases. Overlaps with other medical specialties are beginning to emerge, as recent advances in genetics are revealing etiologies for neurologic, endocrine, cardiovascular, pulmonary, ophthalmologic, renal, psychiatric, and dermatologic conditions.

Subspecialties

In some ways, many of the individual fields within medical genetics are hybrids between clinical care and research. This is due in part to recent advances in science and technology that have enabled an unprecedented understanding of genetic disorders.

Clinical genetics

Clinical genetics is the practice of clinical medicine with particular attention to hereditary disorders. Referrals are made to genetics clinics for a variety of reasons, including birth defects, developmental delay, autism, epilepsy, short stature, and many others. Examples of genetic syndromes that are commonly seen in the genetics clinic include chromosomal rearrangements, Down syndrome, DiGeorge syndrome (22q11.2 Deletion Syndrome), Fragile X syndrome, Marfan syndrome, Neurofibromatosis, Turner syndrome, and Williams syndrome.

Metabolic/biochemical genetics

Metabolic (or biochemical) genetics involves the diagnosis and management of inborn errors of metabolism in which patients have enzymatic deficiencies that perturb biochemical pathways involved in metabolism of carbohydrates, amino acids, and lipids. Examples of metabolic disorders include galactosemia, glycogen storage disease, lysosomal storage disorders, metabolic acidosis, peroxisomal disorders, phenylketonuria, and urea cycle disorders.

Cytogenetics

Cytogenetics is the study of chromosomes and chromosome abnormalities. While cytogenetics historically relied on microscopy to analyze chromosomes, new molecular technologies such as array comparative genomic hybridization are now becoming widely used. Examples of chromosome abnormalities include aneuploidy, chromosomal rearrangements, and genomic deletion/duplication disorders.

Molecular genetics

Molecular genetics involves the discovery of and laboratory testing for DNA mutations that underlie many single gene disorders. Examples of single gene disorders include achondroplasia, cystic fibrosis, Duchenne muscular dystrophy, hereditary breast cancer (BRCA1/2), Huntington disease, Marfan syndrome, Noonan syndrome, and Rett syndrome. Molecular tests are also used in the diagnosis of syndromes involving epigenetic abnormalities, such as Angelman syndrome, Beckwith-Wiedemann syndrome, Prader-willi syndrome, and uniparental disomy.

Mitochondrial genetics

Mitochondrial genetics concerns the diagnosis and management of mitochondrial disorders, which have a molecular basis but often result in biochemical abnormalities due to deficient energy production.

There exists some overlap between medical genetic diagnostic laboratories and molecular pathology.

Genetic Counseling

Genetic counseling is the process of providing information about genetic conditions, diagnostic testing, and risks in other family members, within the framework of nondirective counseling. Genetic counselors are non-physician members of the medical genetics team who specialize in family risk assessment and counseling of patients regarding genetic disorders. The precise role of the genetic counselor varies somewhat depending on the disorder.

History

Although genetics has its roots back in the 19th century with the work of the Bohemian monk Gregor Mendel and other pioneering scientists, human genetics emerged later. It started to develop, albeit slowly, during the first half of the 20th century. Mendelian (single-gene) inheritance was studied in a number of important disorders such as albinism, brachydactyly (short fingers and toes), and hemophilia. Mathematical approaches were also devised and applied to human genetics. Population genetics was created.

Medical genetics was a late developer, emerging largely after the close of World War II (1945) when the eugenics movement had fallen into disrepute. The Nazi misuse of eugenics sounded its death knell. Shorn of eugenics, a scientific approach could be used and was applied to human and medical genetics. Medical genetics saw an increasingly rapid rise in the second half of the 20th century and continues in the 21st century.

