

CHAPTER 2

Comparative Genomics



● LEARNING GOALS:

- To know the three major divisions of living things – archaea, bacteria and eukaryotes – based on analysis of the sequences of 16S rRNA genes.
- To recognize the prevalence of horizontal gene transfer, especially among prokaryotes, and to understand that horizontal gene transfer is inconsistent with the hierarchical 'tree of life' picture that the Linnaean classification scheme suggests.
- To be familiar with major events in the history of life.
- To appreciate the general distribution of genome sizes and numbers of genes.
- To distinguish the characteristics of different types of genome organization in viruses, prokaryotes and eukaryotes.
- To know the evidence for the endosymbiotic origin of organelles.
- To recognize the effects of gene duplication on genome evolution.
- To be able to distinguish the meanings of homologue, orthologue, and paralogue.
- To understand the mechanism of genome change at the levels of individual bases, genes, chromosome segments and whole genomes.
- To understand the limits of what genomes determine and what they do not determine, and the limits of what we can currently explain on the basis of genetics and what we cannot.
- To appreciate as far as possible what makes us human.
- To understand the idea of a model organism in the study of human disease.
- To appreciate the goals and plans of the Encyclopedia of DNA Elements (ENCODE) project.

Introduction

It is likely that life originated on Earth about 3.5 billion years ago. The first cellular life forms were probably **prokaryotes**. **Eukaryotes** appeared about 2 billion years later. There are enough residual similarities among living things to suggest a common ancestor for us all. The great diversity of living forms is, therefore, the result of divergence.

Sequence analysis gives the most unambiguous evidence for the relationships among species. For higher organisms, sequence analysis and the classical tools of comparative anatomy, palaeontology and embryology usually give a consistent picture. Classification of microorganisms is more difficult, partly because it is less obvious how to select the features on which to classify them and partly because a large amount of lateral gene transfer threatens to overturn the picture of the evolutionary tree entirely.

In this chapter, we discuss general approaches to comparative genomics. From a human-centred view, we ask how to compare genomes in a way that illuminates our relationship with other species. In the next chapter, we extend this discussion to a more general view of the interaction and evolution of genomes.

Unity and Diversity of Life

The diversity of life fascinates everyone. The macroscopic life forms most familiar to us come in discrete types called species. (Why living things should be 'quantized' into discrete species is a very subtle question.) Linnaeus, an 18th-century Swedish naturalist, first organized the characteristics of different species into a logical framework. He introduced the system of nomenclature still used today. (We take up biological systematics in more detail in Chapter 3.)

Linnaeus classified living things according to a hierarchy: kingdom, phylum, class, order, family, genus and species. It usually suffices to specify the lowest two levels, as a **binomial**: genus and species – for instance, *Homo sapiens* for humans or *Drosophila melanogaster* for the fruit fly. Each binomial uniquely identifies a species that may also be known by one or more common names; for instance, *Bos taurus* = cow. Conversely, many common names refer to whole groups of species. For example, there are many species of whales, not all in the same genus or even the same family. Of course, most species do not have common names at all.

Classifications of humans and the fruit fly

	Human	Fruit fly
Kingdom	Animalia	Animalia
Phylum	Chordata	Arthropoda
Class	Mammalia	Insecta
Order	Primata	Diptera
Family	Hominidae	Drosophilidae
Genus	<i>Homo</i>	<i>Drosophila</i>
Species	<i>sapiens</i>	<i>melanogaster</i>

For macroscopic organisms, the Linnaean classification is reinterpretable as a phylogenetic tree – a set of ancestor–descendant relationships between species. A thread of continuous family history unites all organisms. But then there is a dissonance between the concept of a continuous evolutionary pathway from a common ancestor to daughter species and the idea that species are fundamentally discrete. It was Darwin’s observation of the diversity among the finches of the Galapagos Islands that provided the resolution of the paradox.

What is a species?

The concept of a species remains a useful one, even though it has proved very difficult to define precisely, even for macroscopic organisms. Many modern biologists define a species as a group of similar organisms that interbreed naturally to produce fertile offspring. When reproductive barriers arise – for instance, when groups of animals are trapped on different islands by rising sea levels – the species divides into two or more populations that, separately, interbreed only within themselves and maintain two gene pools. The separated populations may pursue different evolutionary paths and ultimately diverge to form separate species.

Genome sequences provide an alternative approach to definition of species. Sequences rule microbial taxonomy, but jostle for power with traditional morphological methods in the classification of plants and animals (see Box 2.1).

Microbiologists use Linnaean nomenclature for bacteria but find themselves uncomfortable doing so. Structural characteristics of bacteria lend themselves less well to distinguishing species than the physical features of higher organisms.

Traditional methods for bacterial classification were based on features of morphology (cell size and shape), biochemistry (uptake of stains, carbon and nitrogen sources, fermentation products) and physiology (growth temperature range and optimum, osmotic tolerance). Gram-positive and Gram-negative bacteria differ in their ability to take up crystal violet or methylene blue stain. Gram-positive bacteria contain a thick (20–80 nm) peptidoglycan layer in their cell wall that binds the stain. Immunological cross-reactivity has also been a

Geographical isolation is only one possible cause of speciation.

**BOX
2.1****DNA barcoding**

Field biologists now often characterize populations by DNA sequences. For higher animals, the sequence of the cytochrome *c* oxidase subunit I mitochondrial region (COI) provides a compact index for species identification. In most groups, this region is 648 bp long. The sequence variation within a species is small compared with differences between species.

Biologists describe the identification of species by the sequence of this region as 'barcoding'. The Barcode of Life database (BOLD) collects the information, currently covering 17 293 species. Its query system converts COI sequences to taxonomic assignments (<http://www.barcodinglife.org>).

Barcoding is a focus of the debate over traditional morphology-based taxonomy versus reliance on sequences. Of course, for classification of long-extinct organisms for which no sequences are available, there is no choice. And, despite their utility, barcodes tell us very little about the organisms they identify. No one would deny the phenotypic richness observable in living specimens – not least in their behaviour – that we cannot yet infer, even from complete genome sequences.

basis for classification, especially among infectious species that elicit a clinical motivation for their classification. Before sequencing, hybridization of DNA from two different bacteria was a criterion for similarity. Most bacterial DNAs will form hybrid double-helical structures provided that the similarity in base sequence is >80%.

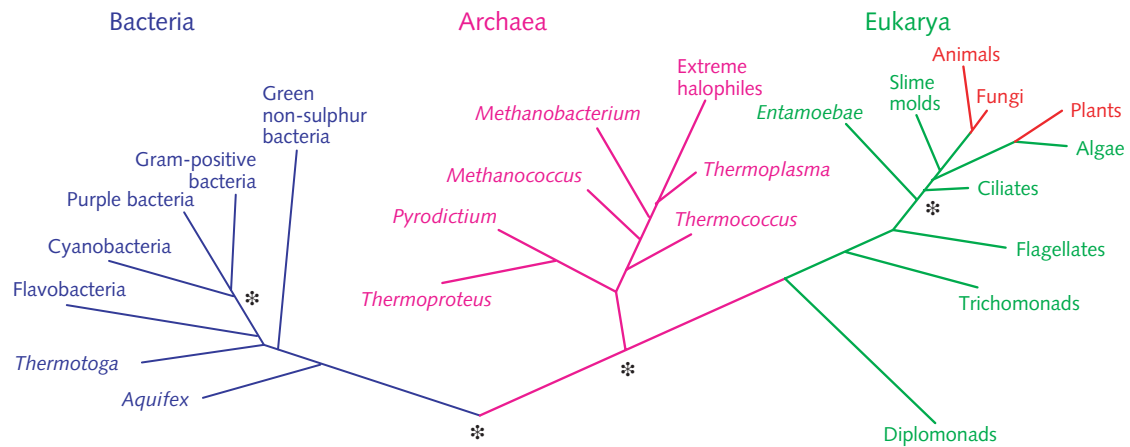


Figure 2.1 Major divisions of the tree of life. **Bacteria** (blue) and **archaea** (magenta) are prokaryotes; their cells do not contain nuclei. Bacteria include the typical microorganisms responsible for many infectious diseases and, of course, *Escherichia coli*, the mainstay of molecular biology. Archaea include, but are not limited to, extreme thermophiles and halophiles, sulphate reducers and methanogens. We ourselves are **eukarya** – organisms containing cells with nuclei (green and red). Asterisks mark crucial splitting points (see Exercise 2.3). This phylogenetic tree was derived by C. Woese from comparisons of ribosomal RNAs. These RNAs are present in all organisms, and show the right degree of divergence. (Too much or too little divergence and relationships become invisible.) Figure 2.2 shows in more detail the group that includes us – animals, fungi and plants (red).

Later, following seminal work by C. Woese, prokaryotic species were defined in terms of variations in 16S ribosomal RNA (rRNA) and other sequences. Bacteria for which the 16S rRNA sequences are more than about 2.5–3% different are considered different species. Typically, this corresponds to no more than 70% similarity in overall genome sequence. If humans and chimpanzees were bacteria, we would *easily* be considered as the same species!

Taxonomy based on sequences

Protein, RNA and DNA sequences have illuminated relationships between species, both for macroscopic organisms and microbes. The sequences have clarified some relationships but have exposed others as simplistic. Major results include the following.

- *All life on Earth has enough general similarity to show that all life forms had a common origin.* Evidence includes the universality of the basic chemical structures of DNA, RNA and proteins, and of their general biological roles and the near-universality of the genetic code.
- *On the basis of 16S rRNAs, C. Woese divided living things most fundamentally into three domains: bacteria, archaea and eukarya.* A domain occupies a level in the hierarchy *above* kingdom.

Figure 2.1 shows the major divisions of the tree of life. At the ends of the eukaryote branch are the metazoa, including yeast and all multicellular organisms – fungi, plants, and animals (see Figure 2.2). We and our closest relatives are in the vertebrate branch of the deuterostomes (see Figure 2.3).

Although archaea and bacteria are both unicellular organisms that lack a nucleus, at the molecular level archaea are in some ways more closely related to eukarya than to bacteria. It is also likely that the archaea are the closest living organisms to the root of the tree of life.

- *Dating of historical events from sequence differences.* As species diverge, their sequences diverge. L. Pauling and E. Zuckerkandl suggested that if sequence divergence occurred at a constant rate, it would provide a ‘molecular clock’ that would allow dating of the splits in lineage between species.

Although the clock is not universal, judicious calibration of rates of sequence change with palaeontological data permits dating of events in the history of life (see Box 2.2 and Figure 2.4).

Figure 2.4 shows major events in the history of life in the context of the standard geological time scale.

BOX 2.2

Molecular phylogeny and chronology

Molecular approaches to phylogeny developed against a background of traditional taxonomy, based on a variety of morphological characters, embryology, geographical distribution and, for fossils, information about the geological context (stratigraphy). The classical methods have some advantages. Traditional taxonomists have much less restricted access to extinct organisms via the fossil record. They can **date** the appearance and extinction of species by geological methods (see Figure 2.4).

Molecular biologists, in contrast, have very limited access to extinct species. Some sub-fossil remains of species that became extinct as recently as within the last two centuries have legible DNA, including specimens of the quagga (a relative of the zebra), the thylacine (Tasmanian 'wolf', a marsupial), the mammoth from the permafrost in Russia, the 'elephant bird' of Madagascar and some New Zealand birds, for instance, moas. It has been possible to sequence mitochondrial DNA from ~10 000-year-old remains of an 'Irish elk'. Some DNA sequences from Neanderthal man have been recovered from an individual who died approximately 30 000 years ago. But *Jurassic Park* remains fiction!

A crucial event in the acceptance of molecular methods occurred in 1967 when V.M. Sarich and A.C. Wilson dated the time of divergence of humans from chimpanzees at 5 million years ago, based on immunological data. At that time palaeontologists dated this split at 15 million years ago and were reluctant to accept the molecular approach. Reinterpretation of the fossil record led to acceptance of a more recent split and broke the barrier to general acceptance of molecular methods. It is now generally accepted that human and chimpanzee lineages diverged between ~6 and 8 million years ago.

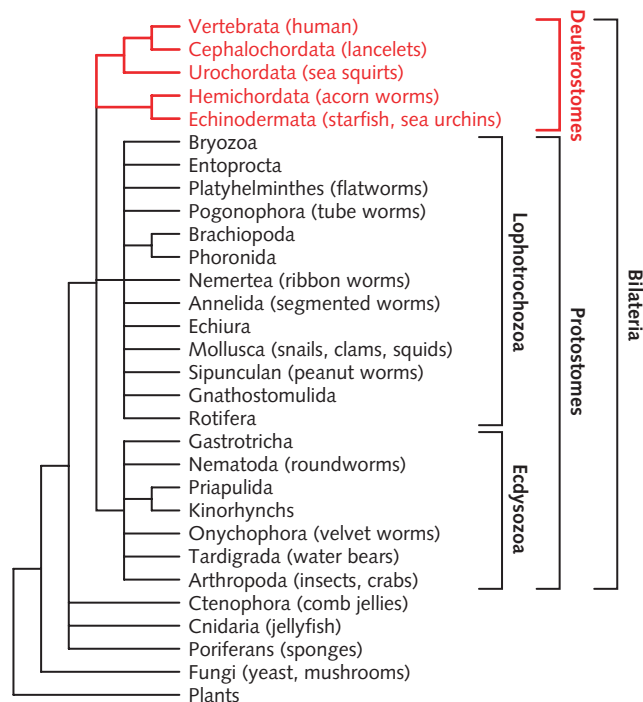


Figure 2.2 Phylogenetic tree of metazoa (multicellular animals). **Bilateria** include all animals that share a left–right symmetry of body plan. **Protostomes** and **deuterostomes** (red) are two major lineages that separated at an early stage of evolution, estimated at 670 million years ago. They show very different patterns of embryological development, including different early cleavage patterns, opposite orientations of the mature gut with respect to the earliest invagination of the blastula, and the origin of the skeleton from mesoderm (deuterostomes) or ectoderm (protostomes). Protostomes comprise two subgroups distinguished on the basis of the sequences of an RNA from the small ribosomal subunit and *HOX* genes. (*HOX* genes govern the development of body plans.) Morphologically, **ecdyssozoa** have a moulting cuticle – a hard outer layer of organic material. **Lophotrochozoa** have soft bodies. Figure 2.3 shows in more detail the group that includes us – the deuterostomes (red).

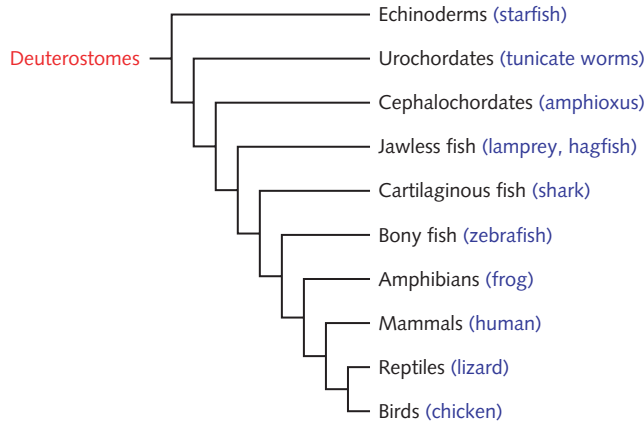
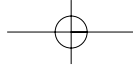


Figure 2.3 Phylogenetic tree of vertebrates and our closest relatives. Chordates, including vertebrates, and echinoderms are all deuterostomes. Examples of each are shown in blue.

Time scale of Earth history

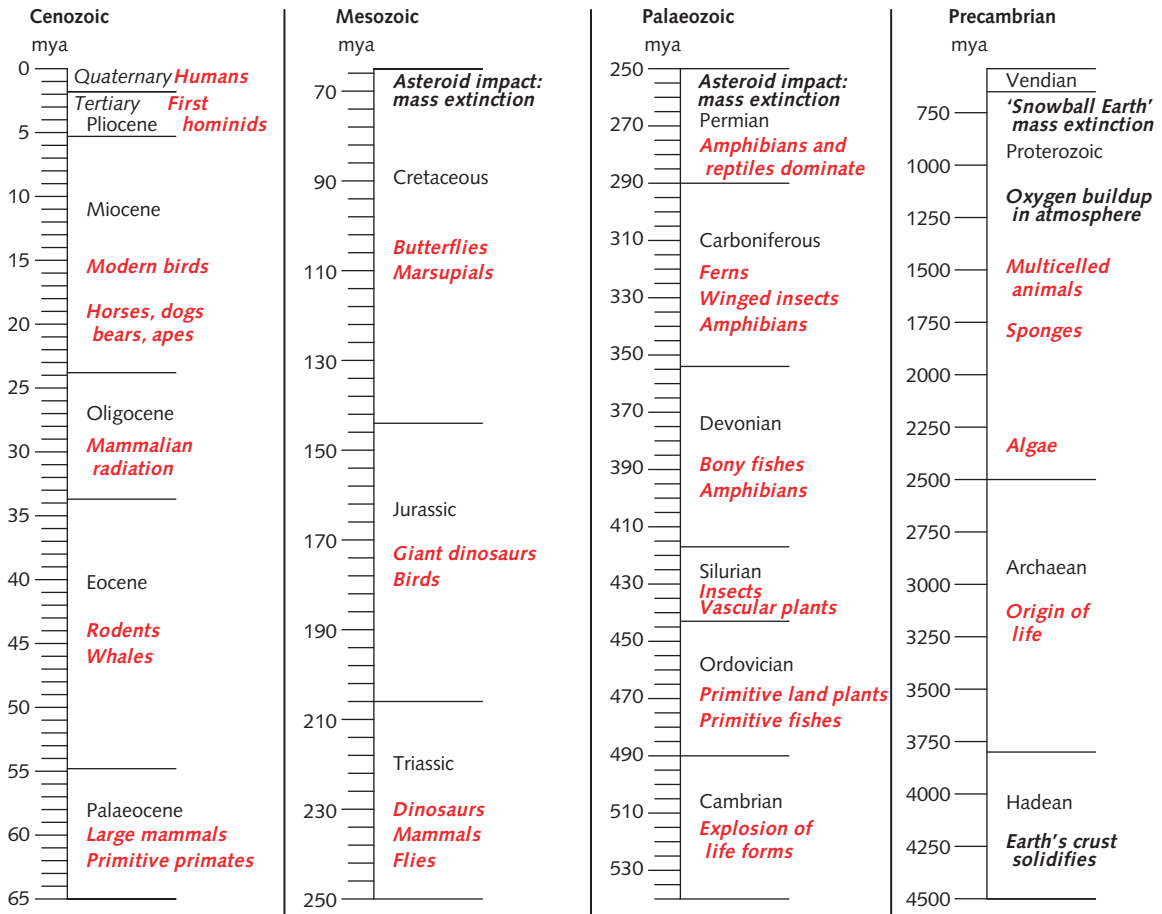
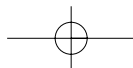


Figure 2.4 Geological time divisions and cataclysmic events (e.g. asteroid impact: mass extinction) are shown in black. The first appearance of, or prevalence of, different life forms is shown in red. mya, millions of years ago.



- *The importance of horizontal gene transfer.* This is the acquisition of genetic material by one organism from another by natural rather than laboratory procedures through some means other than descent from a parent during replication or mating (see Box 2.3). Several mechanisms of horizontal gene transfer are known, including direct uptake, as in Griffith's pneumococcal

**BOX
2.3****Please pass the genes: horizontal gene transfer**

On learning that *Streptomyces griseus* trypsin is more closely related to bovine trypsin than to other microbial proteinases, Brian Hartley commented in 1970 that '... the bacterium must have been infected by a cow'. This was a clear example of lateral or horizontal gene transfer – a bacterium picking up a gene from the soil in which it was growing, and which an organism of another species had deposited there. The classic experiments on pneumococcal transformation by O. Avery, C. MacLeod and M. McCarthy that identified DNA as the genetic material are another example.

Evidence for horizontal transfer includes (1) discrepancies among evolutionary trees constructed from different genes; and (2) direct sequence comparisons between genes from different species.

- In *Escherichia coli*, about 25% of the genes appear to have been acquired by transfer from other species.
- In microbial evolution, horizontal gene transfer is more prevalent among operational genes – those responsible for 'housekeeping' activities such as biosynthesis – than among informational genes – those responsible for organizational activities such as transcription and translation. For example:
 - *Bradyrhizobium japonicum*, a nitrogen-fixing bacterium, symbiotic with higher plants, has two glutamine synthetase genes: one is similar to those of its bacterial relatives; the other is 50% identical to those of higher plants;
 - rubisco (ribulose-1,5-bisphosphate carboxylase/oxygenase), the enzyme that first fixes carbon dioxide at entry to the Calvin cycle of photosynthesis, has been passed around between bacteria, mitochondria and algal plastids, as well as undergoing gene duplication.
 - many phage genes appearing in the *E. coli* genome provide further examples and point to a mechanism of transfer.

