CHAPTER 18

Comparative genomics

Introduction

Comparative genomics is the study of the differences and similarities in genome structure and organization in different organisms. For example, how are the differences between humans and other organisms reflected in our genomes? How similar are the number and types of proteins in humans, fruit flies, worms, plants, yeasts, and bacteria? Essentially, comparative genomics is no more than the application of the bioinformatics methods described in Chapter 9 to the analysis of whole-genome sequences with the objective of identifying biological principles, i.e. biology *in silico*. In a sense this statement greatly underplays the real value of comparative genomics for, as the reader will soon see, it is an extremely powerful technique and provides biological insights that could not be achieved in any other way.

There are two drivers for comparative genetics. One is a desire to have a much more detailed understanding of the process of evolution at the gross level (the origin of the major classes of organism) and at a local level (what makes related species unique). The second driver is the need to translate DNA sequence data into proteins of known function. The rationale here is that DNA sequences encoding important cellular functions are more likely to be conserved between species than sequences encoding dispensable functions or non-coding sequences. Until recently it was thought that the ideal species for comparison are those whose form, physiology, and behavior are as similar as possible but whose genomes have evolved sufficiently that non-functional sequences have had time to diverge. More recently, Bofelli *et al.* (2004) have shown that by comparing genomes that are very distantly related, e.g. mammals and fish, it is possible to identify conserved sequences that, presumably, have a significant function.

The formation of orthologs and paralogs are key steps in gene evolution

In order to compare genome organization in different organisms it is necessary to distinguish between *orthologs* and *paralogs*. Orthologs are homologous genes in different organisms that encode proteins with the same function and which have evolved by direct vertical descent. Paralogs are homologous genes within an organism encoding proteins with related but non-identical functions. Implicit in these definitions is that orthologs evolve simply by the gradual accumulation of mutations, whereas paralogs arise by gene duplication followed by mutation accumulation. Good examples of paralogs are the protein superfamilies described in Chapter 16 (see Fig. 16.6 and Table 16.1).

There are many biochemical activities that are common to most or all living organisms, e.g. the citric acid cycle, the generation of ATP, the synthesis of nucleotides, DNA replication, etc. It might be thought that in each case the key proteins would be orthologs. Indeed, "universal protein families" shared by all archae, eubacteria, and eukaryotes have been described (Kyrpides *et al.* 1999). However, there is increasing evidence that functional equivalence of proteins requires neither sequence similarity nor even common three-dimensional folds (Galperin *et al.* 1998, Huynen *et al.* 1999). The existence of two or more distinct sets of orthologs that are responsible for the same function in different organisms is called non-orthologous gene displacement. Now that close to 200 different genomes have been sequenced it is clear that gene displacement occurs within most essential genes. That is, there are at least two biochemical solutions to each cellular requirement. Only about 60 genes have been identified where gene displacement has not been observed (as yet) and most of these encode components of the transcription and translation systems (Koonin 2003).

Protein evolution occurs by exon shuffling

Analysis of protein sequences and three-dimensional structures has revealed that many proteins are composed of discrete domains. These so-called mosaic proteins are particularly abundant in the metazoa. The majority of mosaic proteins are extracellular or constitute the extracellular parts of membrane-bound proteins and thus they may have played an important part in the evolution of multicellularity. The individual domains of a mosaic protein are often involved in specific functions which contribute to its overall activity. These domains are evolutionarily mobile which means that they have spread during evolution and now occur in otherwise unrelated proteins (Doolittle 1995). Mobile domains are characterized by their ability to fold independently. This is an essential characteristic because it prevents misfolding when they are inserted into a new protein environment. To date, over 60 mobile domains have been identified.

A survey of the genes that encode mosaic proteins reveals a strong correlation between domain organization and intron–exon structure (Kolkman & Stemmer 2001); i.e. each domain tends to be encoded by one or a combination of exons and new combinations of exons are created by recombination within the intervening sequences. This process yields rearranged genes with altered function and is known as *exon shuffling*. Because the average intron is much longer than the average exon and the recombination frequency is proportional to DNA length, the vast majority of crossovers occur in non-coding sequences. The large number of transposable elements and repetitive sequences in introns will facilitate exon shuffling by promoting mismatching and recombination of non-homologous genes. An example of exon shuffling is described in Box 18.1.

Box 18.1 Hemostatic proteins as an example of exon shuffling

The process of blood coagulation and fibrinolysis involves a complex cascade of enzymatic reactions in which inactive zymogens are converted into active enzymes. These zymogens belong to the family of serine proteases and their activation is accompanied by proteolysis of a limited number of peptide bonds. Comparison of the amino acid sequences of the hemostatic proteases with those of archetypal serine proteases such as trypsin shows that the former have large N-terminal extensions (Fig. B18.1). These extensions consist of a number of discrete domains with functions such as substrate recognition, binding of co-factors, etc. and the different domains show a strong correlation with the exon structure of the encoding genes.

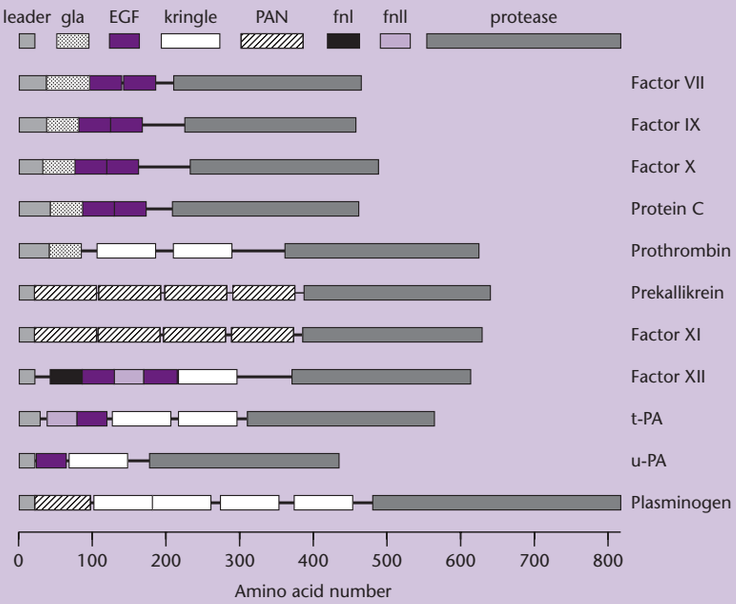

Fig. B18.1 Domain structures of the regulatory proteases of blood coagulation and fibrinolysis. The different domains: gray, serine protease domain; dark purple, EGF-like domain; dotted, Gla domain; cross-hatch, PAN domain; light purple, fibronectin type II domain (fn2); black, fibronectin type I domain (fn1). (Adapted from Kolkman & Stemmer 2001.)

Although mosaic proteins are most common in the metazoa, they are found in unicellular organisms. Because a large number of microbial genomes have been sequenced, including representatives from the three primary kingdoms (Archaea, Eubacteria, and Eukarya), it is possible to determine the evolutionary mobility of domains. With this in mind, Wolf *et al*. (2000a) searched the genomes of 15 bacteria, four archaea, and one eukaryote for genes encoding proteins consisting of domains from the different kingdoms. They found 37 examples of proteins consisting of a "native" domain and a horizontally acquired "alien" domain. In several instances the genome contained the gene for the mosaic protein as well as a sequence encoding a stand-alone version of the alien domain, but more usually the stand-alone counterpart was missing.

Comparative genomics of bacteria

By mid-2004 the website of the National Center for Biotechnology Information listed 173 bacteria (19 Archaea and 154 Eubacteria) whose genomes had been sequenced (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html). Simple analysis of the sequence data reveals two features of note. First, the genome sizes vary from 0.49 Mb (*Nanoarchaeum equitans*) to 9.1 Mb (*Bradyrhizobium japonicum* and two species of *Streptomyces*), i.e. a more than 18-fold difference. Secondly, the gene density is remarkably similar across all species and is about 1 gene per kilobase of DNA. This means that large prokaryotic genomes contain many more genes than smaller ones. By contrast, the human genome contains only twice as many genes as *Drosophila*. So how can we account for the size diversity of prokaryotes?

When the different genomes are arranged in size order (Fig. 18.1) some interesting features emerge. First, the archaebacteria exhibit a very much smaller range of genome sizes. This could be an artifact of the small number of genomes examined but more probably reflects the fact that most of them occupy a specialized environment and have little need for metabolic diversity. The exception is *Methanosarcina acetivorans*. This bacterium is known to thrive in a broad range of environments and at 5.8 Mb has the largest archaeal genome (Galagan *et al.* 2002). Second, the smallest eubacterial genomes are found in those organisms that normally are found associated with animals or humans, e.g. mycoplasmas, rickettsias, chlamydiae, etc. Those organisms that can occupy a greater number of niches have a larger genome size. Not surprisingly, there is a good correlation between genome size and metabolic and functional diversity as demonstrated by the size of the genomes of *Bacillus* and *Streptomyces* (formation of spores, antibiotic synthesis), rhizobia (symbiotic nitrogen fixation), and *Pseudomonas* (degradation of a wide range of aromatic compounds).

The minimal gene set consistent with independent existence can be determined using comparative genomics

The genome of *N. equitans* is the smallest sequenced to date (Waters *et al.* 2003) but this organism is an obligate symbiont. This begs the question, what is the minimal genome that is consistent with a free-living cellular organism? In reality, this is a nonsensical question unless one specifies a defined set of environmental conditions. Conceivably, the absolute minimal set of genes will correspond to the most favorable conditions possible in which all essential nutrients are provided and there are no environmental stress factors. If one ignores functionally important RNA molecules and non-coding sequences, the problem is one of defining the minimal protein set.

The first attempt at identifying the minimal protein set was made by compiling a list of orthologous proteins in *Hemophilus influenzae* and *M. genitalium* (Mushegian & Koonin 1996). The expectation was that this list would predominantly contain proteins integral for cell survival as both bacteria are essentially parasites and thus should have shed auxiliary genes. Altogether 244 orthologs were identified but this list is unlikely to be complete because of the occurrence of non-orthologous gene displacements. Some of these gene displacements can be inferred because both organisms appear to have key metabolic pathways that are incomplete. In this way, Mushegian (1999) extended the minimal protein set to 256 genes.

