INTRODUCTORY BIOTECHNOLOGY (ABG 504) THEORETICAL MODULE

BY

PROFESSOR ‘FUNMI ADEBAMBO

GENETICS

Genetics (from Ancient Greek γενετικός *genetikos*, “genitive” and that from γένεσις *genesis*, “origin”[1][2][3]), a discipline of biology, is the science of heredity and variation in living organisms.[4][5] 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-nineteenth century.[6] Although he did not know the physical basis for heredity, Mendel observed that organisms inherit traits in a discrete manner—these basic units of inheritance are now called genes.

THE DNA

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.

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 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 a person's height, the nutrition and health that person experiences in childhood also have a large effect.

History of genetics

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.

Although the science of genetics began with the applied and theoretical work of Gregor Mendel in the mid-1800s, 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.[7] 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.[8]

Mendelian and classical genetics

The modern science of genetics traces its roots to 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.[9] 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.[10][11] (The adjective *genetic*, derived from the Greek word *genesis* - *γένεσις*, "origin" and that from the word *genno* - *γεννώ*, "to give birth", predates the noun and was first used in a biological sense in 1860.)[12] 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.[13]

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.[14] In 1913, his student Alfred Sturtevant used the phenomenon of genetic linkage to show that genes are arranged linearly on the chromosome.[15]

Molecular genetics

James D. Watson and Francis Crick determined the structure of DNA in 1953.

Although genes were known to exist on chromosomes, chromosomes are composed of both protein and DNA—scientists did not know which of these was responsible for inheritance. In 1928, Frederick Griffith discovered the phenomenon of transformation in which he reported that 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.[16] The Hershey-Chase experiment in 1952 also showed that DNA (rather than protein) was the genetic material of the viruses that infect bacteria, providing further evidence that DNA was the molecule responsible for inheritance.[17]

James D. Watson and Francis Crick determined the structure of DNA in 1953, using the X-ray crystallography work of Rosalind Franklin that indicated DNA had a helical structure (i.e., shaped like a corkscrew).[18][19] 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.[20] 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 worked, it was still not known how DNA influenced 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.[21] 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.[22] 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.[23]

Mendelian inheritance

At its most fundamental level, inheritance in organisms occurs by means of discrete traits, called genes.[24] This property was first observed by Gregor Mendel, who studied the segregation of heritable traits in pea plants.[9][25] In his experiments studying the trait for flower color, Mendel observed that the flowers of each pea plant were either purple or white - and never an intermediate between the two colors. These different, discrete versions of the same gene are called alleles.

In the case of pea plants, each organism has two alleles of each gene, and the plants inherit one allele from each parent.[26] Many organisms, including humans, have this pattern of inheritance. Organisms with two copies of the same allele are called homozygous, while organisms with two different alleles are heterozygous.

The set of alleles for a given organism is called its genotype, while the observable trait that the organism has is called its phenotype. When organisms are heterozygous, 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.[27]

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.

The laws of inheritance were derived by Gregor Mendel, a 19th century Austrian monk, who was conducting plant hybridity experiments. Between 1856 and 1863, he cultivated and tested some 28,000 pea plants. His experiments brought forth two generalizations which later became known as *Mendel's Laws of Heredity* or *Mendelian inheritance*. These are described in his paper "Experiments on Plant Hybridization" that was read to the Natural History Society of Brno on February 8 and March 8, 1865, and was published in 1866.

Mendel's results were largely neglected. Though they were not completely unknown to biologists of the time, they were not seen as being important. Even Mendel himself did not see their ultimate applicability, and thought they only applied to certain categories of species. In 1900, however, the 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, and Correns pointed out Mendel's priority after having read De Vries's paper and realizing that he himself did not have priority, and 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. 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 for individual traits, they could produce the diverse amount of results observed in nature. Thomas Hunt Morgan and his assistants would later integrate the theoretical model of Mendel with the chromosome theory of inheritance, in which the chromosomes of cells were thought to hold the actual hereditary particles, and create what is now known as classical genetics, which was extremely successful and cemented Mendel's place in history.

Mendel's law of segregation

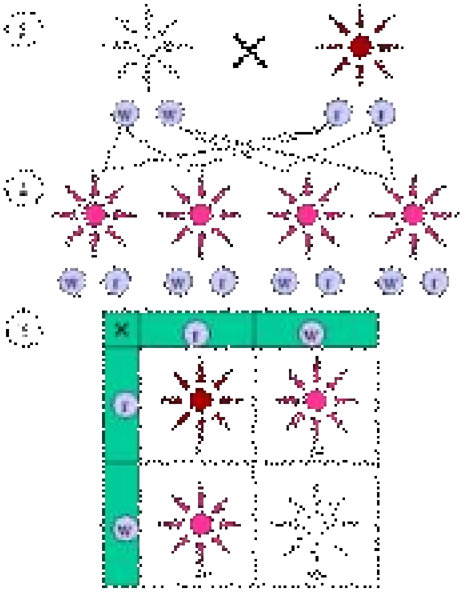


Figure 3 : The color alleles of Mirabilis jalapa are not dominant or recessive.
(1) Parental generation. (2) F_1 generation. (3) F_2 generation. The "red" and "white" allele together make a "pink" phenotype, resulting in a 1:2:1 ratio of red:pink:white in the F_2 generation.

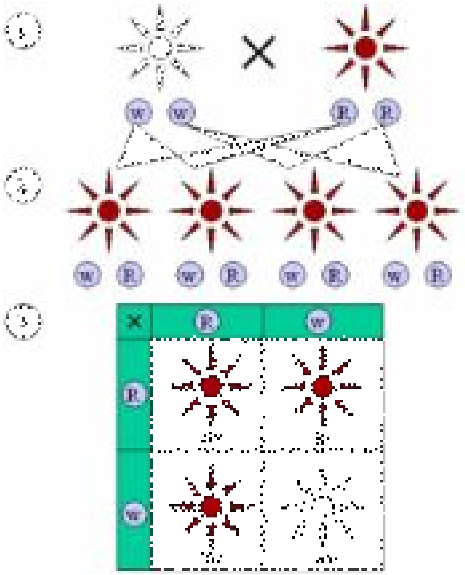


Figure 1 : Dominant and recessive phenotypes.
(1) Parental generation. (2) F$_1$ generation. (3) F$_2$ generation. Dominant (red) and recessive (white) phenotype look alike in the F$_1$ (first) generation and show a 3:1 ratio in the F$_2$ (second) generation

Mendel's law of segregation, also known as Mendel's first law, essentially has four parts.

1. **Alternative versions of genes account for variations in inherited characters.** This is the concept of alleles. Alleles are different versions of genes that impart the same characteristic. Each human has a gene that controls height, but there are variations among these genes in accordance with the specific height the gene "codes" for.
2. **For each character, an organism inherits two genes, one from each parent.** This means that when somatic cells are produced from two gametes, one allele comes from the mother, one from the father. These alleles may be the same (true-breeding organisms, e.g. *ww* and *rr* in Fig. 3), or different (hybrids, e.g. *wr* in Fig. 3).
3. **If the two alleles differ, then one, the dominant allele, is fully expressed in the organism's appearance; the other, the recessive allele, has no noticeable effect on the organism's appearance.** In other words, the dominant allele is expressed in the phenotype of the organism. However this does not always hold true: Today, we know several examples that disprove this "law", e.g. Mirabilis jalapa, the "Japanese wonder flower" (Fig. 3). This is called incomplete dominance. There is also codominance on a molecular level, e.g. people with sickle cell anemia, when normal and sickle-shaped red blood cells mix and prevent malaria.
4. **The two genes for each character segregate during gamete production.** This is the last part of Mendel's generalization. The two alleles of the organism are separated into different gametes, ensuring variation.

During his experiments, Mendel encountered some traits that did not follow the laws he had encountered. These traits did not appear independently, but always together with at least one other trait. Mendel could not explain what happened and chose not to mention it in his work. Today, we know that these traits are linked on the same chromosome.

Mendel's law of independent assortment

Also known as **Mendel's Second Law**

The most important principle of Mendel's law of independent assortment is that the emergence of one trait will not affect the emergence of another. While his 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 showed 9:3:3:1 ratios (Fig. 2). Mendel concluded that each organism carries two sets of information about its phenotype. If the two sets differ on the same phenotype, one of them dominates the other. That way, information can be passed on through the generations, even if the phenotype is not expressed (F_1 generations, figures 1 and 2).

Mendel's findings allowed other scientists to simplify the emergence of traits to mathematical probability. A large portion of Mendel's findings can be traced to his choice to start his experiments only with true breeding plants. He also only measured absolute characteristics such as color, shape, and position of the offspring. His data was expressed numerically and subjected to statistical analysis. This method of data reporting and the large sampling size he used gave credibility to his data. He also had the foresight to look through several successive generations of his pea plants and record their variations. Without his careful attention to procedure and detail, Mendel's work could not have had the impact it made on the world of genetics.

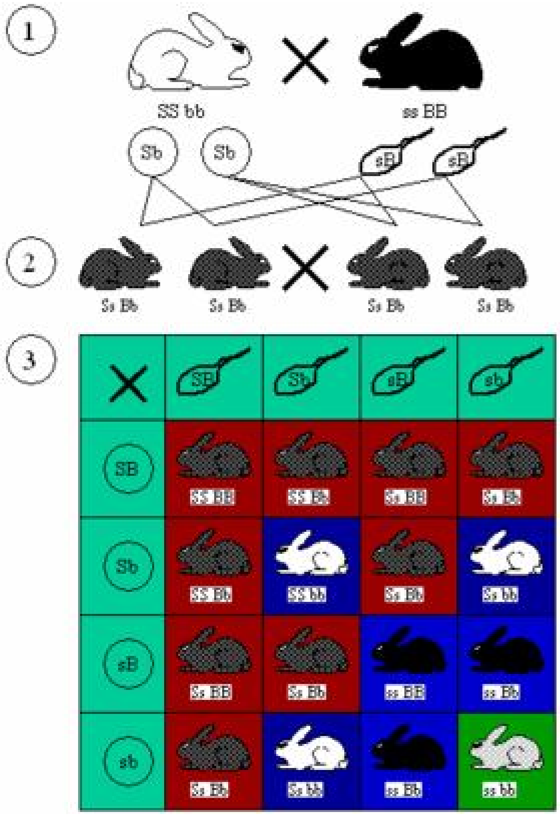


Figure 2 : Two traits (black/white and short/long hair, with black and short dominant) show a 9:3:3:1 ratio in the F_2 generation. (S=short, s=long, B=black, b=white hair)
(1) Parental generation. (2) F_1 generation. (3) F_2 generation.
Results : 9x short black hair, 3x long black hair, 3x short white hair, 1x long white hair.

References

- Peter J. Bowler (1989). *The Mendelian Revolution: The Emergence of Hereditarian Concepts in Modern Science and Society*, Baltimore: John Hopkins University Press.af:Wette van Mendel

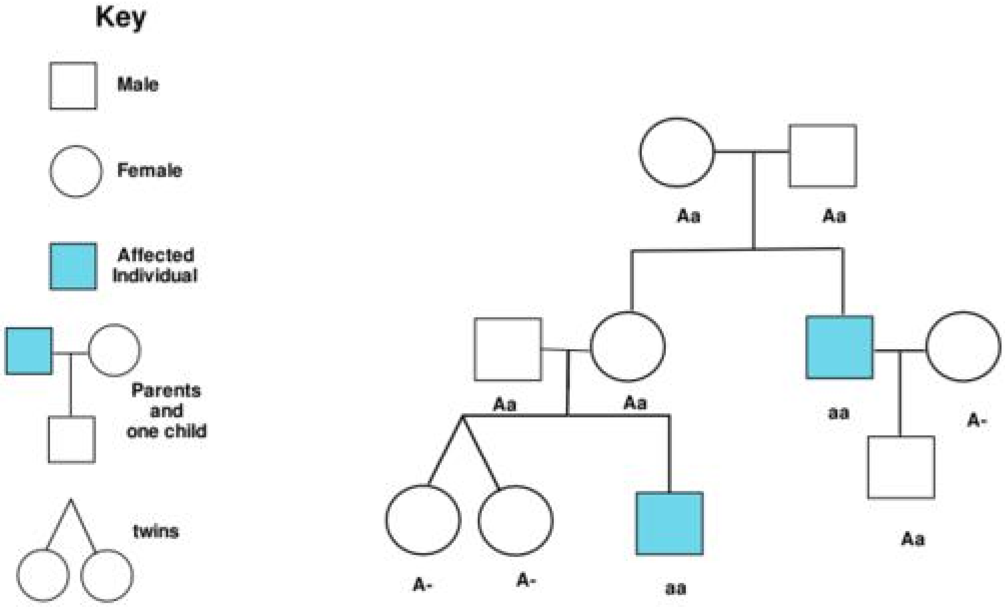
Notation and diagrams

Genetic pedigree charts help track the inheritance patterns of traits.

Geneticists use diagrams and symbols to describe inheritance. A gene is represented by a letter (or letters)—the capitalized letter represents the dominant allele and the recessive is represented by lowercase.[28] 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.

A **pedigree chart** is a chart which tells one all of the known phenotypes for an organism and its ancestors, most commonly humans, show dogs, and race horses. The word pedigree is a corruption of the French "pied de gru" or crane's foot, because the typical lines and split lines (each split leading to different offspring of the one parent line) resemble the thin leg and foot of a crane.



A genetics pedigree chart following a recessive trait.

In animal breeding

In an animal pedigree chart, characteristics are colored in, and all those without that characteristics are left unfilled. A disease may be recessive or dominant. Organisms known to be heterozygous are half colored in, half not. Squares represent males, while circles represent female.

In human genealogy

Pedigree charts are also a common tool in human genealogy studies as a means of displaying the ancestry of a given individual or the descendants of a given individual. A pedigree chart concentrating on all the ancestors of a single individual (including all female lines of ancestry) is also called a "birth brief" in the UK and may number persons on the chart using the Ahnentafel system.

Equally common, however, is the "family tree" form of pedigree chart, which shows the descendants of a particular individual, and thereby highlights sibling and cousin relationships which would not appear on the above form of pedigree chart. A chart on these lines will often concentrate only on the male line of descent, so that the marriages but not the children of female descendants are recorded. In this way, the pedigree will encompass only those who share the same surname.

In addition to the names of the individuals, it is common to include each person's birth date and place, death date and place, and the marriage date and place of each couple.

In England and Wales pedigrees are officially recorded in the College of Arms, which has records going back to the middle ages, including pedigrees collected during roving inquiries by its heralds during the sixteenth and seventeenth century. The purpose of these heraldic visitations was to register and regulate the use of coats of arms. Those who claimed the right to bear arms had to provide proof either of a grant of arms to them by the College, or of descent from an ancestor entitled to arms. It was for this reason that pedigrees were recorded by the visitations. Pedigrees continue to be registered at the College of Arms and kept up to date on a voluntary basis but they are not accessible to the general public without payment of a fee.

More visible, therefore, are the pedigrees recorded in published works, such as Burke's Peerage and Burke's Landed Gentry in the United Kingdom and, in continental Europe by the Almanach de Gotha. Due to space considerations, however, these publications typically use a narrative pedigree, whereby relationships are indicated by numbers (one for each child, a different format for each generation) and by indentations (each generation being indented further than its predecessor). This format is very flexible, and allows for a great deal of information to be included, but it lacks the clarity of the traditional chart pedigree.

U.S. usage

In the United States, the term "pedigree chart" refers to a chart showing the direct ancestors of a given individual. In addition to the names of the individuals, the chart often includes each person's birth date and place, death date and place, and each couple's marriage date and place. It is also common for persons on the chart to be numbered according to the Ahnentafel numbering system.[1]

Pedigree chart numbered on the Ahnentafel system

```
                                            8.  Great-grandfather
                    4. Paternal grandfather-|
                    |                       9.  Great-grandmother
          2. Father-|
          |         |                       10. Great-grandfather
          |         5. Paternal grandmother-|
          |                                 11. Great-grandmother
          |
1. Person-|
          |
          |                                 12. Great-grandfather
          |         6. Maternal grandfather-|
          |         |                       13. Great-grandmother
          3. Mother-|
                    |                       14. Great-grandfather
                    7. Maternal grandmother-|
                                            15. Great-grandmother
```

References

1. ↑ U.S.-style pedigree chart (PDF). Includes lines for birth, marriage, and death data, and utilizes Ahnentafel numbering.

When studying human genetic diseases, geneticists often use pedigree charts to represent the inheritance of traits.[29] These charts map the inheritance of a trait in a family tree.

Interactions of multiple genes

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

Organisms have thousands of genes, and in sexually reproducing organisms, assortment of these genes are generally independent 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. However, some genes do not assort independently, thus demonstrating what is called genetic linkage.

Genetic linkage occurs when particular alleles are inherited together. Typically, an organism can pass on an allele without regard to which allele was passed on for a different gene. This is because chromosomes are sorted randomly during meiosis.

However, alleles that are on the same chromosome are more likely to be inherited together, and are said to be linked.

Because there is some crossing over of DNA when the chromosomes segregate, alleles on the same chromosome can be separated and go to different cells. There is a greater probability of this happening if the alleles are far apart on the chromosome, as it is more likely that a cross-over will occur between them.

The physical distance between two genes can be calculated using the offspring of an organism showing two linked genetic traits, and finding the percentage of the offspring where the two traits don't run together. The higher the percentage of descendence that doesn't show both traits, the further apart on the chromosome they are.

A study of the linkages between many genes enables the creation of a **linkage map** or genetic map.

Among individuals of an experimental population or species, some phenotypes or traits occur randomly with respect to one another in a manner known as independent assortment. Today scientists understand that independent assortment occurs when the genes affecting the phenotypes are found on different chromosomes.

An exception to independent assortment develops when genes appear near one another on the same chromosome. When genes occur on the same chromosome, they are usually inherited as a single unit. Genes inherited in this way are said to be linked. For example, in fruit flies the genes affecting eye color and wing length are inherited together because they appear on the same chromosome.

But in many cases, even genes on the same chromosome that are inherited together produce offspring with unexpected allele combinations. This results from a process called crossing over. Sometimes at the beginning of meiosis, a chromosome pair (made up of a chromosome from the mother and a chromosome from the father) may 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.

Genetic linkage was first discovered by the British geneticists William Bateson and Reginald Punnett shortly after Mendel's laws were rediscovered.

Contents

[show]

-

edit Linkage mapping

The observations by Thomas Hunt Morgan 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. Morgan's student Alfred Sturtevant developed the first genetic map, also called a linkage map.

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. 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 mapping is critical for identifying the location of genes that cause genetic diseases. In a normal population, genetic traits and markers will occur in all possible combinations with the frequencies of combinations determined by the frequencies of the individual genes. For example, if alleles *A* and *a* occur with frequency 90% and 10%, and alleles *B* and *b* at a different genetic locus occur with frequencies 70% and 30%, the frequency of individuals having the combination *AB* would be 63%, the product of the frequencies of *A* and *B*, regardless of how close together the genes are. However, if a mutation in gene *B* that causes some disease happened recently in a particular subpopulation, it almost always occurs with a particur allele of gene *A* if the individual in which the mutation occurred had that variant of gene *A* and there have not been sufficient generations for recombination to happen between them (presumably due to tight linkage on the genetic map). In this case, called linkage disequilibrium, it is possible to search potential markers in the subpopulation and identify which marker the mutation is close to, thus determining the mutation's location on the map and identifying the gene at which the mutation occurred. Once the gene has been identified, it can be targeted to identify ways to mitigate the disease.

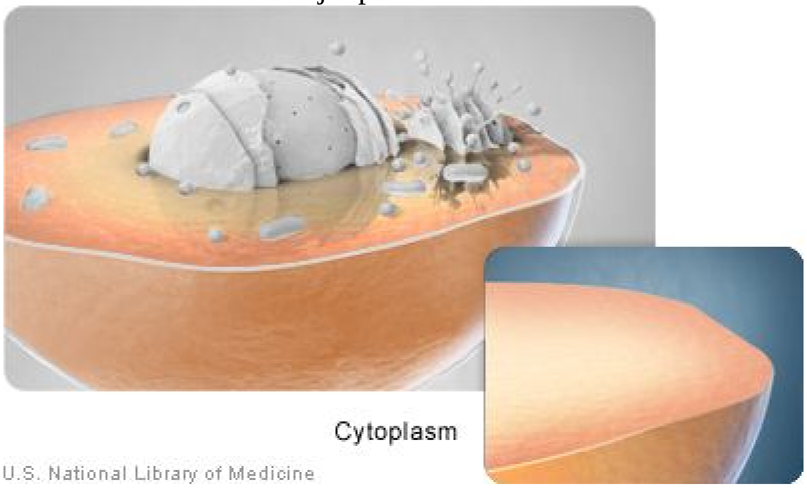
Often different genes can interact in a way that influences the same trait. In the Blue-eyed Mary (*Omphalodes verna*), for example, there exists a gene with alleles that determine the color of flowers: blue or magenta. Another gene, however, controls whether the flowers have color at all: color or white. 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 such as this, is called epistasis, with the second gene epistatic to the first.[30]

Many traits are not discrete features such as purple or white flowers, but are instead continuous features for example human height and skin color.

These complex traits are the product of many genes.[31] 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.[32] 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 complex trait with 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%.[33]**

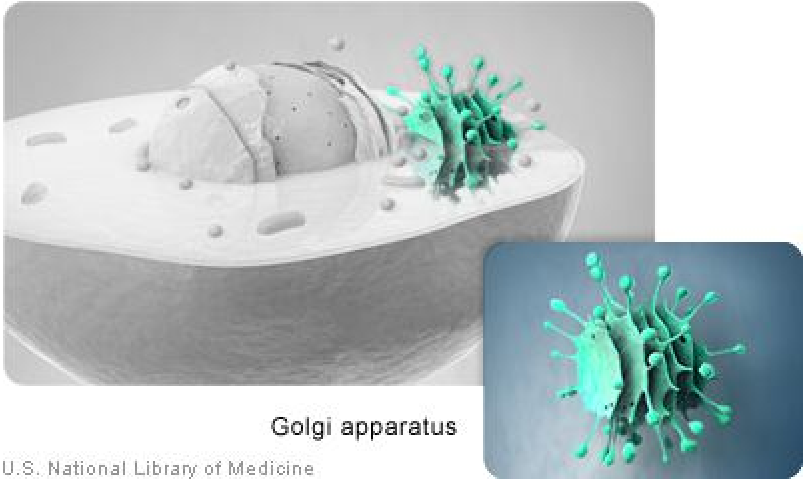
THE CELL

Major parts of a cell



The cytoplasm surrounds the cell’s nucleus and organelles.

Major parts of a cell



The Golgi apparatus is involved in packaging molecules for export from the cell.

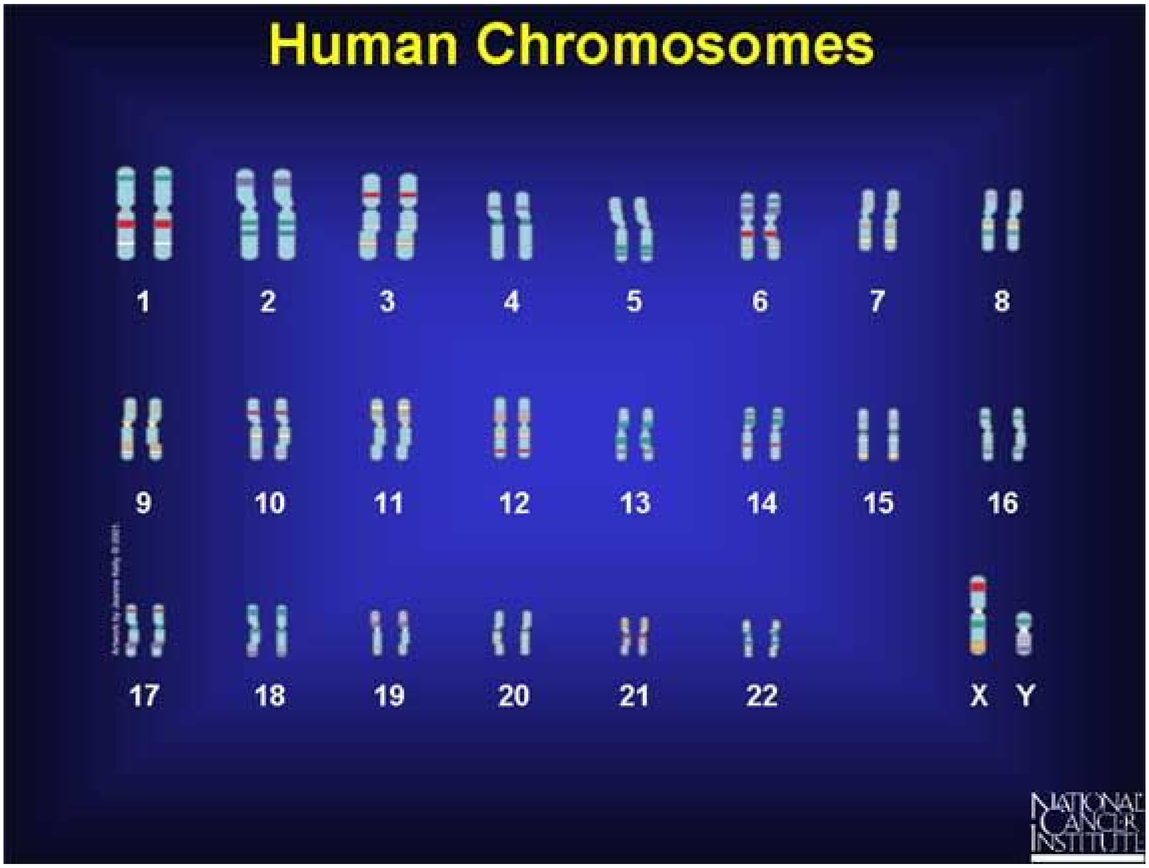
Mitochondria
U.S. National Library of Medicine

Mitochondria provide the cell’s energy.

