National Center for Biotechnology Information

A Basic Introduction to the Science Underlying NCBI Resources

WHAT IS A GENOME?

Life is specified by **genomes**. Every organism, including humans, has a genome that contains all of the biological information needed to build and maintain a living example of that organism. The biological information contained in a genome is encoded in its **deoxyribonucleic acid (DNA)** and is divided into discrete units called **genes**. Genes code for proteins that attach to the genome at the appropriate positions and switch on a series of reactions called gene expression.

> In 1909, Danish botanist Wilhelm Johanssen coined the word **gene** for the hereditary unit found on a chromosome. Nearly 50 years earlier, Gregor Mendel had characterized hereditary units as **factors**— observable differences that were passed from parent to offspring. Today we know that a single gene consists of a unique sequence of DNA that provides the complete instructions to make a functional product, called a protein. Genes instruct each cell type—such as skin, brain, and liver—to make discrete sets of proteins at just the right times, and it is through this specificity that unique organisms arise.

The Physical Structure of the Human Genome

Nuclear DNA

Inside each of our cells lies a **nucleus**, a membrane-bounded region that provides a sanctuary for genetic information. The nucleus contains long strands of DNA that encode this genetic information. A **DNA** chain is made up of four **chemical bases**: **adenine** (A) and **guanine** (G), which are called **purines**, and **cytosine** (C) and **thymine** (T), referred to as **pyrimidines**. Each base has a slightly different composition, or combination of

oxygen, carbon, nitrogen, and hydrogen. In a DNA chain, every base is attached to a sugar molecule (deoxyribose) and a phosphate molecule, resulting in a nucleic acid or **nucleotide**. Individual nucleotides are linked through the phosphate group, and it is the precise order, or sequence, of nucleotides that determines the product made from that gene.

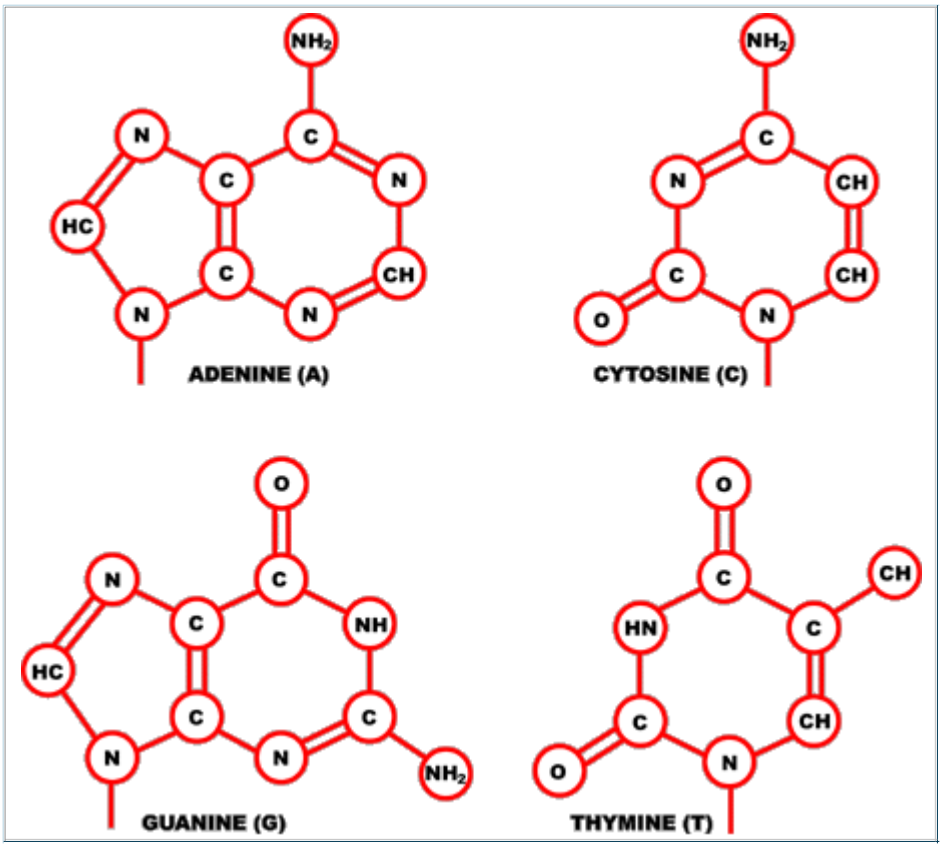

Figure 1. The four DNA bases.

Each DNA base is made up of the sugar 2'-deoxyribose linked to a phosphate group and one of the four bases depicted above: adenine (*top left*), cytosine (*top right*), guanine (*bottom left*), and thymine (*bottom right*).

A DNA chain, also called a strand, has a sense of direction, in which one end is chemically different than the other. The so-called 5' end terminates in a 5' phosphate group (-PO4); the 3' end terminates in a 3' hydroxyl group (-OH). This is important because DNA strands are always synthesized in the 5' to 3' direction.

The DNA that constitutes a gene is a double-stranded molecule consisting of two chains running in opposite directions. The chemical nature of the bases in double-stranded DNA creates a slight twisting force that gives DNA its characteristic gently coiled structure, known as the double helix. The two strands are connected to each other by chemical pairing of each base on one strand to a specific partner on the other strand. Adenine (A) pairs with thymine (T), and guanine (G) pairs with cytosine (C). Thus,

A-T and **G-C base pairs** are said to be **complementary**. This complementary base pairing is what makes DNA a suitable molecule for carrying our genetic information—one strand of DNA can act as a **template** to direct the synthesis of a complementary strand. In this way, the information in a DNA sequence is readily copied and passed on to the next generation of cells.

Organelle DNA

Not all genetic information is found in nuclear DNA. Both plants and animals have an organelle—a "little organ" within the cell—called the **mitochondrion**. Each mitochondrion has its own set of genes. Plants also have a second organelle, the **chloroplast**, which also has its own DNA. Cells often have multiple mitochondria, particularly cells requiring lots of energy, such as active muscle cells. This is because mitochondria are responsible for converting the energy stored in macromolecules into a form usable by the cell, namely, the **adenosine triphosphate (ATP) molecule**. Thus, they are often referred to as the power generators of the cell.