Nor is the phenomenon of horizontal gene transfer limited to prokaryotes. Both eukaryotes and prokaryotes are chimaeras. Eukaryotes derive their informational genes primarily from an organism related to *Methanococcus* and their operational genes primarily from proteobacteria, with some contributions from cyanobacteria and methanogens. Almost all informational genes from *Methanococcus* itself are similar to those in yeast. At least eight human genes appeared in the *Mycobacterium tuberculosis* genome. *S. griseus* trypsin is an example of eukaryote → prokaryote transfer.

The observations hint at the model of a 'global organism', or a genomic World Wide DNA Web from which organisms download genes at will! How can this be reconciled with the fact that the discreteness of species has been maintained? The conventional explanation is that the living world contains ecological 'niches' to which individual species are adapted. It is the discreteness of niches that explains the discreteness of species. But this explanation depends on the stability of normal heredity to maintain the fitness of the species. Why would the global organism not break down the lines of demarcation between species, just as global access to pop culture threatens to break down lines of demarcation among national and ethnic cultural heritages? Perhaps the answer is that it is the informational genes, which appear to be less subject to horizontal transfer, that determine the identity of the species.

transformation experiments, or via a viral carrier. Arrangements of species into phylogenetic trees, in contrast, assumes strict ancestor–descendant relationships between different organisms during evolution.

Horizontal gene transfer among different species has affected most genes in prokaryotes. It requires a change in our thinking from ordinary ‘clonal’ or parental models of heredity. Microorganisms do not easily fit into the structure of the ‘tree’ of life but require a more complex organizational chart.

Sizes and Organization of Genomes

We appeal to genomes to help us to understand ourselves as individuals, and our relationships with all of the other organisms that march in the pageant of life. To make progress, we must integrate several data streams, including:

- genome sequences;
- RNA and protein expression patterns;
- the spatial organization of individual macromolecules, their complexes, organelles, entire cells, tissues and bodies; and
- regulatory networks, the internal structure and logic of adaptive control systems.

Even these may not be enough. History – sometimes observed but more usually inferred – provides essential additional clues. We see, today, a snapshot of one stage in a history of life that extends back in time for at least 3.5 billion years. We must try to read the past in contemporary genomes, which contain records of their own development.

This programme requires development of novel methods. New fields of study require new approaches. S. Luria once suggested that to determine common features of all life one should not try to survey everything, but, rather, identify the organism most different from us and see what we have in common with it. Let us add a complementary idea: to take the most *closely related* organisms and identify the *differences*. That is:

- How do the human genome and the *E. coli* genome express our *common* heritage?
- How do genomes that are over 98% identical create the *differences* between humans and chimpanzees?

If we could answer these questions, we would have achieved a lot.

Genome sizes

One reason for resistance to Darwin’s theory of evolution was its denial to human beings of a special status relative to animals. Genomics threatens

US Supreme Court Justice Felix Frankfurter wrote that ‘. . . the American constitution is not just a document, it is a historical stream.’ Like a genome!

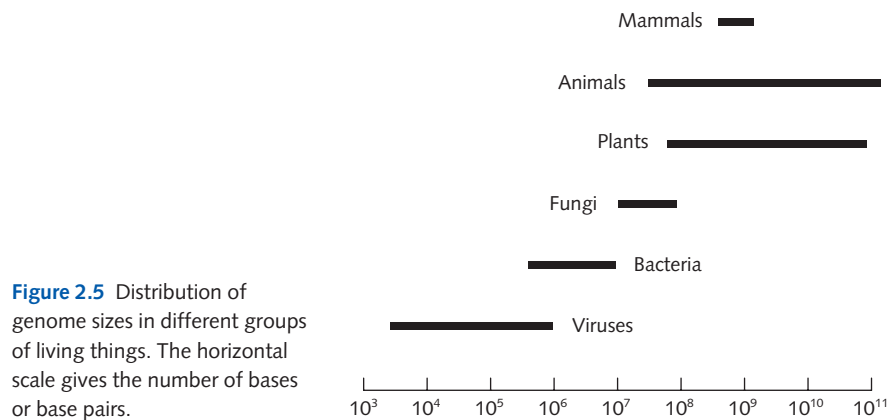


Figure 2.5 Distribution of genome sizes in different groups of living things. The horizontal scale gives the number of bases or base pairs.

The term *C-value* has been used to refer to the amount of DNA in a haploid cell, i.e. a gamete; the letter *C* refers to the **constancy** of the amount of DNA per cell in a species.

to do this all over again. Humans do have unique features. Many people, not excluding molecular biologists, expect these features to be reflected in the genome. And so they must be, although, frankly, not in any obvious way.

The overall size of the human genome is not special. Different organisms have different total amounts of DNA per cell (see Figure 2.5 and Table 2.1):

There is a general correlation between complexity of organism and amount of DNA per cell. Prokaryotes have less DNA per cell than eukaryotes and yeast has less than mammals. However, although humans have more DNA per cell than certain other organisms popular in molecular biology, including *Caenorhabditis elegans* and the fruit fly, many organisms have even greater amounts than we do. The genome of *Amoeba dubia* is 200 times larger than the human genome. The genome of the marbled lungfish (*Protopterus aethiopicus*) a closer relative, is 43 times as large as ours.

Why the different amounts of DNA? Most of the human genome does not encode protein or structural RNA. Regions of genomes without known function are often referred to as 'junk DNA'. Of course, the fact that we may not know the function of much of our genome does not mean that it has none. (Maybe it is junk, but it is certainly not all transcriptionally inert.) Moreover, the amount of space between genes affects the rate of crossing over and recombination and, thereby, rates of evolution. Indeed, the large amount of repetitive sequence between our genes enhances recombination rates by promoting homologous recombination. Rate of evolutionary change is a characteristic of a species that is certainly subject to selective pressure. Features of the genome that affect rate of evolution cannot be dismissed entirely as junk.

If genome size *per se* does not single out humans, what about numbers of genes? Again there is a general correlation between complexity of organism and estimated numbers of genes. Viral genomes encode only a few proteins. Prokaryote genomes contain hundreds or thousands of genes. The simple eukaryote yeast has almost 6000 genes, fewer than twice as many as *E. coli*. Metazoa have tens of thousands of genes.

Table 2.1 Genome sizes

Organism	Number of base pairs	Number of genes	Comment
ϕ X-174	5 386	10	Virus infecting <i>E. coli</i>
Influenza A	13 590	10	Strain A/goose/guangdong/1/96(H5N1)
Human mitochondrion	16 569	37	Subcellular organelle
Epstein–Barr virus (EBV)	172 282	80	Cause of mononucleosis
<i>Nanoarchaeum equitans</i>	490 885	552	Archaeon, smallest known genome of a cellular organism
<i>Mycoplasma pneumoniae</i>	816 394	680	Cause of cyclic pneumonia epidemics
<i>Rickettsia prowazekii</i>	1 111 523	834	Bacterium, cause of epidemic typhus
Mimivirus	1 181 404	1 262	Virus with the largest known genome
<i>Borrelia burgdorferi</i>	1 471 725	1 738	Bacterium, cause of Lyme disease
<i>Aquifex aeolicus</i>	1 551 335	1 749	Bacterium from hot spring
<i>Thermoplasma acidophilum</i>	1 564 905	1 509	Archaeal prokaryote, lacks cell wall
<i>Helicobacter pylori</i>	1 667 867	1 589	Chief cause of stomach ulcers
<i>Methanococcus jannaschii</i>	1 664 970	1 783	Archaeal prokaryote, thermophile
<i>Haemophilus influenzae</i>	1 830 138	1 738	Bacterium, cause of middle-ear infections
<i>Thermotoga maritima</i>	1 860 725	1 879	Marine bacterium
<i>Archaeoglobus fulgidus</i>	2 178 400	2 437	Another archaeon
<i>Deinococcus radiodurans</i>	3 284 156	3 187	Radiation-resistant bacterium
<i>Synechocystis</i>	3 573 470	4 003	Cyanobacterium, 'blue-green alga'
<i>Vibrio cholerae</i>	4 033 460	3 890	Cause of cholera
<i>Mycobacterium tuberculosis</i>	4 411 532	3 959	Cause of tuberculosis
<i>Bacillus subtilis</i>	4 214 814	4 779	Popular in molecular biology
<i>Escherichia coli</i>	4 639 221	4 377	Molecular biologists' all-time favourite
<i>Saccharomyces cerevisiae</i>	12 495 682	5 770	Yeast, first eukaryotic genome sequenced
<i>Caenorhabditis elegans</i>	100 258 171	19 099	'The worm'
<i>Arabidopsis thaliana</i>	115 409 949*	25 498	Flowering plant (angiosperm), 'the weed'
<i>Drosophila melanogaster</i>	122 653 977	13 472	The fruit fly
<i>Takifugu rubripes</i>	3.65×10^8	~38 000	Pufferfish (fugu fish)
Human	3.3×10^9	25 000?	
Wheat	16×10^9	30 000	
Salamander	10^{11}	?	
<i>Psilotum nudum</i>	2.5×10^{11}	?	Whisk fern, a simple plant
<i>Amoeba dubia</i>	6.7×10^{11}	?	Protozoan

*An alternative estimate suggests that the size of the *A. thaliana* genome is 157 Mb.

Species	Genome size (Mb)	Coding (%)	Approximate number of genes	Estimated gene density (kb/gene)
<i>E. coli</i>	4.64	88	4 300	0.95
Yeast	12.5	70	6 000	2.1
Puffer fish	365	15	30 000	10
<i>A. thaliana</i>	115	29	25 000	4.5
Human	3289	1.3	30 000	27

However, within groups of related organisms, including vertebrates, there is no simple correlation between apparent complexity of organism, or even genome size, and numbers of genes. Two vertebrates, the puffer fish and humans, appear to have roughly the same number of genes but differ by almost an order of magnitude in genome size. It was also unexpected to find that the worm *C. elegans* appears to have more genes than the fruit fly.

The phenomenon of alternative splicing shows the situation to be more complicated than simple gene estimates make it appear. (Alternative splicing is the synthesis of different proteins from the same RNA transcript by joining different sets of exons.) This is one reason why it has been difficult to get an accurate count of the number of genes in humans and other higher organisms. In eukaryotes, estimates of gene number refer to maximal sets of exons in units that are coordinately transcribed and translated. In fact, variation in splicing may create many proteins from each gene. It is, therefore, very different to estimate the size – to say nothing of the complexity – of a eukaryote's proteome from its genome. As an extreme example, in the mammalian immune system, billions of distinct antibodies arise from regions in the genome containing fewer than ~100 exons.

The basis of the complexity of expression patterns, metabolic activity and indeed all other phenotypic features is the organization of the genome itself. Different types of organism have experimented with different solutions of the problems of packaging long, narrow strands of DNA and of controlling access of transcriptional machinery to different regions.

Even alternative splicing gives only a static idea of proteome complexity. Cells control gene expression patterns by complex and dynamic regulatory networks. Conclusion: it is difficult to correlate numbers of expressed genes with organismal complexity if one has no good way of measuring either.

Viral Genomes

Viruses infect cells using specialized proteins on their surfaces that effect attachment and invasion. Viral nucleic acid enters the host cell. In some cases, viral proteins required for replication also enter the host cell. Once inside the host cell, the invading viral molecules must (1) make multiple copies of the viral genome; (2) synthesize viral proteins, including both enzymes active only within

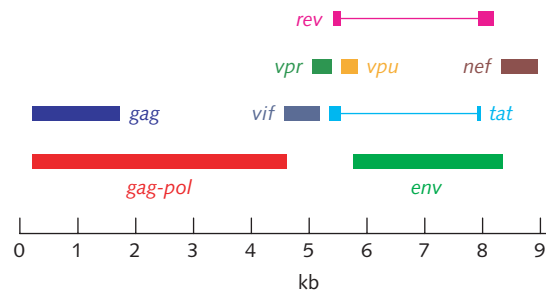


Figure 2.6 Diagram showing the sizes of the individual gene transcripts of HIV-1. The introns of the *rev* and *tat* genes are indicated by a thin line. The proteins encoded by these genes are:

Gene	Proteins	Function
<i>gag</i>	p24, p6, p7, p17	Structural proteins of capsid and matrix
<i>pol</i>	Reverse transcriptase, integrase, protease	Integration into host genome and cleavage of viral-encoded polyproteins
<i>env</i>	Precursors of gp120, gp41	Envelope proteins, active in attachment and fusion to host cells
<i>tat</i>	Tat	Facilitates transcription of viral RNA
<i>rev</i>	Rev	Enhances cytoplasmic export of transcripts
<i>nef</i>	Nef	Interferes with host immune function
<i>vif</i>	Vif	Interferes with host defence
<i>vpr</i>	Vpr	Needed for nuclear import of viral nucleic acid
<i>vpu</i>	Vpu	Promotes assembly and release of progeny virus; also stimulates degradation of host CD4 proteins, disabling the host immune system

the host cell and coat proteins (and others) to be assembled into the progeny virions; and then (3) ‘pack up and leave’.

Viral genomes contain only relatively short stretches of nucleic acids. Some, such as the virus that causes hepatitis C, encode a single polyprotein, cleavage of which produces the few proteins the virus needs to take over the cell. Other viruses, such as human immunodeficiency virus type 1 (HIV-1), contain several genes. The HIV-1 genome is about 9.8 kb long, containing a total of nine genes (see Figure 2.6, and Box 2.4). One gene encodes the Gag–Pol fusion protein, which is cleaved to release Gag (the HIV-1 protease), reverse transcriptase and integrase. Other mRNAs expressed by HIV-1 contain introns and are spliced to express Rev and Tat.

Recombinant viruses

Mixed infections of a cell by different viruses permit genetic recombination. It is even possible to package *unaltered* nucleic acid from one strain into an envelope composed of protein from another strain. In that event, the absorption–penetration–surface antigenicity characters are those of the coat proteins, but the hereditary characteristics are those of the nucleic acid. (Infection with such

**BOX
2.4****Types of viral genome**

As assembled within the virion, a viral genome may consist of:

Nucleic acid

- | <i>Nucleic acid</i> | <i>Examples</i> |
|-----------------------|---|
| • Single-stranded DNA | Bacteriophages ϕ X-174 and M13 |
| • Double-stranded DNA | Adenoviruses, smallpox virus, Epstein-Barr virus, bacteriophage λ |
| • Single-stranded RNA | Bacteriophages MS2, Q β , tobacco mosaic virus, HIV-1 |
| • Double-stranded RNA | Bluetongue virus |

Single-stranded DNA viral genomes are generally converted to double-stranded DNA by the host. Replication of RNA viruses is prone to mutation because the error-correction mechanisms active in host DNA replication do not apply. This helps viruses to evade host immune systems and facilitates their jumping between host species. (Emerging viral diseases, including but not limited to acquired immunodeficiency syndrome (AIDS) and avian flu, are usually based on viruses with RNA genomes.)

A viral genome consisting of single-stranded RNA can be:

- (+)sense = same sequence as protein-translatable mRNA
 (-)sense = complementary sequence to mRNA
 ambisense = mixture of both.

Inside the cell, (+)sense viral RNAs present themselves as messenger RNA (mRNA) and are translated. (-)Sense viral RNAs and double-stranded viral RNAs require specialized polymerases for conversion to mRNA. Retroviral genomes contain (+)sense RNA, which is reverse transcribed into host DNA. These viral polymerases and reverse transcriptases are proteins that are contained in the infecting virion and enter the host cell along with the viral nucleic acid.

Some viral genomes are infectious on their own. For some RNA viruses, a DNA reverse transcript of the viral RNA is infectious (although at a lower rate than the natural virion). This permits preparation of large quantities of viral genomes for vaccines, avoiding the lability and high mutation rate of viral RNA replication.

Both HIV-1 and influenza viruses have become major threats to human health after jumping from animal hosts.

a virus is a kind of natural Hershey-Chase experiment; see p. 17.) These effects can alter host specificity.

In the laboratory, a virus can be constructed as a vector to produce foreign proteins inside a cell. Two applications of this technique are as follows.

1. To produce a vaccine, insert a DNA sequence coding for the immunogen (perhaps the HIV-1 surface glycoprotein gp120) into the vaccinia virus genome. (The HIV-1 protein itself is of course not infectious: the Hershey-Chase experiment again!) Infection by recombinant virus leads to expression of immunogen and elicitation of an immune response, giving the host protective immunity. The immunity created by such an *intracellular* exposure to the immunogen is much more powerful than that achievable simply by injecting the immunogen into the bloodstream.
2. A recombinant virus carrying a normal human gene can be useful for gene therapy. Cystic fibrosis arises from a mutation in the cystic fibrosis transmembrane regulator (*CFTR*) gene. Can a viral vector reintroduce the normal variant into the patient's genome? Recombinant adenoviruses containing the functional version of the (*CFTR*) gene are undergoing clinical trials.

Influenza: a past and current threat

Influenza is a contagious disease caused by a virus that infects the respiratory tract. The virus is passed around a population in droplets created when an infected individual coughs or sneezes. Unlike HIV-1, the virus that causes AIDS, influenza virus can survive outside the host, greatly facilitating its transmission.

Every year, influenza seasonally affects many people worldwide. In a typical year, 38 000 people die in the USA from influenza or related complications and 200 000 are hospitalized. The mortality rate is 0.8%, with most fatalities occurring in the very young or elderly. Worldwide, the usual annual fatality rate is 1–1.5 million.

However, in some years, influenza and associated complications attack more viciously. A famous pandemic occurred at the end of the First World War: within an 18 month period in 1918–1919, influenza killed an estimated 50–100 million people, far more than had died in the war. The mortality rate was 1% of those affected. Such an influenza pandemic today could have an even higher mortality in regions of the world containing many people immunocompromised by AIDS. (The human mortality rate of the current avian strain is over 50%!)

In most influenza seasons, fatalities occur as a result of bacterial infection of lungs weakened by the virus, to which the elderly are more vulnerable. The 1918–1919 epidemic was different, in the higher percentage of fatalities among *young* people. Several explanations have been offered including: (1) the high density of young soldiers in military camps and battlefields, leading to more effective transmission; (2) overcrowding and poor nutrition and health care among refugees; and (3) previous epidemics in the 1850s and 1889, leaving many elderly people with some immunity.

Three types of influenza virus are known, of which type A is the most dangerous. The virion contains a spherical lipoprotein coat enclosing eight nucleoproteins containing the RNA genome, encoding a total of ten proteins. Protruding from the envelope are several hundred ‘spikes’ containing the proteins haemagglutinin (80% of the spikes) and neuraminidase (about 20%) (see Figure 2.7). These proteins are essential to the reproduction of the virus. Haemagglutinin binds to host cell surface glycoprotein receptors to promote viral entry into cells, whereas neuraminidase helps progeny virions to get out. Both haemagglutinin and neuraminidase are targets of drugs.

The virus can evolve by point mutations, also called antigenic drift, or by genetic recombination. Immunologically distinct strains of viruses are called serotypes. Different serotypes vary both in the ease of their spreading and in the mortality of infection. For instance, binding of virus to mucosal respiratory surfaces may be affected by amino acid sequence polymorphisms of both viral proteins and host receptors. Contagion and mortality are also dependent on characteristics of the host population, including density and general health levels. A strain that is both highly contagious and has a high mortality rate would be very dangerous indeed. As part of an effort to understand why the 1918–1919

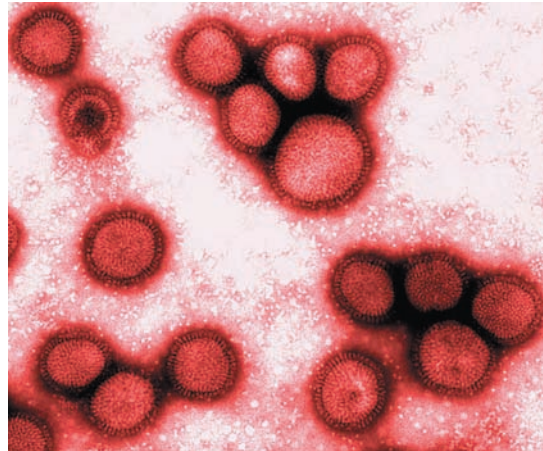


Figure 2.7 Influenza virus.

Picture courtesy Professor Y. Kawaoka, University of Wisconsin, USA, and University of Tokyo, Japan.

strain was so dangerous, scientists have recently reconstructed that virus, based on material recovered from contemporary postmortem specimens.

