Chapter 1. The Human Genome

Learning outcomes

When you have read Chapter 1, you should be able to:

1. Describe the two experiments that led molecular biologists to conclude that genes are made of DNA, and state the limitations of each experiment
2. Give a detailed description of the structure of a polynucleotide and summarize the chemical differences between DNA and RNA
3. Discuss the evidence that Watson and Crick used to deduce the double helix structure of DNA and list the key features of this structure
4. Explain why the DNA double helix has structural flexibility
5. Describe in outline the content of the human nuclear genome
6. Draw the structure of an ‘average' human gene
7. Categorize the human gene catalog into different functional classes
8. Distinguish between conventional and processed pseudogenes and other types of evolutionary relic
9. Give specific examples of the repetitive DNA content of the human genome
10. Give an outline description of the structure and organization of the human mitochondrial genome
11. Discuss the importance of the Human Genome Project

Life as we know it is specified by the genomes of the myriad organisms with which we share the planet. Every organism possesses a genome that contains the biological information needed to construct and maintain a living example of that organism. Most genomes, including the human genome and those of all other cellular life forms, are made of DNA (deoxyribonucleic acid) but a few viruses have RNA (ribonucleic acid) genomes. DNA and RNA are polymeric molecules made up of chains of monomeric subunits called nucleotides.

The human genome, which is typical of the genomes of all multicellular animals, consists of two distinct parts (*Figure 1.1*):

- The nuclear genome comprises approximately 3 200 000 000 nucleotides of DNA, divided into 24 linear molecules, the shortest 50 000 000 nucleotides in length and the longest 260 000 000 nucleotides, each contained in a different chromosome. These 24 chromosomes consist of 22 autosomes and the two sex chromosomes, X and Y.
- The mitochondrial genome is a circular DNA molecule of 16 569 nucleotides, multiple copies of which are located in the energy-generating organelles called mitochondria.

Each of the approximately 10^{13} cells in the adult human body has its own copy or copies of the genome, the only exceptions being those few cell types, such as red blood cells, that lack a nucleus in their fully differentiated state. The vast majority of cells are diploid and so have two copies of each autosome, plus two sex chromosomes, XX for females or XY for males - 46 chromosomes in all. These are called somatic cells, in contrast to sex cells or gametes, which are haploid and have just 23 chromosomes, comprising one of each autosome and one sex chromosome. Both types of cell have about 8000 copies of the mitochondrial genome, 10 or so in each mitochondrion.

This book is about genomes. It explains what genomes are (Part 1), how they are studied (Part 2), how they function (Part 3), and how they replicate and evolve (Part 4). We begin our journey with our own genome, which is quite naturally the one that interests us the most. Later in this chapter we will examine how the human genome is constructed, some of this information dating from the old days when biologists studied genes rather than genomes, but much of it revealed only since the Human Genome Project was completed in the first year of the new millennium. First, however, we must understand the structure of DNA.

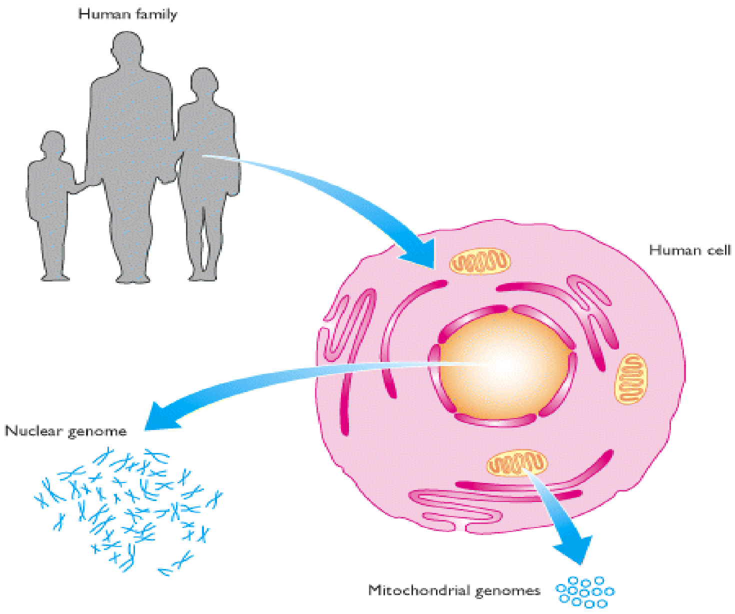


Figure 1.1. The nuclear and mitochondrial components of the human genome. For more details on the anatomy of the human genome, see Section 1.2

1.1. DNA

DNA was discovered in 1869 by Johann Friedrich Miescher, a Swiss biochemist working in Tübingen, Germany. The first extracts that Miescher made from human white blood cells were crude mixtures of DNA and chromosomal proteins, but the following year he moved to Basel, Switzerland (where the research institute named after him is now located) and prepared a pure sample of nucleic acid from salmon sperm. Miescher's chemical tests showed that DNA is acidic and rich in phosphorus, and also suggested that the individual molecules are very large, although it was not until the 1930s when biophysical techniques were applied to DNA that the huge lengths of the polymeric chains were fully appreciated.

Three years before Miescher discovered DNA, Gregor Mendel had published the results of his breeding experiments with pea plants, carried out in the monastery gardens at Brno, a central european city some 550 km from Tübingen in what is now the Czech Republic. Mendel's paper in the *Proceedings of the Society of Natural Sciences in Brno* describes his hypothesis that inheritance is controlled by unit factors, the entities that geneticists today call genes. It is very unlikely that Miescher and Mendel were aware of each other's work, and if either of them had happened to read about the other's discoveries then they certainly would not have made any connection between DNA and genes. To make such a connection - to infer that genes are made of DNA - would have been quite illogical in the late 19th century or indeed for many decades afterwards. The precise biological function of DNA was not known, and the supposition that it was a store of cellular phosphorus seemed entirely reasonable at the time. The chemical nature of genes was equally unknown, and indeed was an irrelevance for most geneticists, who in the years immediately after 1900, when Mendel's work was rediscovered, were able to make remarkable advances in understanding heredity without worrying about what genes were actually made of.

It was not until the 1930s that scientists began to ask more searching questions about genes. In 1944, Erwin Schrödinger, more famous for the wave equation which still terrifies many biology students taking introductory courses in physical chemistry, published a book entitled *What is Life?*, which encapsulated a variety of issues that were being discussed not only by geneticists but also by physicists such as Niels Bohr and Max Delbrück. These scientists were the first molecular biologists and the first to suggest that ‘life' could be explained in molecular terms; our current knowledge of how the genome functions stems directly from their pioneering work. The starting point for the new molecular biology was to discover what genes are made of.

1.1.1. Genes are made of DNA

How could the molecular nature of the genetic material be determined? Back in 1903, WS Sutton had realized that the inheritance patterns of genes paralleled the behavior of chromosomes during cell division. This observation led to the proposal that genes are located in chromosomes and by the 1930s it was universally accepted that the chromosome theory was correct. Examination of cells by cytochemistry, after staining with dyes that bind specifically to just one type of biochemical, had shown that chromosomes are made of DNA and protein, in roughly equal amounts. Some biologists looked on the combination between the two (‘nucleoprotein') as the genetic material, but others argued differently. From today's perspective it can be difficult to understand why these arguments favored the notion that genes were made, not of DNA, but of protein. The explanation is that, at the time, many biochemists thought that all DNA molecules were the same, which meant that DNA did not have the immense variability that was one of the postulated features of the genetic material. Billions of different genes must exist and for each one to have its own individual activity, the genetic material must be able to take many different forms. If every DNA molecule were identical then DNA could not satisfy this requirement and so genes must be made of protein. This assumption made perfect sense because proteins were known, correctly, to be highly variable polymeric molecules, each one made up of a different combination of 20 chemically distinct amino-acid monomers (Section 3.3.1).

The errors that had been made in understanding DNA structure lingered on until the late 1930s. Gradually, however, it was accepted that DNA, like protein, has immense variability. Could DNA therefore be the genetic material? The results of two experiments performed during the middle decades of the 20th century forced biologists to take this possibility seriously.

Bacterial genes are made of DNA

The first molecular biologists realized that the most conclusive way to identify the chemical composition of genes would be to purify some and subject them to chemical analysis. But nothing like this had ever been attempted and it was not clear how it could be done. Ironically, the experiment was performed almost unwittingly by a group of scientists who did not look upon themselves as molecular biologists and who were not motivated by a curiosity to know what genes are made of. Instead, their objective was to find a better treatment for one of the most deadly diseases of the early 20th century, pneumonia.

Before the discovery of antibiotics, pneumonia was mainly controlled by treating patients in the early stages of the disease with an antiserum prepared by injecting dead cells of the causative bacterium (now called *Streptococcus pneumoniae*) into an animal. In order to prepare more effective antisera, studies were made of the immunological properties of the bacterium. It was shown that there is a range of different types of *S. pneumoniae*, each characterized by the mixture of sugars contained in the thick capsule that surrounds the cell and elicits the immunological response (*Figure 1.2A*). In 1923, Frederick Griffith, a British medical officer, discovered that as well as the virulent strains, there were some types of *S. pneumoniae* that did not have a capsule and did not cause pneumonia. This discovery was not a huge surprise because other species of pathogenic bacteria were known to have avirulent, unencapsulated forms. However, 5 years later Griffith obtained some results that were totally unexpected (Griffith, 1928). He performed a series of experiments in which he injected mice with various mixtures of bacteria (*Figure 1.2B*). He showed that, as anticipated, mice injected with virulent *S. pneumoniae* bacteria developed pneumonia and died, whereas those injected with an avirulent type remained healthy. What he did not anticipate, however, was what would happen when mice were injected with a mixture made up of live avirulent bacteria along with some virulent cells that had been killed by heat treatment. The only live bacteria were the harmless ones, so surely the mice would remain healthy? Not so: the mice died. Griffith carried out biopsies of these dead mice and discovered that their respiratory tracts contained virulent bacteria, which were always of the same immunological type as the strain that had been killed by heat treatment before injection. Somehow the living harmless bacteria had acquired the ability to make the capsule sugars of the dead bacteria. This process - the conversion of living harmless bacteria into virulent cells - was called transformation. Although not recognized as such by Griffith, the transforming principle - the component of the dead cells that conferred on the live cells the ability to make the capsular sugars - was genetic material.

Oswald Avery, together with his colleagues Colin MacLeod and Maclyn McCarty, of Columbia University, New York, set out to determine what the transforming principle was made of. The experiments took a long time to carry out and were not completed until 1944 (Avery *et al.,* 1944). But the results were conclusive: the transforming principle was DNA. The transforming principle behaved in exactly the same way as DNA when subjected to various biophysical tests, it was inactivated by enzymes that degraded DNA, and it was not affected by enzymes that attacked protein or any other type of biochemical (*Figure 1.3*).