Unlike **nuclear DNA** (the DNA found within the nucleus of a cell), half of which comes from our mother and half from our father, mitochondrial DNA is only inherited from our mother. This is because mitochondria are only found in the female gametes or "eggs" of sexually reproducing animals, not in the male gamete, or sperm. Mitochondrial DNA also does not recombine; there is no shuffling of genes from one generation to the other, as there is with nuclear genes.

> Large numbers of mitochondria are found in the tail of sperm, providing them with an engine that generates the energy needed for swimming toward the egg. However, when the sperm enters the egg during fertilization, the tail falls off, taking away the father's mitochondria.

Why Is There a Separate Mitochondrial Genome?

The energy-conversion process that takes place in the mitochondria takes place **aerobically**, in the presence of oxygen. Other energy conversion processes in the cell take place **anaerobically**, or without oxygen. The independent aerobic function of these organelles is thought to have evolved from bacteria that lived inside of other simple organisms in a mutually beneficial, or **symbiotic**, relationship, providing them with aerobic capacity. Through the process of evolution, these tiny organisms became incorporated into the cell, and their genetic systems and cellular functions became integrated to form a single functioning cellular unit. Because mitochondria have their own DNA, RNA, and ribosomes, this scenario is quite possible. This theory is also

supported by the existence of a eukaryotic organism, called the amoeba, which lacks mitochondria. Therefore, amoeba must always have a symbiotic relationship with an aerobic bacterium.

Why Study Mitochondria?

There are many diseases caused by mutations in **mitochondrial DNA (mtDNA)**. Because the mitochondria produce energy in cells, symptoms of mitochondrial diseases often involve degeneration or functional failure of tissue. For example, mtDNA mutations have been identified in some forms of diabetes, deafness, and certain inherited heart diseases. In addition, mutations in mtDNA are able to accumulate throughout an individual's lifetime. This is different from mutations in nuclear DNA, which has sophisticated repair mechanisms to limit the accumulation of mutations. Mitochondrial DNA mutations can also concentrate in the mitochondria of specific tissues. A variety of deadly diseases are attributable to a large number of accumulated mutations in mitochondria. There is even a theory, the **Mitochondrial Theory of Aging**, that suggests that accumulation of mutations in mitochondria contributes to, or drives, the aging process. These defects are associated with Parkinson's and Alzheimer's disease, although it is not known whether the defects actually cause or are a direct result of the diseases. However, evidence suggests that the mutations contribute to the progression of both diseases.

In addition to the critical cellular energy-related functions, mitochondrial genes are useful to evolutionary biologists because of their maternal inheritance and high rate of mutation. By studying patterns of mutations, scientists are able to reconstruct patterns of migration and evolution within and between species. For example, mtDNA analysis has been used to trace the migration of people from Asia across the Bering Strait to North and South America. It has also been used to identify an ancient maternal lineage from which modern man evolved.

Ribonucleic Acids

Just like DNA, **ribonucleic acid (RNA)** is a chain, or polymer, of nucleotides with the same 5' to 3' direction of its strands. However, the ribose sugar component of RNA is slightly different chemically than that of DNA. RNA has a 2' oxygen atom that is not present in DNA. Other fundamental structural differences exist. For example, uracil takes the place of the thymine nucleotide found in DNA,

In addition to mRNA, DNA codes for other forms of RNA, including ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and small nuclear RNAs (snRNAs). rRNAs and tRNAs participate in protein assembly whereas snRNAs aid in a process called splicing —the process of editing of mRNA before it can be used as a template for protein synthesis.

and RNA is, for the most part, a single-stranded molecule. DNA directs the synthesis of a variety of RNA molecules, each with a unique role in cellular function. For example, all genes that code for proteins are first made into an RNA strand in the nucleus called a **messenger RNA** (mRNA). The mRNA carries the information encoded in DNA out of the nucleus to the protein assembly machinery, called the **ribosome**, in the cytoplasm. The ribosome complex uses mRNA as a template to synthesize the exact protein coded for by the gene.

Proteins

> "DNA makes RNA, RNA makes protein, and proteins make us."
> **Francis Crick**

Although DNA is the carrier of genetic information in a cell, proteins do the bulk of the work. Proteins are long chains containing as many as 20 different kinds of amino acids. Each cell contains thousands of different proteins: **enzymes** that make new molecules and catalyze nearly all chemical processes in cells; **structural components** that give cells their shape and help them move; **hormones** that transmit signals throughout the body; **antibodies** that recognize foreign molecules; and **transport molecules** that carry oxygen. The genetic code carried by DNA is what specifies the order and number of amino acids and, therefore, the shape and function of the protein.

The "**Central Dogma**"—a fundamental principle of molecular biology—states that genetic information flows from DNA to RNA to protein. Ultimately, however, the genetic code resides in DNA because only DNA is passed from generation to generation. Yet, in the process of making a protein, the encoded information must be faithfully transmitted first to RNA then to protein. Transferring the code from DNA to RNA is a fairly straightforward process called **transcription**. Deciphering the code in the resulting mRNA is a little more complex. It first requires that the mRNA leave the nucleus and associate with a large complex of specialized RNAs and proteins that, collectively, are called the **ribosome**. Here the mRNA is translated into protein by decoding the mRNA sequence in blocks of three RNA bases, called **codons**, where each codon specifies a particular amino acid. In this way, the **ribosomal complex** builds a protein one amino acid at a time, with the order of amino acids determined precisely by the order of the codons in the mRNA.

> In 1961, Marshall Nirenberg and Heinrich Matthaei correlated the first codon (UUU) with the amino acid phenylalanine. After that, it was not long before the genetic code for all 20 amino acids was deciphered.

