CHAPTER 9

Comparative Genomics in Eukaryotes

ALAN FILIPSKI AND SUDHIR KUMAR

Although the word "genome," meaning the total hereditary material of an organism, was coined in 1920 (see Chapter 1), the general concept goes back at least as far as the 4th century BCE, when Aristotle implicated blood as the heredity substance. The blood of the mother, it was thought, supplied matter to the developing fetus whereas the semen (a purified form of blood) of the father conveyed form (Aristotle, 1953). Ironically, although the notions of "blood relations" and characteristics being "in one's blood" persist, it is now known that the blood of mammals actually contains very little genetic material because their erythrocytes contain neither nuclei nor mitochondria (see Chapter 1). As scientific method and technique advanced, heredity eventually came to be associated with bodies called chromosomes in the nuclei of cells (late 19th and early 20th centuries) and finally with the long double-stranded nucleotide polymers called DNA molecules that are wound up within those chromosomes (mid-20th century).

This chapter outlines the development and current status of comparative eukaryotic genomics, from the earliest studies of basic chromosome structure to the sequencing of entire genomes. In the process, a review is provided of the structure, organization, and composition of the primary eukaryotic genomes that have been sequenced thus far. This is a truly exciting time for the biological

The Evolution of the Genome, edited by TR Gregory

sciences, with avenues of research now opening up that had not even been conceived only a few decades ago. Some of the vast possibilities that are already apparent are discussed at the end of the chapter, but this is necessarily a highly truncated list owing to the ever-accelerating rapidity with which the field is advancing.

THE EARLY HISTORY OF COMPARATIVE EUKARYOTIC GENOMICS

The Basics of Eukaryotic Chromosome Structure

Figure 9.1A depicts a typical eukaryotic chromosome in the unreplicated form. Photographic representations often show the chromosome while it is replicating during mitosis, because chromosomes are easier to photograph in this stage. In this latter case (Fig. 9.1B), each chromosome looks more like the letter X, with each arm in two replicates (sister chromatids) emanating from the centromere. The shorter arm is generally depicted at the top of the image and (in humans) is designated as the "p" arm, from the French *petit bras* ("little arm"); the longer arm is labeled the "q" arm, after the French *queue* ("tail"). Chemical stains, described in more detail later, bring out characteristic light and dark banding patterns. Conventions for designating individual chromosomes with numbers or letters are historical (and often idiosyncratic) and reflect usage that evolved within specific communities of investigators.

Before the advent of techniques for reading the sequence of nucleotide bases of a DNA molecule, the chromosome provided the most detailed view available of

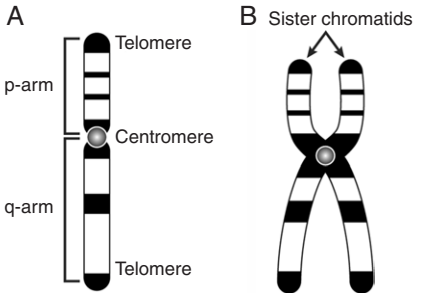

FIGURE 9.1 Schematic representation of a typical human chromosome. In this example, a submetacentric chromosome is shown as it might appear after chemical staining by the Giemsa method. Darkly stained regions are heterochromatic (condensed); lighter regions are euchromatic (uncondensed).

the physical eukaryotic genome. The term "karyotype" refers to a description or depiction (karyogram) of the set of all chromosomes in an organism. It is customary to depict the autosomal (nonsex) chromosomes arranged in homologous pairs in a standard order, usually from largest to smallest, with the shorter arm of each one oriented toward the top of the picture. The sex chromosomes typically are placed last. Sometimes only a haploid chromosome set is depicted. Figure 9.2 shows two examples of eukaryotic karyotypes (from African elephant and Siberian tiger).

Each eukaryotic chromosome is linear with a constriction somewhere along its length called the centromere, and is capped by condensed regions called telomeres. A long DNA double helix molecule stretches from one telomere to the other. Chromosomes consist of a tightly coiled complex of DNA and of proteins such as

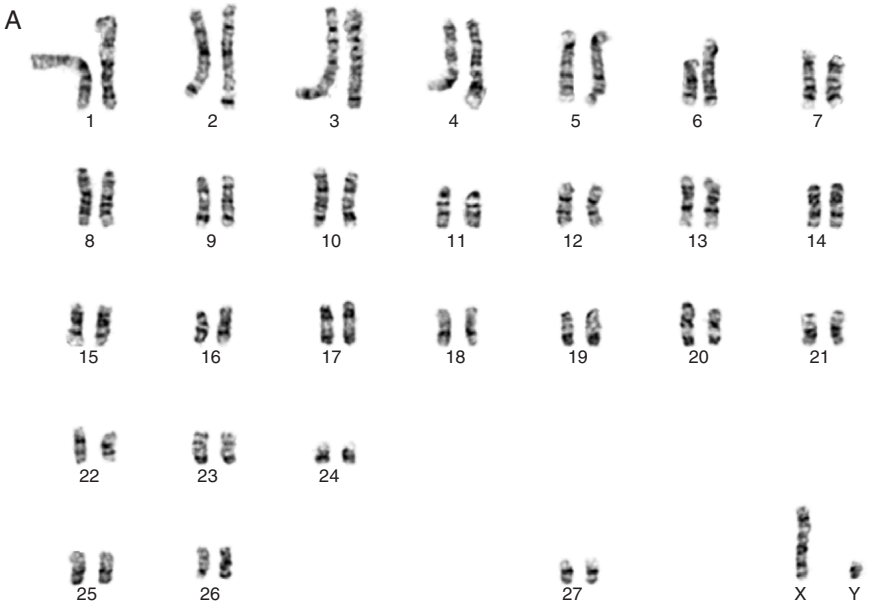

FIGURE 9.2 Representative eukaryote karyotypes from **(A)** the African elephant *Loxodonta africana*, with 2n = 58 (Houck *et al.* 2001), and **(B)** the Siberian tiger *Panthera tigris altaica* (Suedmeyer *et al.* 2003). The normal karyotype for the Siberian tiger is 2n = 38, but this individual has a sex chromosome set of XXY and thus exhibits Klinefelter syndrome. Notice that the chromosome naming and ordering conventions for the two species differ. Felid karyotypes follow the standard established for the common cat in which the chromosomes are labeled with a combination of letters and numbers. The elephant karyotype is arranged with acrocentric/telocentric pairs first, followed by the two metacentric pairs, followed by the pair that distinguishes African from Asian species. Reproduced by kind permission of the Zoological Society of San Diego's Center for Reproduction of Endangered Species Genetics Division and the Kansas City Zoo.

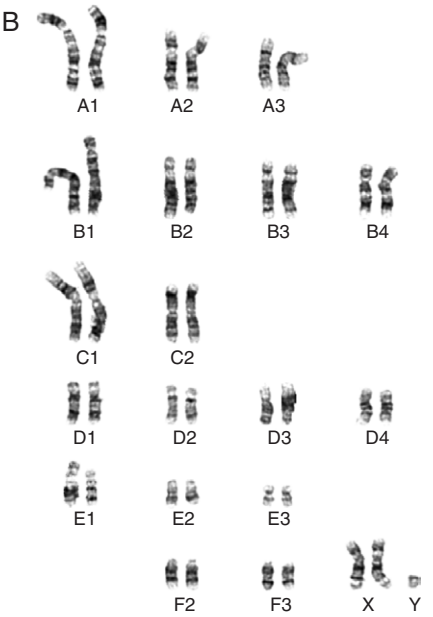

FIGURE 9.2 *(Continued)*.

histones, which means that the DNA molecule, if extended, would be several orders of magnitude longer than the chromosome—in fact, compared to an average mammalian chromosome length of only a few micrometers (μm) during mitosis, the DNA would be a few centimeters long. The DNA, as well as the protein, in and around the centromeres and telomeres has characteristic properties. For example, DNA near the centromere of a human chromosome contains hundreds of thousands of repeats of a characteristic 171-base pair (bp) sequence called an alpha satellite sequence. Telomeres and nearby regions likewise have characteristic repeat sequences. Centromeres and telomeres are heterochromatic (condensed), whereas other regions in the chromosome can be either heterochromatic or euchromatic (uncondensed), as indicated by the banding pattern. Most genes and other single-copy DNA are found in the euchromatic portions.

Karyotyping: The Beginning of Comparative Genomics

Until the 1970s, banding techniques to consistently reveal the fine structure (chromatin patterns) of chromosomes were unavailable, and the only genome

comparisons were based on the number, relative sizes, and shapes of the chromosomes. Even this rough characterization required nontrivial laboratory techniques and generated potentially ambiguous results. For example, the correct chromosome count for humans was not established until 1956 (Ford and Hamerton, 1956; Tjio and Levan, 1956). Even at this level of detail, a great deal of variation among karyotypes of different organisms was apparent. In some organisms, the chromosomes all have the same morphology; for example, all mouse chromosomes are acrocentric (centromere near one end). Other organisms, such as humans, have a mixture of different chromosome morphological types. Chromosomes also vary considerably in size, both within and among genomes. Some chromosomes of fungi and green algae are 1 μm or less in length, whereas some animal and plant chromosomes are more than 30 μm long. Some birds and lizards have a mixture of small and large chromosomes. In terms of numbers of chromosomes, the male of the ant *Myrmecia pilosula* has just one, whereas the fern *Ophioglossum reticulatum* has a diploid chromosome number of 1260. It is remarkable that even closely related and phenotypically similar species such as the Indian muntjac (*Muntiacus muntjak*) and Chinese muntjac (*Muntiacus reevesi*) can differ greatly in chromosome number ("n" represents the haploid number), with 2n = 6 (females) and 2n = 7 (males) for the Indian species as compared to a more typical mammalian value of 2n = 46 for the Chinese species. Even with this extreme karyotypic difference, viable hybrid offspring are known, indicating a high degree of sequence conservation in spite of the radical difference in chromosome number (Levy *et al.,* 1992). Except for frequent evidence of polyploidy in plants (see Chapter 7), there seems to be little phylogenetic pattern or overall trend to chromosome number among major eukaryotic taxa.

Once banding techniques became available in the 1970s (Caspersson *et al.,* 1970; Pardue and Gall, 1970; Seabright, 1971), finer aspects of genomic relationships became visible. Several types of banding can be produced by different dyes and treatments. The most common is G-banding, which produces a characteristic pattern of alternating light and dark regions (note that in plants G-banding does not produce good results). In the human genome up to 850 bands are visible. This method uses trypsin to partially digest the histones of the chromosome prior to staining with the DNA-binding dye called Giemsa, which preferentially stains heterochromatic regions of the chromosome to produce dark bands. The differential staining effect is strongly correlated to locally low G+C content of the DNA, but is not completely explained by the G+C content (Niimura and Gojobori, 2002). Other banding methods in use are R-banding, which gives essentially the reverse of the G-band pattern; Q-banding, which uses fluorescent dye and identifies much the same regions as G-banding; and C-banding, which primarily stains the constitutive heterochromatin of the centromeres.

Following the advent of these powerful techniques for determining chromosomal homology, the genomes of certain groups, perhaps most notably mammals,

began to be studied intensively. With a few exceptions, the content of the mammalian genome is more highly conserved than its karyotype might suggest. For example, although the diploid chromosome number ranges from 2n = 6 or 7 in some varieties of muntjac to 2n = 84 in the black rhino *Diceros bicornis* (Hungerford *et al.,* 1967), haploid genome size varies only from less than 2 billion base pairs (gigabases, Gb) in some bats to more than 8 Gb in one species of (polyploid) rat. Indeed, most groups (such as primates, artiodactyls, marsupials, and monotremes) have much smaller ranges, with sizes close to 3 Gb (Gregory, 2001; Hedges and Kumar, 2002) (see Chapter 1).

With the availability of denser genetic maps and the use of more genomic markers and powerful chromosome painting techniques such as ZOO-FISH (Scherthan *et al.,* 1994; Wienberg and Stanyon, 1995), researchers obtained a more refined picture of relative organization within and among mammalian orders. ZOO-FISH (a modified version of fluorescent *in situ* hybridization [FISH] techniques), for example, is based on interspecies chromosome painting in which DNA from fluorescent-labeled individual chromosomes of one species is hybridized *in situ* to the genome of another species. This method has greatly facilitated the identification of evolutionarily conserved chromosomes, chromosome arms, and segments (Raudsepp *et al.,* 1996; Iannuzzi *et al.,* 1998; Richard *et al.,* 2003).

Comparison of homologous markers in different species showed several common patterns of gross genomic change. Besides duplications and chromosomal fission and fusion, homologous chromosome segments could be identified in different relative locations in different genomes. Chromosomal rearrangement is the transfer of chromosome segments either to other chromosomes (interchromosomal rearrangement) or within a chromosome (intrachromosomal rearrangement). The most common form of interchromosomal rearrangement is the *reciprocal translocation,* in which two chromosomes exchange terminal (end) segments. Other forms of interchromosomal rearrangements are *simple translocation,* in which a terminal segment of one chromosome breaks off and attaches itself to the end of another chromosome, and *intercalary transposition,* in which an internal segment of one chromosome moves to a nonterminal position on another chromosome. Common forms of intrachromosomal rearrangements are *simple transpositions* (a segment moves from one part of a chromosome to another part of the same chromosome) and *in-place inversion* (a genomic segment remains in the same place but its direction is reversed). These are depicted in Figure 9.3.