The problem with the above approach is that if one is too strict in defining the degree of similarity between two proteins required to constitute orthologs then the minimal protein set is greatly underestimated. A variation of the above method is to identify orthologous groups, i.e. clusters of genes that include orthologs and, additionally, those paralogs where there has been selective gene loss following gene duplication. When this approach was taken with four eubacteria, one archaebacterium, and one yeast, 816 clusters of orthologous groups (COGs) were

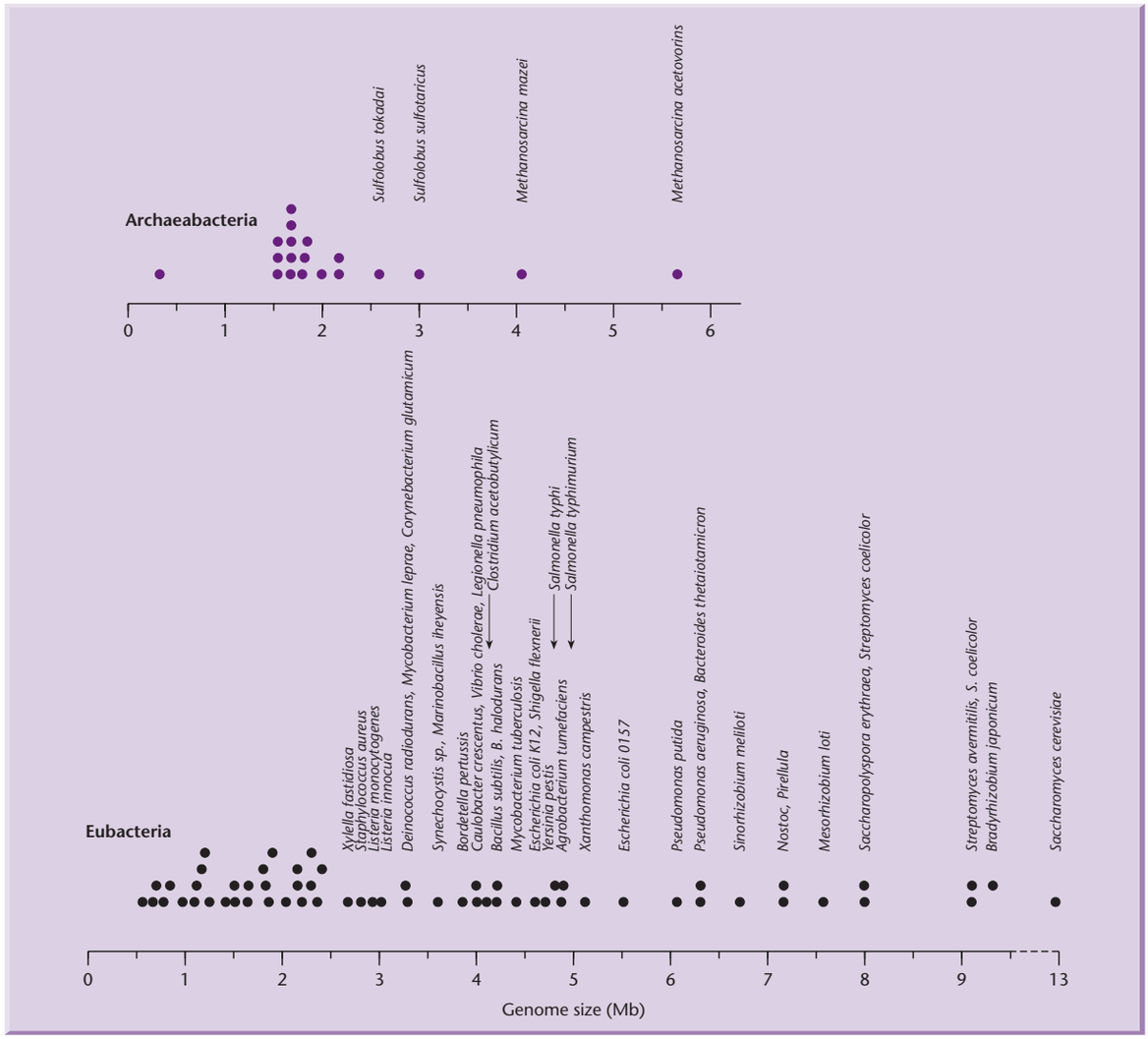

Fig. 18.1 Genome sizes of archaebacteria, some eubacteria, and one eukaryote whose genomes have been completely sequenced.

identified. Of these, 327 contained representatives of all three kingdoms (Mushegian 1999). Based on this set of 327 proteins it was possible to reconstruct all the key biosynthetic pathways. When the analysis was repeated with sequence data from an additional three archaebacteria and 12 eubacteria, the minimal protein set was slightly reduced to 322 COGs.

Larger microbial genomes have more paralogs than smaller genomes

Comparison of the *P. aeruginosa* (6.3 Mb) and *E. coli* (4.5 Mb) genomes indicates that the large genome of *P. aeruginosa* is the result of greater genetic complexity rather than differences in genome organization. Distributions of open-reading frame (ORF) sizes and inter-ORF spacings are nearly identical in the two genomes. If the larger genome of *P. aeruginosa* arose by recent gene duplication one would expect it to have a similar number of paralogous groups compared to the other large bacterial genomes and a larger number of ORFs in each group. In fact, the number of ORFs in the paralogous groups in *Pseudomonas* is similar to the other genomes. Thus selection for environmental versatility (Box 18.2) has favored genetic capability through the development of numerous small paralogous gene families

Box 18.2 Correlation of genome sequence data with the biology of bacteria

Pseudomonas aeruginosa

Pseudomonas aeruginosa is a bacterium that is extremely versatile both ecologically and metabolically. It grows in a wide variety of habitats including soil, water, plant surfaces, biofilms, and both in and on animals including humans. A major problem with *P. aeruginosa* is its resistance to many disinfectants and antibiotics. Pseudomonads are characterized by a limited ability to grow on carbohydrates but a remarkable ability to metabolize many other compounds including an astonishing variety of aromatics.

Analysis of the genome of *P. aeruginosa* (Stover *et al.* 2000) reveals a general lack of sugar transporters and an incomplete glycolytic pathway, both of which explain the poor ability to grow on sugars. By contrast, it has large numbers of transporters for a wide range of metabolites and a substantial number of genes for metabolic pathways not found in many other bacteria such as *E. coli*. As might be expected for an organism with great metabolic versatility, a high proportion of the genes (>8%) are involved in gene regulation. The organism also has the most complex chemosensory system of all the complete bacterial genomes with four loci that encode probable chemotaxis signal-transduction pathways. Finally, sequencing revealed the presence of a large number of undescribed drug efflux systems which probably account for the inherent resistance of the organism to many antibacterial substances.

Caulobacter crescentus

Caulobacter crescentus is a bacterium that is found in oligotrophic (very low nutrient) environments and is not capable of growing in rich media. Not surprisingly, genome sequencing (Nierman *et al.* 2001) has shown that the bacterium possesses a large number of genes for responding to environmental substrates. For example, 2.5% of the genome is devoted to motility, there are two chemotaxis systems and over 16 chemoreceptors. It also has 65 members of the family of outer membrane proteins that catalyze energy-dependent transport across the membrane. By contrast, the metabolically versatile *P. aeruginosa* has 32 and other bacteria fewer than 10.

The bacterium also has an obligatory life cycle involving asymmetric cell division and differentiation (Fig. B18.2). Thus it comes as no surprise that genome sequencing reveals a very high number of two-component signal-transduction proteins, e.g. 34 histidine protein kinase (HPK) genes, 44 response regulator (RR) genes, and 27 hybrid (HPK/RR) genes. In addition, the frequency of the GAnTC target site for DNA methylation was much less than would be expected if it occurred at random.

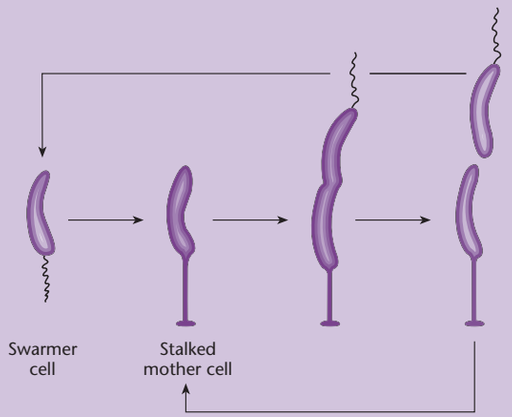

Fig. B18.2 The life cycle of *Caulobacter crescentus*.

Deinococcus radiodurans

This bacterium is remarkable for its ability to survive extremely high doses of ionizing radiation. For example, it can grow in the presence of chronic radiation (6 kilorads/hour) and withstand acute exposures to 1500 kilorads. The organism also is resistant to dessication, oxidizing agents, and ultraviolet radiation. These properties could be the result of one or more of prevention, tolerance, and repair. Genome sequencing (White *et al.* 1999, Makarova *et al.* 2001) has shown that systems for the prevention and tolerance of DNA damage are present but that the key mechanism of resistance is an extremely efficient DNA repair system. Although all of the DNA repair genes identified in *D. radiodurans* have functional homologs in other prokaryotes, no other species has the same high degree of gene redundancy. The bacterium also has multiple genes for proteins involved in exporting oxidation products of nucleotides. Another important component may be the presence of DNA repeat elements scattered throughout the genome. These repeats satisfy several expected requirements for involvement in recombinational repair, including that they are intergenic, they are ubiquitous, and they occur at a frequency that is comparable to the number of double-stranded DNA breaks that can be tolerated.

Organism	Genome size relative to *E. coli*	Percentage of proteins belonging to paralogs
Pseudomonas aeruginosa	1.4	75
Escherichia coli	1	50
Caulobacter crescentus	0.88	48
Hemophilus influenzae	0.38	35
Mycoplasma genitalium	0.12	26

Table 18.1 Relationship between paralogs and genome size.

whose members encode distinct functions. As a general rule, one would expect that as the size of the prokaryotic genome increases then the number of paralogs also would increase, and this is what has been observed (Table 18.1). Furthermore, the biochemical bias in these paralogs reflects the biology of the host organism (Box 18.2).

Analysis of all the prokaryotic genomes sequenced to date has revealed two intriguing observations. First, almost half the ORFs identified are of unknown biological function. This suggests that a number of novel biochemical pathways remain to be identified. Secondly, approximately 25% of all ORFs identified are unique and have no significant sequence similarity to any other available protein sequence. Although this might be an artifact of the small number of bacterial species studied by whole-genome analysis, it does support the observation of incredible biological diversity between bacteria. More importantly, it indicates that there are large numbers of new protein families yet to be discovered, e.g. over 1000 proteins in each of *Bacillus subtilis*, *E. coli*, and *Deinococcus radiodurans*!

Because the DNA and protein sequence databases are updated daily it pays to revisit them from time to time to determine if homologs to previously unidentified proteins have been found. It also pays to re-examine sequence data as new and more sophisticated bioinformatics tools are being developed. The benefits of this can be seen from the work of Robinson *et al.* (1994). They re-examined 18 Mb of prokaryotic DNA sequence and uncovered more than 450 genes that had escaped detection. A more specific example is that of Dandekar *et al.* (2000) who re-examined the sequence data for *Mycoplasma pneumoniae*. They identified an additional 12 ORFs and eliminated one identified previously and found an additional three RNA genes. They also shortened eight protein reading frames and extended 16 others.

