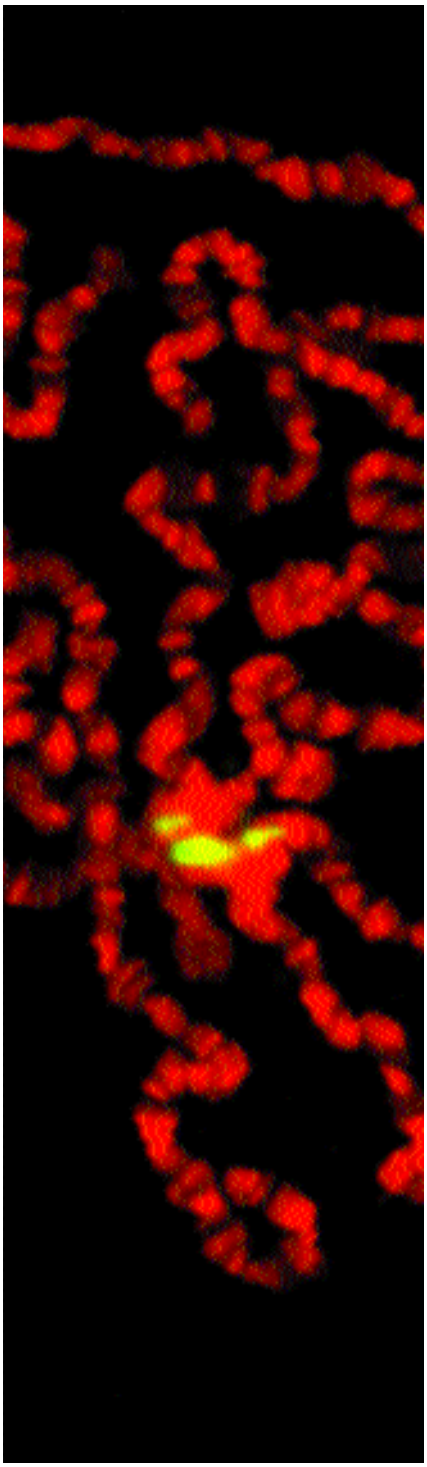
CHAPTER

1 What is a Genome?



Overview

Concepts

- *The genome is the entire DNA content of a cell, including all of the genes and all of the intergenic regions*
- *The human genome contains approximately 80 000 genes but the coding regions of these genes take up only 3% of the genome*
- *The yeast genome contains 6000 genes and has a more compact organization*
- *The genomes of some plants are dominated by repetitive DNA sequences*
- *Prokaryotic genomes are small with very little space between genes*
- *Understanding the information contained in genome sequences will be the major challenge of the early 21st century*

LIFE AS WE KNOW IT is specified by **genomes**. Every organism possesses a genome that contains the **biological information** needed to construct and maintain a living example of that organism. Most genomes, including those for all cellular lifeforms, are made of **DNA** (deoxyribonucleic acid) but a few viruses have **RNA** (ribonucleic acid) genomes. DNA and RNA are polymeric molecules made up of linear, unbranched chains of monomeric subunits called **nucleotides**. Each nucleotide has three parts: a sugar, a phosphate group, and a base (*Figure 1.1*). In DNA, the sugar is 2′-deoxyribose and the bases are adenine (A), cytosine (C), guanine (G) and thymine (T). Nucleotides are linked to one another by **phosphodiester bonds** to form a DNA polymer, or **polynucleotide**, which might be several million nucleotides in length. DNA in living cells is **double-stranded**, two polynucleotides being wound around one another to form the **double helix**. The double helix is held together by **hydrogen bonds** between the base components of the nucleotides in the two strands. The **base-pairing** rules are that A base-pairs with T, and G base-pairs with C. The two DNA molecules in a double helix therefore have **complementary** sequences.

In an RNA nucleotide the sugar is ribose rather than 2′-deoxyribose, and thymine is replaced by the related base called uracil (U). RNA polymers are rarely more than a few thousand nucleotides in length, and RNA in the cell is usually **single-stranded**, though base pairs might form between different parts of a single molecule.

The biological information contained in a genome is encoded in the nucleotide sequence of its DNA or RNA molecules and is divided into discrete units called **genes**. The information contained in a gene is read by proteins that attach to the genome at the appropriate positions and initiate a series of biochemical reactions referred to as **gene expression**. For organisms with DNA genomes, this