A comparison of human (2n = 46) and chimpanzee (2n = 48), for example, reveals an almost perfect correspondence between respective pairs of chromosomes, with metacentric human Chromosome 2 being divided into two acrocentric chimp Chromosomes 12 and 13. This homology is clearly apparent even in banding patterns (Fig. 9.4). Comparison to other primates shows that the divided form is likely the ancestral state, indicating that the human–chimpanzee difference arose by a fusion of the telomeres of the two acrocentric precursors in the

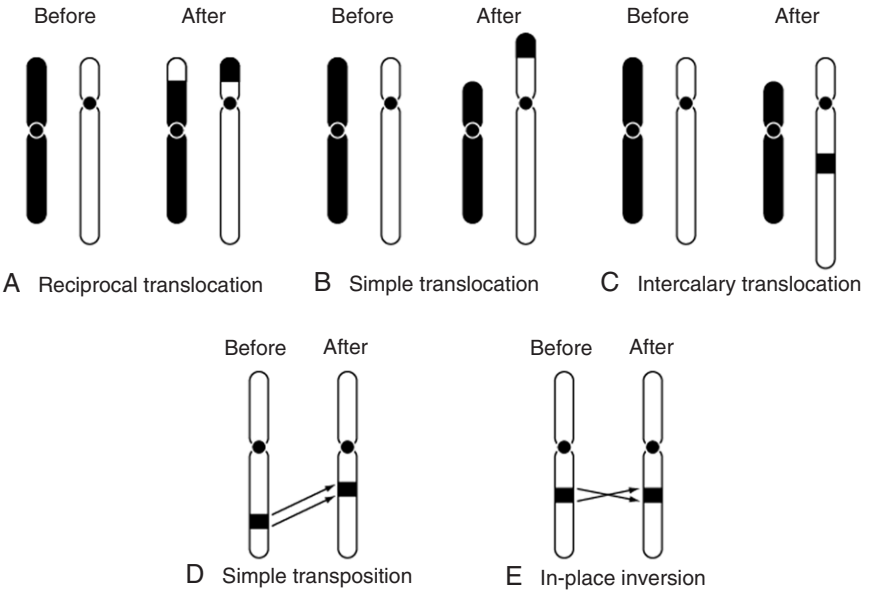

FIGURE 9.3 Diagrams of the most common kinds of chromosomal rearrangements. In each of (**A**), (**B**), and (**C**), an original (nonhomologous) pair of chromosomes is shown on the left and the result after the rearrangement is shown on the right. In (**D**) and (**E**), a single chromosome is shown before and after rearrangement. Reciprocal translocations (**A**) are the most common type of interchromosomal exchanges and involve a swapping of terminal ends between two chromosomes. Simple translocations (**B**) involve the breakage of a terminal end from one chromosome and its fusion to the terminal end of another chromosome, whereas intercalary translocations (**C**) involve the transfer of a part of one chromosome to a nonterminal part of another. Chromosomal segments may also change position within chromosomes, as by simple transposition (**D**), or may stay in the same place but be reversed in direction, as by in-place inversion (**E**).

human line after divergence from chimpanzees (Yunis and Prakash, 1982). Further evidence of this fusion is the presence of remnants of the extra centromere (Avarello *et al.,* 1992) and the extra telomeres (Ijdo *et al.,* 1991). The gibbon (2n = 58) contains many rearrangements with respect to the human genome. For example, the contents of human Chromosome 2 are now dispersed among gibbon Chromosomes 1a, 14, 17, 19, 20, and 22b. If, however, one considers macaques—the immediate outgroup species to the human, chimpanzee, and gibbon clade—it is evident that the chimpanzee pattern that involves separate chromosomal homologs to each arm of human Chromosome 2 again holds. The conclusion is that extensive rearrangements have taken place along the lineage leading to gibbons from their common ancestor with humans and chimps. Similarly, human Chromosome 21 appears to have evolved from two ancestral

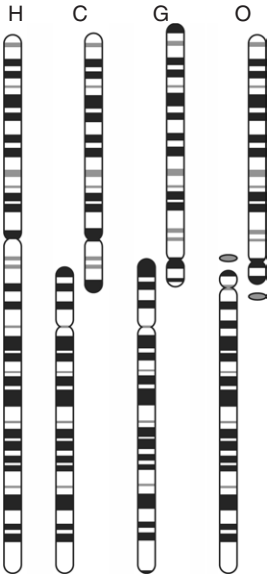

FIGURE 9.4 Schematic representation of evidence for chromosomal evolution. The homologs of the p and q arms of human (H) Chromosome 2 are separate acrocentric chromosomes in other primates (C = chimpanzee, G = gorilla, O = orangutan), indicating a fusion in the human lineage after descent from the most recent common ancestor of human and chimpanzee. Further evidence of this transformation is given by the presence of two inverted arrays of the characteristic vertebrate telomeric repeat in a head-to-head arrangement in the human chromosome at the apparent point of fusion, and the fact that *in situ* hybridization signals the presence of alphoid patterns typical of centromeres at the point in the human chromosome corresponding to the position of the homologous centromere in the other primates. Adapted from Yunis and Prakash (1982), reproduced by permission (© American Association for the Advancement of Science).

blocks present in marsupials and monotremes (Graves, 1996). In this way fissions and fusions can be inferred in a group of closely related lineages. A similar example involves the relation of the nucleoside phosphorylase (NP) gene with the gene complex *PKM2-MP1-HEXA*. In humans, *NP* is on Chromosome 14 and the *PKM2-MP1-HEXA* group lies on Chromosome 15. This separation is preserved in chimpanzees, but the two groups are syntenic (together on the same chromosome) in macaques, rhesus monkeys (Estop *et al.,* 1983), baboons (Thiessen and Lalley, 1986; Thiessen and Lalley, 1987), and pigs (Gellin *et al.,* 1981), indicating general conservation except for a fission in a recent ancestor of chimps and humans (Murphy *et al.,* 2001).

The general pattern of widespread chromosomal conservation with interspersed rapidly evolving lineages is found in many parts of the mammalian phylogeny.

For instance, lemurs (2n = 60), which are basal primates, show an actively evolving genome structure. This time the primary mode is fission, with Chromosome 1 of humans, chimpanzees, and macaques being present as three separate chromosomes (2, 22, and 23) in lemurs. A primatewide comparison of conserved chromosomal regions allows inference of a most parsimonious primate ancestor with 2n = 50 (O'Brien and Stanyon, 1999). The technique is to find large, universally conserved segments and treat these as units that are rearranged through time, while attempting to minimize the number of rearrangements. The reconstruction of chromosome evolution requires about seven major translocation rearrangements to get from the common ancestor of all living primates (60–80 million years ago) (Tavare *et al.,* 2002) to modern humans. This rate of one large-scale rearrangement per 10 million years seems to be a characteristic of the primate lineage. For most extant primate species, fewer than 20 major rearrangements are needed to reconstruct evolution from the common primate ancestor (O'Brien and Stanyon, 1999; Hedges and Kumar, 2003), although many more minor rearrangements are inferred to have taken place (Kumar *et al.,* 2001).

Mouse and human chromosomal homology has been mapped in greater detail owing to the availability of complete genomes. Mice show extensive rearrangements as compared to humans, which are thought to have taken place primarily within the rodent lineage. It has been estimated that chromosomal rearrangements between mouse and rat proceed ten times faster than between far less closely related species such as humans and cats (Stanyon *et al.,* 1999). Indeed, the differences between the cat (2n = 38) and human genomes are not extensive (Nash and O'Brien, 1982; O'Brien and Nash, 1982), and can be accounted for by some 13 translocations and fissions/fusions involving large blocks of genes. This implies a roughly equivalent rate of rearrangements in carnivores as in primates, and indicates that the common mammalian ancestral genome was probably something between that of humans and cats. Dogs (2n = 78) and some bears (Nash *et al.,* 1998) exhibit a somewhat more rapidly evolving genomic architecture with a greater number of karyotypic changes and rearrangements. An ancestral carnivore genome with 2n = 42 has been reconstructed using the same methods discussed previously (Murphy *et al.,* 2001). Among the cetartiodactyls, cows (2n = 60) exhibit a high number of conserved segments with respect to humans, with many of these resulting from intrachromosomal movements such as inversion. Most of the genomic distance (in terms of chromosomal rearrangements) between humans and bovines may be accounted for by 40 to 50 interchromosomal translocations and a similar number of intrachromosomal rearrangements (Band *et al.,* 2000; Jiang *et al.,* 2002).

The banding pattern, morphology (except among ruminants), and gene content of the X chromosome are very highly conserved among eutherian mammals (Chowdhary *et al.,* 1998). A portion called XCR is even identifiable among marsupials as well as eutherians, whereas a more recently added XAR portion has

been created from autosomal material prior to eutherian diversification. A study involving 25 markers revealed complete conservation of order between humans and cats, whereas mice showed seven conserved segments with respect to the others (Murphy *et al.,* 1999). Other studies show more intrachromosomal rearrangement of the X chromosomes of some mammals, however (Nadeau, 1989; Farr and Goodfellow, 1992). Wakefield and Graves (1996) found only one of 42 markers on the human X chromosome with an autosomal homolog in a eutherian (*AMD2* on Chromosome 20 of rat). Thus the X chromosome exhibits the same general pattern in this respect as the autosomes, although at a much slower pace. The evolution of the mammalian Y chromosome is also anomalous in several ways—although homologs of human Y-chromosome genes may sometimes appear on X chromosomes of eutherian mammals, it is very rare to find them in autosomes. But the Y chromosome does tend to have a high degree of activity in terms of both content and organization, especially in primates (Archidiacono *et al.,* 1998; Skaletsky *et al.,* 2003). For these reasons, the mammalian sex chromosomes are usually excluded from generalizations based on the autosomes.

Human Chromosome 17 is conserved as an entire chromosome in chimpanzees, macaques, lemurs, tree shrews, cats, horses, pigs, dolphins, cows, Chinese (but not Indian) muntjacs, and sheep, and as an arm in minks, bats, harbor seals, spectacled bears, and giant pandas. Human Chromosome 20 is conserved as an entire chromosome in chimps, lemurs, horses, and pigs and as an arm in gibbons, macaques, tree shrews, cats, minks, dolphins, bats, spectacled bears, and giant pandas. Murid rodents (mice and rats), as usual, form an exception in which fragmentation prevails, but even there, human Chromosome 20 forms a conserved unit in both rats and mice. Note that the human Chromosome 20 homolog appears in both cows and Indian muntjacs as two segments separated by material from human Chromosome 10, indicating that an inversion took place prior to the divergence of the cervids and bovids. Similar arguments based on distribution of synteny led one group (Chowdhary *et al.,* 1998) to postulate a primordial eutherian karyotype of 2n = 48 consisting of human chromosome segments 1p, 1q, 2pter-q13, 2q13-qter, 3+21, 4, 5, 6, 7, 8, 9, 10, 11, 12+22a, 13, 14+15, 16q+19q, 16p, 17, 18, 19p, 20, 22b, X and Y (where "ter" refers to the terminus of the respective arm, and "a" and "b" refer to portions of Chromosome 22).

The genomes of marsupials and monotremes appear to be more conserved than those of eutherian mammals. Among marsupials (2n ranging from 14 to 22), a primitive genome with 2n = 14 appears to best account for the existing diversity. The rock wallabies, however, have a more actively evolving genome, with some 20 different karyotypes described. Similar activity has been noted in some mouse populations, with six different karyotypic races resulting from multiple Robertsonian (centric) fusions being noted on the island of Madeira from a founding population only 500 years old. Amazingly, some of these races have diploid chromosome numbers as low as the 20s, compared with a more typical

2n = 40 (Britton-Davidian *et al.,* 2000). The monotremes, platypus and echidna, have very similar karyotypes (Graves, 1996), despite having been separated for as long as, or longer than, the major orders of eutherians.

In summary, gene order and synteny on the mammalian genome tends to be, with a few notable exceptions such as murid rodents, rather conserved, even when karyotypic change is rampant. It is difficult to infer an ancestral karyotype because fission and fusion are both fairly frequent, but gene order at a coarse level is probably not very different from what is observed today in humans or cats. Mammals generally display a slow rate of chromosome exchange (one or two major exchanges in 10 million years) punctuated in certain lineages by episodes of radical genome reorganization. The reason for these episodes remains unknown (O'Brien *et al.,* 1999; Kumar *et al.,* 2001).

Chromosome painting results have been used to confirm some phylogenetic hypotheses, such as the close relationship of carnivores with perissodactyls and artiodactyls in the hypothesized superordinal clade ferungulata (O'Brien *et al.,* 1999; Murphy *et al.,* 2001). On the whole, however, there appears to be limited phylogenetic information to be obtained from comparative genomics at the karyotype level. As will be seen, this is not the case with sequence-based comparisons.

Besides chromosome number and structure, the other crude descriptor of genomes is haploid DNA content, or "C-value." By the mid-20th century, techniques had been developed to measure this parameter (see Chapters 1 and 2). It became increasingly clear that (1) genome size varies enormously among species, and (2) except at a very basic level (e.g., prokaryotes versus eukaryotes) there is no correlation between genome size and notions of organismal complexity. For example, genome DNA content is now known to vary over several orders of magnitude among eukaryotes and by a factor of 350 even among vertebrates. On the other hand, some groups of eukaryotes, such as birds, mammals, and teleost fishes, show relatively little variation. The factors correlating with genome content are numerous and diverse and their interactions complex. Despite early hopes that C-value might prove to be a simple characterization of an organism's complexity, it has instead raised many more biological questions than it answered (see Chapters 1 and 2).

Genome Architecture

In the last half of the 20th century, a more detailed, sequence-oriented picture of the overall architecture of the eukaryotic genome began to take shape. In the late 1970s, it was discovered that eukaryotic messenger RNA (mRNA) was shorter than the genomic DNA from which it was transcribed and that sections, then known as intervening sequences, were spliced out in eukaryotes and their viruses after initial transcription (Berget *et al.,* 1977; Chow *et al.,* 1977). It was also found

that a large fraction of eukaryotic DNA was repetitive and did not appear to code for proteins or to have any of the functions known by analogy to prokaryotic DNA (Britten and Kohne, 1968). Masatoshi Nei called such apparently useless DNA "nonsense DNA" (Nei, 1969), whereas Susumu Ohno called it "junk DNA" (Ohno, 1972). One thing was clear: the eukaryotic genome was much messier to deal with than the compact prokaryotic genome. Although a number of theories have since been proposed to explain the presence of noncoding DNA and the interrupted coding sequences, much uncertainty remains about the functional and evolutionary significance of these features (see Chapters 1 and 2). Certain repetitive elements may simply be parasitic or "selfish" DNA (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Scientists have at least now created a taxonomy of the repetitive elements and know something about the means by which they replicate (see Chapter 3). Figure 9.5 shows a breakdown of different DNA types in the human genome; other eukaryotes have similar classes of elements, but in different proportions.