Horizontal gene transfer may be a significant evolutionary force but is not easy to detect

Horizontal, or lateral, gene transfer is the occurrence of genetic exchange between different evolutionary lineages. It is generally recognized that horizontal gene transfer has occurred but there is considerable debate about the extent of its occurrence. For example, Gogarten *et al.* (2002) believe that it occurs much more than has hitherto been recognized whereas Kurland *et al.* (2003) feel that it has had little influence on genome phylogeny. Now that so many microbial genomes have been sequenced it might be thought that detecting lateral gene transfer would be easy but there are doubts about the validity of some of the methods used to detect it. Basically, two methods are used: the detection of sequences with unusual nucleotide composition and the detection of a gene, or genes, for a function that is totally absent in all closely related species. For example, analysis of the genomes of two bacterial thermophiles indicated that 20–25% of their genes were more similar to genes in archaeabacteria than those of eubacteria (Aravind *et al.* 1998, Nelson *et al.* 1999). These archaeal-like genes occurred in clusters in the genome and had a markedly different nucleotide composition and could have arisen by horizontal gene transfer.

Garcia-Vallve *et al.* (2000) have developed a statistical procedure for predicting whether genes of a complete genome have been acquired by horizontal gene transfer. This procedure is based on analysis of G + C content, codon usage, amino acid usage, and gene position. When it was applied to 24 sequenced genomes it suggested that 1.5–14.5% of genes had been horizontally transferred and that most of these genes were present in only one or two lineages. However, Koski *et al.* (2001) have urged caution in the use of codon bias and base composition to predict horizontal gene transfer. They compared the ORFs

of *E. coli* and *Salmonella typhi*, two closely related bacteria that are estimated to have diverged 100 million years ago. They found that many *E. coli* genes of normal composition have no counterpart in *S. typhi*. Conversely, many genes in *E. coli* have an atypical composition and not only are also found in *S. typhi*, but are found at the same position in the genome, i.e. they are *positional* orthologs.

Karlin (2001) has defined genes as "putative aliens" if their codon usage difference from the average gene exceeds a high threshold and codon usage differences from ribosomal protein genes and chaperone genes also are high. Using this method, in preference to variations in G + C content, he noted that stretches of DNA with anomalous codon usage were frequently associated with pathogenicity islands. These are large stretches of DNA (35–200 kb) that encode several virulence factors and are present in all pathogenic isolates of a species and usually absent from non-pathogenic isolates. Of particular relevance is that they encode an integrase, are flanked by direct repeats, and insert into the chromosome adjacent to tRNA genes (Hacker *et al.* 1997). In this respect, pathogenicity islands resemble temperate phages and could have been acquired by new hosts by transduction (Boyd *et al.* 2001). Alternatively, spread could have been achieved by conjugative transposons. There are many other putative examples of horizontal gene transfer (see Gogarten *et al.* 2002, for a list) but the evidence that transmission occurred in this way is much scantier than for pathogenicity islands with the possible exception of RNA polymerase (Iyer *et al.* 2004).

The comparative genomics of closely related bacteria gives useful insights into microbial evolution

Now that so many microbial genomes have been sequenced it is possible to undertake comparative genomic studies between closely related bacteria or distantly related bacteria. Both kinds of studies are valuable because they reveal different kinds of information. Studies on distantly related bacteria are covered in the next section and here we cover only studies on bacteria that are phylogenetically close.

The most detailed comparative analysis of related bacteria has been undertaken on three genera of the Enterobacteriaceae: *Escherichia, Shigella*, and *Salmonella* (Chaudhuri *et al.* 2004). Initially a comparison was made between one laboratory strain of *E. coli* and two O157 enteropathogenic isolates (Hayashi *et al.* 2001, Perna *et al.* 2001) and later this was supplemented by inclusion of a uropathogenic strain (Welch *et al.* 2002). These studies showed that the genomic backbone is homologous but the homology is punctuated by hundreds of lineage-specific islands of introgressed DNA scattered throughout the genome. Also, the pathogenic strains are 590–800 kb larger than the laboratory strain and this size difference is caused entirely by variations in the amount of island DNA. Many of these islands are at the same relative backbone position in the different pathogens but the island sequences are unrelated. A more surprising finding was that only 39% of the proteins that each strain encodes are common to all of the strains. Furthermore, the pathogen genomes are as different from each other as each pathogen is from the benign strain. A later analysis of the genome of *Shigella flexneri*, a major cause of dysentery, indicated that this bacterium has the same genome structure as *E. coli* and even should be considered as a distinct strain of *E. coli* rather than as belonging to a different genus (Wei *et al.* 2003).

As noted earlier, distinctive codon usage is considered to be an indicator of horizontal gene transfer. Analysis of the different *E. coli* genomes showed that the islands had distinctly different codon usage and a 3–4.5 fold higher use of certain rare codons. Of the approximately 2000 genes that were found in islands in the pathogens only about 10% of them were shared. However, many of these shared genes are related to genes associated with bacteriophages or insertion sequences suggesting that they may have been involved in horizontal gene transfer. Many of the other, non-shared island genes encode known pathogenicity determinants. When different uropathogenic strains are compared, e.g. ones responsible for cystitis, pyelonephritis, and urosepsis, many of their island genes are unique to one strain too. These results suggest that both pathogenic and non-pathogenic strains of *E. coli* have evolved through a complex process. The ancestral backbone genes that define *E. coli* have undergone slow accumulation of vertically acquired sequence changes but the remainder of the genes may have been introduced by numerous occurrences of horizontal gene transfer.

Salmonella species are considered to be close relatives of *E. coli* and two serovars (*S. typhi* and *S. typhimurium*) have been completely sequenced (McClelland *et al.* 2001, Parkhill *et al.* 2001a) and compared to the *E. coli* genome (Fig. 18.2), with which they share extensive synteny. As would be expected, the relationship between *S. typhi* and

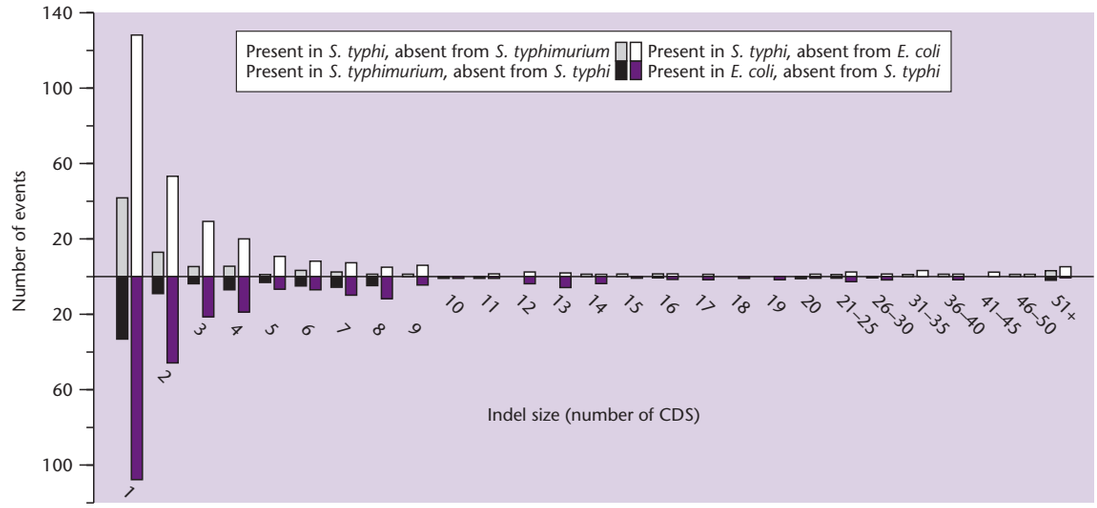

Fig. 18.2 Distribution of insertions and deletions in *S. typhi* relative to *E. coli* and *S. typhimurium*. The graph shows number of insertion–deletion events plotted against the size of the inserted or deleted element (shown as number of genes), clearly indicating that most of the events involve a small number of genes. Values above the lines represent genes present in *S. typhi*; values below the line represent genes absent in *S. typhi*. Dark bars show the comparison with *S. typhimurium*; light bars with *E. coli*. (Redrawn with permission from Parkhill *et al.* 2001b.)

S. typhimurium is very much closer than between *S. typhi* and *E. coli*, although there still are significant differences. There are 601 genes (13.1%) that are unique to *S. typhi* compared with *S. typhimurium* and 479 genes (10.9%) unique to *S. typhimurium* relative to *S. typhi*. By contrast, there are 1505 genes (32.7%) unique to *S. typhi* relative to *E. coli* and 1220 genes (28.4%) unique to *E. coli* relative to *S. typhi*. Another difference between *S. typhi* and *S. typhimurium* is the presence of 204 pseudogenes in the former and only 39 in the latter. In most cases these pseudogenes are relatively recent because they are caused by a single frameshift or stop codon. It is worth noting that complete sequencing of closely related genomes facilitates the detection of pseudogenes. This is because a frame-shift or premature stop codon is only recognizable if the gene is colinear with a functional homologous gene in another genome. One biological difference between the two *Salmonella* serovars is that *S. typhi* only infects humans, whereas *S. typhimurium* can infect a wide range of mammals. This may be related to differences in pseudogene content because many of the pseudogenes in *S. typhi* are in housekeeping functions and virulence components.

The bacterium *Bacillus anthracis* is of much current interest as it is the causative agent of anthrax and has been used as a bioterrorism agent. It has long been considered to be closely related to *B. cereus*, which can cause food poisoning, and *B. thuringiensis*, which is pathogenic for certain insects. A comparative genomic analysis of these three strains has shown that while they differ in their chromosomal backbone the major differences in pathogenicity are due to plasmid-borne genes (Radnedge *et al.* 2003, Rasko *et al.* 2004, Hoffmaster *et al.* 2004). Originally it was thought that *B. cereus* lacked the plasmids pXO1 and pXO2 that respectively encode the lethal toxin complex and the poly-gamma-glutamic acid capsule, both of which are key virulence factors. However, similar plasmids have been found in non-pathogenic *B. cereus* strains and only differ from the corresponding ones from *B. anthracis* by the lack of a pathogenicity island containing various toxin genes.

There have been a number of genomic comparisons made between different species of *Mycobacterium*. Of these, the most interesting is that between *M. tuberculosis* and *M. leprae*, the causative organisms of tuberculosis and leprosy (Table 18.2). Of the 1604 ORFs in *M. leprae*, 1439 had homologs in *M. tuberculosis*. Most of the 1116 pseudogenes were translationally inert but also had functional counterparts in *M. tuberculosis*. Even so, there has still been a massive gene decay in the leprosy bacillus. Genes that have been lost include those for part of the oxidative respiratory chain and most of the microaerophilic and anaerobic ones plus numerous catabolic systems. These losses probably account for the inability of microbiologists to culture *M. leprae* outside of animals. At the genome organization level, 65 segments

Table 18.2 Comparison of the genomes of two *Mycobacterium* spp. (Reproduced from Cole *et al.* 2001.)