A given amino acid can have more than one codon. These redundant codons usually differ at the third position. For example, the amino acid serine is encoded by UCU, UCC, UCA, and/or UCG. This redundancy is key to accommodating mutations that occur naturally as DNA is replicated and new cells are produced. By allowing some of the random changes in DNA to have no effect on the ultimate protein sequence, a sort of genetic safety net is created. Some codons do not code for an amino acid at all but instruct the ribosome when to stop adding new amino acids.

Table 1. RNA triplet codons and their corresponding amino acids.

	U	C	A	G
U	UU**U** Phenylalanine UU**C** Phenylalanine UU**A** Leucine UU**G** Leucine	UC**U** Serine UC**C** Serine UC**A** Serine UC**G** Serine	UA**U** Tyrosine UA**C** Tyrosine UA**A** Stop UA**G** Stop	UG**U** Cysteine UG**C** Cysteine UG**A** Stop UG**G** Tryptophan
C	CU**U** Leucine CU**C** Leucine CU**A** Leucine CU**G** Leucine	CC**U** Proline CC**C** Proline CC**A** Proline CC**G** Proline	CA**U** Histidine CA**C** Histidine CA**A** Glutamine CA**G** Glutamine	CG**U** Arginine CG**C** Arginine CG**A** Arginine CG**G** Arginine
A	AU**U** Isoleucine AU**C** Isoleucine AU**A** Isoleucine AU**G** Methionine	AC**U** Threonine AC**C** Threonine AC**A** Threonine AC**G** Threonine	AA**U** Asparagine AA**C** Asparagine AA**A** Lysine AA**G** Lysine	AG**U** Serine AG**C** Serine AG**A** Arginine AG**G** Arginine
G	GU**U** Valine GU**C** Valine GU**A** Valine GU**G** Valine	GC**U** Alanine GC**C** Alanine GC**A** Alanine GC**G** Alanine	GA**U** Aspartate GA**C** Aspartate GA**A** Glutamate GA**G** Glutamate	GG**U** Glycine GG**C** Glycine GG**A** Glycine GG**G** Glycine

A translation chart of the 64 RNA codons.

The Core Gene Sequence: Introns and Exons

Genes make up about 1 percent of the total DNA in our genome. In the human genome, the coding portions of a gene, called **exons**, are interrupted by intervening sequences, called **introns**. In addition, a eukaryotic gene does not code for a protein in one continuous stretch of DNA. Both exons and introns are **"transcribed"** into mRNA, but before it is transported to the ribosome, the primary mRNA transcript is edited. This editing process removes the introns, joins the exons together, and adds unique features to each end of the transcript to make a **"mature"** mRNA. One might then ask what the purpose of an intron is if it is spliced out after it is transcribed? It is still unclear what all the functions of introns are, but scientists believe that some serve as the site for **recombination**, the process by which progeny derive a combination of genes different from that of either parent, resulting in novel genes with new combinations of exons, the key to evolution.

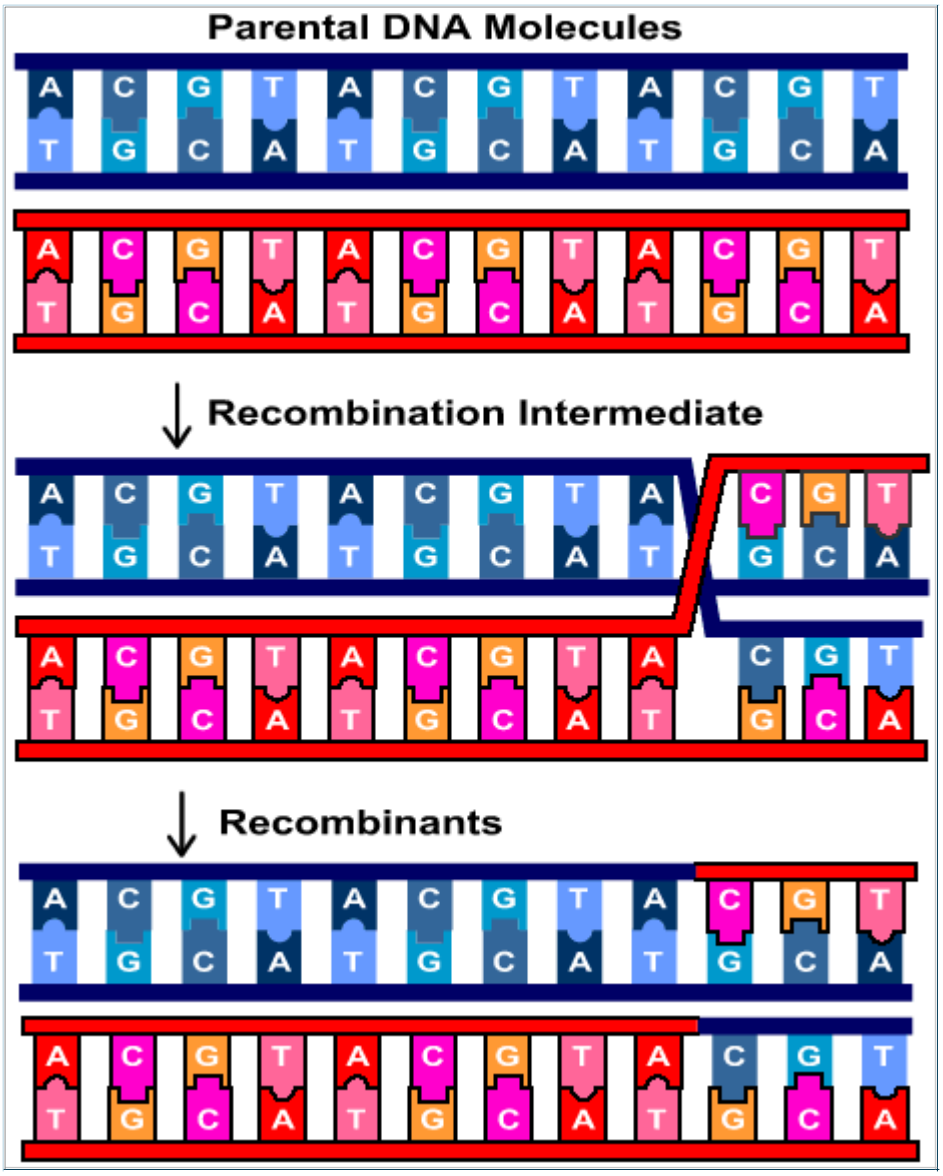

Figure 2. Recombination.