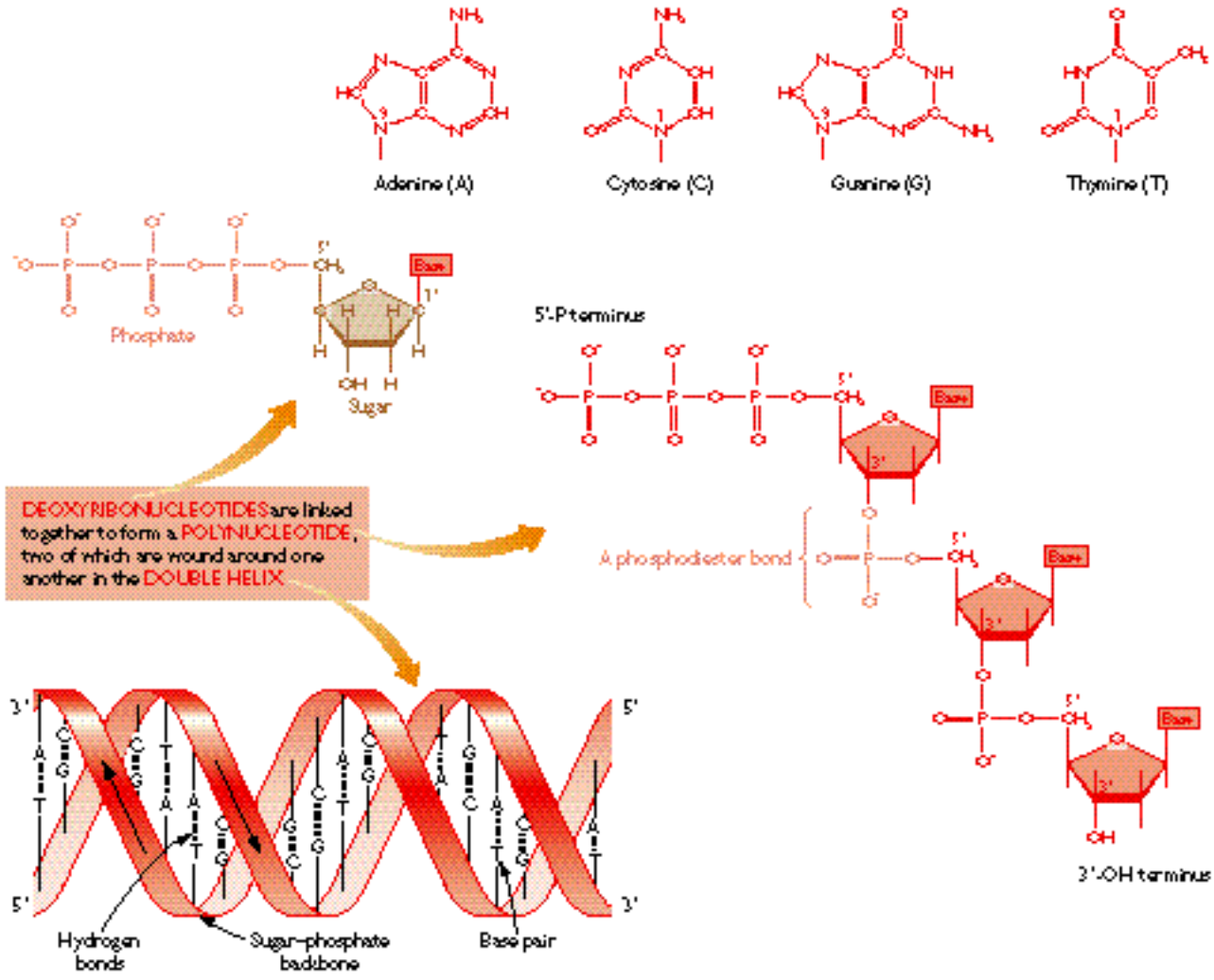


Figure 1.1 The structure of DNA.

Each deoxyribonucleotide consists of the sugar 2′-deoxyribose linked to a phosphate group and one of four bases: adenine, cytosine, guanine or thymine. Nucleotides are linked together by phosphodiester bonds to give a polynucleotide, the two ends of which are chemically different. In the double helix, two polynucleotides, running in different directions, are wound around one another and held together by hydrogen bonds between pairs of bases. See Section 7.1.1 for a more detailed description of DNA and RNA structure.

process was originally looked on as comprising two stages, **transcription** and **translation**, the first producing an RNA copy of the gene and the second resulting in synthesis of a protein whose amino acid sequence is determined, via the **genetic code**, by the nucleotide sequence of the RNA transcript (*Figure 1.2A*). This is still an accurate description of gene expression in simple organisms such as bacteria, but it gives an incomplete picture of the events involved in conversion of genomic information into functional proteins in higher organisms (*Figure 1.2B*). A particular weakness of the transcription–translation interpretation is that it can result in attention being drawn away from the key points in the gene expression pathway at which information flow is regulated.

A complete copy of the genome must be made every time a cell divides. **DNA replication** has to be extremely accurate in order to avoid the introduction of **mutations** into the genome copies. Some mutations do, however, occur, either as errors in replication or due to the effects of chemical and physical mutagens that directly alter the chemical structure of DNA. **DNA repair** enzymes correct many of these errors; those that escape the repair processes become permanent features of the lineage descending from the original mutated genome. These events, along with genome rearrangements resulting from **recombination**, underlie **molecular evolution**, the driving force behind the evolution of living organisms.

1.1 THE HUMAN GENOME

Of all the genomes in existence our own is quite naturally the one that interests us the most. We will therefore begin