Feature	*Mycobacterium leprae*	*Mycobacterium tuberculosis*
Genome size	3,268,203	4,411,532
G + C (%)	57.79	65.61
Protein coding (%)	49.5	90.8
Protein coding genes (No.)	1604	3959
Pseudogenes (No.)	1116	6
Gene density (bp per gene)	2037	1114
Average gene length (bp)	1011	1012
Average unknown gene length (bp)	338	653

showed synteny but differ in their relative order and distribution. These breaks in synteny generally correspond to dispersed repeats, tRNA genes, or gene-poor regions, and repeat sequences occur at the junctions of discontinuity. These data suggest that genome rearrangements are the result of multiple recombination events between related repetitive sequences.

Comparative analysis of phylogenetically diverse bacteria enables common structural themes to be uncovered

Certain structural themes start to emerge as more and more bacterial genomes are sequenced and comparisons made between these sequences. One such theme is the presence of pathogenicity islands in pathogens and their absence from non-pathogens. Another is that chromosomal inversions in closely related bacteria are most likely to occur around the origin or terminus of replication (Eisen *et al.* 2000, Suyama & Bork 2001). Finally, many genomes are littered with prophages and prophage remnants but the exact significance of these is not known.

The systematic comparison of gene order in bacterial and archaeal genomes has shown that there is very little conservation of gene order between phylogenetically distant genomes. A corollary of this is that whenever statistically significant conservation of gene order is observed then it could be indicative of organization of the genes into operons. Wolf *et al.* (2001) undertook a comparison of gene order in all the sequenced prokaryotic genomes and found a number of potential operons. Most of these operons encode proteins that physically interact, e.g. ribosomal proteins and ABC-type transporter cassettes. More important, this analysis enabled functions to be assigned to genes based on predictions of operon function (Chapter 23).

Comparative genomics can be used to analyze physiological phenomena

The bacterium *Deinococcus radiodurans* is characterized by its ability to survive extremely high doses of ionizing radiation. Although the complete genome has been sequenced this has not been sufficient to provide a convincing explanation for the observed physiological phenotype (see Box 18.2, p. 377). Part of the problem is that there are no other organisms that exhibit the same degree of radiation resistance with which to make comparisons. However, Makarova *et al.* (2003) have made more progress with understanding the basis for hyperthermophily. In this context, hyperthermophily is the ability to grow at temperatures exceeding 75°C whereas thermophily is the ability to grow in the range 55–75°C. Complete genome sequences were available for 11 hyperthermophiles including eight archaea from six distinct lineages and three bacteria from diverse phyla. Sequences also were available for 14 thermophiles. Initially a search was made for COGs which met the following criteria:

1. The COGs must encode proteins and be found in at least three hyperthermophiles.
2. The number of hyperthermophiles with a particular COG should be greater than the number of mesophiles.
3. More than 50% of the organisms with a particular COG should be thermophiles.

Altogether, 290 COGs met the above search criteria but most of them were found only in archaeal hyper-

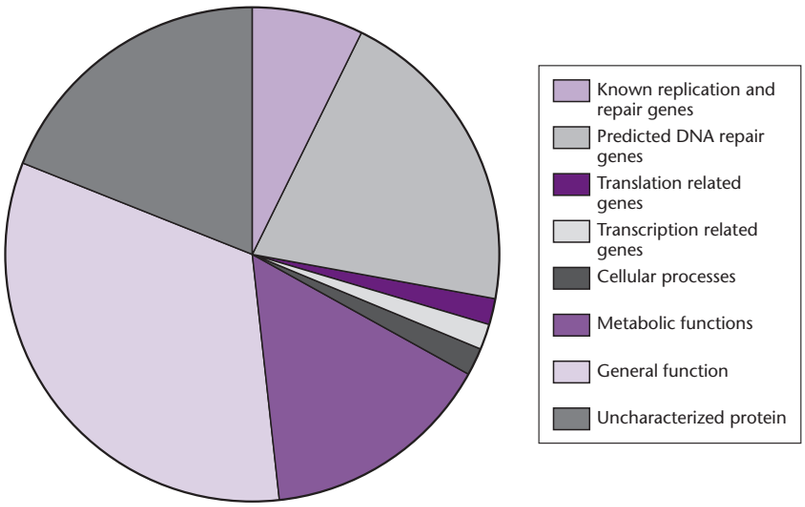

Fig. 18.3 Functions of the 58 COGs associated with hyperthermophily. (Figure reproduced from Makarova *et al.*, 2003, *Trends in Genetics* **19**, 172–6, with permission from Elsevier.)

thermophiles. Therefore the search was refined so that at least one eubacterial hyperthermophile had to encode each COG. In this way 58 COGs were identified as being associated with the hyperthermophilic phenotype. These COGs encode a variety of different cellular functions (Fig. 18.3) and include previously uncharacterized protein families.

Comparative genomics of organelles

Mitochondrial genomes exhibit an amazing structural diversity

Mitochondria are ubiquitous in eukaryotes and play a key role in the generation of ATP through the coupling of electron transport and oxidative phosphorylation. Although the function of mitochondria is highly conserved the structure of the mitochondrial genome exhibits remarkable variation in conformation and size (Fig. 18.4; see Burger *et al.* 2003 for review). Whereas the mtDNAs of animals and fungi are relatively small (15–20 kb), those of plants are very large (200–2000 kb). Plant mitochondria rival the eukaryotic nucleus, and especially the plant nucleus, in terms of the C-value paradox they present: i.e. larger plant mitochondrial genomes do not appear to contain more genes than smaller ones but simply have more spacer DNA. Plant mitochondria also have a large amount of DNA derived from the chloroplast, the nucleus, viruses, and other unknown sources. This process probably is facilitated by the existence of an active, transmembrane potential-dependent mechanism of DNA uptake (Koulintchenko

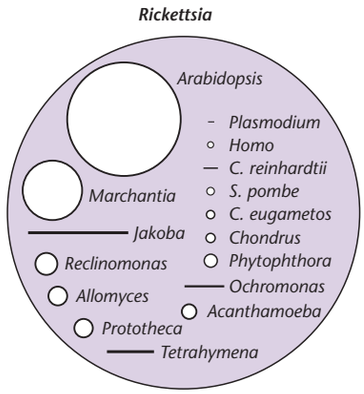

Fig. 18.4 Size and gene content of mitochondrial genomes compared with an α-Proteobacterial (*Rickettsia*) genome. Circles and lines represent circular and linear genome shapes, repectively. (Reprinted from Gray *et al.* 1999 by permission of the American Association for the Advancement of Science.)

et al. 2003). The C-value paradox extends to plant–animal comparisons, where the *Arabidopsis* mtDNA is 20 times larger than human mtDNA but has less than twice the number of genes (Fig. 18.5). Even within a single genus, in this case different species of the yeast *Schizosaccharomyces*, there can be a four-fold variation in the amount of non-coding DNA (Bullerwell *et al.* 2003).

As a result of the steady accumulation of sequence data it now is evident that mtDNAs come in two basic types. These have been designated as "ancestral" and "derived" (Gray *et al.* 1999) and their characteristics are summarized in Table 18.3. It is generally believed that mitochondria are the direct descendants of a bacterial endosymbiont that

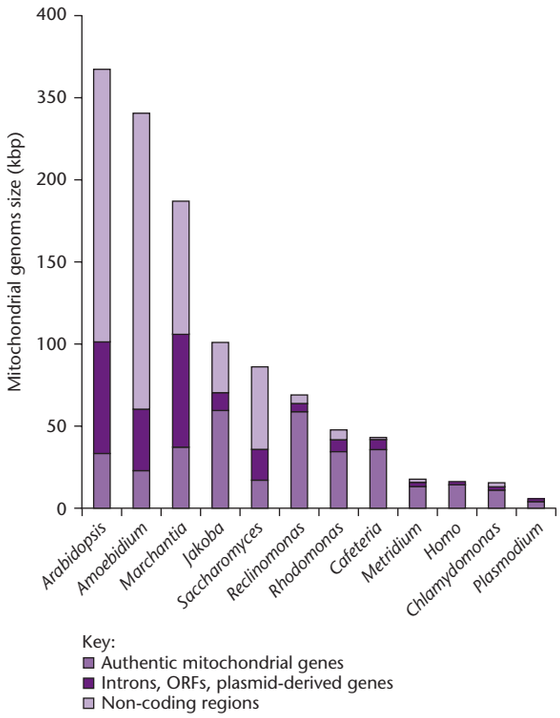

Fig. 18.5 Mitochondrial genome size and coding content across eukaryotes: length of coding regions of authentic mitochondrial genes, introns, intronic ORFs, phage-like reverse transcriptases, and DNA polymerases, and intergenic regions. Species names are: *Reclinomonas americana* (jakobid flagellate); *Rhodomonas salina* (cryptophyte alga); *Marchantia polymorpha* (liverwort, bryophyte); *Cafeteria roenbergensis* (stramenopile flagellate); *Arabidopsis thaliana* (flowering plant, angiosperm); *Homo sapiens* (vertebrate animal); *Metridium senile* (cnidarian animal); *Saccharomyces cerevisiae* (ascomycete fungus); and *Plasmodium falciparum* (apicomplexan protist). *Amoebidium parasiticum* (ichthyosporean protist); *Jakoba libera* (jakobid flagellate); and *Chlamydomonas reinhardtii* (green alga, chlorophyte). Reproduced from Burger *et al.*, 2003, with permission from Elsevier.

became established in a nucleus-containing cell and an ancestral mitochondrial genome is one that has retained clear vestiges of this eubacterial ancestry. The prototypal ancestral mtDNA is that of *Reclinomonas americana*, a heterotrophic flagellated protozoon. The mtDNA of this organism contains 97 genes including all the protein-coding genes found in all other sequenced mtDNAs. Derived mitochondrial genomes are ones that depart radically from the ancestral pattern. In animals and many protists this is accompanied by a substantial reduction in overall size and gene content. In plants, and particularly angiosperms, there has been extensive gene loss but size has increased as a result of frequent duplication of DNA and the capture of sequences from the chloroplast and nucleus (Marienfeld *et al.* 1999).

If mitochondria are derived from a bacterium, what is the closest relative of that bacterium that exists today? The current view is that it is *Rickettsia prowazekii*, the causative agent of epidemic typhus. This organism favors an intracellular lifestyle that could have initiated the endosymbiotic evolution of the mitochondrion. The genome of *R. prowazekii* has been sequenced and the functional profile of its genes shows similarities to mitochondria (Andersson *et al.* 1998). The structure, organization, and gene content of the bacterium most resemble those of the mtDNA of *Reclinomonas americana*.