Current practice

The clinical setting in which patients are evaluated determines the scope of practice, diagnostic, and therapeutic interventions. For the purposes of general discussion, the typical encounters between patients and genetic practitioners may involve:

- Referral to an out-patient genetics clinic (pediatric, adult, or combined) or an in-hospital consultation, most often for diagnostic evaluation.
- Specialty genetics clinics focusing on management of inborn errors of metabolism, skeletal dysplasia, or lysosomal storage diseases.
- Referral for counseling in a prenatal genetics clinic to discuss risks to the pregnancy (advanced maternal age, teratogen exposure, family history of a

genetic disease), test results (abnormal maternal serum screen, abnormal ultrasound), and/or options for prenatal diagnosis (typically amniocentesis or chorionic villus sampling).
- Multidisciplinary specialty clinics that include a clinical geneticist or genetic counselor (cancer genetics, cardiovascular genetics, craniofacial or cleft lip/palate, hearing loss clinics, muscular dystrophy/neurodegenerative disorder clinics).

Diagnostic evaluation

Each patient will undergo a diagnostic evaluation tailored to their own particular presenting signs and symptoms. The geneticist will establish a differential diagnosis and recommend appropriate testing. Increasingly, clinicians use SimulConsult, paired with the National Library of Medicine Gene Review articles, to narrow the list of hypotheses (known as the differential diagnosis) and identify the tests that are relevant for a particular patient. These tests might evaluate for chromosomal disorders, inborn errors of metabolism, or single gene disorders.

Chromosome studies

Chromosome studies are used in the general genetics clinic to determine a cause for developmental delay/mental retardation, birth defects, dysmorphic features, and/or autism. Chromosome analysis is also performed in the prenatal setting to determine whether a fetus is affected with aneuploidy or other chromosome rearrangements. Finally, chromosome abnormalities are often detected in cancer samples. A large number of different methods have been developed for chromosome analysis:

- Chromosome analysis using a karyotype involves special stains that generate light and dark bands, allowing identification of each chromosome under a microscope.
- Fluorescence in situ hybridization (FISH) involves fluorescent labeling of probes that bind to specific DNA sequences, used for identifying aneuploidy, genomic deletions or duplications, characterizing chromosomal translocations and determining the origin of ring chromosomes.
- Chromosome painting is a technique that uses fluorescent probes specific for each chromosome to differentially label each chromosome. This technique is more often used in cancer cytogenetics, where complex chromosome rearrangements can occur.
- Array comparative genomic hybridization is a new molecular technique that involves hybridization of an individual DNA sample to a glass slide or microarray chip containing molecular probes (ranging from large ~200kb bacterial artificial chromosomes to small oligonucleotides) that represent unique regions of the genome. This method is particularly sensitive for detection of genomic gains or losses across the genome but does not detect balanced translocations or distinguish the location of duplicated genetic material (for example, a tandem duplication versus an insertional duplication).

Basic metabolic studies

Biochemical studies are performed to screen for imbalances of metabolites in the bodily fluid, usually the blood (plasma/serum) or urine, but also in cerebrospinal fluid (CSF). Specific tests of enzyme function (either in leukocytes, skin fibroblasts, liver, or muscle) are also employed under certain circumstances. In the US, the newborn screen incorporates biochemical tests to screen for treatable conditions such as galactosemia and phenylketonuria (PKU). Patients suspected to have a metabolic condition might undergo the following tests:

- Quantitative amino acid analysis is typically performed using the ninhydrin reaction, followed by liquid chromatography to measure the amount of amino acid in the sample (either urine, plasma/serum, or CSF). Measurement of amino acids in plasma or serum is used in the evaluation of disorders of amino acid metabolism such as urea cycle disorders, maple syrup urine disease, and PKU. Measurement of amino acids in urine can be useful in the diagnosis of cystinuria or renal Fanconi syndrome as can be seen in cystinosis.
- Urine organic acid analysis can be either performed using quantitative or qualitative methods, but in either case the test is used to detect the excretion of abnormal organic acids. These compounds are normally produced during bodily metabolism of amino acids and odd-chain fatty acids, but accumulate in patients with certain metabolic conditions.
- The acylcarnitine combination profile detects compounds such as organic acids and fatty acids conjugated to carnitine. The test is used for detection of disorders involving fatty acid metabolism, including MCAD.
- Pyruvate and lactate are byproducts of normal metabolism, particularly during anaerobic metabolism. These compounds normally accumulate during exercise or ischemia, but are also elevated in patients with disorders of pyruvate metabolism or mitochondrial disorders.
- Ammonia is an end product of amino acid metabolism and is converted in the liver to urea through a series of enzymatic reactions termed the urea cycle. Elevated ammonia can therefore be detected in patients with urea cycle disorders, as well as other conditions involving liver failure.
- Enzyme testing is performed for a wide range of metabolic disorders to confirm a diagnosis suspected based on screening tests.