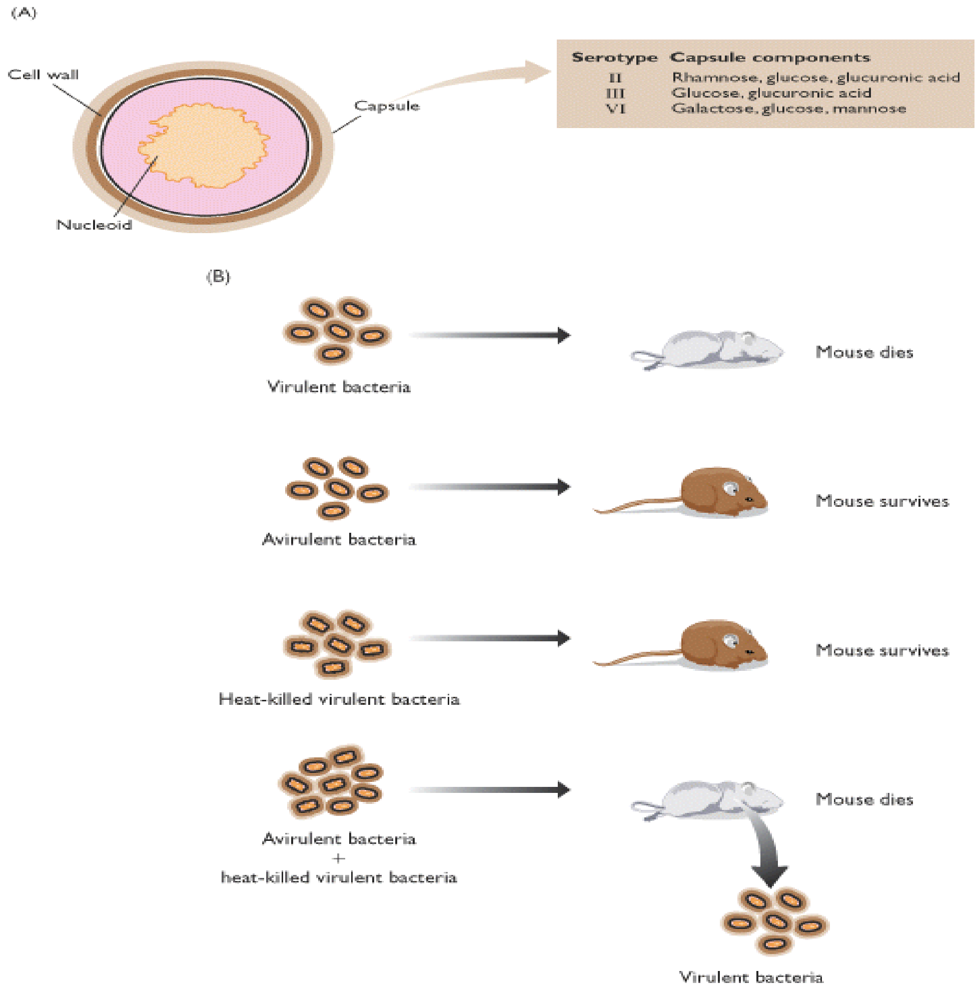

(A)
Cell wall
Capsule
Nucleoid
Serotype Capsule components
II Rhamnose, glucose, glucuronic acid
III Glucose, glucuronic acid
VI Galactose, glucose, mannose
(B)
Virulent bacteria
Mouse dies
Avirulent bacteria
Mouse survives
Heat-killed virulent bacteria
Mouse survives
Avirulent bacteria
+
heat-killed virulent bacteria
Mouse dies
Virulent bacteria

Figure 1.2. Griffith's experiments with virulent and avirulent *Streptococcus pneumoniae* bacteria. (A) Representation of a *S. pneumoniae* bacterium. A serotype is a bacterial type with distinctive immunological properties, conferred in this case by the combination of sugars present in the capsule. Avirulent types have no capsule. (B) The experiments which showed that a component of heat-killed bacteria can transform living avirulent bacteria into virulent cells. Griffith showed that the avirulent bacteria were always transformed into the same serotype as the dead cells. In other words, the living bacteria acquired the genes specifying synthesis of the capsule of the dead cells.

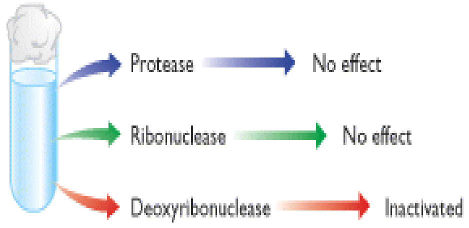


Figure 1.3. The transforming principle is DNA. Avery and his colleagues showed that the transforming principle is unaffected by treatment with a protease or a ribonuclease, but is inactivated by treatment with a deoxyribonuclease.

Avery's experiments were meticulous but, because of several complicating factors, they did not immediately lead to acceptance of DNA as the genetic material. It was not clear in the minds of all microbiologists that transformation really was a genetic phenomenon, and few geneticists really understood the system well enough to be able to evaluate Avery's work. There was also some doubt about the veracity of the experiments. In particular, there were worries about the specificity of the deoxyribonuclease enzyme that he used to inactivate the transforming principle. This result, a central part of the evidence for the transforming principle being DNA, would be invalid if, as seemed possible, the enzyme contained trace amounts of a contaminating protease and hence was also able to degrade protein. These uncertainties meant that a second experiment was needed to provide more information on the chemical nature of the genetic material.

Virus genes are made of DNA

The second experiment was carried out by two *bona fide* molecular biologists, Alfred Hershey and Martha Chase, at Cold Spring Harbor, New York, in 1952 (Hershey and Chase, 1952). Like the work on the transforming principle, the Hershey-Chase experiment was not done specifically to determine the chemical nature of the gene. Hershey and Chase were two of several biologists who were studying the infection cycle of bacteriophages (or ‘phages') - viruses that infect bacteria. Phages are relatively simple structures made of just DNA and protein, with the DNA contained inside the phage, surrounded by a protein capsid (*Figure 1.4A*). To replicate, a phage must enter a bacterial cell and subvert the bacterial enzymes into expressing the information contained in the phage genes, so that the bacterium synthesizes new phages. Once replication is complete, the new phages leave the bacterium, possibly causing its death as they do so, and move on to infect new cells (*Figure 1.4B*). The objective of Hershey and Chase's experiment was to determine if the entire phage particle entered the bacterium at the start of the infection cycle, or if part of the phage stayed outside. If only one component of the phage - the DNA or the protein - entered the cell then that component must be the genetic material.

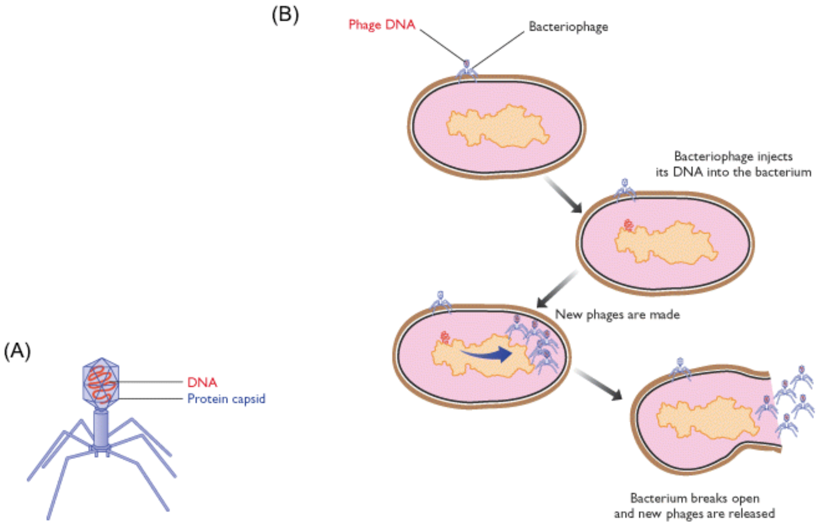

(B)
Phage DNA
Bacteriophage
Bacteriophage injects its DNA into the bacterium
New phages are made
Bacterium breaks open and new phages are released
(A)
DNA
Protein capsid

Figure 1.4. Bacteriophages are viruses that infect bacteria. (A) The structure of a head-and-tail bacteriophage such as T2. The DNA genome of the phage is contained in the head part of the protein capsid. (B) The infection cycle. After injection into an *Escherichia coli* bacterium, the T2 phage genome directs synthesis of new phages. For T2, the infection cycle takes about 20 minutes at 37 °C and ends with lysis of the cell and release of 250–300 new phages. This is the lytic infection cycle. Some phages, such as λ, can also follow a lysogenic infection cycle, in which the phage genome becomes inserted into the bacterial chromosome and remains there, in quiescent form, for several generations of the bacterium (Section 4.2.1)

Their experimental strategy was based on the use of radioactive labels, which had recently been introduced into biology. DNA contains phosphorus, which is absent from protein, so DNA can be labeled specifically with the radioactive phosphorus isotope, ^{32}P. Protein, on the other hand, contains sulfur, which is absent from DNA, and so protein can be labeled with ^{35}S. Hershey and Chase were not the first to use radiolabeling to try to determine which part of the phage entered the cell, but previous experiments by James Watson, Ole Maaløe and others had been inconclusive because of the difficulty in distinguishing between phage material that was actually inside a bacterium and a component that did not enter the cell but remained attached to the outer cell surface. To get round this difficulty, Hershey and Chase made an important modification to the previous experiments. They infected *Escherichia coli* bacteria with radiolabeled T2 phages but, rather than allowing the infection process to go to completion, they left the culture for just a few minutes and then agitated it in a blender. The idea was that the blending would detach the phage material from the surface of the bacteria, enabling this component to be separated from the material inside the cells by centrifuging at a speed that collected the relatively heavy bacteria as a pellet at the bottom of the tube but left the detached material in suspension (*Figure 1.5*). After centrifugation, Hershey and Chase examined the bacterial pellet and found that it contained 70% of the ^{32}P-labeled component of the phages (the DNA) but only 20% of the ^{35}S-labeled material (the phage protein). In a parallel experiment, the bacteria were left for 20 minutes, long enough for the infection cycle to reach completion. With T2 phage, the cycle ends with the bacteria bursting open and releasing new phages into the supernatant. These new phages contained almost half the DNA from the original phages, but less than 1% of the protein.