The structure of a typical eukaryotic gene is depicted in Figure 9.6. A segment of DNA that does not contain any stop codons when interpreted according to its implied reading frame (with the first nucleotide being the first position of the first codon) is called an Open Reading Frame (ORF) (Doolittle, 1986). An ORF is not necessarily part of a protein-coding gene, but it may be. A eukaryotic protein-coding gene may contain several noncontiguous ORFs, possibly in different reading frames from each other. The problem of genes being interrupted by so-called "intervening sequences" is now well known as the intron–exon distinction, and in general the mechanisms by which this occurs are known. These characteristics, interrupted coding sequences and large amounts of DNA whose function is unclear, characterize the eukaryotic genome.

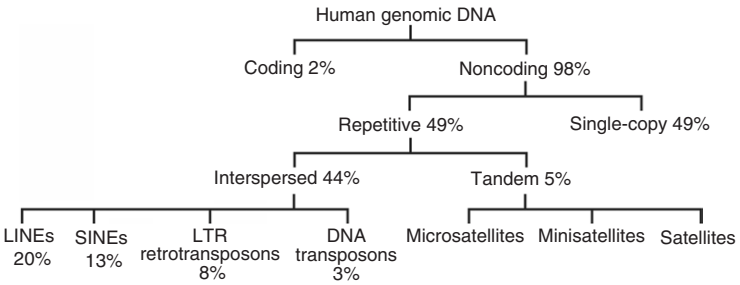

FIGURE 9.5 Composition of the human genome in terms of DNA classes. Percentages are approximate; in particular, exact figures for tandem repeats are not well known because they are most common in difficult-to-sequence constitutive heterochromatic regions (e.g., centromeres). Compositions of other eukaryotes may be radically different. For example, the housefly *Musca domestica* has about 90% single-copy DNA, whereas the toad *Bufo bufo* has only about 20% (John and Miklos, 1988).

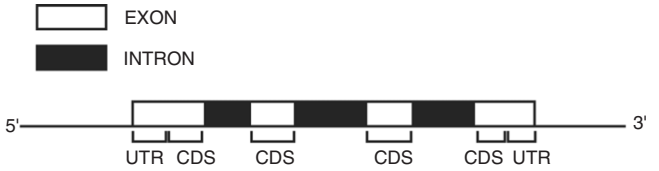

FIGURE 9.6 Architecture of a typical eukaryotic gene. The region from the beginning of the first exon to the end of the last exon is transcribed. Later, the introns are removed during splicing. The Coding DNA Sequences (CDS) are finally translated into a polypeptide. Enhancer and promoter sequences (not indicated) that control transcription may be located near the gene, upstream or downstream, or even within introns. Untranslated regions (UTRs) are not themselves translated into polypeptides, but they control translation in various ways. Lengths and numbers of introns vary widely among different groups of eukaryotes. In humans, the mean intron length is 3400 bp, but the most common (modal) intron length in human genes is around 100 bp. The average number of introns per human gene is about nine.

Another property of the eukaryotic genome is lack of uniformity. Chromosomes, and regions within chromosomes, vary a great deal in almost any parameter one can think of, including gene density, G + C content, and so on; these variations cannot be explained merely by sampling error. Even at a coarse level, the bands revealed by Giemsa staining suggest differences in regions within a chromosome. Bernardi and others have identified five families of regions called *isochores* in the human genome (Bernardi *et al.,* 1985, 1988). Each of these regions is at least 300 kilobases (kb) in length and has a characteristic G + C content. The regions of low G + C content are designated L1 and L2 regions and are gene-poor, whereas regions with high G + C content are designated H1, H2, and H3 and are gene-rich. For example, the H1 regions make up only 3% of the human genome but contain 25% of the genes. It also appears that long genes are less likely to appear in the G + C–rich isochores (but see Duret *et al.,* 1995). The isochores are correlated with G-bands, with the dark G-bands tending to be made up of L1 and L2 isochores, with some contribution from the H1 family (Saccone *et al.,* 1993). Other vertebrates and plants also seem to have an isochore structure. Some controversy, discussed in a later section, arose after the human genome sequence was available about the uniformity of G + C content within isochores, but the concept remains a useful tool for dividing the genome into identifiable regions.

WORKING WITH EUKARYOTIC GENOMES

Today, technology appears to drive the biological sciences as much as hypothesis does (Galison, 1997; Volti, 2001). In contemporary biology, there are already many terabytes of molecular sequence data available and the rate at which it is

accumulating is rapidly accelerating. Fortunately software (including databases and specialized search and analytical tools) has been able to keep up with the data explosion and runs on common, inexpensive hardware in most cases. Turning a eukaryotic genome into grist for this mill usually involves the following steps: mapping, sequencing, and annotation.

Mapping: Genetic and Physical

Mapping involves determining the position of recognizable markers in the genome of interest. The markers may be genes or other sequence features, and the positions may be reckoned crudely in terms of the chromosome or arm on which the marker resides, of its genetic linkage with other markers measured in centiMorgans (linkage map), or of its position on the chromosome specified in terms of base pairs (physical map). Mapping technology began with genetic linkage mapping of *Drosophila melanogaster* and other model organisms in the early 20th century (Sturtevant, 1913), followed by the first rudimentary physical maps two decades later (Bridges, 1935). By the 1970s and 1980s, physical genome mapping underwent great advances. Markers such as Sequence Tagged Sites (STSs) (Olson *et al.,* 1989) based on the Polymerase Chain Reaction (PCR) (Mullis *et al.,* 1986) allowed the accurate mapping of a large number of DNA segments, typically a few hundred base pairs in length, to physical addresses in the genome.

Sequencing: The Holy Grail of Comparative Genomics

It is the ability to read and assemble sequences of nucleotide bases that has enabled the emergence of comparative genomics as it is currently known. Most DNA sequencing has been based on the Sanger method (Sanger *et al.,* 1977). In this technique, the DNA to be sequenced is first denatured into single strands using heat and then a labeled primer sequence is annealed to the strands near the 3′ end of the region of interest. At this point, the solution is divided into four batches corresponding to the four nucleotide bases. Nucleotides and DNA polymerase are added to each batch, and each is additionally given a solution of one of four different kinds of dideoxynucleotides corresponding to the four bases. Dideoxynucleotides are essentially the same as nucleotides except they contain a hydrogen group on the 3′ end instead of a hydroxyl group. These specially modified nucleotides terminate any DNA chain into which they are incorporated because a phosphodiester bond cannot form between the dideoxynucleotide and the next potential nucleotide. DNA synthesis in each batch thus results in a collection of strands of different lengths, each terminating in the same nucleotide.

When these strands are separated by gel electrophoresis, their lengths indicate positions of that nucleotide. Modern automated sequencing methods (Strauss *et al.,* 1986) have streamlined this technique but the same basic principle is used. For example, all four reactions are run together with distinctively labeled dideoxynucleotides so that the sequence can be automatically read from a single lane using gel or capillary methods.

The problem with the basic Sanger chain termination method is that reads are limited to several hundred or at best a few thousand bases, whereas DNA sequences of interest are often far larger. This problem is addressed using the so-called "shotgun method" of obtaining overlapping random sequence reads from the larger sequence and assembling these on the basis of matching overlapping areas into larger contiguous segments called "contigs." This works well except in the presence of low-complexity or repetitive DNA, where matching may not indicate actual overlap. Eukaryotic whole-genome sequencing projects have taken either of two approaches: hierarchical shotgun or whole-genome shotgun. The public National Human Genome Research Institute (NHGRI) Human Genome Project was an example of the former, whereas Celera used the latter approach in its human genome analysis (see later section for more on this). In the hierarchical shotgun approach, the genome is first broken down into a library of cloned regions (e.g., bacterial artificial chromosomes) whose relationship to the entire genome is known through mapping. Each of these clones is then sequenced by the shotgun method, and the resulting sequences are assembled. The generation and mapping of the clone library is a large part of the effort. The whole-genome shotgun method dispenses with the mapped clone library step, and reads are obtained directly from the target genome (see also Chapter 10). This latter method is much more cost-effective, but is more error-prone and requires more sophisticated assembly methods because the maps are not available as a top-down guide.

Other, radically different, sequencing methods are on the horizon. An example is "nanopore sequencing" (Deamer and Branton, 2002). This technique analyzes individual strands of DNA by applying an electric current as they pass through a tiny membrane channel or pore. As charged bases pass through the pores in single file, they block the flow of current in a manner characteristic of the polymer's sequence.

Annotation: Making Biological Sense of the Letters

Once the sequence of a genome has been determined, the job of interpreting the lengthy string of A's, C's, G's, and T's can begin. At a minimum, the goal is to identify the locations of all the functional units such as genes, transposable elements, and regulatory regions in the sequence, and ultimately to determine the functional relationship of these elements to each other and to expression data,

proteins, phenotype, and disease. All of these kinds of information, when attached to sequences, constitute annotation. Generally, annotation is expert-labor intensive, but automated tools for comparing and parsing sequence data are indispensable. All major sequence databases provide record fields for annotating sequence entries.

THE GENESIS OF LARGE-SCALE SEQUENCING PROJECTS FOR EUKARYOTES

SEQUENCING THE HUMAN GENOME: THE MOST AMBITIOUS IDEA

By 1985 the idea of sequencing the human genome began to be discussed. It was an extremely ambitious notion. The first tiny viral genome had been sequenced barely 10 years before, and the completion of even the first prokaryote genome sequence lay a decade in the future (see Chapter 10). Even physical mapping of eukaryotic genomes had been done only for yeast and simple animals. Nevertheless, buoyed by the mounting wave of successful "big science" projects, from the Manhattan project to the Apollo program to the recent Keck telescope, Robert Sinsheimer of the University of California at Santa Cruz and later Walter Gilbert became early proponents of the idea (Gilbert and Bodmer, 1986). It was a controversial as well as a bold idea. Some biologists were appalled by the notion that "assembly line science" might replace the small research group or at least compete with it for funding (Chargaff, 1980). Gilbert countered that the human genome sequence would be the "raw material for the science of the 21st century" (Gruskin and Smith, 1987).

Who would fund the project, even if it were seen as feasible? Private sources did not seem enthusiastic. The most interested agency of the U.S. federal government seemed to be the Department of Energy's (DOE) Office of Health and Environmental Research, which had been studying the genetic consequences of the use of atomic energy. By 1987 the National Institutes of Health (NIH) had also joined the bandwagon and funded a small feasibility study (Roberts, 1987). By 1988 the NIH and DOE had signed a memorandum of cooperation and famed DNA pioneer James Watson was named Associate Director of Human Genome Research at NIH. The project was under way, with high visibility (Goujon, 2001).

Funding for the Human Genome Project (HGP) formally began in 1990. Gilbert estimated that the overall project would cost about $1 per base and would require 15 years, although the cost at the time was closer to $10 per finished base (Collins *et al.,* 2003). The first five-year plan proposed the creation of complete

genetic and physical (STS) maps of the human genome, with sequencing to commence when costs declined to less than $0.50 per base. Ultimate project goals included 100,000 mapped single nucleotide polymorphisms (SNPs, or "snips") as well as the sequencing of 95% of the euchromatic portion of the genome with 99.99% accuracy (Collins *et al.,* 2003). Around the same time, an international coordinating committee, the Human Genome Organization (HUGO), was formed to coordinate international funding and to iron out the anticipated disputes over such issues as intellectual property rights. By 1994, Robert Waterston declared, accurately, that the technology was then available to complete sequencing by 2001, four years ahead of schedule (Boguski, 1995).

In the early 1990s it began to be appreciated that complementary DNA (cDNA) libraries based on expressed sequences (mRNA) would be essential for the discovery and annotation of human genes. To address this, the Expressed Sequence Tag (EST) method was developed (Adams *et al.,* 1991, 1992; Okubo *et al.,* 1992). This allows rapid generation and sequencing of partial mRNA sequences found in a cell. Although this method does not give genomic DNA sequences, it provides an efficient way to characterize expressed genes. Currently the National Center for Biotechnology Information (NCBI) EST Database (www.ncbi.nlm.nih.gov/dbEST) contains nearly 23 million sequences from hundreds of organisms and is a valuable resource for gene prospecting and gene expression studies.

Other organisms were not being neglected. The HGP funded a subsidiary sequencing project for the common gut bacterium *Escherichia coli* as a test bed, and two more eukaryote sequencing projects were also undertaken: the yeast (*Saccharomyces cerevisiae*) genome project (1989) in Europe and the nematode (*Caenorhabditis elegans*) project spearheaded by Sulston and others in the United Kingdom (1990). A sequencing project for thale cress (also called mustard weed), *Arabidopsis thaliana,* was also put forward to the U.S. funding agencies in 1989. All of these were model organisms upon which a great deal of work had already been done, and all were known to have very compact genomes. In a way, the HGP acted as an umbrella to shelter these far more modest projects: if it were indeed feasible to sequence the entire human genome, so the logic went, then surely these more diminutive projects would be relatively simple and would provide a valuable place to develop new techniques.

Private versus Public Efforts

In May 1998, scientist and entrepreneur J. Craig Venter announced plans to form a new private company named Celera that would sequence a large portion of the human genome within three years for a cost of around $300 million, using

whole-genome shotgun methods that were faster and less labor-intensive. The investment was to return a profit by selling access to a database of high-quality well-annotated sequence data. The Celera project ran in parallel to the public NHGRI effort and both announced completion of first drafts at the same time (International Human Genome Sequencing Consortium, 2001; Venter *et al.,* 2001). Some controversy swirled around the manner in which the Celera group released its data, however. Only limited amounts could be freely downloaded without signed nondisclosure agreements. Many felt that that was an unacceptable compromise between scientific openness and commercial interest and that the paper should not have been published in an academic journal. Eric Lander, director of the Whitehead Center for Genome Research and a key figure in the publicly funded International Human Genome Sequencing Consortium, was quoted as saying, "This is the first time in history that a paper reports a scientific result, but tells readers that to see it, they must sign a contract." Further questions were raised about the ability of Celera's whole-genome shotgun to have succeeded at all without building upon the public project's scaffolding (Russo, 2001).

Despite such friction, the overall results of the two projects seemed to be in fairly good accord (Aach *et al.,* 2001), and the two efforts complemented and stimulated each other's progress. From there, Celera went on to involvement in the sequencing of other organisms. Although the private sector has had significant involvement in many sequencing efforts, questions remain about the profitability of the work in the face of publicly funded competition, as well as legal and ethical issues about ownership and restriction of the use of scientific data. It had been argued that, by its nature, the human genome sequence belongs to humanity as a whole (Macer, 1991). In January 2002, Venter left Celera to undertake other projects, including genetic engineering of life forms (see Chapter 10). In April 2002, Venter revealed that the genome sequenced by Celera was not that of a randomly chosen subject, but of Venter himself. One of Venter's announced projects is writing a book analyzing his own genome (Wade, 2002).

