

## CHAPTER 2

# Genome organization and evolution

### Learning goals

1. Knowing the basic sizes and organizing principles of simple and complex genomes.
2. Understanding how genomes are analysed, and the relationship of gene sequences to phenotypic features, including inherited diseases.
3. Recognizing the importance and the difficulty of deriving from a complete genome sequence the amino acid sequences of the proteins encoded, and assigning functions to these proteins.
4. Understanding how to find genes associated with inherited diseases, and how the availability of the complete human genome has changed such investigations.
5. Knowing the general ideas of the contents of particular genomes, and how the genomes of prokaryotes and eukarya differ systematically, and to appreciate the implications of general surveys of genomes of different organisms.
6. Realizing that published genomes record the characteristics of only a single individual; and that there is considerable variation within populations, and great variation between separated populations of organisms belonging to the same species.
7. Appreciating the power of DNA sequences in studying human history, including inference of human migration patterns, and as records of plant and animal domestication.
8. Recognizing the power of DNA sequencing for personal identification, its application in paternity and criminal cases, and the questions of social policy it raises.
9. Appreciating the power of comparative genomics to identify features responsible for differences between species—for instance, what is it that makes us human?

## Genomes and proteomes

.....  
 The numbers of characters in the *E. coli* genome and in the First Folio edition of Shakespeare's plays differ by less than 0.1%.  
 .....

The genome of a typical bacterium comes as a single DNA molecule of about 5 million characters, about the same as a fairly large book. If extended, the bacterial genome would be about 2 mm long. (It has to fit into a cell of diameter of about 0.001 mm). The DNA of higher organisms is organized into chromosomes—normal human cells contain 23 chromosome pairs. The total amount of genetic information per cell—the sequence of nucleotides of DNA—is very nearly constant for all members of a species, but varies widely between species (see Box across for longer list):

Organism	Genome size (base pairs)
Epstein–Barr virus	$0.172 \times 10^6$
Bacterium ( <i>E. coli</i> )	$4.6 \times 10^6$
Yeast ( <i>S. cerevisiae</i> )	$12.5 \times 10^6$
Nematode worm ( <i>C. elegans</i> )	$100.3 \times 10^6$
Thale cress ( <i>A. thaliana</i> )	$115.4 \times 10^6$
Fruit fly ( <i>D. melanogaster</i> )	$128.3 \times 10^6$
Human ( <i>H. sapiens</i> )	$3223 \times 10^6$

Different patterns of proteins also characterize different cells. The relationship between DNA content and protein content is not direct. Not all DNA codes for proteins. Conversely, some genes exist in multiple copies. In eukarya, many genes produce several different proteins by alternative splicing. Therefore, the amount of protein sequence information in a cell, much less the number and pattern of different proteins expressed, cannot easily be estimated from the genome size.

### Genes

A single gene coding for a particular protein corresponds to a sequence of nucleotides along one or more regions of a molecule of DNA. The DNA sequence is collinear with the protein sequence. In species for which the genetic material is double-stranded DNA, genes may appear on either strand. Bacterial genes are continuous regions of DNA. Therefore, the functional protein-coding unit of genetic sequence information from a bacterium is a string of  $3N$  nucleotides encoding a string of  $N$  amino acids, or a string of  $N$  nucleotides encoding a structural RNA molecule. Such a string, equipped with annotations, would form a typical entry in an archive of genetic sequences.

## Genome sizes

Organism	Number of base pairs	Number of genes	Comment
$\phi$ X-174	5386	10	virus infecting <i>E. coli</i>
Human-mitochondrion	16 569	37	subcellular organelle
Epstein-Barr virus (EBV)	172 282	80	cause of glandular fever
<i>Mycoplasma pneumoniae</i>	816 394	680	cause of cyclic pneumonia epidemics
<i>Rickettsia prowazekii</i>	1 111 523	878	bacterium cause of epidemic typhus
<i>Treponema pallidum</i>	1 138 011	1039	bacterium cause of syphilis
<i>Borrelia burgdorferi</i>	1 471 725	1738	bacterium cause of Lyme disease
<i>Aquifex aeolicus</i>	1 551 335	1749	bacterium from hot spring
<i>Thermoplasma acidophilum</i>	1 564 905	1509	archaeal prokaryote lacks cell wall
<i>Campylobacter jejuni</i>	1 641 481	1708	frequent cause of food poisoning
<i>Methanococcus jannaschii</i>	1 664 970	1783	archaeal prokaryote thermophile
<i>Helicobacter pylori</i>	1 667 867	1589	chief cause of stomach ulcers
<i>Haemophilus influenzae</i>	1 830 138	1738	bacterium cause of middle ear infections
<i>Thermotoga maritima</i>	1 860 725	1879	marine bacterium
<i>Archaeoglobus fulgidus</i>	2 178 400	2437	another archaeon
<i>Deinococcus radiodurans</i>	3 284 156	3187	radiation-resistant bacterium
<i>Synechocystis</i>	3 573 470	4003	cyanobacterium 'blue-green alga'
<i>Vibrio cholerae</i>	4 033 460	3890	cause of cholera
<i>Mycobacterium tuberculosis</i>	4 411 529	4275	cause of tuberculosis
<i>Bacillus subtilis</i>	4 214 814	4779	popular in molecular biology
<i>Escherichia coli</i>	4 639 221	4406	molecular biologists' all-time favourite
<i>Pseudomonas aeruginosa</i>	6 264 403	5570	largest prokaryote sequenced as yet
<i>Saccharomyces cerevisiae</i>	$12.1 \times 10^6$	6172	yeast, first eukaryotic genome sequenced
<i>Caenorhabditis elegans</i>	$95.5 \times 10^6$	19 099	the worm
<i>Arabidopsis thaliana</i>	$1.17 \times 10^8$	25 498	flowering plant (angiosperm)
<i>Drosophila melanogaster</i>	$1.8 \times 10^8$	13 601	the fruit fly
<i>Takifugu rubripes</i>	$3.9 \times 10^8$	30 000	puffer fish (fugu fish)
Human	$3.2 \times 10^9$	20 500	
Wheat	$16 \times 10^9$	30 000	
Salamander	$10^{11}$	?	
<i>Psilotum nudum</i>	$10^{11}$	?	whisk fern — a simple plant

In eukarya the nucleotide sequences that encode the amino acid sequences of individual proteins are organized in a more complex manner. Frequently one gene appears split into separated segments in the genomic DNA. An *exon* is a stretch of DNA retained in the mature messenger RNA (mRNA) that the ribosome translates into protein. An *intron* is an **intervening region** between two exons. Cellular machinery splices together the proper segments of initial RNA transcripts, based on signal sequences flanking the exons in the sequences themselves. Many introns are very long—in some cases substantially longer than the exons.

.....  
 The genes that  
 code for  
 proteins, and for  
 structural RNA  
 molecules,  
 present only the  
 static picture of  
 the genome.  
 .....

Control information organizes the expression of genes. Control mechanisms may turn genes on and off, or regulate gene expression more finely. Cascades of controls respond to concentrations of nutrients, or to stress. Regulatory networks orchestrate complex programmes of development during the lifetime of the organism.

Many control regions of DNA lie near the segments coding for proteins. They contain **signal sequences** that serve as binding sites for the molecules that transcribe the DNA sequence, or sequences that bind regulatory molecules that can *block* transcription. Bacterial genomes contain examples of contiguous genes, coding for several proteins that catalyse successive steps in an integrated sequence of reactions, all under the control of the same regulatory sequence. F. Jacob, J. Monod and E. Wollman named these **operons**. One can readily understand the utility of a parallel control mechanism.

Eukaryotic chromosomes contain complexes of DNA with histones. Chromatin remodelling is an important mechanism of transcriptional control. Reversible chemical modification of histones, by a variety of reactions including deacetylation, methylation, decarboxylation, phosphorylation, ubiquitinylation and sumoylation, leads to alterations of the DNA–histone interactions that render transcription initiation sites more or less accessible.

In animals, methylation of DNA provides the signals for tissue-specific expression of developmentally regulated genes. DNA methylation is stable during tissue differentiation, surviving cell division. When a cell divides, enzymes copy the methylation patterns, preserving the settings of the regulatory switches.

Products of certain genes cause cells to commit suicide—a process called **apoptosis**. Defects in the apoptotic mechanism leading to uncontrolled growth are observed in some cancers, and stimulation of these mechanisms is a general approach to cancer therapy.

The conclusion is that to reduce genetic data to individual coding sequences is to disguise the very complex nature of the inter-relationships among them, and to ignore the historical and integrative aspects of the genome. Robbins has expressed the situation unimprovably:

‘... Consider the 3.2 gigabytes of a human genome as equivalent to 3.2 gigabytes of files on the mass-storage device of some computer system of unknown design. Obtaining the sequence is equivalent to obtaining an image of the contents of that mass-storage device. Understanding the sequence is equivalent to reverse engineering that unknown computer system (both the hardware and the 3.2 gigabytes of software) all the way back to a full set of design and maintenance specifications.

...

'Reverse engineering the sequence is complicated by the fact that the resulting image of the mass-storage device will not be a file-by-file copy, but rather a streaming dump of the bytes in the order they were entered into the device. Furthermore, the files are known to be fragmented. In addition, some of the device contains erased files or other garbage. Once the garbage has been recognized and discarded and the fragmented files reassembled, the reverse engineering of the codes can be undertaken with only a partial, and sometimes incorrect, understanding of the CPU on which the codes run. In fact, deducing the structure and function of the CPU is part of the project, since some of the 3.2 gigabytes are the binary specifications for the computer-assisted-manufacturing process that fabricates the CPU. In addition, one must also consider that the huge database also contains code generated from the result of literally millions of maintenance revisions performed by the worst possible set of kludge-using, spaghetti-coding, opportunistic hackers who delight in clever tricks like writing self-modifying code and relying upon undocumented system quirks.'

—Robbins, R.J. (1992). Challenges in the human genome project. *IEEE Engineering in Medicine and Biology*, **11**, 25–34. (©1992 IEEE).

## Proteomics

An organism's genome gives a complete but static set of specifications of the potential life of that individual. The state of development of the organism, and its activity at the molecular level at any moment, depend primarily on the amounts and distribution of its proteins. The *proteome project* deals in an integral way with patterns of expression of proteins in biological systems, in ways that complement and extend genome projects.

What kinds of data would we like to measure, and what mature experimental techniques exist to determine them? The basic goal is a spatio-temporal description of the deployment of proteins in the organism. The rates of synthesis of different proteins vary among different tissues and different cell types and states of activity. Methods are available for efficient analysis of transcription patterns of multiple genes. However, because proteins 'turn over' at different rates, it is also necessary to measure proteins directly. High-resolution two-dimensional polyacrylamide gel electrophoresis (2D PAGE) shows the pattern of protein content in a sample. Mass-spectrometric techniques identify the proteins into which the sample has been separated. We shall develop these topics in Chapter 7.

Application of these methods provides a picture of the protein-based activity of an organism, as the genome provides a complete set of potential proteins. R. Simpson has drawn an analogy: if the genome is a list of the instruments in an orchestra, the proteome is the orchestra in the process of playing a symphony.

Historically the chemical problem of determining amino acid sequences of proteins directly was solved before the genetic code was established and before methods for determination of nucleotide sequences of DNA were developed. However, the amino acid sequences of an organism's proteins are inherent in its genome sequence, by virtue of the genetic code. Indeed, new protein sequence data are now being determined by translation of DNA sequences, rather than by direct sequencing of proteins.

Should any distinction be made between amino acid sequences determined directly from proteins and those determined by translation from DNA? First, we must assume that it is possible correctly to identify within DNA sequences the regions that encode

.....  
 F. Sanger's  
 sequencing of  
 insulin in 1955  
 first proved that  
 proteins had  
 definite amino  
 acid sequences,  
 a proposition  
 that until then  
 was  
 hypothetical.  
 .....

proteins. The pattern-recognition programs that address this question are subject to three types of errors: a genuine protein sequence may be missed entirely, or an incomplete protein may be reported, or a gene may be incorrectly spliced. Several variations on the theme add to the complexity: genes for different proteins may overlap, or genes may be assembled from exons in different ways, in different tissues or even in individual cells. Conversely, some genetic sequences that appear to code for proteins may in fact be defective or not expressed. *A protein inferred from a genome sequence is a hypothetical object until an experiment verifies its existence.*

In many cases the expression of a gene produces a molecule that must be modified within a cell, to make a *mature* protein that differs significantly from the one suggested by translation of the gene sequence. In many cases the missing details of **post-translational modifications**—the molecular analogues of body piercing—are quite important. Post-translational modifications include addition of ligands (for instance the covalently bound haem group of cytochrome *c*), glycosylation, methylation, excision of peptides and many others. Patterns of disulphide bridges—primary chemical bonds between cysteine residues—cannot be deduced directly from the amino acid sequence. In some cases, mRNA is edited before translation, creating changes in amino acid sequences that are not inferrable from the genes.

Cleavage of a peptide is a common post-translational modification. In some cases, cleavage converts an inactive form of a protein to an active one. The proteases active in digestion of our food are examples. In other cases, the effect is to promote correct folding. For instance, insulin is synthesized as a single-chain precursor that folds properly, after which excision of a peptide produces the mature oligomeric form.

Most post-translational cleavage reactions are carried out by proteases. Alternatively, **inteins** are proteins that have a ‘self-splicing’ activity. They autocatalytically excise internal peptides and join the ends. (In contrast, peptide excision from proinsulin leaves two chains that are *not* joined by a peptide bond.)

## Eavesdropping on the transmission of genetic information

How hereditary information is stored, passed on and implemented is perhaps *the* fundamental problem of biology. Three types of maps have been essential (see Box):

1. Linkage maps of genes
2. Banding patterns of chromosomes
3. DNA sequences

These maps represent three very different types of data. Genes, as discovered by Mendel, were entirely abstract entities. Chromosomes are physical objects, banding patterns their visible landmarks. Only with DNA sequences are we dealing directly with stored hereditary information in its physical form.

It was the very great achievement of biology during the last century to forge connections between these three types of data. The first steps—and giant strides

they were indeed—proved that, for any chromosome, the maps are one-dimensional arrays, and indeed that they are collinear. Any schoolchild now knows that genes are strung out along chromosomes, and that each gene corresponds to a DNA sequence. But the proofs of these statements earned a large number of Nobel prizes.

Splitting a long molecule of DNA—for example, the DNA in an entire chromosome—into fragments of convenient size for cloning and sequencing requires additional maps to report the order of the fragments, to facilitate assembly of the entire sequence from the sequences of the fragments. A **restriction endonuclease** is an enzyme that cuts DNA at a specific sequence, usually about 6 bp long. Cutting DNA with several restriction enzymes with different specificities produces sets of overlapping fragments. From the sizes of the fragments it is possible to construct a **restriction map**, stating the order and distance between the restriction endonuclease cleavage sites. A mutation in one of these cleavage sites will change the sizes of the fragments produced by the corresponding enzyme, allowing the mutation to be located in the map.

Restriction enzymes can produce fairly large pieces of DNA. Cutting the DNA into smaller pieces, which are cloned and ordered by sequence overlaps (as in the example using text on pages 19–21) produces a finer dissection of the DNA called a **contig map**.

#### Gene maps, chromosome maps and sequence maps

1. A **gene map** is classically determined by observed patterns of heredity. Linkage groups and recombination frequencies can detect whether genes are on the same or different chromosomes, and, for genes on the same chromosome, how far apart they are. The principle is that the farther apart two linked genes are, the more likely they are to recombine, by crossing over during meiosis. Indeed, two genes on the same chromosome but very far apart will appear to be unlinked. The unit of length in a gene map is the Morgan, defined by the relationship that 1 cM corresponds to a 1% recombination frequency. (We now know that 1 cM  $\sim 1 \times 10^6$  bp in humans, but it varies with the location in the genome and with the distance between genes.)
2. **Chromosome banding pattern maps** Chromosomes are physical objects. Banding patterns are visible features on them. The nomenclature is as follows: in many organisms, chromosomes are numbered in order of size, 1 being the largest. The two arms of chromosomes, separated by the centromere, are called the p (petite = short) arm and q (queue) arm. Regions within the chromosome are numbered p1, p2, ... and q1, q2 ... outward from the centromere. Subsequent digits indicate subdivisions of bands. For example, certain bands on the q arm of human chromosome 15 are labelled 15q11.1, 15q11.2, 15q12. Originally bands 15q11 and 15q12 were defined; subsequently 15q11 was divided into 15q11.1 and 15q11.2.

Deletions of substantial segments of DNA are observable in changes in banding patterns. (Smaller deletions are observable by fluorescent *in-situ* hybridization,

→ **Gene maps, chromosome maps and sequence maps (continued)**

or FISH; see page 77.) The observation of banding patterns was crucial to the identification of chromosomes as the vessels of heredity (see *Introduction to Genomics*, Chapter 1.)

Many deletions are associated with inherited diseases. For instance, deletion in the 15q region in the human are associated with Prader-Willi and Angelman syndromes. These syndromes have the interesting feature that the alternative clinical consequences depend on whether the affected chromosome is paternal or maternal. This observation of **genomic imprinting** shows that the genetic information in a fertilized egg is not simply the bare DNA sequences contributed by the parents. Chromosomes of paternal and maternal origin have different states of methylation, signals for differential expression of their genes. The process of modifying the DNA which takes place during differentiation in development is already present in the zygote.

3. **The DNA sequence itself** Physically a sequence of nucleotides in the molecule, computationally a string of characters A, T, G and C. Genes are regions of the sequence, in many cases interrupted by non-coding regions.

### Identification of genes associated with inherited diseases

In the past, the connections between chromosomes, genes and DNA sequences have been essential for identifying the molecular deficits underlying inherited diseases, such as Huntington disease or cystic fibrosis. Sequencing of the human genome has changed the situation radically.