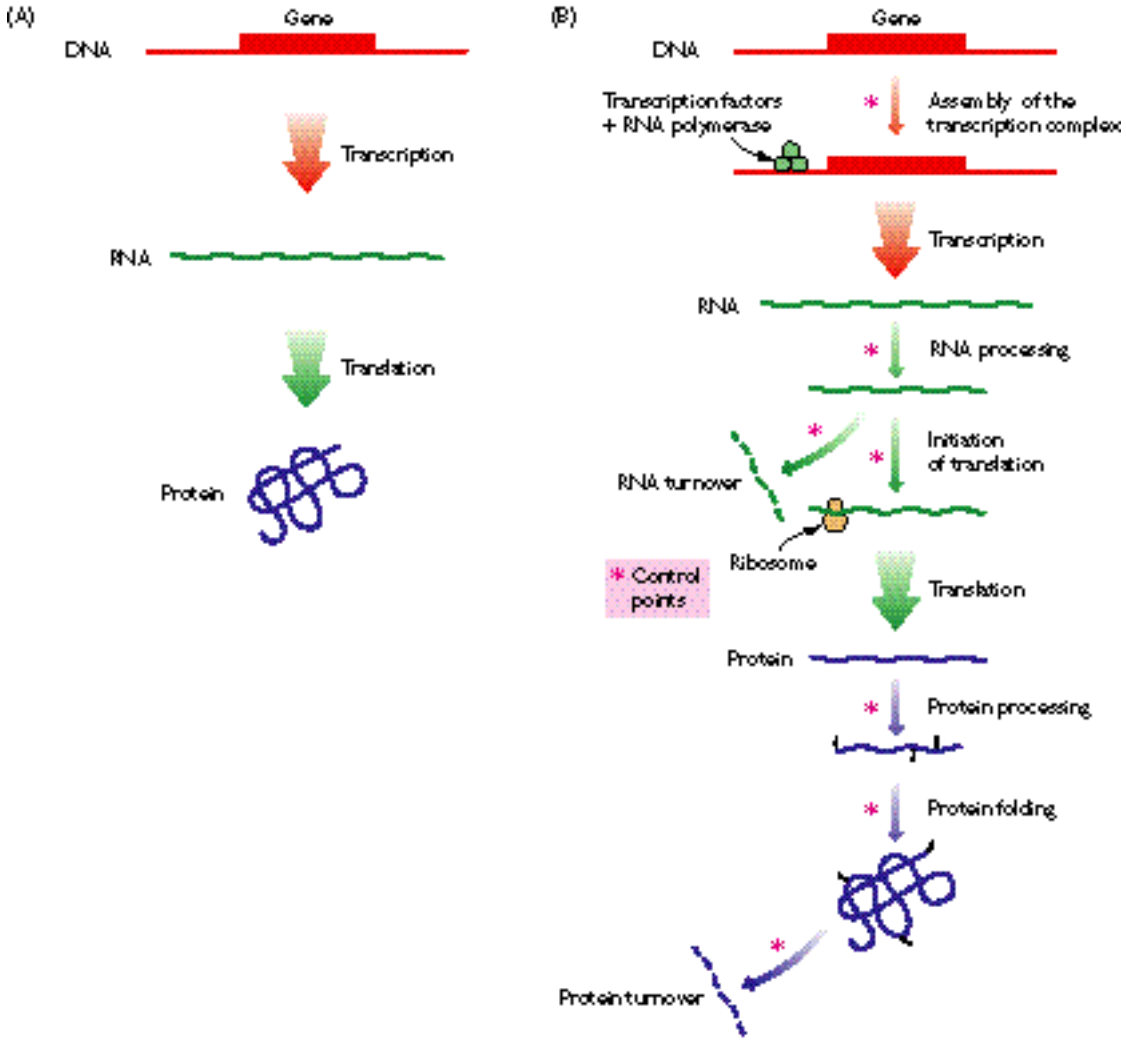


Figure 1.2 Two views of gene expression.

(A) shows the traditional depiction of gene expression, summarized as 'DNA makes RNA makes protein', the two steps being called transcription and translation. (B) gives a more accurate outline of the events involved in gene expression in higher organisms. Key regulatory points are highlighted. Note that these schemes apply only to protein-coding genes. Some genes give rise to noncoding RNAs, such as ribosomal RNA and transfer RNA; these genes are transcribed and processed as shown but the RNAs are not translated (see Section 9.1.1). Full details of all the steps in gene expression are given in Part 2 of this book.

our exploration with an overview of the structure and organization of the human genome. We will then, in the second part of this chapter, consider how similar or different the human genome is to the genomes of other organisms.

1.1.1 The physical structure of the human genome

The human genome is made up of two distinct components (*Figure 1.3*):

- The **nuclear genome**, which comprises approximately 3 000 000 000 bp (**base pairs**) of DNA. This figure is the same as 3 000 000 kb (**kilobase pairs**) or 3000 Mb (**megabase pairs**). The nuclear genome is divided into 24 linear DNA molecules, the shortest 55 Mb in length and the longest 250 Mb, each contained in a different **chromosome**. These 24 chromosomes consist of 22 **autosomes** and the two **sex chromosomes**, X and Y.
- The **mitochondrial genome**, a circular DNA molecule of 16 569 bp, many 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 nuclei 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**, 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.

Before we go any further we should try to gain some appreciation of the immensity of the human genome. Three billion base pairs is such a large number of nucleotides 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 1.4*). The sequence would fill about 3000 books the size of this one. Even the genome of the simplest bacterium would take up a kilometer of this typeface. Such is the enormity of the task facing us if we hope to understand how genomes are constructed and how they work.

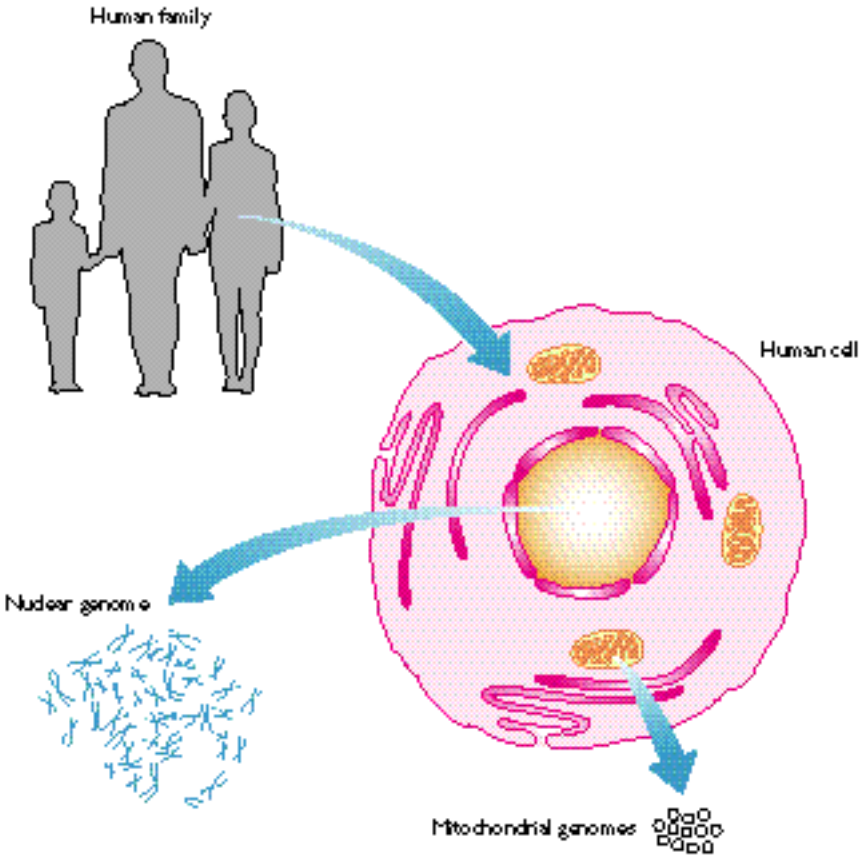


Figure 1.3 The nuclear and mitochondrial components of the human genome.

For more details on the anatomy of the human genome, see Section 6.1.