So far almost all sequencing projects have operated under the guidelines that sequence data, if not annotations and analyses, were to be released to the public with minimal delay. In 1996 this was formalized as the so-called "Bermuda Principles" (named after the location of a meeting convened by the Wellcome trust), which call for automatic, rapid (within 24 hours) release of sequence assemblies to the public domain and which discourage the patenting of genes by sequencing labs (Collins *et al.,* 2003).

The following sections describe whole-genome sequencing efforts, in roughly chronological order, amended somewhat to discuss related projects or organisms together. In each case, the key events in the project are described and the findings discussed in the context of previously sequenced organisms. Table 9.1 lists a few properties of completely sequenced eukaryotic genomes, and Figure 9.7 shows the progress of some major sequencing projects.

TABLE 9.1 Some basic data about eukaryotic organisms that have been fully sequenced (as of spring 2003). In most cases, data were obtained from the publication announcing completion of the initial release of the respective sequencing project. Chr (n) refers to the haploid chromosome number of the organism.

Species	Common name	Taxon	C-value (Mb)	Chr (n)	Genes	G + C	Exons/ gene	Year
Anopheles gambiae	Mosquito	Insect	278	3	13,700	0.35	3.7	2002
Arabidopsis thaliana	Thale cress	Dicot plant	157	5	25,500	0.35	5.2	2000
Caenorhabditis elegans	Roundworm	Nematode	100	6	19,820	0.36	6	1998
Ciona intestinalis	Sea squirt	Chordate	160	14	16,000	0.35	6.8	2002
Drosophila melanogaster	Fruit fly	Insect	180	4	13,600	0.41	4	2000
Encephalitozoon cuniculi	Parasitic microsporidian	Protist	2.9	11	2000	0.50	1.01	2001
Takifugu rubripes	Pufferfish	Fish	400	22	35,000	0.48	9	2002
Homo sapiens	Human	Mammal	3400	23	35,000	0.41	9	2001
Magnaporthe grisea	Rice blast fungus	Fungus	40	7	11,000	0.52	3	2002
Mus musculus	Mouse	Mammal	3250	20	35,000	0.42	9	2002
Neurospora crassa	Bread mold	Fungus	40	7	10,000	0.50	2.7	2003
Oryza sativa	Rice	Monocot plant	490	12	50,000	0.43	5.1	2002
Plasmodium falciparum	Human malaria pathogen	Protist	23	14	5300	0.19	2.4	2002
Plasmodium yoelii	Rodent malaria pathogen	Protist	23	14	5900	0.23	2	2002
Rattus norvegicus	Brown rat	Mammal	3100	21	35,000	0.42	9	2004
Saccharomyces cerevisiae	Yeast	Fungus	12.5	16	6128	0.38	1.04	1996
Schizosaccaromyces pombe	Fission yeast	Fungus	14	3	4824	0.36	2	2002

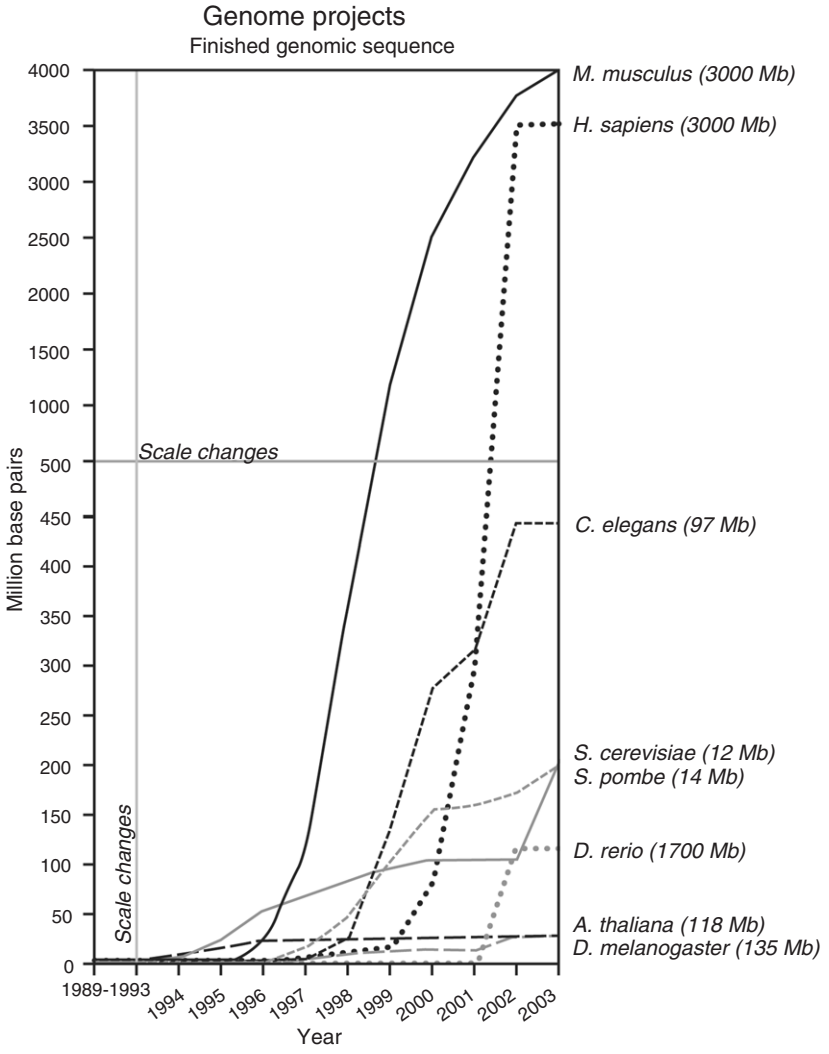

FIGURE 9.7 Cumulative sequencing progress in millions of base pairs for several eukaryotic organisms. Adapted from the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) Genome Monitoring Table at www2.ebi.ac.uk/genomes/mot/.

GENOME SEQUENCING IN FUNGI

SACCHAROMYCES CEREVISIAE: THE FIRST EUKARYOTE TO BE SEQUENCED

As discussed in the previous section, the *S. cerevisiae* project was undertaken around the same time as, and was inspired by, the HGP discussions. At an estimated length of 13 million base pairs (megabases, Mb), the yeast genome is nearly an order of magnitude smaller than other prominent model genomes (e.g., *C. elegans* at 100 Mb and *Arabidopsis* at 157 Mb) (Bennett *et al.,* 2003). Also, a great deal was known about this yeast. For millennia, it has been used in food production for its ability to ferment glucose to ethanol and carbon dioxide. It has been a model organism since the 1970s, and the knowledge of its biochemistry and genetics is highly advanced, in part owing to the early economic importance of the organism to bakers, brewers, and vintners. Moreover, a clone library had already been constructed in the late 1980s (Link and Olson, 1991). The *S. cerevisiae* sequencing project was funded by a mixture of public and private sources primarily centered in Europe and driven by André Goffeau of the Université Catholique de Louvain in Belgium. The philosophy was to supply expertise and coordination among many established laboratories instead of constructing massive new sequencing centers. In 1989, 35 laboratories formed a consortium for the purpose.

In 1992 the complete sequence of Chromosome 3 was published in *Nature* (Oliver *et al.,* 1992) and completion of the entire genome sequence was announced in April 1996 (Goffeau *et al.,* 1996). Soon thereafter, all putative genes (at least those longer than 100 codons) were identified, with the result that for the first time the list of approximately 6000 genes necessary for the functioning of a complete, free-living, eukaryotic organism was known. Most of the yeast genome, about 72%, consists of open reading frames (ORFs), in contrast to the human genome where the figure is less than 2%. Introns tend to be short and almost always located near the 5′ end of the gene; sometimes they occur just after, or even within, the ATG initiation codon. Intron–exon splice sequences are highly conserved. This fortuitous combination of features confirmed the good choice of *S. cerevisiae* as the "practice" model organism with which to begin before being forced to develop more sophisticated methods of exon prediction. Protein-coding genes seem to be randomly oriented on both strands. G+C content varies at many scales, with higher gene density in the broad peaks and a general G+C deficit in subtelomeric and pericentromeric regions. Because so much of the yeast genome codes for proteins, repeats are correspondingly rarer. *Ty* elements (a kind of LTR retrotransposon) account for less than 3% of the genome, whereas short tandem repeats inhabit small regions around the centromeres and telomeres. Thus the *S. cerevisiae* genome has all the major elements of the larger eukaryotic genomes but in different proportions (Dujon, 1996).

The yeast genome provided an opportunity to face the challenge of functional analysis on a small scale. Perhaps the most striking (if not downright surprising) finding of the yeast genome sequencing effort was that at least a third of the putative genes, as identified from ORFs, had no clear-cut previously known homologs. These orphan genes had escaped the notice of traditional genetic methods, indicating that the state of ignorance of genetics was far greater than was realized. Furthermore, alteration of many of these genes had no apparent effect on the phenotype. A systematic program, dubbed EUROFAN, was begun to test each gene by knocking it out, or disrupting its expression. It appeared that about $^2/_3$ of the disrupted genes on Chromosome 3 led to no obvious difference in phenotype (Goujon, 2001). Something about the function of some genes could be inferred by structural clues, such as the presence of transmembrane helices, but clearly significant advancements were needed to elucidate the function of all known genes.

Other Fungal Sequencing Projects

Other relatives of *S. cerevisiae* were also sequenced with a view to comparison. The fungi are a large and diverse kingdom, with at least 100,000 species, including mushrooms, yeasts, and molds (Hawksworth, 1991). Sequencing was begun on another important fungus, the fission yeast *Schizosaccharomyces pombe,* in 1995 and completed in 2002 (Wood *et al.,* 2002), making it the sixth eukaryote to be sequenced. This model organism, first isolated from pombe (an East African beer), is only distantly related to *S. cerevisiae*. Its genome is about 10% larger, but contains ~20% fewer genes and only three chromosomes as opposed to 16 for *S. cerevisiae*. *S. pombe* has on average one intron per gene, much more than *S. cerevisiae*. A pilot deletion project was conducted to determine the apparently essential genes in this organism (Decottignies *et al.,* 2003). The evidence suggests that about 18% of the *S. pombe* genes are in this essential category and that the more phylogenetically widespread a gene is the more likely it is to be required, such that many of the essential eukaryotic genes appeared with the first eukaryotic cell some two billion years ago and have remained strongly conserved.

Genome sequencing work among the fungi has continued and in 2003 the first genome of a filamentous fungus, the intensively studied model mold *Neurospora crassa,* was announced (Galagan *et al.,* 2003) and preliminarily annotated (Mannhaupt *et al.,* 2003). This project has been an important advance in spanning the range of fungal genomes, because the two previously sequenced fungi, *S. cerevisiae* and *S. pombe,* have restricted metabolic and developmental capabilities owing to their specific environmental niches and therefore do not provide a general paradigm for the fungi (Bennett, 1997). An interesting feature of the *N. crassa* genome is the anomalously low fraction of its genes that are members of multigene families (Fig. 9.8). This is thought to be a result of a process called

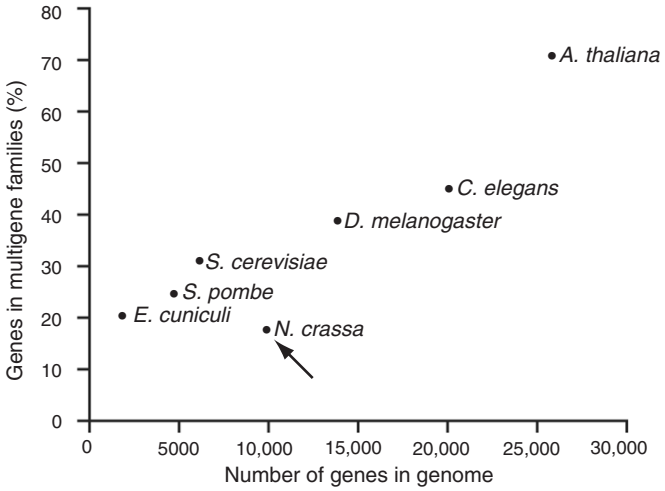

FIGURE 9.8 The proportion of genes in multigene families as a function of the number of genes in the genomes of selected sequenced eukaryotic organisms. The arrow indicates the mold *Neurospora crassa*, which has an especially low proportion of genes in multigene families. Adapted from Galagan *et al.* (2003), reproduced by permission (© Nature Publishing Group).

repeat-induced-point mutation (RIP), which is apparently unique to fungi and mutates and epigenetically silences repetitive DNA. It has been suggested that RIP may be a defense against selfish or mobile DNA (Selker, 1990) (see Chapter 3). As a consequence, *N. crassa* has very little repetitive DNA except for short or highly diverged segments (Krumlauf and Marzluf, 1980). Surprisingly, of the 10,000 predicted protein products of this genome, 41% have no significant matches to known proteins. The fact that new proteins are still being discovered at this rate suggests that the universal proteome is large indeed (Hynes, 2003). It is also interesting that for many *N. crassa* genes, the only known homologs are in prokaryotes (Mannhaupt *et al.,* 2003).

In 2003, three close relatives of *S. cerevisiae* were sequenced (Kellis *et al.,* 2003). These organisms, *Saccharomyces paradoxus, S. bayanus,* and *S. mikatae,* were selected not so much for their intrinsic interest (although two of the three are used in winemaking) but because of their evolutionary similarity to *S. cerevisiae*. In such a case, the whole-genome shotgun method becomes even more efficient because the known genome can be used to help assemble the newly sequenced ones. The real benefit, however, is comparative. When several similar sequences are available, it becomes possible to identify regions that are more conserved than would be expected by chance. Such regions are candidates for small genes or hitherto unknown regulatory regions. Conversely, putative functional genes that cannot be found in the close relatives are suspected to actually be nonfunctional.

Such considerations led to a revision of the gene count for *S. cerevisiae* by an addition of 43 small genes and a deletion of about 500 putative nonconserved genes. Also, many known and newly discovered regulatory motifs were identified. This kind of near-neighbor comparative genomics will be essential for fully parsing any genome, including that of humans.