Human Chromosomes

Human cells contain two sets of chromosomes, one inherited from the mother and one from the father.

Each set has 23 single chromosomes--22 autosomes and a sex-determining chromosome, either X or Y. The set shown here is from a male, since it contains an X and a Y chromosome; if the chromosome set were from a female, it would contain an X and an X.



Let us look at an example of chromosomes (5)

Humans normally have 46 chromosomes in each cell, divided into 23 pairs. Two copies of chromosome 5, one copy inherited from each parent, form one of the pairs. Chromosome 5 spans about 181 million DNA building blocks (base pairs) and represents almost 6 percent of the total DNA in cells.

Identifying genes on each chromosome is an active area of genetic research. Because researchers use different approaches to predict the number of genes on each chromosome, the estimated number of genes varies. Chromosome 5 likely contains between 900 and 1,300 genes.

Genes on chromosome 5 are among the estimated 20,000 to 25,000 total genes in the human genome.

How are changes in chromosome 5 related to health conditions?

Many genetic conditions are related to changes in particular genes such as on chromosome 5. This list of disorders associated with genes on chromosome 5 provides links to additional information.

Changes in the structure or number of copies of a chromosome can also cause problems with health and development. The following chromosomal conditions are associated with such changes in chromosome 5.

cancers

> Changes in the structure of chromosome 5 are associated with certain forms of cancer and conditions related to cancer. These changes are typically somatic, which means they are acquired during a person's lifetime and are present only in tumor cells. Deletions in the long (q) arm of the chromosome have been identified in a form of blood cancer known as acute myeloid leukemia (AML). These deletions also frequently occur in a disorder called myelodysplastic syndrome, which is a disease of the blood and bone marrow. People with this condition have a low number of red blood cells (anemia) and an increased risk of developing AML. When myelodysplastic syndrome is associated with a specific deletion in the long arm of chromosome 5, it is known as 5q- (5q minus) syndrome. Studies are under way to determine which genes in the deleted region of chromosome 5 are related to myelodysplastic syndrome and AML.

cri-du-chat syndrome

> Cri-du-chat syndrome is caused by a deletion of the end of the short (p) arm of chromosome 5. This chromosomal change is written as 5p- (5p minus). The signs and symptoms of cri-du-chat syndrome are probably related to the loss of multiple genes in this region. Researchers are working to identify all of these genes and determine how their loss leads to the features of the disorder. They have discovered that in people with cri-du-chat syndrome, larger deletions tend to result in more severe intellectual disability and developmental delays than smaller deletions. They have also defined regions of the short arm of chromosome 5 that are associated with particular features of cri-du-chat syndrome. A specific region designated 5p15.3 is associated with a cat-like cry, and a nearby region called 5p15.2 is associated with intellectual disability, small head size (microcephaly), and distinctive facial features.

Crohn disease

> Several regions of chromosome 5 have been associated with the risk of developing Crohn disease. For example, a combination of genetic variations in a region of DNA on the long (q) arm of the chromosome (5q31) has been shown to increase a person's chance of developing Crohn disease. Taken together, these variations are known as the inflammatory bowel disease 5 (IBD5) locus. This region of chromosome 5 contains several related genes that may be associated with Crohn disease risk, including SLC22A4 and SLC22A5.
>
> Variations in a region of the short (p) arm of chromosome 5 designated 5p13.1 are also associated with Crohn disease risk. Researchers refer to this part of chromosome 5 as a "gene desert" because it contains no known genes; however, it may contain stretches of DNA that help regulate nearby genes such as PTGER4. Research studies are under way to examine a possible connection between the PTGER4 gene and Crohn disease.

periventricular heterotopia

> In a few cases, abnormalities in chromosome 5 have been associated with periventricular heterotopia, a disorder characterized by abnormal clumps of nerve cells (neurons) around fluid-filled cavities (ventricles) near the center of the brain. In each case, the affected individual had extra genetic material caused by an abnormal duplication of part

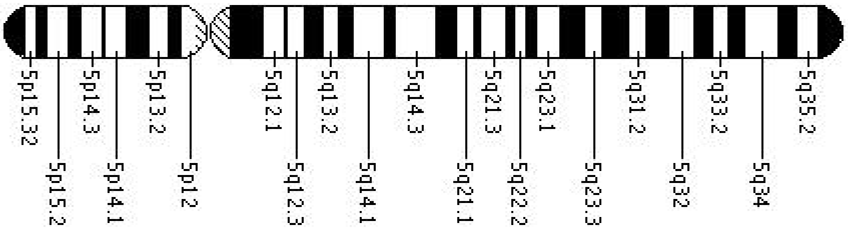
of this chromosome. It is not known how this duplicated genetic material results in the signs and symptoms of periventricular heterotopia.

other chromosomal conditions

Other changes in the number or structure of chromosome 5 can have a variety of effects, including delayed growth and development, distinctive facial features, birth defects, and other medical problems. Changes to chromosome 5 include an extra segment of the short (p) or long (q) arm of the chromosome in each cell (partial trisomy 5p or 5q), a missing segment of the long arm of the chromosome in each cell (partial monosomy 5q), and a circular structure called ring chromosome 5. Ring chromosomes occur when a chromosome breaks in two places and the ends of the chromosome arms fuse together to form a circular structure.

Is there a standard way to draw chromosome 5?

Geneticists use diagrams called ideograms as a standard representation for chromosomes. Ideograms show a chromosome's relative size and its banding pattern. A banding pattern is the characteristic pattern of dark and light bands that appears when a chromosome is stained with a chemical solution and then viewed under a microscope. These bands are used to describe the location of genes on each chromosome.



DNA and chromosomes

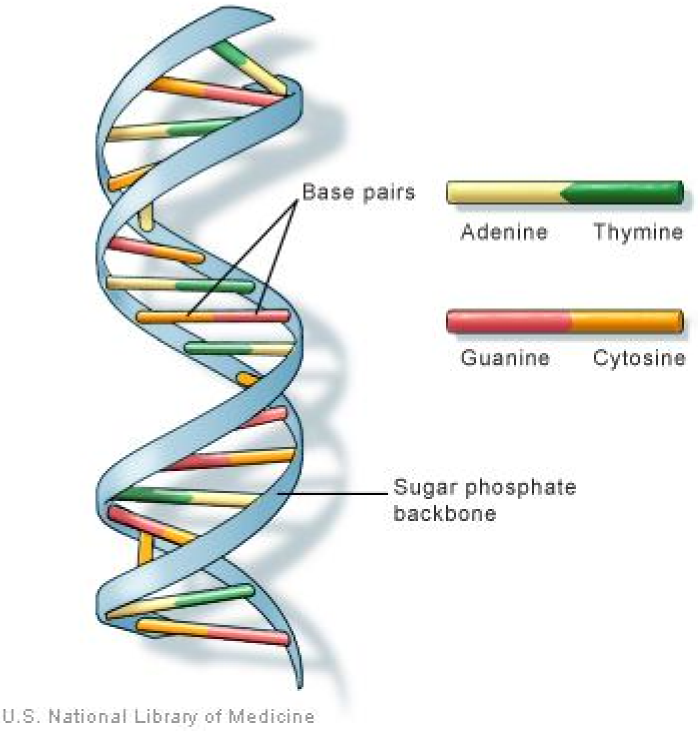
What is DNA?

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria (where it is called mitochondrial DNA or mtDNA).

The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. **The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences.**

DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. **Together, a base, sugar, and phosphate are called a nucleotide.** Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder.

An important property of DNA is that it can replicate, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell.



DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone.

Molecular Structure of the DNA.

In the 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.[34] Viruses are the only exception to this rule—sometimes viruses use the very similar molecule RNA instead of DNA as their genetic material.[35]

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.[36]

Genes are arranged linearly along long chains of DNA sequence, called chromosomes. In bacteria, each cell has a single circular chromosome, while eukaryotic organisms (which includes 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.[37] 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, repeating units of DNA wound around a core of histone proteins.[38] 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.[26] The two alleles for a gene are located on identical loci of sister chromatids, each allele inherited from a different parent.

Chromosomes are copied, condensed, and organized. Then, as the cell divides, chromosome copies separate into the daughter cells.

An exception exists in the sex chromosomes. These are specialized chromosomes in many animals that have evolved to play a role in determining the sex of the organism.[39] In humans and other mammals, the Y chromosome has very few genes and triggers the development of male sexual characteristics, while the X chromosome is similar to the other chromosomes and contains many genes unrelated to sex determination. Females have two copies of the X chromosome, but males have one Y and only one X chromosome - this difference in X chromosome copy numbers leads to the unusual inheritance patterns of sex-linked disorders.

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).[26] 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.

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.[40] Bacteria can also take up raw DNA fragments found in the environment and integrate them into their genome, a phenomenon known as transformation.[41] These processes result in horizontal gene transfer, transmitting fragments of genetic information between organisms that would be otherwise unrelated.

Recombination and linkage

Chromosomal crossover

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.[42] 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 them. 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.[43

How do geneticists indicate the location of a gene?

Geneticists use maps to describe the location of a particular gene on a chromosome. One type of map uses the cytogenetic location to describe a gene’s position. The cytogenetic location is based on a distinctive pattern of bands created when chromosomes are stained with certain chemicals. Another type of map uses the molecular location, a precise description of a gene’s position on a chromosome. The molecular location is based on the sequence of DNA building blocks (base pairs) that make up the chromosome.

Cytogenetic location

Geneticists use a standardized way of describing a gene’s cytogenetic location. In most cases, the location describes the position of a particular band on a stained chromosome:

17q12

It can also be written as a range of bands, if less is known about the exact location:

17q12-q21

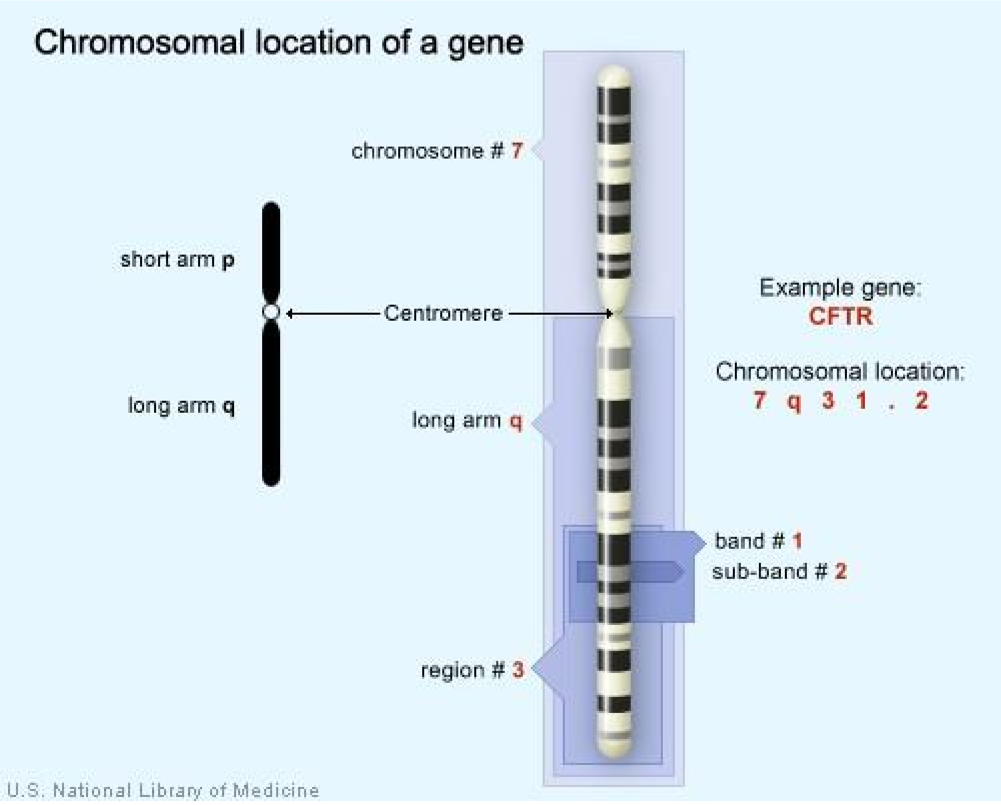
The combination of numbers and letters provide a gene’s “address” on a chromosome. This address is made up of several parts:

- The chromosome on which the gene can be found. The first number or letter used to describe a gene’s location represents the chromosome. Chromosomes 1 through 22 (the autosomes) are designated by their chromosome number. The sex chromosomes are designated by X or Y.

- The arm of the chromosome. Each chromosome is divided into two sections (arms) based on the location of a narrowing (constriction) called the centromere. By convention, the shorter arm is called p, and the longer arm is called q. The chromosome arm is the second part of the gene’s address. For example, 5q is the long arm of chromosome 5, and Xp is the short arm of the X chromosome.

- The position of the gene on the p or q arm. The position of a gene is based on a distinctive pattern of light and dark bands that appear when the chromosome is stained in a certain way. The position is usually designated by two digits (representing a region and a band), which are sometimes followed by a decimal point and one or more additional digits (representing sub-bands within a light or dark area). The number indicating the gene position increases with distance from the centromere. For example: 14q21 represents position 21 on the long arm of chromosome 14. 14q21 is closer to the centromere than 14q22.

Sometimes, the abbreviations “cen” or “ter” are also used to describe a gene’s cytogenetic location. “Cen” indicates that the gene is very close to the centromere. For example, 16pcen refers to the short arm of chromosome 16 near the centromere. “Ter” stands for terminus, which indicates that the gene is very close to the end of the p or q arm. For example, 14qter refers to the tip of the long arm of chromosome 14. (“Tel” is also sometimes used to describe a gene’s location. “Tel” stands for telomeres, which are at the ends of each chromosome. The abbreviations “tel” and “ter” refer to the same location.)



The CFTR gene is located on the long arm of chromosome 7 at position 7q31.2.

Molecular location

The Human Genome Project, an international research effort completed in 2003, determined the sequence of base pairs for each human chromosome. This sequence information allows researchers to provide a more specific address than the cytogenetic location for many genes. A gene’s molecular address pinpoints the location of that gene in terms of base pairs. It describes the gene’s precise position on a chromosome and indicates the size of the gene. Knowing the molecular location also allows researchers to determine exactly how far a gene is from other genes on the same chromosome.

Different groups of researchers often present slightly different values for a gene’s molecular location. Researchers interpret the sequence of the human genome using a variety of methods, which can result in small differences in a gene’s molecular address. Genetics Home Reference presents data from NCBI for the molecular location of genes.

How do genes control the growth and division of cells?

A variety of genes are involved in the control of cell growth and division. The cell cycle is the cell’s way of replicating itself in an organized, step-by-step fashion. Tight regulation of this process ensures that a dividing cell’s DNA is copied properly, any errors in the DNA are repaired, and each daughter cell receives a full set of chromosomes. The cycle has checkpoints (also called restriction points), which allow certain genes to check for mistakes and halt the cycle for repairs if something goes wrong.

If a cell has an error in its DNA that cannot be repaired, it may undergo programmed cell death (apoptosis) (illustration).

Apoptosis is a common process throughout life that helps the body get rid of cells it doesn’t need. Cells that undergo apoptosis break apart and are recycled by a type of white blood cell called a macrophage (illustration). Apoptosis protects the body by removing genetically damaged cells that could lead to cancer, and it plays an important role in the development of the embryo and the maintenance of adult tissues.

Cancer results from a disruption of the normal regulation of the cell cycle. When the cycle proceeds without control, cells can divide without order and accumulate genetic defects that can lead to a cancerous tumor (illustration).

Gene expression

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 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 to one of the twenty possible amino acids in protein - this correspondence is called the genetic code.[44] 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.[45]

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

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 protein are related to their function.[46][47] 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.

A single nucleotide difference within DNA can cause a single 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.[48] 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 - these are called non-coding RNA molecules. In some cases, these products fold into structures which are involved in critical cell functions (eg. ribosomal RNA and transfer RNA). RNA can also have regulatory effect through hybridization interactions with other RNA molecules (eg. 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 dichotomy 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 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.[49] 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.[50] 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.

Gene regulation

Regulation of gene expression

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.[51] 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.[52]

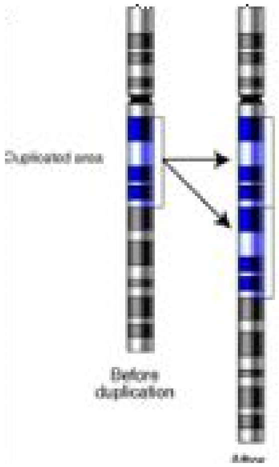
Transcription factors bind to DNA, influencing the transcription of associated genes.

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.[53] 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.[54]

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.[55][56] (Without proofreading error rates are a thousand-fold 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.[57] 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.[58] 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

Evolution

Mutations produce organisms with different genotypes, and those differences can result in different phenotypes. Many mutations have little effect on an organism's phenotype, health, and reproductive fitness. Mutations that do have an effect are often deleterious, but occasionally mutations are beneficial. 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.[59]

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

Population genetics research studies the distributions of these genetic differences within populations and how the distributions change over time.[60] Changes in the frequency of an allele in a population can be influenced by natural selection, where a given allele's higher rate of survival and reproduction causes it to become more frequent in the population over time.[61] Genetic drift can also occur, where chance events lead to random changes in allele frequency.[62]

Over many generations, the genomes of organisms can change, resulting in the phenomenon of evolution. Mutations and the selection for beneficial mutations can cause a species to evolve into forms that better survive their environment, a process called adaptation.[63] New species are formed through the process of speciation, a process often caused by geographical separations that allow different populations to genetically diverge.[64] The application of genetic principles to the study of population biology and evolution is referred to as the modern synthesis.

As sequences diverge and change during the process of evolution, these differences between sequences can be used as a molecular clock to calculate the evolutionary distance between them.[65] Genetic comparisons are generally considered the most accurate method of characterizing the relatedness between species, an improvement over the sometimes deceptive comparison of phenotypic characteristics. The evolutionary distances between species can be combined to form evolutionary trees - these trees represent the common descent and divergence of species over time, although they cannot represent 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.[66] 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.[67] 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.[68] Once a candidate gene is found, further research is often done on the same 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.[69]

Although it is not an inherited disease, cancer is also considered a genetic disease.[70] The process of cancer development in the body is a combination of events. Mutations occasionally occur within cells in the body as they divide. While 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

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

DNA can be manipulated in the laboratory. Restriction enzymes are a commonly used enzyme that cuts DNA at specific sequences, producing predictable fragments of DNA.[71] The use of ligation enzymes allows these fragments to be reconnected, 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.)

DNA can also be amplified using a procedure called the polymerase chain reaction (PCR).[72] 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.[73] With this technology, researchers have been able to study the molecular sequences associated with many human diseases.

As sequencing has become less expensive and with the aid of computational tools, researchers have sequenced the genomes of many organisms by stitching together the sequences of many different fragments (a process called genome assembly).[74] These technologies were used to sequence the human genome, leading to the completion of the Human Genome Project in 2003.[23] 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.[75]

The large amount of sequences 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.

Gene Mutations

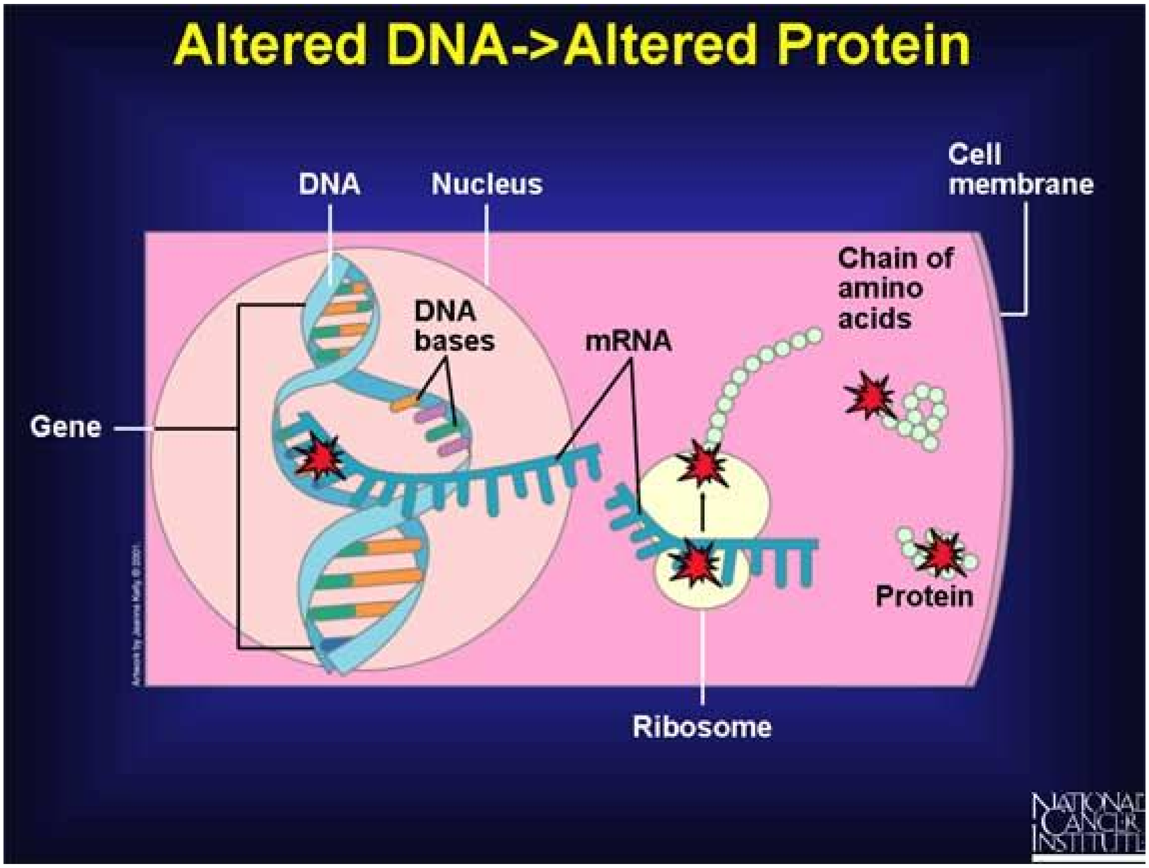
Genes can be altered, or mutated, in many ways.

The most common gene change involves a single base mismatch--a misspelling--placing the wrong base in the DNA. At other times, a single base may be dropped or added. And sometimes large pieces of DNA are mistakenly repeated or deleted.

Altered DNA->Altered Protein

When a gene contains a mutation, the protein encoded by that gene is likely to be abnormal.

Sometimes the protein will be able to function, but imperfectly. In other cases, it will be totally disabled. The outcome depends not only on how it alters a protein's function but also on how vital that particular protein is to survival.