Molecular studies

- DNA sequencing is used to directly analyze the genomic DNA sequence of a particular gene. In general, only the parts of the gene that code for the expressed protein (exons) and small amounts of the flanking untranslated regions and introns are analyzed. Therefore, although these tests are highly specific and sensitive, they do not routinely identify all of the mutations that could cause disease.

- DNA methylation analysis is used to diagnose certain genetic disorders that are caused by disruptions of epigenetic mechanisms such as genomic imprinting and uniparental disomy.
- Southern blotting is an early technique basic on detection of fragments of DNA separated by size through gel electrophoresis and detected using radiolabeled probes. This test was routinely used to detect deletions or duplications in conditions such as Duchenne muscular dystrophy but is being replaced by high-resolution array comparative genomic hybridization techniques. Southern blotting is still useful in the diagnosis of disorders caused by trinucleotide repeats.
- Short tandem repeats are unique markers that can be used to determine haplotypes and are used in identity testing for maternal cell contamination.

Treatments

Each cell of the body contains the hereditary information (DNA) wrapped up in structures called chromosomes. Since genetic syndromes are typically the result of alterations of the chromosomes or genes, there is no treatment currently available that can correct the genetic alterations in every cell of the body. Therefore, there is currently no "cure" for genetic disorders. However, for many genetic syndromes there is treatment available to manage the symptoms. In some cases, particularly inborn errors of metabolism, the mechanism of disease is well understood and offers the potential for dietary and medical management to prevent or reduce the long-term complications. In other cases, infusion therapy is used to replace the missing enzyme. Current research is actively seeking to use gene therapy or other new medications to treat specific genetic disorders.

Management of Metabolic disorders

In general, metabolic disorders arise from enzyme deficiencies that disrupt normal metabolic pathways. For instance, in the hypothetical example:

```
    A ---> B ---> C ---> D          AAAA ---> BBBBBB ---> CCCCCCCCCC ---
> (no D)
       X      Y      Z                    X           Y              (no
Z)
```

Compound "A" is metabolized to "B" by enzyme "X", compound "B" is metabolized to "C" by enzyme "Y", and compound "C" is metabolized to "D" by enzyme "Z". If enzyme "Z" is missing, compound "D" will be missing, while compounds "A", "B", and "C" will build up. The pathogenesis of this particular condition could result from lack of compound "D", if it is critical for some cellular function, or from toxicity due to excess "A", "B", and/or "C". Treatment of the metabolic disorder could be achieved through dietary supplementation of compound "D" and dietary restriction of compounds "A", "B", and/or "C" or by treatment with a medication that promoted disposal of excess "A", "B", or "C". Another approach that can be taken is enzyme replacement therapy, in which a patient is given an infusion of the missing enzyme.

- Diet

Dietary restriction and supplementation are key measures taken in several well-known metabolic disorders, including galactosemia, phenylketonuria (PKU), maple syrup urine disease, organic acidurias and urea cycle disorders. Such restrictive diets can be difficult for the patient and family to maintain, and require close consultation with a nutritionist who has special experience in metabolic disorders. The composition of the diet will change depending on the caloric needs of the growing child and special attention is needed during a pregnancy if a woman is affected with one of these disorders.

- Medication

Medical approaches include enhancement of residual enzyme activity (in cases where the enzyme is made but is not functioning properly), inhibition of other enzymes in the biochemical pathway to prevent buildup of a toxic compound, or diversion of a toxic compound to another form that can be excreted. Examples include the use of high doses of pyridoxine (vitamin B6) in some patients with homocystinuria to boost the activity of the residual cystathione synthase enzyme, administration of biotin to restore activity of several enzymes affected by deficiency of biotinidase, treatment with NTBC in Tyrosinemia to inhibit the production of succinylacetone which causes liver toxicity, and the use of sodium benzoate to decrease ammonia build-up in urea cycle disorders.