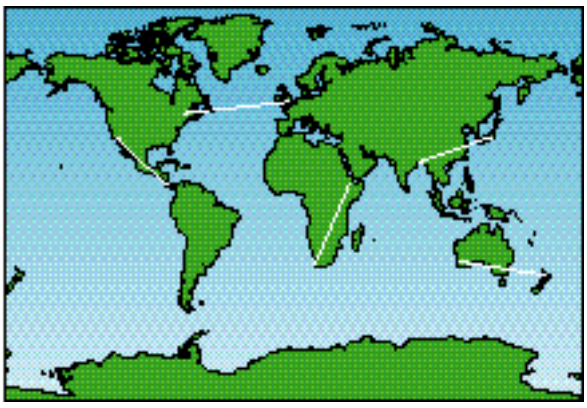


Figure 1.4 The immense length of the human genome.

The map illustrates the distance that would be taken by the human genome sequence if it was printed in the typeface used in this book. There are three billion seconds in 95 years.

1.1.2 The genetic content of the human genome

Understanding how the human nuclear genome is constructed is one of the goals of the **Human Genome Project**. First conceived in 1984 and begun in earnest in 1990, the Project aims to complete the sequence of the human genome by 2005. What will the sequence reveal?

The current consensus is that the human genome contains approximately 80 000 genes, though numbers as low as 50 000 and as high as 150 000 have been suggested by scientists working on the Human Genome Project (Cohen, 1997). In Chapter 6 the evidence on which these estimates are based will be examined (see Box 6.2), and we will also learn that the information contained in these 80 000 or so genes takes up only 3% of the nuclear genome. This is illustrated by *Figure 1.5*, which shows the genetic organization of a 50-kb segment of chromosome 7, 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:

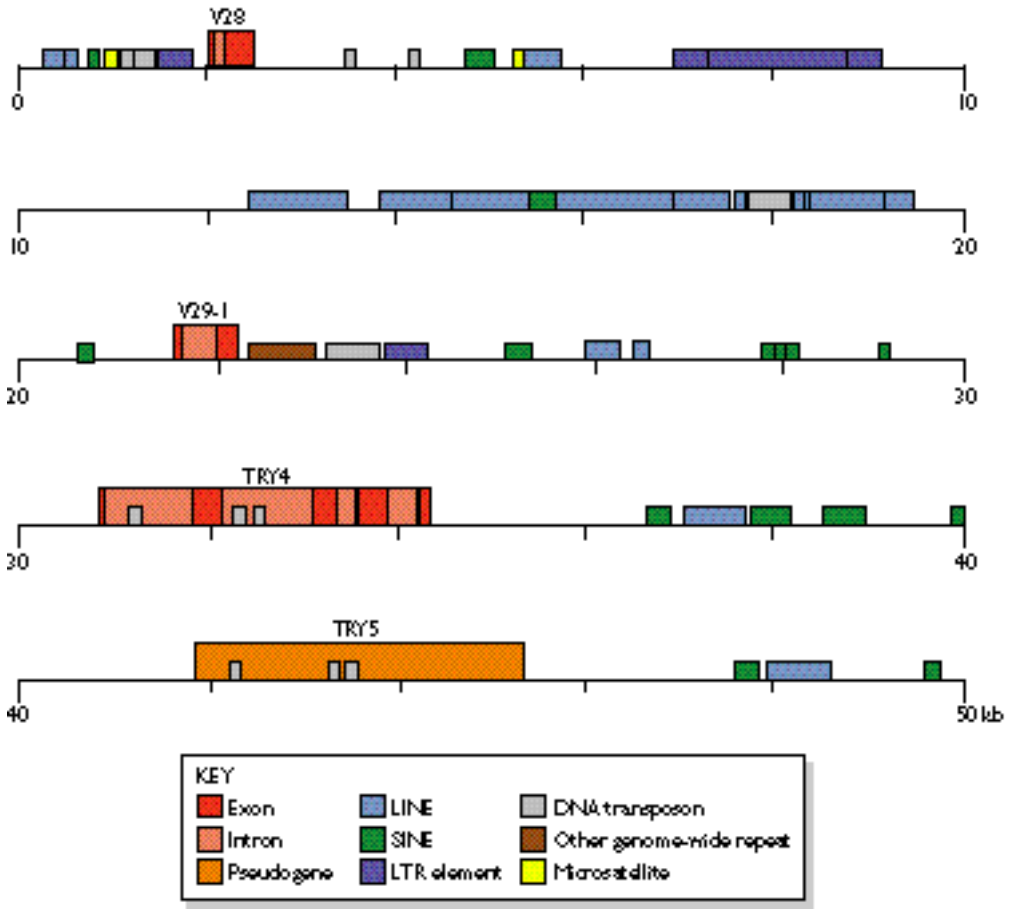


Figure 1.5 A segment of the human genome.

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).

- **One gene**. This gene is called TRY4 and it codes for 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.
- **Two gene segments**. These are V28 and V29-1 and they code for a part of the β T-cell receptor protein after which the locus is named. V28 and V29-1 are quite unusual as they 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 11.2.1). Note that TRY4, V28 and V29-1, like most human genes, are **discontinuous**, being made up of **exons**, containing protein-coding information, separated by noncoding **introns**. The exons, when added together, make up a total of 1414 bp, or 2.8% of the 50-kb segment. The coding capacity of this segment is therefore fairly typical of the genome as a whole.
- **One pseudogene**. A **pseudogene** is a nonfunctional copy of a gene, usually one that has mutated so that its biological information has become unreadable (Section 6.1.1). This particular pseudogene is called TRY5 and it is closely related to the functional members of the trypsinogen gene family.
- **Fifty-two 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**, and examples of each type are seen in this segment, together making up 39.1% of the sequence. We will look at genome-wide repeat sequences in more detail in Section 6.3.2.
- **Two microsatellites**. These are sequences in which a short motif is repeated in tandem (Section 6.3.1). One of the microsatellites seen here has the motif GA repeated sixteen times, giving the sequence

  5′–GAGAGAGAGAGAGAGAGAGAGAGAGAGAGA–3′
  3′–CTCTCTCTCTCTCTCTCTCTCTCTCTCTCT–5′

  The second microsatellite comprises six repeats of TATT. Many microsatellites are polymorphic, the number of repeats being variable in different individuals. Microsatellites are useful marker points in the genome and in Chapters 2 and 3 we will see how they have been used in the construction of genome maps.
- Finally, approximately 50% of our 50-kb segment of the human genome is made up of stretches of nongenic, nonrepetitive, single-copy DNA of no known function or significance.