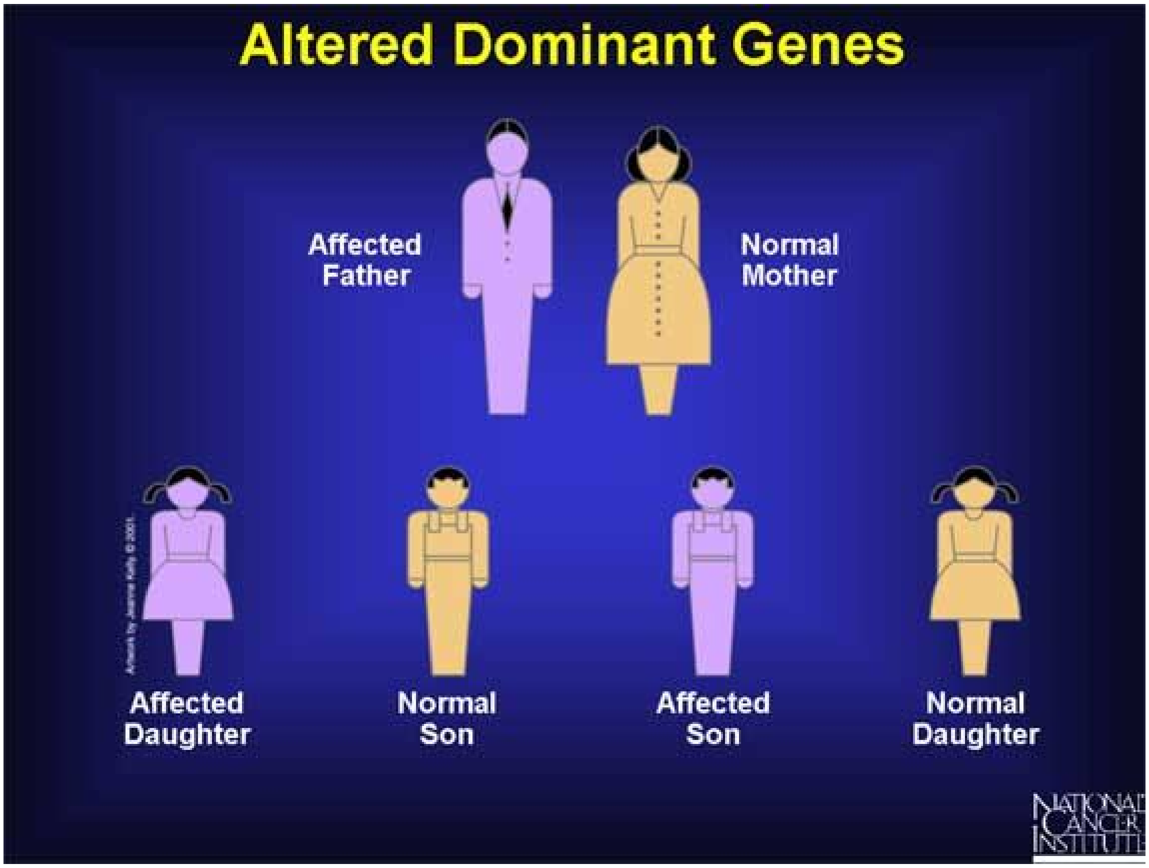


Altered DNA->Altered Protein

Altered Dominant Genes

In dominant genetic disorders, one affected parent has a disease-causing allele that dominates its normal counterpart.

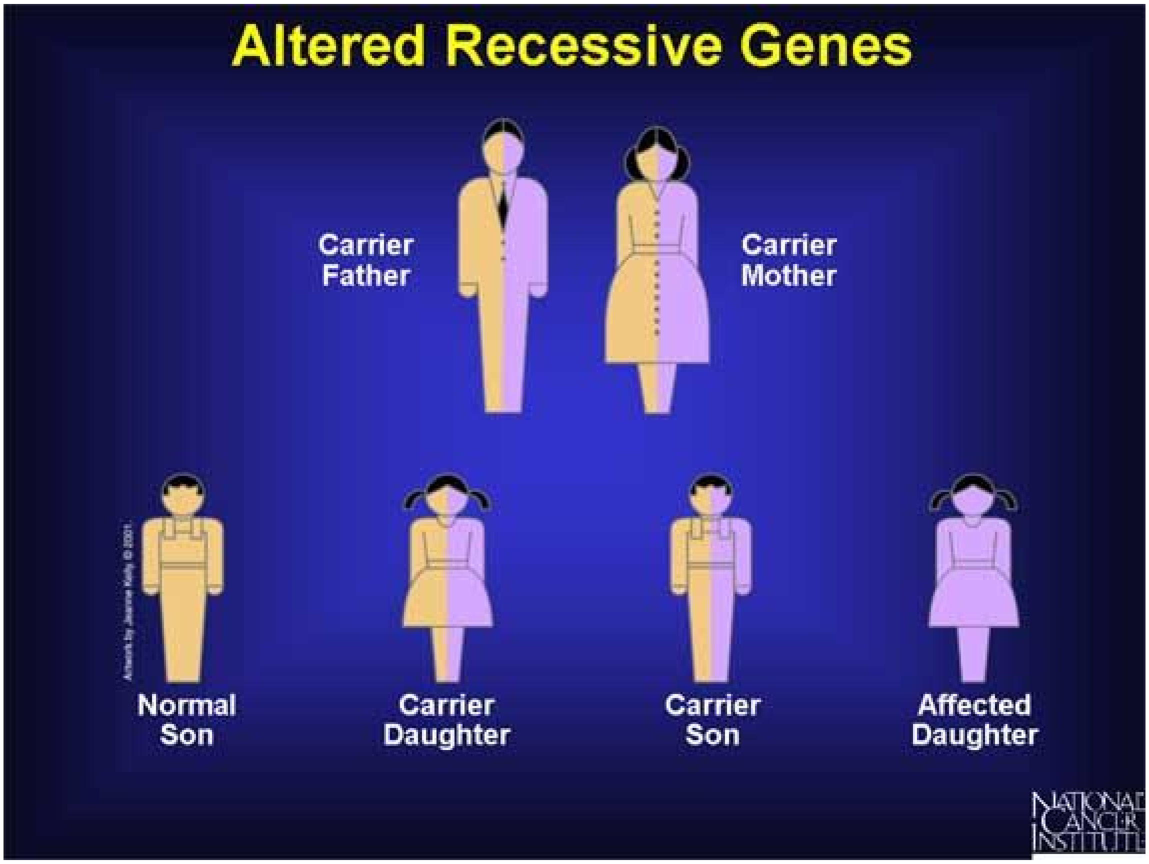
Each child in the family has a 50 percent chance of inheriting the disease allele and the disorder.



Altered Recessive Genes

In diseases associated with altered recessive genes, both parents--though disease free themselves--carry one normal allele and one altered allele.

Each child has one chance in four of inheriting two altered alleles and developing the disorder; one chance in four of inheriting two normal alleles; and two chances in four of inheriting one normal and one altered allele and being a carrier like both parents.



Altered Recessive Genes

Carrier Father

Carrier Mother

Normal Son

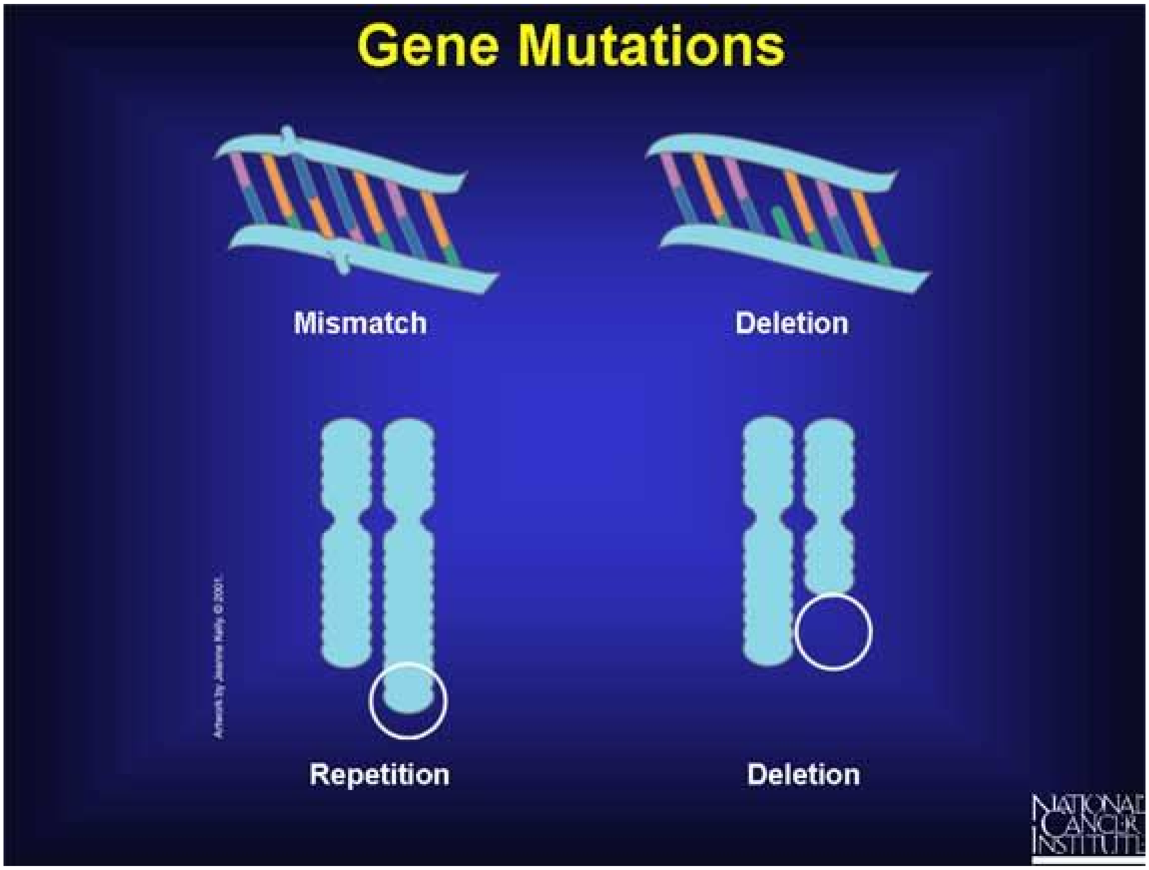
Carrier Daughter

Carrier Son

Affected Daughter

Artwork by Jeanne Kelly © 2001.

NATIONAL CANCER INSTITUTE



Gene Mutations

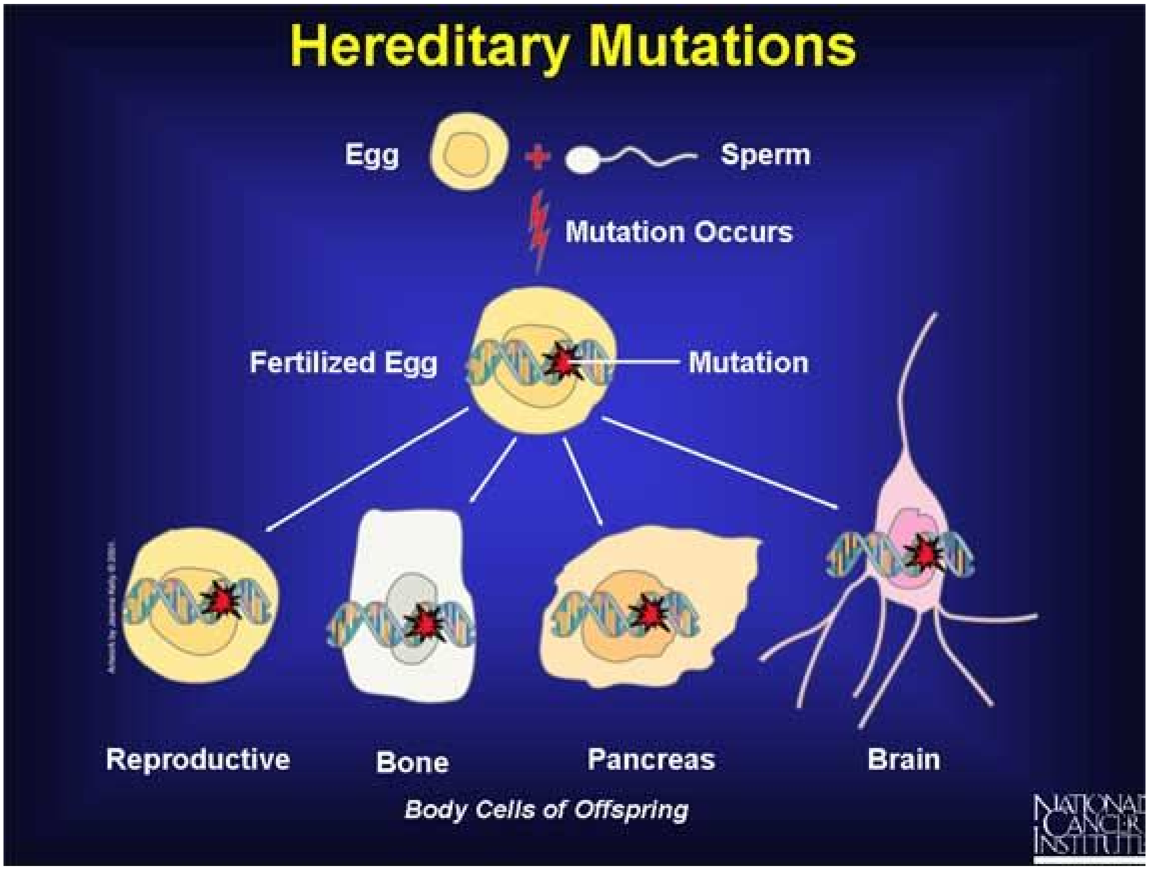
Acquired mutations are changes in DNA that develop throughout a person's lifetime.

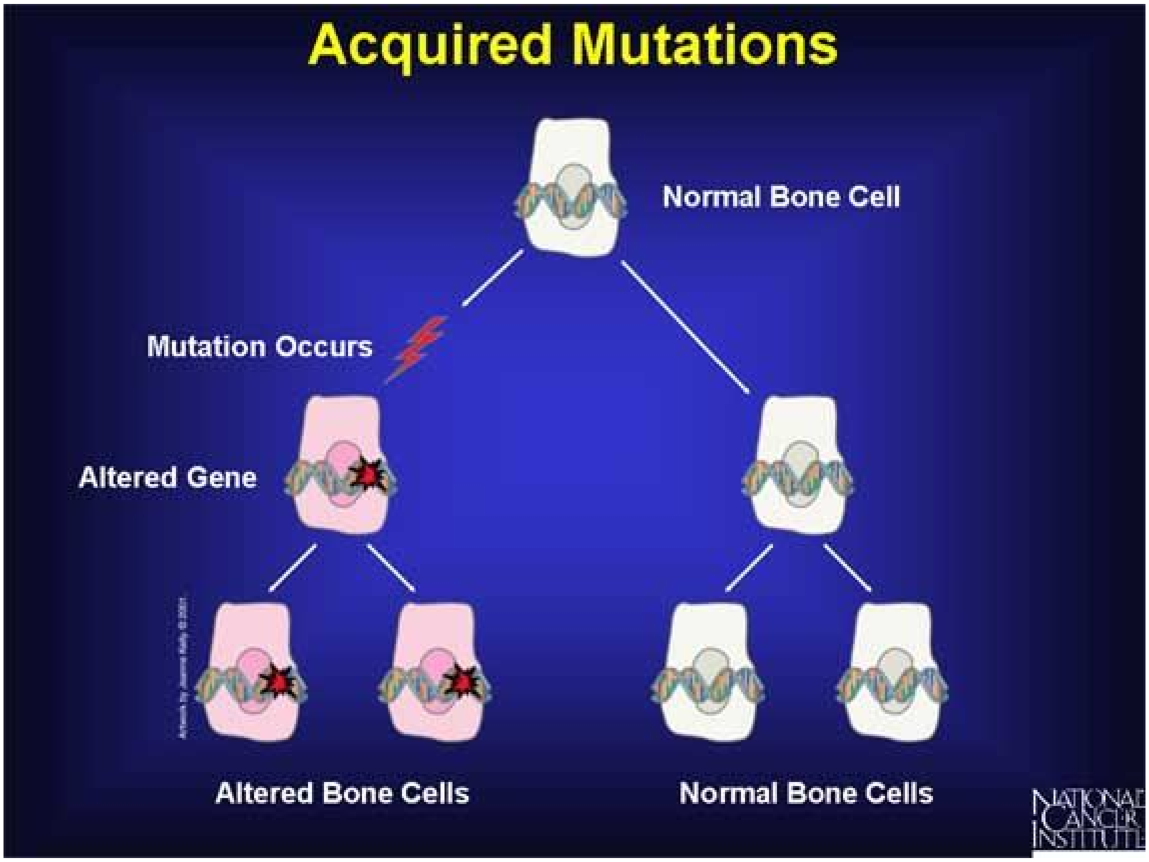
Although mistakes occur in DNA all the time, especially during cell division, a cell has the remarkable ability to fix them. But if DNA repair mechanisms fail, mutations can be passed along to future copies of the altered cell.

Hereditary Mutations

Gene mutations can be either inherited from a parent or acquired.

Hereditary mutations are carried in the DNA of the reproductive cells. When reproductive cells containing mutations combine to produce offspring, the mutation will be in all of the offspring's body cells. The fact that every cell contains the gene change makes it possible to use cheek cells or a blood sample for gene testing.





Acquired Mutations

Normal Bone Cell

Mutation Occurs

Altered Gene

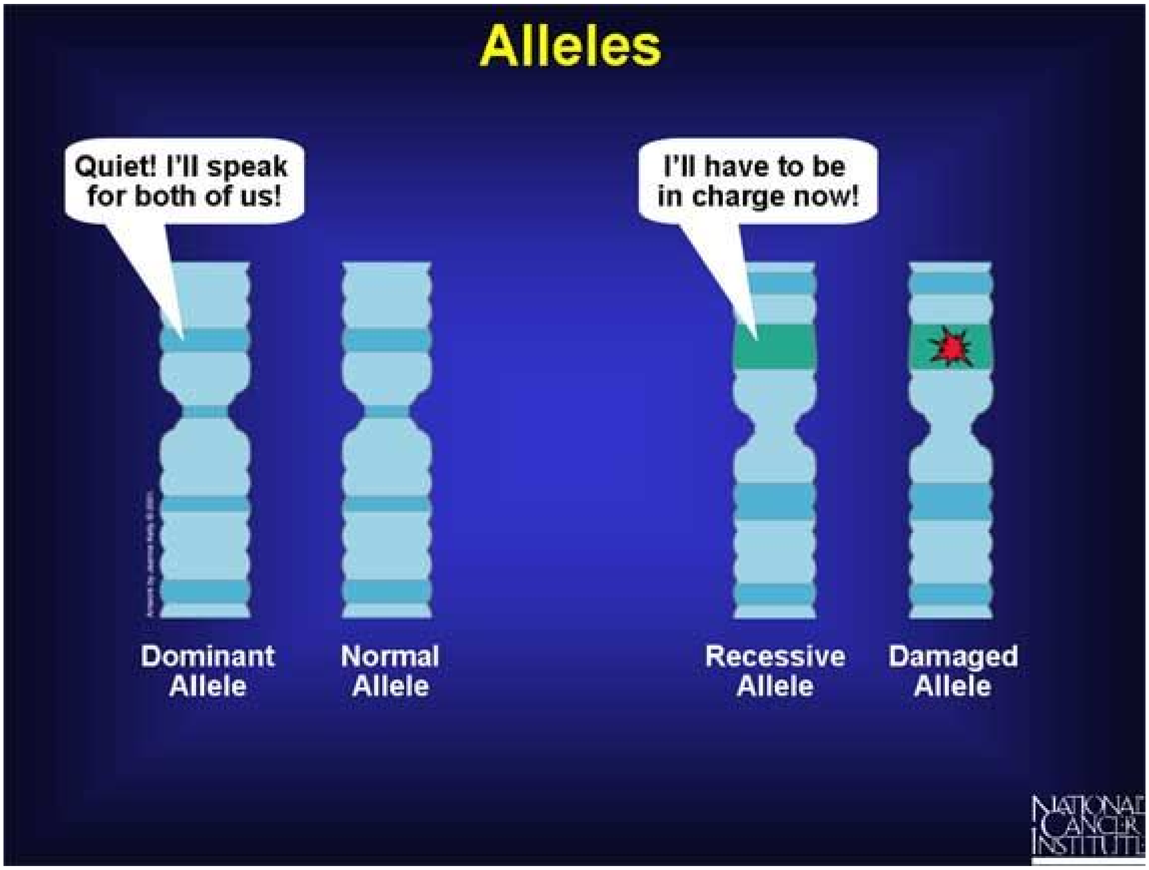
Altered Bone Cells

Normal Bone Cells

< Previous | Index | Next Slide >

Genes come in pairs, with one copy inherited from each parent.

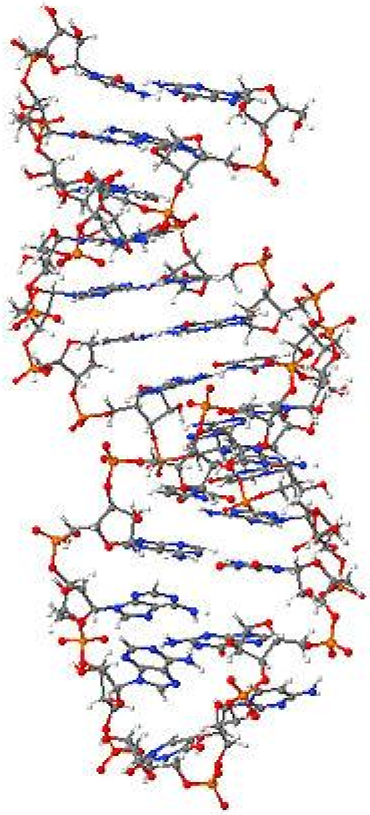
Many genes come in a number of variant forms, known as alleles. A dominant allele prevails over a normal allele. A recessive allele prevails if its counterpart allele on the other chromosome becomes inactivated or lost.



Alleles

Quiet! I'll speak for both of us!
I'll have to be in charge now!
Dominant Allele
Normal Allele
Recessive Allele
Damaged Allele

RNA



A hairpin loop from a pre-mRNA. Notice the single strand with its nitrogen-rich (blue) bases extending from its oxygen-rich (red) backbone.

Ribonucleic acid or **RNA** is a polymer or chain of nucleotide units, each comprising a nitrogenous base (adenine, cytosine, guanine, or uracil), a five-carbon sugar (ribose), and a phosphate group. The sugar and phosphate groups form the polymer's backbone, while the nitrogenous bases extending from the backbone provide RNA's distinctive properties.

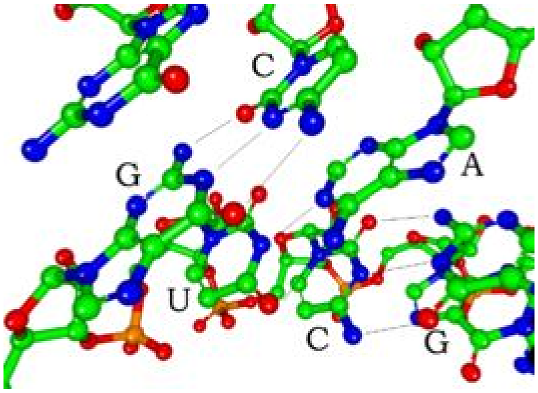
In living cells, RNA in different configurations fulfills several important roles in the process of translating genetic information from deoxyribonucleic acid (DNA) into proteins. One type of RNA (messenger(m) RNA) acts as a messenger between DNA and the protein synthesis complexes known as ribosomes; a second type (ribosomal(r) RNA) forms vital portions of the structure of ribosomes; a third type (transfer(t) RNA) is an essential guide to deliver the appropriate protein building blocks, amino acids, to the ribosome; and other types of RNA, microRNAs (miRNAs) play a role in regulating gene expression, while small nuclear(sn) RNA helps with assuring that mRNA contains no nucleotide units that would lead to formation of a faulty protein. RNA also serves as a genetic blueprint for certain viruses, and some RNA molecules (called ribozymes) are also involved in the catalysis of biochemical reactions.

RNA is very similar to DNA, but differs in a few important structural details. RNA is usually single stranded, while DNA naturally seeks its stable form as a double stranded molecule. RNA nucleotides contain ribose while DNA nucleotides contain the closely related sugar deoxyribose. Furthermore, RNA uses the nucleotide uracil in its composition, instead of the thymine that is present in DNA. RNA is transcribed from DNA by enzymes called RNA polymerases and is generally further processed by other enzymes, some of them guided by non-coding RNAs.

Single-stranded RNA is similar to the protein polymer in its natural propensity to fold back and double up with itself in complex ways assuming a variety of biologically useful configurations.

The connectedness of living organisms can be seen in the ubiquitousness of RNA in living cells and in viruses throughout nature, and in the universal role of RNA in protein synthesis.