Different strains of influenza virus contain 1 of 13 recognized types of gene for haemagglutinin (H) and one of nine recognized types of gene for neuraminidase (N). These types identify different major strains of viruses. For instance, the strain that caused the 1918–1919 pandemic was H1N1. The current avian flu virus is H5N1.

Many strains of the virus infect only a restricted range of species and with different mortality rates. These properties can change as the virus evolves. A strain that infects animals can potentially become infectious to humans. Species range depends on the different forms of sialic acid presented on viral glycoproteins. An important determinant is haemagglutinin residue 226, which is Gln in viruses infectious to birds and Leu in viruses infectious to humans.

Avian flu

In 2006, an H5N1 strain of avian flu characterized by very high mortality infected domestic poultry in several countries. It is considered a particularly dangerous threat to humans because it has a high mutation rate and recombines readily. One way for the virus to jump from birds to humans is for two strains to co-infect pigs and use them as a ‘mixing vessel’ for recombination.

Avian flu can normally infect only birds and, in some cases pigs. Domestic poultry stocks raised under conditions of very high population density are particularly vulnerable. Often migratory birds are carriers but do not get sick. (The current avian flu strain has been spread from southeast Asia to Russia by migratory birds.)

The H5N1 strain prevalent in 2006 was first identified in Hong Kong in 1997 and traced to ducks from Guandong province. It jumped the species barrier to mammals, becoming infectious to pigs, in April 2004. It then became super-virulent, killing rodents, birds and humans. This H5N1 strain is 100% fatal in

domesticated chickens and in 54% of reported human cases. So far human to human transmission is uncommon.

Compared with previous epidemics, the world today is particularly vulnerable because of:

- increased human population densities;
- widespread long-distance travel;
- intensive livestock production (including antibiotic feeding, which may create drug-resistant strains of infectious bacteria); and
- policies of various governments that do not adequately reimburse farmers who must sacrifice animals, creating a disincentive to report disease; the result is a delay or even default of an effective response.

Increased population densities of both humans and animals threaten a greater rate of spread of a dangerous strain of virus, even in comparison with the recent 1968–1969 epidemic:

Year	Population in China (millions):		
	Humans	Pigs	Poultry
1968	790	5.2	12.3
2005	1300	508	13 000

Aggressive approaches to controlling avian flu have involved large-scale culling of stocks. In 1997, the H5N1 strain infected poultry in Hong Kong and caused six human fatalities. The entire poultry population of the island had to be destroyed: 1.5 million birds in three days. An H7N7 epidemic in the Netherlands in 2003 led to the killing of >30 million birds (approximately twice the human population of the country). In Asia in 2004, over 100 million birds were culled.

In birds, the pandemic is already here.

Drugs against influenza

Tamiflu (oseltamivir) and Relenza (zanamivir) are the two major drugs against influenza. Both are inhibitors of the viral neuraminidase.

Relenza (see Figure 2.8) was designed at the Commonwealth Scientific and Industrial Research Organization (CSIRO) laboratory in Melbourne, Australia. Crystal structures of influenza neuraminidase showed that conserved sequences formed a cavity, suggesting a target site for drugs. By targeting the active site of the enzyme, it is harder for the virus to evolve resistance.

Ethical dilemma: publication of reconstructed virulent 1918–1919 strain

Recently, scientists were able to recover and sequence the strain of influenza active in the 1918–1919 pandemic. The journal *Science* published the work and, consistent with editorial policy, required that the sequence be deposited in the nucleic acid databanks.

In *The New York Times* on 17 September 2005, R. Kurzweil and W. Joy wrote an article critical of the decision to make the sequence generally available

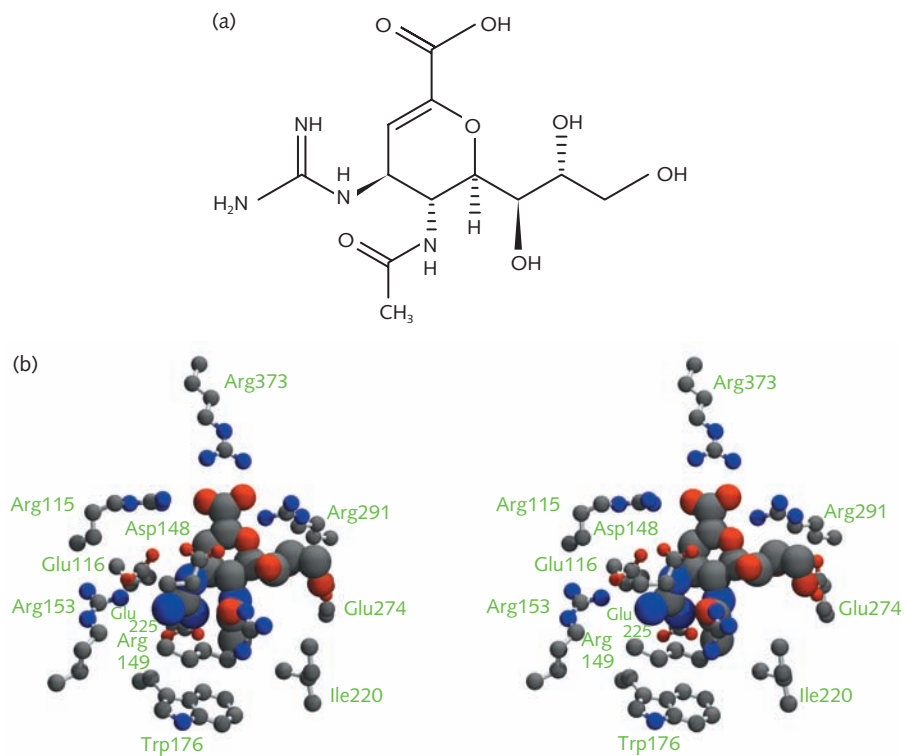


Figure 2.8 (a) The structure of the anti-influenza drug zanamivir (Relenza). (b) Zanamivir is a transition-state analogue that binds to the active site of influenza neuraminidase. Here the atoms of the drug are shown as large spheres and the residues from the neuraminidase are shown in ball-and-stick representation. [1A4G]

in databanks on the grounds that terrorists might use the information to recreate the virus and use it as weapon.

Reactions to the publication of the reconstructed pandemic viral sequence illustrate the conflict between the recognition that free and open access to information is a benefit to the progress of science, and the dangers of its misuse. A precedent occurred before the Second World War when physicist Leo Szilard tried, unsuccessfully, to persuade colleagues not to publish results that might prove useful in the development of atomic weapons. He suggested that journals record dates of receipt and acceptance of manuscripts but then sequester the articles for the duration. This occurred well before the strict secrecy imposed after the Manhattan Project was organized.

Science did make an exception to its mandatory-deposition policy in publishing the Human Genome Draft Sequence by J.C. Venter and co-workers in 2001.

Genome Organization in Prokaryotes

A typical prokaryotic genome has the form of a single circular molecule of double-stranded DNA, between 0.6 and 10 million bp long. For instance, a cell of *E. coli* strain K12 contains a single molecule of double-stranded DNA

4 639 675 bp long, closed into a circle. The DNA is supercoiled and associated with histone-like proteins into a 'chromosome', appearing in a subcellular structure called the nucleoid. Some *E. coli* cells may contain **plasmids**: short, usually circular, double-stranded DNA molecules, ranging from 1 kb to several megabases in length.

Although single circular genomes containing most of the DNA are common in bacteria and archaea, many exceptions are known. Many prokaryotic cells contain plasmids. Some prokaryotes have linear DNA. *Borrelia burgdorferi*, the organism that causes Lyme disease, is an example. *B. burgdorferi* also contains numerous plasmids, some of which are circular and some linear. Other prokaryotes contain more than one chromosome. *Vibrio cholerae*, the organism that causes cholera, contains two circular DNA molecules of 2 961 146 and 1 072 314 bp.

Some but not all prokaryote genomes contain **insertion sequences**, mobile genetic elements similar to eukaryotic transposons.

The 4.6 Mb chromosome of *E. coli* encodes approximately 4400 genes, distributed on both strands. The absence of introns and the shorter intergenic regions account for the higher coding densities. A very large fraction of the DNA, 87.8%, codes for proteins, 0.8% codes for structural RNAs and only 0.7% has no known function (see Figure 2.9).

Species	Coding	Average gene density
<i>E. coli</i>	>90%	1 gene/kb
Pufferfish	15%	1 gene/10 kb
Human	5%	1 gene/30 kb

Many prokaryotic genomes have been sequenced. They illuminate, in a somewhat simpler context than the human genome, how these organisms solve problems common to all cellular life forms. In addition, there are good practical motives for studying features of prokaryotes. Differences between prokaryotic

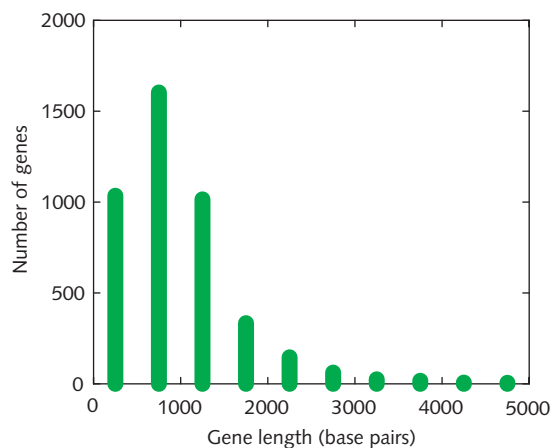
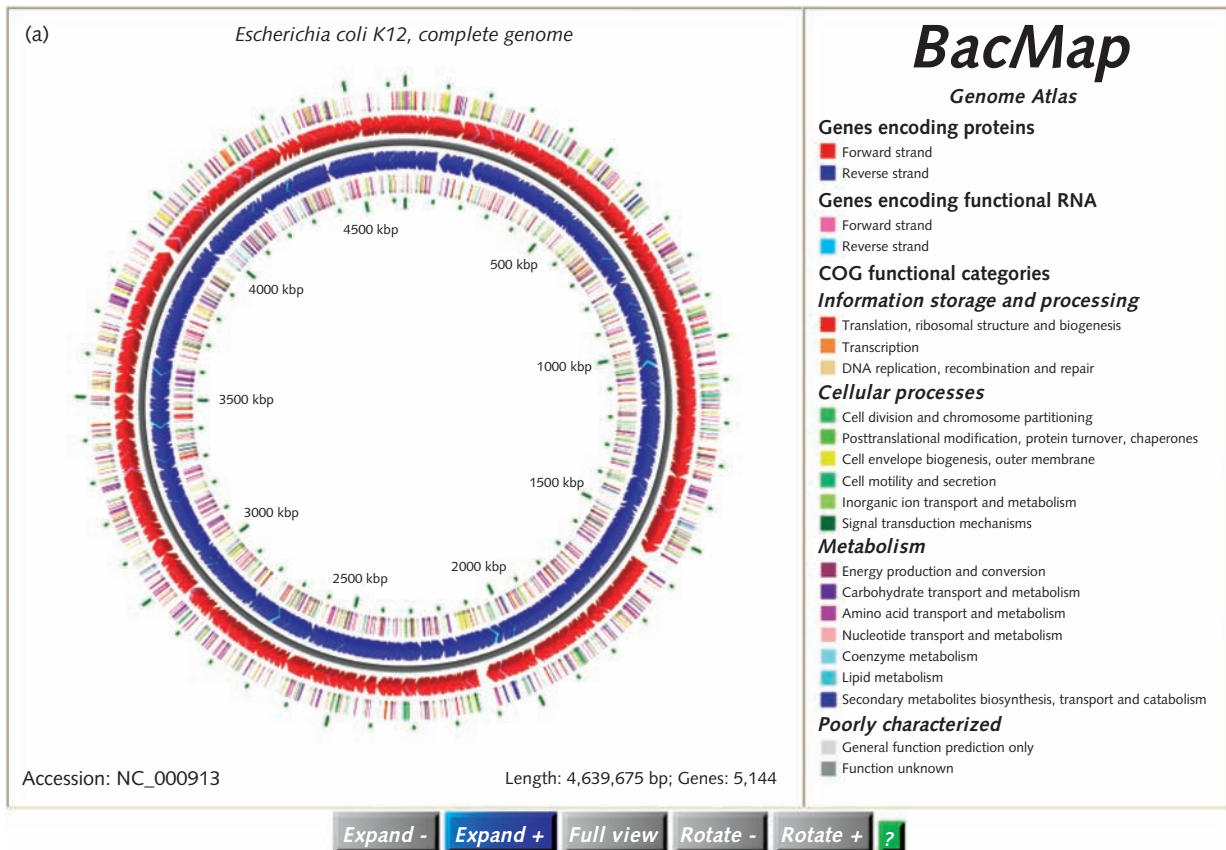


Figure 2.9 Distribution of gene lengths in *E. coli*. Two very long genes for hypothetical proteins, *yeeJ* and *ydbA*, of length 7152 and 8619 bp, respectively, are omitted. The average gene length is 960 bp. Most genes are less than 1500 bp long.



Click tick marks to expand the view.

Valid XHTML 1.0; Valid CSS.

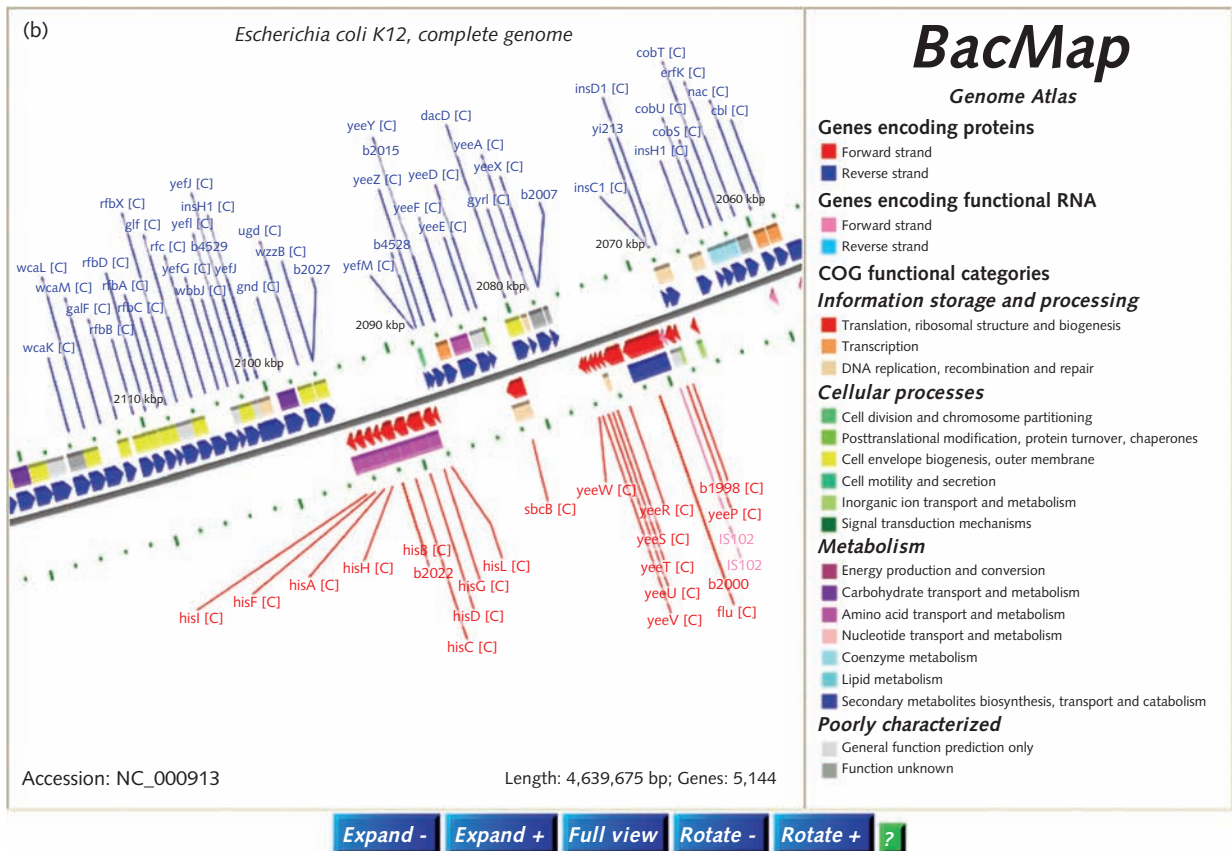
Displayed PNG file size: 186 kb.
Centered on base 1; Zoom = 1.

Figure 2.10 Map of the genome of *E. coli* K12. (a) Full view. Red arrows show protein-coding regions of the forward strand. Blue arrows show protein-coding regions of the reverse strand. Pink arrows show structural RNA-encoding regions of the forward strand. Cyan arrows show structural RNA-encoding regions of the reverse strand. Radial ticks identify individual gene products, colour coded according to function. COG categories refer to the Clusters of Orthologous Groups database (<http://www.ncbi.nlm.nih.gov/COG/>). (b) Expanded view of the region containing the *his* operon. The BacMap site provides access to genomes of bacteria and archaea (<http://wishart.biology.ualberta.ca/BacMap/>).

Pictures reproduced, by permission, from BacMap: An Interactive Atlas for Exploring Bacterial Genomes. See: Stothard, P., Van Domselaar, G., Shrivastava, S., Guo, A., O'Neill, B., Cruz, J., Ellison, M. & Wishart, D.S. (2005). BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucl. Acids Res.* 33, D317–D320.

and eukaryotic metabolism – enzymes unique to prokaryotes – are appropriate targets for drugs against infection. Of great importance to clinical medicine is understanding how prokaryotes evolve to develop pathogenicity and antibiotic resistance.

We visualize the contents of bacterial chromosomes as concentric circular diagrams, looking vaguely like ‘tie-dyed’ patterns (see Figure 2.10).



Click tick marks to expand the view.

Valid XHTML 1.0; Valid CSS.

Displayed PNG file size: 84 kb.
Centered on base 2,090,000; Zoom = 36.

Replication and transcription

In *E. coli*, replication begins at a specific site called *oriC* and proceeds in both directions. This site is the calibration point from which the genome is indexed. Replication ends at the *terC* site, found almost, but not exactly, half way around the circle. (In contrast, archaea often have multiple sites of origin of replication.)

In prokaryotes, many mRNA transcripts contain several tandem genes, which require separate initiation of translation. (In this, they are unlike viral polyproteins, which are translated in one piece and then cleaved.) In bacteria, but less frequently in archaea, co-transcribed genes have related functions, forming an 'operon'.

Timing illuminates the interrelationships among these processes. Under ordinary conditions, it takes *E. coli* 40 minutes to replicate its genome. The full generation time between cell divisions is about an hour. This explains why genes that require high rates of expression tend to be near the origin of replication: the

availability for transcription of partially replicated DNA in effect increases the copy number of such genes. Conversely, the half-life of mRNA is only a few minutes. Therefore, translation must overlap transcription.

Gene transfer

There are three methods of transfer of DNA between prokaryotic cells.

- *Transformation.* The uptake of ‘naked’ DNA, as in the experiments of Griffith and of Avery, MacLeod and McCarthy.
- *Conjugation.* Insertion of some or all of the DNA from one cell into another – the prokaryotic equivalent of ‘mating’, although there is no meiosis or zygote formation. Bacterial conjugation does permit formation of recombinants. The start point for DNA transfer varies with the position in the genome of a mobile site. This is *not* the same as the origin of replication, *oriC*.
- *Transduction.* Transfer of DNA from one cell to another via a bacteriophage. During replication in one cell, a phage can pick up fragments of bacterial DNA and transmit it to another cell subsequently infected by progeny virions.

Bacterial conjugation has proved very useful in genome mapping. Transfer of a complete genome takes 100 minutes. Interrupting the process at different times (by physical agitation) results in partial genome transfer. Identifying which genes have entered the recipient cell after different intervals revealed the order of the genes. Positions in the genetic map of *E. coli*, for example, were classically expressed in minutes. Now, of course, they are specified in terms of the DNA sequence itself.

Genome Organization in Eukaryotes

Genomic information in eukaryotic cells is divided between the main nuclear genome and cytoplasmic organelles: **mitochondria** and **chloroplasts**.

In the nucleus, DNA is complexed with proteins to form chromosomes. We already encountered chromosomes in Chapter 1 and noted that chromatin remodelling is an important component of regulation of gene expression. DNA in organelles also forms nucleoprotein complexes. These resemble bacterial nucleoids, reflecting the endosymbiont origin of organelles. Organelle genomes are circular, double-stranded DNA molecules. Organelles in some species contain more than one DNA molecule.

With a few exceptions, the amount of nuclear DNA per cell is constant in all cells of an organism except for gametes. Although the DNA content per organelle is also at least roughly constant, cells of different tissues contain different numbers of organelles. Mitochondria are more numerous in cells that

consume large amounts of energy, such as brain, heart and eye (about 10 000 mitochondria per cell), than in skin cells (only a few hundred). In plants, a leaf cell may contain up to 100 chloroplasts, the number varying among species; unsurprisingly, root cells have none.