Given a disease attributable to a defective protein:

- ◆ If we know the protein involved, we can pursue rational approaches to therapy.
- ◆ If we know the gene involved, we can devise tests to identify sufferers or carriers.
- ◆ In many cases, knowledge of the chromosomal location of the gene is unnecessary for either therapy or detection; it is required only for identifying the gene, providing a bridge between the patterns of inheritance and the DNA sequence. (This is not true of diseases arising from chromosome abnormalities.)

For instance, in the case of sickle-cell anaemia, we know the protein involved. The disease arises from a single point mutation in haemoglobin. We can proceed directly to drug design. We need the DNA sequence only for genetic testing and counselling. In contrast, if we know neither the protein nor the gene, we must somehow retrace the steps back to the gene from the phenotype, a process called **positional cloning** or **reverse genetics**. Positional cloning used to involve a kind of 'Tinker to Evers to Chance' cascade from the gene map to the chromosome map to the DNA sequence. Later we shall see how recent developments have short-circuited this process.

Patterns of inheritance identify the type of genetic defect responsible for a condition. Simple Mendelian inheritance patterns show, for example, that Huntington disease and cystic fibrosis are caused by single genes. To find the gene associated with cystic fibrosis it was necessary to begin with the gene map, using linkage patterns of heredity in affected families to localize the affected gene to a particular region of a particular chromosome. Knowing the general region of the chromosome, it was then possible to search the DNA of that region to identify candidate genes, and finally to pinpoint the particular gene responsible and sequence it (see Boxes, *Identification of the cystic fibrosis gene*, and *Positional cloning: finding the cystic fibrosis gene*). In contrast, many diseases do not show simple inheritance, or, even if only a single gene is involved, heredity creates only a predisposition, the clinical consequences of which depend on environmental factors. The full human genome sequence, and measurements of expression patterns, will be essential to identify the genetic components of these more complex cases.

### Mappings between the maps

A gene linkage map can be calibrated to chromosome banding patterns through observation of individuals with deletions or translocations of parts of chromosomes. The genes responsible for phenotypic changes associated with a deletion must lie within the deletion. Translocations are correlated with altered patterns of linkage and recombination.

There have been several approaches to coordinating chromosome banding patterns with individual DNA sequences of genes:

- ◆ In **fluorescent *in situ* hybridization** (FISH) a probe sequence is labelled with fluorescent dye. The probe is hybridized with the chromosomes, and the chromosomal location where the probe is bound shows up directly in a photograph (see Plate II). Typical resolution is  $\sim 10^5$  bp, but specialized new techniques can achieve high resolution, down to 1 kbp. Simultaneous FISH with two probes can detect linkage and even estimate genetic distances. This is important in species for which the generation time is long enough to make standard genetic approaches inconvenient. FISH can also detect chromosomal abnormalities.
- ◆ **Somatic cell hybrids** are rodent cells containing few, one, or even partial human chromosomes. (Chromosome fragments are produced by irradiating the human cells prior to fusion. Such lines are called **radiation hybrids**.) Hybridization of a probe sequence with a panel of somatic cell hybrids, detected by fluorescence, can identify which chromosome contains the probe. This approach has been superseded by use of clones of yeast, bacteria or phage containing fragments of human DNA in artificial chromosomes (YACs, BACs and PACs, i.e. yeast, bacterial and bacteriophage P1 artificial chromosomes).

Of course, with sequences from the human or other organisms for which the complete genome sequence is known, these methods are obsolete. Given a DNA sequence, one would just look it up—but mind the gaps!

### Identification of the cystic fibrosis gene

Cystic fibrosis, a disease known to folklore since at least the Middle Ages and to science for about 500 years, is an inherited recessive autosomal condition. Its symptoms include intestinal obstruction; reduced fertility including anatomical abnormalities (especially in males); and recurrent clogging and infection of lungs—the primary cause of death now that there are effective treatments for the gastrointestinal symptoms. Approximately half the sufferers die before age 25 years, and few survive beyond 50. Cystic fibrosis affects 1/2500 individuals in the American and European populations. Approximately 1/25 Caucasians carry a mutant gene, and 1/65 African-Americans. The protein that is defective in cystic fibrosis also acts as a receptor for uptake of *Salmonella typhi*, the pathogen that causes typhoid fever. Increased resistance to typhoid in heterozygotes—who do not develop cystic fibrosis itself but are carriers of the mutant gene—probably explains why the gene has not been eliminated from the population.

The pattern of inheritance showed that cystic fibrosis was the effect of a single gene. However, the actual protein involved was unknown. It had to be found via the gene.

Clinical observations provided the gene hunters with useful clues. It was known that the problem had to do with  $\text{Cl}^-$  transport in epithelial tissues. People had long recognized that children with excessive salt in their sweat—tasteable when kissing an infant on the forehead—were short-lived. Modern physiological studies showed that epithelial tissues of cystic fibrosis patients cannot reabsorb chloride. When closing in on the gene, the expected distribution, among tissues, of its expression, and of the type of protein implicated, were useful guides.

In 1989 the gene for cystic fibrosis was isolated and sequenced. This gene—called *CFTR* (cystic fibrosis transmembrane conductance regulator)—codes for a 1480 amino acid protein that normally forms a cyclic AMP-regulated epithelial  $\text{Cl}^-$  channel. The gene, comprising 24 exons, spans a 250 kb region. For 70% of mutant alleles, the mutation is a 3 bp deletion, deleting the residue Phe508 from the protein. This mutation is denoted del508. The effect of the deletion is defective translocation of the protein, which is degraded in the endoplasmic reticulum rather than transported to the cell membrane.

An *in utero* test for cystic fibrosis is based on recovery of foetal DNA. A PCR primer is designed to give a 154 bp product from the normal allele and a 151 bp product from the del508 allele.

Clinicians have taken advantage of the fact that the affected tissues of the airways are easily accessible, to experiment with gene therapy. Unfortunately, the results have so far been disappointing. Use of genetically engineered adenovirus sprayed into the respiratory passages to deliver the correct gene to epithelial tissues caused inflammation. An alternative approach now being pursued is to deliver the normal *CFTR* gene via liposomes.

### Positional cloning: finding the cystic fibrosis gene

The process by which the gene responsible for cystic fibrosis was found has been called *positional cloning* or *reverse genetics*.

- ◆ A search in family pedigrees for a linked marker showed that the cystic fibrosis gene was close to a known variable number tandem repeat (VNTR), DOCR-917. Somatic cell hybrids placed this on chromosome 7, band q3.
- ◆ Other markers found were linked more tightly to the target gene. It was thereby bracketed by a VNTR in the *MET* oncogene and a second VNTR, D7S8. The target gene lies 1.3 cM from *MET* and 0.9 cM from D7S8—localizing it to a region of approximately 1–2 million bp. A region this long could well contain 100–200 genes.
- ◆ The inheritance patterns of additional markers from within this region localized the target more sharply to within 500 kb. A technique called *chromosome jumping* made the exploration of the region more efficient.
- ◆ A 300 kb region at the correct distance from the markers was cloned. Probes were isolated from the region, to look for active genes, characterized by an upstream CCGG sequence. (The restriction endonuclease *HpaII* is useful for this step; it cuts DNA at this sequence, but only when the second C is not methylated, i.e. when the gene is active.)
- ◆ Identification of genes in this region by sequencing.
- ◆ Checking in animals for genes similar to the candidate genes turned up four likely possibilities. Checking these possibilities against a complementary DNA (cDNA) library from sweat glands of cystic fibrosis patients and healthy controls identified one probe with the right tissue distribution for the expected expression pattern of the gene responsible for cystic fibrosis. One long coding segment had the right properties, and indeed corresponded to an exon of the cystic fibrosis gene. Most cystic fibrosis patients have a common alteration in the sequence of this gene—a 3 bp deletion, deleting the residue Phe508 from the protein.

Proof that the gene was correctly identified included:

- ◆ 70% of cystic fibrosis alleles have the deletion. It is not found in people who are neither sufferers nor likely to be carriers.
- ◆ Expression of the wild-type gene in cells isolated from patients restores normal  $\text{Cl}^-$  transport.
- ◆ Knockout of the homologous gene in mice produces the cystic fibrosis phenotype.
- ◆ The pattern of gene expression matches the organs in which it is expected.
- ◆ The protein encoded by the gene would contain a transmembrane domain, consistent with involvement in transport.

## High-resolution maps

Formerly, genes were the only visible portions of genomes. Now, markers are no longer limited to genes with phenotypically observable effects, which are anyway too sparse for an adequately high-resolution map of the human genome. Now that we can interrogate DNA sequences directly, any features of DNA that vary among individuals can serve as markers, including the following.

- ◆ **Variable number tandem repeats (VNTRs)**, also called minisatellites. VNTRs contain regions 10–100 bp long, repeated a variable number of times—same sequence, different number of repeats. In any individual, VNTRs based on the same repeat motif may appear only once in the genome; or several times, with different lengths on different chromosomes. The distribution of the sizes of the repeats is the marker. Inheritance of VNTRs can be followed in a family and mapped to a disease phenotype like any other trait. VNTRs were the first genetic sequence data used for personal identification—genetic fingerprints—in paternity and in criminal cases.

Formerly, VNTRs were observed by producing **restriction fragment length polymorphisms (RFLPs)** from them. VNTRs are generally flanked by recognition sites for the same restriction enzyme, which will neatly excise them. The results can be spread out on a gel, and the distribution of their lengths detected by Southern blotting. However, it is much easier and more efficient to measure the sizes of VNTRs by amplifying them with PCR, and this method has replaced the use of restriction enzymes.

- ◆ **Short tandem repeat polymorphisms (STRPs)**, also called microsatellites. STRPs are regions of only 2–5 bp but repeated many times; typically 10–30 consecutive copies. They have several advantages as markers over VNTRs, one of which is a more even distribution over the human genome.

There is no reason why these markers need lie within expressed genes, and usually they do not. (The CAG repeats in the gene for huntingtin and certain other disease genes are exceptions.)

Panels of microsatellite markers greatly simplify the identification of genes. It is interesting to compare a recent project to identify a disease gene now that the human genome sequence is available, with such classic studies as the identification of the gene for cystic fibrosis (see Box: *Identification of a gene for Berardinelli–Seip syndrome*).

### Identification of a gene for Berardinelli–Seip syndrome

Berardinelli–Seip syndrome (congenital generalized lipodystrophy) is an autosomal recessive disease. Its symptoms include absence of body fat, insulin-resistant diabetes and enhanced rate of skeletal growth.

To determine the gene involved, a group led by J. Magré subjected DNA from members of affected families to linkage analysis and homozygosity mapping with

.....  
Distinguish:  
VNTRs are  
characteristics  
of genome  
sequences;  
RFLPs are  
artificial  
mixtures of  
short stretches  
of DNA created  
in the laboratory  
in order to  
identify VNTRs.  
.....

→ Identification of a gene for Berardinelli–Seip syndrome (*continued*)

a genome-wide panel containing ~400 microsatellite markers of known genetic location, with an average spacing of ~10 cM. In this procedure, a fixed panel of primers specific for the amplification and analysis of each marker is used to compare whole DNA of affected individuals with that of unaffected relatives. The measurements reveal the lengths of the repeats associated with each microsatellite. For every microsatellite, each observed length is an allele. Identifying microsatellite markers that are closely linked to the phenotype localizes the desired gene. The measurements are done efficiently and in parallel using commercial primer sets and instrumentation.

Two markers in chromosome band 11q13—D11S4191 and D11S987—segregated with the disease, and some affected individuals born from consanguineous families were homozygous for them. Finer probing, mapping with additional markers, localized the gene on chromosome 11 to a region of about 2.5 Mb.

There are 27 genes in the implicated 2.5 Mb region and its vicinity. Sequencing these genes in a set of patients identified a deletion of three exons in one of them. It was proved to be the disease gene by comparing its sequences in members of the families studied, and demonstrating a correlation between the presence of the syndrome and abnormalities in the gene. None of the other 26 genes in the suspect interval showed such correlated alterations.

Previous studies had identified a different gene, *BSCL1*, at 9q34, in other families with the same syndrome. The gene *BSCL1* has not yet been identified. It is possible that abnormalities in these two genes produce the same effect because their products participate in a common pathway which can be blocked by dysfunction of either.

The gene on chromosome 11, *BSCL2*, contains 11 exons spanning 14 kb. It encodes a 398-residue protein, named seipin. Observed alterations in the gene include large and small deletions, and single amino acid substitutions. The effects are consistent with loss of functional protein, either by causing frameshifts or truncation, or a missense mutation Ala212→Pro that credibly interferes with the stability of a helix or sheet in the protein structure.

Seipin has homologues in mouse and *Drosophila*. There are no clues to the function of any of the homologues, although they are predicted to contain transmembrane helices. What does provide suggestions about the aetiology of some aspects of the syndrome is the expression pattern, highest in brain and testis. This might be consistent with earlier endocrinological studies of Berardinelli–Seip syndrome that identified a problem in the regulation of release of pituitary hormones by the hypothalamus. Discovery of a protein of unknown function involved in the syndrome opens the way to investigation of what may well be a new biological pathway.

Additional mapping techniques deal more directly with the DNA sequences, and can short-circuit the process of gene identification:

- ◆ A *contig* or *contiguous clone map*, is a series of overlapping DNA clones of known order along a chromosome from an organism of interest—for instance, human—stored

in yeast or bacterial cells as YACs or BACs. A contig map can produce a very fine mapping of a genome. In a YAC, human DNA is stably integrated into a small extra chromosome in a yeast cell. A YAC can contain up to  $10^6$  bp. In principle, the entire human genome could be represented in 10 000 YAC clones. In a BAC, human DNA is inserted into a plasmid in an *E. coli* cell. (A plasmid is a small piece of double-stranded DNA found in addition to the main genome, usually but not always circular.) A BAC can carry about 250 000 bp. Despite their smaller capacities, BACs are preferred to YACs because of their greater stability and ease of handling.

- ◆ A **sequence tagged site** (STS) is a short, sequenced region of DNA, typically 200–600 bp long, that appears in a unique location in the genome. It need not be polymorphic. An STS can be mapped into the genome by using PCR to test for the presence of the sequence in the cells containing a contig map.

One type of STS arises from an **expressed sequence tag** (EST), a piece of cDNA (complementary DNA, i.e. a DNA sequence derived from the mRNA of an expressed gene). The sequence contains only the exons of the gene, spliced together to form the sequence that encodes the protein. cDNA sequences can be mapped to chromosomes using FISH, or located within contig maps.

How do contig maps and STS facilitate identifying genes? If you are working with an organism for which the full genome sequence is not known, but for which full contig maps are available for all chromosomes, you would identify STS markers tightly linked to your gene, then locate these markers in the contig maps.

## Picking out genes in genomes

Computer programs for genome analysis identify **open reading frames** or **ORFs**. An ORF is a region of DNA sequence that begins with an initiation codon (ATG) and ends with a stop codon. An ORF is a potential protein-coding region.

Approaches to identifying protein coding regions choose from or combine two possible approaches:

### 1. Detection of regions similar to known coding regions from other organisms.

These regions may encode amino acid sequences similar to known proteins, or may be similar to ESTs. Because ESTs are derived from mRNA, they correspond to genes known to be transcribed. It is necessary to sequence only a few hundred initial bases of cDNA to give enough information to identify a gene: characterization of genes by ESTs is like indexing poems or songs by their first lines.

### 2. *Ab initio* methods, that seek to identify genes from the properties of the DNA sequences themselves.

Computer-assisted annotation of genomes is more complete and accurate for bacteria than for eukarya. Bacterial genes are relatively easy to identify because they are contiguous—they lack the introns characteristic of eukaryotic genomes—and the intergene spaces are small. In higher organisms, identifying genes is harder. Identification of exons is one problem, assembling them is another. Alternative splicing patterns present a particular difficulty.

A framework for *ab initio* gene identification in eukaryotic genomes includes the following features:

- ◆ The initial (5') exon starts with a transcription start point, preceded by a core promoter site such as the TATA box typically ~30 bp upstream. It is free of in-frame stop codons, and ends immediately before a dinucleotide GT splice signal. (Occasionally a non-coding exon precedes the exon that contains the initiator codon.)
- ◆ Internal exons, like initial exons, are free of in-frame stop codons. They begin immediately after an AG splice signal and end immediately before a GT splice signal.
- ◆ The final (3') exon starts immediately after an AG splice signal and ends with a stop codon, followed by a polyadenylation signal sequence. (Occasionally a non-coding exon follows the exon that contains the stop codon.)

All coding regions have non-random sequence characteristics, based partly on codon usage preferences. Empirically, it is found that statistics of hexanucleotides perform best in distinguishing coding from non-coding regions. Starting from a set of known genes from an organism as a training set, pattern-recognition programs can be tuned to particular genomes.

Accurate gene detection is a crucial component of genome sequence analysis. This problem is an important focus of current research.

## 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). Of the 459 groups, a few are major players, and 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 more compact, more individual institutions are likely to acquire instruments and do at least their prokaryotic sequencing 'in-house'.