Chemical and stereochemical structure



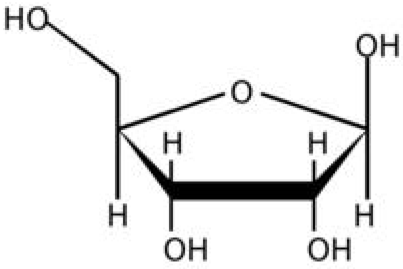
Base-pairing in a siRNA (small interfering RNA) segment, a double-stranded type of RNA. Hydrogen atoms are not shown.

RNA is a **nucleic acid**, a complex, high-molecular-weight macromolecule composed of nucleotide chains whose sequence of bases conveys genetic information.

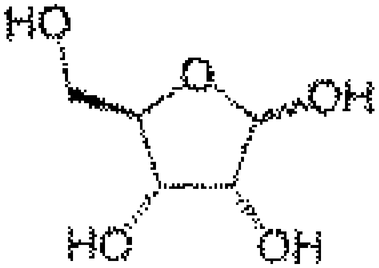
A **nucleotide** is a chemical compound comprising three components: a nitrogen-containing base, a pentose (five-carbon) sugar, and one or more phosphate groups. The nitrogen-containing base of a nucleotide (also called the **nucleobase**) is typically a derivative of either purine or pyrimidine. The most common nucleotide bases are the purines adenine and guanine and the pyrimidines cytosine and thymine (or uracil in RNA).

Nucleic acids are polymers of repeating units (called monomers). Specifically, they often comprise long chains of nucleotide monomers connected by covalent chemical bonds.

RNA molecules may comprise as few as 75 nucleotides or more than 5,000 nucleotides, while a DNA molecule may comprise more than 1,000,000 nucleotide units.



Ribose in acyclic form



A conventional skeletal formula

In RNA, the sugar component, **ribose** is a water-soluable, pentose sugar (monosaccharide with five carbon atoms). Ribose has the chemical formula $C_5H_{10}O_5$.

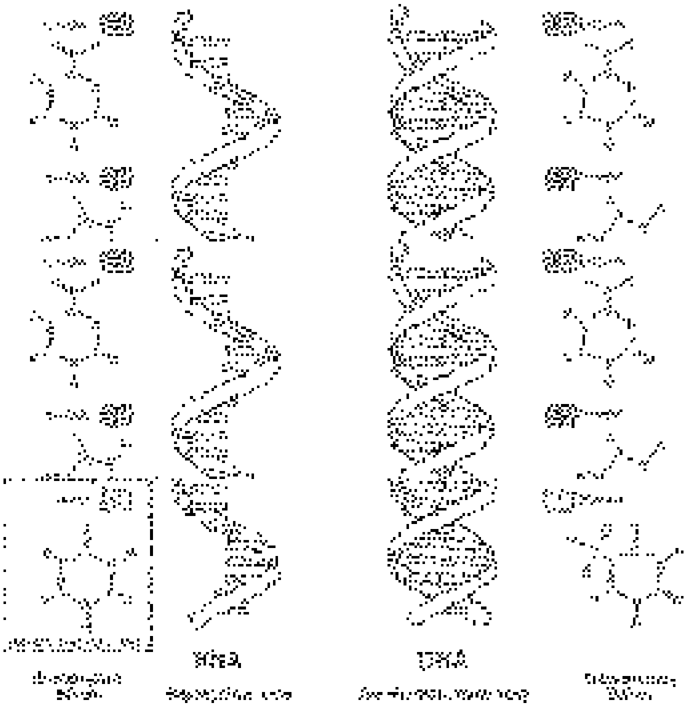
Ribose is an aldopentose, which means a pentose sugar with an aldehyde functional group in position one. An aldehyde group comprises a carbon atom bonded to a hydrogen atom and double-bonded to an oxygen atom (chemical formula O=CH-). Ribose forms a five-member ring with four carbon atoms and one oxygen. Hydroxyl (-OH) groups are attached to three of the carbons. The fourth carbon in the ring (one of the carbon atoms adjacent to the oxygen) has attached to it the fifth carbon atom and a hydroxyl group.

The RNA polymer features a ribose and phosphate backbone with one of four different nucleotide bases—adenine, guanine, cytosine, and uracil—attached to each ribose-phosphate unit.

There are also numerous modified bases and sugars found in RNA that serve many different roles. Pseudouridine (Ψ), in which the linkage between uracil and ribose is changed from a C–N bond to a C–C bond, and ribothymidine (T), are found in various places (most notably in the TΨC loop of tRNA). Another notable modified base is hypoxanthine (a deaminated guanine base whose nucleoside is called inosine). Inosine plays a key role in the Wobble Hypothesis of the genetic code. There are nearly 100 other naturally occurring modified nucleosides, of which pseudouridine and nucleosides with 2'-O-methylribose are by far the most common. The specific roles of many of these modifications in RNA are not fully understood. However, it is notable that in ribosomal RNA, many of the post-translational modifications occur in highly functional regions, such as the peptidyl transferase center and the subunit interface, implying that they are important for normal function.

The most important structural feature of RNA that distinguishes it from DNA is the presence of a hydroxyl group at the 2'-position of the ribose sugar. The presence of this functional group enforces the C3'-endo sugar conformation (as opposed to the C2'-endo conformation of the deoxyribose sugar in DNA) that causes the helix to adopt the A-form geometry rather than the B-form most commonly observed in DNA. This results in a very deep and narrow major groove and a shallow and wide minor groove. A second consequence of the presence of the 2'-hydroxyl group is that in conformationally flexible regions of an RNA molecule (that is, not involved in formation of a double helix), it can chemically attack the adjacent phosphodiester bond to cleave the backbone.

Comparison with DNA



Left: An RNA strand, with its nitrogenous bases. **Right:** Double-stranded DNA.

The most common nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). The main role of DNA is the long-term storage of genetic information. DNA is often compared to a blueprint, since it contains instructions for constructing other components of the cell, such as proteins and RNA molecules. The DNA segments that carry genetic information are called *genes*, but other DNA sequences have structural purposes or are involved in regulating the expression of genetic information. RNA, also, may serve more than one purpose, but it is most commonly identified as the intermediate between the DNA blueprint and the actual workings of the cell, serving as the template for the synthesis of proteins from the genetic information stored in DNA.

RNA and DNA differ in three main ways.

First, unlike DNA which is double-stranded, RNA is intrinsically a single-stranded molecule in most of its biological roles and has a much shorter chain of nucleotides. (While RNA is usually single-stranded, the RNA molecule also quite commonly forms double-helical regions where a given strand has folded back on itself. Double-stranded RNA is found also in certain viruses.)

Secondly, while DNA contains *deoxyribose*, RNA contains *ribose*. There is no hydroxyl group attached to the pentose ring in the 2' position in DNA, whereas RNA has two hydroxyl groups. These hydroxyl groups make RNA less stable than DNA because it is more prone to hydrolysis. ("Deoxy" simply indicates that the sugar lacks an oxygen atom present in ribose, the parent compound.)

Thirdly, the complementary nucleotide to adenine is not thymine, as it is in DNA, but rather uracil, which is an unmethylated form of thymine.

Most biologically active RNAs, including tRNA, rRNA, snRNAs, and other non-coding RNAs (such as the signal recognition particle(SRP) RNAs), contain extensively base paired regions that have folded together to form double stranded helices. Structural analysis of these RNAs reveals that they are highly structured with tremendous variety with collections of short helices packed together into structures much more akin to proteins than to DNA, which is usually limited to long double-stranded helices. Through such a variety of structures, RNAs can achieve chemical catalysis, like enzymes. For instance, determination of the structure of the ribosome—an enzyme that catalyzes peptide bond formation—revealed that its active site is composed entirely of RNA.

Synthesis

Synthesis of RNA is usually catalyzed by an enzyme, RNA polymerase, using DNA as a template. Initiation of synthesis begins with the binding of the enzyme to a promoter sequence in the DNA (usually found "upstream" of a gene). The DNA double helix is unwound by the helicase activity of the enzyme. The enzyme then progresses along the template strand in the 3' -> 5' direction, synthesizing a complementary RNA molecule with elongation occurring in the 5' -> 3' direction. The DNA sequence also dictates where termination of RNA synthesis will occur (Nudler and Gottesman 2002).

There are also a number of RNA-dependent RNA polymerases as well that use RNA as their template for synthesis of a new strand of RNA. For instance, a number of RNA viruses (such as poliovirus) use this type of enzyme to replicate their genetic material (Hansen et al. 1997). Also, it is known that RNA-dependent RNA polymerases are required for the RNA interference pathway in many organisms (Ahlquist 2002).

Biological roles

RNA's great variety of possible structures and chemical properties permits it to perform a much greater diversity of roles in the cell than DNA. Three principal types of RNA are involved in protein synthesis:

- Messenger RNA (mRNA) serves as the template for the synthesis of a protein. It carries information from DNA to the ribosome.
- Transfer RNA (tRNA) is a small chain of nucleotides that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of synthesis. It pairs the amino acid to the appropriate three-nucleotide codon on the mRNA molecule.
- Ribosomal RNA (rRNA) molecules are extremely abundant and make up at least 80 percent of the RNA molecules found in a typical eukaryotic cell. In the cytoplasm, usually three or four rRNA molecules combine with many proteins to perform a structural and essential catalytic role, as components of the ribosome.

RNA also may serve as a catalyst for reactions and as a genetic blueprint, rather than DNA, in various viruses. Some RNA, including tRNA and rRNA, is non-coding in that it is not translated into proteins.

Messenger RNA (mRNA)

Messenger RNA

Messenger RNA is RNA that carries information from DNA to the ribosome sites of protein synthesis in the cell. In eukaryotic cells, once mRNA has been transcribed from DNA, it is "processed" before being exported from the nucleus into the cytoplasm, where it is bound to ribosomes and translated into its corresponding protein form with the help of tRNA. In prokaryotic cells, which do not have nucleus and cytoplasm compartments, mRNA can bind to ribosomes while it is being transcribed from DNA. After a certain amount of time the message degrades into its component nucleotides, usually with the assistance of ribonucleases.

Non-coding RNA

RNA genes (also known as non-coding RNA or small RNA) are genes that encode RNA that is not translated into a protein. The most prominent examples of RNA genes are those coding for transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation. Two other groups of non-coding RNA are microRNAs (miRNA) which regulate the expression of genes through a process called RNA interference (RNAi), and small nuclear RNAs (snRNA), a diverse class that includes for example the RNAs that form spliceosomes that excise introns from pre-mRNA (Berg et al. 2002).

Transfer RNA (tRNA)

Transfer RNA

Transfer RNA is a small RNA chain of about 74-95 nucleotides that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis, during translation. It has sites for amino-acid attachment and an anticodon region for codon recognition that binds to a specific sequence on the messenger RNA chain through hydrogen bonding. It is a type of non-coding RNA.

Ribosomal RNA (rRNA)

Ribosomal RNA

Ribosomal RNA is the catalytic component of the ribosomes, the protein synthesis factories in the cell. Eukaryotic ribosomes contain four different rRNA molecules: 18S, 5.8S, 28S, and 5S rRNA. Three of the rRNA molecules are synthesized in the nucleolus, and one is synthesized elsewhere. rRNA molecules are extremely abundant and make up at least 80 percent of the RNA molecules found in a typical eukaryotic cell.

Catalytic RNA

Ribozyme

Although RNA contains only four bases, in comparison to the twenty-odd amino acids commonly found in proteins, certain RNAs (called ribozymes) are still able to catalyze chemical reactions. These include cutting and ligating other RNA molecules, and also the catalysis of peptide bond formation in the ribosome.

Genetic blueprint in some viruses

Some viruses contain either single-stranded or double-stranded RNA as their source of genetic information. Retroviruses, for example, store their genetic information as RNA, though they replicate in their hosts via a DNA intermediate. Once in the host's cell, the RNA strands undergo reverse transcription to DNA in the cytosol and are integrated into the host's genome. Human immunodeficiency virus (or HIV) is a retrovirus thought to cause acquired immune deficiency syndrome (AIDS), a condition in which the human immune system begins to fail, leading to life-threatening opportunistic infections.

Double-stranded RNA (dsRNA) is RNA with two complementary strands, similar to the DNA found in all cells. dsRNA forms the genetic material of some viruses called dsRNA viruses. In eukaryotes, long RNA such as viral RNA can trigger RNA interference, where short dsRNA molecules called siRNAs (small interfering RNAs) can cause enzymes to break down specific mRNAs or silence the expression of genes. siRNA can also increase the transcription of a gene, a process called RNA activation (Doran 2007). siRNA is often confused with miRNA; siRNAs are double-stranded, whereas miRNAs are single-stranded.

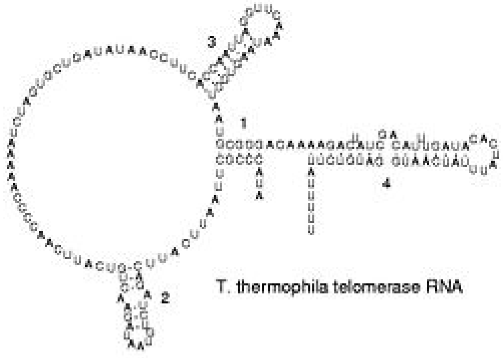
RNA world hypothesis

The RNA world hypothesis proposes that the earliest forms of life relied on RNA both to carry genetic information (like DNA does now) and to catalyze biochemical reactions like an enzyme. According to this hypothesis, descendants of these early lifeforms gradually integrated DNA and proteins into their metabolism.

In the 1980s, scientists discovered that certain RNA molecules (called *ribozymes*) may function as enzymes, whereas previously only proteins were believed to have catalytic ability. Many natural ribozymes catalyze either their own cleavage or the cleavage of other RNAs, but they have also been found to catalyze the aminotransferase activity of the ribosome.

The discovery of ribozymes provides a possible explanation for how early RNA molecules might have first catalyzed their own replication and developed a range of enzymatic activities. Known as the RNA world hypothesis, this explanation posits that RNA evolved before either DNA or proteins from free-floating nucleotides in the early "primordial soup." In their function as enzymes, RNA molecules might have begun to catalyze the synthesis of proteins, which are more versatile than RNA, from amino acid molecules. Next, DNA might have been formed by reverse transcription of RNA, with DNA eventually replacing RNA as the storage form of genetic material. Although there are remaining difficulties with the RNA world hypothesis, it remains as a possible key to understanding the origin and development of the multi-functional nature of nucleic acids, the interconnectedness of life, and its common origins.

RNA secondary structures



Secondary structure of an RNA from a telomerase.

The functional form of single stranded RNA molecules, just like proteins, frequently requires a specific tertiary structure. The scaffold for this structure is provided by secondary structural elements, which arise through the formation of hydrogen bonds within the interfolded molecule. This leads to several recognizable "domains" of secondary structure like hairpin loops, bulges, and internal loops. The secondary structure of RNA molecules can be predicted computationally by calculating the minimum free energies (MFE) structure for all different combinations of hydrogen bondings and domains (Mathews et al. 2004). There has been a significant amount of research directed at the RNA structure prediction problem.

History

Nucleic acids were discovered in 1868 by Johann Friedrich Miescher (1844-1895), who called the material 'nuclein' since it was found in the nucleus. It was later discovered that prokaryotic cells, which do not have a nucleus, also contain nucleic acids.

The role of RNA in protein synthesis had been suspected since 1939, based on experiments carried out by Torbjörn Caspersson, Jean Brachet, and Jack Schultz. Hubert Chantrenne elucidated the messenger role played by RNA in the synthesis of proteins in ribosomes. Finally, Severo Ochoa discovered RNA, winning Ochoa the 1959 Nobel Prize for Medicine. The sequence of the 77 nucleotides of a yeast RNA was found by Robert W. Holley in 1964, winning Holley the 1968 Nobel Prize for Medicine. In 1976, Walter Fiers and his team at the University of Ghent determined the complete nucleotide sequence of bacteriophage MS2-RNA (Fiers et al. 1976).

List of RNA types

Type	Function	Distribution
mRNA	Codes for protein	All cells
rRNA	Translation	All cells
tRNA	Translation	All cells
snRNA	RNA modification	All cells
snoRNA	RNA modification	All cells
miRNA	Gene regulation	Eukaryotes
piRNA	Gene regulation	Animal germline cells
siRNA	Gene regulation	Eukaryotes
Antisense mRNA	Preventing translation	Bacteria
tmRNA	Terminating translation	Bacteria
SRP RNA	mRNA tagging for export	All cells
Ribozyme	Catalysis	All cells
Transposon	Self-propagating	All cells
Viroid	Self-propagating	Infected plants

In addition, the genome of many types of viruses consists of RNA, namely:

- Double-stranded RNA viruses
- Positive-sense RNA viruses

- Negative-sense RNA viruses
- Retroviruses
- Satellite viruses

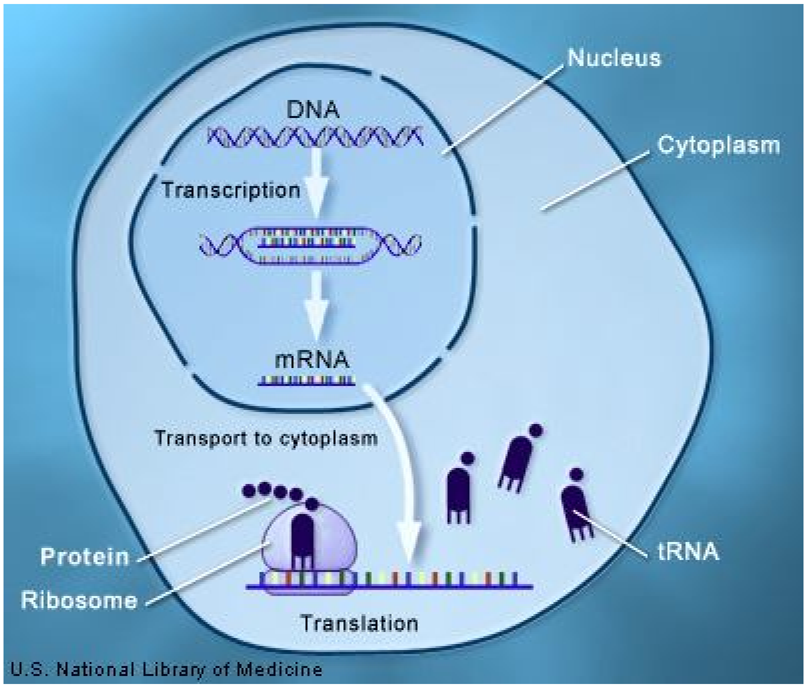
How do genes direct the production of proteins?

Most genes contain the information needed to make functional molecules called proteins. (A few genes produce other molecules that help the cell assemble proteins.) The journey from gene to protein is complex and tightly controlled within each cell. It consists of two major steps: transcription and translation. Together, transcription and translation are known as gene expression.

During the process of transcription, the information stored in a gene's DNA is transferred to a similar molecule called RNA (ribonucleic acid) in the cell nucleus. Both RNA and DNA are made up of a chain of nucleotide bases, but they have slightly different chemical properties. The type of RNA that contains the information for making a protein is called messenger RNA (mRNA) because it carries the information, or message, from the DNA out of the nucleus into the cytoplasm.

Translation, the second step in getting from a gene to a protein, takes place in the cytoplasm. The mRNA interacts with a specialized complex called a ribosome, which "reads" the sequence of mRNA bases. Each sequence of three bases, called a codon, usually codes for one particular amino acid. (Amino acids are the building blocks of proteins.) A type of RNA called transfer RNA (tRNA) assembles the protein, one amino acid at a time. Protein assembly continues until the ribosome encounters a "stop" codon (a sequence of three bases that does not code for an amino acid).

The flow of information from DNA to RNA to proteins is one of the fundamental principles of molecular biology. It is so important that it is sometimes called the "central dogma."



U.S. National Library of Medicine

Through the processes of transcription and translation, information from genes is used to make proteins.

Genetic linkage occurs when particular alleles are inherited together. Typically, an organism can pass on an allele without regard to which allele was passed on for a different gene. This is because chromosomes are sorted randomly during meiosis. However, alleles that are on the same chromosome are more likely to be inherited together, and are said to be linked.

Because there is some crossing over of DNA when the chromosomes segregate, alleles on the same chromosome can be separated and go to different cells. There is a greater probability of this happening if the alleles are far apart on the chromosome, as it is more likely that a cross-over will occur between them.

The physical distance between two genes can be calculated using the offspring of an organism showing two linked genetic traits, and finding the percentage of the offspring where the two traits don't run together. The higher the percentage of descendence that doesn't show both traits, the further apart on the chromosome they are.

A study of the linkages between many genes enables the creation of a **linkage map** or genetic map.

Among individuals of an experimental population or species, some phenotypes or traits occur randomly with respect to one another in a manner known as independent assortment. Today scientists understand that independent assortment occurs when the genes affecting the phenotypes are found on different chromosomes.

An exception to independent assortment develops when genes appear near one another on the same chromosome. When genes occur on the same chromosome, they are usually inherited as a single unit. Genes inherited in this way are said to be linked. For example, in fruit flies the genes affecting eye color and wing length are inherited together because they appear on the same chromosome.

But in many cases, even genes on the same chromosome that are inherited together produce offspring with unexpected allele combinations. This results from a process called crossing over. Sometimes at the beginning of meiosis, a chromosome pair (made up of a chromosome from the mother and a chromosome from the father) may 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.

Genetic linkage was first discovered by the British geneticists William Bateson and Reginald Punnett shortly after Mendel's laws were rediscovered.

Genetic marker

A **genetic marker** is a known DNA sequence (e. g. a gene or part of gene) that can be identified by a simple assay associated with a certain phenotype.

A genetic marker may be a short DNA sequence, such as a sequence surrounding a single base-pair change (single nucleotide polymorphism), or a long one, like microsatellites.

Uses

Genetic markers can be used to study the relationship between an inherited disease and its genetic cause (for example, a particular mutation of a gene that results in a defective protein). It is known that pieces of DNA that lie near each other on a chromosome tend to be inherited together. This property enables the use of a marker, which can then be used to determine the precise inheritance pattern of the gene that has not yet been exactly localized.

Genetic markers have to be easily identifiable, associated with a specific locus, and highly polymorphic, because homozygotes do not provide any information. Detection of the marker can be direct by DNA sequencing, or indirect using allozymes.

Some of the methods used to study the genome or phylogenetics are RFLP, Amplified fragment length polymorphism AFLP, RAPD, SSR.

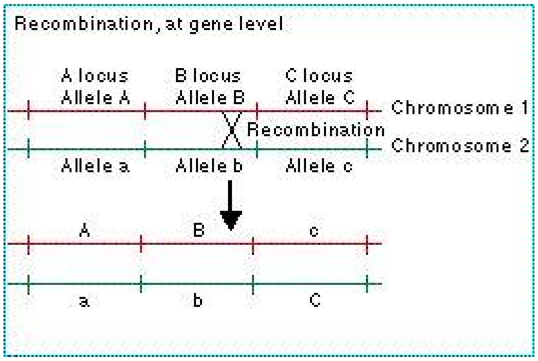
Insulin production

Genetic markers also play a role in genetic engineering, as they can be used to produce normal, functioning proteins to replace defective ones. The damaged or faulty section of DNA is removed and replaced with the identical, but functioning, gene sequence from another source.

This is done by removal of the faulty section of DNA and its replacement with the functioning gene from another source, usually a human donor. These gene sections are placed in solution with bacterial cells, a small number of which take up the genetic material and reproduce the new DNA sequence. Engineers need to know which bacteria have been successful in duplicating these genes so another gene is added, altering the bacteria's resistance to antibiotics. Replica plating or a fermenter is used to grow enough bacteria to test resistance to antibiotics. It is important that the cultures are not mixed.

This process can be used as a treatment for diabetes mellitus. Bacterial DNA often has two resistency genes: one for tetracycline and one for ampicillin. The insulin gene can be inserted in the middle of the ampicillin gene after it has been removed using restriction endonucleases. If the gene has been taken up, the bacteria both produces insulin and is also no longer ampicillin resistant. The bacteria are then allowed to grow on an agar plate containing a culture medium. The bacteria grow and produce colonies on the agar jelly. A piece of filter paper can be placed onto the top of this agar plate so that the exact positions of the colonies are remembered. This produces a copy which can then be transferred onto a second agar plate containing ampicillin. All of the bacteria that are not resistant to ampicillin will die. These locations on the second plate show the places on the first plate where bacteria are not resistant and therefore produce insulin.

Recombination



Recombination is an event, occurring by the crossing-over of chromosomes during meiosis, in which DNA is exchanged between a pair of chromosomes. Thus two genes that were previously unlinked, being on separate chromosomes, can become linked because of recombination; and vice versa: linked genes may become unlinked.

Like mutation, recombination is an important source of new variation for natural selection to work upon. However, also like mutation, recombination places a genetic load upon the population.