Gene transfer has occurred between mtDNA and nuclear DNA

The principal function of the mitochondrion is the generation of ATP via oxidative phosphorylation. At least 21 genes encode proteins critical for oxidative phosphorylation and one would expect all of these

Table 18.3 Properties of ancestral and derived mtDNAs. (Reprinted from Marienfeld *et al.* 1999 by permission of Elsevier Science.)

Ancestral mtDNA	Derived mtDNA
1 Many extra genes compared with animal mtDNA	**1** Extensive gene loss
2 rRNA genes that encode eubacteria-like 23S, 16S, and 5S rRNAs	**2** Marked divergence in rDNA and rRNA structure
3 Complete, or almost complete, set of tRNA genes	**3** Accelerated rate of sequence divergence in both protein-coding and rRNA genes
4 Tight packing of genetic information with few or no introns	**4** Highly biased use of codons including, in some cases, elimination of certain codons
5 Eubacterial-like gene clusters	**5** Introduction of non-standard codon assignments
6 Use standard genetic code	

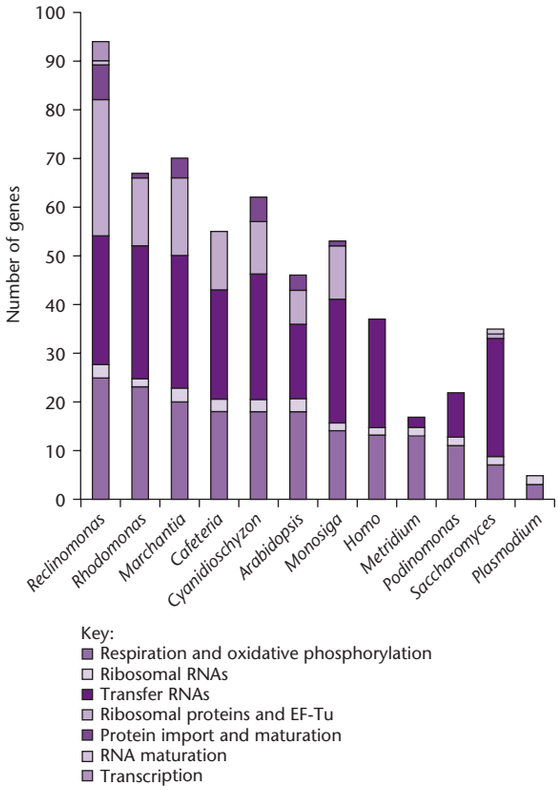

Fig. 18.6 Mitochondrial gene classes and their representation across eukaryotes. Species names are: *Reclinomonas americana* (jakobid flagellate); *Rhodomonas salina* (cryptophyte alga); *Marchantia polymorpha* (liverwort, bryophyte); *Cafeteria roenbergensis* (stramenopile flagellate); *Cyanidioschyzon merolae* (red alga); *Arabidopsis thaliana* (flowering plant, angiosperm); *Monosiga brevicollis* (choanozoan flagellate); *Homo sapiens* (vertebrate animal); *Metridium senile* (cnidarian animal); *Pedinomonas minor* (green alga, chlorophyte); *Saccharomyces cerevisiae* (ascomycete fungus); and *Plasmodium falciparum* (apicomplexan protist). Reproduced from Burger *et al.*, 2003, with permission from Elsevier.

genes to be located in the mtDNA. Similarly, an mtDNA location would be expected for the genes encoding the 14 ribosomal proteins that are required to translate mtRNA. However, sequence data indicate that many mitochondrial genomes lack a number of key genes (Fig. 18.6) and the missing genes can be found in the nucleus. Functional transfer of mitochondrial genes to the nucleus has stopped in animals, hence their consistency in size. Part of the reason for this is that further transfer is blocked by changes in the mitochondrial genetic code. However, this gene transfer continues to occur in plants and protists because there is no genetic code barrier to transfer. Note that it is not just intact genes that are transferred for Woischnik & Moraes (2002) found human mitochondrial pseudogenes in the nuclear genome. Many of these pseudogenes comprised parts of two adjacent mitochondrial genes.

In the case of the mitochondrial *cox2* gene, transfer to the nucleus is still on-going in the case of the legumes (Palmer *et al.* 2000). Analysis of 25 different legumes identified some genera in which the *cox2* gene was located in the mitochondrion, some in which it was nuclear, and some where it was present in both genomes. In most cases where two copies of the gene are present, only one gene is transcriptionally active, although at least one genus was found in which both genes are transcribed.

Adams *et al.* (2000a) studied the distribution of the *rps10* gene in 277 angiosperms and identified 26 cases where the gene has been lost from the mtDNA. In 16 of these loss lineages, the nuclear gene was characterized in detail. To be active in the nucleus, a gene acquired from mtDNA must be inserted into the nuclear genome in such a way that a mature translatable mRNA can be produced. Moreover, the resulting protein is made in the cytoplasm and must be targeted to and imported into mitochondria. What emerged was that in some cases pre-existing copies of other nuclear genes have been parasitized with the *rps10* coding sequence. In several instances a mitochondrial targeting sequence has been co-opted to provide entry for the RPS10 protein back into the mitochondrion but different nuclear genes provide this sequence in different plants. In other cases, the RPS10 protein is imported despite the absence of an obvious targeting sequence. These results, and similar findings for other mitochondrial genes (Adams *et al.* 2001), provide confirmation that nuclear transfer is on-going and has happened on many separate occasions in the past. Nor is nuclear transfer confined to mitochondrial genes for Millen *et al.* (2001) have made similar observations with chloroplast genes. Henze & Martin (2000) have reviewed the mechanisms whereby this transfer can occur.

Horizontal gene transfer has been detected in mitochondrial genomes

In the previous section we discussed intracellular horizontal evolution whereby genes moved between the mitochondrion and the nucleus. However, cross-species acquisition of DNA by plant mitochondrial genomes also has been detected. The first example detected was that of a homing group I intron (Palmer

et al. 2000). These introns encode site-specific endonucleases with relatively long target sites that catalyze their efficient spread from intron-containing alleles to intron-lacking alleles of the same gene in genetic crosses. This intron has been detected in the mitochondrial *cox1* gene of 48 angiosperms out of 281 tested. Based on sequence data for the intron and the host genome, it appears that this intron has been independently acquired by cross-species horizontal transfer to the host plants on many separate occasions. What is not clear are the identities of the donor and recipient in each individual case. By contrast with this group I intron, the 23 other introns in angiosperm mtDNA belong to group II and all are transmitted in a strictly vertical manner.

More recently, Bergthorsson *et al.* (2003) have reported widespread horizontal transfer of mitochondrial genes between distantly related angiosperms, including between monocotyledonous and dicotyledenous plants. The genomic consequences of these mtDNA-to-mtDNA transfers include gene duplication, recapture of genes previously lost through transfer to the nucleus, and a chimeric (half-monocot, half-dicot) ribosomal protein gene.

Comparative genomics of eukaryotes

The minimal eukaryotic genome is smaller than many bacterial genomes

In determining the minimal genome we are seeking to answer a number of different questions. What is the minimal size of the genome of a free-living unicellular eukaryote and how does it compare with the minimal bacterial genome? That is, what are the fundamental genetic differences between a eukaryotic and a prokaryotic cell? Next, what additional genetic information does it require for multicellular coordination? In animals, what are the minimum sizes for a vertebrate genome and a mammalian genome? Finally, what is the minimum size of genome for a flowering plant? Given that many eukaryotic genomes contain large amounts of noncoding DNA these questions have to be answered by considering both genome size and the number of proteins that are encoded.

The smallest eukaryotic genome that has been sequenced is that of the obligate intracellular parasite *Encephalitozoon cuniculi* (Katinka *et al.* 2001). This has a genome size of only 2.9 Mb although its close relative *E. intestinalis* may have a genome that is even smaller (2.3 Mb). Genome compaction in these organisms is achieved by a reduction in the length of intergenic spacers and a shortness of most putative proteins relative to their orthologs in other eukaryotes. Even so, *E. cuniculi* has approximately 2000 ORFs, which is 7–8 times the number in the minimal bacterial genome. The genome of the yeast *Schizosaccharomyces pombe* has about 4800 ORFs (Wood *et al.* 2002) but is unlikely to represent the minimal free-living eukaryotic genome unless the *E. cuniculi* genome has lost many more essential genes than those metabolic and biosynthetic ones already recognized. The multicellular fungus *Neurospora crassa* has approximately 10,000 ORFs (Galagan *et al.* 2003), about 25% fewer than the fruitfly *Drosophila melanogaster* (Adams *et al.* 2000b). Many of these genes do not have homologs in either *Saccharomyces cerevisiae* or *S. pombe* (Borkovich *et al.* 2004) but exactly how many of them are essential for multicellular existence remains to be seen.

Comparative genomics can be used to identify genes and regulatory elements

As noted in Chapter 9 accurately identifying genes in a complete genome sequence can be very difficult and identifying regulatory elements can be even harder still. A powerful method for finding functional elements such as genes and regulatory regions is to align orthologous genomic sequences from different species and search out regions of sequence conservation. The rationale for this approach is that mutations in functional DNA will be deleterious and thus counter-selected thereby resulting in a reduced rate of evolution of functional elements. The two most important factors affecting the results of a comparative analysis are the amount of divergence being captured and the phylogenetic scope of the aligned sequences (Cooper & Sidow 2003). The amount of divergence affects the power and resolution of the analyses. The scope, which is defined as the narrowest taxonomic group that encompasses all analyzed sequences, affects the applicability of conclusions and the generality of the results. For example, a dipteran scope that includes *Drosophila* (fruitfly) and *Anopheles* (mosquito) can be used to find elements that were present in their common ancestor as well as ones present before the diversification of hexapods, arthropods, and metazoa (Fig. 18.7).