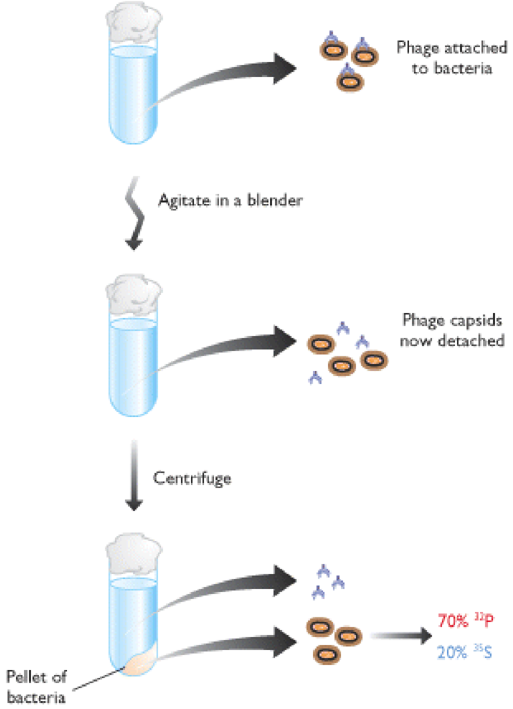


Figure 1.5. The Hershey-Chase experiment. The bacteriophages were labeled with ^{32}P and ^{35}S. A few minutes after infection, the culture was agitated to detach the empty phage capsids from the cell surface. The culture was then centrifuged and the radioactive content of the bacterial pellet determined. This pellet contained most of ^{32}P-labeled component of the phages (the DNA) but only 20% of the ^{35}S-labeled material (the phage protein). In a second experiment, Hershey and Chase showed that new phages produced at the end of an interrupted infection cycle contained less than 1% of the protein from the parent phages.

Hershey and Chase's results suggested that DNA was the major component of the infecting phages that entered the bacterial cell and, similarly, was the major, or perhaps only, component to be passed on to the progeny phages. These observations lent support to the view that DNA is the genetic material, but were they conclusive? Not according to Hershey and Chase (1952) who wrote 'Our experiments show clearly that a physical separation of phage T2 into genetic and non-genetic parts is possible … . The chemical identification of the genetic part must wait, however, until some questions … have been answered.' Even if the experiment had provided compelling evidence that the genetic material of phages was DNA, it would have been erroneous to extrapolate from these unusual life forms (which some biologists contend are not really 'living') to cellular organisms. Indeed, we know that some phage genomes are made of RNA. The Hershey-Chase experiment is important, not because of what it tells us, but because it alerted biologists to the fact that DNA *might* be the genetic material and was therefore worth studying. It was this that influenced Watson and Crick to study DNA and, as we will see below, it was their discovery of the double helix structure, which solved the puzzling question of how genes can replicate, that really convinced the scientific world that genes are made of DNA.

1.1.2. The structure of DNA

The names of James Watson and Francis Crick are so closely linked with DNA that it is easy to forget that, when they began their collaboration in Cambridge, England in October 1951, the detailed structure of the DNA polymer was already known. Their contribution was not to determine the structure of DNA *per se*, but to show that in living cells two DNA chains are intertwined to form the double helix. We will consider the two facets of DNA structure separately.

Nucleotides and polynucleotides

DNA is a linear, unbranched polymer in which the monomeric subunits are four chemically distinct nucleotides that can be linked together in any order in chains hundreds, thousands or even millions of units in length. Each nucleotide in a DNA polymer is made up of three components (*Figure 1.6*):

1. **2′-deoxyribose**, which is a pentose, a type of sugar composed of five carbon atoms. These five carbons are numbered 1′ (spoken as 'one-prime'), 2′, etc. The name '2′-deoxyribose' indicates that this particular sugar is a derivative of ribose, one in which the hydroxyl (-OH) group attached to the 2′-carbon of ribose has been replaced by a hydrogen (-H) group.

2. A nitrogenous base, one of cytosine, thymine (single-ring pyrimidines), adenine or guanine (double-ring purines). The base is attached to the 1′-carbon of the sugar by a **β-*N*-glycosidic bond** attached to nitrogen number 1 of the pyrimidine or number 9 of the purine.
3. A **phosphate group**, comprising one, two or three linked phosphate units attached to the 5′-carbon of the sugar. The phosphates are designated α, β and γ, with the α-phosphate being the one directly attached to the sugar.

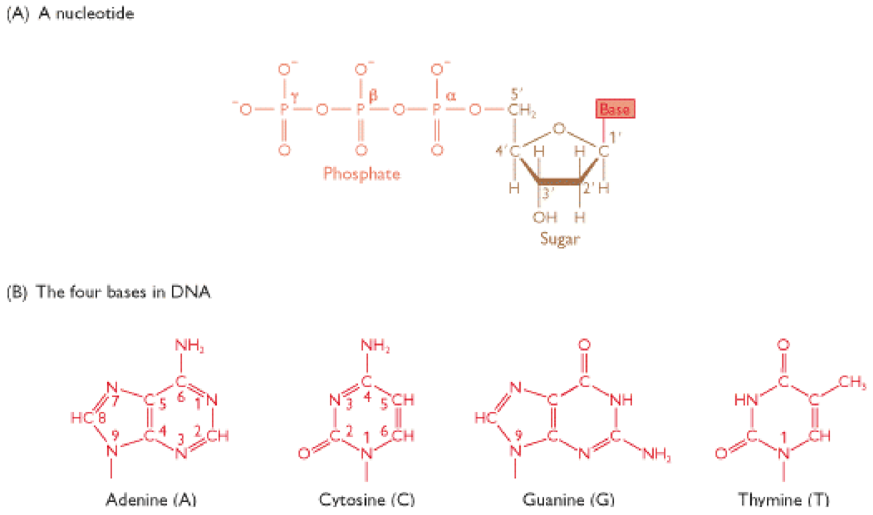


Figure 1.6. The structure of a nucleotide. (A) The general structure of a deoxyribonucleotide, the type of nucleotide found in DNA. (B) The four bases that occur in deoxyribonucleotides.

A molecule made up of just the sugar and base is called a nucleoside; addition of the phosphates converts this to a nucleotide. Although cells contain nucleotides with one, two or three phosphate groups, only the nucleoside triphosphates act as substrates for DNA synthesis. The full chemical names of the four nucleotides that polymerize to make DNA are:

2′-deoxyadenosine 5′-triphosphate
2′-deoxycytidine 5′-triphosphate
2′-deoxyguanosine 5′-triphosphate
2′-deoxythymidine 5′-triphosphate

The abbreviations of these four nucleotides are dATP, dCTP, dGTP and dTTP, respectively, or, when referring to a DNA sequence, A, C, G and T, respectively.

In a polynucleotide, individual nucleotides are linked together by phosphodiester bonds between their 5′- and 3′-carbons (*Figure 1.7*). From the structure of this linkage we can see that the polymerization reaction (*Figure 1.8*) involves removal of the two outer phosphates (the β- and γ-phosphates) from one nucleotide and replacement of the hydroxyl group attached to the 3′-carbon of the second nucleotide. Note that the two ends of the polynucleotide are chemically distinct, one having an unreacted triphosphate group attached to the 5′-carbon (the **5′** or **5′-P terminus**) and the other having an unreacted hydroxyl attached to the 3′-carbon (the **3′** or **3′-OH terminus**). This means that the polynucleotide has a chemical direction, expressed as either 5′→3′ (down in *Figure 1.8*) or 3′→5′ (up in *Figure 1.8*). An important consequence of the polarity of the phosphodiester bond is that the chemical reaction needed to extend a DNA polymer in the 5′→3′ direction is different to that needed to make a 3′→5′ extension. All natural DNA polymerase enzymes are only able to carry out 5′→3′ synthesis, which adds significant complications to the process by which double-stranded DNA is replicated (Section 13.2). The same limitation applies to RNA polymerases, the enzymes which make RNA copies of DNA molecules (Section 3.2.2).

RNA

Although our attention is firmly on DNA, the structure of RNA is so similar to that of DNA that it makes sense to introduce it here. RNA is also a polynucleotide but with two differences compared with DNA (*Figure 1.9*). First, the sugar in an RNA nucleotide is ribose and, second, RNA contains uracil instead of thymine. The four nucleotide substrates for synthesis of RNA are therefore:

adenosine 5′-triphosphate
cytidine 5′-triphosphate
guanosine 5′-triphosphate
uridine 5′-triphosphate

which are abbreviated to ATP, CTP, GTP and UTP, or A, C, G and U, respectively.

As with DNA, RNA polynucleotides contain 3′–5′ phosphodiester bonds, but these phosphodiester bonds are less stable than those in a DNA polynucleotide because of the indirect effect of the hydroxyl group at the 2′-position of the sugar. This may be one reason why the biological functions of RNA do not require the polynucleotide to be more than a few thousand nucleotides in length, at most. There are no RNA counterparts of the million-unit sized DNA molecules found in human chromosomes.

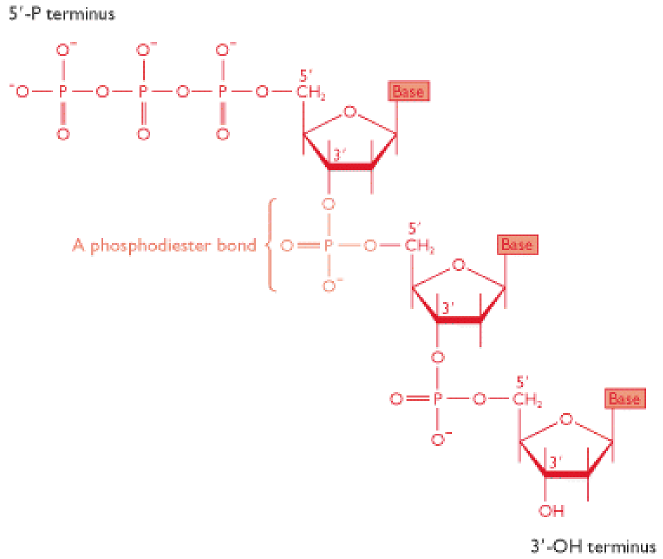


Figure 1.7. A short DNA polynucleotide showing the structure of the phosphodiester bond. Note that the two ends of the polynucleotide are chemically distinct.