Recombination involves pairing between complementary strands of two parental duplex DNAs (*top* and *middle panel*). This process creates a stretch of hybrid DNA (*bottom panel*) in which the single strand of one duplex is paired with its complement from the other duplex.

Gene Prediction Using Computers

When the complete mRNA sequence for a gene is known, computer programs are used to align the mRNA sequence with the appropriate region of the genomic DNA sequence. This provides a reliable indication of the beginning and end of the coding region for that gene. In the absence of a complete mRNA sequence, the boundaries can be estimated by ever-improving, but still inexact, gene prediction software. The problem is the lack of a single sequence pattern that indicates the beginning or end of a eukaryotic gene. Fortunately, the middle of a gene, referred to as the **core gene sequence**--has enough consistent features to allow more reliable predictions.

From Genes to Proteins: Start to Finish

We just discussed that the journey from DNA to mRNA to protein requires that a cell identify where a gene begins and ends. This must be done both during the transcription and the translation process.

Transcription

Transcription, the synthesis of an RNA copy from a sequence of DNA, is carried out by an enzyme called **RNA polymerase**. This molecule has the job of recognizing the DNA sequence where transcription is initiated, called the **promoter site**. In general, there are two "promoter" sequences upstream from the beginning of every gene. The location and base sequence of each promoter site vary for **prokaryotes** (bacteria) and **eukaryotes** (higher organisms), but they are both recognized by RNA polymerase, which can then grab hold of the sequence and drive the production of an mRNA.

Eukaryotic cells have three different RNA polymerases, each recognizing three classes of genes. **RNA polymerase II** is responsible for synthesis of mRNAs from protein-coding genes. This polymerase requires a sequence resembling TATAA, commonly referred to as the **TATA box**, which is found 25-30 nucleotides upstream of the beginning of the gene, referred to as the **initiator sequence**.

Transcription terminates when the polymerase stumbles upon a termination, or stop signal. In eukaryotes, this process is not fully understood. Prokaryotes, however, tend to have a short region composed of G's and C's that is able to fold in on itself and form complementary base pairs, creating a stem in the new mRNA. This stem then causes the polymerase to trip and release the **nascent**, or newly formed, mRNA.

Translation

The beginning of **translation**, the process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids, differs slightly for prokaryotes and eukaryotes, although both processes always initiate at a codon for methionine. For prokaryotes, the ribosome recognizes and attaches at the sequence AGGAGGU on the mRNA, called the **Shine-Delgarno sequence**, that appears just upstream from the methionine (AUG) codon. Curiously, eukaryotes lack this recognition sequence and simply initiate translation at the amino acid methionine, usually coded for by the bases AUG, but sometimes GUG. Translation is terminated for both prokaryotes and eukaryotes when the ribosome reaches one of the three stop codons.

Structural Genes, Junk DNA, and Regulatory Sequences

Structural Genes

Over 98 percent of the genome is of unknown function. Although often referred to as **"junk" DNA**, scientists are beginning to uncover the function of many of these intergenic sequences—the DNA found between genes.

Sequences that code for proteins are called **structural genes**. Although it is true that proteins are the major components of structural elements in a cell, proteins are also the real workhorses of the cell. They perform such functions as transporting nutrients into the cell; synthesizing new DNA, RNA, and protein molecules; and transmitting chemical signals from outside to inside the cell, as well as throughout the cell—both critical to the process of making proteins.

Regulatory Sequences

A class of sequences called **regulatory sequences** makes up a numerically insignificant fraction of the genome but provides critical functions. For example, certain sequences indicate the beginning and end of genes, sites for initiating replication and recombination, or provide landing sites for proteins that turn genes on and off. Like structural genes, regulatory sequences are inherited; however, they are not commonly referred to as genes.

Other DNA Regions

Forty to forty-five percent of our genome is made up of short sequences that are repeated, sometimes hundreds of times. There are numerous forms of this **"repetitive DNA"**, and a few have known functions, such as stabilizing the chromosome structure or inactivating one of the two X chromosomes in developing females, a process called **X-inactivation**. The most highly repeated sequences found so far in mammals are called **"satellite DNA"** because their unusual composition allows them to be easily separated from other DNA. These sequences are associated with chromosome structure and are found at the **centromeres** (or centers) and **telomeres** (ends) of chromosomes. Although they do not play a role in the coding of proteins, they do play a significant role in chromosome structure, duplication, and cell division. The highly variable nature of these sequences makes them an excellent **"marker"** by which individuals can be identified based on their unique pattern of their satellite DNA.

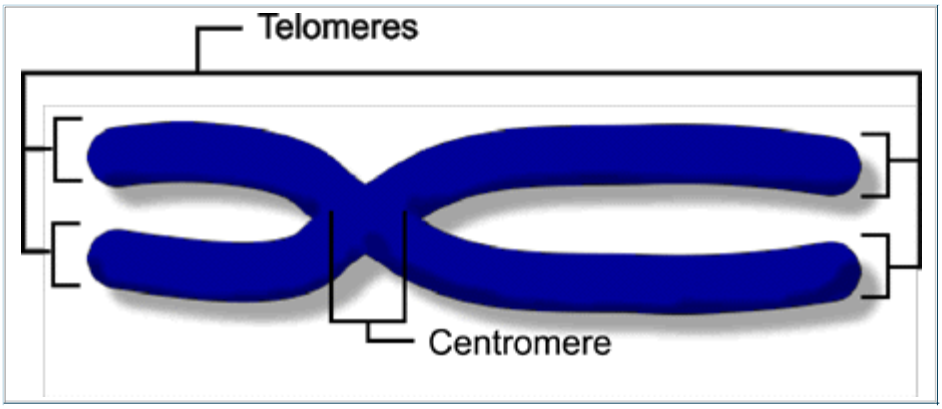

Figure 3. A chromosome.