An example of a comparative analysis with a narrow scope is the genomic comparison of *S. cerevisiae* with three other species of *Saccharomyces* (Kellis *et al.*

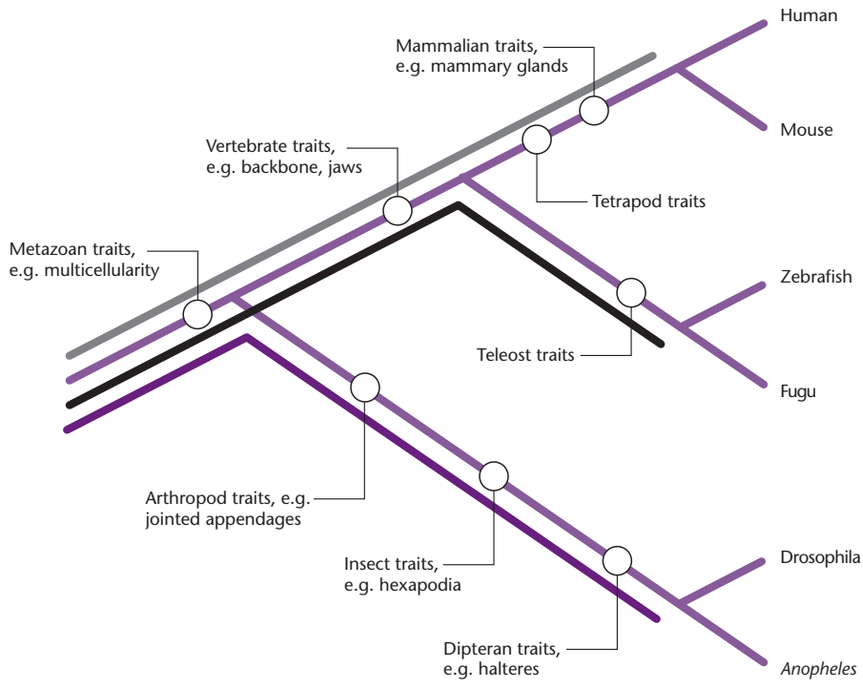

Fig. 18.7 The importance of scope and the impact of shared ancestry on comparative sequence analysis. The tree describing the relationships among six actively studied genomes is drawn in light purple (not to scale). Each colored line indicates the phylogenetic scope that applies to each pair of species at the terminal nodes: gray line, placental mammal scope; black line, teleost scope; dark purple line, dipteran scope. Overlaps of the colored lines indicate shared ancestry and capture traits shared by the indicated scopes and, by implication, shared functional elements. Open circles and associated text show various traits that exemplify the major animal clades and the branch of the tree on which they arose. Reproduced from Cooper *et al.* (2003), with permission from Elsevier.

2003). The gene analysis resulted in a major revision of the *S. cerevisiae* gene catalog that affected 15% of all genes, reduced the total count by about 500 genes, and identified 43 new small ORFs (50–99 amino acids). This latter finding is particularly significant since small ORFs can only be considered putative genes in the absence of function or conservation in different species. A comparative analysis with a more divergent scope is that between the pufferfish (*Fugu rubripes*) and human genomes (Aparicio *et al.* 2002). This identified almost 1000 putative genes that had not been identified in the two published reports on the human genome sequence.

The direct identification of regulatory elements is very difficult since they are short (6–15 bp), tolerate some degree of sequence variation, and follow few known rules. Computational analysis of single genome sequences has been used successfully to identify regulatory elements such as promoters associated with known sets of genes. However, this approach is of relatively little value in identifying other regulatory elements involved in gene expression (enhancers, silencers) and chromatin organization (insulators, matrix attachment regions). As the examples below show, comparative analysis is much more useful in this respect. Comparisons within a narrow scope are particularly useful as they permit almost the entire genome to be scanned for regulatory regions. In this way Kellis *et al.* (2003) were able to recognize an additional nine regulatory protein motifs in addition to the 42 that were already known.

Enhancers are regulatory elements that upregulate gene expression by sequence-specific positioning of transcriptional activators. Enhancers can function independently of position and orientation although they generally are located within hundreds of kilobases of their target genes. Using comparative analysis, Spitz *et al.* (2003) discovered a cluster of enhancer elements that are conserved between mammals and *Fugu*. These enhancers coordinate expression between *Hoxd* genes and nearby genes that are evolutionarily unrelated.

Silencers are elements that are capable of repressing transcription. Many are found near their corresponding promoter but there are other types. Sequencing of the chicken CD4 gene showed that it

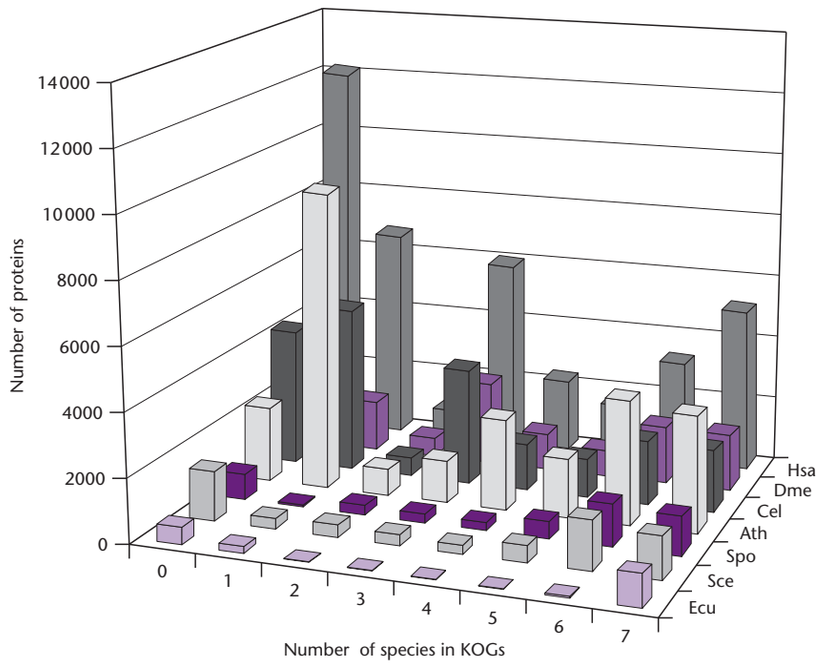

Fig. 18.8
Assignment of proteins from each of the seven analyzed eukaryotic genomes to KOGs with different numbers of species and to LSEs. 0, Proteins without detectable homologs (singletons); 1, LSEs. Species abbreviations: Ath, *Arabidopsis thaliana*; Cel, *Caenorhabditis elegans*; Dme, *Drosophila melanogaster*; Ecu, *Encephalitozoon cuniculi*; Hsa, *Homo sapiens*; Sce, *Saccharomyces cerevisisae*; Spo, *Schizosaccharomyces pombe*.

is similar to the mammalian CD4 gene and has a functional human silencer (Koskinen *et al.* 2002). This level of distant conservation suggests that this silencer has a fundamental role in controlling gene expression.

Insulator elements are barriers that separate domains within chromatin and confine the actions of regulatory elements to their appropriate targets. They can block the action of enhancers as well as prevent the spread of chromatin condensation from nearby regions. Farrell *et al.* (2002) discovered conserved genomic regions that flank the β-globin loci in mouse and man. These regions contain binding sites for CTCF, a protein known to be important for enhancer-blocking insulator activity.

Matrix attachment regions (MARs) are regions of DNA that are involved in the binding to the nuclear matrix. Glazko *et al.* (2003) aligned intergenic sequences from mouse and man and identified conserved segments. Further analysis showed that 11% of these had sequence motifs characteristic of MARs and that many of them precede the 5′ ends of genes. This latter observation suggests a role in regulating transcription.

Comparative genomics gives insight into the evolution of key proteins

Koonin *et al.* (2004) have undertaken a comprehensive evolutionary classification of the proteins encoded in seven completely sequenced eukaryotic genomes: three animals (man, nematode, and fruitfly), one plant (*Arabidopsis*), a budding yeast, a fission yeast, and the microsporidian *E. cuniculi.* In particular, they looked for eukaryotic clusters of orthologous groups (KOGs) and the results are shown in Fig. 18.8. The fraction of proteins assigned to KOGs tends to decrease with increasing genome size, except for the obligate parasite *E. cuniculi*. By contrast, lineage-specific expansions of paralogous groups show the opposite trend with the largest numbers being in the higher eukaryotes. Only a minority of KOGs have readily detectable prokaryotic counterparts, indicating the extent of innovation linked to the origin of eukaryotes.

A total of 131 KOGs were represented by a single gene in each of the seven genomes. Since these KOGs are present in the minimal genome of *E. cuniculi* they must encode core biological functions. Nearly all of them encode subunits of known multiprotein complexes and many of them are involved in rRNA processing, ribosome assembly, intron splicing, transcription, and protein assembly and trafficking.

The evolution of species can be analyzed at the genome level

The yeasts *Saccharomyces paradoxus*, *S. mikatae*, and S. *bayanus* are estimated to have separated from S. *cerevisiae* 5–20 million years ago. The genomes of all four have been sequenced and Kellis *et al.* (2003)

Species	Reciprocal Translocations	Inversions	Segmental Duplications
S. paradoxus	0	4	3
S. mikatae	4	13	0
S. bayanus	5	3	0

Table 18.4 Genomic rearrangements in three yeast species when compared with *S. cerevisiae*.

have undertaken a comparative analysis. They found a high level of "genomic churning" in the vicinity of the telomeres and gene families in these regions showed significant changes in number, order, and orientation. Only a few rearrangements were seen outside of the telomeric regions and these are summarized in Table 18.4. All 20 inversions were flanked by tRNA genes in opposite transcriptional orientation and usually these were of the same isoacceptor type. The role of tRNA genes in genomic inversion has not been noted before. Of the nine translocations, seven occurred between Ty elements and two between highly similar pairs of ribosomal genes.

At the gene level, five genes were unique to *S. paradoxus*, eight genes unique to *S. mikatae*, and 19 unique to *S. bayanus*. Most of them encoded functions involved in sugar metabolism or gene regulation. The majority (86%) of these unique genes were located near a telomere or a Ty element, locations that are consistent with rapid genome evolution. One gene was identified that appears to be evolving very rapidly and across the four species showed 32% nucleotide identity and 13% amino acid identity. Functionally it appears to be involved in sporulation, which in yeast is a stage in sexual reproduction. In this regard, it is consistent with the observation that many of the best-studied examples of positive selection in other organisms are genes related to gamete function. One gene also was identified that showed perfect 100% conservation at the amino acid and the nucleotide level. The latter observation is very unusual given the redundancy of the genetic code and suggests that the gene might encode an antisense RNA.

Analysis of dipteran insect genomes permits analysis of evolution in multicellular organisms

The fruit fly *Drosophila melanogaster* and the malaria mosquito *Anopheles gambiae* are both highly adapted, successful dipteran species that diverged about 250 million years ago. They share a similar body plan and a considerable number of other features but differ in terms of ecology, morphology, and life style. For example, *Drosophila* feeds on decaying fruit while *Anopheles* feeds on the blood of specific hosts. A number of obvious differences can be seen at the whole-genome level (Table 18.5) but these give little insight into the evolutionary process.

Table 18.5 Genome statistics for the mosquito and fruit fly.