As David Haig explains, the reason for recombination has been a puzzle for evolutionary biologists.

Genetic recombination

Genetic recombination is a process by which a molecule of nucleic acid (usually DNA; but can also be RNA) is broken and then joined to a different DNA molecule. Recombination can occur between similar molecules of DNA, as in homologous recombination, or dissimilar molecules of DNA as in non-homologous end joining. Recombination is a common method of DNA repair in both bacteria and eukaryotes. In eukaryotes, recombination occurs in meiosis as a way of facilitating chromosomal crossover. The crossover process leads to offspring having different combinations of genes from their parents, and can occasionally produce new chimeric alleles. In organisms with an adaptive immune system, a type of genetic recombination called V(D)J recombination helps immune cells rapidly diversify and adapt to recognize new pathogens. The shuffling of genes brought about by genetic recombination is thought to have many advantages, as it is a major engine of genetic variation and also allows asexually reproducing organisms to avoid Muller's ratchet.

In genetic engineering, recombination can also refer to artificial and deliberate recombination of disparate pieces of DNA, often from different organisms, creating what is called recombinant DNA. A prime example of such a use of genetic recombination is gene targeting, which can be used to add, delete or otherwise change an organism's genes. This technique is important to biomedical researchers as it allows them to study the effects of specific genes. Techniques based on genetic recombination are also applied in protein engineering to develop new proteins of biological interest.

Genetic recombination is catalyzed by many different enzymes, called *recombinases*. RecA, the chief recombinase found in *Escherichia coli*, is responsible for the repair of DNA double strand breaks (DSBs). In yeast and other eukaryotic organisms there are two recombinases required for repairing DSBs. The RAD51 protein is required for mitotic and meiotic recombination and the DMC1 protein is specific to meiotic recombination.

Chromosomal crossover

Chromosomal crossover

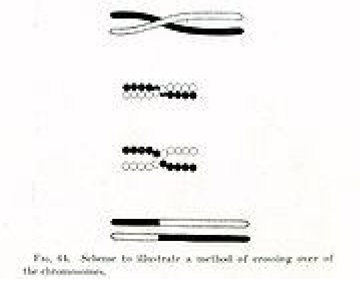


Fig. 64. Scheme to illustrate a method of crossing over of the chromosomes.

Thomas Hunt Morgan's illustration of crossing over (1916)

Chromosomal crossover refers to recombination between the paired chromosomes inherited from each of one's parents, generally occurring during meiosis. During prophase I the four available chromatids are in tight formation with one another. While in this formation, homologous sites on two chromatids can mesh with one another, and may exchange genetic information.

Because recombination can occur with small probability at any location along chromosome, the frequency of recombination between two locations depends on their distance. Therefore, for genes sufficiently distant on the same chromosome the amount of crossover is high enough to destroy the correlation between alleles.

Gene conversion

Gene conversion

In gene conversion, a section of genetic material is copied from one chromosome to another, but leaves the donating chromosome unchanged.

Nonhomologous recombination

Non-homologous end joining

Recombination can occur between DNA sequences that contain no sequence homology. This is referred to as *nonhomologous recombination* or nonhomologous end joining.

In B cells

Immunoglobulin class switching

B cells of the immune system perform genetic recombination, called immunoglobulin class switching. It is a biological mechanism that changes an antibody from one class to another, for example, from an isotype called IgM to an isotype called IgG.

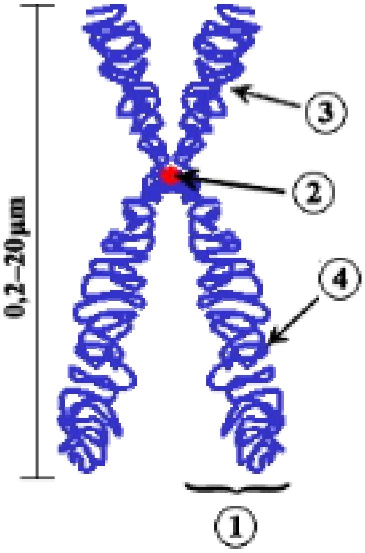
Linkage mapping

The observations by Thomas Hunt Morgan 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. Morgan's student Alfred Sturtevant developed the first genetic map, also called a linkage map.

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. 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 mapping is critical for identifying the location of genes that cause genetic diseases. In a normal population, genetic traits and markers will occur in all possible combinations with the frequencies of combinations determined by the frequencies of the individual genes. For example, if alleles *A* and *a* occur with frequency 90% and 10%, and alleles *B* and *b* at a different genetic locus occur with frequencies 70% and 30%, the frequency of individuals having the combination *AB* would be 63%, the product of the frequencies of *A* and *B*, regardless of how close together the genes are. However, if a mutation in gene *B* that causes some disease happened recently in a particular subpopulation, it almost always occurs with a particur allele of gene *A* if the individual in which the mutation occurred had that variant of gene *A* and there have not been sufficient generations for recombination to happen between them (presumably due to tight linkage on the genetic map). In this case, called linkage disequilibrium, it is possible to search potential markers in the subpopulation and identify which marker the mutation is close to, thus determining the mutation's location on the map and identifying the gene at which the mutation occurred. Once the gene has been identified, it can be targeted to identify ways to mitigate the disease.

Chromosome



A scheme of a condensed (metaphase) 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.

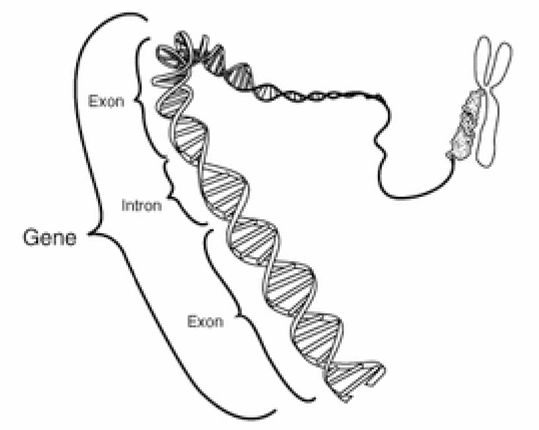
Chromosomes are organized structures of DNA and proteins that are found in cells. Chromosomes contain a single continuous piece of DNA, which contains many genes, regulatory elements and other nucleotide sequences. Chromosomes also contain DNA-bound proteins, which serve to package the DNA and control its functions. The word *chromosome* comes from the Greek χρῶμα (*chroma*, color) and σῶμα (*soma*, body) due to their property of being stained very strongly by some dyes.

Chromosomes vary extensively between different organisms. The DNA molecule may be circular or linear, and can contain anything from tens of kilobase pairs to hundreds of megabase pairs. Typically eukaryotic cells (cells with nuclei) have large linear chromosomes and prokaryotic cells (cells without nuclei) 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 massively-long DNA molecules to fit into the cell nucleus. The structure of chromatin varies through the cell cycle, and is responsible for the organisation of chromosomes into the classic four-arm structure during mitosis and meiosis.

"Chromosome" is a rather loosely defined term. In prokaryotes, a small circular DNA molecule may be called either a plasmid or a small chromosome. These small circular genomes are also found in mitochondria and chloroplasts, reflecting their bacterial origins. The simplest chromosomes are found in viruses: these DNA or RNA molecules are short linear or circular chromosomes that often lack any structural proteins.

Genes



This stylistic schematic diagram shows a gene in relation to the double helix structure of DNA and to a chromosome (right). Introns are regions often found in eukaryote genes which are removed in the splicing process: only the exons encode the protein. This diagram labels a region of only 40 or so bases as a gene. In reality many genes are much larger, as are introns and exons.

Genes are the units of heredity in living organisms. They are encoded in the organism's genetic material (usually DNA or RNA), and control the development and behavior of the organism. During reproduction, the genetic material is passed on from the parent(s) to the offspring. Genetic material can also be passed between un-related individuals (e.g. via transfection, or on viruses). Genes encode the information necessary to construct the chemicals (proteins etc.) needed for the organism to function.

The word "gene" (coined 1909 by Danish botanist Wilhelm Johannsen) comes from the Greek *genos* ("origin") and is shared by many disciplines, including classical genetics, molecular genetics, evolutionary biology and population genetics. Because each discipline models the biology of life differently, the usage of the word gene varies between disciplines. It may refer to either material or conceptual entities.

Following the discovery that DNA is the genetic material, and with the growth of biotechnology and the project to sequence the human genome, the common usage of the word "gene" has increasingly reflected its meaning in molecular biology, namely the segments of DNA which cells transcribe into RNA and translate, at least in part, into proteins. The Sequence Ontology project defines a gene as: "A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions".

In common speech, "gene" is often used to refer to the hereditary cause of a trait, disease or condition—as in "the gene for obesity." Speaking more precisely, a biologist might refer to an allele or a mutation that *has been implicated in* or *is associated with* obesity. This is because biologists know that many factors other than genes decide whether a person is obese or not: eating habits, exercise, prenatal environment, upbringing, culture and the availability of food, for example.

Moreover, it is very unlikely that variations within a single gene—or single genetic locus—fully determine one's genetic predisposition for obesity. These aspects of inheritance—the interplay between genes and environment, the influence of many genes—appear to be the norm with regard to many and perhaps most ("complex" or "multi-factoral") traits. The term phenotype refers to the characteristics that result from this interplay (see genotype-phenotype distinction).

Properties of genes

In molecular biology, a gene is considered to be the region of DNA (or RNA, in the case of some viruses) that determines the structure of a protein (the coding sequence), together with the region of DNA that controls when and where the protein will be produced (the regulatory sequence). The genetic code determines how the coding DNA sequence is converted in to a protein sequence (transcription and translation). The genetic code is essentially the same for all known life, from bacteria to humans.

Through the proteins they encode, genes govern the cells in which they reside. In multicellular organisms they control the development of the individual from the fertilized egg and the day-to-day functions of the cells that make up tissues and organs. The instrumental roles of their protein products range from mechanical support of the cell structure to the transportation and manufacture of other molecules and to the regulation of other proteins' activities.

The genes that exist today are those that have reproduced successfully in the past. Often, many individual organisms share a gene; thus, the death of an individual need not mean the extinction of the gene. Indeed, if the sacrifice of one individual enhances the survivability of other individuals with the same gene, the death of an individual may enhance the overall survival of the gene. This is the basis of the selfish gene view, popularized by Richard Dawkins. He points out in his book, *The Selfish Gene*, that to be successful genes need have no other "purpose" than to propagate themselves, even at the expense of their host organism's welfare. A human that behaved in such a way would be described as "selfish," although ironically a selfish gene may promote altruistic behaviours. According to Dawkins, the possibly disappointing answer to the question "what is the meaning of life?" may be "the survival and perpetuation of ribonucleic acids and their associated proteins".

Types of genes

Due to rare, spontaneous errors (e.g. in DNA replication) mutations in the sequence of a gene may arise. Once propagated to the next generation, this mutation may lead to variations within a species' population. Variants of a single gene are known as alleles, and differences in alleles may give rise to differences in traits, for example eye colour. A gene's most common allele is called the wild type allele, and rare alleles are called mutants.

In most cases, RNA is an intermediate product in the process of manufacturing proteins from genes. However, for some gene sequences, the RNA molecules are the actual functional products. For example, RNAs known as ribozymes are capable of enzymatic function, and small interfering RNAs have a regulatory role. The DNA sequences from which such RNAs are transcribed are known as non-coding RNA, or RNA genes.

Most living organisms carry their genes and transmit them to offspring as DNA, but some viruses carry only RNA. Because they use RNA, their cellular hosts may synthesize their proteins as soon as they are infected and without the delay in waiting for transcription. On the other hand, RNA retroviruses, such as AIDS, require the reverse transcription of their genome from RNA into DNA before their proteins can be synthesized.

Human gene nomenclature

For each known human gene the HUGO Gene Nomenclature Committee (HGNC) approve a gene name and symbol (short-form abbreviation). All approved symbols are stored in the HGNC Database. Each symbol is unique and each gene is only given one approved gene symbol. It is necessary to provide a unique symbol for each gene so that people can talk about them. This also facilitates electronic data retrieval from publications. In preference each symbol maintains parallel construction in different members of a gene family and can be used in other species, especially the mouse.

Typical numbers of genes in an organism

organism	genes	base pairs
Plant	<50,000	$<10^{11}$
Human, mouse or rat	25,000	3×10^{9}
Fruit Fly	13,767	1.3×10^{8}
Honey bee	15,000	3×10^{8}
Worm	19,000	9.7×10^{7}
Fungus	6,000	1.3×10^{7}
Bacterium	500–6,000	5×10^{5}–10^{7}
Mycoplasma genitalium	500	580,000
DNA virus	10–900	5,000–800,000
RNA virus	1–25	1,000–23,000
Viroid	0–1	~500

The shown table gives typical numbers of genes and genome size for some organisms. Estimates of the number of genes in an organism are somewhat controversial because they depend on the discovery of genes, and no techniques currently exist to prove that a DNA sequence contains no gene. (In early genetics, genes could be identified only if there were mutations, or alleles.) Nonetheless, estimates are made based on current knowledge.

Chemistry and function of genes

Chemical structure of a gene

Four kinds of sequentially linked nucleotides compose a DNA molecule or strand (more at DNA). These four nucleotides constitute the genetic alphabet. A sequence of three consecutive nucleotides, called a codon, is the protein-coding vocabulary. The sequence of codons in a gene specifies the amino-acid sequence of the protein it encodes.

In most eukaryotic species, very little of the DNA in the genome encodes proteins, and the genes may be separated by vast sequences of so-called junk DNA. Moreover, the genes are often fragmented internally by non-coding sequences called introns, which can be many times longer than the coding sequence. Introns are removed on the heels of transcription by splicing. In the primary molecular sense, they represent parts of a gene, however.

All the genes and intervening DNA together make up the genome of an organism, which in many species is divided among several chromosomes and typically present in two or more copies. The location (or locus) of a gene and the chromosome on which it is situated is in a sense arbitrary. Genes that appear together on the chromosomes of one species, such as humans, may appear on separate chromosomes in another species, such as mice. Two genes positioned near one another on a chromosome may encode proteins that figure in the same cellular process or in completely unrelated processes. As an example of the former, many of the genes involved in spermatogenesis reside together on the Y chromosome.

Many species carry more than one copy of their genome within each of their somatic cells. These organisms are called diploid if they have two copies or polyploid if they have more than two copies. In such organisms, the copies are practically never identical. With respect to each gene, the copies that an individual possesses are liable to be distinct alleles, which may act synergistically or antagonistically to generate a trait or phenotype. The ways that gene copies interact are explained by chemical dominance relationships (more at genetics, allele).

Expression of molecular genes

For various reasons, the relationship between DNA strand and a phenotype trait is not direct. The same DNA strand in two different individuals may result in different traits because of the effect of other DNA strands or the environment.

- The DNA strand is expressed into a trait only if it is transcribed to RNA. Because the transcription starts from a specific base-pair sequence (a promoter) and stops at another (a terminator), our DNA strand needs to be correctly placed between the two. If not, it is considered as junk DNA, and is not expressed.

- Cells regulate the activity of genes in part by increasing or decreasing their rate of transcription. Over the short term, this regulation occurs through the binding or unbinding of proteins, known as transcription factors, to specific non-coding DNA sequences called regulatory elements. Therefore, to be expressed, our DNA strand needs to be properly regulated by other DNA strands.

- The DNA strand may also be silenced through DNA methylation or by chemical changes to the protein components of chromosomes (see histone). This is a permanent form of regulation of the transcription.

- The RNA is often edited before its translation into a protein. Eukaryotic cells splice the transcripts of a gene, by keeping the exons and removing the introns. Therefore, the DNA strand needs to be in an exon to be expressed. Because of the complexity of the splicing process, one transcribed RNA may be spliced in alternate ways to produce not one but a variety of proteins (alternative splicing) from one pre-mRNA. Prokaryotes produce a similar effect by shifting reading frames during translation.

- The translation of RNA into a protein also starts with a specific start and stop sequence.

- Once produced, the protein interacts with the many other proteins in the cell, according to the cell metabolism. This interaction finally produces the trait.

This complex process helps explain the different meanings of "gene":

- a nucleotide sequence in a DNA strand;
- or the transcribed RNA, prior to splicing;
- or the transcribed RNA after splicing, i.e. without the introns

The latter meaning of gene is the result of more "material entity" than the first one.

Mutations and evolution

Just as there are many factors influencing the expression of a particular DNA strand, there are many ways to have genetic mutations.

For example, natural variations within *regulatory sequences* appear to underlie many of the heritable characteristics seen in organisms. The influence of such variations on the trajectory of evolution through natural selection may be as large as or larger than variation in sequences that encode proteins. Thus, though regulatory elements are often distinguished from genes in molecular biology, in effect they satisfy the shared and historical sense of the word. Indeed, a breeder or geneticist, in following the inheritance pattern of a trait, has no immediate way to know whether this pattern arises from coding sequences or regulatory sequences. Typically, he or she will simply attribute it to variations within a gene.

Errors during DNA replication may lead to the duplication of a gene, which may diverge over time. Though the two sequences may remain the same, or be only slightly altered, they are typically regarded as separate genes (i.e. not as alleles of the same gene). The same is true when duplicate sequences appear in different species. Yet, though the alleles of a gene differ in sequence, nevertheless they are regarded as a single gene (occupying a single locus).

History

The existence of genes was first suggested by Gregor Mendel, who, in the 1860s, studied inheritance in pea plants and hypothesized a factor that conveys traits from parent to offspring. Although he did not use the term *gene*, he explained his results in terms of inherited characteristics. Mendel was also the first to hypothesize independent assortment, the distinction between dominant and recessive traits, the distinction between a heterozygote and homozygote, and the difference between what would later be described as genotype and phenotype. Mendel's concept was finally named when Wilhelm Johannsen coined the word *gene* in 1909.

In the early 1900s, Mendel's work received renewed attention from scientists. In 1910, Thomas Hunt Morgan showed that genes reside on specific chromosomes. He later showed that genes occupy specific locations on the chromosome. With this knowledge, Morgan and his students began the first chromosomal map of the fruit fly *Drosophila*. In 1928, Frederick Griffith showed that genes could be transferred. In what is now known as Griffith's experiment, injections into a mouse of a deadly strain of bacteria that had been heat-killed transferred genetic information to a safe strain of the same bacteria, killing the mouse.

In 1941, George Wells Beadle and Edward Lawrie Tatum showed that mutations in genes caused errors in certain steps in metabolic pathways. This showed that specific genes code for specific proteins, leading to the "one gene, one enzyme" hypothesis. Oswald Avery, Collin Macleod, and Maclyn McCarty showed in 1944 that DNA holds the gene's information. In 1953, James D. Watson and Francis Crick demonstrated the molecular structure of DNA. Together, these discoveries established the central dogma of molecular biology, which states that proteins are translated from RNA which is transcribed from DNA. This dogma has since been shown to have exceptions, such as reverse transcription in retroviruses.

Evolutionary concept of gene

George C. Williams first explicitly advocated the gene-centric view of evolution in his book *Adaptation and Natural Selection*. Also, he proposed an evolutionary concept of gene to be used when we are talking about natural selection favoring some gene. The definition is: "that which segregates and recombines with appreciable frequency." According to this definition, even an asexual genome could be considered a gene, insofar it have an appreciable permanency through many generations.

The difference is: the molecular gene *transcribes* as a unit, and the evolutionary gene *inherits* as a unit.

Richard Dawkins' *The Selfish Gene* and *The Extended Phenotype* defended the idea that the gene is the only replicator in living systems. This means that only genes transmit their structure largely intact and are potentially immortal in the form of copies. So, genes should be the unit of selection.

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.

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) is a statistical test often used for linkage analysis in human populations, and also in 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 score favor the presence of linkage, whereas negative LOD scores indicate that linkage is less likely. The test was developed by Newton E. Morton. 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 [1]. 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:

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

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 chromosomal crossover will take place between two loci (or genes) during meiosis. 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%.

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 color (*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**.

Genetic association

From Wikipedia, the free encyclopedia

Jump to: navigation, search

Studies concerning **genetic association** aim to test whether single-locus alleles or genotype frequencies (or more generally, multilocus haplotype frequencies) are different between 2 groups (usually diseased subjects and healthy controls). Genetic association studies are based on the principle that genotypes can be compared "directly", i.e. with the sequences of the actual genomes.

Contents

[hide]

[edit] What is genetic association?

The occurrence together in a population, more often than can be readily explained by chance, of two or more traits of which at least one is known to be genetic. This can be between phenotypes, e.g. visible characteristics such as flower colour or height, between a phenotype and a genetic polymorphism, such as a single nucleotide polymorphism, or between two genetic polymorphisms. Association between genetic polymorphisms occurs when there is non-random association of their alleles as a result of their proximity on the same chromosome; this is known as genetic linkage.

Linkage disequilibrium (LD) is a term used in the study of population genetics for the non-random association of alleles at two or more loci, not necessarily on the same chromosome. It is not the same as linkage, which describes the phenomenon whereby two or more loci on a chromosome have reduced recombination between them because of their physical proximity to each other. LD describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies.

Genetic association studies are performed to determine whether a genetic variant is associated with a disease or trait: if association is present, a particular allele, genotype or haplotype of a polymorphism or polymorphism(s) will be seen more often than expected by chance in an individual carrying the trait. Thus, a person carrying one or two copies of a high-risk variant is at increased risk of developing the associated disease or having the associated trait.

Genetic association studies

Case-control designs

- Case control studies are a classical epidemiological tool. Case-control studies use subjects who already have a disease, trait or other condition and determine if there are characteristics of these patients that differ from those who don’t have the disease or trait. In genetic case-control studies, the frequency of alleles or genotypes is compared between the cases and controls. The cases will have been diagnosed with the disease under study, or have the trait under test; the controls, who are either known to be unaffected, or who have been randomly selected from the population. A difference in the frequency of an allele or genotype of the polymorphism under test between the two groups indicates that the genetic marker may increase risk of the disease or likelihood of the trait, or be in linkage disequilibrium with a polymorphism which does. Haplotypes can also show association with a disease or trait.

One problem with the case-control design is that genotype and haplotype frequencies vary between ethnic or geographic populations. If the case and control populations are not well matched for ethnicity or geographic origin then false positive association can occur because of the confounding effects of population stratification.

Family based designs

Family based association designs aim to avoid the potential confounding effects of population stratification by using the parents as controls for the case, which is their affected offspring. The most commonly used test is the transmission disequilibrium test, or TDT. Two similar tests are used, the transmission disequilibrium test (TDT) and haploid-relative-risk (HRR). Both measure association of genetic markers in nuclear families by transmission from parent to offspring. If an allele increases the risk of having a disease then that allele is expected to be transmitted from parent to offspring more often in populations with the disease.

Quantitative trait association

A quantitative trait (see quantitative trait locus) is a measurable trait that shows continuous variation, such as height or weight. Quantitative traits often have a 'normal' distribution in the population. In addition to the case control design, quantitative trait association can also be performed using an unrelated population sample or family trios in which the quantitative trait is measured in the offspring.

Statistical programs of association analysis

There are many computer packages for analyzing genetic association, such as UNPHASED, WHAP, FBAT, Merlin, PLINK, and Golden Helix's *HelixTree Software*. However simple genotypic or alleleic association with a dichotomous trait can be measured using the chi-squared test for significance.

Linkage disequilibrium

In population genetics, **linkage disequilibrium** is the non-random association of alleles at two or more loci, not necessarily on the same chromosome. It is not the same as linkage, which describes the association of two or more loci on a chromosome with limited recombination between them. Linkage disequilibrium describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies. Non-random associations between polymorphisms at different loci are measured by the degree of linkage disequilibrium (LD).

An example is the prevalence of two rare diseases in Finland: there, compared to elsewhere in Europe, cystic fibrosis is less prevalent but congenital chloride diarrhea is more prevalent (see Finnish disease heritage). Both diseases are due to mutations on chromosome 7, in adjacent genes.[1]

The level of linkage disequilibrium is influenced by a number of factors including genetic linkage, the rate of recombination, the rate of mutation, genetic drift, non-random mating, and population structure. For example, some organisms (such as bacteria) may show linkage disequilibrium because they reproduce asexually and there is no recombination to break down the linkage disequilibrium.