A chromosome is composed of a very long molecule of DNA and associated proteins that carry hereditary information. The centromere, shown at the center of this chromosome, is a specialized structure that appears during cell division and ensures the correct distribution of duplicated chromosomes to daughter cells. Telomeres are the structures that seal the end of a chromosome. Telomeres play a critical role in chromosome replication and maintenance by counteracting the tendency of the chromosome to otherwise shorten with each round of replication.

Another class of non-coding DNA is the "**pseudogene**", so named because it is believed to be a remnant of a real gene that has suffered mutations and is no longer functional. Pseudogenes may have arisen through the duplication of a functional gene, followed by inactivation of one of the copies. Comparing the presence or absence of pseudogenes is one method used by evolutionary geneticists to group species and to determine relatedness. Thus, these sequences are thought to carry a record of our evolutionary history.

How Many Genes Do Humans Have?

In February 2001, two largely independent draft versions of the human genome were published. Both studies estimated that there are 30,000 to 40,000 genes in the human genome, roughly one-third the number of previous estimates. More recently scientists estimated that there are less than 30,000 human genes. However, we still have to make guesses at the actual number of genes, because not all of the human genome sequence is annotated and not all of the known sequence has been assigned a particular position in the genome.

So, how do scientists estimate the number of genes in a genome? For the most part, they look for tell-tale signs of genes in a DNA sequence. These include: **open reading frames**, stretches of DNA, usually greater than 100 bases, that are not interrupted by a **stop codon** such as TAA, TAG or TGA; **start codons** such as ATG; specific sequences found at **splice junctions**, a location in the DNA sequence where RNA removes the non-coding areas to

form a continuous gene transcript for translation into a protein; and **gene regulatory sequences**. This process is dependent on computer programs that search for these patterns in various sequence databases and then make predictions about the existence of a gene.

From One Gene–One Protein to a More Global Perspective

Only a small percentage of the 3 billion bases in the human genome becomes an expressed gene product. However, of the approximately 1 percent of our genome that is expressed, 40 percent is alternatively spliced to produce multiple proteins from a single gene. **Alternative splicing** refers to the cutting and pasting of the primary mRNA transcript into various combinations of mature mRNA. Therefore the one gene–one protein theory, originally framed as "one gene–one enzyme", does not precisely hold.

With so much DNA in the genome, why restrict transcription to a tiny portion, and why make that tiny portion work overtime to produce many alternate transcripts? This process may have evolved as a way to limit the deleterious effects of mutations. Genetic mutations occur randomly, and the effect of a small number of mutations on a single gene may be minimal. However, an individual having many genes each with small changes could weaken the individual, and thus the species. On the other hand, if a single mutation affects several alternate transcripts at once, it is more likely that the effect will be devastating—the individual may not survive to contribute to the next generation. Thus, alternate transcripts from a single gene could reduce the chances that a mutated gene is transmitted.

Gene Switching: Turning Genes On and Off

The estimated number of genes for humans, less than 30,000, is not so different from the 25,300 known genes of *Arabidopsis thaliana*, commonly called mustard grass. Yet, we appear, at least at first glance, to be a far more complex organism. A person may wonder how this increased complexity is achieved. One answer lies in the regulatory system that turns genes on and off. This system also precisely controls the amount of a gene product that is produced and can further modify the product after it is made. This exquisite control requires multiple regulatory input points. One very efficient point occurs at transcription, such that an mRNA is produced only when a gene product is needed. Cells also regulate gene expression by **post-transcriptional modification**; by allowing only a subset of the mRNAs to go on to translation; or by restricting translation of specific mRNAs to only when the product is needed. At other levels, cells regulate gene expression through DNA folding, chemical modification of the nucleotide bases, and intricate "**feedback mechanisms**" in which

some of the gene's own protein product directs the cell to cease further protein production.

Controlling Transcription

Promoters and Regulatory Sequences

Transcription is the process whereby RNA is made from DNA. It is initiated when an enzyme, **RNA polymerase**, binds to a site on the DNA called a **promoter sequence**. In most cases, the polymerase is aided by a group of proteins called "**transcription factors**" that perform specialized functions, such as DNA sequence recognition and regulation of the polymerase's enzyme activity. Other regulatory sequences include **activators**, **repressors**, and **enhancers**. These sequences can be ***cis*-acting** (affecting genes that are adjacent to the sequence) or ***trans*-acting** (affecting expression of the gene from a distant site), even on another chromosome.

The Globin Genes: An Example of Transcriptional Regulation

An example of transcriptional control occurs in the family of genes responsible for the production of globin. **Globin** is the protein that complexes with the iron-containing heme molecule to make hemoglobin. **Hemoglobin** transports oxygen to our tissues via red blood cells. In the adult, red blood cells do not contain DNA for making new globin; they are ready-made with all of the hemoglobin they will need.

During the first few weeks of life, embryonic globin is expressed in the yolk sac of the egg. By week five of gestation, globin is expressed in early liver cells. By birth, red blood cells are being produced, and globin is expressed in the bone marrow. Yet, the globin found in the yolk is not produced from the same gene as is the globin found in the liver or bone marrow stem cells. In fact, at each stage of development, different globin genes are turned on and off through a process of transcriptional regulation called **"switching"**.

To further complicate matters, globin is made from two different protein chains: an alpha-like chain coded for on chromosome 16; and a beta-like chain coded for on chromosome 11. Each chromosome has the embryonic, fetal, and adult form lined up on the chromosome in a sequential order for developmental expression. The developmentally regulated transcription of globin is controlled by a number of *cis*-acting DNA sequences, and although there remains a lot to be learned about the interaction of these sequences, one known control sequence is an enhancer called the **Locus Control Region (LCR)**. The LCR sits far upstream on the sequence and controls the alpha genes on chromosome 16. It may also interact with other factors to determine which alpha gene is turned on.