CAENORHABDITIS ELEGANS AND *DROSOPHILA MELANOGASTER*: THE FIRST ANIMAL GENOMES TO BE SEQUENCED

The Worm Project

The *C. elegans* sequencing project was initiated by two groups: John Sulston and Alan Coulson at the Medical Research Council (MRC) Laboratory of Molecular Biology in Cambridge in the United Kingdom, and Robert Waterston at the Washington University School of Medicine in St. Louis in the United States. The genome of this organism is much larger than that of *S. cerevisiae* (100 Mb as opposed to 13 Mb) and contains the genes needed for development of a multicellular animal, in addition to genes for the housekeeping functions common to all eukaryotic cells. An advantage of using a nematode as a model organism, other than its simplicity and the knowledge of all developmental cell lineages (Sulston *et al.,* 1983), is that nematodes are thought to have diverged early among the animals, so that genes with homologs in both *C. elegans* and another animal are likely to be ancestral to the entire kingdom.

Results of the *C. elegans* sequencing project were reported in 1998 (*C. elegans* Sequencing Consortium, 1998). This provided the first opportunity for genome comparison between two species from different eukaryotic kingdoms: the metazoan *C. elegans* and the unicellular yeast *S. cerevisiae*. Again, the *C. elegans* genome is about eight times as large as the yeast genome and contains about three times as many genes (20,000 vs. 6000). It was anticipated that the two organisms would contain a common core of genes associated with basic eukaryotic cell maintenance, and this was indeed the case. About 20% of *C. elegans* proteins and 40% of yeast proteins had a very similar homolog (BLASTp P-value $< 10^{-10}$; see www.ncbi.nlm.nih.gov/blast) in the other organism (Chervitz *et al.,* 1998). Figure 9.9 shows the distribution of functions of genes that had identifiable homologs in both species. In most cases, the classification was obtained from yeast annotations. The numbers in each section represent the ratio of worm to yeast genes in each category. These data support the idea that core eukaryotic cellular functions are performed by a highly conserved group of genes without many paralogs, even in larger genomes. It is encouraging for the process of understanding genomes that annotations of core functions seem to be transferable between disparate organisms.

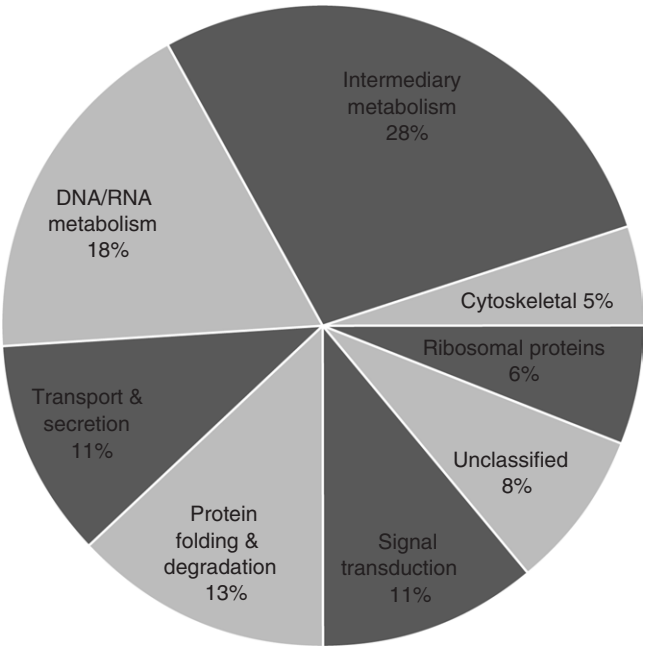

FIGURE 9.9 Distribution of core biological function genes conserved in both yeast (*Saccharomyces cerevisiae*) and nematode worm (*Caenorhabditis elegans*). Yeast and worm protein sequences were clustered into closely related groups. Each sequence group (including groups with two or more sequences) was assigned into a single functional category, relying primarily on the functional annotations for the yeast genes when available. The unclassified category contains groups of sequences without annotation. The number (in %) within each category reflects the ratio of worm to yeast proteins for that category. Adapted from Chervitz *et al*. (1998), reproduced by permission (© American Association for the Advancement of Science).

The complementary aspect of the comparison is to look for genes or families of genes that do not appear to have homologs in both organisms. Although there are some regulatory and signal transduction domains in *S. cerevisiae* genes that do not appear in *C. elegans,* it is primarily the other way around, as expected. Gene types present in *C. elegans* and not in *S. cerevisiae* include those involved in extracellular signaling and adhesion, such as epidermal growth factor (EGF) and factors involved in programmed cell death. In general, comparisons between *C. elegans* and *S. cerevisiae* are consistent with the understanding that common core eukaryotic functions are preserved whereas most disparities can be ascribed to obvious differences in organismal requirements, in this case multicellularity.

A surprising feature of the *C. elegans* genome is that about 15% of its genes are grouped into operons containing from two to eight genes each (Spieth *et al.,* 1993).

In the manner of prokaryotes, all genes in an operon are controlled by a single promoter and produce a single pre-mRNA transcript. However, in contrast to the situation in prokaryotes, genes from a single operon need not be functionally related and are translated separately. They are apparently not ancestrally related to prokaryote operons, but are evolutionarily conserved, as most also appear in *Caenorhabditis briggsae,* a species that last shared a common ancestor with *C. elegans* around 50–100 million years ago. Operon organization in *C. elegans* is thought to be facilitated by a type of mRNA splicing called "trans-splicing" (Nilsen, 1989). Trans-splicing is known to occur in other animals, such as flatworms and hydra, but the extent of occurrence of operons within or beyond the nematodes is not yet known (Blumenthal and Gleason, 2003).

The Fruit Fly Project

Despite its early lead as a genetic model organism and the wealth of genetic mapping and functional studies that had been done on it, *D. melanogaster* was not the first choice for an animal to be sequenced. There are a number of reasons for this. For one thing, a large part of the *D. melanogaster* genome is heterochromatic, making it more difficult to map.

A large portion of the *D. melanogaster* genome was sequenced in 2000 by a consortium of private and public research groups lead by Celera Genomics (Adams *et al.,* 2000). The project was notable in that it was the first eukaryote project to use the whole-genome shotgun sequencing method that Celera would later use on the human genome. The initial draft sequence contained many gaps and regions of low sequence quality, but these deficiencies were rectified by the third release (Celniker *et al.,* 2002). Although only about 120 Mb of the 180 Mb genome was sequenced (i.e., the euchromatic portion), it is thought that this includes the vast majority of the protein-coding genes. This result was built upon by a sequencing project for the euchromatic portion of the sister species *D. pseudoobscura,* using a comparative sequence approach. The *Drosophila* genome will be discussed in more detail in relevant later sections.

THE HUMAN GENOME PROJECT

The "completion" of the human genome sequencing project was announced in February 2001 by simultaneous publication of special issues of the journals *Nature* and *Science* describing results of the publicly and privately funded efforts, respectively (International Human Genome Sequencing Consortium, 2001; Venter *et al.,* 2001). This accomplishment surely ranks with the moon landing as a major achievement of the "big science" paradigm. Although the precise point

chosen as the completion date was, unlike setting foot on the moon, somewhat arbitrary, at the time of the big announcement more than 90% of the genome was sequenced, the remainder being mostly highly repetitive heterochromatic DNA, which is difficult to sequence and thought not to be very informative. At that time, draft sequences were publicly available, and, although the job of annotation had hardly begun (some 40% of open reading frames were of unknown function), enough was known to paint a broad picture of the human genomic landscape and to compare it to the only truly completely sequenced animal at the time, *C. elegans*.

Annotation of the human genome was greatly facilitated by homology searches for genes known to exist in other vertebrates such as the pufferfishes (genera *Tetraodon* and *Takifugu* [Fugu]) (Brenner *et al.,* 1993; Roest Crollius *et al.,* 2000a,b, 2002). Although the Fugu sequencing project was not officially completed until October of 2001, extensive clone libraries were available.

Even without annotation, however, many simple statistical tests could be performed. One controversial question relates to variation in G+C content and its correlation with other local properties of the genome such as gene density. Bernardi and coworkers (Bernardi *et al.,* 1985; Bernardi, 1995) had postulated that the genome is a mosaic of five different types of compositionally homogeneous regions known as isochores. More recent examination showed that most of the G+C content variance among small (20 kb) regions can be explained by the average G+C content of larger (300 kb) windows that contain them (International Human Genome Sequencing Consortium, 2001), meaning that the hypothesis of strict homogeneity among the small regions was not supportable, and leading to the conclusion that isochores are not as strict or as homogeneous as some expected. On the other hand, further analyses showed that for other choices of region size, the hypothesis of homogeneity may not be rejected (Li, 2002). In any case, regions of distinctive composition certainly exist in the human genome, and Bernardi (2001) added that the original description of isochores did not specify strict statistical homogeneity.

Another interesting feature of the human genomic landscape is the density of CpG islands (Bird, 1986). The notation "CpG" refers to a guanine nucleotide immediately following a cytosine in a DNA strand: 5′ . . . CG . . . 3′. (The "p" in CpG refers to the phosphodiester bond that connects adjacent nucleotides in a strand as distinguished from the hydrogen bonds between the C and G in complementary strands.) A CpG island is a DNA region, usually a few hundred nucleotides in length, with a higher-than-usual G+C content and much higher density of the usually underrepresented CpG dinucleotide. About 30,000 CpG islands were detected in the human genome and it was noted that CpG island density correlates with gene density (International Human Genome Sequencing Consortium, 2001). Other studies showed that about half of human and mouse genes are associated with upstream CpG islands, so the feature becomes important for gene detection (Antequera and Bird, 1993).

A strikingly large portion of the human genome consists of transposable elements. In humans, identified transposable elements make up about 44% of the genome, as compared to about 7% for *C. elegans,* 10 to 22% for *D. melanogaster,* and 14% for *A. thaliana* (see Chapter 3). Short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), LTR retrotransposons, and DNA transposon copies make up approximately 13%, 20%, 8%, and 3% of the sequence, respectively, in the human genome (see Chapters 1 and 3). More interesting is the comparative age distribution of these elements. Figure 9.10 shows the distribution of ages of these elements for both mice and humans, revealing a marked decline in all transposon activity for the human lineage, going back at least as far as the eutherian radiation, to the extent that only *Alu* and *LINE1* elements show any recent activity; interestingly, mice show no similar decline in TE activity (International Human Genome Sequencing Consortium, 2001).

Figure 9.11 shows a similar comparison extended to cover humans, *D. melanogaster, C. elegans,* and *A. thaliana*. The interspersed repeats in the human

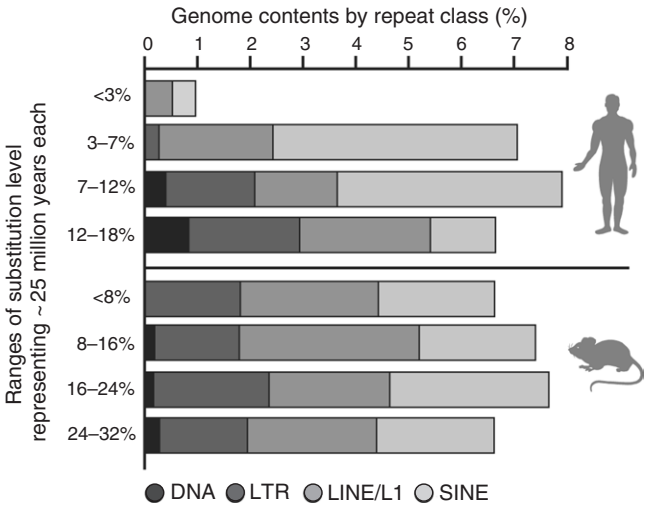

FIGURE 9.10 Age distribution of interspersed repeats in the human and mouse genomes. Bases covered by interspersed repeats were sorted by their divergence from their consensus sequence (which approximates the repeat's original sequence at the time of insertion). CpG dinucleotides in the consensus were excluded from the substitution level calculations because the CT transition rate in CpG pairs is about 10-fold higher than other transitions and causes distortions in comparing transposable elements with high and low CpG content. The data are grouped into bins representing roughly equal time periods of 25 million years. There is a different correspondence between substitution levels and time periods owing to different rates of nucleotide substitution in the two species. Adapted from the International Human Genome Sequencing Consortium (2001), reproduced by permission (© Nature Publishing Group).

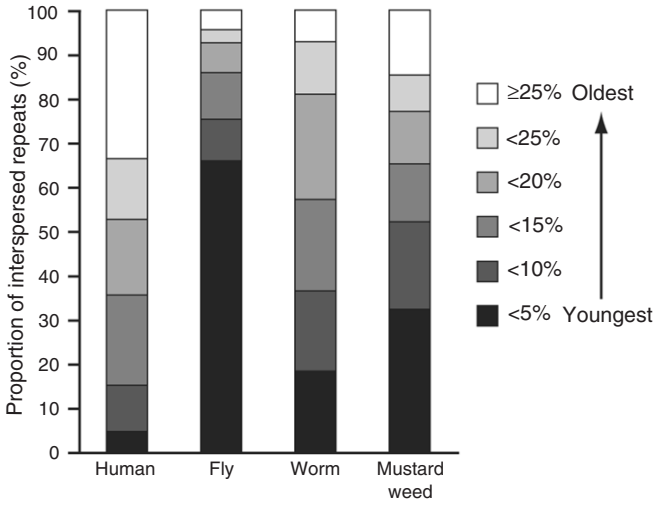

FIGURE 9.11 Comparison of the age of interspersed repeats in four eukaryotic genomes. The copies of repeats were pooled by their nucleotide substitution level from the consensus. Adapted from the International Human Genome Sequencing Consortium (2001), reproduced by permission (© Nature Publishing Group).

genome tend to be older than those of *C. elegans* and *A. thaliana,* and much older than those of *D. melanogaster*. One reason for this may be the increased rate of genome "cleaning" in flies owing to small deletions (Petrov *et al.,* 1996). In humans, most interspersed repeats belong to one of two families, *Alu* and *LINE1*. In *D. melanogaster, C. elegans,* and *A. thaliana,* on the other hand, there are no such dominant families, but a greater diversity of repeat types. This may be because of the much higher fraction of the shorter-lived DNA transposons in *D. melanogaster, A. thaliana,* and *C. elegans* (25%, 49%, and 87%, respectively) than in the human genome. This relative lack of DNA transposons in humans (and probably mammals as a whole) may be related to the improved immune system of this group.