- Enzyme replacement therapy

Certain lysosomal storage diseases are treated with infusions of a recombinant enzyme (produced in a laboratory), which can reduce the accumulation of the compounds in various tissues. Examples include Gaucher disease, Fabry disease, Mucopolysaccharidoses and Glycogen storage disease type II. Such treatments are limited by the ability of the enzyme to reach the affected areas (the blood brain barrier prevents enzyme from reaching the brain, for example), and can sometimes be associated with allergic reactions. The long-term clinical effectiveness of enzyme replacement therapies vary widely among different disorders.

Other examples

- Angiotensin receptor blockers in Marfan syndrome & Loeys-Dietz
- Bone marrow transplantation
- Gene therapy

Career paths and training

There are a variety of career paths within the field of medical genetics, and naturally the training required for each area differs considerably. It should be noted that the information included in this section applies to the typical pathways in the United States and there may be differences in other countries. US Practitioners in clinical, counseling,

or diagnostic subspecialties generally obtain board certification through the American Board of Medical Genetics.

Career	Degree	Description	Training
Clinical Geneticist	MD or MD/PhD	A **Clinical geneticist** is typically a physician who evaluates patients in the office or as a hospital consultation. This process includes a medical history, family history (pedigree), a detailed physical examination, reviewing objective data such as imaging and test results, establishing a differential diagnosis, and recommending appropriate diagnostic tests.	College (4 yrs) → Medical school (4 yrs) → Primary residency (2-3 yrs) → Residency in Clinical genetics (2 yrs). Some Clinical geneticists also obtain a PhD degree (4-7 yrs). A new residency track offers a 4 yr primary residency in Clinical genetics immediately after finishing Medical school.
Genetic Counselor	MS	A **Genetic counselor** specializes in communication of genetic information to patients and families. Genetic counselors often work closely with Clinical geneticists or other physicians (such as Obstetricians or Oncologists) and often convey the results of the recommended tests.	College (4 yrs) → Graduate program in Genetic counseling (2 yrs).
Metabolic nurse and/or nutritionist	BA/BS, MS, RN	One of the critical aspects of the management of patients with metabolic disorders is the appropriate nutritional intervention (either restricting the compound that cannot be metabolized, or supplementing compounds that are deficient as the result of an enzyme deficiency). The metabolic nurse and nutritionist play important roles in coordinating the dietary management.	College (4 yrs) → Nursing school or graduate training in nutrition.
Biochemical Diagnostics	PhD, MD, or MD/PhD	Individuals who specialize in **Biochemical genetics** typically work in the diagnostic laboratory, analyzing and interpreting specialized biochemical tests that measure amino acids, organic acids, and enzyme activity. Some	College (4 yrs) → Graduate school (PhD, usually 4–7 years) and/or Medical school (MD, 4 years)

		Clinical Geneticists are also board certified in Biochemical Genetics.	
Cytogenetic Diagnostics	PhD, MD, or MD/PhD	Individuals who specialize in **Cytogenetics** typically work in the diagnostic laboratory, analyzing and interpreting karyotypes, FISH, and comparative genomic hybridization tests. Some Clinical Geneticists are also board certified in Cytogenetics.	College (4 yrs) → Graduate school (PhD, usually 4–7 years) and/or Medical school (MD, 4 years)
Molecular Diagnostics	PhD, MD, or MD/PhD	Individuals who specialize in **Molecular genetics** typically work in the diagnostic laboratory, analyzing and interpreting specialized genetic tests that look for disease-causing changes (mutations) in the DNA. Some examples of molecular diagnostic tests include DNA sequencing and Southern blotting.	College (4 yrs) → Graduate school (PhD, usually 4–7 years) and/or Medical school (MD, 4 years)
Research Geneticist	PhD, MD, or MD/PhD	Any researcher who studies the genetic basis of human disease or uses model organisms to study disease mechanisms could be considered a Research Geneticist. Many of the clinical career paths also include basic or translational research, and thus individuals in the field of medical genetics often participate in some form of research.	College (4 yrs) → Graduate school (PhD, usually 4–7 years) and/or Medical school (MD, 4 years) → Post-doctoral research training (usually 3+ years)
Laboratory Technician	BS or MS	Technicians in the diagnostic or research labs handle samples and run the assays at the bench. Often these individuals are promoted to supervisory positions.	College (4 yrs), may have higher degree (MS, 2+ years)