### Major genome sequencing centres associated with 20 or more projects

Institution	Number of projects
US Department of Energy Joint Genome Institute	435
J. Craig Venter Institute	302
The Institute for Genomic Research (TIGR)	296
Washington University	184
Broad Institute	157

→ Major genome sequencing centres associated with 20 or more projects (*continued*)

<b>Institution</b>	<b>Number of projects</b>
Gordon and Betty Moore Foundation Marine Microbiology Initiative	114
Wellcome Trust Sanger Institute	96
Genoscope	70
Baylor College of Medicine	47
Institut Pasteur	34
University of Tokyo	33
Integrated Genomics	24

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

### 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 below). Almost all the results are freely available on the Web. Many others are in progress (not counting assemblies from metagenomics sequencing projects):

#### A sample of 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>

##### Other chordates

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

→ A sample of completed eukaryotic genomes (*continued*)

**Higher 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 eukarya**

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>

Sequencing of the genomes of many other organisms is in progress. 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 (e.g. 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 and published	676
Prokaryotic completed or in progress	3000
Eukaryotic completed or in progress	1000

as of 1 January 2008

Groups involved in many full-genome sequencing projects create and maintain databases focused on individual species. Scientists with specialized 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 Center for Protein Sequences (MIPS), organizes and presents information on sequence, structure, function and molecular interactions in *Saccharomyces 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 *Arabidopsis 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>, <http://genome.ucsc.edu>).



### Web resources Genome databases

#### Lists of completed genomes:

<http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/allorg.html>

<http://www.ebi.ac.uk/genomes/mot/index.html>

<http://pir.georgetown.edu/pirwww/search/genome.html>

#### Organism-specific databases:

<http://www.unl.edu/stc-95/ResTools/biotools/biotools10.html>

<http://www-fp.mcs.anl.gov/~gaasterland/genomes.html>

<http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-db.html>

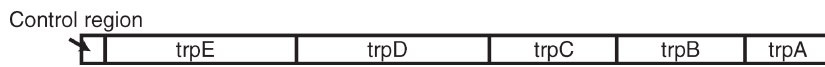
[http://www.bioinformatik.de/cgi-bin/browse/Catalog/Databases/Genome\\_Projects/](http://www.bioinformatik.de/cgi-bin/browse/Catalog/Databases/Genome_Projects/)

## Genomes of prokaryotes

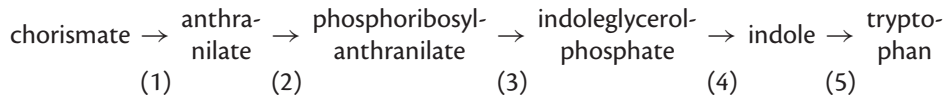
The genetic material of most prokaryotic cells takes the form of a large single circular piece of double-stranded DNA, usually less than 5 Mb long. In addition, cells may contain plasmids.

The protein-coding regions of bacterial genomes do not contain introns. In many prokaryotic genomes the protein-coding regions are partially organized into *operons*—tandem genes transcribed into a single mRNA molecule, under common transcriptional control. In bacteria, the genes of many operons code for proteins with related functions. For instance, successive genes in the *trp* operon of *E. coli* code for proteins that catalyse successive steps in the biosynthesis of tryptophan (see Fig. 2.1). In archaea, a metabolic relationship between genes in operons is less frequently observed.

The typical prokaryotic genome contains only a relatively small amount of non-coding DNA (in comparison with eukarya), distributed throughout the sequence. In *E. coli*, only ~11% of the DNA is non-coding.



**Fig. 2.1** The *trp* operon in *E. coli* begins with a control region containing promoter, operator and leader sequences. Five structural genes encode proteins that catalyse successive steps in the synthesis of the amino acid tryptophan from its precursor chorismate:



Reaction step (1): *trpE* and *trpD* encode two components of anthranilate synthase. This tetrameric enzyme, comprising two copies of each subunit, catalyses the conversion of chorismate to anthranilate. Reaction step (2): the protein encoded by *trpD* also catalyses the subsequent phosphoribosylation of anthranilate. Reaction step (3): *trpC* encodes another bifunctional enzyme, phosphoribosylanthranilate isomerase—indoleglycerolphosphate synthase. It converts phosphoribosyl anthranilate to indoleglycerolphosphate, through the intermediate, carboxyphenylaminodeoxyribose phosphate. Reaction steps (4) and (5): *trpB* and *trpA* encode the  $\beta$  and  $\alpha$  subunits, respectively, of a third bifunctional enzyme, tryptophan synthase (an  $\alpha_2\beta_2$  tetramer). A tunnel within the structure of this enzyme delivers, without release to the solvent, the intermediate produced by the  $\alpha$  subunit—indoleglycerolphosphate  $\rightarrow$  indole—to the active site of the  $\beta$  subunit—which converts indole  $\rightarrow$  tryptophan.

A separate gene, *trpR*, not closely linked to this operon, codes for the trp repressor. The repressor can bind to the operator sequence in the DNA (within the control region) only when binding tryptophan. Binding of repressor blocks access of RNA polymerase to the promoter, turning the pathway off when tryptophan is abundant. Further control of transcription in response to tryptophan levels is exerted by the attenuator element in the mRNA, within the leader sequence. The attenuator region (a) contains two tandem Trp codons and (b) can adopt alternative secondary structures, one of which terminates transcription. Levels of tryptophan govern levels of Trp-tRNAs, which govern the rate of progress of the tandem Trp codons through the ribosome. Stalling on the ribosome at the tandem Trp codons in response to low tryptophan levels reduces the formation of the mRNA secondary structure that terminates transcription.

### The genome of the bacterium *Escherichia coli*

*Escherichia coli*, strain K-12, has long been the workhorse of molecular biology. The genome of strain MG1655, published in 1997 by the group of F. Blattner at the University of Wisconsin, contains 4 639 221 bp in a single circular DNA molecule, with no plastids. Approximately 89% of the sequence codes for proteins or structural RNAs. An inventory reveals:

- ◆ 4284 protein-coding genes
- ◆ 122 structural RNA genes
- ◆ non-coding repeat sequences
- ◆ regulatory elements
- ◆ transcription/translation guides

- ◆ transposases
- ◆ prophage remnants
- ◆ insertion sequence elements
- ◆ patches of unusual composition, likely to be foreign elements introduced by horizontal transfer.

Analysis of the genome sequence required identification and annotation of protein-coding genes and other functional regions. Many *E. coli* proteins were known before the sequencing was complete, from many years of intensive investigation: 1853 proteins had been described before publication of the genome sequence. Other genes could be assigned functions from identification of homologues by searching in sequence databanks. The narrower the range of specificity of the function of the homologues, the more precise could be the assignment. Currently, over 60% of proteins can be assigned at least a general function (see Box). Other regions of the genome are recognized as regulatory sites, or mobile genetic elements, also on the basis of similarity to homologous sequences known in other organisms.

We visualize the contents of bacterial and organelle genomes as concentric circular diagrams, looking vaguely like 'tie-dyed' patterns. Complex patterns of colour coding serve as a visual 'feature table.' The site <http://wishart.biology.ualberta.ca/BacMap/index.html> contains an atlas of bacterial genome diagrams.<sup>1</sup>

.....  
*Introduction to Genomics*  
 contains several examples.  
 .....

#### Distribution of *E. coli* proteins among 22 functional groups

Functional class	Number	%
Regulatory function	45	1.05
Putative regulatory proteins	133	3.10
Cell structure	182	4.24
Putative membrane proteins	13	0.30
Putative structural proteins	42	0.98
Phage, transposons and plasmids	87	2.03
Transport and binding proteins	281	6.55
Putative transport proteins	146	3.40
Energy metabolism	243	5.67
DNA replication, recombination, modification and repair	115	2.68
Transcription, RNA synthesis, metabolism and modification	55	1.28

<sup>1</sup> Stothard, P., Van Domselaar, G., Shrivastava, S., Guo, A., O'Neill, B., Cruz, J., Ellison, M. and Wishart, D.S. (2005). BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucleic Acids Research*, **33**, D317–D320.

→ Distribution of *E. coli* proteins among 22 functional groups (continued)

Functional class	Number	%
Translation, post-translational protein modification	182	4.24
Cell processes (including adaptation, protection)	188	4.38
Biosynthesis of cofactors, prosthetic groups and carriers	103	2.40
Putative chaperones	9	0.21
Nucleotide biosynthesis and metabolism	58	1.35
Amino acid biosynthesis and metabolism	131	3.06
Fatty acid and phospholipid metabolism	48	1.12
Carbon compound catabolism	130	3.03
Central intermediary metabolism	188	4.38
Putative enzymes	251	5.85
Other known genes (gene product or phenotype known)	26	0.61
Hypothetical, unclassified, unknown	1632	38.06

Source: F.R. Blattner *et al.* (1997). The complete genome sequence of *Escherichia coli* K12. *Science*, **277**, 1453–1462.

The distribution of protein-coding genes over the genome of *E. coli* does not seem to follow any simple rules, either along the DNA or on different strands. Indeed, comparison of strains suggests that the genes are mobile.

The *E. coli* genome is relatively gene dense. Genes coding for proteins or structural RNAs occupy ~89% of the sequence. The average size of an ORF is 317 amino acids. If the genes were evenly distributed, the average intergenic region would be 130 bp; the observed average distance between genes is 118 bp. However, the sizes of intergenic regions vary considerably. Some intergenic regions are large. These contain sites of regulatory function, and repeated sequences. The longest intergenic region, 1730 bp, contains non-coding repeat sequences.

Approximately three-quarters of the transcribed units contain only one gene; the rest contain several consecutive genes, or operons. It is estimated that the *E. coli* genome contains 630–700 operons. Operons vary in size, although few contain more than five genes. The genes in operons tend to have related functions.

In some cases, the same DNA sequence encodes parts of more than one polypeptide chain. One gene codes for both the  $\tau$  and  $\gamma$  subunits of DNA polymerase III. Translation of the entire gene forms the  $\tau$  subunit. The  $\gamma$  subunit corresponds approximately to the N-terminal two-thirds of the  $\tau$  subunit. A frameshift on the ribosome at this point leads to chain termination 50% of the time, causing a 1:1 ratio of expressed  $\tau$  and  $\gamma$

subunits. There do not appear to be any overlapping genes in which different reading frames both code for expressed proteins.

In other cases, the same polypeptide chain appears in more than one enzyme. A protein that functions on its own as lipoate dehydrogenase is also an essential subunit of pyruvate dehydrogenase, 2-oxoglutarate dehydrogenase and the glycine cleavage complex.

Having the complete genome, we can examine the protein repertoire of *E. coli*. The largest class of proteins are the enzymes—approximately 30% of the total genes. Many enzymatic functions are shared by more than one protein. Some of these sets of functionally similar enzymes are very closely related, and appear to have arisen by duplication, either in *E. coli* itself, or in an ancestor or gene donor species. Other sets of functionally similar enzymes have very dissimilar sequences, and differ in specificity, regulation, or intracellular location.

Several features of *E. coli*'s generous endowment of enzymes give it a versatile metabolic competence, which allows it to grow and compete under varying conditions:

- ◆ It can synthesize all components of proteins and nucleic acids (amino acids and nucleotides), and cofactors.
- ◆ It has metabolic flexibility: Both aerobic and anaerobic growth are possible, utilizing different pathways of energy capture. It can grow on many different carbon sources. Not all metabolic pathways are active at any given time; the alternatives allow response to changes in conditions.
- ◆ Even for specific metabolic reactions there are many cases of multiple enzymes. These provide redundancy, and contribute to an ability to tune metabolism to varying conditions, through complementary control mechanisms.
- ◆ However, *E. coli* does not possess a complete range of enzymatic capacity. It cannot fix CO<sub>2</sub> or N<sub>2</sub>.

We have described here some of the *static* features of the *E. coli* genome and its protein repertoire. Current research has elucidated dynamic aspects, including the mechanisms that govern protein expression patterns in time and space.

### The genome of the archaeon *Methanococcus jannaschii*

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. The assumption was that the way to do this would be to find an organism adapted to the most different environment.

Deep-sea exploration has revealed environments as far from the familiar as those portrayed in science fiction. Hydrothermal vents are underwater volcanoes emitting hot lava and gases through cracks in the ocean floor. They create niches for communities of living things disconnected from the surface, which depend on the minerals exuded from the vent as inorganic nutrients. They support living communities of microorganisms that are the only known forms of life not dependent on sunlight, directly or indirectly, for their energy source.

The microorganism *Methanococcus jannaschii* was collected from a hydrothermal vent 2600 m deep off the coast of Baja California, Mexico, in 1983. It is a thermophilic

organism, surviving at temperatures from 48 to 94°C, with an optimum at 85°C. It is a strict anaerobe, capable of self-reproduction from inorganic components. Its overall metabolic equation is to synthesize methane from H<sub>2</sub> and CO<sub>2</sub>.

*Methanococcus jannaschii* belongs to the archaea, one of the three major divisions of life along with the bacteria and eukarya (see Fig. 1.2). The archaea comprise groups of prokaryotes, including organisms adapted to extreme environmental conditions such as high temperature and pressure, or high salt concentration. However, many archaea are not extremophiles.

The genome of *M. jannaschii* was sequenced in 1996 by The Institute for Genomic Research (TIGR). It was the first archaeal genome sequenced. It contains a large chromosome containing a circular double-stranded DNA molecule 1 664 976 bp long, and two extrachromosomal elements of 58 407 and 16 550 bp. There are 1784 predicted protein-coding regions, of which 1728 are on the chromosome, and 44 and 12 on the large and small extrachromosomal elements, respectively. Some RNA genes contain introns. As in other prokaryotic genomes there is little non-coding DNA.

*Methanococcus jannaschii* would appear to satisfy Luria's goal of finding our most distant extant relative. Comparison of its genome sequence with others shows that it is distantly related to other forms of life. Only 42% of the genes have been assigned a function. However, to everyone's great surprise, archaea are in some ways more closely related to eukarya than to bacteria! They are a complex mixture. In archaea, proteins involved in transcription, translation and regulation are more similar to those of eukarya. Archaeal proteins involved in metabolism are more similar to those of bacteria.

### The genome of one of the simplest organisms: *Mycoplasma genitalium*

*Mycoplasma genitalium* is an infectious bacterium, the cause of nongonococcal urethritis. Its genome was sequenced in 1995 by a collaboration of groups at TIGR, The Johns Hopkins University and The University of North Carolina. The genome is a single DNA molecule containing 580 070 bp. At the time, this was the smallest cellular genome yet sequenced. So far, *M. genitalium* is the closest we have to a **minimal organism**, the smallest capable of independent life. (Viruses, in contrast, require the cellular machinery of their hosts.)

The genome is dense in coding regions. A total of 468 genes have been identified as expressed proteins. Some regions of the sequence are gene rich, others gene poor, but overall 85% of the sequence is coding. The average length of a coding region is 1040 bp. As in other bacterial genomes, the coding regions do not contain introns. Further compression of the genome is achieved by overlapping genes. It appears that many of these have arisen through loss of stop codons.

The gene repertoire of *M. genitalium* includes some that encode proteins essential for independent reproduction, such as those involved in DNA replication, transcription and translation, plus ribosomal and transfer RNAs (rRNAs and tRNAs). Other genes are specific for the infectious activity, including adhesins that mediate binding to infected cells, other molecules for defence against the host's immune system, and a large number of transport proteins. As an adaptation to the parasitic lifestyle of the organism, there has been widespread loss of metabolic enzymes, including those

responsible for amino acid biosynthesis—indeed, one of the 20 amino acids is absent from all *M. genitalium* proteins (see Weblem 2.7).

## Genomes of eukarya

It is rare in science to encounter a completely new world containing phenomena entirely unsuspected. The complexity of the eukaryotic genome is such a world (see Box).

### Inventory of a eukaryotic genome

#### Moderately repetitive DNA

- ◆ Functional
  - dispersed gene families
    - e.g. actin, globin
  - tandem gene family arrays
    - rRNA genes (250 copies)
    - tRNA genes (50 sites with 10–100 copies each in human)
    - histone genes in many species
- ◆ Without known function
  - short interspersed elements (SINEs)
    - Alu is an example
    - 200–300 bp long
    - 100 000s of copies (300 000 Alu)
    - scattered locations (not in tandem repeats)
  - long interspersed elements (LINEs)
    - 1–5 kb long
    - 10–10 000 copies per genome
  - pseudogenes

#### Highly repetitive DNA

- ◆ Minisatellites
  - composed of repeats of 14–500 bp segments
  - 1–5 kb long
  - many different ones
  - scattered throughout the genome
- ◆ Microsatellites
  - composed of repeats of up to 13 bp
  - ~100s of kb long
  - ~10<sup>6</sup> copies/genome
  - most of the heterochromatin around the centromere

→ **Inventory of a eukaryotic genome (continued)**

◆ **Telomeres**

- contain a short repeat unit (typically 6 bp: TTAGGG in human genome, TTGGGG in *Paramecium*, TAGGG in trypanosomes, TTTAGGG in *Arabidopsis*)
- 250–1000 repeats at the end of each chromosome

In eukaryotic cells, the majority of DNA is in the nucleus, separated into bundles of nucleoprotein, the chromosomes. Each chromosome contains a single double-stranded DNA molecule. Smaller amounts of DNA appear in organelles—mitochondria and chloroplasts. The organelles originated as intracellular parasites. Organelle genomes usually have the form of circular double-stranded DNA, but are sometimes linear and sometimes appear as multiple circles. The genetic code by which organelle genes are translated differs from that of nuclear genes.

Nuclear genomes of different species vary widely in size (see page 71). The correlation between genome size and complexity of the organism is very rough. It certainly does not support any preconception that humans stand on a pinnacle. In many cases differences in genome size reflect different amounts of simple repetitive sequences, often referred to as 'junk DNA'.

In addition to variation in DNA content, eukaryotic species vary in the number of chromosomes and distribution of genes among them. Some differences in the distribution of genes among chromosomes involve translocations, or chromosome fragmentations or joinings. For instance, humans have 23 pairs of chromosomes; chimpanzees have 24. Human chromosome 2 is equivalent to a fusion of chimpanzee chromosomes 12 and 13 (see Fig. 2.2). The difficulty of chromosome pairing during mitosis in a zygote after such an event can contribute to the reproductive isolation associated with species separation.