One of the simplest-sounding questions to ask about the human genome is "how many genes does it encode?" It turns out that this is still a very difficult question to answer, even with the complete sequence in hand. In contrast to bacteria, for which precise gene counts can usually be determined (see Chapter 10), there is still extensive uncertainty about the exact number of protein-coding genes in the genomes of humans and other mammals. The reason is that, in the human genome, less than 2% of the DNA codes for proteins (Fig. 9.5). Small exons (see Fig. 9.6 and Table 9.2) are easily lost in the noncoding material. Conversely, some of the nongenic DNA has features in common with protein-coding genes. About 0.5% of the human genome, for example, is thought to consist of pseudogenes,

TABLE 9.2 Some characteristics of human genes. The sample was obtained by aligning genomic DNA from the Human Genome Project with a curated set of full-length mRNA sequences called RefSeq (Pruitt and Maglott, 2001). Some lengths may be underestimated, particularly for the untranslated regions (UTR), because of the currently poor ability in detecting these. Adapted from the International Human Genome Sequencing Consortium (2001), reproduced by permission (© Nature Publishing Group).

	Median	Mean	Sample size
Internal exon size	122 bp	145 bp	43,317 exons
Number of exons	7	8.8	3501 genes
Intron size	1023 bp	3365 bp	27,238 introns
3′ UTR	400 bp	770 bp	689
5′ UTR	240 bp	300 bp	463
Coding sequence	1100 bp (367 aa)	1340 bp (447 aa)	1804
Genomic extent	14 kb	27 kb	1804

remnants of genes that are no longer functional. One study suggests that 1/5 of *C. elegans* annotated genes may in fact be pseudogenes (Mounsey *et al.,* 2002). Thus any count may be either too high or too low and it is difficult even to establish a tight upper or lower bound. Protein-coding genes were identified in the human genome sequence by comparisons with expression libraries, genes known from other organisms, and the use of gene-finding programs to detect ORFs. To the surprise of most experts, the best informed estimates of the total number of protein-coding genes in the human genome settled in the neighborhood of 30,000 to 35,000, much lower than most previous estimates, many of which favored a figure at least twice as large and which ranged up to 140,000.[1]

The exact number of genes in the human genome is still unknown because it remains possible that some small genes have been missed by gene-finding programs and/or that some identified genes are really only pseudogenes. That said, a total of 30,000 to 35,000 currently seems to be a fair number. This means that the human genome, although about 30 times as large as that of *C. elegans,* contains only around twice as many genes. Part of the explanation for this disparity is that human genes are much more extended by introns. Although the most common intron length in humans (around 90 bp) is only around twice

[1]A wager conducted between 2000 and 2003 among participants at the Cold Spring Harbor Genome meetings yielded several hundred guesses ranging from 25,947 to over 150,000 with a mean of around 60,000. All submitted guesses were higher than the official provisional total of 24,847 based on the Ensembl database, so the lowest entry was declared the winner on 30 May, 2003 (http://www.ensembl.org/Genesweep/). See the June 2000 editorial in *Nature Genetics*, "The Nature of the Number" (vol. 28, p.127–128) and Pearson (2003) for details.

as great as the corresponding figure for *C. elegans,* the mean value for humans (3300 bp) is more than 10 times the mean length for *C. elegans,* indicating that some human genes are very extended indeed. Typically, these sprawling genes are found in G + C-poor regions of the human genome.

Another issue raised by this low gene count relates to complexity. Humans have only two or three times as many genes as *D. melanogaster* or *C. elegans*—are not humans more than two or three times more complex? In reference to the old "C-value paradox" (Thomas, 1971), which expressed similar concern about raw genome size, Hahn and Wray (2002) called this the "G-value paradox" and Claverie (2001) dubbed it the "N-value paradox." Although "paradox" is perhaps too strong a word to express a subjective discomfort of this sort (see Chapter 1), mammals are known to be quite complex in some areas, such as the immune system, number of cell types, nervous system, and so on. One solution to this apparent discrepancy lies in the use of alternative splicing and alternative polyadenylation (Edwalds-Gilbert *et al.,* 1997). Alternative splicing allows a single form of pre-mRNA transcript to be spliced into a number of different forms by skipping exons or by recognizing alternative splice sites (see Fig. 9.12). The old idea of "one gene, one protein" is long dead, but the extent to which a gene can produce different products is not easy to estimate. Early methods based on EST alignments suggested that at least 35% of human genes may be involved in alternative splicing (Hanke *et al.,* 1999; Mironov *et al.,* 1999; Brett *et al.,* 2000). Refined estimates based on the complete sequence of several human chromosomes put the fraction at closer to 60%, with an average of at least two or three transcripts per gene. This is much higher than estimates for *C. elegans* of around 22% alternatively spliced, with an average of less than two transcripts per gene. Thus the human transcriptome may be several times larger, in comparison to invertebrates, than the gene count would suggest. It would seem that the initially high estimates of gene number arose, at least in part, by a failure to appreciate this. The interaction of genes through chains of transcriptional regulation may also allow great complexity

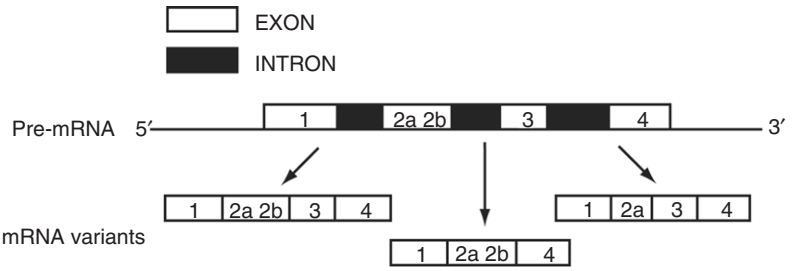

FIGURE 9.12 An illustration of alternative splicing. The same transcribed pre-mRNA strand can be spliced into several variant mature mRNAs. Exons can be included or excluded as units, or alternative splice sites can be used within a single exon (exon 2 in this case).

to arise from a limited number of basic forms (Huang *et al.,* 1999; Fickett and Wasserman, 2000). About 10% of human genes code for transcription factors (proteins that bind to DNA and affect how it is transcribed) whereas only about 5% of yeast genes do. Such a difference, coupled with a more complex network of transcription enhancers and promoters, can result in a much larger set of gene expression patterns leading to a nonlinear increase in organismal complexity (Levine and Tjian, 2003). Unfortunately, the identification and understanding of transcriptional control regions in the human genome lags behind the ability to identify ORFs. In short, the resolution of the G- or N-value paradox may be simply that the relation of gene number to complexity should have not been expected to be linear. Of course, the same was true with genome size and the C-value paradox (see Chapter 1).

Some have argued, both before and after the announcement of the estimates based on the human draft sequence, that there are fundamental limits on the number of genes. George (2002) suggested that the number of genes is limited in organisms with an adaptive immune system by the burden of self-recognition. Pal and Hurst (2000) argued that increase in gene number may be limited by increasing probability of error, both heritable (accumulation of deleterious mutations) and especially nonheritable (e.g., regulatory failure). Another important aspect of comparative genomics is the identification of new genes in humans (taken as a representative vertebrate) that do not have homologs in other sequenced species. It appears that less than 10% of the proteome is in this category, which includes immune and nervous system proteins. Figure 9.13 shows a distribution of where homologs to human genes have been found.

Genome Variation in Human Populations

Another aspect of comparative genomics relates to sequence differences within a species or population. The most common variation of this kind is the single nucleotide polymorphism (SNP), defined as occurring when different nucleotide bases (single nucleotide alleles) appear at a homologous site in a population. Usually, a less frequent allele must occur at an arbitrarily specified frequency, say 1% of the population, to qualify a site as polymorphic, but disease-causing alleles are obviously also of interest at much lower levels of frequency. With the initial draft of the human genome sequence it became possible to assess SNP distribution in a comprehensive manner. Data of this kind are important not only for studies in population genetics and the history of the human species, but also for medical applications, as many known SNPs are associated with heritable diseases (Taylor *et al.,* 2001). In an initial analysis of 1.42 million SNPs, mostly collected by the Human Genome Project and a nonprofit consortium called TSC ("The SNP Consortium"), it was found that two homologous chromosomes randomly selected from the population can be expected to differ in one site out of 1331

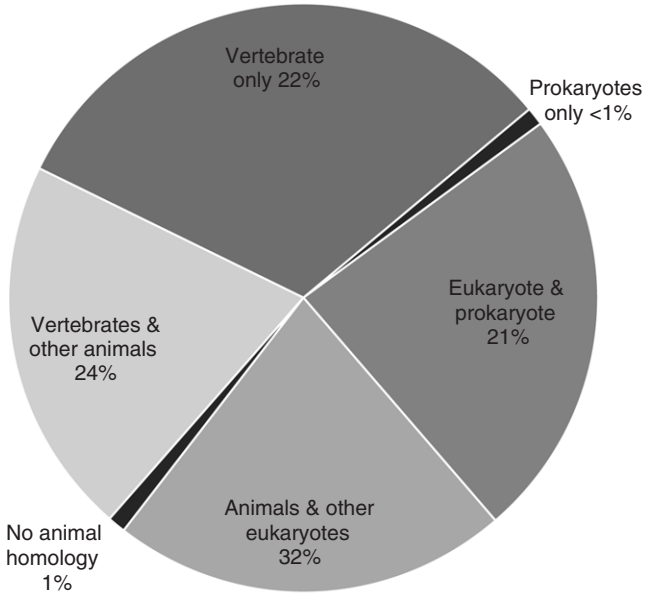

FIGURE 9.13 Distribution of the homologs of the predicted human proteins. For each protein, a homolog to a phylogenetic lineage was considered present if a search of the NCBI nonredundant protein sequence database, using the gapped BLASTp program (www.ncbi.nlm.nih.gov/blast), gave a random expectation (*E*) value of less than or equal to 0.001. Additional searches for probable homologs with lower sequence conservation were performed using the PSI-BLAST program, run for three iterations using the same cutoff for inclusion of sequences into the profile. Adapted from the International Human Genome Sequencing Consortium (2001), reproduced by permission (© Nature Publishing Group).

(Sachidanandam *et al.,* 2001). What is the total number of SNPs (at the 1% level) in the human population? Using classical neutral population genetic methods, Kruglyak (2001) placed the figure at 11 million sites, so that less than 15% have been identified. Thus increased depth (number of individuals assayed) and breadth (genome coverage) will be needed before the catalog of human genotype variation can be said to be complete. Of course, most SNPs are believed to be neutral, so only functionally relevant changes need be considered for many biomedical applications (Kruglyak and Nickerson, 2001). The number of these is expected to be much smaller than the total given above, but it is difficult to separate them out, especially those that may lie in unidentified control regions of the genome. Another simplification is to use not raw SNPs in association studies, but linked groups of alleles called haplotypes (Seltman *et al.,* 2003). As the database of human SNPs and haplotypes grows, the hope is that it may be used to help develop somatic gene therapies for specific diseases and to predict an individual's

reaction to therapeutic drugs ("pharmacogenetics" or "pharmacogenomics") (Stephens, 1999; March, 2000; McCarthy and Hilfiker, 2000). Finding these associations is statistically challenging because although there is a great deal of data (Huang *et al.,* 2003), many of the associations are quite weak (Ioannidis, 2003).

Other potentially important forms of genomic variation in human populations have been described recently. Notably, Sebat *et al.* (2004) found evidence that copy-number polymorphisms involving large segments (more than 100 kb) of DNA contribute substantially to genomic variation between normal humans. However, the significance of such polymorphisms for human health, for example through gene dosage variation, remains largely unknown.

The genome sequence of the chimpanzee, the closest relative of humans, is seen as an important key to understanding exactly what "makes us human" and is therefore of great interest to evolutionary biologists and to the public at large. The mouse– human comparison provides a broad perspective on the mammalian genome, but the high number of rearrangements between the two species requires intensive searching for homologies. The human–chimp comparison will allow researchers to focus on differences. It had been reported (Sibley and Ahlquist, 1984, 1987; Ebersberger *et al.,* 2002) that sequence divergence between chimp and human was less than a few percent, although a recent study has found that aligned regions of human and chimp genomic DNA differ by around 5% when indels are accounted for (Britten, 2002). On the other hand, when attention is restricted to a sampling of genes themselves, the percent nonsynonymous DNA difference was found to be just 0.6% (Wildman *et al.,* 2003). It will be of great interest to isolate these differences, not only in protein-coding exons, but also in regulatory regions such as promoters and enhancers, and relate them to function. The prevailing hypothesis is that differences between human and chimpanzee are primarily owing to differences in gene expression during development, so expression studies are also essential. For example, preliminary mRNA studies show that central nervous system expression patterns diverge more between humans and primate relatives than do patterns for other organs (Normile, 2001; Enard *et al.,* 2002). Large-scale chimpanzee sequencing is already under way in the United States and in Japan by a group called the International Chimpanzee Genome Sequencing Project. In terms of biomedical research, however, there are stronger arguments for sequencing the rhesus macaque (*Macaca mulatta*) than the chimp (Cyranoski, 2002). Although differences between chimp and human pathologies are of significant interest, the chimp is no longer a common laboratory animal, compared to the rhesus macaque, which provides models for many human diseases.

Pufferfish Synergy

The human genome project was aided by other concurrent sequencing projects. Although not a traditional model organism for genetics, the pufferfish

Takifugu rubripes ("Fugu") was an ideal sequencing subject because of its remarkably compact genome and so was the second vertebrate to be completely sequenced. Fugu appears to have approximately the same number of genes as humans and a very similar exon–intron pattern, but has much shorter introns and intergenic regions, resulting in a genome about 1/9 as big as that of humans (Brenner *et al.,* 1993; Hedges and Kumar, 2002). The combined factors of compact genome and improved sequencing methods (whole-genome shotgun, in this case) allowed the Fugu genome to be completed for a cost of only about 12 million dollars, less than 1% of the total spent on the human genome project (Aparicio *et al.,* 2002). Besides the evolutionary insights to be gained from comparison of two distantly related vertebrates, the Fugu genome was used to help find functional elements and annotate the human genome (Aparicio *et al.,* 2002). The Fugu draft sequence was produced by a consortium led by the U.S. Department of Energy's Joint Genome Institute (JGI) in Walnut Creek, California, and the Singapore Biomedical Research Council's Institute for Molecular and Cell Biology (IMCB). The consortium's sequencing efforts were aided by two U.S. companies, Celera Genomics, Rockville, Maryland, and Myriad Genetics, Inc., Salt Lake City, Utah. Completion was announced at the 13th International Genome Sequencing and Analysis Conference in San Diego, California, on October 26, 2001.