Thalassemias are a group of diseases characterized by the absence or decreased production of normal globin, and thus

hemoglobin, leading to decreased oxygen in the system. There are alpha and beta thalassemias, defined by the defective gene, and there are variations of each of these, depending on whether the embryonic, fetal, or adult forms are affected and/or expressed. Although there is no known cure for the thalassemias, there are medical treatments that have been developed based on our current understanding of both gene regulation and cell differentiation. Treatments include blood transfusions, iron chelators, and bone marrow transplants. With continuing research in the areas of gene regulation and cell differentiation, new and more effective treatments may soon be on the horizon, such as the advent of gene transfer therapies.

The Influence of DNA Structure and Binding Domains

Sequences that are important in regulating transcription do not necessarily code for transcription factors or other proteins. Transcription can also be regulated by subtle variations in DNA structure and by chemical changes in the bases to which transcription factors bind. As stated previously, the chemical properties of the four DNA bases differ slightly, providing each base with unique opportunities to chemically react with other molecules. One chemical modification of DNA, called **methylation**, involves the addition of a **methyl group (-CH3)**. Methylation frequently occurs at cytosine residues that are preceded by guanine bases, oftentimes in the vicinity of promoter sequences. The methylation status of DNA often correlates with its functional activity, where inactive genes tend to be more heavily methylated. This is because the methyl group serves to inhibit transcription by attracting a protein that binds specifically to methylated DNA, thereby interfering with polymerase binding. Methylation also plays an important role in **genomic imprinting**, which occurs when both maternal and paternal alleles are present but only one allele is expressed while the other remains inactive. Another way to think of genomic imprinting is as **"parent of origin differences"** in the expression of inherited traits. Considerable intrigue surrounds the effects of DNA methylation, and many researchers are working to unlock the mystery behind this concept.

Controlling Translation

Translation is the process whereby the genetic code carried by an mRNA directs the synthesis of proteins. **Translational regulation** occurs through the binding of specific molecules, called **repressor proteins**, to a sequence found on an RNA molecule. Repressor proteins prevent a gene from being expressed. As we have just discussed, the default state for a gene is that of being expressed via the recognition of its promoter by RNA polymerase. Close to the promoter region is another *cis*-acting site called the **operator**, the target for the repressor

protein. When the repressor protein binds to the operator, RNA polymerase is prevented from initiating transcription, and gene expression is turned off.

Translational control plays a significant role in the process of embryonic development and cell differentiation. Upon fertilization, an egg cell begins to multiply to produce a ball of cells that are all the same. At some point, however, these cells begin to **differentiate**, or change into specific cell types. Some will become blood cells or kidney cells, whereas others may become nerve or brain cells. When all of the cells formed are alike, the same genes are turned on. However, once differentiation begins, various genes in different cells must become active to meet the needs of that cell type. In some organisms, the egg houses store immature mRNAs that become translationally active only after fertilization. Fertilization then serves to trigger mechanisms that initiate the efficient translation of mRNA into proteins. Similar mechanisms serve to activate mRNAs at other stages of development and differentiation, such as when specific protein products are needed.

Mechanisms of Genetic Variation and Heredity

Does Everyone Have the Same Genes?

When you look at the human species, you see evidence of a process called **genetic variation**, that is, there are immediately recognizable differences in human traits, such as hair and eye color, skin pigment, and height. Then there are the not so obvious genetic variations, such as blood type. These expressed, or **phenotypic**, traits are attributable to **genotypic** variation in a person's DNA sequence. When two individuals display different phenotypes of the same trait, they are said to have two different **alleles** for the same gene. This means that the gene's sequence is slightly different in the two individuals, and the gene is said to be **polymorphic**, "**poly**" meaning many and "**morph**" meaning shape or form. Therefore, although people generally have the same genes, the genes do not have exactly the same DNA sequence. These polymorphic sites influence gene expression and also serve as markers for genomic research efforts.

Genetic Variation

The cell cycle is the process that a cell undergoes to replicate.

Most genetic variation occurs during the phases of the cell cycle when DNA is duplicated. Mutations in the new DNA strand can manifest as **base substitutions**, such as when a single base gets replaced with another; **deletions**, where one or more bases are left out; or **insertions**, where one or more bases are added. Mutations can

either be **synonymous**, in which the variation still results in a codon for the same amino acid or **non-synonymous**, in which the variation results in a codon for a different amino acid. Mutations can also cause a **frame shift**, which occurs when the variation bumps the reference point for reading the genetic code down a base or two and results in loss of part, or sometimes all, of that gene product. DNA mutations can also be introduced by toxic chemicals and, particularly in skin cells, exposure to ultraviolet radiation.

The manner in which a cell replicates differs with the various classes of life forms, as well as with the end purpose of the cell replication. Cells that compose tissues in multicellular organisms typically replicate by organized duplication and spatial separation of their cellular genetic material, a process called **mitosis**. **Meiosis** is the mode of cell replication for the formation of sperm and egg cells in plants, animals, and many other multicellular life forms. Meiosis differs significantly from mitosis in that the cellular progeny have their complement of genetic material reduced to half that of the parent cell.

Mutations that occur in **somatic cells**—any cell in the body except gametes and their precursors—will not be passed on to the next generation. This does not mean, however, that somatic cell mutations, sometimes called **acquired mutations**, are benign. For example, as your skin cells prepare to divide and produce new skin cells, errors may be inadvertently introduced when the DNA is duplicated, resulting in a daughter cell that contains the error. Although most defective cells die quickly, some can persist and may even become cancerous if the mutation affects the ability to regulate cell growth.

Mutations and the Next Generation

There are two places where mutations can be introduced and carried into the next generation. In the first stages of development, a sperm cell and egg cell fuse. They then begin to divide, giving rise to cells that differentiate into tissue-specific cell types. One early type of differentiated cell is the germ line cell, which may ultimately develop into mature gametes. If a mutation occurs in the developing germ line cell, it may persist until that individual reaches reproductive age. Now the mutation has the potential to be passed on to the next generation.