No short segment can be considered truly representative of the human genome as a whole. In some respects the 50-kb sequence shown in *Figure 1.5* is probably fairly atypical, but it illustrates the range of genetic features contained in the human genome. The next question that we will address is how similar the human genome is to the genomes of other organisms.

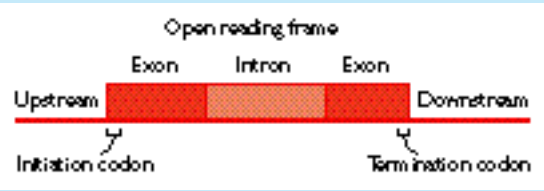
Box 1.1: Some key facts about genes

A gene is a segment of the genome that is transcribed into RNA. If the RNA is a transcript of a protein-coding gene then it is called a **messenger RNA (mRNA)** and is translated into protein. If the RNA is noncoding, such as **ribosomal RNA (rRNA)** and **transfer RNA (tRNA)** then it is not translated. Noncoding RNAs have various functions in the cell (Section 9.1.1).

The part of a protein-coding gene that is translated into protein is called the **open reading frame (ORF)**. Each triplet of nucleotides in the ORF is a **codon** that specifies an amino acid in accordance with the rules of the **genetic code** (Section 10.1.2). The ORF is read in the 5′ to 3′ direction along the mRNA. The ORF starts with an **initiation codon** (usually ATG) and ends with a **termination codon** (TAA, TAG or TGA). The part of the mRNA before the ORF is called the **leader** segment, and that following the ORF is the **trailer** segment.

Many genes in eukaryotes are discontinuous, being split into **exons** and **introns**. The introns are removed from the primary transcript by **splicing** to produce the functional RNA molecule (Section 9.2.3).

'Upstream' refers to the region of DNA before a gene; **'downstream'** is after the gene.



The anatomy of a eukaryotic protein-coding gene

1.2 GENOMES OF OTHER ORGANISMS

Biologists divide the living world into two types of organism (*Figure 1.6*):

- **Eukaryotes**, whose cells contain membrane-bound compartments, including a nucleus and organelles such as mitochondria and, in the case of plant cells, chloroplasts. Eukaryotes include animals, plants, fungi and protozoa.
- **Prokaryotes**, whose cells lack extensive internal

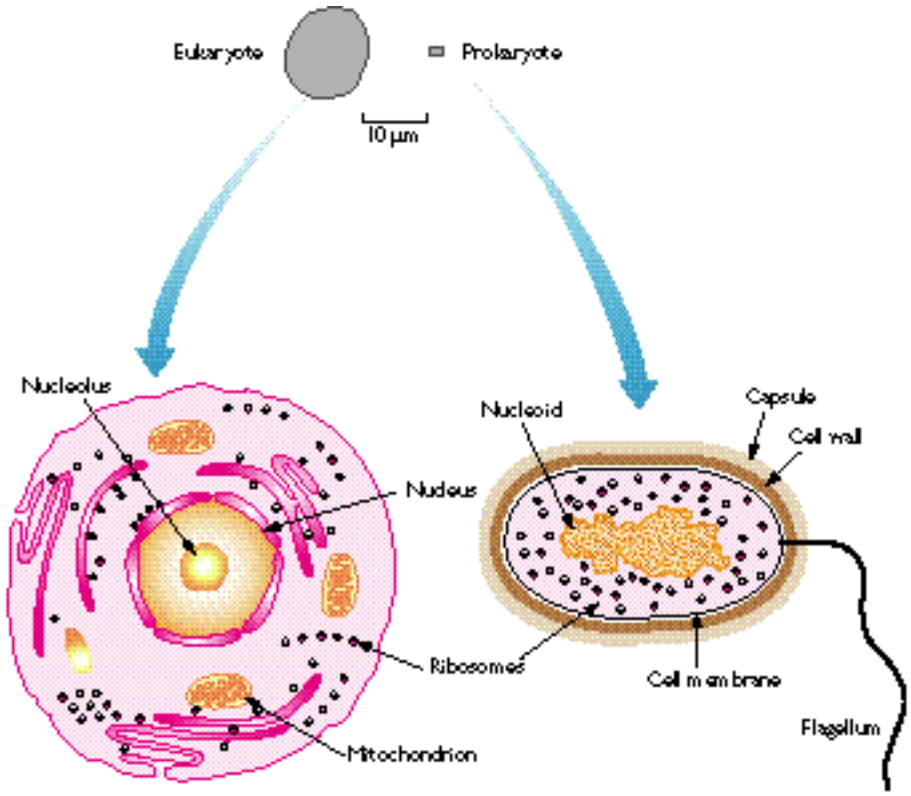


Figure 1.6 Cells of eukaryotes (left) and prokaryotes (right).

The top part of the figure shows a typical human cell and typical bacterium drawn to scale. The human cell is 10 µm in diameter and the bacterium is rod-shaped with dimensions of 1 × 2 µm. The lower drawings show the internal structures of eukaryotic and prokaryotic cells. Eukaryotic cells are characterized by their membrane-bound compartments, which are absent in prokaryotes. The bacterial DNA is contained in the structure called the nucleoid.

compartments. There are two very different groups of prokaryotes, distinguished from one another by characteristic genetic and biochemical features:

(i) the **bacteria**, which include most of the commonly encountered prokaryotes such as the gram-negatives (e.g. *Escherichia coli*), the gram-positives (e.g. *Bacillus subtilis*), the cyanobacteria (e.g. *Anabaena*) and many more;
(ii) the **archaea**, which are less well-studied, and have mostly been found in extreme environments such as hot springs, brine pools and lake bottoms.