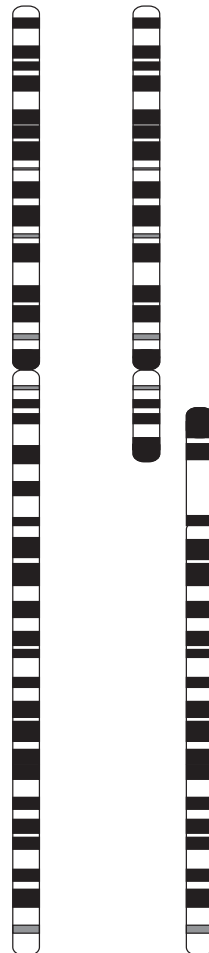
Other differences in chromosome complement reflect duplication or hybridization events. The wheat first used in agriculture, in the Middle East at least 10 000–15 000 years ago, is a diploid called einkorn (*Triticum monococcum*), containing 14 pairs of chromosomes. Emmer wheat (*T. turgidum* ssp. *dicoccum*), also cultivated since Palaeolithic times, and durum wheat (*T. turgidum* ssp. *durum*), are merged hybrids of relatives of einkorn with other wild grasses, to form tetraploid species. Additional hybridizations, with different wild grasses, gave hexaploid forms, including spelt (*T. aestivum* ssp. *spelta*), and modern common wheat *T. aestivum* ssp. *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* ssp. *durum*) 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 turgidum</i> ssp. <i>dicoccum</i>	AABB
Durum wheat	<i>Triticum turgidum</i> ssp. <i>durum</i>	AABB

Variety of wheat	Classification	Chromosome complement
Spelt	<i>Triticum aestivum</i> ssp. <i>spelta</i>	AABBDD
Common wheat	<i>Triticum aestivum</i> ssp. <i>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 *Triticum speltoides*, or a relative, D = genome of another wild grass, *Triticum tauschii* or a relative, R = genome of rye *Secale cereale*.

All 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 from *T. aestivum* ssp. *aestivum*.



**Fig. 2.2** Left: human chromosome 2. Right: matching chromosomes from chimpanzee.

Recent investigations of the history of wheat go beyond simple chromosome counts, to studies of relationships between species and subspecies at the genomic level. General results have measured the decay of synteny between orthologous regions after polyploidization, and mapping of insertions and deletions. Particular results include identification of mutations that confer properties favourable for agriculture. These properties include survival under stressful climate or soil conditions; and firmer attachment of grains to spikes, preserving them for harvesting against dispersal by wind.

A species that undergoes a revolutionary genomic change such as polyploidization is threatened with a penalty in the form of loss of genetic diversity. For the change must have occurred initially in only one or a few individuals, founders of new populations. Evidence for gene flow between domestic and wild forms of wheat suggests a mechanism for recovery and maintenance of genetic diversity.

.....  
 For a corresponding discussion of maize domestication see *Introduction to Genomics*, chapter 3.  
 .....

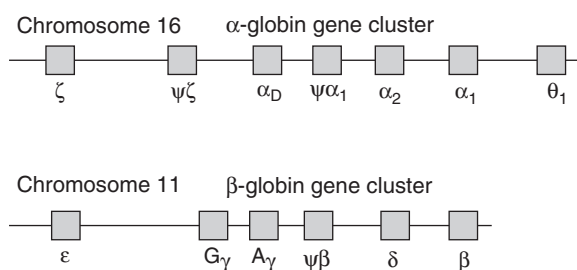
## Gene families

In addition to duplications of entire chromosomes, duplications of individual genes are common, as a result of unequal crossing-over. As a result, gene families within single chromosomes are common in eukarya.

Some family members are *paralogues*—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, human  $\alpha$  and  $\beta$  globin 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. The human globin gene cluster is a good example (see Box).

### The globin gene cluster

Human haemoglobin genes and pseudogenes appear in clusters on chromosomes 11 and 16. The normal adult human synthesizes primarily three types of globin chains:  $\alpha$  and  $\beta$  chains, which assemble into haemoglobin  $\alpha_2\beta_2$  tetramers; and myoglobin, a monomeric protein found in muscle. Other forms of haemoglobin, encoded by different genes, are synthesized in the embryonic and foetal stages of life. Other globins are unlinked; they arose long before this cluster diverged.



The globin gene cluster (continued)

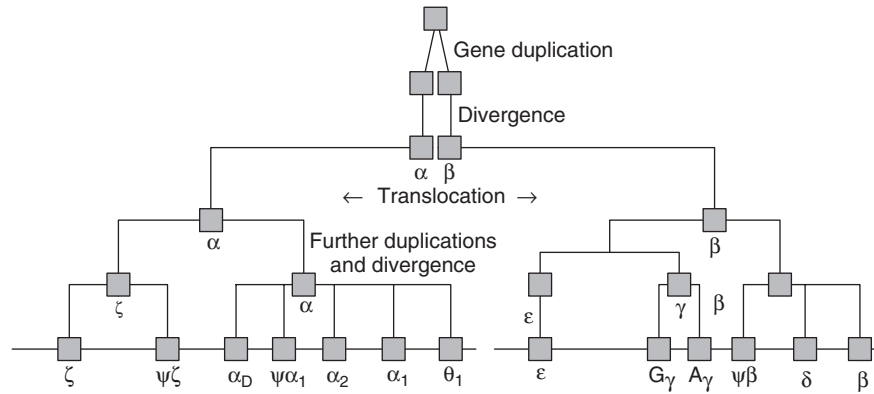
The  $\alpha$  gene cluster on chromosome 16 extends over 28 kbp. It contains three functional genes:  $\zeta$ , and two  $\alpha$  genes identical in their coding regions,  $\alpha_1$  and  $\alpha_2$ ; three pseudogenes,  $\psi\zeta$ ,  $\psi\alpha_1$  and  $\psi\alpha_2$ ; and another homologous gene the function of which is obscure,  $\theta_1$ . The  $\beta$  gene cluster on chromosome 11 extends over 50 kbp. It includes five functional genes:  $\epsilon$ ; two  $\gamma$  genes ( $G_\gamma$  and  $A_\gamma$ ), which differ in one amino acid,  $\delta$ ;  $\beta$ ; and one pseudogene,  $\psi\beta$ . The genes for myoglobin, neuroglobin and cytoglobin are unlinked from both of these clusters.

All human haemoglobin and myoglobin genes have the same intron/exon structure. They contain three exons separated by two introns:



Here E = exon and I = intron. The lengths of the regions in this figure reflect the human  $\beta$ -globin gene. This exon/intron pattern is conserved in most expressed vertebrate globin genes, including haemoglobin  $\alpha$  and  $\beta$  chains and myoglobin. In contrast, the genes for plant globins have an additional intron, genes for *Paramecium* globins have one fewer intron, and genes for insect globins contain none. The gene for human neuroglobin, a homologue expressed at low levels in the brain, contains three introns, like plant globin genes.

The distribution of haemoglobin genes and pseudogenes on the chromosomes appears to reflect their evolution via duplication and divergence.



The expression of these genes follows a strict developmental pattern. In the embryo (up to 6 weeks after conception) two haemoglobin chains are primarily synthesized— $\zeta$  and  $\epsilon$ —which form a  $\zeta_2\epsilon_2$  tetramer. Between 6 weeks after conception until about 8 weeks after birth, foetal haemoglobin— $\alpha_2\gamma_2$ —is the predominant species. This is succeeded by adult haemoglobin— $\alpha_2\beta_2$ .

Thalassaemias are genetic diseases associated with defective or deleted haemoglobin genes. Most Caucasians have four genes for the  $\alpha$  chain of normal adult haemoglobin, two alleles of each of the two tandem genes  $\alpha_1$  and  $\alpha_2$ . Therefore  $\alpha$ -thalassaemias can present clinically in different degrees of severity, depending on

→ **The globin gene cluster (continued)**

how many genes express normal  $\alpha$  chains. Only deletions leaving fewer than two active genes present as symptomatic under normal conditions. Observed genetic defects include deletions of both genes (a process made more likely by the tandem gene arrangement and repetitive sequences which make crossing-over more likely); and loss of chain termination leading to transcriptional 'readthrough', creating extended polypeptide chains which are unstable.

$\beta$ -Thalassaemias are usually point mutations, including missense mutations (amino acid substitutions), or nonsense mutations (changes from a triplet coding for an amino acid to a stop codon) leading to premature termination and a truncated protein, mutations in splice sites, or mutations in regulatory regions. Certain deletions including the normal termination codon and the intergenic region between  $\delta$  and  $\beta$  genes create  $\delta$ - $\beta$  fusion proteins.

### The genome of *Saccharomyces cerevisiae* (Baker's yeast)

Yeast is one of the simplest known eukaryotic organisms. Its cells, like our own, contain a nucleus and other specialized intracellular compartments. The sequencing of its genome, by an unusually effective international consortium involving ~100 laboratories, was completed in 1992. The yeast genome contains 12 057 500 bp of nuclear DNA, distributed over 16 chromosomes. The chromosomes range in size over an order of magnitude, from the 1352 kbp chromosome IV to the 230 kbp chromosome I.

The yeast genome contains 6172 predicted protein-coding genes, ~140 genes for rRNAs, 40 genes for small nuclear RNAs (snRNAs) and 275 tRNA genes. In two respects, the yeast genome is denser in coding regions than the known genomes of the more complex eukarya *Caenorhabditis elegans*, *Drosophila melanogaster* and human: (1) Introns are relatively rare, and relatively small. Only 231 genes in yeast contain introns. (2) There are fewer repeat sequences compared with more complex eukarya.

A duplication of the entire yeast genome appears to have occurred ~150 Mya. This was followed by translocations of pieces of the duplicated DNA and loss of one of the copies of most (~92%) of the genes.

Of the 6172 protein-coding genes, 4778 correspond to molecules to which a function can be assigned. About 1000 more contain some similarity to known proteins in other species. Another ~800 are similar to ORFs in other genomes that correspond to unknown proteins. Many of these homologues appear in prokaryotes. Only about one-third of yeast proteins have identifiable homologues in the human genome.

In taking censuses of genes, it has been useful to classify their functions into broad categories. The following classification of yeast protein functions is taken from <http://mips.gsf.de/genre/proj/yeast/Search/Catalogs/catalog.jsp>

Functional category	Number of proteins
Metabolism	1514
Energy	367
Cell cycle and DNA processing	1007
Transcription	1078
Protein synthesis	480
Protein fate (folding, modification, destination)	1154
Protein with binding function or cofactor requirement (structural or catalytic)	1048
Regulation of metabolism and protein function	249
Cellular transport, transport facilities and transport routes	1038
Cellular communication/signal transduction mechanism	234
Cell rescue, defence and virulence	554
Interaction with the environment	463
Transposable elements, viral and plasmid proteins	120
Cell fate	273
Development (systemic)	69
Biogenesis of cellular components	862
Cell type differentiation	452
Total functionally classified proteins	4778
Functionally unclassified proteins	1394

Yeast is a testbed for development of methods to assign functions to gene products. The search for homologues has been exhaustive and continues. Collections of mutants exist containing a knockout of every gene. (A unique sequence 'bar code' introduced into each mutant facilitates identification of the ones that grow under selected conditions.) Cellular localization and expression patterns are being investigated. Several types of measurements, including those based on activation of transcription by pairs of proteins that can form dimers, are producing catalogues of interprotein interactions.

### The genome of *Caenorhabditis elegans*

The nematode worm *C. elegans* entered biological research in the 1960s, at the express invitation of Sydney Brenner. He recognized its potential as an organism sufficiently complex to be interesting, yet simple enough to permit complete analysis, at the cellular level, of its development and neural circuitry.

The *C. elegans* genome, completed in 1998, provided the first full DNA sequence of a multicellular organism. The *C. elegans* genome contains ~97 Mbp of DNA distributed on paired chromosomes I, II, III, IV, V and X (see Table). There is no Y chromosome. Different genders in *C. elegans* appear in the XX genotype, a self-fertilizing hermaphrodite; and the XO genotype, a male.

#### 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

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 of yeast. The gene density is relatively high for a eukaryote, with ~1 gene/5 kb of DNA. Exons cover ~27% of the genome; the genes contain an average of five introns each. Approximately 25% of the genes are in clusters of related genes.

Many *C. elegans* proteins are common to other life forms. Others are apparently specific to nematodes: 42% of proteins have homologues outside the phylum; 34% are homologous to proteins of other nematodes; and 24% have no known homologues outside *C. elegans* itself. Many of the proteins have been classified according to structure and function (see Box).

#### *C. elegans*: 20 most common protein domains

Type of domain	Number
Seven-transmembrane spanning chemoreceptor	650
Eukaryotic protein kinase domain	410
Two-domain, C4 type zinc finger	240
Collagen	170

→ *C. elegans*: 20 most common protein domains (continued)

Type of domain	Number
Seven-transmembrane spanning receptor (rhodopsin-family)	140
C <sub>2</sub> H <sub>2</sub> -type zinc finger	130
C-type lectin	120
RNA recognition motif	100
C <sub>3</sub> HC <sub>4</sub> -type (RING finger) zinc fingers	90
Protein tyrosine phosphatase	90
Ankyrin repeat	90
WD domain G-β repeat	90
Homeobox domain	80
Neurotransmitter-gated ion channel	80
Cytochrome P450	80
Conserved C-terminal helicase	80
Short chain and alcohol dehydrogenases	80
UDP-glucuronosyl and UDP-glucosyl transferases	70
EGF-like domain	70
Immunoglobulin superfamily	70

Source: *C. elegans* genome consortium paper in the 11 December 1998 issue of *Science*.

Several kinds of RNA genes have been identified. The *C. elegans* genome contains 659 genes for tRNA, almost half of them (44%) on the X chromosome. Spliceosomal RNAs appear in dispersed copies, often identical. (*Spliceosomes* are the organelles that convert pre-mRNA transcripts to mature mRNA by excising introns, and stitching the exons together.) RNAs appear in a long tandem array at the end of chromosome I. 5S RNAs appear in a tandem array on chromosome V. Some RNA genes appear in introns of protein-coding genes.

The *C. elegans* genome contains many repeat sequences. Approximately 2.6% of the genome consists of tandem repeats. Approximately 3.6% of the genome contains inverted repeats; these appear preferentially within introns, rather than between genes. Repeats of the hexamer sequence TTAGGC appear in many places. There are also simple duplications, involving hundreds to tens of thousands of kilobases.

### The genome of *Drosophila melanogaster*

*Drosophila melanogaster*, the fruit fly, has been the subject of detailed studies of genetics and development for almost a century. Its genome sequence, the product of a collaboration between Celera Genomics and the Berkeley *Drosophila* Genome Project, was announced in 1999.

The chromosomes of *D. melanogaster* are nucleoprotein complexes, with variation in their structure along their lengths. Approximately one-third of the genome is contained in heterochromatin, highly coiled and compact (and therefore densely staining) regions flanking the centromeres. The other two-thirds is euchromatin, a relatively uncoiled, less compact form. Most of the active genes are in the euchromatin. The heterochromatin in *D. melanogaster* contains many tandem repeats of the sequence AATAACATAG, and relatively few genes.

The total chromosomal DNA of *D. melanogaster* contains ~180 Mbp. The sequence released in 1999 consists of the euchromatic portion, ~120 Mbp. In 2007 an additional ~15 Mbp of heterochromatin sequence was assembled.

The genome is distributed over five chromosomes: three large autosomes, a Y chromosome and a fifth tiny chromosome containing only ~1 Mbp of euchromatin. The fly's 13 601 genes are approximately double the number in yeast, but are fewer than in *C. elegans*, perhaps a surprise. The average density of genes in the euchromatin sequence is 1 gene/8 kb, much lower than the typical 1 gene/kb densities of prokaryotic genomes.

The heterochromatin contains at least ~250 protein-coding genes. They differ from typical euchromatic protein-coding genes by containing longer introns. Most of the intron sequences are repetitive, predominantly fragmented transposable elements.

Despite the fact that insects are not very closely related to mammals, the fly genome is useful in the study of human disease. It contains homologues of 289 human genes implicated in various diseases, including cancer, and 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 disease and malaria.

The non-coding regions of the *D. melanogaster* genome must contain regions controlling spatiotemporal patterns of development. The developmental biology of the fly has been studied very intensively. It is therefore an organism in which the study of the genomics of development should prove extremely informative.

### The genome of *Arabidopsis thaliana*

As a flowering plant, *A. thaliana* is a very distant relative of the other higher eukaryotic organisms for which genome sequences are available. It invites comparative analysis to identify common and specialized features.

*Arabidopsis thaliana* has a relatively small genome—146 Mb—distributed over five chromosomes. (The maize genome is almost 20 times larger.) The *Arabidopsis* Genome

.....  
 The compact genome was one reason for adopting *Arabidopsis* by the research community. *A. thaliana* is called 'the fruit fly of botany'.  
 .....

Initiative reported 115.4 Mbp of genomic DNA sequence in 2000. There are five pairs of chromosomes, containing 25 498 predicted genes. The genome is relatively compact, with 1 gene/4.6 kb on the average. This figure is intermediate between prokaryotes and *Drosophila*, and roughly similar to *C. elegans*. The genes of *Arabidopsis* are relatively small. Exons are typically 250 bp long, and introns relatively small, with mean length 170 bp. Typical of plant genes is an enrichment of coding regions in GC content.

### The *Arabidopsis thaliana* genome

	Chromosome					Total
	1	2	3	4	5	
Length (bp)	29 105 111	19 646 945	23 172 617	17 549 867	25 353 409	115 409 949
Number of genes	6543	4036	5220	3825	5874	25 498
Density (kb/gene)	4.0	4.9	4.5	4.6	4.4	
Mean gene length	2078	1949	1925	2138	1974	

Most *Arabidopsis* proteins have homologues in animals, but some systems are unique, among higher organisms, to plants. These include cell wall production and photosynthesis. It might be expected that these need special proteins that might not be shared with animals. Many proteins shared with animals have diverged widely since the last common ancestor. Typical of another difference between plants and animals, 25% of the nuclear genes have signal sequences governing their transport into organelles—mitochondria and chloroplasts—compared with 5% of mitochondrial-targeted nuclear genes in animals.