	Anopheles	*Drosophila*
Genome size	278 Mb	165 Mb
Total exon length	10 Mb	13.6 Mb
Total intron length	22.6 Mb	12.9 Mb
Average introns per gene	3.5	4.7
Average amino acids per protein	548	649

When the two genomes are compared at the protein level (Zdobnov *et al.* 2002) five classes of protein can be recognized (Fig. 18.9). A total of 6089 orthologs were identified in the two species and their average sequence identity was 56%. By contrast there is 61% sequence identity of orthologs between the pufferfish and humans, which diverged 450 million years ago. This indicates that insect proteins diverge at a higher rate than vertebrate proteins. This could be because insects have a much shorter life cycle and may experience different selective pressures. When the orthologs are classified according to gene ontology it is not surprising to find that the proteins involved in immunity show the greatest divergence and structural proteins are the most conserved.

The "many-to-many" orthologs shown in Fig. 18.9 represent groups of genes in which gene duplication has occurred in one or both species after divergence,

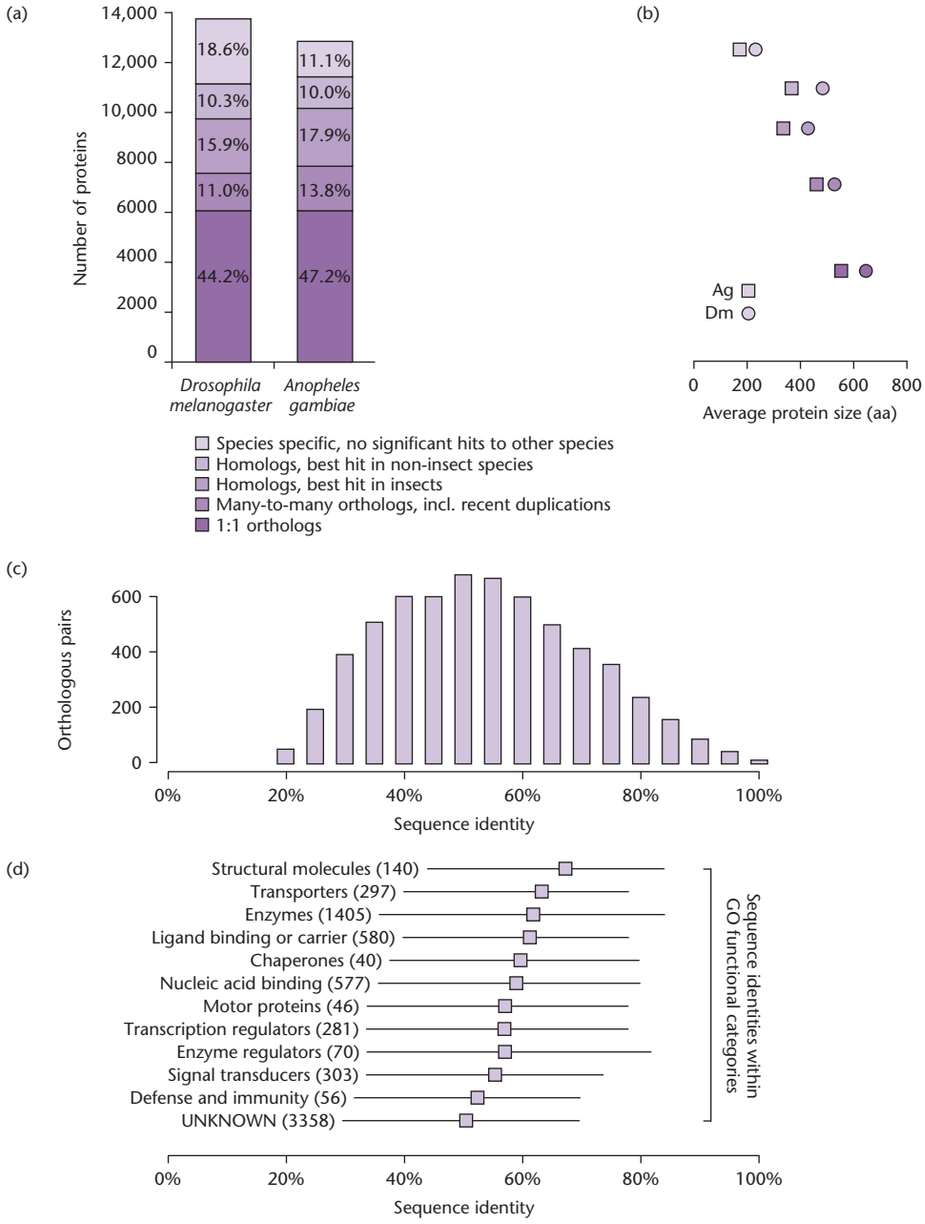

Fig. 18.9 Analysis of the proteome of *Anopheles gambiae* (Ag) and *Drosophila melanogaster* (Dm). (a) Classification of the proteins according to their conservation. (b) Plot of the average protein length for each protein class in A. (c) Histogram of the sequence identities of the 1:1 orthologs. (d) Sequence conservation of the 1:1 orthologs by functional category. Reprinted with permission from Zdobnov *et al.* (2002), Copyright (2002) AAAS.

i.e. paralogy. These, and the homologs, probably represent adaptations to environment and life strategies leading to changes in cellular and phenotypic features. For example, four *Anopheles* paralogs without a counterpart in *Drosophila* are similar to the human gene encoding leukotriene B4 12-hydroxy dehydrogenase, an enzyme that can inactivate the proinflammatory leukotriene B4. The anopheline mosquito may have acquired this gene to facilitate the taking of a blood meal. A total of 579 orthologs were restricted to *Anopheles* and *Drosophila* and did not even share domains with proteins identified in the other organisms whose genomes have been sequenced. Most of those that could be annotated encoded specific odorant and taste receptors, cuticle proteins, pheromone and pheromone-binding proteins, and insect-specific defense molecules.

The dynamics of gene evolution can be analyzed by comparing the intron and exon structure of the 1:1 orthologs. For example, equivalent introns in *Drosophila* have only half the length of those in *Anopheles* whereas exon lengths and intron

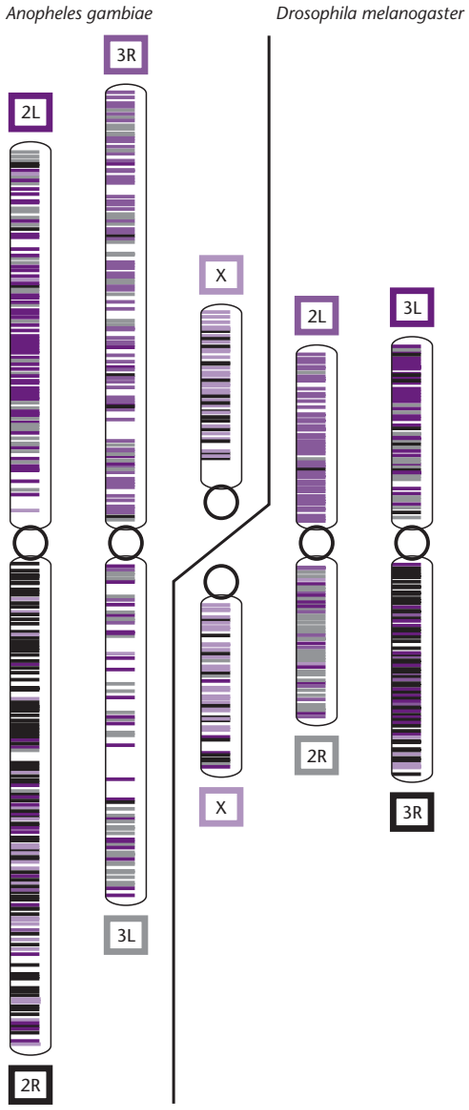

Fig. 18.10 Homology of chromosomal arms in insects. Each chromosomal arm is marked by a color shown around its name (pairs of chromosomes with significant homology, such as *Dm*2L/*Ag*3R, use the same color). Coloring inside the schematic chromosome arms denotes microsynteny matches to a region in the other species; the color shown is the color of the chromosome containing the matching region in the other species. Reprinted with permission from Zdobnov *et al.* (2002). Copyright (2002) AAAS. Full-color version available at www.blackwellpublishing.com/primrose

frequencies are roughly similar. Approximately 55% of *Anopheles* introns in 1:1 orthologs have equivalent positions in *Drosophila* but almost 10,000 introns have been lost or gained between the two species. The rate of gain or loss of introns has been calculated to be one per gene per 125 million years.

Given that the two diptera being studied are estimated to have diverged 250 million years ago one would expect that, in addition to changes in exon/intron structure, there would be significant variations in genome structure. Indeed, the gene order of the 1:1 orthologs has only been retained over very small distances and this is referred to as microsynteny. However, at the macro level, chromosomal arms exhibit significant remnants of homology between the two species and major inter-arm transfers and intra-arm shuffling of gene order can be detected (Fig. 18.10).

A number of mammalian genomes have been sequenced and the data is facilitating analysis of evolution

The genomes of humans, the mouse, and the rat have been completely sequenced and good progress is being made with the genome of the chimpanzee (International Human Genome Sequencing Consortium 2001, Venter *et al.* 2001, Mouse Genome Sequencing Consortium 2002, Rat Genome Sequencing Project Consortium 2004, The International Chimpanzee Chromosome 22 Consortium 2004). Figure 18.11 shows an analysis of the three completely sequenced genomes. About 1 billion nucleotides (40% of rat genome) align in all three species and this "ancestral core" contains 94–95% of the known coding exons and regulatory regions, which in turn represent 1–2% of the genome. A further 30% of the rat genome aligns only with the mouse genome and consists largely of rodent-specific repeats. A further 15% of the rat genome comprises rat-specific repeats. More genomic changes have been detected in the rodent lineages than in the human. These include approximately 250 large rearrangements between a hypothetical rodent ancestor and human, approximately 50 between this ancestor and rat, and a similar number between the ancestor and the mouse.