Eukaryotes and prokaryotes have quite different types of genome and we must therefore consider them separately.

1.2.1 Genomes of eukaryotes

Humans are fairly typical eukaryotes and the human genome is in many respects a good model for eukaryotic genomes in general. All of the eukaryotic nuclear genomes that have been studied are, like the human version, divided into two or more linear DNA molecules, each contained in a different chromosome, and all eukaryotes also possess smaller, usually circular, mitochondrial genomes. The only general eukaryotic feature not illustrated by the human genome is the presence in plants and other photosynthetic organisms of a third genome, located in the chloroplasts. The two organellar genomes – mitochondrial and chloroplast – will be examined in our detailed survey of genome organization in Chapter 6.

Although the basic physical structures of all eukaryotic nuclear genomes are similar, one important feature is very different in different organisms. This is genome size, the smallest eukaryotic genomes being less than 10 Mb in length, and the largest over 100 000 Mb. As can be seen in *Table 1.1*, this size range coincides to a certain extent with the complexity of the organism, the simplest eukaryotes such as fungi having the smallest genomes, and higher eukaryotes such as vertebrates and flowering plants having the largest ones. This makes sense as one would expect the complexity of an organism to be related to the number of genes in its genome, so higher eukaryotes need larger genomes to accommodate the extra genes. But the correlation is not precise: if it was, then the nuclear genome of the yeast *Saccharomyces cerevisiae*, which at 12 Mb is 0.004 times the size of the human nuclear genome, would be expected to contain 0.004 × 80 000 genes, which is just 320. In fact the *S. cerevisiae* genome contains about 6000 genes.

For many years the lack of precise correlation between the complexity of an organism and the size of its genome

Table 1.1 Sizes of genomes

Organism	Genome size (Mb)
Prokaryotes	0.5–50
Mycoplasma genitalium	0.58
Escherichia coli	4.64
Bacillus megaterium	30
Eukaryotes	
Fungi	9.4–175
Saccharomyces cerevisiae (yeast)	12.1
Aspergillus nidulans	25.4
Protozoa	37.5–330 000
Tetrahymena pyriformis	190
Invertebrates	56.5–21 250
Caenorhabditis elegans (nematode worm)	100
Drosophila melanogaster (fruit fly)	140
Bombyx mori (silkworm)	490
Strongylocentrotus purpuratus (sea urchin)	845
Locusta migratoria (locust)	5000
Vertebrates	375–140 000
Fugu rubripes (pufferfish)	400
Homo sapiens (humans)	3000
Mus musculus (mouse)	3300
Plants	95–120 000
Arabidopsis thaliana (vetch)	100
Oryza sativa (rice)	565
Pisum sativum (pea)	4800
Zea mays (maize)	5000
Triticum aestivum (wheat)	17 000
Fritillaria assyrica (fritillary)	120 000

Data taken from Brown (1998)

was looked on as a bit of a puzzle, the so-called C-value paradox. In fact the answer is quite simple: space is saved in the genomes of less complex organisms as the genes are more closely packed together. The *S. cerevisiae* genome, the sequence of which was completed in 1996, illustrates this point, as we can see from the top two parts of *Figure 1.7*, where the 50-kb segment of the human genome that we looked at earlier in the chapter is compared with a 50-kb segment of the yeast genome. The yeast genome segment, which comes from chromosome III (the first eukaryotic chromosome to be completely sequenced; Oliver *et al.*, 1992), has the following distinctive features:

- ***It contains more genes than the human segment.*** This region of yeast chromosome III contains 26 genes thought to code for proteins and two coding for **transfer RNAs** (**tRNAs**), short molecules involved in reading the genetic code during translation (Section 10.1). These 28 genes take up 66.4% of the 50-kb sequence.
- ***Relatively few of the yeast genes are discontinuous.*** In this segment of chromosome III none of the genes are discontinuous. In the entire yeast genome there are only 239 introns, which is a remarkably small number compared with the genomes of higher eukaryotes, in which some individual genes can have more than 100 introns.
- ***There are fewer genome-wide repeats.*** This part of chromosome III contains a single LTR element, called Ty2, and four truncated LTR elements called delta sequences. These five genome-wide repeats make up 13.5% of the 50-kb segment, but this figure is not entirely typical of the yeast genome as whole. When all 16 yeast chromosomes are considered the total amount of sequence taken up by genome-wide repeats is only 3.4% of the total.

The picture that emerges is that the genetic organization of the yeast genome is much more economical than that of the human version. The genes themselves are more compact, having fewer introns, and the spaces between the genes are relatively short with much less space taken up by genome-wide repeats and other noncoding sequences. We will see in Chapter 6 that the hypothesis that the more complex eukaryotes have less compact genomes holds when other species are examined. We will also see that the genome-wide repeats appear to play an intriguing role in dictating the compactness or otherwise of a genome, a general rule being that the smaller genomes lack extensive repetition, but the larger ones display a proliferation in the number of repeat sequences that are present. This is strikingly illustrated by the maize genome, which at 5000 Mb is larger than the human genome but still relatively small for a flowering plant. Only a few limited regions of the maize genome have been sequenced, but some remarkable results have been obtained, revealing a genome dominated by repetitive elements. *Figure 1.7C* shows a 50-kb segment of this genome, either side of one member of a family of genes coding for the alcohol dehydrogenase enzymes (SanMiguel *et al.*, 1996). This is the only gene in this 50-kb region, though there is a second one, of unknown function, approximately 100 kb beyond the right-hand end of the sequence shown here. Instead of genes, the dominant feature of this genome segment is the genome-wide repeats. The majority of these are of the LTR element type, which comprise virtually all of the noncoding part of the segment, and on their own are estimated to make up approximately 50% of the maize genome.

These examples show that some of the features of eukaryotic genomes are surprisingly unusual. We will fill in the details concerning the contents and organizations of eukaryotic genomes in Chapter 6, and in Chapter 15 we will examine what little is known about the evolutionary events that led to these diverse designs.