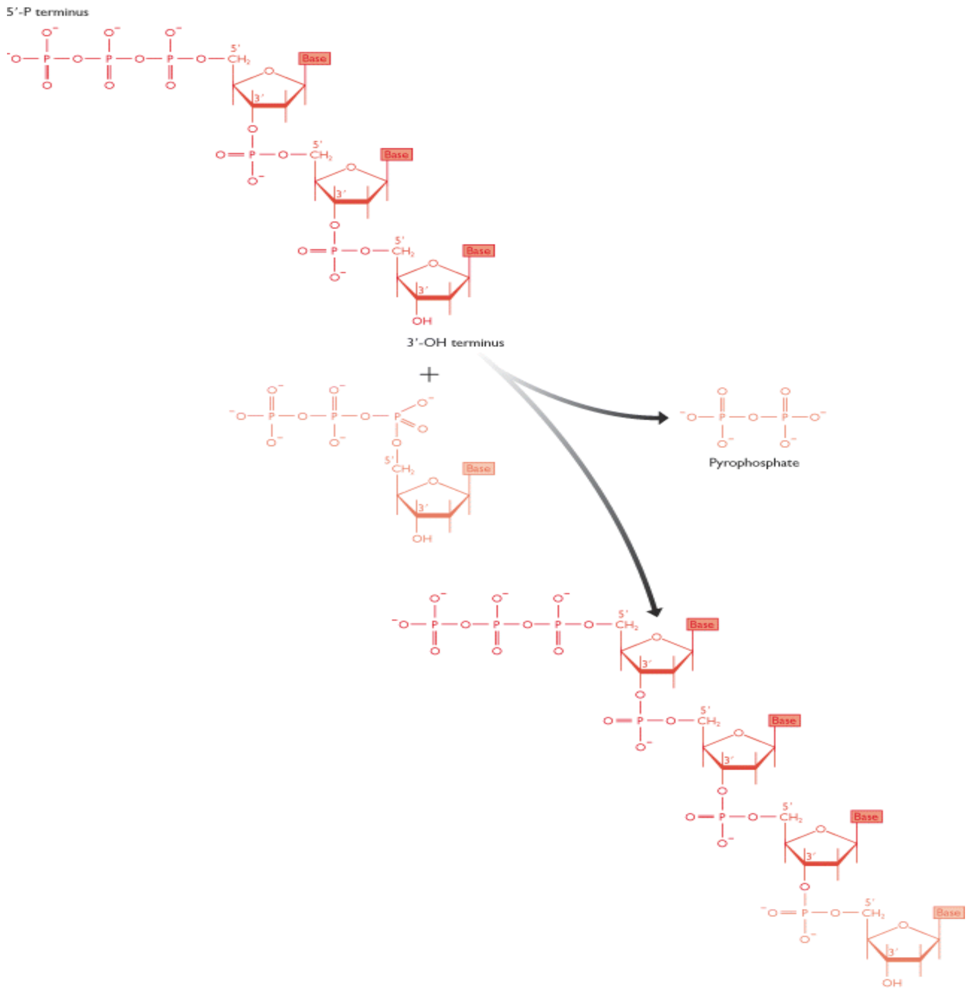


Figure 1.8. The polymerization reaction that results in synthesis of a DNA polynucleotide. Synthesis occurs in the 5′→3′ direction, with the new nucleotide being added to the 3′-carbon at the end of the existing polynucleotide. The β- and γ-phosphates of the nucleotide are removed as a pyrophosphate molecule.

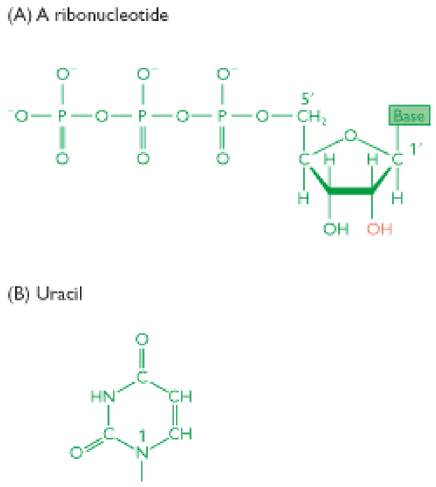


Figure 1.9. The chemical differences between DNA and RNA. (A) RNA contains ribonucleotides, in which the sugar is ribose rather than 2′-deoxyribose. The difference is that a hydroxyl group rather than hydrogen atom is attached to the 2′-carbon. (B) RNA contains the pyrimidine called uracil instead of thymine.

1.1.3. The double helix

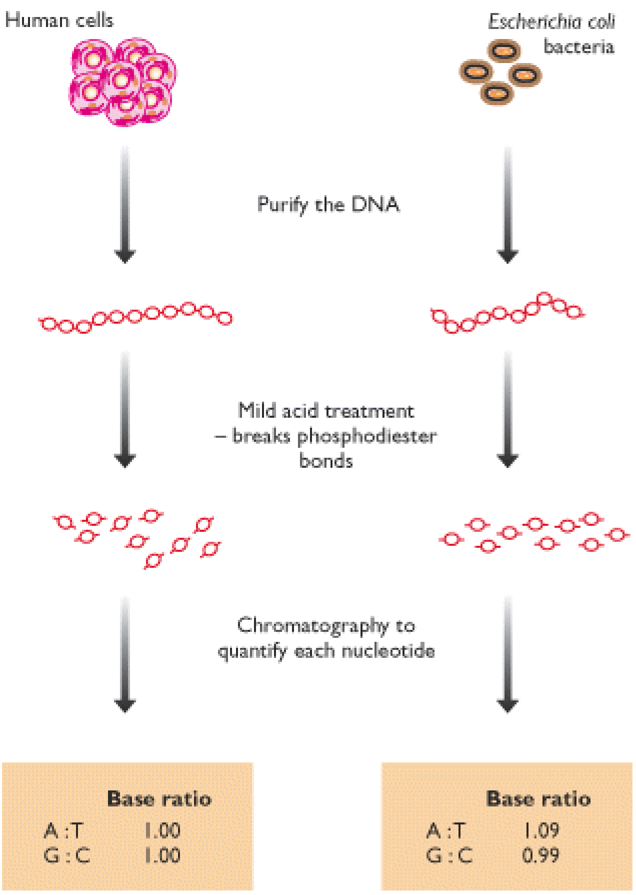
In the years before 1950, various lines of evidence had shown that cellular DNA molecules are comprised of two or more polynucleotides assembled together in some way. The possibility that unraveling the nature of this assembly might provide insights into how genes work prompted Watson and Crick, among others, to try to solve the structure. According to Watson in his book *The Double Helix* (see Further Reading), their work was a desperate race against the famous American biochemist, Linus Pauling (Section 3.3.1), who initially proposed an incorrect triple helix model, giving Watson and Crick the time they needed to complete the double helix structure (Watson and Crick, 1953). It is now difficult to separate fact from fiction, especially regarding the part played by Rosalind Franklin, whose X-ray diffraction studies provided the bulk of the experimental data in support of the double helix and who was herself very close to solving the structure. The one thing that is clear is that the double helix, discovered by Watson and Crick on Saturday 7 March 1953, was the single most important breakthrough in biology during the 20th century.

The evidence that led to the double helix

Watson and Crick used four types of information to deduce the double helix structure:

- ***Biophysical data*** of various kinds. The water content of DNA fibers was particularly important because it enabled the density of the DNA in a fiber to be estimated. The number of strands in the helix and the spacing between the nucleotides had to be compatible with the fiber density. Pauling's triple helix model was based on an incorrect density measurement which suggested that the DNA molecule was more closely packed than it actually is.

- X-ray diffraction patterns (Section 9.1.3), most of which were produced by Rosalind Franklin of Kings College, London, and which revealed the helical nature of the structure and indicated some of the key dimensions within the helix.
- ***The base ratios***, which had been discovered by Erwin Chargaff of Columbia University, New York. Chargaff carried out a lengthy series of chromatographic studies of DNA samples from various sources and showed that, although the values are different in different organisms, the amount of adenine is always the same as the amount of thymine, and the amount of guanine equals the amount of cytosine (*Figure 1.10*). These base ratios led to the base-pairing rules, which were the key to the discovery of the double helix structure.
- ***Model building***, which was the only major technique that Watson and Crick made use of themselves. Scale models of possible DNA structures enabled the relative positioning of the various atoms to be checked, to ensure that pairs of groups that formed bonds were not too far apart, and that other groups were not so close together as to interfere with one another.


Human cells
Escherichia coli bacteria
Purify the DNA
Mild acid treatment – breaks phosphodiester bonds
Chromatography to quantify each nucleotide

Base ratio	
A : T	1.00
G : C	1.00

Base ratio	
A : T	1.09
G : C	0.99

Figure 1.10. The base ratio experiments performed by Chargaff. DNA was extracted from various organisms and treated with acid to hydrolyze the phosphodiester bonds and release the individual nucleotides. Each nucleotide was then quantified by chromatography. The data show some of the actual results obtained by Chargaff. These indicate that, within experimental error, the amount of adenine is the same as that of thymine, and the amount of guanine is the same as that of cytosine.

The key features of the double helix

The double helix is right-handed, which means that if it were a spiral staircase and you were climbing upwards then the rail on the outside of the staircase would be on your right-hand side. The two strands run in opposite directions (*Figure 1.11A*). The helix is stabilized by two types of chemical interaction:

- Base-pairing between the two strands involves the formation of hydrogen bonds between an adenine on one strand and a thymine on the other strand, or between a cytosine and a guanine (*Figure 1.11B*). Hydrogen bonds are weak electrostatic attractions between an electronegative atom (such as oxygen or nitrogen) and a hydrogen atom attached to a second electronegative atom. Hydrogen bonds are longer than covalent bonds and are much weaker, typical bond energies being 1–10 kcal mol^{-1} at 25 °C, compared with up to 90 kcal mol^{-1} for a covalent bond. As well as their role in the DNA double helix, hydrogen bonds stabilize protein secondary structures. The two base-pair combinations - A base-paired with T, and G base-paired with C - explain the base ratios discovered by Chargaff. These are the only pairs that are permissible, partly because of the geometries of the nucleotide bases and the relative positions of the groups that are able to participate in hydrogen bonds, and partly because the pair must be between a purine and a pyrimidine; a purine-purine pair would be too big to fit within the helix, and a pyrimidine-pyrimidine pair would be too small.
- Base-stacking, sometimes called π-π **interactions**, involves hydrophobic interactions between adjacent base pairs and adds stability to the double helix once the strands have been brought together by base-pairing. These hydrophobic interactions arise because the hydrogen-bonded structure of water forces hydrophobic groups into the internal parts of a molecule.