Mutations may also be introduced during meiosis, the mode of cell replication for the formation of sperm and egg cells. In this case, the germ line cell is healthy, and the mutation is introduced during the actual process of gamete replication. Once again, the sperm or egg will contain the mutation, and during the reproductive process, this mutation may then be passed on to the offspring.

One should bear in mind that not all mutations are bad. Mutations also provide a species with the opportunity to adapt to new environments, as well as to protect a species from new pathogens. Mutations are what lie behind the popular saying of "**survival of the fittest**", the basic theory of evolution proposed by Charles Darwin in 1859. This theory proposes that as new environments arise, individuals carrying certain mutations that enable an evolutionary advantage will survive to pass this mutation on to its offspring. It does not suggest that a mutation is derived from the environment, but that survival in that environment is enhanced by a particular mutation. Some genes, and even some organisms, have evolved to tolerate mutations better than others. For example, some viral genes are known to have high mutation rates. Mutations serve the virus well by enabling adaptive traits, such as changes in the outer protein coat so that it can escape detection and thereby destruction by the host's immune system. Viruses also produce certain enzymes that are necessary for infection of a host cell. A mutation within such an enzyme may result in a new form that still allows the virus to infect its host but that is no longer blocked by an anti-viral drug. This will allow the virus to propagate freely in its environment.

Mendel's Laws—How We Inherit Our Genes

In 1866, Gregor Mendel studied the transmission of seven different pea traits by carefully test-crossing many distinct varieties of peas. Studying garden peas might seem trivial to those of us who live in a modern world of cloned sheep and gene transfer, but Mendel's simple approach led to fundamental insights into genetic inheritance, known today as **Mendel's Laws**. Mendel did not actually know or understand the cellular mechanisms that produced the results he observed. Nonetheless, he correctly surmised the behavior of traits and the mathematical

predictions of their transmission, the independent segregation of alleles during gamete production, and the independent assortment of genes. Perhaps as amazing as Mendel's discoveries was the fact that his work was largely ignored by the scientific community for over 30 years!

Mendel's Principles of Genetic Inheritance
Law of Segregation: Each of the two inherited factors (alleles) possessed by the parent will segregate and pass into separate gametes (eggs or sperm) during meiosis, which will each carry only one of the factors. **Law of Independent Assortment:** In the gametes, alleles of one gene separate independently of those of another gene, and thus all possible combinations of alleles are equally probable. **Law of Dominance:** Each trait is determined by two factors (alleles), inherited one from each parent. These factors each exhibit a characteristic dominant, co-dominant, or recessive expression, and those that are dominant will mask the expression of those that are recessive.

How Does Inheritance Work?

Our discussion here is restricted to sexually reproducing organisms where each gene in an individual is represented by two copies, called **alleles**—one on each chromosome pair. There may be more than two alleles, or **variants**, for a given gene in a population, but only two alleles can be found in an individual. Therefore, the probability that a particular allele will be inherited is 50:50, that is, alleles randomly and independently segregate into daughter cells, although there are some exceptions to this rule.

The term **diploid** describes a state in which a cell has two sets of homologous chromosomes, or two chromosomes that are the same. The maturation of germ line stem cells into gametes requires the diploid number of each chromosome be reduced by half. Hence, gametes are said to be **haploid**—having only a single set of homologous chromosomes. This reduction is accomplished through a process called meiosis, where one chromosome in a diploid pair is sent to each daughter gamete. Human gametes, therefore, contain 23 chromosomes, half the number of somatic cells—all the other cells of the body.

Because the chromosome in one pair separates independently of all other chromosomes, each new gamete has the potential for a totally new combination of chromosomes. In humans, the independent segregation of the 23 chromosomes can lead to as many as 16 to 17 million different combinations in one individual's gametes. Only one of these gametes will combine with one of the nearly 17 million possible combinations from the other parent, generating a staggering potential for individual variation. Yet, this is just the beginning. Even more variation is possible when you

consider the recombination between sections of chromosomes during meiosis as well as the random mutation that can occur during DNA replication. With such a range of possibilities, it is amazing that siblings look so much alike!

Expression of Inherited Genes

Gene expression, as reflected in an organism's phenotype, is based on conditions specific for each copy of a gene. As we just discussed, for every human gene there are two copies, and for every gene there can be several variants or alleles. If both alleles are the same, the gene is said to be **homozygous**. If the alleles are different, they are said to be **heterozygous**. For some alleles, their influence on phenotype takes precedence over all other alleles. For others, expression depends on whether the gene appears in the homozygous or heterozygous state. Still other phenotypic traits are a combination of several alleles from several different genes. Determining the allelic condition used to be accomplished solely through the analysis of pedigrees, much the way Mendel carried out his experiments on peas. However, this method can leave many questions unanswered, particularly for traits that are a result of the interaction between several different genes. Today, molecular genetic techniques exist that can assist researchers in tracking the transmission of traits by pinpointing the location of individual genes, identifying allelic variants, and identifying those traits that are caused by multiple genes.

The Nature of Alleles

A **dominant allele** is an allele that is almost always expressed, even if only one copy is present. Dominant alleles express their phenotype even when paired with a different allele, that is, when heterozygous. In this case, the phenotype appears the same in both the heterozygous and homozygous states. Just how the dominant allele overshadows the other allele depends on the gene, but in some cases the dominant gene produces a gene product that the other allele does not. Well-known dominant alleles occur in the human genes for Huntington disease, a form of dwarfism called achondroplasia, and polydactylism (extra fingers and toes).

On the other hand, a **recessive allele** will be expressed only if there are two identical copies of that allele, or for a male, if one copy is present on the X chromosome. The phenotype of a recessive allele is only seen when both alleles are the same. When an individual has one dominant allele and one recessive allele, the trait is not expressed because it is overshadowed by the dominant allele. The individual is said to be a carrier for that trait. Examples of recessive disorders in humans include sickle cell anemia, Tay-Sachs disease, and phenylketonuria (PKU).