Mitochondrial genomes vary in size among species. Human mitochondrial DNA is 16 569 bp long. Yeast mitochondrial DNA is 75 kb and that of plants is considerably larger: the DNA of muskmelon (cantaloupe) mitochondria is 2.4 Mb! Chloroplast genomes range from about 110 to 160 kb, larger than animal mitochondrial DNAs. In some species, such as the protozoan *Cryptosporidium*, the mitochondria contain no DNA at all!

Mitochondria and chloroplasts carry out their own protein synthesis. Chloroplasts and plant mitochondria translate their genes according to the standard genetic code, but animal mitochondria use variants.

There is active traffic between organelle and nuclear genomes (see Box 2.5). Approximately 90% of chloroplast proteins are encoded by nuclear genes and gene transfer is still going on. However, the differences in genetic code inhibit mitochondrial → nuclear transfer in animals. In the *Arabidopsis thaliana* genome, a 620 kb insertion of mitochondrial DNA in nuclear chromosome 2 contains much of the mitochondrial genome including some duplicated material. (The full mitochondrial genome is only 366 924 bp long.)

Leaves may contain 10^6 chloroplasts per mm^2 of surface area.

Organelle	RNA encoded	Proteins encoded
Animal mitochondria	Two ribosomal RNAs, 22 tRNAs	12 or 13
Plant mitochondria	Three ribosomal RNAs, ~22 tRNAs	30–39
Chloroplast	Four ribosomal RNAs (two copies), 37 tRNAs	50–57

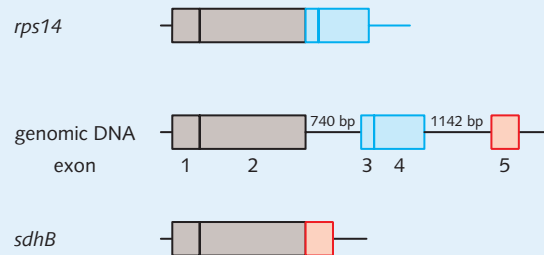
Photosynthetic sea slugs: endosymbiosis of chloroplasts

The endosymbiotic origin of mitochondria and plant chloroplasts is a well-accepted theory about events that happened 1–2 billion years ago. Acquisition of endosymbiotic chloroplasts by sea slugs is observable today (see Figure 2.11). The slugs, which are molluscs – i.e. animals – eat algae. They open the algal cells and discard the contents – including the nucleus – except for the chloroplasts. The chloroplasts are taken up into host cells, where they carry out photosynthesis. The slug can live for months on the molecules synthesized using solar energy.

The mollusc does not get an entirely ‘free lunch’. In algae typical of the slug’s food, the chloroplast genome encodes only 13% of the organelle proteins. During the active life of the chloroplast within the animal’s cells, its proteins turn over and must be synthesized. Genes from the algal chloroplast have entered the mollusc nuclear genome and are expressed by the host.

BOX
2.5

Traffic between the mitochondrial and nuclear genomes



Rps14, a protein from the small subunit of the rice mitochondrial ribosome, is encoded by a nuclear gene, *rps14*, on chromosome 8. In fact, by alternative splicing, the five exons of this region (centre strip) encode both Rps14 (ribosomal protein 14 of the small subunit) and SdhB (the B subunit of succinate dehydrogenase).

Both Rps14 and SdhB are synthesized in the cytoplasm and transported into the mitochondria. It is likely that both genes were originally in the mitochondrial genome. A region similar to the nuclear gene *rps14* remains in the rice mitochondrial genome. This mitochondrial gene is translated, but the product has become non-functional as a result of four single-nucleotide deletions that destroy the reading frame. Certain other higher plants have functional mitochondrial *rps14* genes (broadbean, rapeseed), whereas others resemble rice in containing non-functional mitochondrial *rps14* genes but functional nuclear genes (potato, *Arabidopsis*).

Genes similar to *sdhB* have not been observed in plant mitochondrial genomes. It is likely, therefore, that the move of the *sdhB* gene from the mitochondrial to the nuclear genome is an old event and the move of the *rps14* gene is a relatively recent one.

Moving a mitochondrial gene to the nucleus moves the site of its expression to the cytoplasm. It needs a leader sequence containing a proper targeting signal to direct the protein to mitochondria. The protein encoded by the rice nuclear gene for mitochondrial *rps14* appears to have borrowed a mitochondrial targeting signal from *sdhB*, part of an earlier generation of immigrants, by alternative splicing. Compared with products of mitochondrial genes for Rps14, the nuclear-encoded version has an N-terminal extension derived from the *sdhB* exons. This extension is cleaved off in the mitochondria.

Figure 2.11 A lettuce sea slug (*Elysia crispata*) on a patch of the alga *Bryopsis*. The slug eats algae, extracts and endocytoses the chloroplasts and then basks in the sun, as in the picture, while the chloroplasts photosynthesize organic compounds.

Photograph by William Capman, Augsburg College, Minneapolis, MN, USA.



How Genomes Differ

There is a growing consensus that the dynamics of expression patterns contain the most interesting features of genomes. It is nevertheless prudent to begin less ambitiously, with static aspects – the sequences themselves. Similarities and differences among genome sequences appear (1) at the levels of individual bases; (2) at the level of genes (see Box 2.6); (3) in larger-scale blocks; and (4) at the level of whole genomes that have undergone complete duplications.

Variation at the level of individual nucleotides

Closely related genomes tend to contain regions encoding closely related proteins. Alignments of the sequences of homologous genes reveal differences, mostly in the form of single-site mutations or insertions and deletions. Typically, there is reasonable correlation between overall species divergence and divergence of sequences of individual genes and the corresponding proteins. Comparisons of amino acid and gene sequences of thioredoxins provide a typical example (see Figure 2.12).

Compare these protein sequences with the corresponding gene sequences from human, chicken and *Staphylococcus aureus* (see Figure 2.13). Note the large gap in the bacterial gene corresponding to the intron in the human and chicken genes. The asterisks under the sequences indicate positions containing the same base in all three genomes. The colons indicate positions containing two identical bases among the three; in most cases, two common bases appear in the human and chicken sequences, even in the non-coding regions. Note the frequent occurrence of patterns ‘**.’ and ‘** -blank’. What is the likely reason for this?

Duplications

Duplications of individual genes, of regions containing many genes and of complete genomes have been an important mechanism of evolution. They are a prolific source of variation, the raw material of both selection and genetic drift.

BOX 2.6

What can happen to a gene?

During evolution:

1. A gene may pass to descendants, accumulating favourable (or unfavourable) mutations or drifting neutrally.
2. A gene may be lost.
3. A gene may be duplicated, followed by divergence or by loss of one of the pair.
4. A gene may undergo horizontal transfer to an organism of another species.
5. A gene may undergo complex patterns of fusion, fission or rearrangement, perhaps involving regions encoding individual protein domains.

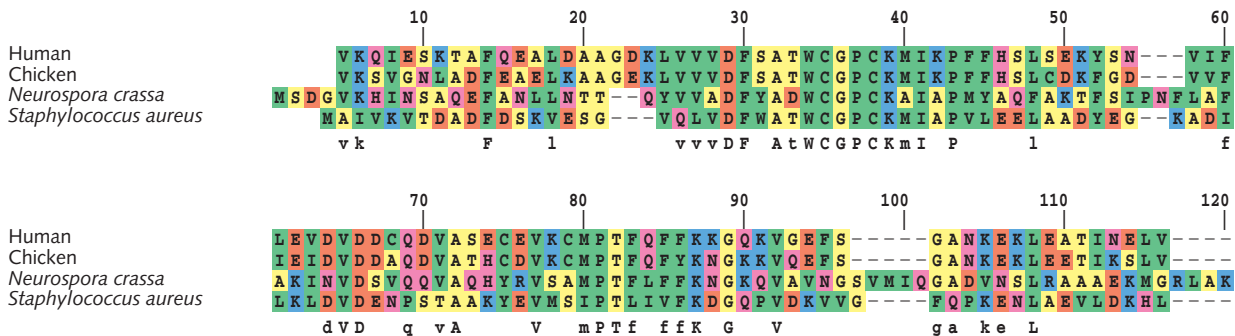


Figure 2.12 Thioredoxins are proteins that catalyse disulphide-exchange reactions, contributing to the speed and accuracy of the protein-folding process. The human thioredoxin gene extends over 13 kb and consists of five exons. This figure shows the alignment of amino acid sequences of thioredoxins from two vertebrates (human and chicken), a fungus (*Neurospora crassa*) and a bacterium (*Staphylococcus aureus*). Colour coding: green, amino acids with medium-sized and large hydrophobic sidechains; yellow, small sidechains; magenta, polar sidechains; blue, positively charged sidechains; red, negatively charged sidechains. Upper-case letters in black on the line below the sequences indicate amino acids conserved in all four sequences. Lower-case letters in the line below the sequences indicate amino acids conserved in three of the four sequences.

Duplications are seen in archaea, bacteria, and eukaryotes. Estimates of amounts of duplication vary, but there is agreement that it is substantial. The *A. thaliana* genome, for instance, contains over 60% duplications.

Duplication of genes

Organisms in all three domains of life show duplication of individual genes.

Species	Duplicate genes (%)
Bacteria	
<i>Mycoplasma pneumoniae</i>	44
<i>Helicobacter pylori</i>	17
<i>Haemophilus influenzae</i>	17
Archaea	
<i>Archaeoglobus fulgidus</i>	30
Eukarya	
<i>Saccharomyces cerevisiae</i>	30
<i>Caenorhabditis elegans</i>	49
<i>Drosophila melanogaster</i>	41
<i>Arabidopsis thaliana</i>	65
<i>Homo sapiens</i>	38

From: Zhang, J. (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. 18, 292–298.

**BOX
2.7****Homologues, orthologues and paralogues**

Homologues are regions of genomes, or portions of proteins, that are derived from a common ancestor. Because only in rare cases can we actually observe the ancestor–descendant relationship, most assignments of homology are inferences from similarity in sequence, structure, and/or genomic context.

Paralogues are related genes that have diverged to provide *separate* functions in the *same* species. **Orthologues**, in contrast, are homologues that perform the *same* function in *different* species. (For instance, the α and β chains of human haemoglobin are paralogues, and human and horse myoglobin are orthologues.)

Other related sequences may be pseudogenes, which may have arisen by duplication or by retrotransposition from mRNA, followed by the accumulation of mutations to the point of loss of function or expression.

After duplication, both copies of a gene may survive and diverge. Alternatively, one copy may turn into a pseudogene or be deleted, leaving only one functional copy.

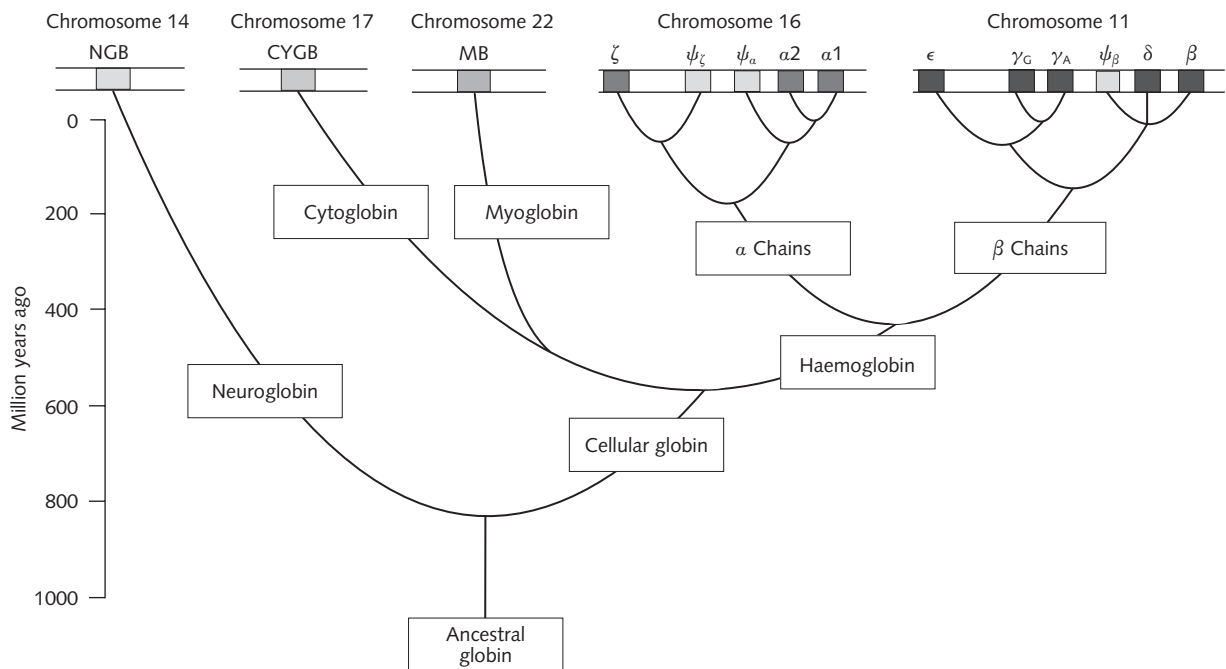
As first proposed by S. Ohno in 1970, duplication followed by divergence is an important source of proteins with novel functions. It is generally easier to ‘recruit’ and adapt an already active molecule to a new function than to invent a new protein from scratch. The course of evolution of proteins descended from a common ancestor will differ, depending on whether they are retaining or changing their function (see Box 2.7).

In analysing the divergence of related genes, how can we distinguish the effect of selection from genetic drift? Given two aligned gene sequences, we can calculate K_s , the number of synonymous substitutions, and K_a , the number of non-synonymous substitutions. Most but not all synonymous substitutions are changes in the third position of codons. (The calculation of K_a and K_s involves more than simple counting because of the need to estimate and correct for possible multiple changes.) The ratio of K_a/K_s distinguishes the role of selective pressure and drift in the divergence of genes after duplication:

- $K_a/K_s \approx 1$ **Neutral evolution:** silent and substitution mutations have occurred to approximately equal extents.
- $K_a/K_s \gg 1$ **Positive selection:** substitution mutations are more prevalent than silent mutations, implying that selective pressures are active and the substitutions are advantageous.
- $K_a/K_s \ll 1$ **Purifying selection:** substitution mutations are underrepresented, implying that the sequence is optimized fairly rigidly, with relatively little tolerance for mutation.

The many globins in the human genome provide a good example of gene duplication and divergence (see Figure 1.5). Genes for several versions of the haemoglobin α and β chains form clusters on chromosomes 16 and 11.

‘Walls supply stones more easily than quarries, and palaces and temples will be demolished to make stables of granite, and cottages of porphyry.’ – Johnson, Rasselas



Other, isolated, loci contain genes for neuroglobin, cytoglobin and myoglobin. Closely linked genes, as in the α - and β -globin regions, suggest relatively recent divergence. Yet even within these clusters, the proteins encoded have diverged in function, showing small but significant variations in oxygen affinity and responses to allosteric effectors. Also, these proteins appear at different stages of our development, implying divergence in the control of their expression.

We can date the globin duplications by looking back into evolutionary history (see Figure 2.14). Neuroglobin split off from other globins before the last common ancestor of the vertebrates, perhaps 10^9 years ago. The divergence of myoglobin and cytoglobin from haemoglobin occurred before the emergence of the jawless fishes, during the Cambrian about 500 million years ago. The divergence of α - and β -globins occurred early in the vertebrate lineage, approximately 450 million years ago.

Within mammals, the α - and β -globin regions are quite variable in content and extent (see Figure 2.15). Even within the primate lineage, we can date a duplication of the γ -globin gene (see Figure 2.16).

Duplication can affect individual exons

Fibronectin, a large extracellular protein involved in cell adhesion and migration, is a **modular protein** (see Box 2.8) containing multiple tandem repeats of three types of domain called F1, F2 and F3. It is a linear array of the form:

Figure 2.14 Duplication and dispersal through the genome of globin genes during animal evolution.

From: Burmester, T., Ebner, B., Weich, B. & Hankeln, T. (2002). Cytoglobin: a novel globin type ubiquitously expressed in vertebrate tissues. *Mol. Biol. Evol.* 19, 416–421.

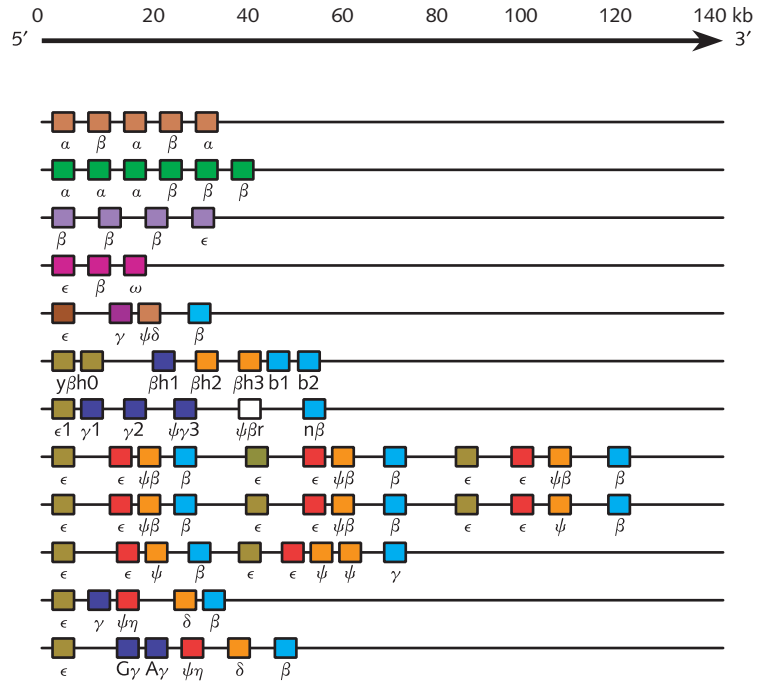


Figure 2.15 Layout of the β -globin locus in selected vertebrates. Colour coding: Light brown, fish; green, amphibian; purple, avian; magenta, marsupial; dark brown, ϵ -like; dark blue, γ -like; orange, δ -like; cyan, β -like; white, rat-specific pseudogene; red, η -like.

The zebrafish and *Xenopus* regions illustrate the organization of the region prior to the separation of α - and β -globins. They alone of the species illustrated here contain only α - and β -globin genes.

Goat and sheep are closely related species that show similar patterns. Rat and mouse are closely related species that show different patterns.

After: Aguilera, G., Bielawski, J.P. & Yang, Z. (2004). Gene conversion and functional divergence in the β -globin gene family. *J. Mol. Evol.* 59, 177–189.

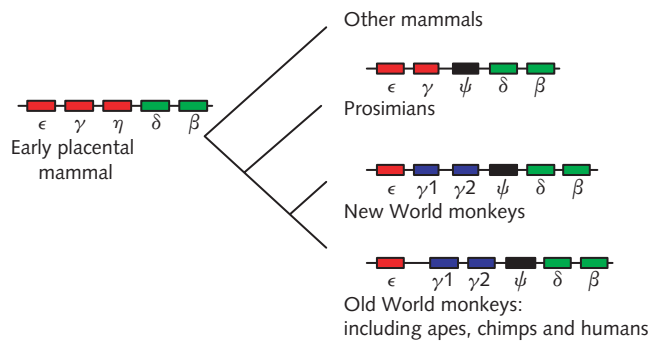


Figure 2.16 Evolution of primate β -globin region (not drawn to scale). Red boxes, embryonically expressed genes; green boxes, post-embryonically expressed genes; blue boxes, fetally expressed genes; black boxes, pseudogenes. Mammals ancestral to the groups in this chart had an embryonic ϵ -globin and a post-embryonic β -globin gene. Marsupials retain this pair (see Figure 2.15). By the time this diagram begins, the ϵ gene had duplicated to form three embryonic genes, ϵ , γ and η , and the β gene had duplicated to form δ and β . The η gene fell into desuetude, mutating into a pseudogene. Subsequent duplication of the γ gene in anthropoids produced γ_1 and γ_2 , and a change in the control of expression to convert the γ genes from embryonic to fetal expression.

**BOX
2.8****Modular proteins**

A **modular protein** contains a linear string of compact units, called **domains**. Domains appear to have independent stability and can be 'mixed and matched' with one another in different proteins. Domains sometimes, but by no means always, correspond to single exons.

Modular proteins are common in eukaryotes. Individual domains of eukaryotic modular proteins are often separately homologous to single-domain prokaryotic proteins (and, less commonly, vice versa).

(F1)₆(F2)₂(F1)₃(F3)₁₅(F1)₃ (see Figure 2.17). In the human genome, each domain of fibronectin is encoded by either one or two tandem exons.