Both base-pairing and base-stacking are important in holding the two polynucleotides together, but base-pairing has added significance because of its biological implications. The limitation that A can only base-pair with T, and G can only base-pair with C, means that DNA replication can result in perfect copies of a parent molecule through the simple expedient of using the sequences of the pre-existing strands to dictate the sequences of the new strands. This is template-dependent DNA synthesis and it is the system used by all cellular DNA polymerases (Section 4.1.1). Its counterpart, template-dependent RNA synthesis, is used by RNA polymerases to make RNA copies of genes, these copies preserving the biological information contained in the sequence of the genomic DNA molecule (Section 3.2.2). The only difference between DNA and RNA syntheses is that when RNA is made, the adenines in the DNA template do not specify thymines in the RNA copy. This is because RNA does not contain thymine; instead adenine pairs with uracil in DNA-RNA hybrids and in double-stranded RNA structures.

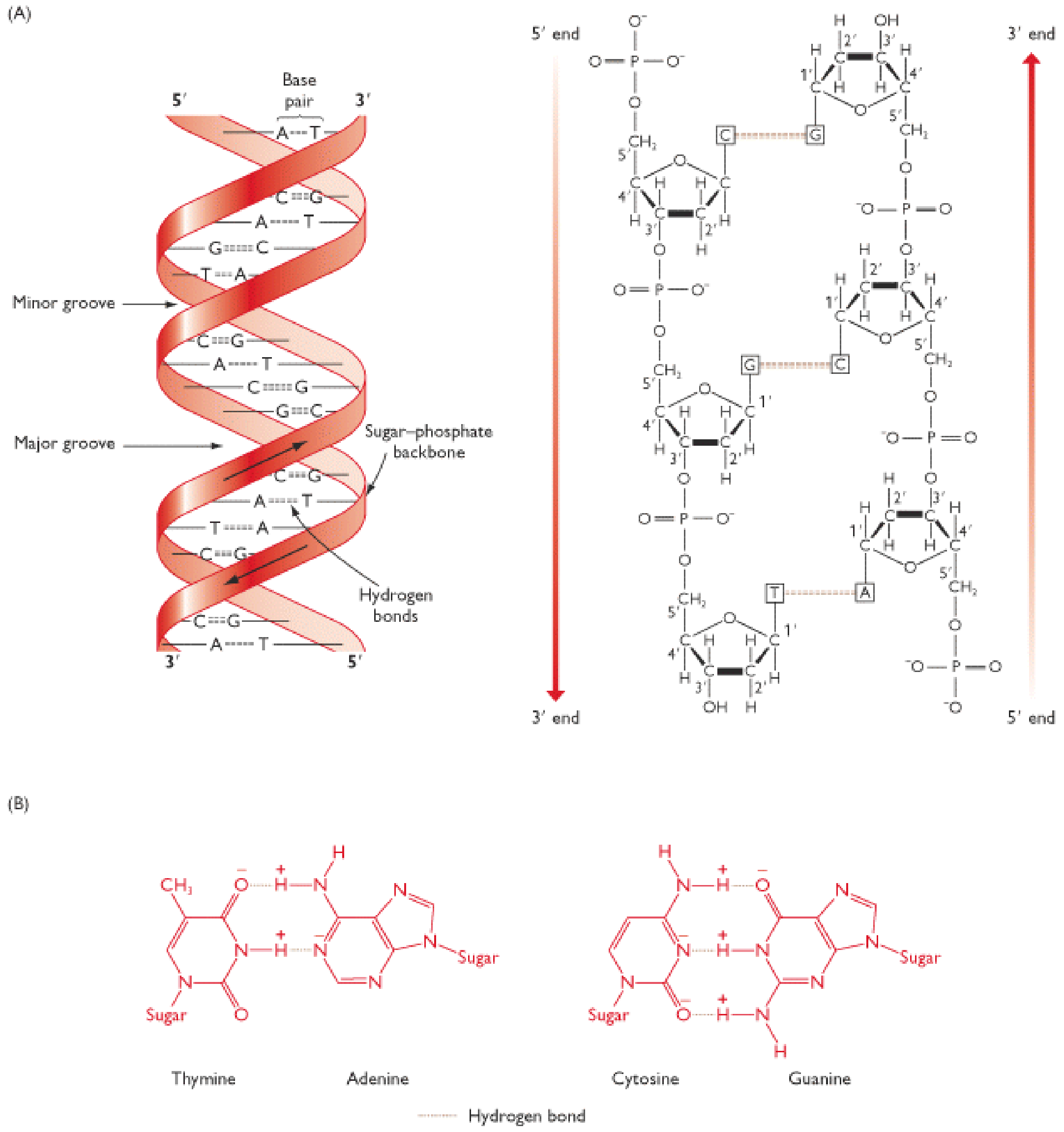


Figure 1.11. The double helix structure of DNA. (A) Two representations of the double helix. On the left the structure is shown with the sugar-phosphate ‘backbones' of each polynucleotide drawn as a red ribbon with the base pairs in black. On the right the chemical structure for three base pairs is given. (B) A base-pairs with T, and G base-pairs with C. The bases are drawn in outline, with the hydrogen bonding indicated by dotted lines. Note that a G-C base pair has three hydrogen bonds whereas an A-T base pair has just two. The structures in part (A) are redrawn from Turner *et al.* (1997) (left) and Strachan and Read (1999) (right).

The double helix has structural flexibility

The double helix described by Watson and Crick, and shown in *Figure 1.11A* , is called the B-form of DNA. Its characteristic features lie in its dimensions: a helical diameter of 2.37 nm, a rise of 0.34 nm per base pair, and a pitch (i.e. distance taken up by a complete turn of the helix) of 3.4 nm, this corresponding to ten base pairs per turn. The DNA in living cells is thought to be predominantly in this B-form, but it is now clear that genomic DNA molecules are not entirely uniform in structure. This is mainly because each nucleotide in the helix has the flexibility to take up slightly different molecular shapes. To adopt these different conformations, the relative positions of the atoms in the nucleotide must change slightly. There are a number of possibilities but the most important conformational changes involve rotation around the β-*N*-glycosidic bond, changing the orientation of the base relative to the sugar, and rotation around the bond between the 3′- and 4′-carbons. Both rotations have a significant effect on the double helix: changing the base orientation influences the relative positioning of the two polynucleotides, and rotation around the 3′–4′ bond affects the conformation of the sugar-phosphate backbone.

Rotations within individual nucleotides therefore lead to major changes in the overall structure of the helix. It has been recognized since the 1950s that changes in the dimensions of the double helix occur when fibers containing DNA molecules are exposed to different relative humidities. For example, the modified version of the double helix called the A-form (*Figure 1.12*) has a diameter of 2.55 nm, a rise of 0.29 nm per base pair and a pitch of 3.2 nm, corresponding to 11 base pairs per turn (*Table 1.1*). Other variations include B′-, C-, C′-, C″-, D-, E- and T-DNAs. All these are right-handed helices like the B-form. A more drastic reorganization is also possible, leading to the left-handed Z-DNA (*Figure 1.12*), a slimmer version of the double helix with a diameter of only 1.84 nm.

The bare dimensions of the various forms of the double helix do not reveal what are probably the most significant differences between them. These relate not to diameter and pitch, but the extent to which the internal regions of the helix are accessible from the surface of the structure. As shown in *Figures 1.11* and *1.12* , the B-form of DNA does not have an entirely smooth surface; instead, two grooves spiral along the length of the helix. One of these grooves is relatively wide and deep and is called the major groove; the other is narrow and less deep and is called the minor groove. A-DNA also has two grooves (*Figure 1.12*), but with this conformation the major groove is even deeper, and the minor groove shallower and broader. Z-DNA is different again, with one groove virtually non-existent but the other very narrow and deep. In each form of DNA part of the internal surface of at least one of the grooves is formed by chemical groups attached to the nucleotide bases. In Chapter 9 we will see that expression of the biological information contained within a genome is mediated by DNA-binding proteins which attach to the double helix and regulate the activity of the genes contained within it. To carry out their function, each DNA-binding protein must attach at a specific position, near to the gene whose activity it must influence. This can be achieved, with a greater or lesser degree of ambiguity, by the protein reaching down into a groove, within which the DNA sequence can be ‘read' without the helix being opened up by breaking the base pairs. A corollary of this is that a DNA-binding protein whose structure enables it to recognize a specific nucleotide sequence within, say, B-DNA might not be able to recognize that sequence if the DNA has taken up a different conformation. As we will see in Chapter 9, conformational variations along the length of a DNA molecule, together with other structural polymorphisms caused by the nucleotide sequence, could be important in determining the specificity of the interactions between the genome and its DNA-binding proteins.

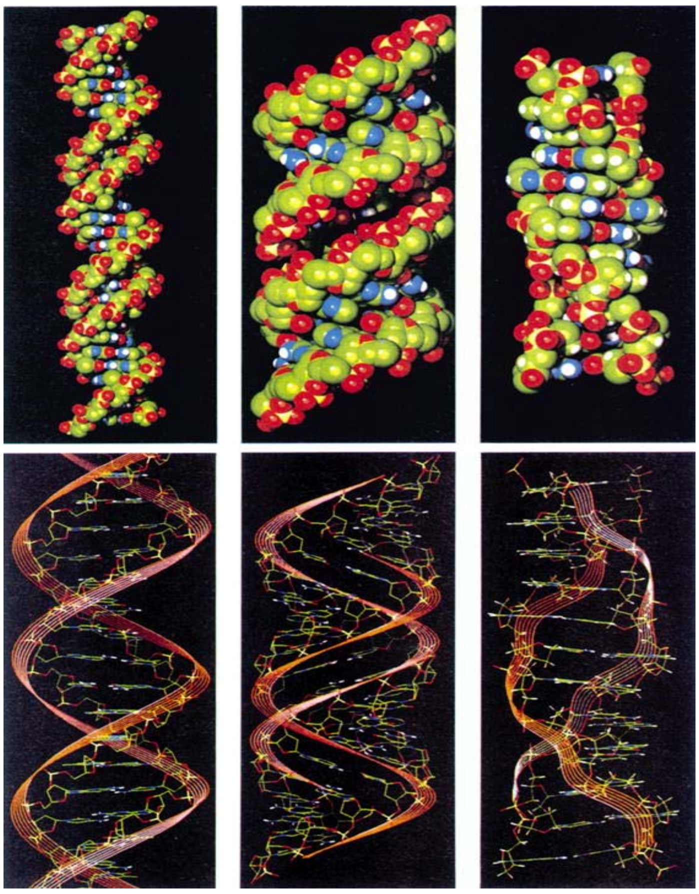


Figure 1.12. Computer-generated images of B-DNA (left), A-DNA (center) and Z-DNA (right). Reprinted with permission from Kendrew A (ed.), *The Encyclopaedia of Molecular Biology*, Plate 1. Copyright 1994 Blackwell Science