Sequencing of the freshwater, nonpoisonous pufferfish *Tetraodon nigroviridis* was also announced by Genoscope (The French National Sequencing Center) in Paris, and the Whitehead Institute Center for Genome Research in Cambridge, Massachusetts, at around the same time as the report of the pufferfish genome. These two bony fishes with similarly compact genomes provide a useful contrast of vertebrate genome divergence. The pufferfish and human lineages have been separated for more than 400 million years, whereas *Tetraodon* and *Takifugu* are thought to have diverged 20–30 million years ago.

THE MOUSE AND RAT GENOMES: THE RISE OF MODERN MAMMALIAN COMPARATIVE GENOMICS

Approximately a year after the "completion" of the human genome project, when about 95% of the euchromatic sequence was available in finished form, the first draft sequence of the mouse genome was released (Mouse Genome Sequencing Consortium, 2002). Although the mouse genome is perhaps atypical of mammalian genomes in some ways, it is among the most valuable to use to shed light on human biology. This is largely owing to the status of mouse as the preeminent model mammalian organism in genetics. A great deal is known about the function of many mouse genes and many more can be elucidated using knockout studies that would be impossible to conduct in humans. Homologies between mouse and

human genes are readily determined by sequence comparisons and provide an initial key for functional studies. Further clues are provided by relative conservation and divergence of different genes and DNA stretches that lie outside of known genes. This latter category of conserved extragenic DNA provides valuable pointers to the location of promoters, enhancers, and other hard-to-detect but extremely important functional elements. Sensitive methods have been developed for determining genomewide homology mapping, even covering regions apparently not under selection (Schwartz *et al.,* 2003).

In many such applications of comparative genomics, it is necessary to distinguish carefully between homology and orthology of sequences. For example, if a pair of sequences is taken, one from mouse and one from human, homology (common ancestry) can be inferred based on a high degree of similarity at protein or DNA sequence level. However, it is not known whether the two sequences first diverged from their common ancestral sequence at the time of the rodent-primate split, or earlier. In the first case, when the sequences first diverge via a speciation event, they are called orthologous; in the second case, they must have first diverged by a gene duplication event and are said to be paralogous (see Chapter 5). The importance of the difference is that orthologous sequences can be used as proxies for their respective species in phylogenetic and timing analyses, whereas paralogous sequences cannot. This is illustrated in Figure 9.14.

For this reason, care was taken to identify mouse–human orthologs as quickly as possible. Although the orthology of two sequences cannot be determined with absolute certainty without extensive species and genome sampling and the use of a phylogenetic approach (Zmasek and Eddy, 2001), some useful methods are being employed to compare completely sequenced species like mouse and human. Specifically, a pair of sequences, *h* from human and *m* from mouse, is considered orthologous if *h*'s closest match in the mouse genome is *m* and *m*'s closest match in the human genome is *h*. In this way, human orthologs can be found for about 80% of mouse genes (Mouse Genome Sequencing Consortium, 2002).

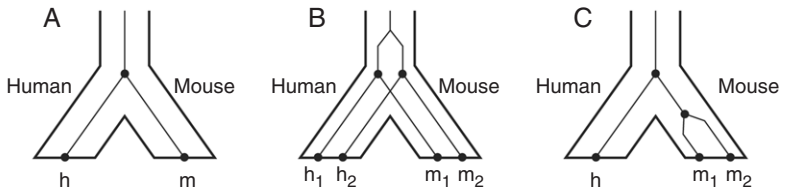

FIGURE 9.14 Orthology versus paralogy. In (**A**), the sequences *h* and *m* are orthologous because their most recent ancestral sequence coincides with the divergence of the mouse and human lineages. In (**B**), a gene duplication occurred before the divergence of these lineages. In this case, h_1 and m_1 are orthologs, as are h_2 and m_2 whereas h_1 and m_2 are paralogs, as are h_2 and m_1. Note that orthology is not necessarily a one-to-one relationship. In (**C**), *h* is orthologous to both m_1 and m_2.

Although the mouse genome is somewhat smaller than that of humans (2.5 Gb euchromatic DNA for mouse vs. 2.9 Gb for human) and has a slight but statistically significant difference in G+C content (Fig. 9.15), it appears to contain about the same number of genes. The main resource for mammalian gene detection and verification is the use of expression data such as cDNAs and ESTs (Hubbard *et al.,* 2002). Results from such searches are then integrated with results from *de novo* gene prediction, producing a final catalog. This process is far from clear-cut, as it relies on complete transcript libraries and uses gene-finding programs that have difficulties separating noise from data when genes are spread out as much as they are in mammals. This process can be facilitated when two genomes are available, because conserved regions within homologies may lead to discovery and validation of splice sites and other genomic elements (Korf *et al.,* 2001; Wiehe *et al.,* 2001).

The fraction of genes in mice or humans that do not appear to have homologs (orthologs or paralogs) in the other is less than 2%. The expansions of certain gene families are readily apparent in the mouse lineage, such as those involving immunity and olfaction, relative to their presence in humans, suggesting either mouse duplications or human losses in these functional areas. An example is the oligoadenylate synthetase (OAS) gene family involved in interferon-induced antiviral response, which shows many recent murine gene duplications (Kumar *et al.,* 2000). Although the gene sequences have been generally well conserved, their positions in the genome have not—that is, genes that are syntenic in one genome

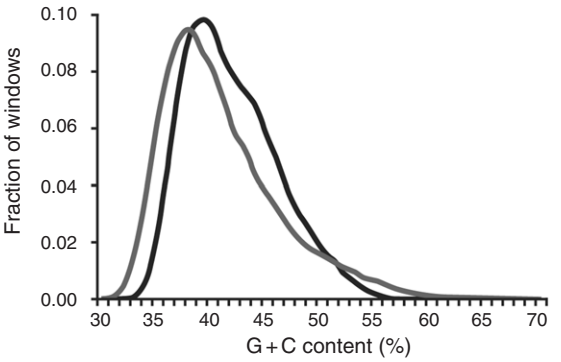

FIGURE 9.15 Distribution of G+C content of the human (gray line) and mouse (black line) genomes. Mice have a slightly higher mean G+C content than humans (42% versus 41%), but humans have a larger fraction of windows with either high or low G+C content. The distribution was determined using the unmasked genomes in 20 kb nonoverlapping windows, with the fraction of windows (y-axis) in each percentage bin (x-axis) plotted for both humans and mice. Adapted from the Mouse Genome Sequencing Consortium (2002), reproduced by permission (© Nature Publishing Group).

are not so in the other. Detailed analysis shows that chromosomal rearrangements have moved segments of DNA both within and among chromosomes, breaking many syntenies. The exact number of relocated segments is difficult to observe because some may be very small, particularly near the centromeres and telomeres, but the number of large segments conserved between humans and mice is at least several hundred (Kumar *et al.,* 2001; Eichler and Sankoff, 2003). This means that the average length of conserved segments between human and mouse genomes is rather small.

On a finer level, gene structure seems to be highly conserved. Analysis of 1506 pairs of genes for which there is strong evidence of orthology shows that 86% genes have the same number of exons and 91% of orthologous exon pairs have the same length in humans and mice. Only about 1% of introns, however, have identical length, and the average length of the mouse introns in this set was 3888 bp compared to 4661 bp for human. This reflects the smaller euchromatic C-value for mice (Mouse Genome Sequencing Consortium, 2002).

In April 2004, the complete genome sequence was published for another important (and indeed, the first) mammal used in medical research, the brown rat *Rattus norvegicus*. The rat genome sequence was obtained using a new approach that combines aspects of the traditional mapping and whole-genome shotgun methods used in the public and private human genome sequencing projects, respectively. Unlike previous mammalian analyses (i.e., between mice and humans), the mouse-rat comparison allows inferences to be made regarding genome evolution over a relatively short time scale (i.e., only 12–40 million years) (Kumar and Hedges, 1998; Rat Genome Sequencing Project Consortium, 2004). The euchromatic portion of the rat genome appears to be intermediate in size (2.75 Gb) relative to that of mice (2.5 Gb) and humans (2.9 Gb), and contains a similar number of genes. The preponderance of segmental duplications (which occur primarily in pericentromeric regions) is also intermediate in rats, and there are signs that some gene families have expanded by duplication in rats but not in mice (Rat Genome Sequencing Project Consortium, 2004). The activity of *L1* transposable elements (a LINE) also seems to be higher in rats than in mice, although a roughly similar number of SINE copies (~ 300,000) appears to have been inserted into the genomes of both rodents after the divergence of their respective lineages.

Sequences comprising about one billion nucleotides (39% of the rat genome) appear to be common to all three mammals, representing an "ancestral eutherian core," which includes around 95% of the known protein-coding and regulatory regions. About 28% of the rat sequence aligns only with mice, not humans, and another 29% aligns with neither of the two mammals. As compared with primates, rodents appear to have much more dynamic genomes, experiencing a faster rate of both molecular (base substitutions) and chromosomal (rearrangements) evolution (Kumar *et al.,* 2001; Kumar and Subramanian, 2002; Rat Genome Sequencing Project Consortium, 2004).

GENOME SEQUENCING IN PLANTS AND THEIR PATHOGENS

COMPARATIVE GENOMICS OF *ARABIDOPSIS*

Arabidopsis thaliana was a natural choice for the first plant genome to be sequenced. It had the smallest known genome and highest gene density of any plant, plus it was already extensively studied, is easy to grow, and has a short life cycle. Although the genome is only 157 Mb, it contains homologs to nearly all genes found in flowering plants but with much less repetitive DNA than most.

The *A. thaliana* project began in 1990 and involved researchers from many countries. Sequencing itself had begun by 1993, funded primarily by the European Union. By 1996, funding agencies from the United States, as well as Europe and Japan, also contributed to the work, and the *Arabidopsis* Genome Initiative (AGI) was set up with the intention of completing the sequencing (clone libraries were already available) by 2004. The agreement was a model of international scientific cooperation. A coordinating committee was to assign different portions of the genome as needed to prevent duplication of work. No sequence information was to be withheld to benefit any private group, and partial sequences were to be released as soon as available to one of the major databanks. As with all genome projects, it was becoming increasingly clear that ongoing annotation was vital to the value of the data and an organization was formed to help curate annotations. Another important part of the AGI project was a parallel effort to sequence gene expression data in the form of cDNA. Because of advancing sequencing technology, the project was completed well ahead of schedule, with the first report released in 2000 (*Arabidopsis* Genome Initiative, 2000).

The genome size, gene content, and gene family diversity of *A. thaliana* are comparable to that of *C. elegans,* but differences of gene content in different functional classes are significant. For example, less than 20% of *A. thaliana* proteins involved in transcription have strict homologs within *C. elegans* (BLASTp *E*-value less than 10^{-30}). In contrast, more than 40% of proteins involved in protein synthesis and signal transduction have such homologs, suggesting common ancestry. It is particularly interesting that relatively high proportions (15–30%) of proteins in the energy and metabolism categories have close matches in *E. coli*. This may result from lateral transfers or unusually extreme conservation. About 35% of genes in *A. thaliana* are apparently unique, or at least are not present in the animal and yeast genomes sequenced thus far. About 150 families of genes, including structural proteins and enzymes, appear to be unique to plants (*Arabidopsis* Genome Initiative, 2000).

The proportion of proteins belonging to families of more than five members is substantially higher in *A. thaliana* (37.4%) than in *C. elegans* (24.0%), as is the proportion of gene families with more than two members. These features of

A. thaliana, and presumably other plant genomes, may indicate less constraint on genome size in plants and/or a higher propensity for genome duplication. In fact, most (58%) of the *A. thaliana* genome is in the form of large (at least 100 kb) duplicated segments with more than 50% sequence identity. This indicates that the genome was structured by a past polyploidy event (see Chapter 6), as is known to be very common in plant evolution (see Chapter 7). Tandem segment duplications also appear to be common in *A. thaliana,* with about 1500 tandemly duplicated arrays of genes, containing an average of around three genes each, but up to a maximum of 23. This suggests that unequal crossing over may be an important factor in plant genome evolution as well.

The Rice Genome

Rice (*Oryza sativa*) is the most important food crop in the world, providing staple nourishment for half the world's population, and is also the cereal crop with the smallest genome, about 490 Mb in size. Rice genetics has been intensively studied and comprehensive genetic and physical maps have been available for some time. For many mapped traits (Gale and Devos, 1998), the rice genome exhibits a strong colinearity, or preservation of genetic linkage relationships, with the much larger genomes of other grain crops such as wheat, barley, oats, and corn (Fig. 9.16). In the same way, *A. thaliana* serves as a genomic key to the Brassica group of crop plants (cabbage, cauliflower, mustard, rape, rutabaga, and turnip, among others). These qualities made rice a practically ideal next choice for sequencing after the tiny-genomed model plant *A. thaliana*. Two strains of rice, the *japonica* and *indica* varieties, were sequenced by different groups (Goff *et al.,* 2002; Yu *et al.,* 2002). The *japonica* announcement by the private company Syngenta raised controversy because the full results were not deposited in a public data bank such as Genbank, although the public did have limited access to a private database (Brickley, 2002). This marked the second time the journal *Science* allowed authors of a scientific paper to withhold full access to sequence data described in a publication, the other instance being the Celera human genome paper.