The *Arabidopsis* nuclear genome is relatively compact. Protein-coding genes contain an average of 5.4 exons, of average length 276 bp, separated by relatively short introns ~165 bp long. The intergenic spacing is also short, ~4.6 kb. A feature of plant genes is that the G+C content of exons (44%) is higher than that of introns (32%).

The structure of the *A. thaliana* genome reveals both local and genome-wide duplications. There were probably *three* polyploidizations, estimates of the dates of which vary widely. The ranges: 225–300 million years ago for the first, 150–170 million years ago for the second, and 25–40 million years ago for the most recent, have been suggested. In addition, local duplications have affected ~17% of genes. Close relatives, such as cabbage and cauliflower, have undergone additional polyploidizations during the 12 million years since they diverged from *Arabidopsis*.

Higher plants must integrate the effects of three genomes—nuclear, chloroplast and mitochondrial. The organelle genomes are much smaller:

**Gene distribution in *A. thaliana* between nucleus and organelles**

	<b>Nucleus</b>	<b>Chloroplast</b>	<b>Mitochondrion</b>
Genome size (kb)	125 100	154	367
Protein genes	25 498	79	58
Density (kb/protein gene)	4.5	1.2	6.25

Many genes for proteins synthesized by nuclear genes and transported to organelles appear to have originated in the organelles and been transferred to the nucleus.

Genome analysis must address questions of divisions of labour. Relative to animal cells, organelles in plant cells bear a greater metabolic burden, if only because of the activities of chloroplasts. Chloroplast genomes are relatively gene dense, with preserved gene order. In plant mitochondria, genes are more widely spaced, and recombination is more common. Mitochondrial and chloroplast genes contain fewer introns:

	<b>Nucleus</b>	<b>Chloroplast</b>	<b>Mitochondrial</b>
Genome:			
Genes containing introns:	80%	18.4%	12%

The *Arabidopsis* proteome contains many proteins specific to plants, including those involved in photosynthesis, and metabolism of components of cell walls:

- ◆ Plants have many special metabolic pathways, for photosynthesis; and for metabolism of cell wall components, alkaloids and growth regulators such as auxins and gibberellins. Complex metabolism requires the genome to encode a large and varied set of enzymes.
- ◆ Plants are threatened by pathogens, and have evolved defence mechanisms dissimilar from our immune system. One weapon against pathogens involves production of reactive oxygen species. Plants synthesize some defence molecules against animals, and others that attract pollinators. These have provided useful sources of flavours, fragrances, and drugs, encompassing traditional 'herbal medicine' and modern pharmacology.
- ◆ In keeping with the essential role of light in plant life, *Arabidopsis* has many light sensors, that regulate development and circadian responses.
- ◆ *Arabidopsis* is rich in genes that encode water-transporting channels, peptide-hormone transporters, metabolic and biosynthetic enzymes, and proteins involved in defence, detoxification and environmental sensing.

Comparing the proteins encoded in the nuclear genome of *Arabidopsis* with human proteins, the fraction of homologues observed varies with functional category. For protein synthesis, 60% of nuclear-encoded *Arabidopsis* genes have human homologues.

For transcription regulation, the figure is only 30%. It is not that transcription is poorly represented in plant genomes; it is just that plants do it differently. In fact, plants have several times as many transcription factors as the fruit fly. Although many components of the signal transduction pathways familiar from animals are absent in plants, plants have developed specific transcription factor families unknown in animals.

Many *Arabidopsis* genes are homologous to human genes implicated in disease. For instance, plants and animals have similar DNA repair systems, and *Arabidopsis* has a homologue of *BRCA2*. For some human disease-associated genes, the plant homologue is more similar to the human protein than those from fruit fly or *C. elegans*. Study of the function of the plant homologues will be illuminating, even though it is unlikely that *Arabidopsis* will be suitable for clinical trials of drugs intended for human use!

## The genome of *Homo sapiens* (The Human Genome)

Notice

PERSONS attempting to find a motive in this narrative will be prosecuted; persons attempting to find a moral in it will be banished; persons attempting to find a plot in it will be shot.

Mark Twain, Preface to *The Adventures of Huckleberry Finn*

In February 2001, the International Human Genome Sequencing Consortium and Celera Genomics published, separately, drafts of the human genome. On April 14, 2003 the finishing of the genome was announced—reduced error rate, closure of most gaps. This date was within a few days of the 50th anniversary of the publication of the Watson–Crick model for the structure of DNA.

The sequence amounts to  $\sim 3.2 \times 10^9$  bp, 30 times larger than the genomes of *C. elegans* or *D. melanogaster*. One reason for this disparity in size is that coding sequences form less than 5% of the human genome; repeat sequences over 50%. Perhaps the most surprising feature was the small number of genes identified. The finding of only about 20 000–25 000 genes suggests that alternative splicing patterns make a very significant contribution to our protein repertoire. It is estimated that  $\sim 35\%$  of genes have alternative splicing patterns.

The human genome is distributed over 22 chromosome pairs plus the X and Y chromosomes. The DNA contents of the autosomes range from 279 Mbp down to 48 Mbp. The X chromosome contains 163 Mbp and the Y chromosome only 51 Mbp.

The exons of human protein-coding genes are relatively small compared with those in other known eukaryotic genomes. The introns are relatively long. As a result many protein-coding genes span long stretches of DNA. For instance, the dystrophin gene, coding for a 3685 amino acid protein, is  $>2.4$  Mbp long.

### Protein-coding genes

Analysis of the human protein repertoire implied by the genome sequence has proved difficult because of the problems in reliably detecting genes, and because of alternative splicing patterns. Of the estimated 20 000–25 000 genes, the top categories in a functional classification are:

Function	Number	% of genome
Nucleic acid binding	2207	14.0%
DNA binding	1656	10.5%
DNA repair protein	45	0.2%
DNA replication factor	7	0.0%
Transcription factor	986	6.2%
RNA binding	380	2.4%
Structural protein of ribosome	137	0.8%
Translation factor	44	0.2%
Transcription factor binding	6	0.0%
Cell cycle regulator	75	0.4%
Chaperone	154	0.9%
Motor	85	0.5%
Actin binding	129	0.8%
Defence/immunity protein	603	3.8%
Enzyme	3242	20.6%
Peptidase	457	2.9%
Endopeptidase	403	2.5%
Protein kinase	839	5.3%
Protein phosphatase	295	1.8%
Enzyme activator	3	0.0%
Enzyme inhibitor	132	0.8%
Apoptosis inhibitor	28	0.1%
Signal transduction	1790	11.4%
Receptor	1318	8.4%
Transmembrane receptor	1202	7.6%
G-protein-linked receptor	489	3.1%
Olfactory receptor	71	0.4%
Storage protein	7	0.0%
Cell adhesion	189	1.2%
Structural protein	714	4.5%
Cytoskeletal structural protein	145	0.9%
Transporter	682	4.3%
Ion channel	269	1.7%
Neurotransmitter transporter	19	0.1%
Ligand binding or carrier	1536	9.7%
Electron transfer	33	0.2%
Cytochrome P450	50	0.3%

Function	Number	% of genome
Tumour suppressor	5	0.0%
Unclassified	4813	30.6%
Total	15 683	100.0%

Source: <http://www.ebi.ac.uk/protome/>

[under Functional classification of *H. sapiens* using Gene Ontology (GO):  
General statistics (InterPro proteins with GO hits)]

A classification based on structure revealed the most common types:

Protein	Number
Immunoglobulin and major histocompatibility complex domain	591
Zinc finger, C2H2 type	499
Eukaryotic protein kinase	459
Rhodopsin-like GPCR superfamily	346
Serine/Threonine protein kinase family active site	285
EGF-like domain	259
RNA-binding region RNP-1 (RNA recognition motif)	214
G-protein beta WD-40 repeats	196
Src homology 3 (SH3) domain	194
Pleckstrin homology (PH) domain	188
EF-hand family	185
Homeobox domain	179
Tyrosine kinase catalytic domain	173
Immunoglobulin V-type	163
RING finger	159
Proline rich extensin	156
Fibronectin type III domain	151
Ankyrin-repeat	135
KRAB box	133
Immunoglobulin subtype	128
Cadherin domain	118
PDZ domain (also known as DHR or GLGF)	117

Protein	Number
Leucine-rich repeat	113
Serine proteases, trypsin family	108
Ras GTPase superfamily	103
Src homology 2 (SH2) domain	100
BTB/POZ domain	99
TPR repeat	92
AAA ATPase superfamily	92
Aspartic acid and asparagine hydroxylation site	91

Source: <http://www.ebi.ac.uk/proteome/>

### Repeat sequences

Repeat sequences comprise >50% of the genome:

- ◆ Transposable elements, or interspersed repeats—almost half the entire genome! These include the LINES and SINEs (see Box).

#### Type of transposable elements in the human genome

Element	Size (bp)	Copy number	Fraction of genome
Short interspersed nuclear elements (SINEs)	100–300	1 500 000	13%
Long interspersed nuclear elements (LINEs)	6000–8000	850 000	21%
Long terminal repeats	15 000–110 000	450 000	8%
DNA transposon fossils	80–3000	300 000	3%

- ◆ Retroposed pseudogenes
- ◆ Simple ‘stutters’—repeats of short oligomers. These include the minisatellites and microsatellites. Trinucleotide repeats such as CAG, corresponding to glutamine repeats in the corresponding protein, are implicated in numerous diseases.
- ◆ Segmental duplications, of blocks of ~10–300 kb. Interchromosomal duplications appear on non-homologous chromosomes, sometimes at multiple sites. Some intra-chromosomal duplications include closely spaced duplicated regions many kilobases long of very similar sequence implicated in genetic diseases; for example Charcot-Marie-Tooth syndrome type 1A, a progressive peripheral neuropathy resulting from duplication of a region containing the gene for peripheral myelin protein 22.
- ◆ Blocks of tandem repeats, including gene families

## RNA

RNA genes in the human genome include:

1. A total of 497 RNA genes. One large cluster contains 140 tRNA genes within a 4 Mbp region on chromosome 6.
2. Genes for 28S and 5.8S rRNAs appear in a 44 kb tandem repeat unit of 150–200 copies. 5S RNA genes also appear in tandem arrays containing 200–300 genes, the largest of which is on chromosome 1.
3. Small nucleolar RNAs include two families of molecules that cleave and process rRNAs.
4. Spliceosomal snRNAs, including the U1, U2, U4, U5 and U6 snRNAs, many of which appear in clusters of tandem repeats of nearly identical sequences, or inverted repeats.



online  
resource  
centre

### Web resources Human genome information

#### Interactive access to DNA and protein sequences

<http://www.ensembl.org/>

#### Images of chromosomes, maps, loci

<http://www.ncbi.nlm.nih.gov/genome/guide/>

#### Gene map 99

<http://www.ncbi.nlm.nih.gov/genemap99/>

#### Overview of human genome structure

<http://hgrep.ims.u-tokyo.ac.jp>

#### Single-nucleotide polymorphisms

<http://snp.cshl.org/>

#### Human genetic disease

<http://www.ncbi.nlm.nih.gov/0mim/>

<http://www.geneclinics.org/profiles/all.html>

#### Social, ethical, legal issues

<http://www.nhgri.nih.gov/ELSI/>

## Single-nucleotide polymorphisms (SNPs) and haplotypes

All people, except identical siblings, have unique DNA sequences. Comparisons between unrelated individuals reveal overall differences between whole-genome sequences of ~0.1%. Many of the differences between individuals have the form of *Single-nucleotide polymorphisms*, or SNPs. There are also many short deletions.

A SNP (pronounced 'snip') is a genetic variation between individuals, limited to a single base pair which can be substituted, inserted or deleted. Sickle-cell anaemia is an example of a disease caused by a specific SNP: an A→T mutation in the β-globin gene changes a Glu→Val, creating a sticky surface on the haemoglobin molecule that leads to polymerization of the deoxy form.

Not all SNPs are linked to diseases. Many are not within functional regions (although the density of SNPs is higher than the average in regions containing genes). Some SNPs that occur within exons are mutations to synonymous codons, or cause substitutions that do not significantly affect protein function. Other types of SNPs can cause more than local perturbation to a protein: (1) A mutation from a sense codon to a stop codon, or vice versa, will cause either premature truncation of protein synthesis or 'readthrough'. (2) A deletion or insertion will cause a phase shift in translation.

The A, B and O alleles of the genes for blood groups illustrate these possibilities. A and B alleles differ by four SNP substitutions. They code for related proteins that add different saccharide units to an antigen on the surface of red blood cells.

Allele	Sequence	Saccharide
A	...gctggtgaccctt...	n-Acetylgalactosamine
B	...gctcgtcaccgcta...	Galactose
O	...cgtggt-accctt...	—

The O allele has undergone a mutation causing a phase shift, and produces no active enzyme. The red blood cells of type O individuals contain neither the A nor the B antigen. This is why people with type O blood are universal donors in blood transfusions. The loss of activity of the protein does not seem to carry any adverse consequences. Indeed, individuals of blood types B and O have greater resistance to smallpox.

Strong correlation of a disease with a specific SNP is advantageous in clinical work, because it is relatively easy to test for affected people or carriers. But if a disease arises from dysfunction of a specific protein, there ought to be many sites of mutations that could cause inactivation. However, a particular site may predominate if (1) all bearers of the gene are descendants of a single individual in whom the mutation occurred, and/or (2) the disease results from a *gain* rather than loss of a specific property, such as in the ability of sickle-cell haemoglobin to polymerize, and/or (3) the mutation rate at a particular site is unusually high, as in the Gly380→Arg mutation in the fibroblast growth receptor gene *FGFR3*, associated with achondroplasia (a syndrome including short stature).

In contrast, many independent mutations have been detected in the *BRCA1* and *BRCA2* genes, loci associated with increased disposition to early-onset breast and ovarian cancer. The normal gene products function as tumour suppressors. Insertion or deletion mutants causing phase shifts generally produce a missing or inactive protein. But it cannot be deduced *a priori* whether a novel *substitution* mutant in *BRCA1* or *BRCA2* confers increased risk or not.

Treatments of diseases caused by defective or absent proteins include:

1. **Providing normal protein** We have mentioned insulin for diabetes, and Factor VIII for the most common type of haemophilia. Another example is the administration of human growth hormone in patients with an absence or severe reduction in normal levels. Use of recombinant proteins eliminates the risk of transmission of AIDS through blood transfusions or of Creutzfeldt–Jakob disease from growth hormone isolated from crude pituitary extracts.

2. **Lifestyle adjustments that make the function unnecessary** Phenylketonuria (PKU) is a genetic disease caused by deficiency in phenylalanine hydroxylase, the enzyme that converts phenylalanine to tyrosine. Accumulation of high levels of phenylalanine causes developmental defects, including mental retardation. The symptoms can be avoided by a phenylalanine-free diet. Screening of newborns for high blood phenylalanine levels is legally required in the USA and many other countries.

3. **Gene therapy** to replace absent proteins is an active field of research.

Other clinical applications of SNPs reflect correlations between genotype and reaction to therapy (pharmacogenomics). For example, a SNP in the gene for *N*-acetyl transferase (*NAT-2*) is correlated with peripheral neuropathy—weakness, numbness and pain in the arms, legs, hands or feet—as a side effect of treatment with isoniazid (isonicotinic acid hydrazide), a common treatment for tuberculosis. Patients who test positive for this SNP are given alternative treatment.

SNPs are distributed throughout the genome, occurring on average every 2000 bp. Although they arose by mutation, many positions containing SNPs have low mutation rates, and provide stable markers for mapping genes.

Each of us bears an accumulated collection of SNPs reflecting mutations that occurred in our ancestors. Some constellations of SNPs are co-inherited as blocks. Others are not: mutations in *different* DNA molecules of diploid chromosomes become separated within a single generation, by assortment. Mutations on the *same* chromosome become separated more slowly, by recombination. Haploid sequences, such as most of the human Y chromosome or mitochondrial DNA, are not subject to recombination. Mutations in these sequences remain together.

Mutations in the same DNA molecule in diploid chromosomes will become unlinked by recombination events that occur between their loci. The greater the separation between two sites, the greater the frequency of recombination. However, recombination rates vary widely along the genome, by several orders of magnitude. SNPs on opposite sides of recombinational ‘hot spots’ are more likely to be separated in any generation. SNPs lying within recombination-poor (‘cold’) regions will tend to stay together.

In humans, many 100 kb regions tend to remain intact. They show the expected number of SNPs, but relatively few of the possible combinations. An average SNP density of 0.1%, or 1 SNP/kb, suggests ~100 SNPs per 100 kb. The genome of any individual may possess, or may lack, each of them, giving a very large number ( $2^{100}$ ) of possible combinations. However, many 100 kb regions show fewer than five combinations of SNPs. These discrete combinations of SNPs in recombination-poor regions define an individual’s **haplotype**, or ‘**haploid genotype**’ (see Box).

.....  
**Haplotypes are local combinations of genetic polymorphisms that tend to be co-inherited.**  
 .....

#### Haplotype distributions

Our individual genomes are characterized by a distribution of genetic markers. SNPs are convenient features to observe, and to study within and across populations. Although the overall density of SNPs in our genomes is ~1 SNP/5 kb,



#### Haplotype distributions (*continued*)

many 100 kb regions show only a few (typically 2–4) of the possible combinations of SNPs, suggesting that recombination is rare within the region. These segments, which remain intact, are separated by intervals in which recombination is more frequent.