Fibronectin domains also appear in other modular proteins. The duplication of the exon(s) encoding a domain, followed by transfer to another protein, is called 'exon shuffling'.

Family expansion: G-protein-coupled receptors

Repeated duplications can generate large numbers of homologues. G-protein-coupled receptors (GPCRs) are a large superfamily of eukaryotic cell-surface receptors active in signal recognition and processing, including the senses of sight, taste and smell.

The human genome contains about 700 active GPCRs. They are integral membrane proteins with a common structure comprising seven transmembrane helices (see Figure 2.18). GPCRs interact with G proteins within the cell. Some GPCRs mediate responses to extracellular chemical signals. Reception of the signal triggers *intracellular* signalling cascades, which may reach as far as the nucleus to affect gene expression.

The interaction patterns of large families of functionally related proteins can create great complexity. Odorant receptors are GPCRs expressed on sensory neurons in the nasal cavities of humans and animals. The human genome has about 1000 odorant-receptor genes, of which only 40% are active. The mouse genome has about 1300, of which 80% are active. These genes are distributed around mammalian genomes, arranged in clusters of up to 100 genes, in >40 locations in mouse and >100 locations in humans. Formation of many of the clusters appears to antedate the divergence of humans and mice, but individual gene duplication and divergence within clusters has led to specialization.

Although mammals have of the order of 1000 odorant-receptor proteins, and each neuron in the nasal epithelium expresses only one odorant-receptor allele, over ten times as many scent molecules can be distinguished. How is this achieved? Each scent molecule interacts with several different receptors. Conversely, each receptor protein binds several related scent molecules. Each neuron signals the detection of one *group* of odours, and the brain compares the outputs and performs the required computation. Identification of a specific scent depends on its detection by a *combination* of receptors.

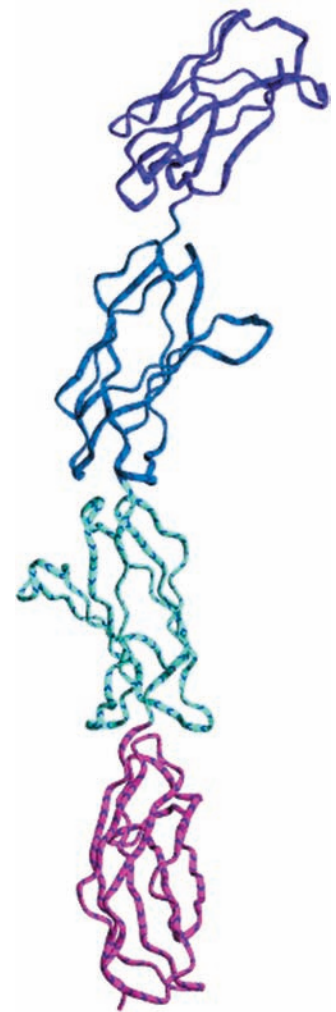


Figure 2.17 A fragment of fibronectin, a modular protein, showing four tandem domains [1FNF].

GPCRs are important in the pharmaceutical industry, for both therapeutic effectiveness and financial reward. About half of all prescription drugs target GPCRs.

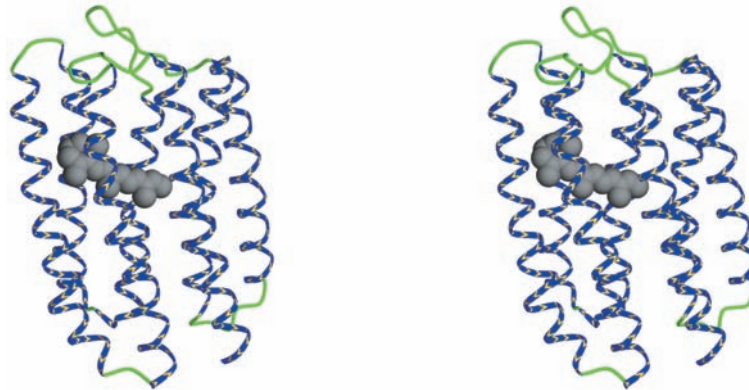


Figure 2.18 G-protein-coupled receptors (GPCRs) are a large family of transmembrane proteins involved in signal transduction into cells. They share a substructure containing seven transmembrane helices, arranged in a common topology. This figure shows the only experimentally determined mammalian GPCR structure, bovine opsin [1H68]. This molecule senses light and generates a nerve impulse.

The seven-helical structure is common to the family of GPCRs. The helices traverse the membrane, with loops protruding outside and inside the cell. This figure shows a view parallel to the membrane, with the extracellular side at the top. The transmembrane region is generally flanked by N- and C-terminal domains. The N-terminal domain is always outside the cell and the C-terminal domain always inside.

GPCRs constitute the largest known family of receptors. The family is as old as the eukaryotes and is large and diverse. Mammalian genomes contain ~1500–2000 GPCRs, accounting for about 3–5% of the genome. A similar fraction of the *C. elegans* genome codes for GPCRs.

Some GPCRs are involved in sensory reception, including vision, smell and taste. Some, like opsin and bacteriorhodopsin, bind chromophores. (Bacteriorhodopsin is not a signalling molecule but a light-driven proton pump.) Others respond to extracellular ligands including hormones and neurotransmitters.

As expected from the structure, in many groups of GPCRs the sequences of the helical regions diverge less than the sequences of the loops. It is the loops that determine the specificity of the ligand, and of the G-protein partner.

The common mechanism of function of GPCRs is a conformational change, induced by receptor binding or light absorption. The activated state of the GPCR interacts with an intracellular G protein, triggering a signal cascade. As there are substantially more GPCRs than G proteins, many GPCRs must interact with a single G protein. For instance, all odorant receptors interact with the same G protein α -subunit.

GPCRs are the targets for many drugs used in the treatment of high blood pressure, asthma, allergies and other conditions. The large number of related GPCRs is a challenge to the design of drugs that bind to a unique target. Many drugs have undesired side effects because of imperfect specificity.

Large-scale duplications

The genomes of many species contain duplications of multigene regions, the length varying from species to species.

Large-scale segmental duplications are an important component of the difference between human and chimpanzee genomes, affecting about 2.7% of the genome. Some duplications are found in chimpanzees but not humans, some in humans but not chimpanzees, and some in both.

We depend less on our sense of smell than mice and dogs do, but we have about half the number of expressed odorant receptor proteins. Would you have expected a larger discrepancy?

Some duplications that appear in the human but not in the chimpanzee genome involve segments associated with developmental disorders, including a region on human chromosome 15 involved in Prader–Willi and Angelman syndromes (see Box 1.5, p. 33). These syndromes arise from microdeletions. In humans, the duplication contributes to the frequency of the disease, by presenting sites for homologous recombination during meiosis, which show up in some of the gametes as deletions. It is, therefore, likely that Prader–Willi and Angelman syndromes are *less* common in chimpanzees than in humans.

Whole-genome duplication

Genomes can duplicate if the chromosomes replicate but do not segregate properly into separate progeny cells upon mitosis.

The mere appearance of two copies of many genes does not prove whole-genome duplication. One must adduce (1) the genome-wide occurrence of pairs of homologous genes *appearing in the same order*; or (2) ‘molecular clock’ evidence showing equal divergence times in many pairs of homologues. (However, different genes diverge at different rates, and homologues may be under different selective pressures. Clock arguments are, therefore, relatively weak.)

The yeast genome underwent a duplication about 10^8 years ago. The effects are obscured by subsequent chromosomal rearrangements and by massive loss of duplicated material. The duplication has nevertheless left its traces in multiple homologues that retain their genomic order. The yeast genome contains 55 duplicated regions, on average 55 kb long, together covering ~50% of the genome and including 376 pairs of homologous genes.

Animals appear to have undergone two whole-genome duplications. One occurred after the divergence of chordates from cephalochordates, before the emergence of jawless fishes, and another before the divergence of jawless and jawed fishes. These both occurred about 400–600 million years ago. More recently, a third duplication occurred in the lineage leading to ray-finned fishes, such as zebrafish and medaka (see Box 2.9).

BOX 2.9

HOX (homeobox) genes

Organisms with bilateral symmetry, including insects and vertebrates, contain *HOX* genes, which encode a family of DNA-binding proteins. The expression of these genes varies along the head-to-tail body axis and controls the setting out of the body plan. *HOX* genes have overlapping domains of expression within the body. Different regions of the embryo develop into anatomically distinct regions of the adult, based on the subset of genes expressed.

HOX genes reveal the duplications that have occurred during vertebrate evolution. Insects and amphioxus have a single *HOX* cluster. Humans have four *HOX* clusters. Zebrafish have seven *HOX* clusters, interpretable in terms of a series of duplications: $1 \rightarrow 2 \rightarrow 4 \rightarrow 8$ followed by loss of one to reduce $8 \rightarrow 7$.

Indeed, there is a mapping between (1) the order of the genes on the chromosome; (2) the relative times during development of the onset of their activity; and (3) the order of their action along the body. (The β -globin locus shares the first two of these but not the third; see Chapter 1.)

Plant genomes are very susceptible to duplication. The sequence of the *Arabidopsis* genome reveals at least two and possibly three successive duplication events.

Most plants are **polyploids**, i.e. they contain multiple sets of entire chromosomes. **Autopolyploids** contain multiple copies of genomes from the same parent. **Allopolyploids** contain multiple copies of genomes from different parents. Many crop species are polyploids, relative to the wild species from which they were domesticated, including wheat, alfalfa, oats, coffee, potatoes, sugar cane, cotton, peanuts and bananas. Often polyploidy increases the size of the fruit or grain, a useful property for agriculture (see Box 2.10).

Polyploidy may have other advantages. In studies of Arctic flora, it is observed that the fraction of diploid and polyploid plant species increases towards higher latitudes. Many arctic plants tend to exist in small, separated populations and frequently go through ‘bottlenecks’ of marginal survival, for instance during glaciations. After recession of the ice, deglaciated areas may be repopulated by a few or even one dispersed seed. Carrying many copies of the genome in the cells of each individual may help to preserve genetic diversity, even in tiny populations.

**BOX
2.10****Polyploidy in wheat**

The wheat first used in agriculture, in the Middle East at least 10 000–15 000 years ago, is a diploid called **einkorn** (*Triticum monococcum*), containing 14 pairs of chromosomes. Emmer wheat (*T. dicoccum*), also cultivated since palaeolithic times, and durum wheat (*T. turgidum*), are merged hybrids of relatives of einkorn with other wild grasses to form tetraploid species. Additional hybridizations, to different wild wheats, gave hexaploid forms, including spelt (*T. spelta*) and modern common wheat (*T. aestivum*). Triticale, a robust crop developed in modern agriculture and currently used primarily for animal feed, is an artificial genus arising from crossing durum wheat (*T. turgidum*) and rye (*Secale cereale*). Most triticale varieties are hexaploids.

Variety of wheat	Classification	Chromosome complement
Einkorn	<i>Triticum monococcum</i>	AA
Emmer wheat	<i>Triticum dicoccum</i>	AABB
Durum wheat	<i>Triticum turgidum</i>	AABB
Spelt	<i>Triticum spelta</i>	AABBDD
Common wheat	<i>Triticum aestivum</i>	AABBDD
Triticale	<i>Triticosecale</i>	AABBRR

A, genome of original diploid wheat or a relative; B, genome of a wild grass, *Aegilops speltoides* or a relative; D, genome of another wild grass, *T. tauschii* or a relative; R, genome of rye *S. cereale*.

All of these species are still cultivated – some to only minor extents – and have their individual uses in cooking. Spelt, or *farro* in Italian, is the basis of a well-known soup; pasta is made from durum wheat; and bread is made from *T. aestivum*.

Although polyploidization is much more common in plants than in animals, related species of frogs (genus *Xenopus*) are diploid, tetraploid, octaploid and dodecaploid. One tetraploid mammal is known, the rat *Tympanoctomys barrerae* from the Monte Desert in west-central Argentina. These species provide a model for control of expression of duplicated genes. For example, in 'polyploid' frogs, silencing is non-syntenic. Each copy of the genome contains some expressed and some silenced genes. This is a different model from the silencing of an entire X chromosome in cells of mammalian females (see Chapter 1).

There are a number of examples of tissue-specific polyploidization. The endosperm of maize kernels undergoes repeated cycles of **endoreplication** (replication of nuclear DNA in the absence of mitosis) to produce cells that can have as many as 96 copies of the haploid genome. In mammals, it is observed that the number of polyploid cells in the liver increases with age, or in response to disease or surgery even in children. This may be a defence against oxidative stress. In the bone marrow of mammals, very large polyploid cells called megakaryocytes 'bud off' portions of their cytoplasm to form platelets. (Platelets are enucleate cells in the blood involved in clotting.) In a related condition, called **polyteny**, replicated chromosomes remain in alignment rather than separate as in polyploids. This is the origin of the giant salivary gland chromosomes of *Drosophila*, which played such an important role in the history of cytogenetics (see Figure 1.17).

Comparisons at the chromosome level: synteny

Comparison of chromosome banding patterns provide snapshots of similarities and differences in large-scale organization among eukaryotic genomes. Synteny literally means 'on the same band', that is, on the same chromosome. (The chromosome exchange that causes chronic myeloid leukaemia (see Chapter 1) is a *breaking* of synteny.) Closely related species generally show a correspondence between large syntenic blocks. The similarity of the banding patterns reveals the underlying similarity of the patterns in the DNA sequences themselves.

Figure 2.19 shows the relationships among the karyotypes of human, chimpanzee, gorilla and orangutan.

What makes us human?

It is too difficult to look only at the human genome and try to deduce . . . ourselves. Two approaches help in understanding the genome.

- *Comparative genomics*. We can compare the human and chimpanzee genomes and ask how *differences* between these genomes might give rise to *differences* between the species.

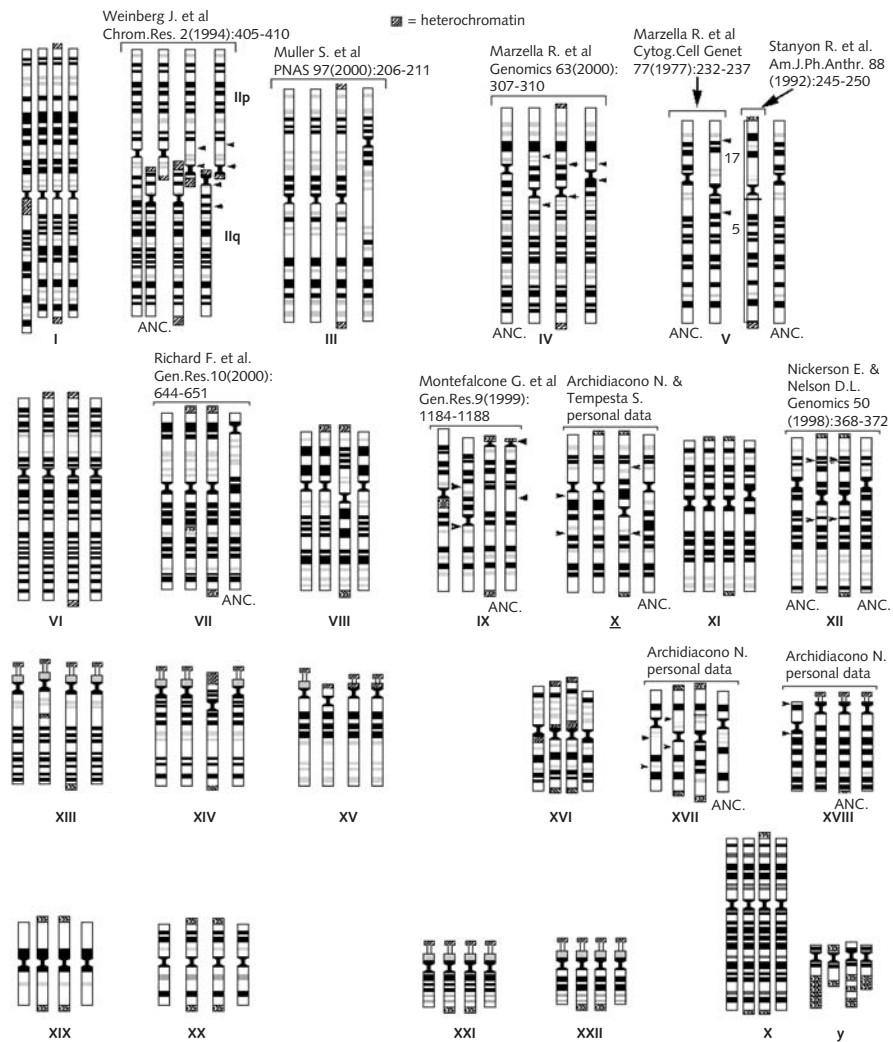
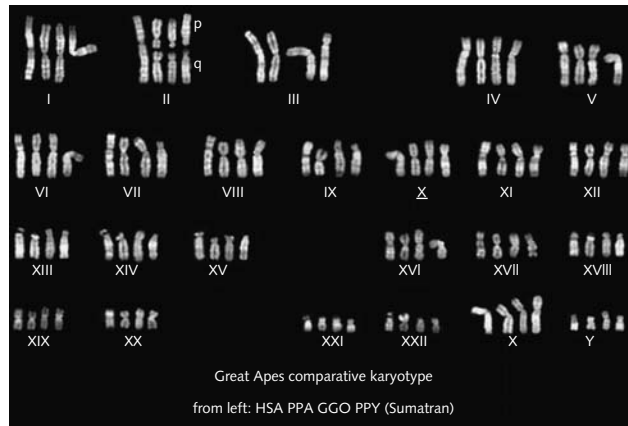


Figure 2.19 Top: photograph of banding patterns. Bottom: ideograms. HSA, *Homo sapiens*; PTR, *Pan troglodytes* (chimpanzee); GGO, *Gorilla gorilla*; PPY, *Pongo pygmaeus* (orangutan).

Photographs courtesy of Prof. M. Rocchi, Università di Bari, Italy.

- *Study of human disease.* Many mutations cause disease and give clues to the functions of the affected regions. These regions may encode enzymes or regulatory proteins or RNAs, or they may be DNA sequences that are targets of regulatory mechanisms. Understanding the effects of mutations both illuminates human biology and, often, has immediate clinical applications.

Comparative genomics

The human and chimpanzee genomes are about 96% identical. To understand what makes us human – or at least what makes us not chimpanzees – we can focus on 13 Mb of different sequence, rather than the full 3.2 billion. There are even fewer differences in our amino acid sequences. Humans and chimpanzees express very similar sets of proteins, and most of the homologous proteins of chimpanzee and human are identical or very similar. About 30% of homologous human and chimpanzee proteins show no differences at all. On average, there are only two amino acid differences.

How, then, do humans and chimpanzees develop differently? The ultimate answer *must* lie within the static sequence of the genome. However, a satisfactory answer will require understanding of the dynamics, specifically of patterns of regulation of gene expression.

There is a paradox here. On the one hand, living systems are fairly robust to perturbations. Yeast, for example, survives individual knockout of 80% of its genes. On the other hand, the 4% differences between chimpanzee and human genomes make profound differences in phenotype. This suggests a *chaotic system*, one in which tiny perturbations can lead to large changes in the subsequent trajectory. *Superposed on the robustness are specific changes that exert immense leverage.*

There are two ways to find these crucial sequences. One is to look closely at the differences between human and chimpanzee genomes and try to figure out what the changed loci are doing. Another is to examine human mutations that affect phenotypic properties that chimpanzees do *not* share with humans, such as language and reasoning. Study of human mutations has focused on the genetics of disease (see Box 2.11).

Combining the approaches: the FOXP2 gene

Language is a unique feature of our species. It should show up as a genetic difference between our genomes and those of other species, including chimpanzee.

Many people suffer from diseases that interfere with production or comprehension of language, or both. Some of these are associated with trauma or with complex genetics (dyslexia is an example). However, one abnormality with simple Mendelian inheritance appears in a family in London. Members of the 'KE' family have a severe disorder affecting both the facial motor control involved in producing speech and also the mental processing of language.