Table 1.1. Features of different conformations of the DNA double helix

Feature	Conformation		
	B-DNA	A-DNA	Z-DNA
Type of helix	Right-handed	Right-handed	Left-handed
Helical diameter (nm)	2.37	2.55	1.84
Rise per base pair (nm)	0.34	0.29	0.37
Distance per complete turn (pitch) (nm)	3.4	3.2	4.5
Number of base pairs per complete turn	10	11	12
Topology of major groove	Wide, deep	Narrow, deep	Flat
Topology of minor groove	Narrow, shallow	Broad, shallow	Narrow, deep

1.2. The Human Genome

The critical feature of a DNA molecule is its nucleotide sequence. If the sequence of a DNA molecule is known then the genes that it contains can be identified and the activities of those genes can be studied in detail. Since the mid-1970s, molecular biologists have been able to obtain the sequences of longer and longer stretches of DNA, culminating in the 1990s with completion of the first complete sequences of entire genomes. The most important of these projects has been the one devoted to the human genome.
The Human Genome Project was conceived in 1984 and begun in earnest in 1990 with the primary aim of determining the nucleotide sequence of the entire human nuclear genome. The much smaller mitochondrial genome had been sequenced in the early 1980s (Anderson *et al.*, 1981). The project has been funded by governments and charities from across the world and has been the largest and most complex international collaboration ever attempted in any area of science. A second human genome project was set up by a private company - Celera Genomics of Maryland, USA - in 1998. Both projects completed a draft of the human genome sequence in 2001 and the results were published in the scientific journals *Nature* and *Science* in February of that year (IHGSC, 2001; Venter *et al.,* 2001). These drafts were not complete sequences, each representing only 83–84% of the entire genome, but their coverage was thought to include all of the most important parts of the genome, most of the remaining 16–17% being made up of sequences at the very ends of chromosomes (the telomeres) and around the centromeres (Section 2.2.1), where few, if any, genes are located (Bork and Copley, 2001).

Although only incomplete drafts, each of the genome projects has produced over 2.6 billion base pairs of sequence. This is such a large number that it is difficult to grasp the scale that it represents; an analogy is helpful. The typeface used for the text of this book enables approximately 60 nucleotides of DNA sequence to be written in a line 10 cm in length. If printed out in this format, the human genome sequence would stretch for 5000 km, the distance from Montreal to London, Los Angeles to Panama, Tokyo to Calcutta, Cape Town to Addis Ababa, or Auckland to Perth (*Figure 1.13*). The sequence would fill about 3000 books the size of this one. Understanding the sequence is clearly going to be an enormous task.

We should also bear in mind that although it is standard practice to refer to *the* human genome sequence, there are in fact many human genome sequences because every individual, except pairs of identical twins, have their own version. The differences between individual genomes are largely due to **single nucleotide polymorphisms (SNPs)**, positions in the genome where some individuals have one nucleotide (e.g. an A) and others have a different nucleotide (e.g. a G). Over 1.4 million SNPs have been identified, an average of one for every 2.0 kb of sequence (SNP Group, 2001). On average, every 2 kb also contains a microsatellite (also called a short tandem repeat or **STR**), which is a series of repeated nucleotides (e.g. CACACACA) in which the number of repeats is variable in different individuals. Many of these SNPs and microsatellites have no effect on the function of the genome but many others do. For example, 60 000 SNPs lie within genes and at least some of these have an impact on the activities of these genes, leading to the variations that give each of us our own individual biological characteristics.

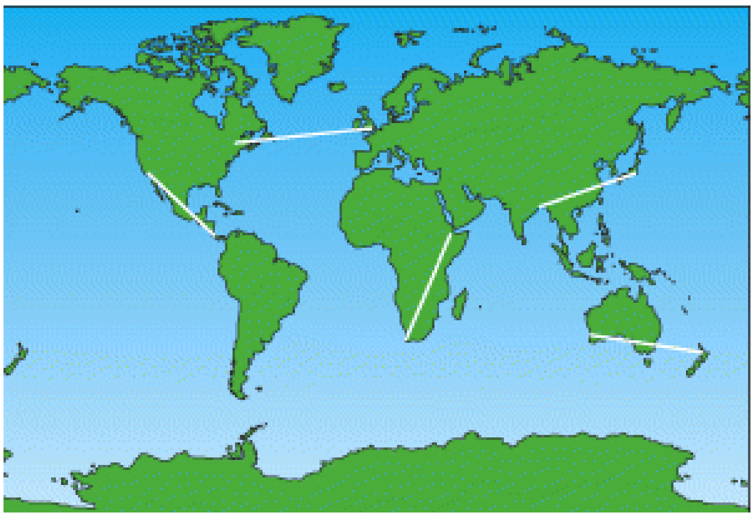


Figure 1.13. The immense length of the human genome. The map illustrates the distance that would be covered by the human genome sequence if it were printed in the typeface used in this book.

1.2.1. The content of the human nuclear genome

What do the DNA sequences reveal about the composition of the human nuclear genome? To begin we will examine a 50-kb segment of chromosome 7 (*Figure 1.14*), this segment forming part of the 'human β T-cell receptor locus', a much larger (685 kb) region of the genome that specifies proteins involved in the immune response (Rowen *et al.*, 1996). Our 50-kb segment contains the following genetic features:

- ***One gene***. This gene is called TRY4 and it contains information for synthesis of the protein called trypsinogen, the inactive precursor of the digestive enzyme trypsin. TRY4 is one of a family of trypsinogen genes present in two clusters at either end of the β T-cell receptor locus. These genes have nothing to do with the immune response, they simply share this part of chromosome 7 with the β T-cell receptor locus. TRY4 is an example of a **discontinuous** gene, the information used in synthesis of the trypsinogen protein being split between five exons, separated by four non-coding introns.
- ***Two gene segments***. These are V28 and V29-1, and each specifies a part of the β T-cell receptor protein after which the locus is named. V28 and V29-1 are not complete genes, only segments of a gene, and before being expressed they must be linked to other gene segments from elsewhere in the locus. This occurs in T lymphocytes and is an example of how a permanent change in the activity of the genome can arise during cellular differentiation (see Section 12.2.1). Like TRY4, both V28 and V29-1 are discontinuous.
- ***One pseudogene***. A pseudogene is a non-functional copy of a gene, usually one whose nucleotide sequence has changed so that its biological information has become unreadable (see page 22). This particular pseudogene is called TRY5 and it is closely related to the functional members of the trypsinogen gene family.
- ***52 genome-wide repeat sequences***. These are sequences that recur at many places in the genome. There are four main types of genome-wide repeat, called LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements), **LTR** (long terminal repeat) **elements** and DNA transposons. Examples of each type are seen in this short segment of the genome.
- ***Two microsatellites***, which, as mentioned above, are sequences in which a short motif is repeated in tandem. One of the microsatellites seen here has the motif GA repeated 16 times, giving the sequence:
  ```
  5′- GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA-3′
  3′- CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT-5′
  ```
- The second microsatellite comprises six repeats of TATT.
- Finally, approximately 50% of our 50-kb segment of the human genome is made up of stretches of non-genic, non-repetitive, single-copy DNA of no known function or significance.

Now we will look at these different components of the genome in greater detail.

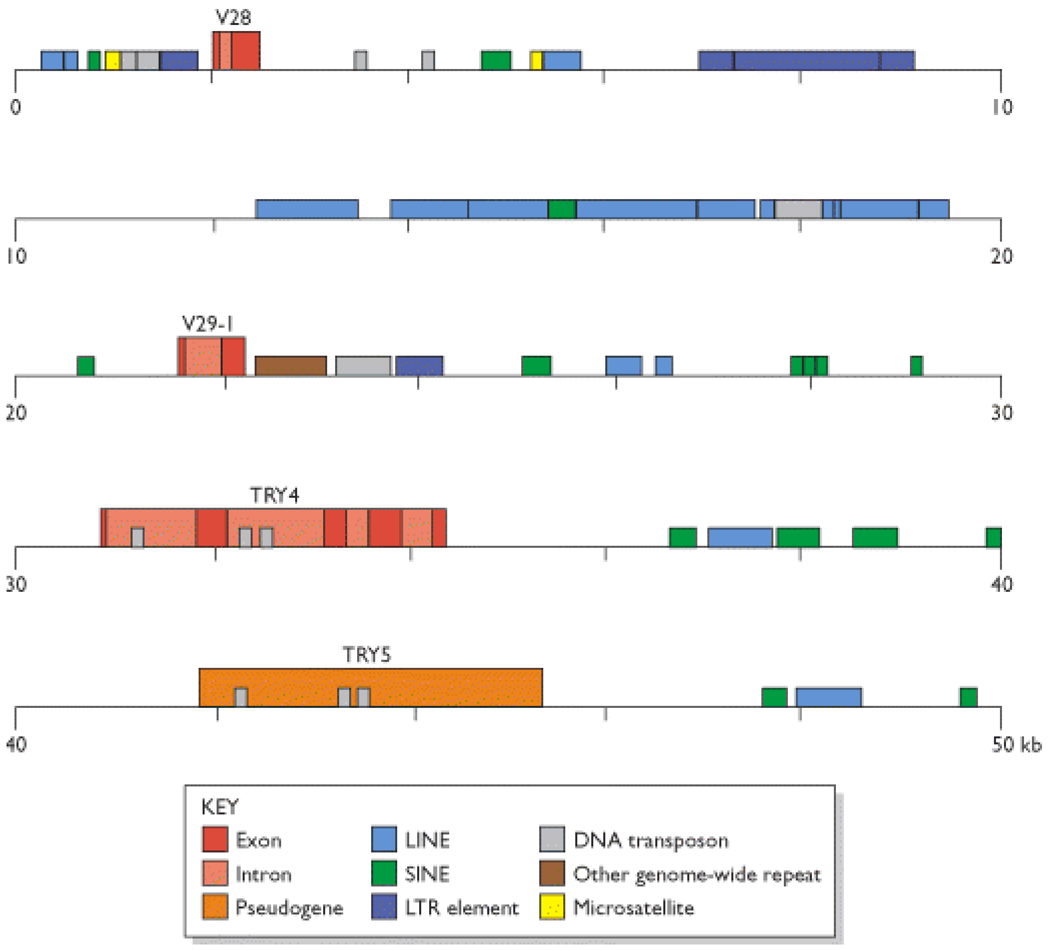


Figure 1.14. A segment of the human genome. This map shows the location of genes, gene segments, pseudogenes, genome-wide repeats and microsatellites in a 50-kb segment of the human β T-cell receptor locus on chromosome 7. Redrawn from Rowen *et al*. (1996).