1.2.2 Genomes of prokaryotes

Prokaryotic genomes are very different from eukaryotic ones. There is some overlap between the largest prokaryotic and smallest eukaryotic genomes, but on the whole prokaryotic genomes are much smaller (see *Table 1.1*). The *E. coli* genome, for example, is just 4639 kb, two-fifths the size of the yeast genome. The next difference concerns physical organization. In prokaryotes most if not all of the genome is contained in a single DNA molecule, and this

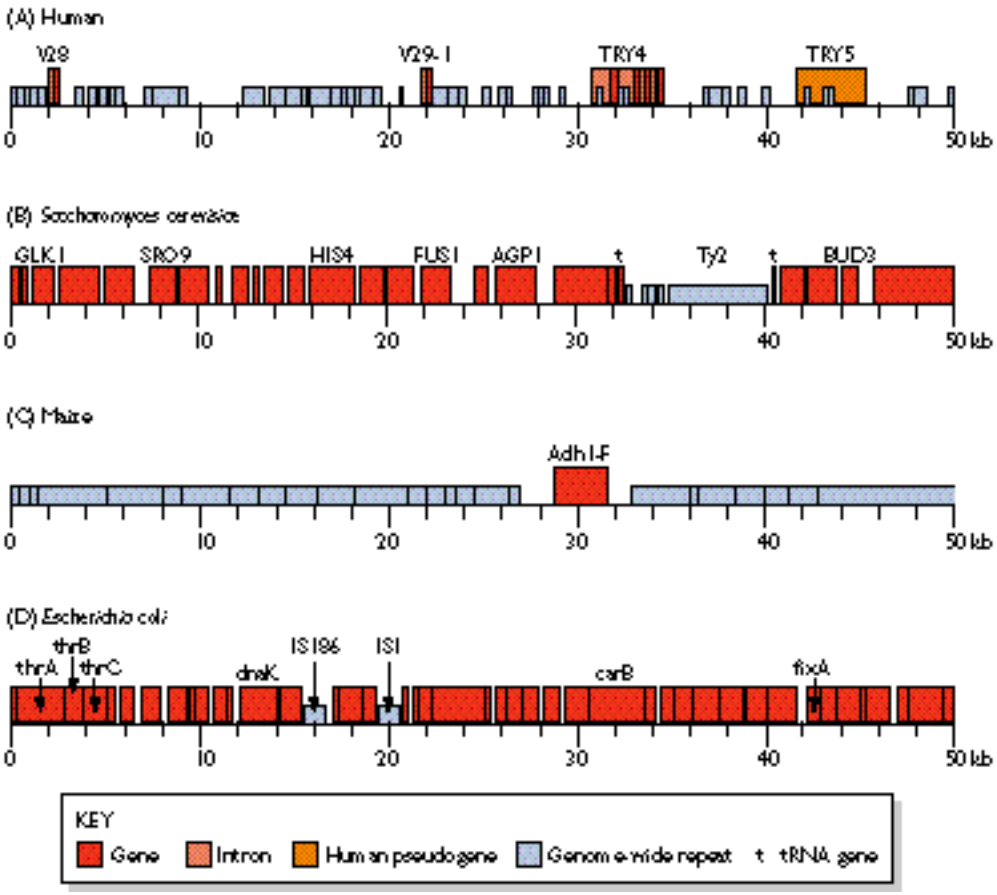


Figure 1.7 Comparison of the genomes of humans, yeast, maize and *E. coli*.

(A) is the 50-kb segment of the human β T-cell receptor locus shown in *Figure 1.5*. This is compared with 50-kb segments from the genomes of *S. cerevisiae* (chromosome III; redrawn from Oliver *et al.*, 1992), maize (redrawn from SanMiguel *et al.*, 1996) and *E. coli* (redrawn from Blattner *et al.*, 1997). See the text for more details.

molecule is circular rather than linear. In addition to their single 'chromosome', prokaryotes may also have additional genes on independent, smaller, circular or linear DNA molecules called **plasmids** (*Figure 1.8*). Genes carried by plasmids are useful, coding for properties such as antibiotic resistance or the ability to utilize complex compounds such as toluene as a carbon source, but plasmids appear to be dispensable: a prokaryote can exist quite effectively without them. For this reason, a bacterial or archaeal 'genome' is usually defined as just the main DNA molecule, with plasmids looked on as ancillary components, not parts of the genome itself (for exceptions to this rule, see Section 6.2.1).

As well as having smaller genomes, prokaryotes generally have fewer genes than eukaryotes. *E. coli* has just 4397. After our discussion regarding eukaryotic gene organization, it will probably come as no surprise to learn that prokaryotic genomes are even more compact than those of lower eukaryotes. If we return to *Figure 1.7* we can see this illustrated by part *D*, which shows a 50-kb segment of the *E. coli* genome (Blattner *et al.*, 1997). It is immediately obvious that there are more genes and less space between them, 43 genes taking up 85.9% of the segment. Some genes have virtually no space between them at all: *thrA* and *thrB*, for example, are separated by a single nucleotide, and *thrC* begins at the nucleotide immediately following the last nucleotide of *thrB*. These three genes are an example of an **operon**, a group of genes involved in a single biochemical pathway (in this case, synthesis of the amino acid threonine) and expressed in conjunction with one another. Operons have been used as model systems for understanding how gene expression is regulated (Section 8.3.1).

Two other features of prokaryotic genomes can be deduced from *Figure 1.7D*. First, there are no introns in the genes present in this segment of the *E. coli* genome. In fact *E. coli* has no discontinuous genes at all and it is generally believed that this type of gene structure is absent in prokaryotes with a few exceptions, mainly among the archaea. The second feature is the infrequency of repetitive sequences. Prokaryotic genomes do not have anything equivalent to the high-copy-number, genome-wide repeat families found in eukaryotic genomes. They do, however, possess certain sequences that might be repeated elsewhere in the genome, examples being the **insertion sequences**

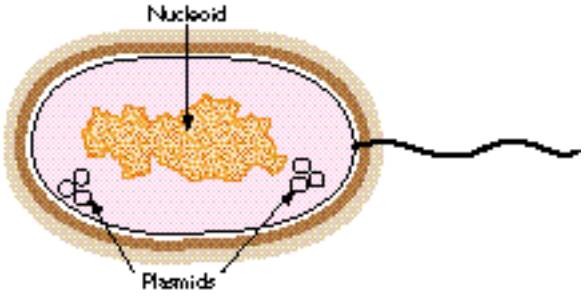


Figure 1.8 Plasmids are small circular DNA molecules that are found inside some prokaryotic cells.