The few discrete combinations of SNPs define the *haplotype* of an individual. The International HapMap project collects and curates haplotype distributions from several human populations.

Haplotypes are difficult to measure, because it is essential to determine which SNPs appear in the *same* DNA strand. Clearly, study of mixed samples from several individuals can determine the frequencies of individual SNPs but not their correlation into haplotypes. Even a sample containing both chromosomes from a diploid cell mixes the contributions of both copies of the region. However, mass spectral studies of amplified single-copy DNA molecules, produced by dilution, can identify the *combination* of SNPs appearing together on the same chromosome, allowing unambiguous haplotyping.

Haplotypes provide a very economical characterization of entire genomes. They simplify the search for genes responsible for diseases—or any other phenotype–genotype correlations. For field biologists, including anthropologists, haplotypes permit detection of migratory and interbreeding patterns in populations.

In looking for genes responsible for diseases or other phenotypic traits, haplotypes provide a magnifying glass. The goal is to correlate phenotype with genetic sequence. The target may be to identify one base out of  $3 \times 10^9$ . By correlating phenotype with *haplotype*, much less sequencing data must be collected to localize the site to within the typical length of a haplotype block, perhaps  $\sim 100$  kb, containing only a few genes. Another way to look at it is to regard boundaries between haplotype blocks as like the grooves in a bar of chocolate that permit it to be broken easily into bite-size fragments.

### Systematic measurements and collections of SNPs

Variations in human genomes are the subject of several large-scale projects.

The SNP consortium (<http://snp.cshl.org>) collects human SNPs. Its database currently contains nearly 5 million SNPs.

The International HapMap project collects and curates haplotype distributions from several human populations. SNPs are its raw material, from which it identifies the correlations among them. Phase I of the project, published in October 2005, had the goal of measuring the distributions of at least one SNP every 5 kb across the whole human genome. Blood samples were provided by 269 individuals from four continents (see Box). Over 1 million SNPs of significant frequency ( $>5\%$ ) were documented. In addition, 10 selected 500 kb regions were fully sequenced from 48 of the samples. Phase II will extend the analysis of the samples to determine an additional 4.6 million SNPs from the same individuals.

### Origin of samples for the International HapMap Project

Population origin	Location	Number of individuals	Relationships
Yoruba	Ibadan, Nigeria	90	30 parent–offspring trios
Northern and western European descent	Utah, USA	90	30 parent–offspring trios
Han Chinese	Beijing, China	45	
Japanese	Tokyo, Japan	44	

Why the choice of parent–offspring combinations? A difficulty in determining haplotypes in heterozygous regions of diploid chromosomes, is how to determine which SNPs lie on the *same* DNA. Comparison of parental and child sequences can sort the observed SNPs into haploid contributions.

Source: The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

The work of the International HapMap Consortium, together with other studies, show that:

- ◆ Most of the variations appear in all populations sampled. Some of the inter–population differences reflect different relative amounts of the same SNPs.
- ◆ However, a very few SNPs are unique to particular populations. For example, out of over 1 million SNPs, only 11 are consistently different between all individuals of European origin in the sample studied, and all individuals of Chinese or Japanese origin in the sample studied.
- ◆ The genomes of individuals from Japan and China are very similar, suggesting more recent common ancestry than other population pairs in the study.
- ◆ The X chromosome varies more between different populations than other chromosomes. This may arise from the fact that males contain only one X chromosome, the genes on which are therefore more subject to selective pressure. Recombinations of X chromosomes can occur, but only in females.
- ◆ Lengths of haplotype blocks vary among the different sources of samples. They tend to be shorter among populations from Africa, consistent with the idea of African origin of the human species.

The International HapMap Consortium paid due attention to ethical, legal and social issues. Informed consent of the donors preceded collection of samples. The procedure for informed consent involved not only individual agreement, but also community engagement, including interactive explanation of the project. Samples were labelled anonymously. In fact, more samples were collected than used (similar in some ways to the principle of issuing blank cartridges to a firing squad). Nevertheless, characteristics of a population constitutes personal information, the release of which may affect all individuals in the population, including those who were never asked to contribute a

sample, and even those who refused. For this reason, the HapMap Consortium did not collect medical information about the sample contributors, even under the protections of consent and anonymity.

## Genetic diversity in anthropology

SNP data are of great utility in anthropology, giving clues to historical variations in population size, and migration patterns.

Degrees of genetic diversity are interpretable in terms of the size of the founding population. Founders are the original set of individuals from whom an entire population is descended. Founders can be either original colonists, such as the Polynesians who first settled New Zealand, or merely the survivors of a near-extinction. Cheetahs show the effects of a population bottleneck, estimated to have occurred 10 000 years ago. All living cheetahs are as closely related to one another as siblings. Extrapolations of mitochondrial DNA variation in contemporary humans suggest a single maternal ancestor who lived 140 000–200 000 years ago. Calling her Eve suggests that she was the first woman. But fossil evidence for human ancestors goes back much longer. Mitochondrial Eve was the founder of a surviving population following a near-extinction.

There is now consensus that our species, *Homo sapiens*, arose in Africa approximately 100 000–150 000 years ago. The evidence for human origins in Africa is that contemporary genetic diversity is highest there. The mitochondrial DNA haplogroup L1 (see Box), believed to be the oldest haplotype that survives, is found in the KhoiSan of the Kalahari Desert in southern Africa, and in the Biaka pygmies of the central African rainforest.

### Human mitochondrial DNA haplogroups

Human mitochondrial DNA is a double-stranded closed circular molecule 16 569 bp long. It is inherited almost exclusively through maternal lines. A fertilized egg contains the mother's mitochondria. Although sperm contain mitochondria—essential to provide energy for their motility—the few paternal mitochondria that enter the egg are selectively eliminated. As a haploid entity, mitochondrial DNA is therefore not subject to recombination, and changes only by mutation.

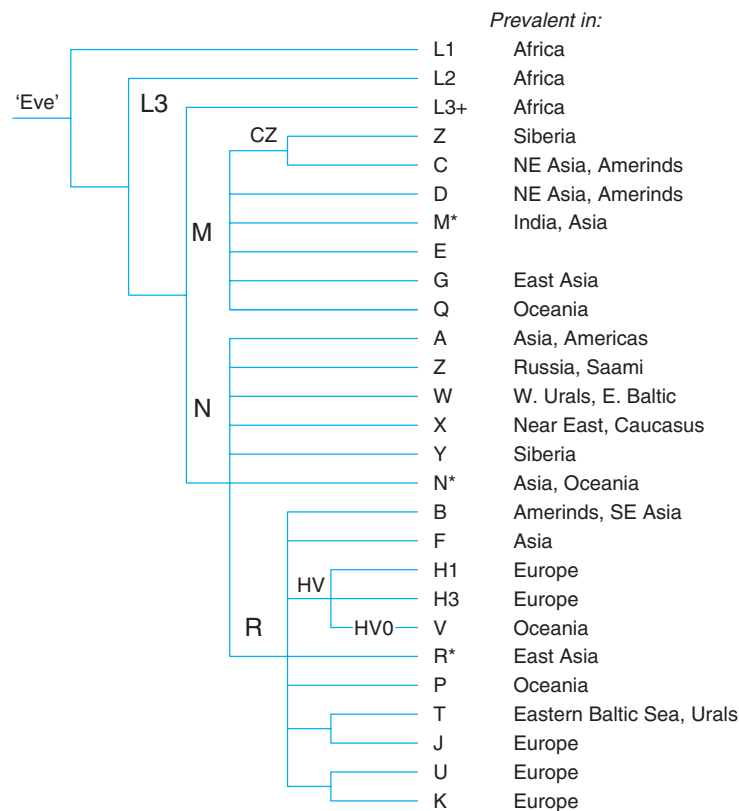
Mitochondrial DNA is estimated to adopt 1 mutation every 25 000 years. This gives a reasonable rate of divergence to trace human migration patterns. [Nuclear DNA mutates ~10 times more slowly than mitochondrial DNA because (1) histones protect it, (2) active repair mechanisms edit out some mutations and (3) the activity of mitochondria in respiration exposes the DNA to mutagenic oxygen radicals.]

Human mitochondrial DNA contains genes for 22 tRNAs, two rRNAs and 13 proteins. The major non-coding region is the control region, or D-loop, involved in regulation and initiation of replication. This region is ~1 kb long. It shows a higher rate of substitution than the rest of the mitochondrial genome, by a factor of about 4.

### Human mitochondrial DNA haplogroups (*continued*)

Different mitochondrial DNA sequences are associated with different populations. Mutations are referred to the first human mitochondrial DNA sequence determined, called the Cambridge Reference Sequence. Groups of related sequences are called haplogroups. (The distribution of the number of sequence differences between different individuals has peaks at ~70 for Africans and ~30 for non-Africans.) The original classification of sequence variants depended on changes in restriction sites (see Fig. 2.3). This was followed by explicit sequencing of the control region, focusing on its two highly polymorphic segments. For finest resolution, contemporary studies are now more frequently determining full mitochondrial DNA sequences, except in cases of ancient DNA where the best recoverable material may be fragmentary.

Several databases focus on human mitochondrial genomes, including MITOMAP (<http://www.mitomap.org>) and mtDB (<http://www.genpat.uu.se/mtDB>).



**Fig. 2.3** Phylogenetic tree of major mitochondrial haplogroups. The nomenclature began with a study of Native Americans, or Amerinds, and the letters A, B, C and D were assigned to them. Other letters were introduced, and (as more detailed sequencing data appeared) were subdivided as needed (HV0 was formerly called pre-V).

Migrations beginning approximately 60 000 years ago took our ancestors around the world, and continue to do so. Unlike modern population flows, documented in historical records, we depend on archaeological relics, contemporary genomics and linguistics to infer the timing, the routes, the numbers of individuals and even perhaps the motivation, of ancient migrations.

Population-specific SNPs are informative about migrations. Mitochondrial sequences provide information about female ancestors, and Y chromosome sequences provide information about male ancestors. For example, it has been suggested that the population of Iceland—first inhabited about 1100 years ago—is descended from Scandinavian males, and from females from both Scandinavia and the British Isles. Mediaeval Icelandic writings refer to raids on settlements in the British Isles.

Other crucial transitions in human social organization, such as turning from hunting to agriculture, can be seen in domestications of other species such as maize and dog. Genomic data are joining classical archaeological evidence to illuminate times and places of domestications (see Box).

.....  
[See Introduction to Genomics, chapter 3.](#)  
.....

#### Genetic analysis of cattle domestication

Animal resources are an integral and essential aspect of human culture. Analysis of DNA sequences sheds light on their historical development and on the genetic variety characterizing modern breeding populations.

Contemporary domestic cattle include those familiar in Western Europe and North America, *Bos taurus*; and the zebu of Africa and India, *Bos indicus*. The most obvious difference in external appearance is the humpback of the zebu. It has been widely believed that the domestication of cattle occurred once, about 8000–10 000 years ago, and that the two species subsequently diverged.

Analyses of mitochondrial DNA sequences from European, African and Asian cattle suggest, however, that (1) all European and African breeds are more closely related to each other than either is to Indian breeds, and (2) the two groups diverged about 200 000 years ago, implying recent independent domestications of different species. The similarity in physical appearance of the African and Indian zebu (and other similarities at the molecular level; for instance, VNTR markers in nuclear DNA) must then be attributable to importation of cattle from India to East Africa.

A fascinating relationship between human DNA sequences and language families has been investigated by L.L. Cavalli-Sforza and colleagues. These studies have proved useful in working out inter-relationships among American Indian languages. They confirm that the Basques, known to be a linguistically isolated population, have also been genetically isolated.

With the study of isolated populations, anthropological genetics provides data useful in medicine, as mapping disease genes is easier if the background variation is low. Genetically isolated populations in Europe include, in addition to the Basques, the Finns, Icelanders, Welsh and Lapps.

### Genetic diversity and personal identification

Variations in our DNA sequences give us individual fingerprints, useful for identification and for establishment of relationships, including but not limited to questions of paternity. The use of DNA analysis as evidence in criminal trials is now well established.

Genetic fingerprinting techniques were originally based on patterns of VNTRs, but have been extended to include analysis of other features including mitochondrial DNA sequences.

For most of us, all our mitochondria are genetically identical, a condition called homoplasmy. However, some individuals have mitochondria with different DNA sequences; called heteroplasmy. Such sequence variation in a disease gene can complicate the observed inheritance pattern of the disease.

The most famous case of heteroplasmy involved Tsar Nicholas II of Russia. After the revolution in 1917 the Tsar and his family were taken to exile in Yekaterinburg in Central Russia. During the night of 16–17 July 1918, the Tsar, Tsarina Alexandra, at least three of their five children, plus their physician and three servants who had accompanied the family, were killed, and their bodies buried in a secret grave. When the remains were rediscovered, assembly of the bones and examination of the dental work suggested—and sequence analysis confirmed—that the remains included an expected family group. The identity of the remains of the Tsarina were proved by matching the mitochondrial DNA sequence with that of a maternal relative, Prince Philip, Chancellor of the University of Cambridge, Duke of Edinburgh and grandnephew of the Tsarina.

However, comparisons of mitochondrial DNA sequences of the putative remains of Nicholas II with those of two maternal relatives revealed a difference at base 16 169: the Tsar had a C and the relatives a T. Extreme political and even religious sensitivities mandated that no doubts were tolerable. Further tests showed that the Tsar was heteroplasmic; T was a minor component of his mitochondrial DNA at position 16 169. To confirm the identity beyond any reasonable question, the body of Grand Duke Georgij, brother of the Tsar, was exhumed, and was shown to have the same rare heteroplasmy.

### Evolution of genomes

The availability of complete information about genomic sequences has redirected research (see Box). A general challenge in analysis of genomes is to identify ‘interesting events’. A background mutation rate in coding sequences is reflected in *synonymous* nucleotide substitutions: changes in codons that do not alter the amino acid. With this as a baseline, one can search for instances in which there are significantly higher rates of *non-synonymous* nucleotide substitutions: changes in codons that cause mutations in the corresponding protein. (Note, however, that synonymous changes are not necessarily selectively neutral.)

**Distribution of genome projects**

Organism	Complete	Draft assembly	In progress	Total
Prokaryotes	622	453	457	1532
Archaea	49	4	29	82
Bacteria	573	449	428	1450
Eukarya	22	128	174	324
Animals	4	53	82	139
Mammals	2	20	23	45
Birds		1	2	3
Fishes		3	6	9
Insects	1	19	17	37
Flatworms	1	1	3	4
Roundworms	1	4	12	17
Amphibians			2	2
Reptiles			1	1
Other animals		6	19	25
Plants		6	32	40
Land plants	2	4	25	31
Green Algae		2	7	9
Fungi	10	49	29	88
Ascomycetes	8	41	20	69
Basidiomycetes	1	6	4	11
Other fungi	1	2	5	8
Protists	6	18	27	51
Apicomplexans	1	9	7	17
Kinetoplasts	1	2	5	8
Other protists	4	7	14	25
Total:	644	581	631	1856

These projects have the goal of a high-quality full-genome sequence for a new species. Other high-throughput sequencing efforts involve: (1) Resequencing: sequencing a new individual of a species for which a full reference sequence is available. J.D. Watson's genome is an example. Related projects probe for specific variations, e.g. testing for mutations in BRCA1 and BRCA2 to assess risk of breast or ovarian cancer. (2) Comparative genomics: selection of interesting regions and sequencing homologous segments from many species, e.g. the ENCODE project (see p. 123). (3) Metagenomics, including Environmental Shotgun Sequencing, to measure variation in the biota at a point in space and time (see p. 128).

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

Given two aligned gene sequences, we can calculate  $K_s$  = the number of synonymous substitutions, and  $K_a$  = the number of non-synonymous substitutions. (The calculation involves more than simple counting because of the need to estimate and correct for possible multiple changes.) A high ratio of  $K_s/K_a$  identifies pairs of sequences apparently showing positive selection, possibly even functional changes.

The new field of comparative genomics treats questions that can only now be addressed, such as:

- ◆ What *genes* do different phyla share? What genes are unique to different phyla? Do the arrangements of these genes in the genome vary from phylum to phylum?
- ◆ What homologous **proteins** do different phyla share? What proteins are unique to different phyla? Does the integration of the activities of these proteins vary from phylum to phylum? Do the mechanisms of control of expression patterns of these proteins vary from phylum to phylum?
- ◆ What **biochemical functions** do different phyla share? What biochemical functions are unique to different phyla? Does the integration of these biochemical functions vary from phylum to phylum? If two phyla share a function, and the protein that carries out this function in one phylum has a homologue in the other, does the homologous protein carry out the same function?

The same questions could be asked about different species within each phylum.

M.A. Andrade, C. Ouzounis, C. Sander, J. Tamames and A. Valencia compared the protein repertoire of species from the three major domains of life: *Haemophilus influenzae* represented the bacteria, *Methanococcus jannaschii* the archaea, and *Saccharomyces cerevisiae* (yeast) the eukarya. Their classification of protein functions contained as major categories processes involving energy, information, and communication and regulation:

#### General functional classes

- ◆ Energy
  - Biosynthesis of cofactors, amino acids
  - Central and intermediary metabolism
  - Energy metabolism
  - Fatty acids and phospholipids
  - Nucleotide biosynthesis
  - Transport
- ◆ Information
  - Replication
  - Transcription
  - Translation
- ◆ Communication and regulation
  - Regulatory functions
  - Cell envelope/cell wall
  - Cellular processes

The number of genes in the three species, known at the time of that study, are:

Species	Number of genes
<i>Haemophilus influenzae</i>	1680
<i>Methanococcus jannaschii</i>	1735
<i>Saccharomyces cerevisiae</i>	6278

Are there, among these, shared proteins for shared functions? In the category of energy, proteins are shared across the three domains. In the category of communication, proteins are unique to each domain. In the categories of regulation and information, archaea share some proteins with bacteria and others with eukarya.