The mutation responsible for this condition has been identified. It is a single-nucleotide polymorphism (SNP) in a gene called *FOXP2*, which encodes a

**BOX
2.11****Mutations and disease**

- *Some diseases involve single genes encoding metabolic enzymes.* Phenylketonuria (PKU) is a genetic disease caused by a deficiency of phenylalanine hydroxylase, the enzyme that converts phenylalanine to tyrosine (see p. 379). Accumulation of high levels of phenylalanine causes developmental defects, including mental retardation. The disease cannot be cured, but symptoms can be avoided by lifestyle control: a phenylalanine-free diet. Screening of newborns for high blood phenylalanine levels is legally required in the USA and many other countries.
Some deleterious mutations have not been eliminated from populations by selection because they carry some compensating advantages. The genes for sickle-cell anaemia and for glucose-6-phosphate dehydrogenase deficiency confer resistance to malaria.*
- *Some diseases involve multiple genes and complex interactions with the environment.* Over 100 genes are implicated in susceptibility to asthma. Several environmental factors also affect risk: life history of challenge to the immune system plays an important role. Thus, breast feeding and bouts of common childhood diseases appear to have protective effects. Occupational exposure to allergens and toxins contributes to development of the disease, as well as triggering attacks in asthma sufferers.
- *Cancer is a disease of genomic instability.* Several genes act as tumour suppressors, with mutations correlating with increased risk of developing cancer. Some of these act to repair DNA damage.

*For more information, see Lesk, A.M. (2004). *Introduction to Protein Science*. Oxford University Press, Oxford, UK.

transcription factor. The major protein encoded by *FOXP2* is 715 amino acids long. It is quite a stable protein from the evolutionary point of view, with only one substitution between mouse and the identical chimpanzee, rhesus macaque and gorilla sequences. However, the human protein has two mutations relative to the other primate sequences.

This example illustrates the power of a combination of studies of human phenotypes and comparative sequence analysis. And yet the observation that the phenotype shown by members of the KE family arises from two SNPs in a single gene may be deceptively simple. We cannot conclude that the expression of only one gene is involved in creating the phenotype, as is the case for phenylketonuria, for example. The *FOXP2* gene product is a transcription factor. Its activity affects the expression of many genes. The effectiveness of their coordinated expression required co-evolution, i.e. sequence changes in other genes.

Genes and minds: neurogenomics

Higher mental processes unique to humans are the last major unexplored territory in our understanding of life. Mental illness is a major cause of disability in Europe and the USA, but it has been difficult to integrate human mental achievements and diseases into mainstream biology and medicine. This is partly because the physical correlates of mental events are so complex and partly

because of a clinical tradition of treating mental disease by conversation between therapist and patient, and by counselling. That tradition, with its origin in the pervasive influence of Freud, was based on the assumption that the causes of mental disease were either emotional trauma alone or emotional consequences of physical trauma. Even though we now know that the genetic component of mental disease is much more important than formerly suspected and that even environmentally caused mental diseases have biochemical effects, the fact remains that treatments alternative to counselling, based on a profound biological understanding, are not available. Drugs now play an important role in treating symptoms of mental disease. But Macbeth's anguished cry, 'Cans't thou not minister to a *mind* diseased?' is still largely an unmet challenge.

One approach to mental processes in humans is to work our way up from simpler nervous systems. This was Sydney Brenner's motive for choosing *C. elegans* as a model organism. The adult hermaphrodite form has exactly 302 nerve cells and 7000 synapses (see Figure 2.20). J. White worked out the complete 'wiring diagram' by tracing individual cells through serial sections. We know not only the static structure and organization of the adult system, but the details of how it develops.

Some people espouse this approach. ('You have to walk before you can run.') Others retort that human mental achievements and diseases go far beyond those of *C. elegans*. It is undeniable that understanding the minds of worms – or even of chimpanzees – must have limited applicability to humans. Nevertheless, many surprising similarities have appeared, even in such distantly related organisms.

What kinds of human mental phenomenon do model organisms show? The basic mechanisms of sensation, such as vision and olfaction, are similar in many animals. Learning and memory are widespread throughout the animal kingdom. Flies learn; even worms learn. Language is one uniquely human attribute and its genetic component has been traced through clinical disorders. One might assume that 'higher' emotions and sophisticated talents – love, despair, the ability to play chess or a talent for art or music – would also be absent from simple model organisms, but in many cases tantalizing analogies do exist. Fruit flies show courtship behaviour under the control of known genes. The objection that fruit flies are merely displaying unsophisticated instinctive

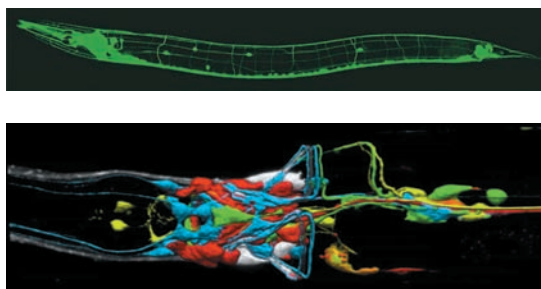


Figure 2.20 Top: the nervous system of *Caenorhabditis elegans*, head at left, labelled with green fluorescent protein. Bottom: the brain of *C. elegans*, with different groups of neurons labelled with fluorescent proteins of different colours.

Reproduced by permission of Prof. H. Hutter, Simon Fraser University, Canada.

behaviour raises the question of how rational and sophisticated is most human activity, including courtship.

Genomics provides the bridge between the minds of the different species. Two ways to use model organisms are the study of homologues to illuminate functions of human proteins and the insertion of human genes into the model organisms to study the effects of their expression.

Models for neurological disease, in natural or transgenic *Drosophila* and *C. elegans*, include Parkinson's and Huntington's disease, Friedrich's ataxia (a degenerative disease of the nervous system) and early-onset dystonia (neuromuscular dysfunction producing sustained involuntary and repetitive muscle contractions or abnormal postures). Species more closely related to humans, including zebrafish and mice, provide models for most mental diseases.

Traditionally, there was a fairly rigid distinction between neurological/physiological/biochemical diseases affecting cognitive performance and psychiatric diseases believed to have emotional causes and effects. However, even if an illness is caused by an unhealthy emotional environment (and leaving aside the genetic component of *susceptibility* to disease in response to such an environment), organic changes – down to the molecular level – underlie the psychological manifestations.

There is now good evidence for a substantial genetic component in numerous psychological conditions, including dyslexia, autism, schizophrenia, attention-deficit hyperactivity disorder and others. A number of conditions appear in both **familial** forms, showing relatively simple patterns of inheritance, and **sporadic** forms, with more complex genetic components and greater influence of environment. Familial forms often show early onset.

For instance, familial Alzheimer's disease is a relatively early-onset form of the condition, with early onset defined in this case as appearing before the age of 65. It affects about 10% of those with Alzheimer's disease. The familial form is associated with mutations in genes on chromosomes 1, 14 and 21. In contrast, a mutation in the gene for ApoE on chromosome 19 causes increased risk of late-onset Alzheimer's disease.

Genetics of behaviour

- *Different strains of C. elegans show different social behaviour in dining.* *C. elegans* can be grown on an agar lawn in a Petri dish. The wild type collected from Australia will congregate in groups to feed. The wild type collected from Britain will eat separately (see Figure 2.21). The difference has been traced to a single amino acid change in a seven-transmembrane helix protein, NPR-1.
- *In fruit flies, T. Tully and co-workers have identified a gene associated with memory.* CREB (cyclic AMP response element-binding protein) encodes a transcription factor, part of a large family of paralogues in mammals and distributed widely in eukaryotes and prokaryotes. Some engineered changes in CREB produced flies that could learn but not store memories; other changes

The unity of life is a theme of this book, but perhaps it would be wrong to read too much into this example.

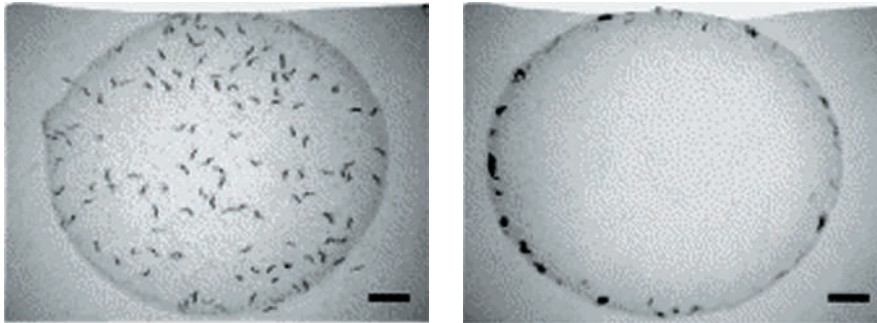


Figure 2.21 Alternative social behaviours of *Caenorhabditis elegans*, feeding on a lawn of agar in a Petri dish. Left: solitary feeding. Right: clumping. The agar is slightly thicker at the edges of the dish, which is why the worms congregate there. Bars, 2.5 mm.

From: de Bono, M. & Bargmann, C.I. (1998). Natural variation in a neuropeptide Y receptor homolog modifies social behaviour and food response in *C. elegans*. *Cell* 94, 679–689.

produced flies that learned substantially faster than normal. Alterations in mouse *CREB* have produced memory impairment.

- *The Lesch–Nyhan syndrome was the first correlation discovered between a specific human genetic defect and behavioural anomalies.* It is an X-linked deficiency of a single protein, the enzyme hypoxanthine–guanine phosphoribosyltransferase. The consequent inability to metabolize uric acid properly leads to physical symptoms including gout and kidney stones. Patients also show poor muscle control, mental retardation, facial grimacing, writhing and repetitive limb movements, and uncontrollable lip and finger biting to the point of severe self-mutilation.
- *Seasonal affective disorder, a mood change in response to prolonged darkness, is related to the regulation of circadian rhythms.* Jet lag is a related condition. The system of genes involved in circadian rhythms was originally worked out in fruit flies and has homologues in many animals, including humans and *C. elegans*, and in plants.
- *Mutations in an X-linked human gene, DLG3, cause severe learning disability.* Knockout of the mouse homologue *PSD95* produces learning-impaired mice. The protein encoded by *PSD95* binds to the NMDA receptor, a protein involved in synaptic plasticity. Overexpression of NMDA receptor gives mice superior learning and memory abilities.
- *Several genes are implicated in schizophrenia.* A common effect, hypersensitivity to the neurotransmitter dopamine in the brain, plays a role in schizophrenia. Some antipsychotic drugs block dopamine receptors on neuronal surfaces; conversely, amphetamines, which release dopamine, aggravate the symptoms.
- *Depression is a common response to stress.* Although we all suffer stressful episodes in our lives, the development of debilitating mental disease depends on our alleles in the promoter region of the gene for the serotonin transporter

(5-HTT). Individuals with one or two copies of the shorter allele of this gene are more likely to exhibit depression and suicidal tendencies than individuals homozygous for the longer allele.

- *Maltreatment in childhood is a cause of antisocial behaviour.* Obviously. However, the likelihood of development of antisocial behaviour depends on the allele for monoamine oxidase A (MAOA). Maltreated children with a genotype producing high expression levels of MAOA are less likely to become antisocial or violent offenders. A. Caspi and colleagues found that, although only 12% of a sample of people had the combination of low-activity MAOA genotype and childhood maltreatment, these individuals accounted for 44% of subsequent convictions for violent offences.

Genomes of Chimpanzees and Humans

The genome sequence of Clint, a male chimpanzee from the Yerkes National Primate Research Center at Emory University, in Atlanta, Georgia, USA, was reported in 2005. Clint represented the West African subspecies *Pan troglodytes verus*.

As expected from so closely related a species:

- There is close alignment of the genome: 96% is alignable with the human genome.
- The sequences of the alignable regions differ at 1.23% of the positions. Recognizing that there is intraspecies divergence among humans and among chimpanzees, it is likely that the true interspecies difference amounts to about 1% of the alignable sequence.
- The 4% non-alignable regions represent insertions and deletions. It is estimated that about 45 Mb of human sequence do not correspond to chimpanzee sequence and a similar amount of chimpanzee sequence does not correspond to human sequence. More positions in the genomes differ as a result of insertions/deletions than differ as a result of base substitutions.
- The distribution of differences is variable across the genome. For all syntenic 1 Mb segments across the genomes, the range in difference is about 0.005–0.025%. Looking at the distribution with respect to the chromosomes, divergence tends to be higher near the telomeres. The divergence is lowest for the X chromosome and highest for the Y.
- The proteins encoded are also very similar in sequence. Of 13 454 orthologous proteins, 29% have identical sequences. On average, there are one to two amino acid residue differences between corresponding chimpanzee and human proteins.
- Although most proteins are very similar, a few show large K_a/K_s ratios, suggesting that they are under positive selection. These include two proteins,

Sadly and unexpectedly, Clint died a few weeks before the paper on his genome was submitted to Nature. He was 24 years old. Even in the wild, chimpanzees can live for 40–45 years. Cheetah, the chimpanzee who starred with Johnny Weissmuller in the Tarzan films during the 1930s and 1940s, is still alive at 74.

glycophorin C and granulysin (involved in combating infection) and other proteins involved in reproduction. (Selection can act most directly on reproduction itself, producing high rates of evolutionary change.)

- Changes in gene expression patterns show that genes active in the brain have changed more rapidly in humans.

Genomes of Mice and Rats

The mouse and rat are by far the most common mammalian laboratory animals. Knowledge of these species and correlation with human biology is encyclopaedic. Determination of mouse and rat complete genome sequences was clearly a high-priority goal. The genome of the laboratory mouse (*Mus musculus*) appeared in December 2002. The genome of the brown or Norway rat (*Rattus norvegicus*) appeared in April 2004.

Mice, rats and humans are closely related mammals and illuminate one another. Laboratory studies on rodents are useful guides to the biochemistry and molecular biology of humans. Mice and rats provide the first test of tolerance to, and effectiveness of, novel drugs aimed ultimately at human therapy. Outside the laboratory, however, the close relationship has been tragic for humans: shared parasites permit rats to transmit disease. (An epidemic of bubonic plague in 1347–1352 killed a third of the population of Europe.) Shared diets make food supplies vulnerable to rodent infestation.

The last common ancestor of humans and rodents lived approximately 75 million years ago. Rats and mice separated much more recently: 12–24 million years ago. The genomes of all three species are approximately the same size. The rat genome is ~5% smaller than the human genome. The mouse genome is about ~15% smaller than the human genome. Sequence divergence and chromosome segment rearrangement appear to have been faster in the rodent lineage. The human genome shows more duplication – one reason why it is larger.

Human, mouse and rat genomes encode similar numbers of genes. Most proteins have homologues in all three species, with very similar amino acid sequences (see Figure 2.22). (Transgenic animals – substituting rodent genes with human genes – can equip model organisms with exact human sequences if necessary.) The genes for most rodent–human homologues have a common exon–intron structure. Gene duplications create protein families, which may be of different sizes in different species. For instance, consistent with their greater dependence on a sense of smell, rodents have more odorant receptors than we do.

Some genomic variation is observable at the chromosomal level. The mouse has 19 chromosomes, plus X/Y. The rat has 20 chromosomes, plus X/Y. Synteny between the mouse and rat genomes is high. Synteny between the mouse and human genomes is variable. Most human chromosomes contain

An EMBL facility has accommodation for 35 000 mice.

Cytochrome c

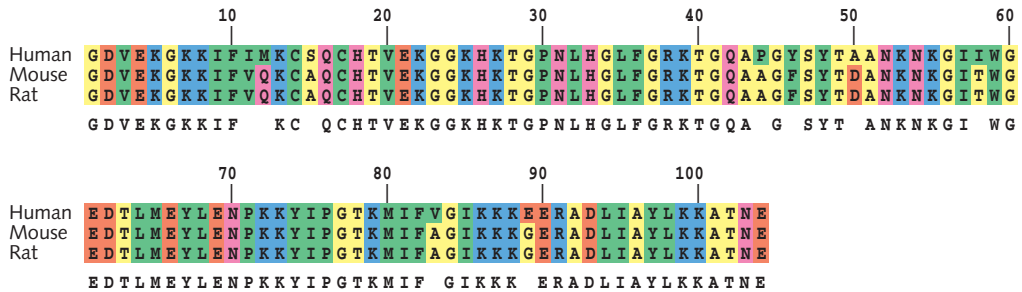


Figure 2.22 Alignment of the amino acid sequences of cytochrome c from human, mouse and rat. All have 104 residues. The mouse and rat sequences are identical. Letters below the sequences show residues conserved in all three species. The human sequence shows nine substitutions, most of which are conservative. For colour coding, see Figure 2.12.

many small blocks that correspond separately to regions distributed among several mouse chromosomes. However, almost all of human chromosome 20 corresponds almost continuously with a region of mouse chromosome 2. Almost all of human chromosome X appears in mouse chromosome X, differing by rearrangement of nine contiguous blocks, including some reversals. Almost all of human chromosome 17 appears within a region of mouse chromosome 11; however, the sequence is broken into 16 segments that are rearranged, including some reversals.

The blocks are identified by matching sequences of genetic markers. Even though most of the correspondences are distributed among different chromosomes, most of the genomes can be partitioned into syntenic blocks, making up a total of 2.35 Gb (over 90% of the mouse genome). These regions include almost all known exons and regulatory regions.

Alignment of genetic maps is less stringent than alignment of sequences. The rearrangements within and among chromosomes make it impossible to align the three genomes as one long linear sequence. However, at the nucleotide level, 40% of human and mouse genomes are block-alignable, about 1 Gb in all.

The differences among the human, mouse and rat genomes arise from selection or neutral drift. Regions changing under selection rather than drift are likely to be functional. This is a powerful way to search for regulatory regions, which are harder to identify in genomes than protein-encoding genes.

Genomics confirms the utility of the rat and mouse for clinical research. Of a set of ~1000 genes for which known mutations are associated with human disease, almost all have homologues in rodents. In certain interesting cases, the sequence of the human disease-associated mutant is identical to the mouse and rat wild type. There has probably been co-evolution, with a compensatory change in some other gene or genes. This can be a source of clues to the function and interactions involved in the disease.

Flies and Worms are Model Organisms for the Study of Human Diseases

Above the molecular level, the differences between humans and flies and between humans and worms are more obvious than the similarities. At the biochemical and genomic level, the situation is reversed. The underlying common features of the structure, organization and development of different species is of both academic interest and practical importance, as flies and worms provide models for human diseases.

The genome of C. elegans

The nematode worm *C. elegans* entered molecular biology in 1963, at the invitation of S. Brenner. Brenner recognized its potential as a sufficiently complex organism to be interesting but simple enough to permit complete analysis of its development and neural circuitry, *at least* at the cellular level (see Figure 2.20).

The *C. elegans* genome, completed in 1998, was the first full DNA sequence of a multicellular organism. *C. elegans* contains ~97 Mb of DNA distributed on six paired chromosomes (see Box 2.12). There is an X but no Y chromosome: different genders in *C. elegans* are a self-fertilizing hermaphrodite, genotype XX, and a male, genotype XO (i.e. a single unpaired X chromosome).

The *C. elegans* genome is about eight times larger than that of yeast, and its 19 099 predicted genes are approximately three times the number in yeast. Exons cover ~27% of the genome. The genes contain an average of five introns. The gene density is relatively low, for a eukaryote, with ~1 gene/5 kb of DNA. Approximately 25% are in clusters of related genes.

BOX 2.12

Distribution of *C. elegans* genes

Chromosome	Size (Mb)	Number of protein genes	Density of protein genes (kb/gene)	Number of tRNA genes
I	7.9	2803	5.06	13
II	8.5	3259	3.65	6
III	7.6	2508	5.40	9
IV	9.2	3094	5.17	7
V	9.8	4082	4.15	5
X	10.1	2631	6.54	3

From: The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018.

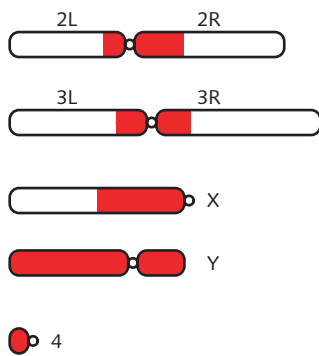


Figure 2.23 The chromosomes of *Drosophila melanogaster*. Heterochromatin is shown in red.

The genome of *Drosophila melanogaster*

The total chromosomal DNA of *Drosophila melanogaster* contains about 180 Mb. Approximately two-thirds is euchromatin, a relatively uncoiled and non-compact form, containing most of the active genes. The euchromatic portion, about 120 Mb, was the first segment of the sequence released. The other one-third of the *Drosophila* genome appears as heterochromatin, highly compact regions flanking the centromeres. Heterochromatin contains many tandem repeats of the sequence AATAACATAG and relatively few genes. Determination of the heterochromatin sequence is in progress.

The genome is distributed over five chromosomes: three large autosomes, a tiny chromosome containing only ~1 Mb of euchromatin and an X/Y chromosome pair, of which the Y chromosome is heterochromatic and relatively gene poor. The fly's ~14 000 genes are approximately double the number in yeast, but fewer than in *C. elegans*, perhaps a surprise. The average density of genes in the euchromatin sequence is 1 gene/9 kb; about half that of *C. elegans* (see Box 2.13). The genes of the metacentric chromosomes 2 and 3 are reported separately for the two arms, arbitrarily designated left (L) and right (R). The other chromosomes are telocentric (see Figure 2.23).