Linkage disequilibrium measure, *D*

If we look at haplotypes for two loci A and B with two alleles each—a two-locus, two-allele model—the following table denotes the frequencies of each combination:

Haplotype	Frequency
A_1B_1	x_{11}
A_1B_2	x_{12}
A_2B_1	x_{21}
A_2B_2	x_{22}

Note that these are relative frequencies. One can use the above frequencies to determine the frequency of each of the alleles:

Allele	Frequency
A_1	$p_1 = x_{11} + x_{12}$
A_2	$p_2 = x_{21} + x_{22}$
B_1	$q_1 = x_{11} + x_{21}$
B_2	$q_2 = x_{12} + x_{22}$

If the two loci and the alleles are independent from each other, then one can express the observation A_1B_1 as "A_1 is found and B_1 is found". The table above lists the frequencies for A_1, p_1, and for B_1, q_1, hence the frequency of A_1B_1 is x_{11}, and according to the rules of elementary statistics $x_{11} = p_1q_1$.

The deviation of the observed frequency of a haplotype from the expected is a quantity[2] called the linkage disequilibrium[3] and is commonly denoted by a capital D:

$D = x_{11} - p_1 q_1$

In the genetic literature the phrase "two alleles are in LD" usually means that $D \neq 0$. Contrariwise, "linkage equilibrium" denotes the case $D = 0$.

The following table illustrates the relationship between the haplotype frequencies and allele frequencies and D.

	A_1	A_2	Total
B_1	$x_{11} = p_1 q_1 + D$	$x_{21} = p_2 q_1 - D$	q_1
B_2	$x_{12} = p_1 q_2 - D$	$x_{22} = p_2 q_2 + D$	q_2
Total	p_1	p_2	1

D is easy to calculate with, but has the disadvantage of depending on the frequency of the alleles. This is evident since frequencies are between 0 and 1. There can be no D observed if any locus has an allele frequency 0 or 1 and is maximal when frequencies are at 0.5. Lewontin (1964) suggested normalising D by dividing it with the theoretical maximum for the observed allele frequencies. Thus $D' = \frac{D}{D_{\mathbf{max}}}$ when $D \geq 0$When $D < 0$, $D' = \frac{D}{D_{\mathbf{min}}}$.

D_{max} is given by the smaller of $p_1 q_2$ and $p_2 q_1$. D_{min} is given by the larger of $-p_1 q_1$ and $-p_2 q_2$

Another measure of LD which is an alternative to D' is the correlation coefficient between pairs of loci, denoted as $r = \frac{D}{\sqrt{p_1 p_2 q_1 q_2}}$. This is also adjusted to the loci having different allele frequencies. There is some relationship between r and D'. [4]

In summary, linkage disequilibrium reflects the difference between the expected haplotype frequencies under the assumption of independence, and observed haplotype frequencies. A value of 0 for D' indicates that the examined loci are in fact independent of one another, while a value of 1 demonstrates complete dependency.

Role of recombination

In the absence of evolutionary forces other than random mating and Mendelian segregation, the linkage disequilibrium measure D converges to zero along the time axis at a rate depending on the magnitude of the recombination rate c between the two loci.

Using the notation above, $D = x_{11} - p_1q_1$, we can demonstrate this convergence to zero as follows. In the next generation, x_{11}', the frequency of the haplotype A_1B_1, becomes

$$x'_{11} = (1-c)\,x_{11} + c\,p_1q_1$$

This follows because a fraction $(1 - c)$ of the haplotypes in the offspring have not recombined, and are thus copies of a random haplotype in their parents. A fraction x_{11} of those are A_1B_1. A fraction c have recombined these two loci. If the parents result from random mating, the probability of the copy at locus A having allele A_1 is p_1 and the probability of the copy at locus B having allele B_1 is q_1, and as these copies are initially on different haplotypes, these are independent events so that the probabilities can be multiplied.

This formula can be rewritten as

$$x'_{11} - p_1q_1 = (1-c)\,(x_{11} - p_1q_1)$$

so that

$$D_1 = (1-c)\;D_0$$

where D at the n-th generation is designated as D_n. Thus we have

$$D_n = (1-c)^n\;D_0.$$

If $n \rightarrow \infty$, then $(1-c)^n \rightarrow 0$ so that D_n converges to zero.

If at some time we observe linkage disequilibrium, it will disappear in future due to recombination. However, the smaller the distance between the two loci, the smaller will be the rate of convergence of D to zero.

Linkage disequilibrium appears frequently in genetic systems

Human leucocyte antigen (HLA)

HLA constitutes a group of cell surface antigens as MHC of humans. Because HLA genes are located at adjacent loci on the particular region of a chromosome and presumed to exhibit epistasis with each other or with other genes, a sizable fraction of alleles are in linkage disequilibrium.

An example of such linkage disequilibrium is between HLA-A1 and B8 alleles in unrelated Danes[5] referred to by Vogel and Motulsky (1997).[6]

Table 1. Association of HLA-A1 and B8 in unrelated Danes[5]

			No. of individuals		
			Antigen j		
			+	**−**	**Total**
			B8 $^+$	***B8*** $^-$	
Antigen i	**+**	***A1*** $^+$	$a = 376$	$b = 237$	C
	−	***A1*** $^-$	$c = 91$	$d = 1265$	D
Total			A	B	N

Because HLA is codominant and HLA expression is only tested locus by locus in surveys, LD measure is to be estimated from such a 2x2 table to the right.[6][7][8][9]

pf_i = frequency of antigen $i = C / N = 0.311$,
$pf_i = 0.237$,

$$gf_i = \text{frequency of gene } i = 1 - \sqrt{1 - pf_i} = 0.170,$$

and

$$hf_{ij} = \text{estimated frequency of haplotype } ij = gf_i \, gf_j = 0.0215.$$

Denoting the '—' alleles at antigen i to be 'x,' and at antigen j to be 'y,' the observed frequency of haplotype xy is

$$o[hf_{xy}] = \sqrt{d/N}$$

and the estimated frequency of haplotype xy is

$$e[hf_{xy}] = \sqrt{(D/N)(B/N)}.$$

Then LD measure Δ_{ij} is expressed as

$$\Delta_{ij} = o[hf_{xy}] - e[hf_{xy}] = \frac{\sqrt{Nd} - \sqrt{BD}}{N} = 0.0769.$$

Standard errors *SEs* are obtained as follows:

$$SE \text{ of } gf_i = \sqrt{C}/(2N) = 0.00628,$$

$$SE \text{ of } hf_{ij} = \sqrt{\frac{(1-\sqrt{d/B})(1-\sqrt{d/D}) - hf_{ij} - hf_{ij}^2/2}{2N}} = 0.00514$$

$$SE \text{ of } \Delta_{ij} = \frac{1}{2N}\sqrt{a - 4N\Delta_{ij}\left(\frac{B+D}{2\sqrt{BD}} - \frac{\sqrt{BD}}{N}\right)} = 0.00367 .$$

Then, if

$t = \Delta_{ij} / (SE \text{ of } \Delta_{ij})$

exceeds 2 in its absolute value, the magnitude of Δ_{ij} is large statistically significantly. For data in Table 1 it is 20.9, thus existence of statistically significant LD between A1 and B8 in the population is admitted.

Table 2. Linkage disequilibrium among HLA alleles in Caucasians[9]

HLA-A alleles i	**HLA-B alleles j**	Δ_{ij}	t
A1	B8	0.065	16.0
A3	B7	0.039	10.3
A2	Bw40	0.013	4.4
A2	Bw15	0.01	3.4
A1	Bw17	0.014	5.4
A2	B18	0.006	2.2
A2	Bw35	-0.009	-2.3
A29	B12	0.013	6.0
A10	Bw16	0.013	5.9

Table 2 shows some of the combinations of HLA-A and B alleles where significant LD was observed among Caucasians.[9]

Vogel and Motulsky (1997)[6] argued how long would it take that linkage disequilibrium between loci of HLA-A and B disappeared. Recombination between loci of HLA-A and B was considered to be of the order of magnitude 0.008. We will argue similarly to Vogel and Motulsky below. In case LD measure was observed to be 0.003 in Caucasians in the list of Mittal[9] it is mostly non-significant. If Δ_0 had reduced from 0.07 to 0.003 under recombination effect as shown by $\Delta_n = (1 - c)^n\Delta_0$, then $n \approx 400$. Suppose a generation took 25 years, this means 10,000 years. The time span seems rather short in the history of humans. Thus observed linkage disequilibrium between HLA-A and B loci might indicate some sort of interactive selection.[6]

Further information: HLA A1-B8 haplotype

Between an HLA locus and a presumed major gene locus having disease susceptibility

Presence of linkage disequilibrium between an HLA locus and a presumed major gene of disease susceptibility corresponds to any of the following phenomena:

- Relative risk for the person having a specific HLA allele to become suffered from a particular disease is larger than one.[10]
- The HLA antigen frequency among patients exceeds more than that among a healthy population. This is evaluated by δ value[11] to exceed 0.

Table 3. Association of ankylosing spondylitis with HLA-B27 allele[12]

		Ankylosing spondylitis		**Total**
		Patients	**Healthy controls**	
HLA alleles	***B27*** $^{+}$	*a* = 96	*b* = 77	*C*
	B27 $^{-}$	*c* = 22	*d* = 701	*D*
Total		*A*	*B*	*N*

- 2x2 association table of patients and healthy controls with HLA alleles shows a significant deviation from the equilibrium state deduced from the marginal frequencies.

(1) Relative risk

Relative risk of an HLA allele for a disease is approximated by the odds ratio in the 2x2 association table of the allele with the disease. Table 3 shows association of HLA-B27 with ankylosing spondylitis among a Dutch population.[12] Relative risk *x* of this allele is approximated by

$$x = \frac{a/b}{c/d} = \frac{ad}{bc} \ (= 39.7, \text{ in Table 3 }).$$

Woolf's method[13] is applied to see if there is statistical significance. Let

$$y = \ln(x) \ (= 3.68)$$

and

$$\frac{1}{w} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \ (= 0.0703).$$

Then

$$\chi^2 = wy^2 \ \left[= 193 > \chi^2(p = 0.001,\ df = 1) = 10.8\right]$$

follows the chi-square distribution with *df* = 1. In the data of Table 3, the significant association exists at the 0.1% level. Haldane's[14] modification applies to the case when either of $a,\ b,\ c,\ \text{and}\ d$is zero, where replace *x* and 1 / *w* with

$$x = \frac{(a + 1/2)(d + 1/2)}{(b + 1/2)(c + 1/2)}$$

and

$$\frac{1}{w} = \frac{1}{a+1} + \frac{1}{b+1} + \frac{1}{c+1} + \frac{1}{d+1},$$

respectively.

Table 4. Association of HLA alleles with rheumatic and autoimmune diseases among white populations[10]

Disease	HLA allele	Relative risk (%)	FAD (%)	FAP (%)	δ
Ankylosing spondylitis	B27	90	90	8	0.89
Reiter's syndrome	B27	40	70	8	0.67
Spondylitis **in inflammatory bowel disease**	B27	10	50	8	0.46
Rheumatoid arthritis	DR4	6	70	30	0.57
Systemic lupus erythematosus	DR3	3	45	20	0.31
Multiple sclerosis	DR2	4	60	20	0.5
Juvenile diabetes mellitus **(type 1)**	DR4	6	75	30	0.64

In Table 4, some examples of association between HLA alleles and diseases are presented.[10]

(1a) Allele frequency excess among patients over controls

Even high relative risks between HLA alleles and the diseases were observed, only the magnitude of relative risk would not be able to determine the strength of association.[11] δ value is expressed by

$$\delta = \frac{FAD - FAP}{1 - FAP}, \quad 0 \leq \delta \leq 1$$

,

where *FAD* and *FAP* are HLA allele frequencies among patients and healthy populations, respectively.[11] In Table 4, δ column was added in this quotation. Putting aside 2 diseases with high relative risks both of which are also with high δ values, among other diseases, juvenile diabetes mellitus (type 1) has a strong association with DR4 even with a low relative risk = 6.

(2) Discrepancies from expected values from marginal frequencies in 2x2 association table of HLA alleles and disease

This can be confirmed by χ^2 test calculating

$$\chi^2 = \frac{(ad-bc)^2 N}{ABCD} \; (= 336, \text{ for data in Table 3; } P < 0.001).$$

where *df* = 1. For data with small sample size, such as no marginal total is greater than 15 (and consequently $N \leq 30$), one should utilize Yates' correction for continuity or Fisher's exact test.[15]

Resources

A comparison of different measures of LD is provided by Devlin & Risch [16]

The International HapMap Project enables the study of LD in human populations online. The Ensembl project integrates HapMap data and such from dbSNP in general with other genetic information.

Analysis software

- LDHat
- Haploview
- LdCompare[17] — open-source software for calculating LD.
- PyPop
- SNP and Variation Suite - commercial software with interactive LD plot.
- GOLD - Graphical Overview of Linkage Disequilibrium
- TASSEL - software to evaluate linkage disequilibrium, traits associations, and evolutionary patterns

References

1. ^ Höglund P, Haila S, Socha J, Tomaszewski L, Saarialho-Kere U, Karjalainen-Lindsberg ML, Airola K, Holmberg C, de la Chapelle A, Kere J (November 1996). "Mutations of the Down-regulated in adenoma (DRA) gene cause congenital chloride diarrhoea". *Nature Genetics* **14** (3): 316–9. doi:10.1038/ng1196-316. PMID 8896562.

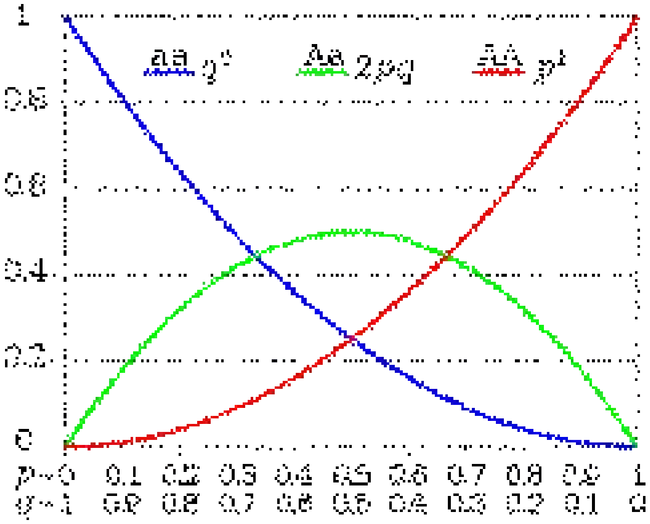
2. ^ Robbins, R.B. (07/01/1918). "Some applications of mathematics to breeding problems III". *Genetics* **3** (4): 375–389. http://www.genetics.org/cgi/reprint/3/4/375.
3. ^ R.C. Lewontin and K. Kojima (1960). "The evolutionary dynamics of complex polymorphisms". *Evolution* **14** (4): 458–472. doi:10.2307/2405995. http://links.jstor.org/sici?sici=0014-3820%28196012%2914%3A4%3C458%3ATEDOCP%3E2.0.CO%3B2-4.
4. ^ P.W. Hedrick and S. Kumar (2001). "Mutation and linkage disequilibrium in human mtDNA". *Eur. J. Hum. Genet.* **9**: 969–972. doi:10.1038/sj.ejhg.5200735.
5. ^ *a* *b* Svejgaard A, Hauge M, Jersild C, Plaz P, Ryder LP, Staub Nielsen L, Thomsen M (1979). *The HLA System: An Introductory Survey, 2nd ed.* Basel; London; Chichester: Karger; Distributed by Wiley, ISBN 3805530498(pbk).
6. ^ *a* *b* *c* *d* Vogel F, Motulsky AG (1997). *Human Genetics : Problems and Approaches, 3rd ed.* Berlin; London: Springer, ISBN 3540602909.
7. ^ Mittal KK, Hasegawa T, Ting A, Mickey MR, Terasaki PI (1973). "Genetic variation in the HL-A system between Ainus, Japanese, and Caucasians," *In* Dausset J, Colombani J, eds. *Histocompatibility Testing, 1972,* pp. 187-195, Copenhagen: Munksgaard, ISBN 8716011015.
8. ^ Yasuda N, Tsuji K (1975). "A counting method of maximum likelihood for estimating haplotype frequency in the HL-A system." *Jinrui Idengaku Zasshi* **20**(1): 1-15, PMID 1237691.
9. ^ *a* *b* *c* *d* Mittal KK (1976). "The HLA polymorphism and susceptibility to disease." *Vox Sang* **31**: 161-173, PMID 969389.
10. ^ *a* *b* *c* Gregersen PK (2009). "Genetics of rheumatic diseases," *In* Firestein GS, Budd RC, Harris ED Jr, McInnes IB, Ruddy S, Sergent JS, eds. (2009). *Kelley's Textbook of Rheumatology,* pp. 305-321, Philadelphia, PA: Saunders/Elsevier, ISBN 9781416032854.
11. ^ *a* *b* *c* Bengtsson BO, Thomson G (1981). "Measuring the strength of associations between HLA antigens and diseases." *Tissue Antigens* **18**(5): 356-363, PMID 7344182.
12. ^ *a* *b* Nijenhuis LE (1977). "Genetic considerations on association between HLA and disease." *Hum Genet* **38**(2): 175-182, PMID 908564.
13. ^ Woolf B (1955). "On estimating the relation between blood group and disease." *Ann Hum Genet* **19**(4): 251-253, PMID 14388528.
14. ^ Haldane JB (1956). "The estimation and significance of the logarithm of a ratio of frequencies." *Ann Hum Genet* **20**(4): 309-311, PMID 13314400.
15. ^ Sokal RR, Rohlf FJ (1981). *Biometry: The Principles and Practice of Statistics in Biological Research.* Oxford: W.H. Freeman, ISBN 0716712547.
16. ^ Devlin B., Risch N. (1995). "A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping". *Genomics* **29**: 311–322. doi:10.1006/geno.1995.9003. http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6WG1-45S9156-30-1&_cdi=6809&_user=128590&_orig=browse&_coverDate=09%2F30%2F1995&_sk=999709997&view=c&wchp=dGLbVtb-zSkzk&md5=71c2158ad4c51ae80b12a68c68814f78&ie=/sdarticle.pdf.

17. ^ Hao K., Di X., Cawley S. (2007). "LdCompare: rapid computation of single- and multiple-marker r2 and genetic coverage". *Bioinformatics* **23**: 252–254. doi:10.1093/bioinformatics/btl574. PMID 17148510. http://bioinformatics.oxfordjournals.org/cgi/reprint/23/2/252.

Further reading

- Hedrick, Philip W. (2005). *Genetics of Populations* (3rd ed.). Sudbury, Boston, Toronto, London, Singapore: Jones and Bartlett Publishers. ISBN 0763747726.
- Bibliography: Linkage Disequilibrium Analysis : a bibliography of more than one thousand articles on Linkage disequilibrium published since 1918.

Hardy–Weinberg principle



Hardy–Weinberg principle for two alleles: the horizontal axis shows the two allele frequencies *p* and *q* and the vertical axis shows the genotype frequencies. Each graph shows one of the three possible genotypes.

The **Hardy–Weinberg principle** states that both allele and genotype frequencies in a population remain constant—that is, they are in equilibrium—from generation to generation unless specific disturbing influences are introduced. Those disturbing influences include *non-random mating*, *mutations*, *selection*, *limited population size*, "overlapping generations", *random genetic drift* and *gene flow*. It is important to understand that outside the lab, one or more of these "disturbing influences" are always in effect. That is, Hardy–Weinberg equilibrium is impossible in nature. Genetic equilibrium is an ideal state that provides a baseline to measure genetic change against.

Static allele frequencies in a population across generations assume: random mating, no mutation (the alleles don't change), no migration or emigration (no exchange of alleles between populations), infinitely large population size, and no selective pressure for or against any traits.

In the simplest case of a single locus with two alleles: the dominant allele is denoted **A** and the recessive **a** and their frequencies are denoted by *p* and *q*; freq(**A**) = *p*; freq(**a**) = *q*; $p + q = 1$. If the population is in equilibrium, then we will have freq(**AA**) = p^2 for the **AA** homozygotes in the population, freq(**aa**) = q^2 for the **aa** homozygotes, and freq(**Aa**) = $2pq$ for the heterozygotes.

Based on these equations, we can determine useful but difficult-to-measure facts about a population. For example, a patient's child is a carrier of a recessive mutation that causes cystic fibrosis in homozygous recessive children. The parent wants to know the probability of her grandchildren inheriting the disease. In order to answer this question, the genetic counselor must know the chance that the child will reproduce with a carrier of the recessive mutation. This fact may not be known, but disease frequency is known. We know that the disease is caused by the homozygous recessive genotype; we can use the Hardy–Weinberg principle to work backward from disease occurrence to the frequency of heterozygous recessive individuals.

This concept is also known by a variety of names: HWP**,** Hardy–Weinberg equilibrium**,** Hardy–Weinberg Theorem**,** HWE**, or** Hardy–Weinberg law**. It was named after G. H. Hardy and Wilhelm Weinberg.**

Derivation

A better, but equivalent, probabilistic description for the HWP is that the alleles for the next generation for any given individual are chosen randomly and independent of each other. Consider two alleles, A and a, with frequencies *p* and *q*, respectively, in the population. The different ways to form new genotypes can be derived using a Punnett square, where the fraction in each is equal to the product of the row and column probabilities.

Table 1: Punnett square for Hardy–Weinberg equilibrium

		Females	
		A (*p*)	**a (*q*)**
Males	**A (*p*)**	AA (p^2)	Aa (*pq*)
	a (*q*)	Aa (*pq*)	aa (q^2)

The final three possible genotypic frequencies in the offspring become:

- $f(\mathbf{AA}) = p^2$
- $f(\mathbf{Aa}) = 2pq$
- $f(\mathbf{aa}) = q^2$

These frequencies are called Hardy–Weinberg frequencies (or Hardy–Weinberg proportions). This is achieved in one generation, and only requires the assumption of random mating with an infinite population size.

Sometimes, a population is created by bringing together males and females with different allele frequencies. In this case, the assumption of a single population is violated until after the first generation, so the first generation will not have Hardy–Weinberg equilibrium. Successive generations will have Hardy–Weinberg equilibrium.

Deviations from Hardy–Weinberg equilibrium

Violations of the Hardy–Weinberg assumptions can cause deviations from expectation. How this affects the population depends on the assumptions that are violated. Generally, deviation from the Hardy–Weinberg equilibrium denotes the evolution of a species.

- Random mating. The HWP states the population will have the given genotypic frequencies (called Hardy–Weinberg proportions) after a single generation of random mating within the population. When violations of this provision occur, the population will not have Hardy–Weinberg proportions. Three such violations are:
    - Inbreeding, which causes an increase in homozygosity for all genes.
    - Assortative mating, which causes an increase in homozygosity only for those genes involved in the trait that is assortatively mated (and genes in linkage disequilibrium with them).
    - Small population size, which causes a random change in genotypic frequencies, particularly if the population is very small. This is due to a sampling effect, and is called genetic drift.

The remaining assumptions affect the allele frequencies, but do not, in themselves, affect random mating. If a population violates one of these, the population will continue to have Hardy–Weinberg proportions each generation, but the allele frequencies will change with that force.

- Selection, in general, causes allele frequencies to change, often quite rapidly. While directional selection eventually leads to the loss of all alleles except the favored one, some forms of selection, such as balancing selection, lead to equilibrium without loss of alleles.
- Mutation will have a very subtle effect on allele frequencies. Mutation rates are of the order 10^{-4} to 10^{-8}, and the change in allele frequency will be, at most, the same order. Recurrent mutation will maintain alleles in the population, even if there is strong selection against them.
- Migration genetically links two or more populations together. In general, allele frequencies will become more homogeneous among the populations. Some models for migration inherently include nonrandom mating (Wahlund effect, for example). For those models, the Hardy–Weinberg proportions will normally not be valid.

How these violations affect formal statistical tests for HWE is discussed later.

Unfortunately, violations of assumptions in the Hardy–Weinberg principle does not mean the population will violate HWE. For example, balancing selection leads to an equilibrium population with Hardy–Weinberg proportions. This property with selection *vs*. mutation is the basis for many estimates of mutation rate (call mutation-selection balance).

Sex linkage

Where the **A** gene is sex linked, the heterogametic sex (*e.g.*, mammalian males; avian females) have only one copy of the gene (and are termed hemizygous), while the homogametic sex (*e.g.*, human females) have two copies. The genotype frequencies at equilibrium are *p* and *q* for the heterogametic sex but p^2*, 2*pq *and* q^2 *for the homogametic sex.*

For example, in humans red–green colorblindness is an X-linked recessive trait. In western European males, the trait affects about 1 in 12, (q = 0.083) whereas it affects about 1 in 200 females (0.005, compared to q^2 = 0.007), very close to Hardy–Weinberg proportions.

If a population is brought together with males and females with different allele frequencies, the allele frequency of the male population follows that of the female population because each receives its X chromosome from its mother. The population converges on equilibrium very quickly.