IS1 and IS186 that can be seen in the 50-kb segment shown in *Figure 1.7D*. These are examples of **transposable elements**, sequences that have the ability to move around the genome and to transfer from one organism to another, even sometimes between two different species (Section 6.3.2). The positions of the IS1 and IS186 elements shown in *Figure 1.7D* therefore refer only to the particular *E. coli* strain from which this sequence was obtained: if a different strain is examined then the IS sequences could well be in different positions or might be entirely absent from the genome.

1.3 WHY ARE GENOME PROJECTS IMPORTANT?

As we begin the new millennium, the major goal of molecular biology is to obtain the complete sequences of as many genomes as possible. Considerable effort is being devoted not only to the Human Genome Project but also to equivalent projects aimed at the genomes of a wide range of other organisms (*Table 1.2*). Why all this activity devoted to genome sequences? There are many reasons.

First, our current perception is that genome sequences are the key to the continued development of not only molecular biology and genetics, but also of those areas of biochemistry, cell biology and physiology now described as the **molecular life sciences**. A catalog containing a description of the sequence of every gene in a genome is immensely valuable, even if at first the functions of many of the genes are 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. The genome sequence therefore opens the way to a comprehensive description of the molecular activities of living cells and the ways in which these activities are controlled.

Gene catalogs also aid the isolation and utilization of important genes, such as those human genes responsible

Table 1.2 Some of the organisms for which complete genome sequences should be available by 2005

Organism	Genome size (Mb)	Internet address for latest news
Archaea[†]		
Methanococcus jannaschii	1.66	http://www.tigr.org/tdb/mdb/mjdb/mjdb.html
Methanobacterium thermoautotrophicum	1.75	http://www.genomecorp.com/htdocs/sequences/methanobacter/abstract.html
Archaeoglobus fulgidus	2.18	ftp://ftp.tigr.org/pub/data/a_fulgidus
Bacteria[†]		
Mycoplasma genitalium	0.58	http://www.tigr.org/tdb/mdb/mgdb/mgdb.html
Mycoplasma pneumoniae	0.81	http://www.zmbh.uni-heidelberg.de/M-pneumoniae/MP_Home.html
Treponema pallidum	1.14	http://www.tigr.org/tdb/mdb/tpdb/tp_bg.html
Borrelia burgdorferi	1.44	ftp://ftp.tigr.org/pub/data/b_burgdorferi
Aquifex aeolicus	1.55	
Helicobacter pylori	1.66	http://www.tigr.org/tdb/mdb/hpdb/hpdb.html
Haemophilus influenzae	1.83	http://www.tigr.org/tdb/mdb/mdb.html
Synechocystis sp.	3.57	http://kazusa.or.jp/cyano/cyano.html
Bacillus subtilis	4.20	http://www.pasteur.fr/Bio/SubtiList.html
Mycobacterium tuberculosis	4.40	http://www.sanger.ac.uk/Projects/M_tuberculosis/
Escherichia coli	4.60	http://www.genetics.wisc.edu:80/index.html
Eukaryotes		
Saccharomyces cerevisiae	12.1	http://www.mips.biochem.mpg.de/
Arabidopsis thaliana	100	http://genome-www.stanford.edu/Arabidopsis/
Caenorhabditis elegans	100	http://moulon.inra.fr/acedb/acedb.html
Drosophila melanogaster	140	http://flybase.bio.indiana.edu/
Oryza sativa	565	http://www.staff.or.jp/
Homo sapiens	3000	http://gdbwww.gdb.org/
Mus musculus	3300	http://www.informatics.jax.org/

[†]Most of these archaeal and bacterial sequences have already been completed.

for inherited disease, or bacterial genes whose protein products have industrial value. These genes can be isolated from a genome even if the complete genome sequence is not known, but the process is time-consuming and costly, and a different project has to be devised for each gene that is sought. It will be very much easier to obtain a copy of the desired gene if the sequence of its genome is already known, so that the gene can simply be withdrawn from the catalog.

Genome sequences 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 all higher eukaryotes, contains extensive amounts of noncoding DNA. We assume that most of the noncoding DNA has no function, but perhaps this is because we do not know enough about it. Could the noncoding 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 noncoding 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 sequencing 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.

REFERENCES

Blattner FR, Plunkett G, Bloch CA, *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.

Cohen J (1997) How many genes are there? *Science*, **275**, 769.

Oliver SG, van der Aart QJM, Agostini-Carbone ML, *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature*, **357**, 38–46.

Rowen L, Koop BF and Hood L (1996) The complete 685-kilobase DNA sequence of the human β T cell receptor locus. *Science*, **272**, 1755–1762.

SanMiguel P, Tikhonov A, Jin Y-K, *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.

FURTHER READING

Alberts B, Bray D, Lewis J, Raff M, Roberts K and Watson JD (1992) *Molecular Biology of the Cell, 3rd edition.* Garland Publishing, New York. — *The best source of general information on eukaryotic cell biology.*

Brown TA (1998) *Molecular Biology Labfax, 2nd edition, Volume 1.* Academic Press, London. — *Chapter 4 gives extensive data on genome sizes and genome sequencing projects.*

McKusick VA (1989) The Human Genome Organisation: history, purposes and membership. *Genomics*, **5**, 385–387. — *Describes the goals of the Human Genome Project.*

Prescott LM, Harley JP and Klein DA (1993) *Microbiology, 2nd edition.* W.C. Brown, Dubuque — *One of the best books on prokaryotic biology.*