Analysis of shared functions among all domains of life has led people to ask whether it might be possible to define a *minimal organism*—that is, an organism with the smallest gene complement consistent with independent life based on the central DNA→RNA→protein dogma (i.e. excluding protein-free life forms based solely on RNA). The minimal organism must have the ability to reproduce, but not be required to compete in growth and reproductive rate with other organisms. One may reasonably assume a generous nutrient medium, relieving the organism of biosynthetic responsibility, and dispense with stress response functions including DNA repair.

The smallest known independent organism is *Mycoplasma genitalium*, with 468 predicted protein sequences. In 1996, A.R. Mushegian and E.V. Koonin compared the genomes of *M. genitalium* and *H. influenzae*. (At the time, these were the only completely sequenced bacterial genomes.) The last common ancestor of these widely diverged bacteria lived about 2 billion years ago. Of 1703 protein-coding genes of *H. influenzae*, 240 are homologues of proteins in *M. genitalium*. Mushegian and Koonin reasoned that all of these must be essential, but might not be sufficient for autonomous life—because some essential functions might be carried out by unrelated proteins in the two organisms. For instance, the common set of 240 proteins left gaps in essential pathways, which could be filled by adding 22 enzymes from *M. genitalium*. Finally, removing functional redundancy and parasite-specific genes gave a list of 256 genes as the proposed necessary *and* sufficient minimal set.

What is in the proposed minimal genome? Functional classes include:

- ◆ Translation, including protein synthesis
- ◆ DNA replication
- ◆ Recombination and repair—a second function of essential proteins involved in DNA replication.
- ◆ Transcription apparatus
- ◆ Chaperone-like proteins
- ◆ Intermediary metabolism—the glycolytic pathway
- ◆ No nucleotide, amino acid or fatty acid biosynthesis

- ◆ Protein export machinery
- ◆ Limited repertoire of metabolite transport proteins

It should be emphasized that the viability of an organism with these proteins has not been proven. Moreover, even if experiments proved that some minimal gene content—the proposed set or some other set—is necessary and sufficient, this does not answer the related question of identifying the gene complement of the common ancestor of *M. genitalium* and *H. influenzae*, or of the earliest cellular forms of life. Only 71% of the proposed set of 256 proteins have recognizable homologues among eukaryotic or archaeal proteins.

Nevertheless, identification of functions necessarily common to all forms of life allows us to investigate the extent to which different forms of life accomplish these functions in the same ways. Are similar reactions catalysed in different species by homologous proteins? Genome analysis has revealed families of proteins with homologues in archaea, bacteria and eukarya. The assumption is that these have evolved from an individual ancestral gene through a series of speciation and duplication events, although some may be the effects of horizontal transfer. The challenge is to map common functions and common proteins.

Several thousand protein families have been identified with homologues in archaea, bacteria and eukarya. Different species contain different amounts of these common families: in bacteria, the range is from *Aquifex aeolicus*, 83% of the proteins of which have archaeal and eukaryotic homologues, to *Borrelia burgdorferi*, in which only 52% of the proteins have archaeal and eukaryotic homologues. Archaeal genomes have somewhat higher percentages (62–71%) of proteins with bacterial and eukaryotic homologues. But only 35% of the proteins of yeast have bacterial and archaeal homologues.

Does the common set of proteins carry out the common set of functions? Among the proteins of the minimal set identified from *M. genitalium*, only ~30% have homologues in all known genomes. Other essential functions must be carried out by unrelated proteins, or possibly by unrecognized homologues. The protein families for which homologues carry out common functions in archaea, bacteria and eukarya are enriched in those involved in translation and biosynthesis:

Protein functional class	Number of families appearing in all known genomes
Translation, including ribosome structure	53
Transcription	4
Replication, recombination, repair	5
Metabolism	9
Cellular processes: (chaperones, secretion, cell division, cell wall biosynthesis)	9

The picture is emerging that evolution has explored the vast potential of proteins to different extents for different types of functions. It has been most conservative in the area of protein synthesis.

#### Web resources Databases of aligned gene families

**Pfam: Protein families database**

<http://www.sanger.ac.uk/Software/Pfam/>

**COG: Clusters of Orthologous Groups**

<http://www.ncbi.nlm.nih.gov/COG/>

**HOBACGEN: Homologous Bacterial Genes Database**

<http://pbil.univ-lyon1.fr/databases/hobacgen.html>

**HOVERGEN: Homologous Vertebrate Genes Database**

<http://pbil.univ-lyon1.fr/databases/hovergen.html>

**TAED: The Adaptive Evolution Database**

<http://www.sbc.su.se/liberles/TAED.html>

Many other sites contain data on individual families.



online  
resource  
centre

### Please pass the genes: horizontal gene transfer

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'. It was a clear case of lateral or horizontal gene transfer—a bacterium picking up a gene from the soil in which it was growing, which an organism of another species had deposited there. The classic experiments on bacterial transformation by Griffith and by O. Avery, C. MacLeod and M. McCarthy that identified DNA as the genetic material are another example. In general, horizontal gene transfer 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. Several mechanisms of horizontal gene transfer are known, including direct uptake, as in the pneumococcal transformation experiments, or via a viral carrier.

Analysis of genome sequences has shown that horizontal gene transfer is not a rare event, but has affected most genes in microorganisms. It requires a change in our thinking from ordinary 'clonal' or parental models of heredity. 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 *E. coli*, 755 ORFs (a total of 547.8 kb, ~18% of the genome) appear to have entered the genome by horizontal transfer after divergence from the *Salmonella* lineage 100 Mya.
- ◆ 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 50% identical to those of higher plants. Rubisco (ribulose-1,5-bisphosphate carboxylase/oxygenase), the enzyme that first fixes carbon dioxide at the 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 eukarya and prokaryotes are chimeras. Eukarya 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. Nor is gene transfer limited to ancient ancestors. The human genome revealed hundreds of bacterial proteins among our genes. Conversely, at least eight human genes appeared in the *M. tuberculosis* genome.

The observations hint at the model of a 'global organism', a genetic common market, or even a 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 wouldn't the global organism break down the lines of demarcation between species, just as global access to pop culture threatens to break down lines of demarcation between 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. Metagenomic sampling is illuminating these questions (see page 128).

It is interesting that although evidence for the importance of horizontal gene transfer is overwhelming, it was dismissed for a long time as rare and unimportant. The source of the intellectual discomfort is clear: parent-to-child transmission of genes is at the heart of the Darwinian model of biological evolution whereby selection (differential reproduction) of parental phenotypes alters gene frequencies in the next generation. For offspring to gain genes from elsewhere than their parents smacks of Lamarck and other discredited alternatives to the paradigm. The evolutionary tree as an organizing principle of biological relationship is a deeply ingrained concept: scientists display an environmentalist-like fervour in their commitment to trees, even when trees are not an appropriate model of a network of relationships (see Chapter 5). Perhaps it is well to recall that Darwin knew nothing of genes, and the mechanism that generated the variation on which selection could operate was a mystery to him. Maybe he would have accepted horizontal gene transfer more easily than his followers!

## Comparative genomics of eukarya

A comparison of the genomes of yeast, fly, worm and human revealed 1308 groups of proteins that appear in all four. These form a conserved core of proteins for basic functions, including metabolism, DNA replication and repair, and translation.

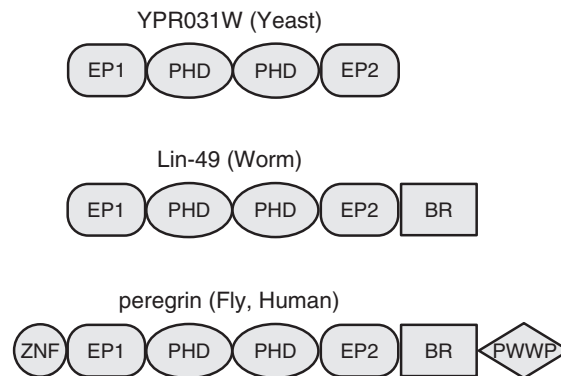
These proteins are made up of individual protein domains, including single-domain proteins, oligomeric proteins, and modular proteins containing many domains (the biggest, the muscle protein titin, contains 250–300 domains.) The proteins of the worm and fly are built from a structural repertoire containing about three times as many domains as the proteins of yeast. Human proteins are built from about twice as many as those of the worm and fly. Most of these domains also appear in bacteria and archaea, but some are specific to (probably, invented by) vertebrates (see Table). These include proteins that mediate activities unique to vertebrates, such as defence and immunity proteins, and proteins in the nervous system; only one of them is an enzyme, a ribonuclease.

### Distribution of probable homologues of predicted human proteins

Vertebrates only	22%
Vertebrates and other animals	24%
Animals and other eukarya	32%
Eukarya and prokaryotes	21%
No homologues in animals	1%
Prokaryotes only	1%

To create new proteins, inventing new domains is an unusual event. It is far more common to create different combinations of existing domains in increasingly complex ways. A common mechanism is by accretion of domains at the ends of modular proteins (see Fig. 2.4). This process can occur independently, and take different courses, in different phyla.

Gene duplication followed by divergence is a mechanism for creating protein families. For instance, there are 906 genes + pseudogenes for olfactory receptors in the human genome. These are estimated to bind ~10 000 odour molecules. Homologues have been demonstrated in yeast and other fungi (some comparisons *are* odorous), but it is the need of vertebrates for a highly developed sense of smell that multiplied and specialized the family to such a great extent. Eighty percent of the human olfactory receptor genes are in clusters. Compare the small size of the globin gene cluster (pages 95–6), which did not require such great variety.



**Fig. 2.4** Evolution by accretion of domains, of molecules related to perigrin, a human protein that probably functions in transcription regulation. The *C. elegans* homologue, lin-49, is essential for normal development of the worm. The yeast homologue is involved in histone acetylation. The proteins contain these domains: ZNF = C<sub>2</sub>H<sub>2</sub>-type zinc finger (not to be confused with acetylene; C and H stand for cysteine and histidine); EP1 and EP2 = Enhancer of polycomb 1 and 2, PHD = plant homeodomain, a repressor domain containing the C<sub>4</sub>H<sub>3</sub>C<sub>3</sub> type of zinc finger, BR = bromo domain; PWWP = domain containing sequence motif Pro-Trp-Trp-Pro.

### 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—has been the initial focus. The basic approaches have been comparative genomics and expression profiling, and involved both laboratory and computational analysis.

Regions corresponding to the selected human genome segments from 28 vertebrates have been sequenced (see Table and Fig. 2.5). These data illuminate each other. The ENCODE project will apply, improve, and develop as necessary, a variety of experimental and computational methods. Lessons learned from work with the selected subset will guide the scaling up of successful methods to analysis of entire genomes (see <http://www.genome.gov/10005107>).

#### Species targeted by the ENCODE project

Class:	Quality of sequencing		
	High	Medium	Unfinished
<i>Actinopterygii</i>	Zebrafish		
<i>Actinopterygii</i>	Fugu fish		
<i>Amphibia</i>	Frog		
<i>Aves</i>	Chicken		

Species targeted by the ENCODE project (*continued*)

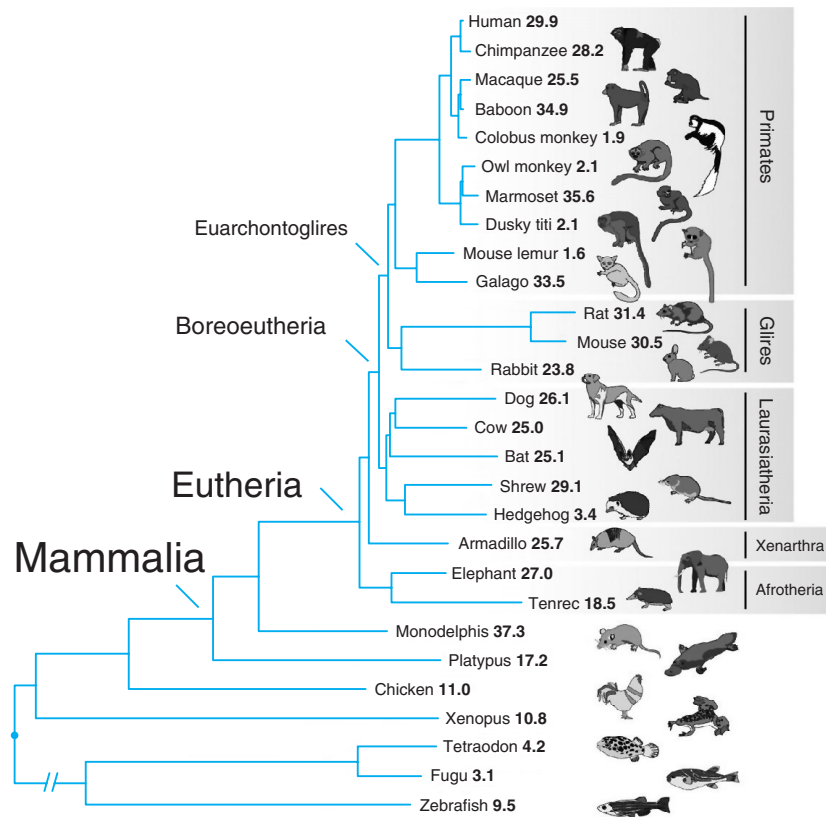
		Quality of sequencing		
		High	Medium	Unfinished
<b>Class Mammalia</b>				
<b>Order:</b>	<b>Suborder:</b>			
<i>Monotremata</i>				Platypus
<i>Marsupialia</i>			Opossum	
<i>Proboscidea</i>				African elephant
<i>Insectivora</i>				Tenrec
<i>Xenarthra</i>				Armadillo
<i>Insectivora</i>				Hedgehog
<i>Insectivora</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>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>	Chimpanzee		
Primates	<i>Hominidae</i>	Human		

High-quality sequence will be finished to state-of-the-art standards, including resolving difficult regions. Medium-quality sequence 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.

Coordinating with ENCODE, the HapMap project focuses on variations among humans in 10 of the ENCODE regions. Sequences from 48 individuals from different geographic origins, yielded 30 000 SNPs.

Analysis of function involves two steps: deciding whether a segment has functional significance, and, if so, identifying what it does. Approximately 5% of the human genome is conserved with respect to mouse and rat sequences. The original idea was to use conservation as a filter to identify functional regions. We shall see that this was only a partial success, which itself is interesting. Only about a 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



**Fig. 2.5** Phylogenetic tree of species treated in the ENCODE project. Numbers are the length of sequence examined, in Mb. (Reproduced and modified with permission from Margulies E.H. *et al.* (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Research*, **17**, 760–774. Copyright 2007. Cold Spring Harbor Laboratory Press).

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  $\alpha$ - and  $\beta$ -globin loci, and the region containing the gene for the cystic fibrosis transmembrane conductance regulator (CFTR), for which sequence information from different species is known.

Chromosome	Approximate sizes of ENCODE regions in 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)

Chromosome	Approximate sizes of ENCODE regions in Mb (gene of interest)
6	0.5, 0.5, 0.5, 0.5
7	0.5, 1.0, 1.1, 1.2, 1.9 (CFTR)
8	0.5
9	0.5
10	0.5
11	0.5, 0.5, 0.6, 0.5 (Apo cluster), 1.0 ( $\beta$ -globin)
12	0.5
13	0.5, 0.5
14	0.5, 0.5
15	0.5
16	0.5, 0.5, 0.5 ( $\alpha$ -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

The results of the ENCODE measurements confirmed many expectations, but turned up a number of surprises:

- ◆ The correlation between conservation and function of segments was imperfect: 60% of the sequences that appear to be under evolutionary constraint map to functional elements. Will functions ultimately be discovered for the remaining 40%? Conversely, many regions of known function appear *not* to be conserved, among the different vertebrate genomes and among different human individuals.
- ◆ Almost all the human genome is transcribed. Transcripts detected form an overlapping cover of nearly the entire genome. Of course most of this involves regions that do not code for proteins, or for known structural RNAs. Some of these transcripts function in regulation of gene expression, for instance small interfering RNAs (siRNAs). But most of the transcripts do not have a known function.
- ◆ The relationship between chromatin structure and regulation of transcription and replication has been exposed in greater detail. In local regions of the sequence, there is a good correlation between chromatin accessibility and histone modification patterns and the locations and activity of transcriptional start sites. On a larger

scale—on the order of 1 Mb—there is a relationship between histone modifications and the temporal pattern of DNA replication.

## Metagenomics: the collection of genomes in a coherent environmental sample

Classically, microbiologists studied prokaryotes by growing them in culture, producing pure strains for detailed study. Powerful as the methods were, and useful as they were for clinical applications and research, they were also blinders that hindered full appreciation of the variety and interactions of species in natural environments. Microbiologists have described and named only a very small fraction of bacterial species. Viruses present even greater variety, by far.

DNA sequencing has made it possible to:

- ◆ Clarify evolutionary relationships through comparison of full genomes.
- ◆ Use high-throughput sequencing methods to study a cross-section of the life in a natural sample.
- ◆ Study the majority of strains that are difficult to grow in culture.
- ◆ Understand the relationships and interactions among different species that share an ecosystem. The ecosystem could be a region of the ocean, a sample of acid mine drainage, or the gut of a person or animal.
- ◆ Compare the populations that occupy neighbouring and distant locations. For example, studying the variation of microbial populations with depth at some location in the ocean, or surveying how populations vary with latitude and longitude across a region of ocean or land.
- ◆ Reveal unsuspected levels of complexity—of both individual species and their interactions—for example, the very great variety of viruses inhabiting the ocean.