Generalizations

The simple derivation above can be generalized for more than two alleles and polyploidy.

Generalization for more than two alleles

Consider an extra allele frequency, *r*. The two-allele case is the binomial expansion of $(p + q)^2$, and thus the three-allele case is the trinomial expansion of $(p + q+ r)^2$.

$$(p+q+r)^2 = p^2 + r^2 + q^2 + 2pq + 2pr + 2qr$$

More generally, consider the alleles **A**$_1$, ... **A**$_i$ given by the allele frequencies p_1 to p_i;

$$(p_1 + \cdots + p_i)^2$$

giving for all homozygotes:

$$f(A_iA_i) = p_i^2$$

and for all heterozygotes:

$$f(A_iA_j) = 2p_ip_j$$

Generalization for polyploidy

The Hardy–Weinberg principle may also be generalized to polyploid systems, that is, for organisms that have more than two copies of each chromosome. Consider again only two alleles. The diploid case is the binomial expansion of:

$$(p+q)^2$$

and therefore the polyploid case is the polynomial expansion of:

$$(p+q)^c$$

where *c* is the ploidy, for example with tetraploid (*c* = 4):

Table 2: Expected genotype frequencies for tetraploidy

Genotype	Frequency
AAAA	p^4
AAAa	$4p^3q$
AAaa	$6p^2q^2$
Aaaa	$4pq^3$
aaaa	q^4

Depending on whether the organism is a 'true' tetraploid or an amphidiploid will determine how long it will take for the population to reach Hardy–Weinberg equilibrium.

Complete generalization

For *n* distinct alleles in *c*-ploids, the genotype frequencies in the Hardy–Weinberg equilibrium are given by individual terms in the multinomial expansion of $(p_1+\cdots+p_n)^c$:

$$(p_1+\cdots+p_n)^c = \sum_{k_1,\ldots,k_n \ \in\mathbb{N}:k_1+\cdots+k_n=c} \binom{c}{k_1,\ldots,k_n} p_1^{k_1}\cdots p_n^{k_n}$$

Applications

The Hardy–Weinberg principle may be applied in two ways, either a population is assumed to be in Hardy–Weinberg proportions, in which the genotype frequencies can be calculated, or if the genotype frequencies of all three genotypes are known, they can be tested for deviations that are statistically significant.

Application to cases of complete dominance

Suppose that the phenotypes of **AA** and **Aa** are indistinguishable, *i.e.*, there is complete dominance. Assuming that the Hardy–Weinberg principle applies to the population, then q can still be calculated from f(**aa**):

$$q = \sqrt{f(aa)}$$

and p can be calculated from q. And thus an estimate of f(**AA**) and f(**Aa**) derived from p^2 and $2pq$ respectively. Note however, such a population cannot be tested for equilibrium using the significance tests below because it is assumed *a priori*.

Significance tests for deviation

Testing deviation from the HWP is generally performed using Pearson's chi-squared test, using the observed genotype frequencies obtained from the data and the expected genotype frequencies obtained using the HWP. For systems where there are large numbers of alleles, this may result in data with many empty possible genotypes and low genotype counts, because there are often not enough individuals present in the sample to adequately represent all genotype classes. If this is the case, then the asymptotic assumption of the chi-square distribution, will no longer hold, and it may be necessary to use a form of Fisher's exact test, which requires a computer to solve. More recently a number of MCMC methods of testing for deviations from HWP have been proposed (Guo & Thompson, 1992; Wigginton *et al.* 2005)

Example χ^2 test for deviation

These data are from E.B. Ford (1971) on the Scarlet tiger moth, for which the phenotypes of a sample of the population were recorded. Genotype-phenotype distinction is assumed to be negligibly small. The null hypothesis is that the population is in Hardy–Weinberg proportions, and the alternative hypothesis is that the population is not in Hardy–Weinberg proportions.

Table 3: Example Hardy–Weinberg principle calculation

Genotype	White-spotted (**AA**)	Intermediate (**Aa**)	Little spotting (**aa**)	**Total**
Number	1469	138	5	1612

From which allele frequencies can be calculated:

$$p = \frac{2 \times \mathrm{obs}(AA) + \mathrm{obs}(Aa)}{2 \times (\mathrm{obs}(AA) + \mathrm{obs}(Aa) + \mathrm{obs}(aa))}$$

$$= \frac{1469 \times 2 + 138}{2 \times (1469 + 138 + 5)}$$

$$= \frac{3076}{3224}$$

$$= 0.954$$

and

$$\begin{aligned} q &= 1 - p \\ &= 1 - 0.954 \\ &= 0.046 \end{aligned}$$

So the Hardy–Weinberg expectation is:

$$\begin{aligned} \mathrm{Exp}(AA) &= p^2 n = 0.954^2 \times 1612 = 1467.4 \\ \mathrm{Exp}(Aa) &= 2pqn = 2 \times 0.954 \times 0.046 \times 1612 = 141.2 \\ \mathrm{Exp}(aa) &= q^2 n = 0.046^2 \times 1612 = 3.4 \end{aligned}$$

Pearson's chi-square test states:

$$\begin{aligned} \chi^2 &= \sum \frac{(O - E)^2}{E} \\ &= \frac{(1469 - 1467.4)^2}{1467.4} + \frac{(138 - 141.2)^2}{141.2} + \frac{(5 - 3.4)^2}{3.4} \\ &= 0.001 + 0.073 + 0.756 \\ &= 0.83 \end{aligned}$$

There is 1 degree of freedom (degrees of freedom for test for Hardy–Weinberg proportions are # genotypes − # alleles). The 5% significance level for 1 degree of freedom is 3.84, and since the χ^2 value is less than this, the null hypothesis that the population is in Hardy–Weinberg frequencies is not rejected.

Fisher's exact test (probability test)

Fisher's exact test can be applied to testing for Hardy–Weinberg proportions. Because the test is conditional on the allele frequencies, *p* and *q*, the problem can be viewed as testing for the proper number of heterozygotes. In this way, the hypothesis of Hardy–Weinberg proportions is rejected if the number of heterozygotes are too large or too small. The conditional probabilities for the heterozygote, given the allele frequencies are given in Emigh (1980) as

$$prob[n_{12}|n_1] = \frac{\binom{n}{n_{11}, n_{12}, n_{22}}}{\binom{2n}{n_1}} 2^{n_{12}},$$

where n_{11}, n_{12}, n_{22} are the observed numbers of the three genotypes, **AA**, **Aa**, and **aa**, respectively, and n_1 is the number of **A** alleles, where $n_1 = 2n_{11} + n_{12}$.

An example Using one of the examples from Emigh (1980)[1], we can consider the case where $n = 100$, and $p = 0.34$. The possible observed heterozygotes and their exact significance level is given in Table 4.

Table 4: Example of Fisher's Exact Test for n=100, p=0.34.[1]

Number of heterozygotes	Significance level
0	0.000
2	0.000
4	0.000
6	0.000
8	0.000
10	0.000
12	0.000
14	0.000
16	0.000
18	0.001
20	0.007
22	0.034
34	0.067
24	0.151
32	0.291
26	0.474
30	0.730
28	1.000

Using this table, you look up the significance level of the test based on the observed number of heterozygotes. For example, if you observed 20 heterozygotes, the significance level for the test is 0.007. As is typical for Fisher's exact test for small samples, the gradation of significance levels is quite coarse.

Unfortunately, you have to create a table like this for every experiment, since the tables are dependent on both *n* and *p*.

Inbreeding coefficient

The inbreeding coefficient, *F* (see also *F*-statistics), is one minus the observed frequency of heterozygotes over that expected from Hardy–Weinberg equilibrium.

$$F = \frac{\operatorname{E}(f(\mathbf{Aa})) - \operatorname{O}(f(\mathbf{Aa}))}{\operatorname{E}(f(\mathbf{Aa}))} = 1 - \frac{\operatorname{O}(f(\mathbf{Aa}))}{\operatorname{E}(f(\mathbf{Aa}))},$$

where the expected value from Hardy–Weinberg equilibrium is given by

$$\operatorname{E}(f(\mathbf{Aa})) = 2pq$$

For example, for Ford's data above;

$$F = 1 - \frac{138}{141.2}$$
$$= 0.023.$$

For two alleles, the chi square goodness of fit test for Hardy–Weinberg proportions is equivalent to the test for inbreeding, $F = 0$.

The inbreeding coefficient is unstable as the expected value approaches zero, and thus not useful for rare and very common alleles. For: *E=0, O>0, F=-infinity and E=0, O=0, F=undefined*.

History

Mendelian genetics were rediscovered in 1900. However, it remained somewhat controversial for several years as it was not then known how it could cause continuous characteristics. Udny Yule (1902) argued against Mendelism because he thought that dominant alleles would increase in the population. The American William E. Castle (1903) showed that without selection, the genotype frequencies would remain stable. Karl Pearson (1903) found one equilibrium position with values of $p = q = 0.5$. Reginald Punnett, unable to counter Yule's point, introduced the problem to G. H. Hardy, a British mathematician, with whom he played cricket. Hardy was a pure mathematician and held applied mathematics in some contempt; his view of biologists' use of mathematics comes across in his 1908 paper where he describes this as "very simple".

> *To the Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making...*
> *Suppose that Aa is a pair of Mendelian characters, A being dominant, and that in any given generation the number of pure dominants (AA), heterozygotes (Aa), and pure recessives (aa) are as p:2q:r. Finally, suppose that the numbers are fairly large, so that mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show that in the next generation the numbers will be as* $(p+q)^2:2(p+q)(q+r):(q+r)^2$*, or as* $p_1:2q_1:r_1$*, say.*
> *The interesting question is — in what circumstances will this distribution be the same as that in the generation before? It is easy to see that the condition for this is* $q^2 = pr$*. And since* $q_1^2 = p_1r_1$*, whatever the values of p, q, and r may be, the distribution will in any case continue unchanged after the second generation*

The principle was thus known as *Hardy's law* in the English-speaking world until 1943, when Curt Stern pointed out that it had first been formulated independently in 1908 by the German physician Wilhelm Weinberg.[2][3] Others have attempted to associate Castle's name with the Law because of his work in 1903, but it is only rarely seen as the Hardy–Weinberg–Castle Law.

Derivation of Hardy's Equations

The derivation of Hardy's equations is illustrative. He begins with a population of genotypes consisting of pure dominants (AA), heterozygotes (Aa), and pure recessives (aa) in the relative proportions p:2q:r with the conditions noted above, that is,

$$p + 2q + r = 1.$$

Rewriting this as (p + q) + (q + r) = 1 and squaring both sides yields Hardy's result:

$$p_1 + 2q_1 + r_1 = (p + q)^2 + 2(p + q)(q + r) + (q + r)^2 = 1$$

Hardy's equivalence condition is

$$\begin{aligned} E_1 &= q_1^2 - p_1 r_1 \\ &= [(p+q)(q+r)]^2 - (p+q)^2(q+r)^2 = 0 \end{aligned}$$

for generations after the first. For a putative third generation,

$$
\begin{aligned}
p_2 &= (p_1 + q_1)^2 \\
&= p_1^2 + 2p_1q_1 + q_1^2
\end{aligned}
$$

Substituting for p_1 and q_1 and factoring out $(p+q)^2$ yields,

$$p_2 = (p + q)^2[(p + q)^2 + 2(p + q)(q + r) + (q + r)^2]$$

The quantity in brackets is equal to 1, therefore, $p_2 = p_1$ and will remain so for succeeding generations. The result will be the same for the other two genotypes.

Numerical Example

An example computation of the genotype distribution given by Hardy's original equations is instructive. The phenotype distribution from Table 3 above will be used to compute Hardy's initial genotype distribution. Note that the p and q values used by Hardy are not the same as those used above.

$$
\begin{aligned}
sum &= \mathrm{obs}(AA) + 2 \times \mathrm{obs}(Aa) + \mathrm{obs}(aa) = 1469 + 2 \times 138 + 5 \\
&= 1750 \\
p &= \frac{1469}{1750} = 0.83943 \\
2q &= \frac{2 \times 138}{1750} = 0.15771 \\
r &= \frac{5}{1750} = 0.00286
\end{aligned}
$$

As checks on the distribution, compute

$$p + 2q + r = 0.83943 + 0.15771 + 0.00286 = 1.00000$$

and

$$E_0 = q^2 - pr = 0.00382$$

For the next generation, Hardy's equations give,

$$q = \frac{0.15771}{2} = 0.07886$$

$$\begin{aligned} p_1 &= (p+q)^2 = 0.84325 \\ 2q_1 &= 2(p+q)(q+r) = 0.15007 \\ r_1 &= (q+r)^2 = 0.00668 \end{aligned}$$

Again as checks on the distribution, compute

$$p_1 + 2q_1 + r_1 = 0.84325 + 0.15007 + 0.00668 = 1.00000$$

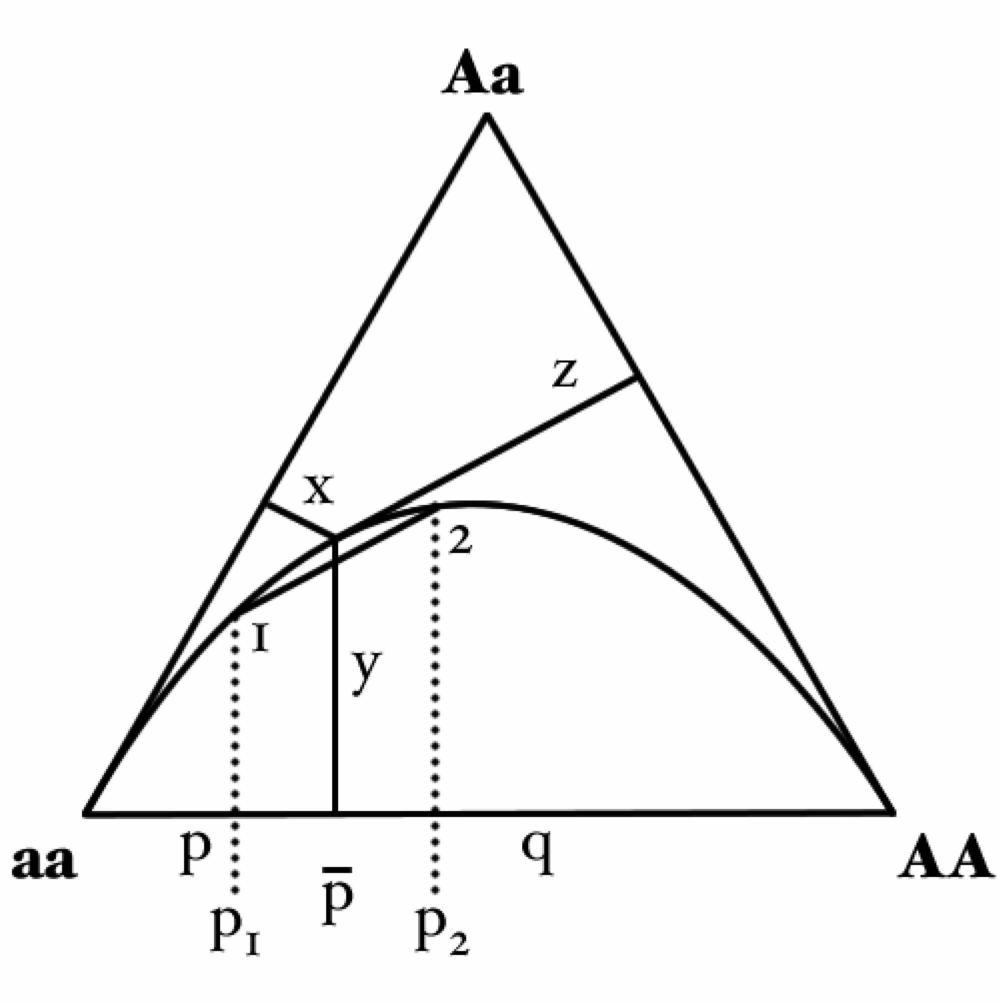
and

$$E_1 = q_1^2 - p_1 r_1 = 0.00000$$

which are the expected values. The reader may demonstrate that subsequent use of the second-generation values for a third generation will yield identical results.

Graphical representation

It is possible to represent the distribution of genotype frequencies for a bi-allelic locus within a population graphically using a de Finetti diagram. This uses a triangular plot (also known as trilinear, triaxial or ternary plot) to represent the distribution of the three genotype frequencies in relation to each other. Although it differs from many other such plots in that the direction of one of the axes has been reversed.



The curved line in the above diagram is the Hardy–Weinberg parabola and represents the state where alleles are in Hardy–Weinberg equilibrium.

It is possible to represent the effects of Natural Selection and its effect on allele frequency on such graphs (e.g. Ineichen & Batschelet 1975)

The De Finetti diagram has been developed and used extensively by A.W.F. Edwards in his book *Foundations of Mathematical Genetics*.

References

- Castle, W. E. (1903). The laws of Galton and Mendel and some laws governing race improvement by selection. *Proc. Amer. Acad. Arts Sci.*. **35**: 233–242.
- Crow, Jf (Jul 1999). "Hardy, Weinberg and language impediments." (Free full text). *Genetics* **152** (3): 821–5. ISSN 0016-6731. PMID 10388804. PMC 1460671. http://www.genetics.org/cgi/pmidlookup?view=long&pmid=10388804.

- Edwards, A.W.F. 1977. *Foundations of Mathematical Genetics*. Cambridge University Press, Cambridge (2nd ed., 2000). ISBN 0-521-77544-2
- Emigh, T.H. (1980). A comparison of tests for Hardy–Weinberg equilibrium. *Biometrics* **36**: 627–642.
- Ford, E.B. (1971). *Ecological Genetics*, London.
- Guo, Sw; Thompson, Ea (Jun 1992). "Performing the exact test of Hardy-Weinberg proportion for multiple alleles.". *Biometrics* **48** (2): 361–72. doi:10.2307/2532296. ISSN 0006-341X. PMID 1637966.
- Hardy, Gh (Jul 1908). "MENDELIAN PROPORTIONS IN A MIXED POPULATION.". *Science (New York, N.Y.)* **28** (706): 49–50. doi:10.1126/science.28.706.49. ISSN 0036-8075. PMID 17779291.
- Ineichen, Robert; Batschelet, Eduard (1975). "Genetic selection and de Finetti diagrams". *Journal of Mathematical Biology* **2**: 33. doi:10.1007/BF00276014.
- Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London, Ser. A* **200**: 1–66.
- Stern, C. (1943). "The Hardy–Weinberg law". *Science* **97**: 137–138. JSTOR stable url
- Weinberg, W. (1908). "Über den Nachweis der Vererbung beim Menschen". *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg* **64**: 368–382.
- Wigginton, Je; Cutler, Dj; Abecasis, Gr (May 2005). "A note on exact tests of Hardy-Weinberg equilibrium.". *American journal of human genetics* **76** (5): 887–93. doi:10.1086/429864. ISSN 0002-9297. PMID 15789306.
- Yule, G. U. (1902). Mendel's laws and their probable relation to intra-racial heredity. *New Phytol.* **1**: 193–207, 222–238.

Notes

1. ^ *a* *b* Emigh, Ted H. (1980). "A Comparison of Tests for Hardy–Weinberg Equilibrium". *Biometrics* **4** (4): 627–642. doi:10.2307/2556115. http://www.jstor.org/stable/2556115.
2. ^ Crow, James F. (1999). "Hardy, Weinberg and language impediments". *Genetics* **152** (3): 821–825. PMID 10388804. PMC 1460671. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1460671.
3. ^ Stern, Curt (1962). "Wilhelm Weinberg". *Genetics* **47**: 1–5.
4. ^ "evolution :: Genetic equilibrium: the Hardy–Weinberg law — Britannica Online Encyclopedia". http://www.search.eb.com/eb/article-49862.

Quantitative trait locus

Inheritance of quantitative traits **or** polygenic inheritance **refers to the inheritance of a phenotypic characteristic that varies in degree and can be attributed to the interactions between two or more genes and their environment. Though not necessarily genes themselves,** quantitative trait loci **(QTLs) are stretches of DNA that are closely linked to the genes that underlie the trait in question. QTLs can be molecularly identified (for example, with AFLP) to help map regions of the genome that contain genes involved in specifying a quantitative trait. This can be an early step in identifying and sequencing these genes.**

Quantitative traits

Polygenic inheritance, also known as **quantitative** or **multifactorial inheritance** refers to inheritance of a phenotypic characteristic (trait) that is attributable to two or more genes and their interaction with the environment. Unlike monogenic traits, polygenic traits do not follow patterns of Mendelian inheritance (qualitative traits). Instead, their phenotypes typically vary along a continuous gradient depicted by a bell curve.[1]

An example of a polygenic trait is human skin color. Many genes factor into determining a person's natural skin color, so modifying only one of those genes changes the color only slightly. Many disorders with genetic components are polygenic, including autism, cancer, diabetes and numerous others. Most phenotypic characteristics are the result of the interaction of multiple genes.

Examples of disease processes generally considered to be results of *multifactorial etiology*:

Congenital malformation

- Cleft palate[2] [3]
- Congenital dislocation of the hip[4]
- Congenital heart defects
- Neural tube defects
- Pyloric stenosis
- Talipes

Adult onset diseases

- Diabetes Mellitus[3]
- Cancer[3]
- Epilepsy
- Glaucoma
- hypertension
- Ischaemic heart disease
- Manic depression
- Schizophrenia

Multifactorially inherited diseases are said to constitute the majority of all genetic disorders affecting humans which will result in hospitalization or special care of some kind[5] [6].

Multifactorial traits in general

Generally, multifactorial traits outside of illness contribute to what we see as **continuous characteristics** in organisms, such as height[5], skin color, and body mass[7]. All of these phenotypes are complicated by a great deal of interplay between genes and environment[5]. While some authors[5] [7] include intelligence in the same vein, and it is tempting to do so, the problem with intelligence is that it is so ill-defined. Indeed, the entry on intelligence offers so many definitions, that the point is easily made that there is no single, agreed-upon entity that one could say amounts to a definable cluster of heritable traits.

The continuous distribution of traits such as height and skin colour described above reflects the action of genes that do not quite show typical patterns of dominance and recessiveness. Instead the contributions of each involved locus are thought to be additive. Writers have distinguished this kind of inheritance as *polygenic*, or *quantitative inheritance*[8].

Thus, due to the nature of polygenic traits, inheritance will not follow the same pattern as a simple monohybrid or dihybrid cross[6]. Polygenic inheritance can be explained as Mendelian inheritance at many loci[5], resulting in a trait which is normally-distributed. If n is the number of involved loci, then the coefficients of the binomial expansion of $(a + b)^{2n}$ will give the frequency of distribution of all n allele combinations. For a sufficiently high n, this binomial distribution will begin to resemble a normal distribution. From this viewpoint, a disease state will become apparent at one of the tails of the distribution, past some threshold value. Disease states of increasing severity will be expected the further one goes past the threshold and away from the mean[8].

Heritable disease and multifactorial inheritance

A mutation resulting in a disease state is often recessive, so both alleles must be mutant in order for the disease to be expressed phenotypically. A disease or syndrome may also be the result of the expression of mutant alleles at more than one locus. When more than one gene is involved with or without the presence of environmental triggers, we say that the disease is the result of multifactorial inheritance.

The more genes involved in the cross, the more the distribution of the genotypes will resemble a normal, or Gaussian distribution[5]. This shows that multifactorial inheritance is polygenic, and genetic frequencies can be predicted by way of a polyhybrid Mendelian cross. Phenotypic frequencies are a different matter, especially if they are complicated by environmental factors.

The paradigm of polygenic inheritance as being used to define multifactorial disease has encountered much disagreement. Turnpenny (2004) discusses how simple polygenic inheritance cannot explain some diseases such as the onset of Type I diabetes mellitus, and that in cases such as these, not all genes are thought to make an equal contribution[8].

The assumption of polygenic inheritance is that all involved loci make an equal contribution to the symptoms of the disease. This should result in a normal curve distribution of genotypes. When it does not, then idea of polygenetic inheritance cannot be supported for that illness.

A cursory look at some examples

Examples of such diseases are not new to medicine. The above examples are well-known examples of diseases having both genetic and environmental components. Other examples involve atopic diseases such as eczema or dermatitis[5]; also spina bifida (open spine) and anencephaly (open skull) are other examples[2]

While schizophrenia is widely believed to be multifactorially genetic by biopsychiatrists, no characteristic genetic markers have been determined with any certainty.

Is it multifactorially heritable?