The rat, mouse, and human genomes encode similar numbers of genes and the majority have persisted without deletion or duplication since the last common ancestor. About 90% of the genes have strict orthologs in all three genomes but, compared with humans, the rodents have expanded gene families for functions associated with reproduction, immunity, olfaction, and metabolism of xenobiotics. These features are not surprising given what we know about rodent biology! Almost all the human genes known

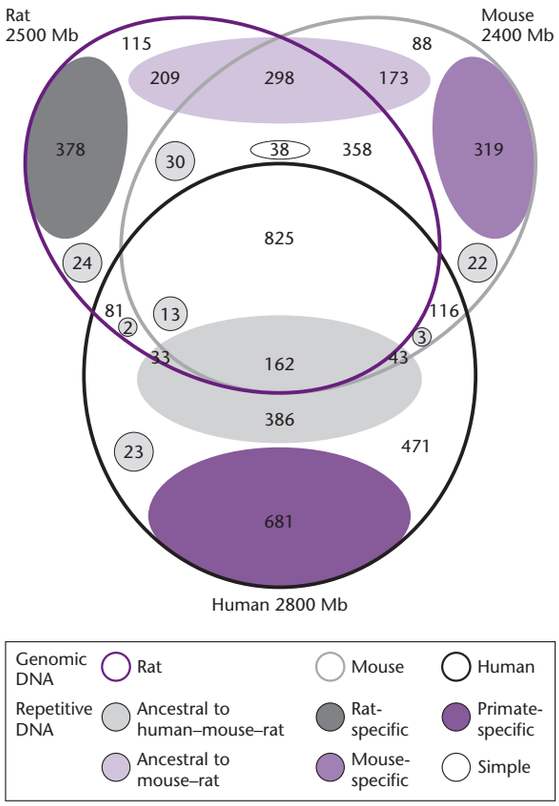

Fig. 18.11 Aligning portions and origins of sequences in rat, mouse, and human genomes. Each outlined ellipse is a genome, and the overlapping areas indicate the amount of sequence that aligns in all three species (rat, mouse, and human) or in only two species. Non-overlapping regions represent sequence that does not align. Types of repeats classified by ancestry: those that predate the human–rodent divergence, those that arose on the rodent lineage before the rat–mouse divergence, species-specific, those that are rat-specific, mouse-specific, human-specific and simple, each indicated as shown in the key and placed to illustrate the approximate amount of each type in each alignment category. Uncolored areas are non-repetitive DNA – the bulk is assumed to be ancestral to the human–rodent divergence. Numbers of nucleotides (in Mb) are given for each sector (type of sequence and alignment category). Reproduced with permission from *Nature*.

to be associated with disease have orthologs in the rat and mouse genomes but there is one surprising finding. Many SNPs causing disease in man are found in mice but these mice are phenotypically normal.

The comparison of the human genome with that of the chimpanzee is perhaps the most interesting of all the genomic comparisons that can be made, as the chimpanzee is our closest living relative. In particular, comparative analysis should help to uncover the genetic basis of cognitive function, bipedalism, and speech development. At the time of writing the complete chimpanzee genome was not available but the 33.3 Mb sequence of chromosome 22 had been completed (The International Chimpanzee Chromosome 22 Consortium 2004). Nearly 1.5% of the chimpanzee genome had single base substitutions when compared with its human equivalent (chromosome 21) in addition to approximately 68,000 insertions or deletions. These differences are sufficient to generate changes in most of the 231 coding sequences. In addition, different expansion of particular subfamilies of retrotransposons was observed between the different lineages, suggesting different impacts of retrotransposition on human and chimpanzee evolution. The full impact of these changes remains to be deciphered.

Comparative genomics can be used to uncover the molecular mechanisms that generate new gene structures

The comparative analyses described in the previous sections indicate that there is a general process of new gene origination. This raises the question of the origin of these new genes. Several molecular mechanisms are known to be involved in the creation of new gene structures (Fig. 18.12) and can operate singly or in combination (Long *et al.* 2003). A good example is *jingwei*, the first identified gene that has originated recently (2 My) in the evolutionary timescale (Fig. 18.13). This gene arose in the common ancestor of two *Drosophila* species. The starting point was the *yellow emperor* gene that duplicated to give the *yellow emperor* and *yande* genes. Whereas *yellow emperor* maintained its original functions, *yande* underwent modification. In particular, mRNA of the alcohol dehydrogenase gene retroposed into the third intron of *yande* as a fused exon and recombined with the first three *yande* exons. This formed *jingwei*, a gene that is translated into a chimeric protein.

Once created, new genes such as *jingwei* may become modified beyond recognition. Examples of this kind of change include domains involved in protein–protein interactions such as von Willebrand A, fibronectin type III, immunoglobulin, and SH3 modules (Ponting *et al.* 2000). These domains show extensive proliferation in higher eukaryotes but have only a distant relationship to homologs in prokaryotes and lower eukaryotes.

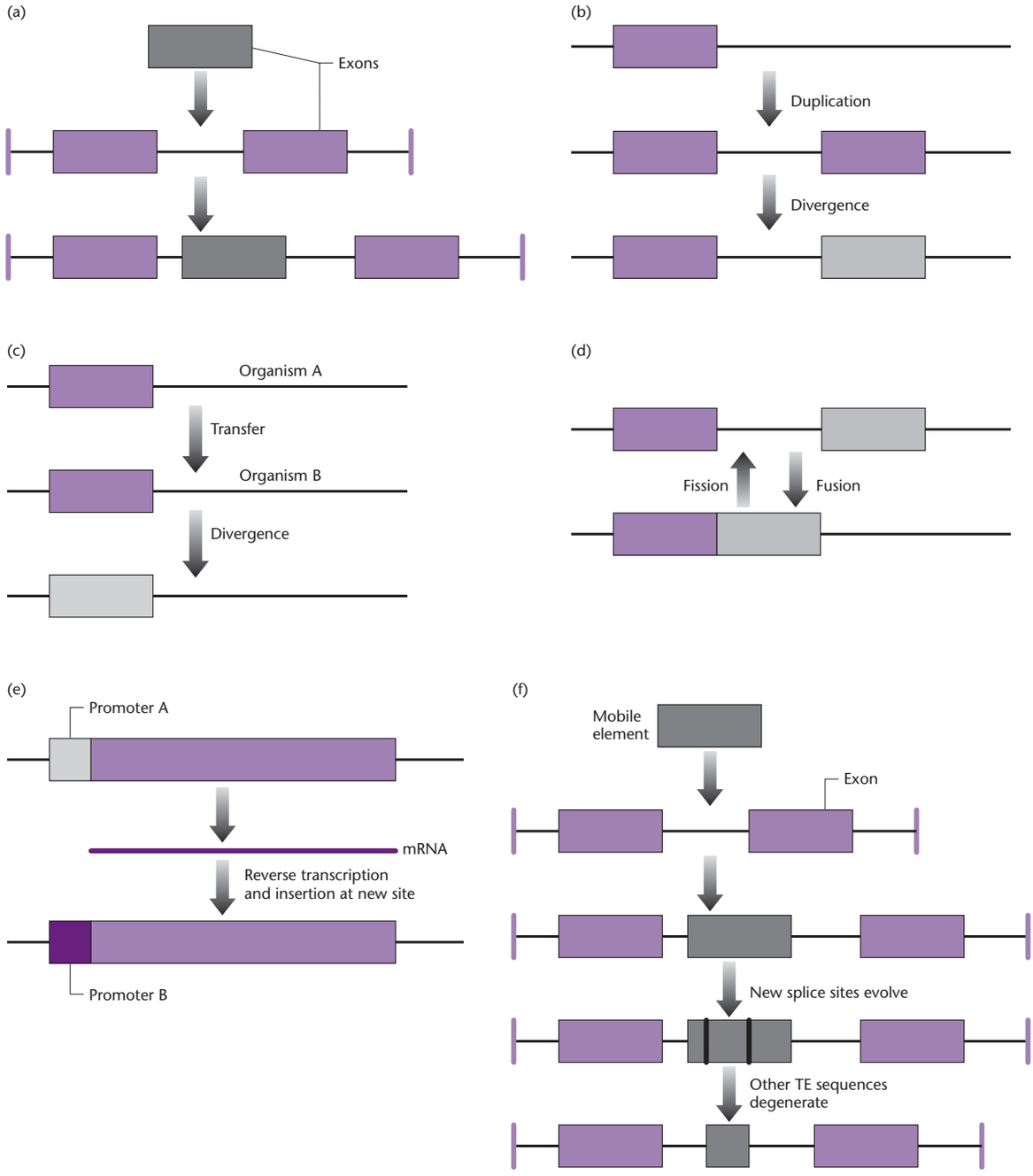

Fig. 18.12 Mechanisms whereby new genes arise. (a) exon capture (exon shuffling); (b) duplication of a gene followed by sequence divergence of the duplicate; (c) divergence of a gene following transfer to a new host; (d) fusion of two separate genes or separation of two fused activities; (e) movement of a gene sequence via an mRNA intermediate followed by coupling to a promoter; (f) capture of a transposable element (TE) followed by degeneration of the TE sequences. Real examples of these mechanisms can be found in the review of Long *et al.* (2003).

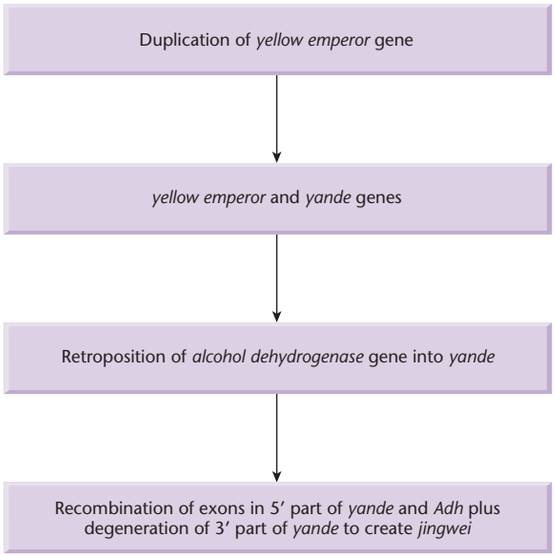

Fig. 18.13 Genomic events leading to the formation of the new gene *jingwei*.

Suggested reading

Kellis M., Patterson N., Endrizzi M., *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–54.
This is rapidly becoming a classic paper on the use of comparative genomics to decipher genome sequences but it also provides insights to the genomic changes that exist between species.

Koonin E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology* **1**, 127–36.

Koonin E.V., Federova N.D., Jackson J.D., *et al.* (2003) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* **5**, R7.
Eugene Koonin probably knows more than anyone about extracting evolutionary information from sequence databases. The two papers cited above are but a tiny sample of his analyses.

Koonin E.V. (2005) Virology: Gulliver among the Lilliputians. *Current Biology* **15**, R167–9.
An analysis of the genome of a virus that is much bigger than many parasitic bacteria.

Long M., Betran E., Thornton K. & Wang W. (2003) The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics* **4**, 865–75.
This is one of the few reviews that attempt to discuss where new genes come from.

Paterson A.H., Bowers J.E., Chapman B.A., *et al.* (2004) Comparative genome analysis of monocots and dicots, towards characterization of angiosperm diversity. *Current Opinion in Biotechnology* **15**, 120–5.

Pedulla M.L., Ford M.E., Houtz J.M., *et al.* (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**, 171–82.
These two papers cover topics not discussed in this chapter, the comparative genomics of plants and viruses respectively, and are well worth reading.

Each year, the January 1 issue of *Nucleic Acids Research* is devoted to short reviews of the different molecular biology and genomics databases. A considerable number of these databases are for the purposes of comparative genomics and all are linked to relevant websites. An example is given below.

Useful website

http://colibase.bham.ac.uk
This is the website for *coli*BASE, an online database for the comparative genomics of *E. coli* and its close relatives. Now that a number of different strains of *E. coli* have been completely sequenced it is clear that there is much more genomic heterogeneity than expected.