From natural samples containing complex mixtures, it is possible to amplify and determine sequences directly, without culturing individual organisms. This is called *environmental shotgun sequencing*. With modern equipment, it is possible to generate very large amounts of sequence data. It is correspondingly difficult to assemble the data into coherent genomes, a problem aggravated by horizontal gene transfer. In fact, a lot of horizontal gene transfer threatens to break down the idea of discrete species. This would be a real revolution in biology.<sup>2</sup>

Alternatively, one can focus on a few selected molecules, and survey organisms for these sequences. This would offer some understanding of the diversity and distribution of organisms in the sample. The molecule of choice continues to be 15S rRNA. This is partly because of its traditional role as a molecule that varies at the appropriate rate

2 See: D. M. Ward *et al.* (2008). Genomics, environmental genomics and the issue of microbial species. *Heredity*, 99, 1–13.

.....  
A millilitre of ocean water may contain 100–200 species.

A gram of soil may contain 4000. How many strains? Don't ask!

.....  
In a set of shotgun sequences of DNA from a soil sample, <1% of the reads showed overlaps.  
.....

to distinguish phylogenetic branching patterns. In addition, rRNA is not very prone to horizontal gene transfer. It thereby preserves the distinctions between taxa. In fact, it may well disguise the mixing that has taken place by horizontal transfer of other genes.

Characterizing an environmental sample by its rRNAs has other serious limitations:

- ◆ **Ribosomal RNA does not reveal any details of the metabolism or other adaptations of the species or strains** K-12 MG1655 and O157:H7 are strains of *E. coli*.

Strain	Genome size size (bp)	Number of genes
K-12 MG1655	4 639 221	4406
O157:H7	5 594 477	5416

.....  
*E. coli* O157:H7 is a virulent strain, responsible for outbreaks of disease.  
 .....

The two strains conserve a common 4.1 Mb sequence, with 98.3% base identity, with 2027 gaps. This is the core of the species genome. The remaining 1.4 Mb of O157:H7 is rich in foreign DNAs acquired by horizontal transfer, including 24 prophages and prophage-like elements.

Despite the high similarity of much of these genomes, there is substantial divergence in the proteome. Strain O157:H7 encodes 1632 proteins and 20 tRNAs not present in K-12 MG1655. This is not exclusively from genes in the extra 1.4 Mb: K-12 MG1655 encodes 528 genes not present in O157:H7. These proteomes differ much more than intraspecific variations observed in higher animals.

- ◆ **Viruses do not contain ribosomes and therefore are invisible to probes for 16S rRNA** Viruses appear to be the ‘dark matter’ of nature: they exist in unsuspected numbers and variety. Many viral proteins are very different from the sets of molecules familiar in cellular organisms. Anyone with the ambition of deriving a catalogue of protein folding patterns, on the basis of the results of current structural genomics projects, should live in dire fear of what the combined viral proteome will reveal.

Perhaps the most ambitious harvesting of metagenomics data came from the *Sorcerer II* Global Ocean Sampling Expedition.<sup>3</sup> During a round-the-world trip between 8 August 2003 and 22 May 2004, samples were collected at ~320 km intervals along a >8000 km route that started in Halifax, Nova Scotia, Canada, along the East Coast of the USA, the Gulf of Mexico, the Galapagos Islands, across the Pacific Ocean to Australia, through the Indian Ocean, South Africa and back across the Atlantic to the USA.

Selected fractions with cells of size 0.1–0.8 μm were filtered to focus on bacteria. Of 7.7 million sequencing reads from these samples, amounting in total to 6.3 × 10<sup>9</sup> bp, there remained, upon counting for overlaps, almost 6 Gb of unique sequence.

.....  
 The expedition was inspired by the *HMS Challenger* expedition of 1872–1876, a survey of ocean geology, climate and biology.  
 .....

<sup>3</sup> S. Yooseph *et al.* (2007). The *Sorcerer II* Global Ocean Sampling Expedition: expanding the universe of protein families. *PLoS Biology*, **13**, e16; Rusch, D.B. *et al.* (2007). The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, **13**, e77.

Over half the reads were unique; that is, they had  $\leq 98\%$  sequence similarity to previously reported sequences. Contigs were assemblable into  $>3$  million whole-genome scaffolds.

The 16S RNA data from the shotgun sequencing revealed 811 distinct sequence types (below 97% identity). Over half represented putative novel species. Note the absence of Archaea from the most highly represented taxa.

Phylum or class	Fraction
$\alpha$ -Proteobacteria	0.32
Unclassified proteobacteria	0.155
$\gamma$ -Proteobacteria	0.132
Bacteroidetes	0.13
Cyanobacteria	0.079
Firmicutes	0.075
Actinobacteria	0.046
Marine group A	0.022
$\beta$ -Proteobacteria	0.017
OP11	0.008
Unclassified bacteria	0.008
$\delta$ -Proteobacteria	0.005
Planctomycetes	0.002
$\epsilon$ -Proteobacteria	0.001

Source: Rusch, D.B. *et al.* (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, **13**, e77.

Translations of gene sequences identified over 6 million bacterial and viral proteins. Of these, 1700 have no detectable similarity to previously known proteins. If all are indeed novel—remember that sequence-based tools do not always successfully identify structural similarity in distantly related proteins—the results will almost double the number of known protein families.

The data flow from metagenomics, already large, will only increase by leaps and bounds, as sequencing methods become even more powerful. The challenges to bioinformatics will be not only in the quantity of the data but in their novelty and variety. In our understanding of the organization of life, the paradigm will shift from the hierarchical model to which since Linnaeus we have been accustomed, to a higher

dimension of complexity. Coming to terms with this will require conceptual as well as computational breakthroughs.

### Recommended reading

- C. elegans* Sequencing Consortium (1999). How the worm was won. The *C. elegans* genome sequencing project. *Trends in Genetics*, **15**, 51–58. [Description of the project in which high-throughput DNA sequencing was originally developed, and its results, the first metazoan genome to be sequenced.]
- Ashburner, M. (2006). *Won for All: How the Drosophila Genome Was Sequenced*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. [A personal description—a ‘blow-by-blow’ account—of the fly genome project.]
- Ureta-Vidal, A., Ettwiller, L. and Birney, E. (2003). Comparative genomics: genomewide analysis in metazoan eukarya. *Nature Reviews Genetics*, **4**, 251–262. [Introduction to the repertoire of solved genomes and how genomes from different species illuminate one another.]
- Zhang, R. and Zhang, C.-T. (2006). The impact of comparative genomics on infectious disease research. *Microbes and Infection*, **8**, 1613–1622. [How comparative genomics of prokaryotes is elucidating infectious disease, including identification of virulence determinants, drug targets, candidate vaccines and diagnostic markers.]
- Doolittle, W.F. (1999). Lateral genomics. *Trends in Cell Biology*, **9**, M5–M8. [How the discovery of horizontal gene transfer has upset traditional views of evolution.]
- Koonin, E.V. (2000). How many genes can make a cell?: the minimal-gene-set concept. *Annual Review of Genomics and Human Genetics*, **1**, 99–116. [A summary of work on comparative genomics.]
- Kwok, P.-Y. and Gu, Z. (1999). SNP libraries: why and how are we building them? *Molecular Medicine Today* **5**, 538–543. [Progress and rationale for databases of single-nucleotide polymorphisms.]
- Southan, C. (2004). Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics*, **4**, 1712–1726. [How many genes are there in the human genome? Estimates have been changing.]
- Bentley, D.R. (2004). Genomes for medicine. *Nature*, **429**, 440–445. [Summary and discussion of the results of genomic sequencing and their applications.]
- Dubcovsky, J. and Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**, 1862–1866.
- Publications of the drafts of the human genome sequences appeared in special issues of *Nature*, **15** February 2001, containing the results of the publicly supported Human Genome Project, and *Science*, **16** February 2001, containing the results produced by Celera Genomics. These are landmark issues.
- The May 2001 issue of *Genome Research*, Volume 11, Number 5, is devoted to the human genome.
- The completion of the human genome in 2003 was announced in various press releases, and described in issues of *Nature* and *Science* magazines:
- Collins, F.S., Green, E.D., Guttmacher, A.E. and Guyer, M.S. (2003). A vision for the future of genomics research. *Nature*, **422**, 835–847.
- Building on the DNA Revolution. (2003). Special section, *Science*, **300**, 11 April 2003, pp. 277ff.
- Xu, J. (2006). Microbial ecology in the age of genomics and metagenomics: concepts, tools and recent advances. *Molecular Ecology*, **15**, 1713–1731.
- Eisen, J.A. (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biology*, **5**, e82 [Two papers discussing the current status and conceptual problems of coming to terms with metagenomics.]

## Exercises, Problems and Weblems

### Exercises

**Exercise 2.1** The overall base composition of the *E. coli* genome is  $A = T = 49.2\%$ ,  $G = C = 50.8\%$ . In a random sequence of 4 639 221 nucleotides with these proportions, what is the expected number of occurrences of the sequence CTAG?

**Exercise 2.2** The *E. coli* genome contains a number of pairs of enzymes that catalyse the same reaction. How would this affect the use of knockout experiments (deletion or inactivation of individual genes) to try to discern function?

**Exercise 2.3** Which of the categories used to classify the functions of yeast proteins (see page 98) would be appropriate for classifying proteins from a prokaryotic genome?

**Exercise 2.4** Which occurred first, a man landing on the moon or the discovery of deep-sea hydrothermal vents? Guess first, then look it up.

**Exercise 2.5** Gardner syndrome is a condition in which large numbers of polyps develop in the lower gastrointestinal tract, leading inevitably to cancer if untreated. In every observed case, one of the parents is also a sufferer. What is the mode of the inheritance of this condition?

**Exercise 2.6** The gene for retinoblastoma is transmitted along with a gene for esterase D to which it is closely linked. However, either of the two alleles for esterase D can be transmitted with either allele for retinoblastoma. How do you know that retinoblastoma is not the direct effect of the esterase D genotype?

**Exercise 2.7** If all somatic cells of an organism have the same DNA sequence, why is it necessary to have cDNA libraries from different tissues?

**Exercise 2.8** Suppose you are trying to identify a gene causing a human disease. You find a genetic marker 0.75 cM from the disease gene. To within approximately how many base pairs have you localized the gene you are looking for? Approximately how many genes is this region likely to contain?

**Exercise 2.9** Leber hereditary optic neuropathy (LHON) is an inherited condition that can cause loss of central vision, resulting from mutations in mitochondrial DNA. You are asked to counsel a woman who has normal mitochondrial DNA and a man with LHON, who are contemplating marriage. What advice would you give them about the risk to their offspring of developing LHON?

**Exercise 2.10** Glucose-6-phosphate dehydrogenase deficiency is a single-gene recessive X-linked genetic defect affecting hundreds of millions of people. Clinical consequences include haemolytic anaemia and persistent neonatal jaundice. The gene has not been eliminated from the population because it confers resistance to malaria. In this case, general knowledge of metabolic pathways identified the protein causing the defect. Given the amino acid sequence of the protein, how would you determine the chromosomal location(s) of the corresponding gene?

**Exercise 2.11** Before DNA was recognized as the genetic material, the nature of a gene—in detailed biochemical terms—was obscure. In the 1940s, G. Beadle and E. Tatum observed that single mutations could knock out individual steps in biochemical pathways. On this basis they proposed the *one gene—one enzyme* hypothesis. On a photocopy of Fig. 2.1, draw lines linking genes in the figure to numbered steps in the sequence of reactions in the pathway. To what extent do the genes of the *trp* operon satisfy the one gene–one enzyme hypothesis and to what extent do they present exceptions?

**Exercise 2.12** The figure shows human chromosome 5 (left) and the matching chromosome from a chimpanzee. On a photocopy of this figure indicate which regions show an inversion of the banding pattern.



**Exercise 2.13** Describe in general terms how the FISH picture in Plate II would appear if the affected region of chromosome 20 were not deleted but translocated to another chromosome.

**Exercise 2.14** 755 open reading frames entered the *E. coli* genome by horizontal transfer in the 14.4 million years since divergence from *Salmonella*. What is the average rate of horizontal transfer in kb/year? To how many typical proteins (~300 amino acids) would this correspond? What percentage of known genes entered the *E. coli* genome via horizontal transfer?

**Exercise 2.15** To what extent is a living genome like a database? Which of the following properties are shared by living genomes and computer databases? Which are properties of living genomes but not databases? Which are properties of databases but not living genomes?

- (a) Serve as repositories of information.
- (b) Are self-interpreting.
- (c) Different copies are not identical.
- (d) Scientists can detect errors.
- (e) Scientists can correct errors.
- (f) There is planned and organized responsibility for assembling and disseminating the information.

---

### Problems

**Problem 2.1** Summarize the experimental evidence that shows that the genetic linkage map on any single chromosome is linearly ordered.

**Problem 2.2** For *M. genitalium* and *H. influenzae*, what are the values of (a) gene density in genes/kb, (b) average gene size in bp, (c) number of genes. Which factor contributes most to the reduction of genome size in *M. genitalium* relative to *H. influenzae*?

**Problem 2.3** It is estimated that the human immune system can produce  $10^{15}$  antibodies. Would it be feasible for such a large number of proteins each to be encoded entirely by a separate gene, the diversity arising from gene duplication and divergence? A typical gene for an IgG molecule is  $\sim 2000$  bp long.

---

### Weblems

**Weblem 2.1** On photocopies of Figs. 1.2, 1.3 and 1.4, indicate the positions of species for which full-genome sequences are known. (<http://www.ebi.ac.uk/genomes/>)

**Weblem 2.2** What are the differences between the standard genetic code and the vertebrate mitochondrial genetic code?

**Weblem 2.3** What is the chromosomal location of the human myoglobin gene?

**Weblem 2.4** Find examples of additional completely sequenced eukaryotic genomes not listed in the table on pages 84–5. Find at least two in each of the categories: Mammals, Other chordates, Higher plants, Other eukarya.

**Weblem 2.5** The table on page 117 contains the statistics of current status of genome projects, as of 1 January 2008. What are the current numbers?

**Weblem 2.6** What is the number of occurrences of the tetrapeptide CTAG in the *E. coli* genome? Is it over-represented or under-represented relative to the expectation for a random sequence of the same length and base composition as the *E. coli* genome? (See Exercise 2.1.)

**Weblem 2.7** Plot a histogram of the cumulative number of completed genome sequences in each year since 1995.

**Weblem 2.8** (a) How many predicted ORFs are there on *Saccharomyces cerevisiae* chromosome X? (b) How many tRNA genes?

**Weblem 2.9** Which amino acid is entirely lacking in the proteins of *M. genitalium*? How does the genetic code of *M. genitalium* differ from the standard one?

**Weblem 2.10** In the human, 1 cM  $\sim 10^6$  bp. In yeast, approximately how many base pairs correspond to 1 cM?

**Weblem 2.11** Sperm cells are active swimmers, and contain mitochondria. At fertilization the entire contents of a sperm cell enters the egg. How is it therefore that mitochondrial DNA is inherited from the mother only?

**Weblem 2.12** The box on pages 95–6 show the duplications and divergences leading to the current human  $\alpha$ - and  $\beta$ - globin gene clusters. (a) In which species, closely related to ancestors of humans, did these divergences take place? (b) In which species related to ancestors of humans did the developmental pattern of expression pattern ( $\zeta_2\epsilon_2$  = embryonic;  $\alpha_2\gamma_2$  = foetal;  $\alpha_2\gamma_2$  = adult) emerge?

**Weblem 2.13** Are language groups more closely correlated with variations in human mitochondrial DNA or Y chromosome sequences? Suggest an explanation for the observed result.

**Weblem 2.14** The mutation causing sickle-cell anaemia is a single base change A $\rightarrow$ T, causing the change Glu $\rightarrow$ Val at position 6 of the  $\beta$  chain of haemoglobin. The base change occurs in the sequence 5'-GTGAG-3' (normal)  $\rightarrow$  GTGTG (mutant). What restriction enzyme is used to distinguish between these sequences, to detect carriers? What is the specificity of this enzyme?

**Weblem 2.15** What mutation is the most common cause of phenylketonuria (PKU)?

**Weblem 2.16** Find three examples of mutations in the CFTR gene (associated with cystic fibrosis) that produce reduced but not entirely absent chloride channel function. What are the clinical symptoms of these mutations?

**Weblem 2.17** Find an example of a genetic disease that is: (a) autosomal dominant, (b) autosomal recessive (other than cystic fibrosis), (c) X-linked dominant, (d) X-linked recessive, (e) Y-linked, (f) the result of abnormal mitochondrial DNA (other than Leber's hereditary optic neuropathy).

**Weblem 2.18** (a) Identify a state of the USA in which newborn infants are routinely tested for homocystinuria. (b) Identify a state of the USA in which newborn infants are *not* routinely tested for homocystinuria. (c) Identify a state of the USA in which newborn infants are routinely tested for biotinidase. (d) Identify a state of the USA in which newborn infants are *not* routinely tested for biotinidase. (e) What are the clinical consequences of failure to detect homocystinuria, or biotinidase deficiency?

**Weblem 2.19** (a) What is the normal function of the protein that is defective in Menke disease? (b) Is there a homologue of this gene in the *A. thaliana* genome? (c) If so, what is the function of this gene in *A. thaliana*?

**Weblem 2.20** *Duchenne muscular dystrophy (DMD)* is an X-linked inherited disease causing progressive muscle weakness. DMD sufferers usually lose the ability to walk by the age of 12, and life expectancy is no more than about 20–25 years. *Becker muscular dystrophy (BMD)* is a less severe condition involving the same gene. Both conditions are usually caused by deletions in a single gene, dystrophin. In DMD there is complete absence of functional protein; in BMD there is a truncated protein retaining some function. Some of the deletions in cases of BMD are longer than others that produce DMD. What distinguishes the two classes of deletions causing these two conditions?

**Weblem 2.21** What chromosome of the cow contains a region homologous to human chromosome region 8q21.12?