Determination of the *D. melanogaster* genome sequence was a collaboration between industry (Celera Genomics) and the academic *Drosophila* Genome Projects based in Berkeley, California, USA, and in Europe. The project was a methodological testbed.

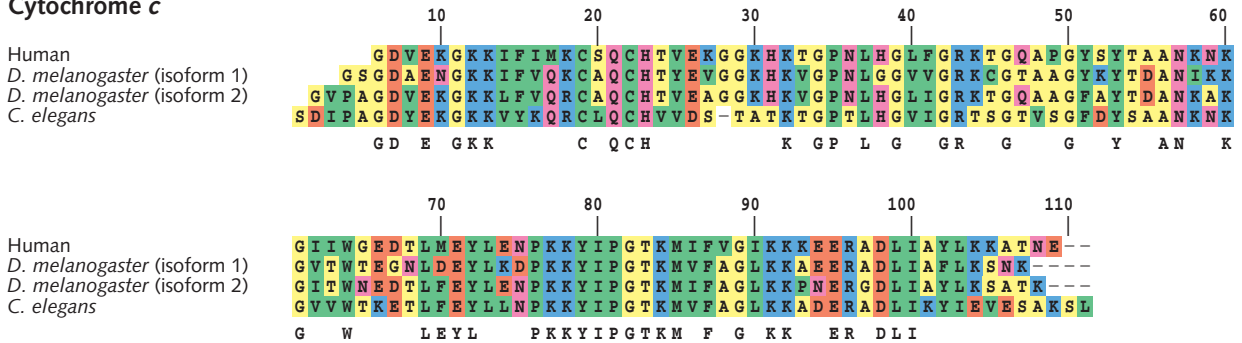
- First, it showed that a relatively large eukaryotic genome could be completed by the method of whole-genome shotgun sequencing (see Chapter 4).
- Second, the annotation of the sequence took place in a burst of activity: the intensive sessions of an 11-day 'jamboree' meeting held at Celera in November 1999. The ~45 participants included experts representing the inherited knowledge of a century of fruit-fly biology and local computer experts

BOX 2.13

Distribution of *D. melanogaster* genes

Chromosome arm	Size (Mb)	Number of protein genes	Density of protein genes (kb/gene)	Number of tRNA genes
X	22.2	2279	9.7	25
2L	22.4	2537	8.8	40
2R	20.8	2947	7.1	100
3L	23.8	2718	8.8	49
3R	27.9	3501	8.0	80
4	1.28	83	15.4	0

Data from Release 4.1: http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7227

Cytochrome c

involved in the sequencing. The flavour of the meeting is well described from the personal point of view by one of the participants, M. Ashburner.*

Homologous genes in humans, worms and flies

Once the genomes were sequenced and annotated, comparisons showed that homologues of many genes appear in all three species. Forty-four percent of protein-coding genes from *D. melanogaster* have human homologues, 25% of protein-coding genes from *C. elegans* have human homologues, and 23% of protein-coding genes from *D. melanogaster* have homologues in *C. elegans*.

Some proteins with common functions, such as cytochrome *c*, are expected to be quite similar in the different species (see Figure 2.24).

In other cases, different species have adapted homologous proteins to slightly different functions.

C. elegans and *D. melanogaster* are favourite subjects of developmental biologists. It is therefore of interest that a number of transcription regulators involved in developmental control are common to human, *D. melanogaster* and *C. elegans*. These include PAX (paired box domain) and HOX (homeobox domain) proteins.

Models of human disease

A model organism is a species in which an interesting feature of human biology – especially a disease – can be studied (Table 2.2). Ideally, a model organism is small and robust, has a relatively simple genome, is easy to maintain and manipulate (both physically and genetically) in the laboratory, has a short generation time, is safe to humans and comes with an extensive knowledge of its biology. As each model organism has its own strengths and limitations, different organisms are useful for different investigations.

In principle, biologists will use the simplest organism that illustrates the human feature of interest. It is easier to do experiments with yeast than with fruit flies. However, sometimes there is no choice. The only animal other than humans that is susceptible to leprosy is the armadillo, which satisfies few, if any, of the criteria of an ideal model organism.

Figure 2.24 Alignments of the amino acid sequences of cytochrome *c* from human, *D. melanogaster* (two isoforms) and *C. elegans*. For colour coding, see Figure 2.12.

A cross-table of numbers and percentages of homologous genes appears at <http://eugenics.org/all/hgsummary.html>

* Ashburner, M. (2006). *Won for All: How the Drosophila Genome Was Sequenced*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA.

Table 2.2 Human disease-associated genes shared with worms, flies and yeast

Affected area	Disease	Description	Gene	Similarity in		
				Worm	Fly	Yeast
Bones	Multiple exostoses	Ossification at tips of femur, pelvis, or ribs	<i>EXT1</i>	***	**	—
Blood	Leukaemia	Chronic myelogenous leukaemia, a blood cell cancer	<i>ABL1</i>	***	***	*
	Bruton agammaglobulinaemia	Lack of mature B cells	<i>BTK</i>	***	**	*
	Glucose-6-phosphate dehydrogenase deficiency	Drug- and stress-induced rupture of red blood cells	<i>G6PD</i>	****	****	****
Brain	Early-onset Alzheimer's disease	Common cause of mental retardation	<i>PS1</i>	**	**	—
	Fragile X syndrome		<i>FMR1</i>	**	—	—
	Juvenile Parkinson's disease		<i>PARK2</i>	***	**	*
Colon	Hereditary non-polyposis cancer	Polyps that become malignant	<i>MSH2</i>	***	***	***
	Adenomatous polyposis		<i>APC</i>	***	*	—
Ears	Hereditary deafness		<i>MYO15</i>	***	***	***
Eyes	Retinoblastoma	Cancer of the eye	<i>RB1</i>	*	*	—
Heart	Familial cardiac myopathy	Inherited cardiac disease Sometimes fatal cardiac arrhythmias	<i>MYH7</i>	***	***	***
	Long QT syndrome		<i>3-SCN5A</i>	***	**	*
Kidney	Polycystic kidney disease 2		<i>PKD2</i>	**	**	—
Liver	Wilson's disease	Build up of copper in cells, causing liver disease and other symptoms	<i>ATP7B</i>	***	***	***
Lung	Cystic fibrosis	Progressive disease of lungs and pancreas Caused by defects in <i>p53</i> gene, which can also cause cancer of the oesophagus, colon, brain, lung, breast and skin	<i>CFTR</i>	***	***	—
	Lung cancer		<i>p53</i>	*	—	—
Muscles	Duchenne's muscular dystrophy	Progressive atrophy of muscles	<i>DMD</i>	***	***	—
Pancreas	Pancreatic cancer		<i>MADH4</i>	***	*	—
	Pancreatic cancer		<i>RAS</i>	**	**	**
Prostate	Advanced cancer of the prostate	Caused by mutations in the <i>PTEN</i> gene, which can also cause cancer of the brain, endometrium and breast	<i>PTEN</i>	**	**	*
Skin	Xeroderma pigmentosum D	Early-onset skin cancer Soft tumours at many sites, plus skeletal and neurological defects	<i>XPD</i>	***	**	***
	Neurofibromatosis 1		<i>NF1</i>	***	*	**
Thyroid	Cancer of the thyroid	Multiple endocrine neoplasia type 2	<i>MEN2</i>	***	**	*

Based on data from Rubin, G.M. *et al.* (2000). Comparative genomics of the eukaryotes. *Science* 287, 2204–2215. Presentation adapted from <http://www.hhmi.org/genesweshare/e400.html>

There are two ways in which model organisms can contribute to understanding and treatment of human disease. The first is to observe homologues in the model organisms of genes implicated in human diseases. One can then study the effect on the model organism of mutation or knockout of the homologues. The second is to introduce a human gene into a model organism and discover its phenotypic effect. A model animal containing an active human gene makes it possible to screen libraries of compounds for potential drugs.

Table 2.2 shows some of the human disease-associated genes with homologues in *D. melanogaster*, *C. elegans* and *Saccharomyces cerevisiae*. The database Homophila (<http://superfly.ucsd.edu/homophila>) provides links between human disease-associated genes and *Drosophila* homologues.

Despite the fact that insects are not very closely related to mammals, fruit flies are useful in the study of human disease. The *D. melanogaster* genome contains homologues of human genes implicated in cancer and in cardiovascular, neurological, endocrinological, renal, metabolic and haematological diseases. Some of these homologues have different functions in humans and flies. Other human disease-associated genes can be introduced into, and studied in, the fly. For instance, the gene for human spinocerebellar ataxia type 3, when expressed in the fly, produces similar neuronal cell degeneration. There are now fly models for Parkinson's disease and malaria.

C. elegans also provides human disease models. Mutations in the human gene for presenilin-1 (*PS1*) are associated with familial early-onset Alzheimer's disease. Mutations in the homologous gene in *C. elegans*, *sel-12* (Figure 2.25) do show neurological defects, but in only a few neurons. Mutants do show more profound defects in egg laying, but this may be a secondary effect.

Although there are greater differences between the nervous systems of humans and *C. elegans* than between their machineries for respiratory energy transduction, the difference between the homologues shown here is not greater than the difference between the cytochrome *c* proteins. The relationship between sequence and function in proteins is full of surprises.

Genome Sequencing Projects

A list of active genome consortia and centres appears on a web page of the US National Library of Medicine (see Box 2.14). Of the 459 groups, a few are major players, whereas most are specialized to only a few projects. Thirteen institutions participate in more than 20 genome sequencing projects. Four hundred groups, distributed around the world, work on four or fewer genomes. As sequencing becomes less expensive and the equipment becomes more compact, more individual institutions are likely to acquire instruments and do at least their prokaryotic sequencing 'in house'.

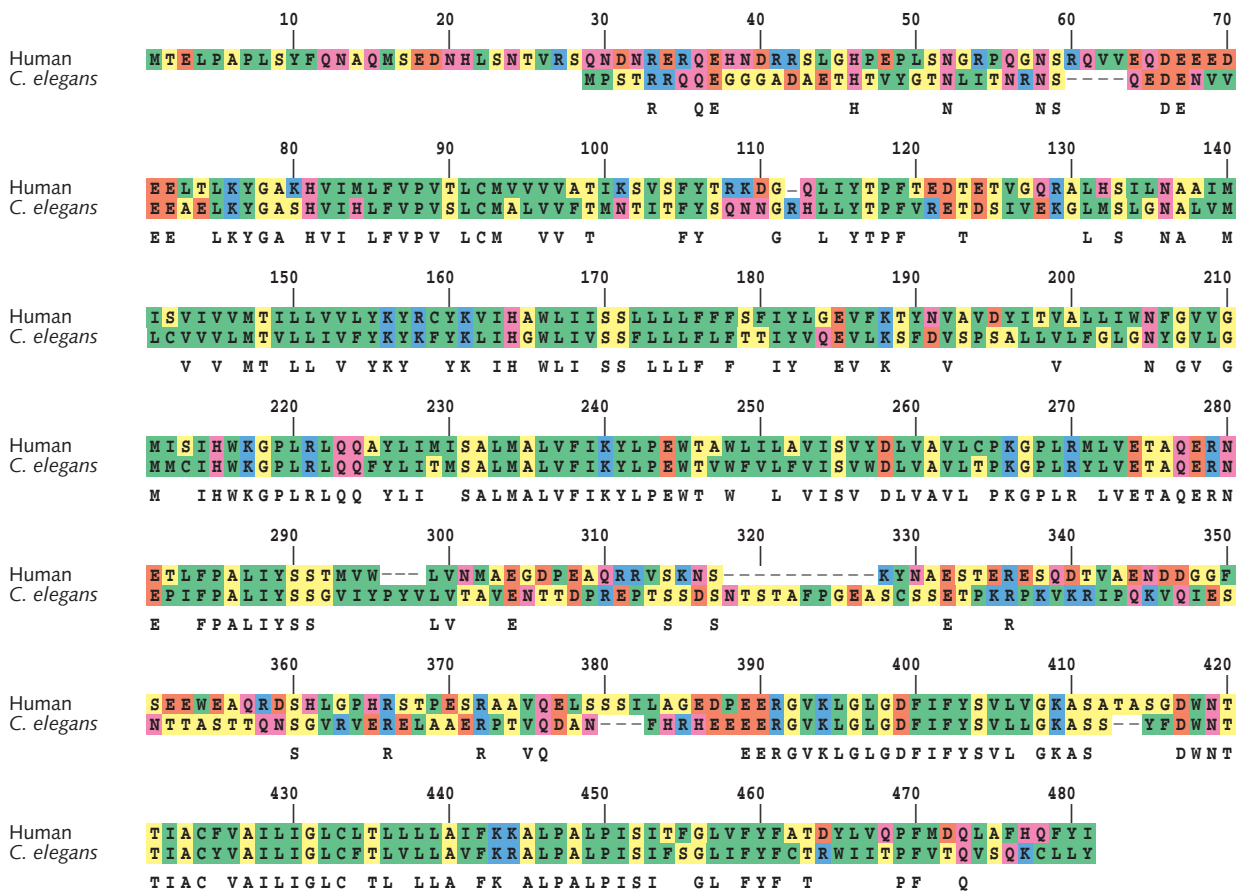


Figure 2.25 Alignment of the amino acid sequences of the human protein presenilin-1 and the *C. elegans* homologue SEL-12.

Genomes on the web

Completely sequenced genomes currently include several hundred bacteria, over 20 archaea, many viruses and organelles and over 30 eukarya (see Table 2.3). Almost all of the results are freely available on the web.

The site <http://www.ebi.ac.uk/2can/genomes/genomes.html> gives brief descriptions of the species represented and their scientific, clinical and/or practical (for example, baker's yeast) significance. A more complete description of the current status of genome projects appears at the site <http://www.genomesonline.org/>:

Current status of genome projects

Total completed	492
Prokaryotic completed or in progress	1093
Eukaryotic completed or in progress	640

Full-genome sequences are included in the general nucleic acid sequence archives.

Groups involved in many full-genome sequencing projects create and maintain databases focused on individual species. Scientists with specialized

**BOX
2.14****Major genome sequencing centres associated with 20 or more projects**

Institution	Number of projects
US Department of Energy Joint Genome Institute	240
The Institute for Genomic Research (TIGR)	204
Wellcome Trust Sanger Institute	93
J. Craig Venter Institute	89
Gordon and Betty Moore Foundation Marine Microbiology Initiative	86
Broad Institute	69
Genoscope	52
Washington University	50
Institut Pasteur	32
University of Oklahoma	24
Protist EST Program	22
Integrated Genomics	21

Source: <http://www.ncbi.nlm.nih.gov/genomes/static/lcenters.html>

expertise assume responsibility for curation and annotation of the data. The analysis includes identification of genes and assignment of function to their products. The results embed the genome in the context of other information about the individual species, arising from other data streams such as proteomics.

For instance, the Comprehensive Yeast Genome Database (CYGD), based at the Munich Information Centre for Protein Sequences (MIPS), organizes and presents information on sequence, structure, function and molecular interactions in *S. cerevisiae* (<http://mips.gsf.de/genre/proj/yeast/>). The MIPS group, one of the leading bioinformatics groups in Europe, has provided the nexus of computational support for numerous collaborative sequencing projects, including yeast and *A. thaliana*.

Several groups, including MIPS, have developed tools specialized for information retrieval and comparative analysis of genomes. Others include the ENSEMBL (at the Wellcome Trust Sanger Centre, Hinxton, UK) and University of California at Santa Cruz genome browsers (<http://www.ensembl.org> and <http://genome.ucsc.edu>).

The ENCODE project

The ENCODE project (*Encyclopedia of DNA elements*) has the ultimate goal of developing methods for comprehensive identification of functional regions of the human genome, including coding and regulatory regions. A selected portion of the human genome – 1%, about 30 Mb – will be the initial focus. The basic

Table 2.3 Completed eukaryotic genomes**Mammals**

Human	<i>Homo sapiens</i>
Chimpanzee	<i>Pan troglodytes</i>
Macaque	<i>Macaca mulatta</i>
Mouse	<i>Mus musculus</i>
Norway or brown rat	<i>Rattus norvegicus</i>
Dog	<i>Canis familiaris</i>
Cow	<i>Bos taurus</i>
African elephant	<i>Loxodonta africana</i>
Opossum	<i>Monodelphis domestica</i>
Platypus	<i>Ornithorhynchus anatinus</i>

Other chordates

Chicken	<i>Gallus gallus</i>
Frog	<i>Xenopus tropicalis</i>
Zebrafish	<i>Danio rerio</i>
Fugu fish	<i>Takifugu rubripes</i>
Green spotted pufferfish	<i>Tetraodon nigroviridis</i>
Sea squirt (tunicate)	<i>Ciona intestinalis</i>
Tunicate	<i>Ciona savignyi</i>

Plants

Thale cress	<i>Arabidopsis thaliana</i>
Rice	<i>Oryza sativa</i>
Maize (corn)	<i>Zea mays</i>
Lotus	<i>Lotus japonicus</i>
Barrel medic	<i>Medicago truncatula</i>
Tomato	<i>Lycopersicon esculentum</i>
Black cottonwood	<i>Populus trichocarpa</i>

Other eukaryotes

Fruit fly	<i>Drosophila melanogaster</i>
Anopheles mosquito	<i>Anopheles gambiae</i>
Dengue mosquito	<i>Aedes aegypti</i>
Honeybee	<i>Apis mellifera</i>
Nematode worm	<i>Caenorhabditis elegans</i>
Baker's yeast	<i>Saccharomyces cerevisiae</i>
Fission yeast	<i>Schizosaccharomyces pombe</i>
Fungus	<i>Candida glabrata</i> CBS138
Fungus	<i>Debaryomyces hansenii</i> CBS767
Microsporidian	<i>Encephalitozoon cuniculi</i>
Sea urchin	<i>Strongylocentrotus purpuratus</i>

approach will be comparative genomics and will involve both laboratory and computational analysis.

Regions corresponding to the selected human genome segments from 29 vertebrates will be sequenced (see Table 2.4). These data will illuminate each other. The ENCODE project will apply, improve and develop, as necessary, a variety of experimental and computational methods. Lessons learned from

Table 2.4 Species targeted by the ENCODE project

		Quality of sequencing		
		High	Medium	Unfinished
Class:				
<i>Actinopterygii</i>		Zebrafish		
<i>Amphibia</i>		Frog		
<i>Aves</i>		Chicken		
Class: Mammalia				
Order:	Suborder:			
<i>Monotremata</i>				Platypus
<i>Marsupialia</i>			Opossum	
<i>Proboscidea</i>				African elephant
<i>Insectivora</i>				Tenrec
<i>Xenarthra</i>				Armadillo
<i>Insectivora</i>				Hedgehog
<i>Insectivora</i>				Shrew
<i>Chiroptera</i>				Bat
<i>Artiodactyla</i>		Cow		
<i>Carnivora</i>		Dog		
<i>Carnivora</i>				Cat
<i>Rodentia</i>		Mouse		
<i>Rodentia</i>		Rat		
<i>Rodentia</i>				Guinea pig
<i>Lagomorpha</i>				Rabbit
Primates	<i>Prosimii</i>			Galago
Primates	<i>Prosimii</i>			Mouse lemur
Primates	<i>Platyrrhini</i>			Duski titi
Primates	<i>Platyrrhini</i>			Owl monkey
Primates	<i>Platyrrhini</i>			Marmoset
Primates	<i>Catarrhini</i>			Colobus
Primates	<i>Catarrhini</i>	Macaque		
Primates	<i>Catarrhini</i>			Baboon
Primates	<i>Hominidae</i>		Orangutan	
Primates	<i>Hominidae</i>	Chimpanzee		
Primates	<i>Hominidae</i>	Human		

High-quality sequences will be finished to state-of-the-art standards, including resolving difficult regions. Medium-quality sequences will have >8-fold coverage, with manual refinement of assembly. Unfinished sequences are whole-genome shotguns; the coverage may vary and assembly may be incomplete.

work with the selected subset will guide the scaling up of successful methods to analysis of entire genomes (see <http://www.genome.gov/10005107>).

Coordinating with ENCODE, the HapMap Project (see Chapter 1) focuses on variations among humans in ten of the ENCODE regions. Sequences from 48 individuals from different geographic origins have yielded 30 000 SNPs.

Analysis of function involves two steps: deciding whether a segment has functional significance and, if so, identifying what it does (see Figure 2.26). Approximately 5% of the human genome is conserved with respect to mouse and rat sequences. This 5% should have interesting functions (without implying that the other 95% does not). Only about one-third of this 5% is predicted to encode protein. Analysis of function will require treatment of both protein-coding and non-protein-coding regions.