A particularly important category of genetic linkage has to do with the **X and Y sex chromosomes**. These chromosomes not only carry the genes that determine male and female traits, but also those for some other characteristics as well. Genes that are carried by either sex chromosome are said to be **sex linked**. Men normally have an X and a Y combination of sex chromosomes, whereas women have two X's. Because only men inherit Y chromosomes, they are the only ones to inherit **Y-linked traits**. Both men and women can have **X-linked traits** because both inherit X chromosomes.

X-linked traits not related to feminine body characteristics are primarily expressed in the phenotype of men. This is because men have only one X chromosome. Subsequently, genes on that chromosome that do not code for gender are expressed in the male phenotype, even if they are recessive. In women, a recessive allele on one X chromosome is often masked in their phenotype by a dominant normal allele on the other. This explains why women are frequently carriers of X-linked traits but more rarely have them expressed in their own phenotypes. In humans, at least 320 genes are X-linked. These include the genes for **hemophilia**, **red–green color blindness**, and **congenital night blindness**. There are at least a dozen Y-linked genes, in addition to those that code for masculine physical traits.

It is now known that one of the X chromosomes in the cells of human females is completely, or mostly, inactivated early in embryonic life. This is a normal self-preservation action to prevent a potentially harmful double dose of genes. Recent research points to the **"Xist" gene** on the X chromosome as being responsible for a sequence of events that silences one of the X chromosomes in women. The inactivated X chromosomes become highly compacted structures known as **Barr bodies**. The presence of Barr bodies has been used at international sport competitions as a test to determine whether an athlete is a male or a female.

Exceptions to Mendel's Laws

There are many examples of inheritance that appear to be exceptions to Mendel's laws. Usually, they turn out to represent complex interactions among various allelic conditions. For example, **co-dominant alleles** both contribute to a phenotype. Neither is dominant over the other. Control of the human blood group system provides a good example of co-dominant alleles.

The Four Basic Blood Types

There are four basic blood types, and they are O, A, B, and AB. We know that our blood type is determined by the "alleles" that we inherit from our parents. For the blood type gene, there are three basic blood type alleles: A, B, and O. We all have two alleles, one inherited from each parent. The possible combinations of the three alleles are OO, AO, BO, AB, AA, and BB. **Blood types A and B are "co-dominant" alleles**, whereas **O is "recessive"**. A codominant allele is apparent even if only one is present; a recessive allele is apparent only if two recessive alleles are present. Because blood type O is recessive, it is not apparent if the person inherits an A or B allele along with it. So, the possible allele combinations result in a particular blood type in this way:

OO = blood type O
AO = blood type A
BO = blood type B
AB = blood type AB
AA = blood type A
BB = blood type B

You can see that a person with blood type B may have a B and an O allele, or they may have two B alleles. If both parents are blood type B and both have a B and a recessive O, then their children will either be BB, BO, or OO. If the child is BB or BO, they have blood type B. If the child is OO, he or she will have blood type O.

Pleiotropism, or pleotrophy, refers to the phenomenon in which a single gene is responsible for producing multiple, distinct, and apparently unrelated phenotypic traits, that is, an individual can exhibit many different phenotypic outcomes. This is because the gene product is active in many places in the body. An example is **Marfan's syndrome**, where there is a defect in the gene coding for a connective tissue protein. Individuals with Marfan's syndrome exhibit abnormalities in their eyes, skeletal system, and cardiovascular system.

Some genes mask the expression of other genes just as a fully dominant allele masks the expression of its recessive counterpart. A gene that masks the phenotypic effect of another gene is called an **epistatic gene**; the gene it subordinates is the **hypostatic gene**. The gene for **albinism** in humans is an epistatic gene. It is not part of the interacting skin-color genes. Rather, its dominant allele is necessary for the development of any skin pigment, and its recessive homozygous state results in the albino condition, regardless of how many other pigment genes may be present. Because of the effects of an epistatic gene, some individuals who inherit the dominant, disease-causing gene show only partial symptoms of the disease. Some, in fact, may show no expression of the disease-causing gene, a condition referred to as

nonpenetrance. The individual in whom such a nonpenetrant mutant gene exists will be phenotypically normal but still capable of passing the deleterious gene on to offspring, who may exhibit the full-blown disease.

Then we have traits that are **multigenic**, that is, they result from the expression of several different genes. This is true for human eye color, in which at least three different genes are responsible for determining eye color. A brown/blue gene and a central brown gene are both found on chromosome 15, whereas a green/blue gene is found on chromosome 19. The interaction between these genes is not well understood. It is speculated that there may be other genes that control other factors, such as the amount of pigment deposited in the iris. This multigenic system explains why two blue-eyed individuals can have a brown-eyed child.

Speaking of eye color, have you ever seen someone with one green eye and one brown eye? In this case, **somatic mosaicism** may be the culprit. This is probably easier to describe than explain. In multicellular organisms, every cell in the adult is ultimately derived from the single-cell fertilized egg. Therefore, every cell in the adult normally carries the same genetic information. However, what would happen if a mutation occurred in only one cell at the two-cell stage of development? Then the adult would be composed of two types of cells: cells with the mutation and cells without. If a mutation affecting melanin production occurred in one of the cells in the cell lineage of one eye but not the other, then the eyes would have different genetic potential for melanin synthesis. This could produce eyes of two different colors.

Penetrance refers to the degree to which a particular allele is expressed in a population phenotype. If every individual carrying a dominant mutant gene demonstrates the mutant phenotype, the gene is said to show complete penetrance.

Molecular Genetics: The Study of Heredity, Genes, and DNA

As we have just learned, DNA provides a blueprint that directs all cellular activities and specifies the developmental plan of multicellular organisms. Therefore, an understanding of DNA, gene structure, and function is fundamental for an appreciation of the molecular biology of the cell. Yet, it is important to recognize that progress in any scientific field depends on the availability of experimental tools that allow researchers to make new scientific observations and conduct novel experiments. The last section of the genetic primer concludes with a discussion of some of the laboratory tools and technologies that allow researchers to study cells and their DNA.

Back to top