Ethical, legal and social implications

Genetic information provides a unique type of knowledge about an individual and his/her family, fundamentally different than a typically laboratory test that provides a "snapshot" of an individual's health status. The unique status of genetic information and inherited disease has a number of ramifications with regard to ethical, legal, and societal concerns.

Societies

The more empirical approach to human and medical genetics was formalized by the founding in 1948 of the American Society of Human Genetics. The Society first began annual meetings that year (1948) and its international counterpart, the International Congress of Human Genetics, has met every 5 years since its inception in 1956. The Society publishes the American Journal of Human Genetics on a monthly basis.

Medical genetics is now recognized as a distinct medical specialty in the U.S. with its own approved board (the American Board of Medical Genetics) and clinical specialty college (the American College of Medical Genetics). The College holds an annual scientific meeting, publishes a monthly journal, Genetics in Medicine, and issues position papers and clinical practice guidelines on a variety of topics relevant to human genetics.

Research

The broad range of research in medical genetics reflects the overall scope of this field, including basic research on genetic inheritance and the human genome, mechanisms of genetic and metabolic disorders, translational research on new treatment modalities, and the impact of genetic testing

Basic genetics research

Basic research geneticists usually undertake research in universities, biotechnology firms and research institutes.

Allelic architecture of disease

Sometimes the link between a disease and an unusual gene variant is more subtle. The genetic architecture of common diseases is an important factor in determining the extent to which patterns of genetic variation influence group differences in health outcomes. According to the common disease/common variant hypothesis, common variants present in the ancestral population before the dispersal of modern humans from Africa play an important role in human diseases. Genetic variants associated with Alzheimer disease, deep venous thrombosis, Crohn disease, and type 2 diabetes appear to adhere to this model. However, the generality of the model has not yet been established and, in some cases, is in doubt. Some diseases, such as many common cancers, appear not to be well described by the common disease/common variant model.

Another possibility is that common diseases arise in part through the action of combinations of variants that are individually rare. Most of the disease-associated alleles discovered to date have been rare, and rare variants are more likely than common variants to be differentially distributed among groups distinguished by ancestry. However, groups could harbor different, though perhaps overlapping, sets of rare variants, which would reduce contrasts between groups in the incidence of the disease.

The number of variants contributing to a disease and the interactions among those variants also could influence the distribution of diseases among groups. The difficulty that has been encountered in finding contributory alleles for complex diseases and in replicating positive associations suggests that many complex diseases involve numerous variants rather than a moderate number of alleles, and the influence of any given variant may depend in critical ways on the genetic and environmental background. If many alleles are required to increase susceptibility to a disease, the odds are low that the necessary combination of alleles would become concentrated in a particular group purely through drift.

Population substructure in genetics research

One area in which population categories can be important considerations in genetics research is in controlling for confounding between population substructure, environmental exposures, and health outcomes. Association studies can produce spurious results if cases and controls have differing allele frequencies for genes that are not related to the disease being studied, although the magnitude of this problem in genetic association studies is subject to debate. Various methods have been developed to detect and account for population substructure, but these methods can be difficult to apply in practice.

Population substructure also can be used to advantage in genetic association studies. For example, populations that represent recent mixtures of geographically separated ancestral groups can exhibit longer-range linkage disequilibrium between susceptibility alleles and genetic markers than is the case for other populations. Genetic studies can use this admixture linkage disequilibrium to search for disease alleles with fewer markers than would be needed otherwise. Association studies also can take advantage of the contrasting experiences of racial or ethnic groups, including migrant groups, to search for interactions between particular alleles and environmental factors that might influence health.