Accordingly, the criteria for selection of regions for the ENCODE project included choosing regions with ranges of gene density and of non-exonic conservation with respect to the mouse sequence. The result is a set of 44 discrete regions, spread around different human chromosomes and the syntenic regions in other species. These include well-studied regions such as the α - and β -globin loci and the region containing *CFTR*, the gene for the cystic fibrosis transmembrane conductance regulator, for which sequence information from different species is known. Sequences of the ENCODE target regions can be aligned and compared (see Figure 2.26)

Chromosome	Approximate sizes of ENCODE regions (Mb) (gene of interest)
1	0.5
2	0.5, 0.5, 0.5, 0.5
4	0.5
5	0.5, 0.5, 1.0 (interleukin)
6	0.5, 0.5, 0.5, 0.5
7	0.5, 1.0, 1.1, 1.2, 1.9 (<i>CFTR</i>)
8	0.5
9	0.5
10	0.5
11	0.5, 0.5, 0.6, 0.5 (Apo cluster), 1.0 (β -globin)
12	0.5
13	0.5, 0.5
14	0.5, 0.5
15	0.5
16	0.5, 0.5, 0.5 (α -globin)
18	0.5, 0.5
19	1.0
20	0.5
21	0.5, 1.7
22	1.7
X	0.5, 1.2

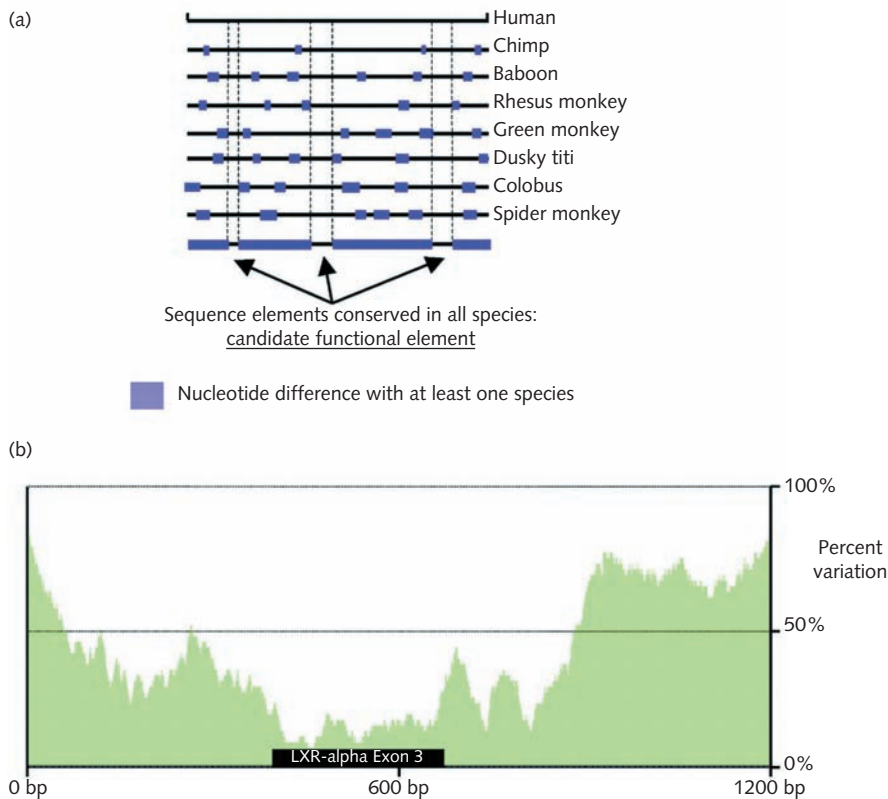
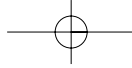
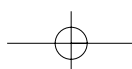


Figure 2.26 Patterns of variation in multiple sequence alignments can suggest regions of likely function. This diagram shows analysis of a 1200 bp region in primate genomes containing an exon of the liver X receptor α gene and flanking regions. This gene encodes a nuclear receptor responsive to elevated levels of intracellular cholesterol. (a) Human, reference sequence; purple, regions in which the sequence of the indicated region differs from at least one other species. The columns with no purple are conserved in all species and define regions likely to be functional. (b) Plot of % variation along the sequence. Regions of lowest variability correspond to the known exon. (A similar but not identical approach was used by E.A. Kabat and T.T. Wu in their classic work identifying complementarity-determining regions of antibodies from regions of hypervariability.)

From: Nobrega, M.A. & Pennacchio, L.A. (2004). Comparative genomic analysis as a tool for biological discovery. *J. Physiol.* 554, 31–39.

● RECOMMENDED READING

- Preface to a set of papers from a colloquium in honour of 100-year-old Ernst Mayr. Together these papers present a synoptic view of the field:
Hey, J., Fitch, W.M. & Ayala, F.J. (2005). Systematics and the origin of species: an introduction. *Proc. Natl. Acad. Sci. USA* 102, 6515–6519.
- Describes some of the current debate in biological taxonomy and the strengths and weaknesses of barcoding:
Moritz, C. & Cicero, C. (2004). DNA barcoding: promise and pitfalls. *PLoS Biol.* 2, e354.



- Not a scientific journal, but this issue contains a number of articles about influenza, placing the biomedical aspects of the problem in a more general context:
Foreign Affairs, issue of July/August 2005.
- Detailed review of work on genome comparisons and what they tell us about genome contents and evolution:
Miller, W., Makova, K.D., Nekrutenko, A. & Hardison, R.C. (2004). Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* 5, 15–56.
- Review of approaches to understanding the mysteries of higher mental processes:
Sforza, D.M. & Smith, D.J. (2003). Genetic and genomic strategies in learning and memory. *Curr. Genomics* 4, 475–485.
- One of the few articles to focus on this particular combination of disciplines:
Boguski, M.S. & Jones, A.R. (2004) Neurogenomics: at the intersection of neurobiology and genome sciences. *Nat. Neurosci.* 7, 429–433.
- How modern developments in biological data collection affect the use of model organisms in the study of human biology and disease:
Barr, M.M. (2003). Super models. *Physiol. Genomics* 13, 15–24.

● EXERCISES, PROBLEMS, AND WEBLEMS

Exercise 2.1 On a photocopy of Figure 2.5, (a) indicate the position of the human genome; (b) indicate the position of the pufferfish genome; (c) indicate the position of the rice (*Oryza sativa*) genome; (d) indicate the position of the yeast *S. cerevisiae* genome; (e) indicate the position of the *E. coli* genome; (f) indicate the range of sizes of mitochondrial genomes; and (g) indicate the range of sizes of chloroplast genomes.

Exercise 2.2 In H.G. Wells' 1898 novel *The War of the Worlds*, invaders from Mars are overcome by disease. Assuming that life on Mars developed independently of life on Earth, why is it unlikely that the Martians died of viral infections?

Exercise 2.3 On a photocopy of Figure 2.1, indicate estimates of the dates of the events at the points marked by asterisks.

Exercise 2.4 What RNA molecule is most closely linked to the *his* operon in *E. coli* (see Figure 2.10)?

Exercise 2.5 Human mitochondrial DNA is 16 569 bp long. A brain cell may contain 10 000 mitochondria. What fraction of the DNA in a brain cell is mitochondrial?

Exercise 2.6 Editing of mRNA transcribed from mitochondrial DNA often converts C → U. Consider conversion of codons by a single C → U change. In the standard genetic code, what changes in amino acids could this produce?

Exercise 2.7 Some antibiotics, for example streptomycin, block protein synthesis in bacteria but not cytoplasmic protein synthesis in eukaryotes. Mitochondria and chloroplasts contain their own protein-synthesizing machinery. Would you expect streptomycin to block mitochondrial protein synthesis? Explain your answer.

Exercise 2.8 On a photocopy of Figure 2.1, indicate by an arrow linking them the approximate positions of the source and destination organisms for the endosymbiotic origins of mitochondria and chloroplasts.

Exercise 2.9 Why could a mollusc that extracts chloroplasts from algae that it eats not simply let the chloroplasts reside in some body cavity (like the symbiotic bacteria in the guts of ruminant animals) rather than endocytosing them?

Exercise 2.10 The main *E. coli* chromosome contains 4 639 221 bp. The cell is roughly a cylinder about 0.1 μm in diameter and 0.2 μm long. If the length of an extended segment of DNA in the B conformation is 3.4 \AA per base pair (0.34 nm), what would be the diameter of the chromosome if it were geometrically a circle?

Exercise 2.11 On a photocopy of Figure 2.12, mark the positions that (a) contain the same amino acid in all three eukaryotes but differ in *S. aureus*; (b) contain the same amino acid in humans and chickens but are not the same in *Neurospora* and/or *S. aureus*; (c) contain the same amino acid in humans and *S. aureus*, but this amino acid does not appear at this position in both of the other two species.

Exercise 2.12 On a photocopy of Figure 2.12, indicate which regions of the amino acid sequences are encoded by which blocks of the nucleotide sequences in Figure 2.13. (This exercise is similar to Exercise 1.5.)

Exercise 2.13 On a photocopy of Figure 2.14, draw horizontal lines at the beginning and end of the Devonian (see Figure 2.4). What is earliest type of species to emerge after the split between neuroglobin and cytoglobin?

Exercise 2.14 Which animals in Figure 2.15 have the largest number of functional β -globin genes?

Exercise 2.15 Which animals in Figure 2.15 show a triplication of part of the β -region? Which genes make up the repeating unit?

Exercise 2.16 Which animals in Figure 2.15 show a pseudogene most closely related to a δ -globin?

Exercise 2.17 For the standard genetic code (p. 11), give an example of a pair of codons related by a synonymous single-site substitution (a) at the third position and (b) at the first position. (c) Give an example of a pair of codons related by a non-synonymous single-site substitution at the third position. (d) Can a change in the second position of a codon ever produce a synonymous mutation?

Exercise 2.18 The human and chimpanzee genomes are 96% identical. (a) How many individual bases of the human genome differ from the corresponding positions of the chimp genome? (b) Assume that the human genome is 3% coding and contains about 20 000 genes, and that all of the sequence differences are independent, single-base changes distributed randomly

throughout the genome (these assumptions are definitely *not true*). Estimate the fraction of genes mutated between humans and chimpanzees.

Exercise 2.19 On a photocopy of Figure 2.1, indicate where three whole-genome duplications are believed to have occurred.

Exercise 2.20 Why might the prescription of monoamine oxidase inhibitors to a child presenting with symptoms of anxiety, in case of suspected maltreatment, be contraindicated?

Problem 2.1 Replication of RNA viruses is error prone. It is estimated that the replication of HIV-1 introduces one mistake per replication of its $\sim 10^5$ bp genome. If the estimated generation time of HIV-1 in a human body is ~ 1 day and 10^{10} progeny viruses are produced per patient per day, and assuming that an AIDS patient is initially infected by viruses with identical genomes, estimate whether the patient will (a) generate a mutation at every possible site in the genome every day; and (b) generate mutations at every possible *pair* of sites in the genome every day.

Problem 2.2 On a copy of Figure 2.8(b), indicate the following interactions: (a) the positively charged guanidino group in zanamivir (left in Figure 2.8b) forms salt bridges with Glu116 and Glu225 in the neuraminidase active site; (b) hydroxyl groups of the glycerol moiety of zanamivir (at the right in Figure 2.8b) are hydrogen bonded to Glu274; (c) the carbonyl oxygen of the *N*-acetyl sidechain is hydrogen bonded to Arg149; (d) the methyl group of the *N*-acetyl sidechain makes hydrophobic interactions with Ile220 and Trp176.

Problem 2.3 Read the letters to *The New York Times* (17 September 2005) discussing the decision to publish the genome sequence of the 1918 pandemic strain of influenza virus. Summarize the arguments for and against publication.

Problem 2.4 From the partial sequences in Figure 2.13. (a) How many positions (necessarily in the coding regions) contain the same base in all three genomes? What percentage of positions contain the same base in all three genomes? (b) How many positions in the coding regions are common to human and chicken but different in *S. aureus*? To what percentage of coding positions does this correspond? (c) How many positions in the non-coding regions are common to human and chicken? To what percentage of non-coding positions does this correspond?

Problem 2.5 To what time frame can you date the duplication of the γ gene in the β -globin locus of primates? (See Figures 2.4, 2.15 and 2.16.)

Problem 2.6 Figure 2.27 shows an evolutionary tree of hominids. With reference to Figure 2.19, can you find similarities in the chromosome structures that confirm these relationships between human, common chimpanzee, gorilla and orangutan?

Problem 2.7 Which substitutions between rodent and human cytochrome *c* would *not* be considered conservative mutations?

Problem 2.8 For the following pairs of homologous proteins, what are the percentages of identical amino acids in an optimal alignment and what are the

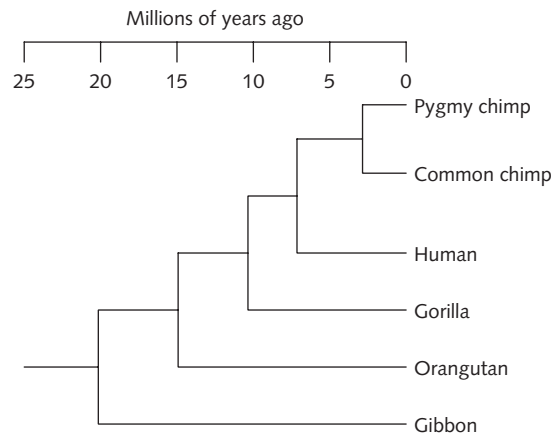


Figure 2.27 Phylogenetic tree of hominids.

percentages of identical residues or conservative substitutions in an optimal alignment: (a) the cytochrome *c* proteins of humans and *C. elegans*; and (b) presenilin-1 homologues of humans and *C. elegans*?

Weblem 2.1 Identify a virus and a prokaryote such that the genome of the virus is larger than the genome of the prokaryote (see Figure 2.5).

Weblem 2.2 Find examples of viruses with (a) a double-stranded circular DNA genome; (b) a double-stranded linear DNA genome; (c) a single-stranded (+)sense RNA genome; and (d) a single-stranded (–)sense RNA genome.

Weblem 2.3 What was the most common serotype of influenza in the USA during the 2004–2005 season (www.cdc.gov/flu/weekly)?

Weblem 2.4 Give an example of a species that represents the simplest known form of (a) metazoan; (b) deuterostome; (c) placoderm; and (d) eutherian.

Weblem 2.5 Mitochondrial DNA is often edited before translation. When a mitochondrial gene is transferred to the nucleus, there are two possibilities: (1) the copying is DNA→DNA. In this case, the nuclear version will initially have the mitochondrial DNA sequence and will require mutation to effect the changes introduced in the mitochondrion by the editing; or (2) the nuclear gene is reverse transcribed from the edited mitochondrial mRNA.

Compare the sequences of the nuclear gene for cytochrome oxidase II from mung bean (*Vigna radiata*) with mitochondrial sequences in related legumes, before and after editing. Do the data support either hypothesis?

Weblem 2.6 Align the following sequences: (1) the nuclear-encoded *rps14* gene from rice (*Oryza sativa*); (2) the mitochondrial-encoded *rps14* gene from broadbean (*Vicia faba*); (3) the nuclear-encoded *sdhB* gene from rice. Identify the leader sequence in the rice *rps14* gene, targeting mitochondrial import, that appears to have been borrowed from the rice *sdhB* gene.

Weblem 2.7 Make two histograms of *E. coli* genes, similar to that of Figure 2.9, showing in one of them the size distribution of genes appearing clockwise in the

genome and in the other the size distribution of genes appearing counterclockwise in the genome. Describe any systematic differences that appear.

Weblem 2.8 Many people believe that Rickettsiae are the closest extant relatives of the organism that, after endosymbiosis, gave rise to mitochondria. Rickettsiae are obligate aerobes. (a) Can you identify an enzyme that (1) catalyses an anaerobic function in mitochondria and (2) lacks a known homologue in Rickettsiae. (b) If you can find one, and if the rickettsial origin of mitochondria is a valid hypothesis, what are reasonable explanations of the presence of the anaerobic enzymatic function in mitochondria? (c) How would you test these explanations?

Weblem 2.9 In eukaryotes, the recombination rate per kilobase – the physical distance that corresponds to the genetic distance – varies among species by several orders of magnitude overall. (It also varies within each genome.) It depends primarily on overall genome size. You can find the data in computer-readable form through this book's Online Resource Centre (www.oxfordtextbooks.co.uk/orc/leskgenomics). Draw graphs of recombination rate against genome size, distinguishing data from different groups of organisms. (a) What relationship do you observe, i.e. colloquially, what is the shape of the curve? (b) Can you plot the data in a way that gives a linear relationship? (c) Do the data from the different groups of organisms follow the same relationship? If not, how do they differ? (d) What general conclusions can you draw?

Weblem 2.10 Add to Figure 2.13 the corresponding partial gene sequence from *N. crassa*. Describe its relationship to human, chicken and *S. aureus* sequences. In particular, answer questions analogous to those in Exercise 2.11 for the DNA sequences.

Weblem 2.11 Plot a histogram of the cumulative number of completed genome sequences in each year since 1995.

Weblem 2.12 What full-genome project is in progress, but not yet complete, for an organism in each of following categories: (a) fungus; (b) amphibian; (c) land plant; (d) insect (*not* a species of *Drosophila*); (e) primate (see: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj>).

Weblem 2.13 Collect and align sequences of the protein HSP70 from about six (of each) Gram-positive bacteria, proteobacteria, other Gram-negative bacteria and archaea. Identify an insertion common to proteobacteria and other Gram-negative bacteria but absent from Gram-positive bacteria and archaea. On this basis, sketch the topology of a phylogenetic tree relating Gram-positive bacteria, proteobacteria, other Gram-negative bacteria and archaea. Where do these results suggest placing the root of the prokaryotic tree?

Weblem 2.14 The genetic code used in translation of genes in animal mitochondria differs from the standard one. Was this feature simply inherited from the original symbiont that gave rise to the organelle, or did it arise subsequently by divergence? Determine (a) what genetic code is used by the living organism that appears to be the closest extant relative of the original

symbiont, *Rickettsia prowazekii*. (b) Do all animal mitochondria use the same variant of the standard genetic code? Based on these data, suggest an answer to the question.

Weblem 2.15 What is significant about the following species that justified sequencing their entire genome sequences? (a) *Plasmodium falciparum*; (b) *Aspergillus fumigatus*; (c) *Tropheryma whipplei*; (d) *Sulfolobus tokodaii*; (e) sea urchin; (f) *Ciona intestinalis*.

Weblem 2.16 In 1976, swine flu killed Private David Lewis, a soldier at Fort Dix, in central New Jersey South of Princeton, USA. What was the response of the US Government, under President Gerald Ford? What was the course of the potential epidemic? In retrospect, does the response appear to have been the right one? Could you have advised President Ford to make a different response? If so, based on what facts known at the time of Private Lewis's illness?

Weblem 2.17 Compute K_a/K_s ratios for the genes for the two isoforms of *D. melanogaster* cytochrome *c*. Does the divergence appear to have arisen from selection or drift?

Weblem 2.18 Compute K_a/K_s ratios for the genes for human presenilin-1 and the *C. elegans* homologue *sel-12*. Does the divergence appear to have arisen from selection or drift?

Weblem 2.19 The classification of tarsiers among primates is ambiguous. Palaeontology places tarsiers with lemurs, lorises and galagos, in the suborder prosimians. Some genetic relationships place tarsiers with monkeys, apes and humans. It is also possible that tarsiers form a separate prosimian infraorder. Find the structure of the β -globin region of the tarsier genome and explain what these data suggest about where, in Figure 2.16, the tarsiers belong.

Weblem 2.20 The human homologues of *C. elegans* NPR-1 are the neuropeptide Y receptors NPY1R, NYP2R, . . . What mutations are known in these human receptors and what, according to Online Mendelian Inheritance in Man (OMIM™), are their effects?

Weblem 2.21 In humans, cholesteryl ester transfer protein is important in controlling blood levels of high-density lipoproteins. Do homologues of this protein exist in (a) mouse; (b) rat; and (c) hamster?

Weblem 2.22 Mouse and rat cytochrome *c* have identical amino acid sequences. Do they have identical gene sequences also? If not, show the differences between the cytochrome *c* genes in mouse and rat. Distinguish between exons and introns.

Weblem 2.23 How many cytochrome *c* pseudogenes are there in the human genome?

Weblem 2.24 Which of the following genes have the same number of exons in human and mouse orthologues? (a) Haemoglobin A (human gene *HBA*); (b) cytochrome *c* (human gene *HCS*); (c) spermidine synthase (human gene *SRM*).

Weblem 2.25 Are any regions containing globin genes included in the choices of ENCODE regions studied by the HapMap project? If so, which ones?