It is difficult to ascertain if any particular disease is multifactorially genetic. If a pedigree chart is taken of the patient's family and relations, and it is shown that the brothers and sisters of the patient have the disease, then there is a strong chance that the disease is genetic and that the patient will also be a genetic carrier. But this is not quite enough. It also needs to be proven that the pattern of inheritance is non-Mendelian. This would require studying dozens, even hundreds of different family pedigrees before a conclusion of multifactorial inheritance is drawn. This often takes several years.

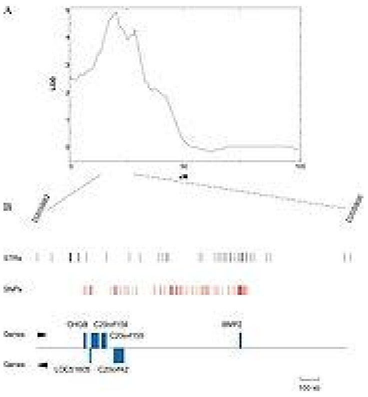
If multifactorial inheritance is indeed the case, then the chance of the patient contracting the disease is reduced if only cousins and more distant relatives have the disease[2]. It must be stated that while multifactorially-inherited disease tends to run in families, inheritance will not follow the same pattern as a simple monohybrid or dihybrid cross[6].

If a genetic cause is suspected and little else is known about the illness, then it remains to be seen exactly how many genes are involved in the phenotypic expression of the disease. Once that is determined, the question must be answered: if two people have the required genes, why some people still don't express the disease. Generally, what makes the two individuals different are likely to be environmental factors. Due to the involved nature of genetic investigations needed to determine such inheritance patterns, this is not usually the first avenue of investigation one would choose to determine etiology.

Psychiatry has determined, often without sufficient evidence, that mental illness follows this pattern. The problem with mental illness itself is that most diagnoses are largely subjective, and even the nosologies in DSM-IV are not widely agreed upon. The example most often cited as an example of a multifactorial mental illness is schizophrenia, however, no genes have been isolated to date. It has been said that genetic causes of mental illness are being emphasised at the expense of paying sufficient attention to environmental factors, especially in the field of biopsychiatry.[9]

More often than not, investigators will hypothesise that a disease is *multifactorially heritable,* along with a cluster of other hypotheses when it is not known what causes the disease.

Quantitative trait locus



A QTL for osteoporosis on the human chromosome 20

Typically, QTLs underlie continuous traits (those traits that vary continuously, e.g. height) as opposed to discrete traits (traits that have two or several character values, e.g. red hair in humans, a recessive trait, or smooth vs. wrinkled peas used by Mendel in his experiments).

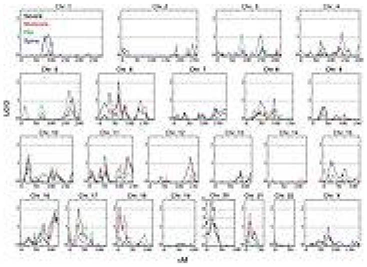
Moreover, a single phenotypic trait is usually determined by many genes. Consequently, many QTLs are associated with a single trait.

A **quantitative trait locus** (**QTL**) is a region of DNA that is associated with a particular phenotypic trait - these QTLs are often found on different chromosomes. Knowing the number of QTLs that explains variation in the phenotypic trait tells us about the genetic architecture of a trait. It may tell us that plant height is controlled by many genes of small effect, or by a few genes of large effect.

Another use of QTLs is to identify candidate genes underlying a trait. Once a region of DNA is identified as contributing to a phenotype, it can be sequenced. The DNA sequence of any genes in this region can then be compared to a database of DNA for genes whose function is already known.

In a recent development, classical QTL analyses are combined with gene expression profiling i.e. by DNA microarrays. Such expression QTLs (eQTLs) describe cis- and trans-controlling elements for the expression of often disease-associated genes. Observed epistatic effects have been found beneficial to identify the gene responsible by a cross-validation of genes within the interacting loci with metabolic pathway- and scientific literature databases.

QTL mapping



Example of a genome-wide scan for QTL of osteoporosis

QTL mapping is the statistical study of the alleles that occur in a locus and the phenotypes (physical forms or traits) that they produce. Because most traits of interest are governed by more than one gene, defining and studying the entire locus of genes related to a trait gives hope of understanding what effect the genotype of an individual might have in the real world.

Statistical analysis is required to demonstrate that different genes interact with one another and to determine whether they produce a significant effect on the phenotype. QTLs identify a particular region of the genome as containing a gene that is associated with the trait being assayed or measured. They are shown as intervals across a chromosome, where the probability of association is plotted for each marker used in the mapping experiment.

The QTL techniques were developed in the late 1980s and can be performed on inbred strains of any species.

To begin, a set of genetic markers must be developed for the species in question. A marker is an identifiable region of variable DNA. Biologists are interested in understanding the genetic basis of phenotypes (physical traits). The aim is to find a marker that is significantly more likely to co-occur with the trait than expected by chance, that is, a marker that has a statistical association with the trait. Ideally, they would be able to find the specific gene or genes in question, but this is a long and difficult undertaking. Instead, they can more readily find regions of DNA that are very close to the genes in question. When a QTL is found, it is often not the actual gene underlying the phenotypic trait, but rather a region of DNA that is closely linked with the gene.

For organisms whose genomes are known, one might now try to exclude genes in the identified region whose function is known with some certainty not to be connected with the trait in question. If the genome is not available, it may be an option to sequence the identified region and determine the putative functions of genes by their similarity to genes with known function, usually in other genomes. This can be done using BLAST, an online tool that allows users to enter a primary sequence and search for similar sequences within the BLAST database of genes from various organisms.

Another interest of statistical geneticists using QTL mapping is to determine the complexity of the genetic architecture underlying a phenotypic trait. For example, they may be interested in knowing whether a phenotype is shaped by many independent loci, or by a few loci, and do those loci interact. This can provide information on how the phenotype may be evolving.

Analysis of variance

The simplest method for QTL mapping is analysis of variance (ANOVA, sometimes called "marker regression") at the marker loci. In this method, in a backcross, one may calculate a t-statistic to compare the averages of the two marker genotype groups. For other types of crosses (such as the intercross), where there are more than two possible genotypes, one uses a more general form of ANOVA, which provides a so-called F-statistic. The ANOVA approach for QTL mapping has three important weaknesses. First, we do not receive separate estimates of QTL location and QTL effect. QTL location is indicated only by looking at which markers give the greatest differences between genotype group averages, and the apparent QTL effect at a marker will be smaller than the true QTL effect as a result of recombination between the marker and the QTL. Second, we must discard individuals whose genotypes are missing at the marker. Third, when the markers are widely spaced, the QTL may be quite far from all markers, and so the power for QTL detection will decrease.

Interval mapping

Lander and Botstein developed interval mapping, which overcomes the three disadvantages of analysis of variance at marker loci. Interval mapping is currently the most popular approach for QTL mapping in experimental crosses. The method makes use of a genetic map of the typed markers, and, like analysis of variance, assumes the presence of a single QTL. Each location in the genome is posited, one at a time, as the location of the putative QTL.

Composite interval mapping (CIM)

In this method, one performs interval mapping using a subset of marker loci as covariates. These markers serve as proxies for other QTLs to increase the resolution of interval mapping, by accounting for linked QTLs and reducing the residual variation. The key problem with CIM concerns the choice of suitable marker loci to serve as covariates; once these have been chosen, CIM turns the model selection problem into a single-dimensional scan. The choice of marker covariates has not been solved, however. Not surprisingly, the appropriate markers are those closest to the true QTLs, and so if one could find these, the QTL mapping problem would be complete anyway.

Non-traditional methods: Family-pedigree based mapping

Plant geneticists are attempting to incorporate some of the methods pioneered in human genetics.[10] There are some successful attempts to do so. One of quick method of QTL mapping was recently discussed.[11]

References

1. ^ Ricki Lewis (2003), *Multifactorial Traits*, McGraw-Hill Higher Education, http://highered.mcgraw-hill.com/sites/007246268x/student_view0/chapter7/
2. ^ *a b c* "Medical Genetics: Multifactorial Inheritance". Children's Hospital of the King's Daughters. 31 December 2005. http://www.chkd.org/HealthLibrary/Content.aspx?pageid=P02134. Retrieved 2007-01-06.
3. ^ *a b c* "Multifactorial Inheritance". *Pregnancy and Newborn Health Education Centre*. The March of Dimes. http://www.marchofdimes.com/pnhec/4439_4138.asp. Retrieved 2007-01-06.
4. ^ Emery's Elements of Medical Genetics
5. ^ *a b c d e f g* Tissot, Robert. "Human Genetics for 1st Year Students: Multifactorial Inheritance". http://www.uic.edu/classes/bms/bms655/lesson11.html. Retrieved 2007-01-06.
6. ^ *a b c* "Multifactorial Inheritance". *Clinical Genetics: A Self-Study Guide for Health Care Providers*. University of South Dakota School of Medicine.

http://www.usd.edu/med/som/genetics/curriculum/1GMULTI5.htm. Retrieved 2007-01-06.

7. ^ a b "Definition of Multifactorial inheritance". *MedicineNet.com MedTerms Dictionary*. MedicineNet.com. http://www.medterms.com/script/main/art.asp?articlekey=4453. Retrieved 2007-01-06.
8. ^ a b c Turnpenny, Peter (2004). "Emery's Elements of Medical Genetics, 12th Edition, Chapter 9" (PDF). Elsevier. http://www.fleshandbones.com/readingroom/viewchapter.cfm?ID=1041. Retrieved 2007-01-06.
9. ^ Douthit, Kathryn. "Preserving the Role of Counseling in the Age of Biopsychiatry: Critical Reflections on the DSM-IV-TR". VISTAS Online. http://counselingoutfitters.com/Douthit2.htm. Retrieved 2007-08-27.
10. ^ Jannink, J; Bink, Mc; Jansen, Rc (Aug 2001). "Using complex plant pedigrees to map valuable genes". *Trends in plant science* **6** (8): 337–42. ISSN 1360-1385. PMID 11495765.
11. ^ Rosyara, U. R.; Maxson-stein, K.L.; Glover, K.D.; Stein, J.M.; Gonzalez-hernandez, J.L. (2007), "Family-based mapping of FHB resistance QTLs in hexaploid wheat", *Proceedings of National Fusarium head blight forum*

Epistasis

Epistasis is the phenomenon where the effects of one gene are modified by one or several other genes, which are sometimes called **modifier genes**. The gene whose phenotype is expressed is said to be **epistatic**, while the phenotype altered or suppressed is said to be **hypostatic**. Epistasis can be contrasted with dominance, which is an interaction between alleles at the same gene locus. Epistasis is often studied in relation to Quantitative Trait Loci (QTL) and polygenic inheritance.

In general, the fitness increment of any one allele depends in a complicated way on many other alleles; but, because of the way that the science of population genetics was developed, evolutionary scientists tend to think of epistasis as the exception to the rule. In the first models of natural selection devised in the early 20th century, each gene was considered to make its own characteristic contribution to fitness, against an average background of other genes. Some introductory college courses still teach population genetics this way.

Epistasis and **genetic interaction** refer to different aspects of the same phenomenon. The term **epistasis** is widely used in population genetics and refers especially to the statistical properties of the phenomenon, and does not necessarily imply biochemical interaction between gene products. However, in general epistasis is used to denote the departure from 'independence' of the effects of different genetic loci. Confusion often arises due to the varied interpretation of 'independence' between different branches of biology. For further discussion of the definitions of epistasis, and the history of these definitions, see [1].

Examples of tightly linked genes having epistatic effects on fitness are found in supergenes and the human major histocompatibility complex genes. The effect can occur directly at the genomic level, where one gene could code for a protein preventing transcription of the other gene. Alternatively, the effect can occur at the phenotypic level. For example, the gene causing albinism would hide the gene controlling color of a person's hair. In another example, a gene coding for a widow's peak would be hidden by a gene causing baldness. Fitness epistasis (where the affected trait is fitness) is one cause of linkage disequilibrium.

Studying genetic interactions can reveal gene function, the nature of the mutations, functional redundancy, and protein interactions. Because protein complexes are responsible for most biological functions, genetic interactions are a powerful tool.

Classification by fitness or trait value

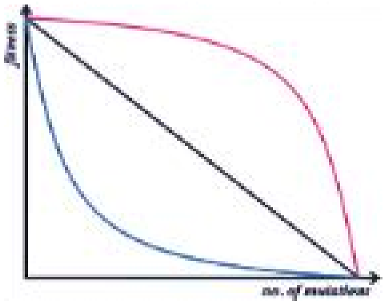


Diagram illustrating different relationships between numbers of mutations and fitness. *Synergistic* epistasis is the blue line - each mutation has a disproportionately large effect on the organism's fitness. *Antagonistic* epistasis is the red line. See Evolution of Sex

Two-locus epistatic interactions can be either synergistic (enhancing the effectiveness) or antagonistic (reducing the activity).[2][3] In the example of a haploid organism with genotypes (at two loci) *AB*, *Ab*, *aB* or *ab*, we can think of the following trait values where higher values suggest greater expression of the characteristic (the exact values are simply given as examples):

	AB	*Ab*	*aB*	*ab*
No epistasis (additive across loci)	2	1	1	0
Synergistic epistasis	3	1	1	0

Antagonistic epistasis	1	1	1	0

Hence, we can classify thus:

Trait values	**Type of epistasis**
$AB = Ab + aB - ab$	No epistasis, additive inheritance
$AB > Ab + aB - ab$	Synergistic epistasis
$AB < Ab + aB - ab$	Antagonistic epistasis

Understanding whether the majority of genetic interactions are synergistic or antagonistic will help solve such problems as the evolution of sex.

Epistasis and sex

Negative epistasis and sex are thought to be intimately correlated. Experimentally, this idea has been tested in using digital simulations of asexual and sexual populations. Over time, sexual populations move towards more negative epistasis, or the lowering of fitness by two interacting alleles. It is thought that negative epistasis allows individuals carrying the interacting deleterious mutations to be removed from the populations efficiently. This removes those alleles from the population, resulting in an overall more fit population. This hypothesis was proposed by Alexey Kondrashov, and is sometimes known as the *deterministic mutation hypothesis*[4] and has also been tested using artificial gene networks.[2]

However, the evidence for this hypothesis has not always been straightforward and the model proposed by Kondrashov has been criticized for assuming mutation parameters far from real world observations. For example, see [5]. In addition, in those tests which used artificial gene networks, negative epistasis is only found in more densely connected networks[2], whereas empirical evidence indicates that natural gene networks are sparsely connected[6], and theory shows that selection for robustness will favor more sparsely connected and minimally complex networks.[6]

Functional or mechanistic classification

- **Genetic suppression** - the double mutant has a less severe phenotype than either single mutant.
- **Genetic enhancement** - the double mutant has a more severe phenotype than one predicted by the additive effects of the single mutants.
- **Synthetic lethality** or **unlinked non-complementation** - two mutations fail to complement and yet do not map to the same locus.
- **Intragenic complementation**, **allelic complementation**, or **interallelic complementation** - two mutations map to the same locus, yet the two alleles complement in the heteroallelic diploid. Causes of intragenic complementation include:

- homology effects such as transvection, where, for example, an enhancer from one allele acts in *trans* to activate transcription from the promoter of the second allele.
- trans-splicing of two mutant RNA molecules to produce a functional RNA.
- At the protein level, another possibility involves proteins that normally function as dimers. In a heteroallelic diploid, two different abnormal proteins could form a functional dimer if each can compensate for the lack of function in the other.

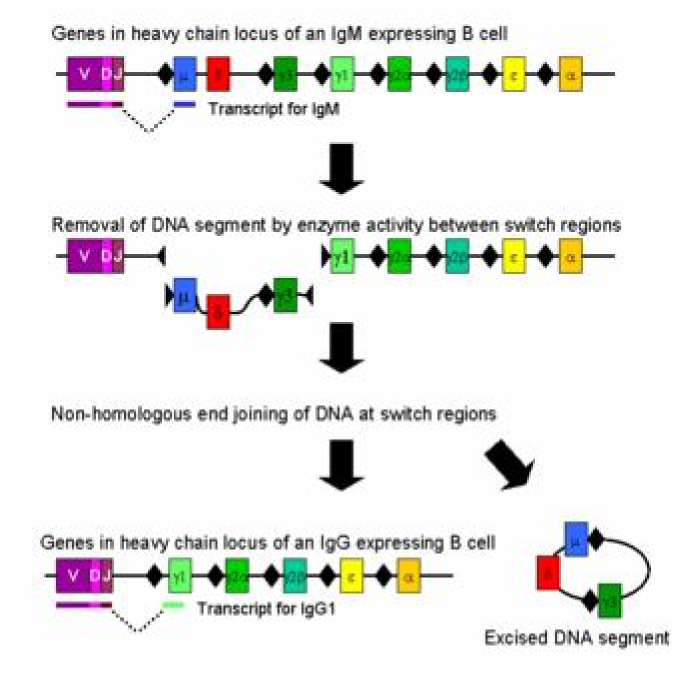
References

1. ^ Cordell, Heather J. (2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". *Human Molecular Genetics* **11** (20): 2463–8. doi:10.1093/hmg/11.20.2463.
2. ^ *a b c* Azevedo R, Lohaus R, Srinivasan S, Dang K, Burch C (2006). "Sexual reproduction selects for robustness and negative epistasis in artificial gene networks". *Nature* **440** (7080): 87–90. doi:10.1038/nature04488. PMID 16511495.
3. ^ Bonhoeffer S, Chappey C, Parkin NT, Whitcomb JM, Petropoulos CJ (2004). "Evidence for positive epistasis in HIV-1". *Science* **306** (5701): 1547–50. doi:10.1126/science.1101786. PMID 15567861.
4. ^ A. S. Kondrashov (1988). "Deleterious mutations and the evolution of sexual reproduction". *Nature* **336**: 435–440. doi:10.1038/336435a0.
5. ^ MacCarthy T, Bergman A. (July 2007). "Coevolution of robustness, epistasis, and recombination favors asexual reproduction". *Proc Natl Acad Sci U S A* **104** (31): 12801–6. doi:10.1073/pnas.0705455104. PMID 17646644.
6. ^ *a b* Leclerc R. (August 2008). "Survival of the sparsest: robust gene networks are parsimonious". *Mol Syst Biol.* **4** (213).

Immunoglobulin class switching

From Wikipedia, the free encyclopedia

Jump to: navigation, search


Genes in heavy chain locus of an IgM expressing B cell
V
DJ
μ
δ
γ3
γ1
γ2a
γ2b
ε
α
Transcript for IgM
Removal of DNA segment by enzyme activity between switch regions
Non-homologous end joining of DNA at switch regions
Genes in heavy chain locus of an IgG expressing B cell
Transcript for IgG1
Excised DNA segment

Mechanism of class switch recombination that allows isotype switching in activated B cells

Immunoglobulin class switching (or **isotype switching** or **isotypic commutation** or **class switch recombination**) is a biological mechanism that changes a B cell's production of antibody from one class to another, for example, from an isotype called IgM to an isotype called IgG. During this process, the constant region portion of the antibody heavy chain is changed, but the variable region of the heavy chain stays the same (the terms "constant" and "variable" refer to changes or lack thereof between antibodies that target different epitopes). Since the variable region does not change, class switching does not affect antigen specificity. Instead, the antibody retains affinity for the same antigens, but can interact with different effector molecules.

Contents

[hide]

[edit] Mechanism

Class switching occurs after activation of a mature B cell via its membrane-bound antibody molecule (or B cell receptor) to generate the different classes of antibody, all with the same variable domains as the original antibody generated in the immature B cell during the process of V(D)J recombination, but possessing distinct constant domains in their heavy chains.[1]

Naïve mature B cells produce both IgM and IgD, which are the first two heavy chain segments in the immunoglobulin locus. After activation by antigen, these B cells proliferate and begin to produce high levels of these antibodies. If these activated B cells are also activated via their CD40 and IL-4 receptors (both modulated by T helper cells), they undergo antibody class switching to produce IgG, IgA or IgE antibodies. During class switching, the constant region of the immunoglobulin heavy chain changes but the variable regions, and therefore antigen specificity, stay the same. This allows different daughter cells from the same activated B cell to produce antibodies of different isotypes or subtypes (e.g. IgG1, IgG2 etc.).[2]

The order of the heavy chain exons are as follows:

- μ - IgM
- δ - IgD
- γ3 - IgG3
- γ1 - IgG1
- pseudogene similar to ε gene that is not used
- α1 - IgA1
- γ2 - IgG2
- γ4 - IgG4
- ε - IgE
- α2 - IgA2

Class switching occurs by a mechanism called class switch recombination (CSR) binding. Class switch recombination is a biological mechanism that allows the class of antibody produced by an activated B cell to change during a process known as isotype or class switching. During CSR, portions of the antibody heavy chain locus are removed from the chromosome, and the gene segments surrounding the deleted portion are rejoined to retain a functional antibody gene that produces antibody of a different isotype. Double-stranded breaks are generated in DNA at conserved nucleotide motifs, called switch (S) regions, which are upstream from gene segments that encode the constant regions of antibody heavy chains; these occur adjacent to all heavy chain constant region genes with the exception of the δ-chain. DNA is nicked and broken at two selected S-regions by the activity of a series of enzymes, including Activation-Induced (Cytidine) Deaminase (AID), uracil DNA glycosylase and apyrimidic/apurinic (AP)-endonucleases.[3][4] The intervening DNA between the S-regions is subsequently deleted from the chromosome, removing unwanted μ or δ heavy chain constant region exons and allowing substitution of a γ, α or ε constant region gene segment. The free ends of the DNA are rejoined by a process called non-homologous end joining (NHEJ) to link the variable domain exon to the desired downstream constant domain exon of the antibody heavy chain.[5] In the absence of non-homologous end joining, free ends of DNA may be rejoined by an alternative pathway biased toward microhomology joins.[6] With the exception of the μ and δ genes, only one antibody class is expressed by a B cell at any point in time.

[edit] Cytokines responsible for class switching

T cell cytokines are responsible for class switching in mouse (Table 1) and human (Table 2).[7] [8] These cytokines may have suppressive effect on production of IgM.

Table 1. Class switching in mouse

T cells	Cytokines	Immunoglobulin classes					
		IgG1	IgG2a	IgG2b	IgG3	IgA	IgE
Th2	IL-4	↑	↓	↓	↓	↓	↑
	IL-5					↑	
Th1	IFNγ	↓	↑	↓	↓	↓	↓
Treg	TGFβ			↑	↓	↑	

Table 2. Class switching in human

T cells	Cytokines	Immunoglobulin classes					
		IgG1	IgG2	IgG3	IgG4	IgA	IgE
Th2	IL-4				↑		↑
	IL-5					↑	
Th1	IFNγ						
Treg	TGFβ					↑	

[edit] See also

- immunogenetics
- antibody

[edit] References

1. ^ Eleonora Market, F. Nina Papavasiliou (2003) *V(D)J Recombination and the Evolution of the Adaptive Immune System* PLoS Biology 1(1): e16.
2. ^ Stavnezer J, Amemiya CT (2004). "Evolution of isotype switching". *Semin. Immunol.* **16** (4): 257–75. doi:10.1016/j.smim.2004.08.005. PMID 15522624.
3. ^ Durandy A (2003). "Activation-induced cytidine deaminase: a dual role in class-switch recombination and somatic hypermutation". *Eur. J. Immunol.* **33** (8): 2069–73. doi:10.1002/eji.200324133. PMID 12884279.
4. ^ Casali P, Zan H (2004). "Class switching and Myc translocation: how does DNA break?". *Nat. Immunol.* **5** (11): 1101–3. doi:10.1038/ni1104-1101. PMID 15496946.

5. ^ Lieber MR, Yu K, Raghavan SC (2006). "Roles of nonhomologous DNA end joining, V(D)J recombination, and class switch recombination in chromosomal translocations". *DNA Repair (Amst.)* **5** (9-10): 1234–45. doi:10.1016/j.dnarep.2006.05.013. PMID 16793349.
6. ^ Yan CT, Boboila C, Souza EK, Franco S, Hickernell TR, Murphy M, Gumaste S, Geyer M, Zarrin AA, Manis JP, Rajewsky K, Alt FW (2007). "IgH class switching and translocations use a robust non-classical end-joining pathway". *Nature* **449**: 478–82. doi:10.1038/nature06020. PMID 17713479.
7. ^ Janeway CA Jr., Travers P, Walport M, Shlomchik MJ (2001). *Immunobiology*. (5th ed.). Garland Publishing. (via NCBI Bookshelf) ISBN 0-8153-3642-X.
8. ^ Male D, Brostoff J, Roth DB, Roitt I (2006). *Immunology, 7th ed.* Philadelphia: Mosby Elsevier, ISBN 9780323033992 (pbk.)

[edit]