Genes and related sequences

We look on the genes as the most important part of the human genome because these are the parts that contain biological information. Most genes specify one or more protein molecules, the 'expression' of these genes involving an RNA intermediate, called **messenger** or **mRNA**, which is transported from the nucleus to the cytoplasm where it directs synthesis of the protein coded by the gene (*Figure 1.15*). Other genes do not specify proteins, the end-products of their expression being non-coding RNA, which plays various roles in the cell (Section 3.2.1).

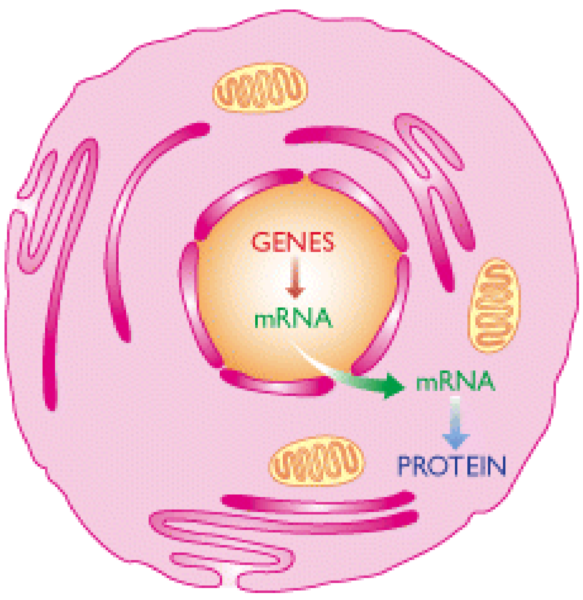


Figure 1.15. Messenger RNA (mRNA) is the intermediate between the genes and their protein products.

Each of the genes and gene segments present in the 50-kb sequence shown in *Figure 1.14* is discontinuous, the biological information being divided into a series of exons separated by non-coding introns. Most human genes are discontinuous, with an average of nine exons per gene, although some genes have many more than this. The record is held by the gene for a large muscle protein called titin, which has 178 exons and is also the longest known human gene at 80 780 bp. During gene expression, the initial RNA that is synthesized is a copy of the entire gene, including the introns as well as the exons. The process called splicing removes the introns from this pre-mRNA and joins the exons together to make the mRNA which eventually directs protein synthesis. At one time it was thought that splicing was a straightforward process, each exon being joined to its neighbor to produce a single mRNA from each discontinuous gene. Now it is known that many pre-mRNAs undergo **alternative** or differential splicing, giving rise to a series of mRNAs containing different combinations of exons and each specifying a different protein (*Figure 1.16*). As well as the gene itself, the pre-mRNA transcribed from a gene also contains sequences from the regions preceding the first exon and following the last exon. These are called the **5′-untranslated region** (**5′-UTR**) and **3′-untranslated region** (**3′-UTR**), respectively.

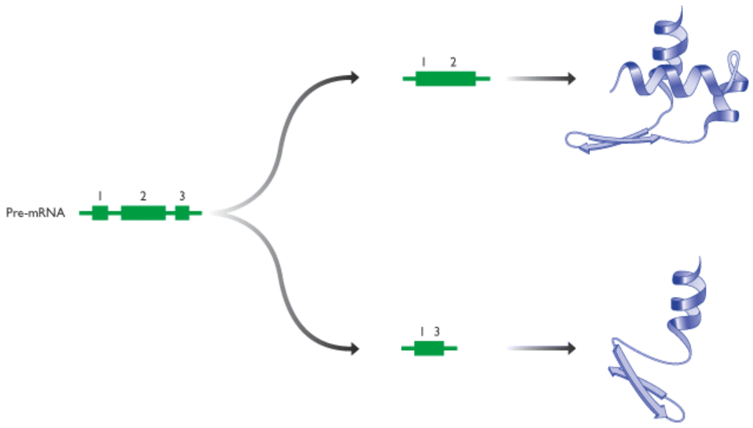


Figure 1.16. Alternative splicing. Alternative splicing results in different combinations of exons becoming linked together, resulting in different proteins being synthesized from the same pre-mRNA.

The various features of an ‘average' human gene are shown in *Figure 1.17* . This diagram is useful for illustrating the components of a gene, but it should be remembered that many human genes do not conform with this ‘average' structure. One type of variation is illustrated by V28 and V29-1 (see *Figure 1.14*). These are not intact genes but just gene segments, two of over 100 similar segments present at the β T-cell receptor locus. In individual T cells, the gene segments are linked together in various combinations to produce different functional receptor genes. This is not the same as splicing because the rearrangements do not involve the RNA transcribed from the genes, but the genes themselves. The final product is a gene that resembles the ‘average' shown in *Figure 1.17* , but this gene is only assembled in the T cells. In other cells segments of the gene are scattered throughout the β T-cell receptor locus. There is also an α T-cell receptor locus, which contains 116 gene segments, and smaller γ and δ loci. Gene rearrangements also occur at the three loci that specify the various components of immunoglobulin proteins. We will examine these rearrangements in more detail in Section 12.2.1 when we look at their impact on the regulation of genome expression. The point that we should bear in mind at the moment is that it is tempting to look on the gene segments at the T-cell receptor and immunoglobulin loci as ‘unusual' because they do not conform to our standard representation of a gene as shown in *Figure 1.17* , but this is not a productive way to approach the study of genomes. By concentrating attention on a simplistic view of the ‘average' gene we risk losing sight of the variations and embellishments that are critical to the overall activity of the genome.

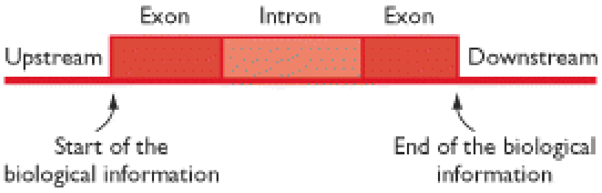

Exon
Intron
Exon
Upstream
Downstream
Start of the biological information
End of the biological information

Figure 1.17. The structure of an ‘average' human gene. This gene has a two exons split by a single intron. For a protein-coding gene, the start of the biological information corresponds to the position of the initiation codon, and the end of the biological information is marked by the termination codon (Section 3.3.2). **‘Upstream'** and **‘downstream'** are two useful terms used to indicate the DNA sequences to either side of the gene.

The functions of human genes

The functions of about half of the 30 000–40 000 human genes are known or can be inferred with a reasonable degree of certainty. The vast majority code for proteins; less than 2500 specify the various types of non-coding RNA. Almost a quarter of the protein-coding genes are involved in expression, replication and maintenance of the genome (*Figure 1.18*) and another 20% specify components of the signal transduction pathways that regulate genome expression and other cellular activities in response to signals received from outside of the cell (Section 12.1). All of these genes can be looked on as having a function that is involved in one way or another with the activity of the genome. Enzymes responsible for the general biochemical functions of the cell account for another 17.5% of the known genes; the remainder are involved in activities such as transport of compounds into and out of cells, the folding of proteins into their correct three-dimensional structures, the immune response, and synthesis of structural proteins such as those found in the cytoskeleton and in muscles. It is possible that as the human gene catalog is made more complete the relative proportions of the genes in the three major categories in *Figure 1.18* will decrease. This is because these major categories represent the most studied areas of cell biology, which means that many of the relevant genes can be recognized because their protein products are known. Genes whose products have not yet been identified are more likely to be involved in the less well studied areas of cellular activity.

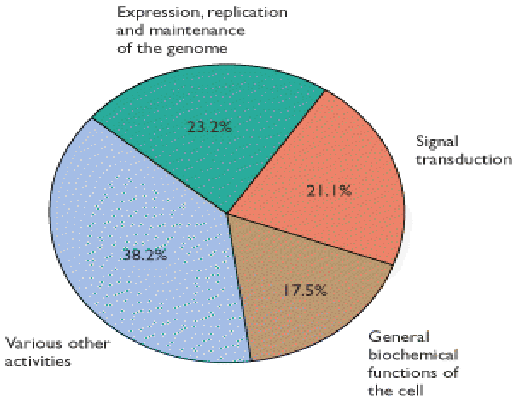


Figure 1.18. Categorization of the human gene catalog. The pie chart shows a categorization of the identified human protein-coding genes. It omits approximately 13 000 genes whose functions are not yet known. The segment labeled ‘various other activities' includes, among others, proteins involved in biochemical transport processes and protein folding, immunological proteins, and structural proteins. Based on *Figure 15* of Venter *et al*. (2001).

One thing that the gene catalog cannot tell us, and will not be able to tell us even when it is complete, is what makes a human being. The minimalist approach to molecular biology, whereby the study of individual genes or groups of genes is expected to lead eventually to a full biomolecular description of how a human being is constructed and functions, has been dealt a severe blow by the draft genome sequences. There are no amazing revelations about what makes humans different from apes. Even when the chimpanzee genome has been completely sequenced (which will not be for several years) it may still not be possible simply from genome comparisons to determine what makes us human (Baltimore, 2001). On the basis of gene number we are only three times more complex than a fruit fly and only twice as complex as the microscopic worm *Caenorhabditis elegans*. More detailed studies of how the human genome functions may reveal key features that underlie some of the special attributes of human beings, but genomics will never explain why a human was able to compose Mozart's 40th symphony, or indeed why it was composed by Mozart and not by an ordinary human.

Pseudogenes and other evolutionary relics

The segment of chromosome 7 shown in *Figure 1.14* contains a single pseudogene, a non-functional copy of a gene. Pseudogenes are a type of evolutionary relic, an indication that the human genome is continually undergoing change. There are two main types of pseudogene:

- A conventional pseudogene is a gene that has been inactivated because its nucleotide sequence has changed by mutation (Section 14.1). Many mutations have only minor effects on the activity of a gene but some are more important and it quite possible for a single nucleotide change to result in a gene becoming completely non-functional. Once a pseudogene has become non-functional it will degrade through accumulation of more mutations and eventually will no longer be recognizable as a gene relic. TRY5 is an example of a conventional pseudogene.
- A processed pseudogene arises not by evolutionary decay but by an abnormal adjunct to gene expression. A processed pseudogene is derived from the mRNA copy of a gene by synthesis of a

DNA copy which subsequently re-inserts into the genome (*Figure 1.19*). Because a processed pseudogene is a copy of an mRNA molecule, it does not contain any introns that were present in its parent gene. It also lacks the nucleotide sequences immediately upstream of the 5′-UTR of the parent gene, which is the region in which the signals used to switch on expression of the parent gene are located. The absence of these signals means that a processed pseudogene is inactive.

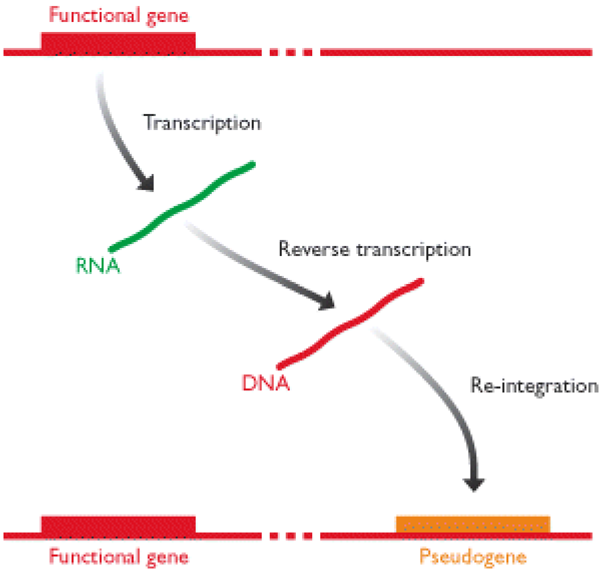


Figure 1.19. The origin of a processed pseudogene. A processed pseudogene is thought to arise by integration into the genome of a copy of the mRNA transcribed from a functional gene. The process by which mRNA is copied into DNA is called reverse transcription and the product is called complementary DNA (cDNA). The cDNA may integrate into the same chromosome as its functional parent, or possibly into a different chromosome.

As well as pseudogenes, genomes also contain other evolutionary relics in the form of truncated genes, which lack a greater or lesser stretch from one end of the complete gene, and gene fragments, which are short isolated regions from within a gene (*Figure 1.20*). We will return to these relics when we look at multigene families in Chapter 2.

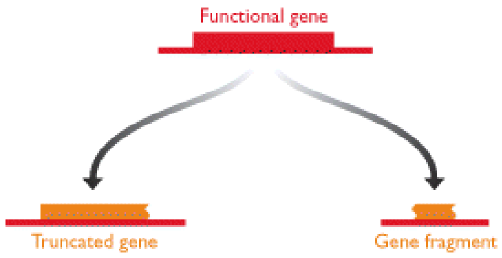


Figure 1.20. A truncated gene and a gene fragment

Genome-wide repeats and microsatellites

The draft sequences have shown that approximately 62% of the human genome comprises intergenic regions, the parts of the genome that lie between genes and which have no known function. These sequences used to be called junk DNA but the term is falling out of favor, partly because the number of surprises resulting from genome research over the last few years has meant that molecular biologists have become less confident in asserting that any part of the genome is unimportant simply because we do not currently know what its function might be. One thing that is clear is that the bulk of the intergenic DNA is made up of repeated sequences of one type or another. Because repeated sequences are important features of all genomes we will deal with them in detail during our general survey of genome anatomies in Chapter 2. Here we will limit ourselves to the key features of the human repeats.

Repetitive DNA can be divided into two categories (*Figure 1.21*): genome-wide or interspersed repeats, whose individual repeat units are distributed around the genome in an apparently random fashion, and tandemly repeated DNA, whose repeat units are placed next to each other in an array. All four types of human genome-wide repeat - SINEs, LINEs, LTR elements and DNA transposons - are represented in

Figure 1.14 . An interesting feature of these genome-wide repeats is that each type appears to be derived from a transposable element, a mobile segment of DNA which is able to move around the genome from one place to another. Many of these elements leave copies of themselves when they move, which explains how they propagate and become common throughout the genome. There are two main classes of transposable element: those that transpose via an RNA intermediate and those that do not (Section 2.4.2). LINEs, SINEs and LTR elements are examples of the first class, and DNA transposons are examples of the second class. The four types of genome-wide repeat are distinguished because of their characteristic sequence features, but there are many variations and each type can be divided into a number of subcategories (*Table 1.2*). SINEs, for example, which are the most numerous genome-wide repeat, comprise three subtypes: Alu elements (approximately 1 090 000 copies in the genome), MIR (393 000 copies) and MIR3 (75 000 copies). Altogether, genome-wide repeats make up 44% of the draft genome sequences, some 1400 Mb of DNA.

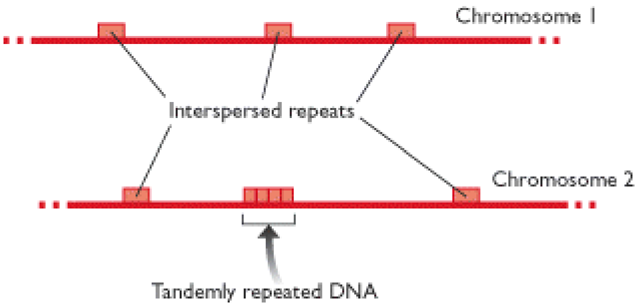


Figure 1.21. The two types of repetitive DNA: interspersed repeats and tandemly repeated DNA.

Table 1.2. The types of genome-wide repeats in the human genome

Type of repeat	Subtype	Approximate number of copies in the human genome
SINEs		1 558 000
	Alu	1 090 000
	MIR	393 000
	MIR3	75 000
LINEs		868 000
	LINE-1	516 000
	LINE-2	315 000
	LINE-3	37 000
LTR elements		443 000
	ERV class I	112 000
	ERV(K) class II	8000
	ERV(L) class III	83 000
	MaLR	240 000
DNA transposons		294 000
	hAT	195 000
	Tc-1	75 000
	PiggyBac	2000
	Unclassified	22 000
Taken from IHGSC (2001). The numbers are approximate and are likely to be under-estimates (Li *et al.*, 2001).		

The microsatellites shown in *Figure 1.14* are examples of tandemly repeated DNA. In a microsatellite the repeat unit is short - up to 13 bp in length - but other types of tandemly repeated sequence have longer units (Section 2.4.1). The commonest type of human microsatellite are dinucleotide repeats, with approximately 140 000 copies in the genome as whole (*Table 1.3*), about half of these being repeats of the motif ‘CA'. Single-nucleotide repeats (e.g. AAAAA) are the next most common (about 120 000 in total). As with genome-wide repeats, it is not clear if microsatellites have a function. It is known that they arise through an error in the process responsible for copying of the genome during cell division (Section 14.1.1), and they might simply be unavoidable products of genome replication.

Table 1.3. Microsatellites in the human genome

Length of repeat unit	Approximate number of copies in the human genome
1	120 000
2	140 000
3	37 500
4	105 000
5	56 000
6	49 000
7	27 000
8	35 500
9	27 500
10	27 500
11	28 000
From IHGSC (2001).	

1.2.2. The human mitochondrial genome

The complete sequence of the human mitochondrial genome has been known for over 20 years (Anderson *et al.*, 1981). At just 16 569 bp, it is much smaller than the nuclear genome, and it contains just 37 genes. Thirteen of these genes code for proteins involved in the respiratory complex, the main biochemical component of the energy-generating mitochondria; the other 24 specify non-coding RNA molecules that are required for expression of the mitochondrial genome. The genes in this genome are much more closely packed than in the nuclear genome (*Figure 1.22*) and they do not contain introns. In many respects, the human mitochondrial genome is typical of the mitochondrial genomes of other animals. We will consider it in more detail when we look at organelle genomes in Chapter 2.

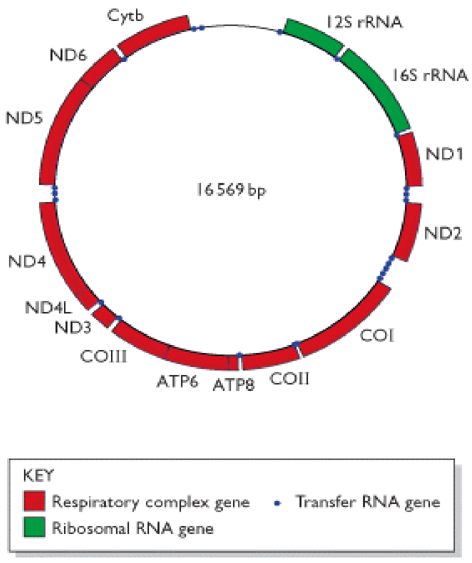


Figure 1.22. The human mitochondrial genome. The human mitochondrial genome is small and compact, with little wasted space, so much so that the ATP6 and ATP8 genes overlap. Abbreviations: ATP6, ATP8, genes for ATPase subunits 6 and 8; COI, COII, COIII, genes for cytochrome *c* oxidase subunits I, II and III; Cytb, gene for apocytochrome *b*; ND1–ND6, genes for NADH hydrogenase subunits 1–6. Ribosomal RNA and transfer RNA are two types of non-coding RNA (Section 3.2.1).

1.3. Why is the Human Genome Project Important?

The human genome has been the focus of biological research for the last decade and will continue to be the center of attention for many years to come. Why is all this activity being devoted to the human genome? There are many reasons.
First, the human gene catalog, containing a description of the sequence of every gene in the genome, will be immensely valuable, even if for many years the functions of some of the genes remain unknown. Not only will the catalog contain the sequences of the coding parts of every gene, it will also include the regulatory regions for these genes. Some of these genes are the ones that, when they function incorrectly, give rise to a genetic disease. The human gene catalog will provide rapid access to these genes, enabling the underlying basis to these diseases to be studied, hopefully leading to strategies for treatment and management.

While the catalog is being completed, attention will focus more and more on the transcriptome and proteome (Chapter 3), which are the keys to understanding how the information contained in the genome is utilized by the cell. The Human Genome Project, and the similar projects currently being carried out with other species' genomes, therefore opens the way to a comprehensive description of the molecular activities of human cells and the ways in which these activities are controlled. This is central to the continued development, not only of molecular biology and genetics, but also of those areas of biochemistry, cell biology and physiology now described as the molecular life sciences.

The genome projects will have additional benefits that at present can only be guessed at. We have seen that the human genome, in common with the genomes of many other organisms, contains extensive amounts of intergenic DNA. We think that most of the intergenic DNA has no function, but perhaps this is because we do not know enough about it. Could the intergenic DNA have a role, but one that at present is too subtle for us to grasp? The first step in addressing this possibility is to obtain a complete description of the organization of the intergenic DNA in different genomes, so that common features, which might indicate a role for some or all of these sequences, can be identified.

There is one final reason for genome projects. The work stretches current technology to its limits. Genome analysis therefore represents the frontier of molecular biology, territory that was inaccessible just a few years ago and which still demands innovative approaches and a lot of sheer hard work. Scientists have always striven to achieve the almost impossible, and the motivation for many molecular biologists involved in genome projects is, quite simply, the challenge of the unknown.