Although estimates of the number of genes in rice are subject to the uncertainties that apply to all large-genomed eukaryotes, plus some unique problems, there seem to be many more genes than in *A. thaliana,* and quite probably more than in mammals such as *H. sapiens*. This high gene number was not unanticipated (Messing, 2001). A remarkable feature of the rice genome is the presence of a strong G + C content gradient within genes. Often, the 5′ end is up to 25% richer in G + C content than the 3′ end (Wong *et al.,* 2002). No comparable gradient is seen in *A. thaliana*. A consequence of this gradient is that codon usage also varies from one end of a gene to the other, complicating the work of

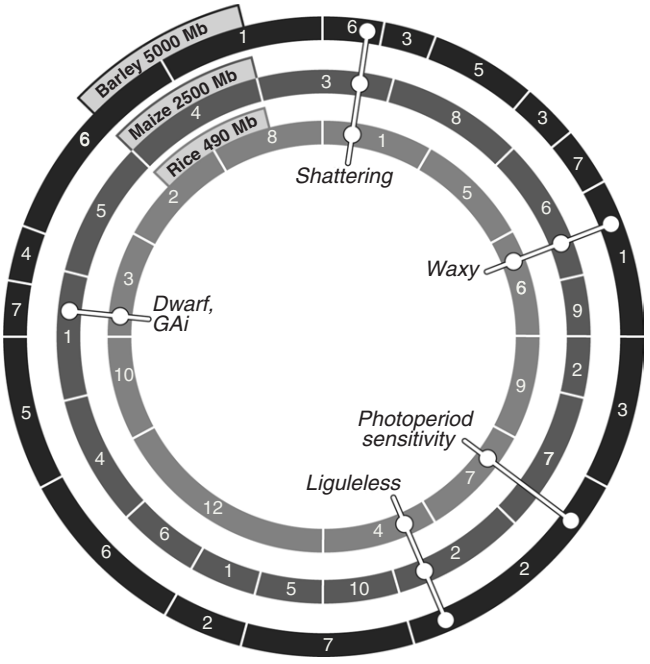

FIGURE 9.16 Genome colinearity among the grasses. The rice genome is the smallest among the grasses most commonly grown as crops. However, enough genomic similarity exists that the genomes of these grain species can be approximated as concentric circles, and the information from the smaller genome of rice provides insight into where to find genes of interest in the larger genomes of the other grasses. Each circle represents a single grain species, with its chromosomes collected end-to-end in a parsimonious manner to best match the structure of the rice genome. A few examples of genetic loci associated with particular traits are shown to illustrate that similar genes occur in similar portions of the genomes across these grasses. Other grain crops that show colinearity with rice include wheat, oats, pearl millet, sorghum, sugar cane, and foxtail millet. Adapted from www.ipw.agrl.ethz.ch/~mbucher/useful/riceposter.pdf, reproduced by permission (© American Association for the Advancement of Science).

gene-detection programs. Therefore gene predictions and counts have an unusual amount of uncertainty in this case.

It appears that about 80% of *A. thaliana* genes have a homolog in rice, whereas only 50% of predicted rice genes have a homolog in *A. thaliana*. The latter figure extends to other sequenced organisms as well, so that about half of rice's genes seem to be novel, without known homologs or functions. Some of this figure may be an effect of the G+C content gradient in rice making homology determination difficult, but it still suggests a great deal of innovation in rice, perhaps amplified by duplications, since the monocot–eudicot divergence.

The Rice Blast Fungus: *Magnaporthe grisea*

In addition to the sequencing of this most important cereal crop, rice's major fungal pathogen, the rice blast fungus, *Magnaporthe grisea*, has also been sequenced (Martin *et al.,* 2002). *M. grisea* represents an excellent model organism for studying fungal phytopathogenicity and host–parasite interactions. Like many other fungal pathogens, *M. grisea* is a haploid, filamentous Ascomycete. It has a fairly small genome of around 40 Mb partitioned into seven chromosomes. *M. grisea* is also closely related to the nonpathogen *Neurospora crassa,* an important model organism for the study of eukaryotic genetics and biology. The main line of defense against this fungus has recently been genetic, via host resistance, although this entails a constant arms race against a rapidly evolving pathogen. Having both sequenced genomes should provide valuable insight for understanding questions of fungal–host interactions. For example: What genes come into play in both species during infection? How does the pathogen recognize when it is on a suitable surface to initiate the infection process? Which genes control host specificity? Mitigating the effects of this fungus could directly help feed tens of millions of people.

OTHER INVERTEBRATE ANIMAL GENOMES

The Mosquito Genome

Anopheles gambiae is the major vector of the human malaria pathogen *Plasmodium falciparum* in Africa. Although malaria has been eliminated in Europe, new malaria control techniques are urgently needed in sub-Saharan Africa, and improved understanding of the ecological relationship of the pathogen and its hosts may provide a key to its elimination. *A. gambiae* is also of interest in that it provides a comparison to *D. melanogaster*. The lineages diverged more than 250 million years ago (Zdobnov *et al.,* 2002) and initial studies, prior to full sequencing, showed considerable divergence in terms of genome rearrangements, although broad conservation of synteny on chromosomal arms was noted (Bolshakov *et al.,* 2002).

As shown in Table 9.1, the mosquito genome is more than twice the size of the *D. melanogaster* genome, although they contain a very similar number of genes (Holt *et al.,* 2002). The difference is mostly owing to a greater amount of intergenic material in the mosquito. For example, the transposable element content of *A. gambiae* is approximately 16% and 60% of euchromatin and heterochromatin, respectively (Rizzon *et al.,* 2002). The fact that most dipterans, including many *Drosophila* species, have genomes closer in size to the mosquito's, suggests that the lineage containing *D. melanogaster* experienced a reduction during recent

evolutionary times (Petrov *et al.,* 1996). Some mechanisms for this loss of noncoding material have been proposed (Hartl, 2000; Petrov, 2001).

An interesting contrast between arthropod and vertebrate genomes is provided by the frequency of large duplicated blocks. The number of blocks in *A. gambiae* containing at least three genes that also appear elsewhere was only about 100, compared to more than 1000 for the human genome (Holt *et al.,* 2002). The presence of such repeats in the mouse genome seem comparable to that for human (Mural *et al.,* 2002), whereas the *C. elegans* and Fugu genomes show little evidence of such duplications (*C. elegans* Sequencing Consortium, 1998; Aparicio *et al.,* 2002).

Preliminary analyses of the *A. gambiae* genome and expressed proteome suggest several strategies for reducing human disease associated with this animal. For example, comparisons of gene expression profiles before and after the female's blood meal reveal that certain products (lipid synthesis and transport proteins, egg melanization factors, lysosomal enzymes) are up-regulated, whereas others (involving cytoskeletal and muscle contractile machinery, glycolysis, and proteins associated with vision) are down-regulated. Understanding these changes may provide opportunities for intervention to disrupt reproduction. Other approaches involve disrupting the mechanism by which the mosquito finds the human (e.g., odor receptors [Hill *et al.,* 2002]), or by interfering at some point in *P. falciparum*'s complex life cycle within its host, possibly using the *A. gambiae* immune response (Dimopoulos *et al.,* 2001; Christophides *et al.,* 2002).

The Sea Squirt: A Primitive Chordate

Ciona intestinalis is a urochordate, the most basal branching group of chordates, and therefore was considered an important target for complete genome sequencing. The adult is a sessile filter feeder, but the tadpole has a notochord. This invertebrate chordate has approximately half as many genes as sequenced vertebrates and gives a perspective on the evolution of this group (Dehal *et al.,* 2002). The difference in gene content is thought to result from the proliferation of gene families involved in vertebrate development, so that the *C. intestinalis* genome provides a view into vertebrate ancestry. On the other hand, some genes known to be in both insects and vertebrates are missing in *C. intestinalis*. The *Hox* gene family provides an interesting example. Invertebrates have a single cluster of up to 13 homologs, whereas vertebrates have several clusters. *C. intestinalis* is in the invertebrate camp in that it has a single (albeit widely spread) cluster of nine *Hox* genes, although *Hox* 7, 8, 9, and 11 are apparently absent (Gionti *et al.,* 1998). A remarkable example of an apparent *C. intestinalis* innovation is a set of genes related to the substance tunicin (Krishnan, 1975). Tunicin is a cellulose-like carbohydrate present in the urochordate (or, "tunicate") body-casing. *C. intestinalis* contains at least

one potential cellulose synthase and several endoglucanases related to the synthesis and degradation of this material. Homologs are found in *A. thaliana* and in termites and wood-eating cockroaches, although horizontal gene transfer from symbionts may be involved (Lo *et al.,* 2000; Nakashima *et al*., 2004).

Genomewide Duplications in Vertebrates?

As discussed in detail in Chapter 6, comparisons of vertebrate and invertebrate genomes have also shed much light on the debate about genomewide duplication events in the early history of vertebrates. In addition to the *Hox* genes mentioned previously, other gene families have been found to suggest two rounds of whole genome duplication in early vertebrates (Wolfe, 2001; Gu *et al.,* 2002) (see Chapter 6). Ohno (1970) argued several decades ago that such events may have given vertebrates a sudden leap in body-plan complexity, which, if true, would have significant implications for the understanding of large-scale vertebrate evolution (see Chapter 11).

PROTIST GENOMES

One should not forget that animals, plants, and fungi form only a fraction of total eukaryote diversity. The "Protista" form a polyphyletic collection of unicellular eukaryotes whose genomes are also interesting and important for several reasons. First, many are pathogens that cause an enormous amount of human disease throughout the world, particularly in the tropics. Their early divergence relative to other eukaryotes (Baldauf, 2003) (Fig. 9.17) also makes them important for major evolutionary questions such as those concerning the evolution of organelles. Some, such as *Dictyostelium* (discussed in a later section), are important as simple models of intercellular communication.

Encephalitozoon cuniculi: A Parasitic Eukaryote with a Tiny Genome

Encephalitozoon cuniculi is a member of the microsporidia, a group of obligate parasites that infest many animal hosts, including rabbits and immunocompromised humans. Because they lack mitochondria, the microsporidia were at first thought to have originated from a deep eukaryotic branch before the organelles were acquired, but closer inspection showed that their nuclear genomes contain some mitochondrial-type genes. This fact, along with further phylogenetic analysis,

indicates that they are more properly classified as fungi that have lost their mitochondria (Katinka *et al.,* 2001).

Most obligate parasites are degenerate in various ways, and *E. cuniculi* is no exception. Its tiny genome (2.9 Mb) is smaller than that of many prokaryotes, including *E. coli*. Sequencing announced in 2001 showed that this small size results from a rarity of introns, lack of transposable elements, reduced metabolic and synthetic activities, reduced intergenic spacers, and even reduced protein lengths relative to eukaryotic orthologs. The mean protein length is only 363 amino acids (aa), compared to 472, 435, 543, and 461 aa for *S. cerevisiae, C. elegans, D. melanogaster,* and human, respectively, and is more comparable to that of prokaryotes such as *E. coli* (315 aa) and *Mycoplasma pneumoniae* (348 aa) (figures from www.ebi.ac.uk/proteome). In a comparison of 350 proteins with *S. cerevisiae* homologs, the *E. cuniculi* sequence was shorter in 85% of the cases, with an average difference of 14.6% (Katinka *et al.,* 2001). Zhang (2000) has suggested that the longer proteins of higher eukaryotes allow more sophisticated gene regulation networks.

Plasmodium: The Malaria Pathogen

In 1996, the International Malaria Genome Sequencing Consortium was formed to sequence the genome of the protist *Plasmodium falciparum*. This organism is of great importance because of its devastating effect as a pathogen—it is responsible for more than 2.5 million deaths each year, a large proportion of them children—but also presents special problems in sequencing. Its genome is twice the size of that of yeast, and its extremely low G + C content (~ 20%) creates technical problems for sequencing. Despite these complications, the complete sequence was announced in 2002 (Gardner *et al.,* 2002). The sequence for the related pathogen in mice, *Plasmodium yoelii,* was announced at the same time for comparative analysis (Carlton *et al.,* 2002).

Compared to free-living protists, *P. falciparum* has fewer genes for enzymes and membrane transporters, but contains an extensive apparatus for evading host defenses. More than 60% of its proteins do not appear to have homologs in previously sequenced eukaryotes. It is not known whether this high figure is a result of *Plasmodium*'s phylogenetic position, high A+T content, or parasite status. Comparison with other eukaryotic genomes reveals that, in terms of overall genome content, *P. falciparum* is slightly more similar to *Arabidopsis thaliana* than to other taxa (Gardner *et al.,* 2002). However, the implied affinity with the plant kingdom may be related to horizontal gene transfer.

Now that the genomes of *P. falciparum, Anopheles gambiae,* and *Homo sapiens* have been completed, the raw data are available to break or control this devastating parasitic cycle. The *P. yoelii* sequence is of great importance here as well, because

it has been used as a proxy in laboratory studies for the human parasite, whose life cycle cannot be completed *in vitro*.

DICTYOSTELIUM: THE "SLIME MOLD"

Dictyostelium discoideum is a haploid protist that has been intensively studied, primarily because of its social life. A group of free-living cells can aggregate into a motile mass that exhibits morphological and biochemical development for the purpose of common reproduction. For this reason, *Dictyostelium* is an excellent model organism for studying intercellular signaling, specialization, and cooperation in a simple context. The *Dictyostelium* genome has six chromosomes totaling about 34 Mb. Like *Plasmodium,* its low G+C content (30%, down to 10% in some regions) challenges conventional sequencing methods. Currently *Dictyostelium* is being sequenced through a collaborative effort among American, British, French, and German laboratories.

In general, most protist genomes can now be sequenced quickly and relatively cheaply. Given the immense impact of these organisms on human health throughout the world, fruits of this research hold the promise of great medical benefits.

COMPARATIVE GENOMICS AND PHYLOGENETICS IN EUKARYOTES

The availability of more sequence data is also refining ideas regarding within-eukaryote relationships and their times of divergence (see review in Hedges, 2002). The tree in Figure 9.17 reveals several important aspects of the current understanding of early eukaryote evolution. It shows the close relationship between animals and fungi, the relationships of the several forms of algae with plants, and that the "protists" are indeed a group with very diverse origins (Baldauf, 2003).

Figure 9.18 shows typical positions where eukaryotic proteins are found in phylogenetic reconstruction in relation to their prokaryotic homologs. Eukaryotic proteins involved in transcription and translation often cluster with archael homologs, whereas metabolic proteins often cluster with bacteria (Rivera *et al.,* 1998). This pattern is thought to result from the symbiotic origin of eukaryotes and horizontal gene transfer from organelles to the nucleus (Margulis, 1996; Gupta, 1998; Doolittle, 1999).

Associated with the phylogeny question is the issue of timing: How long ago did certain lineages diverge from one another? Traditionally, this was settled by the dating of fossils, giving lower bounds on the age of divergences. The molecular

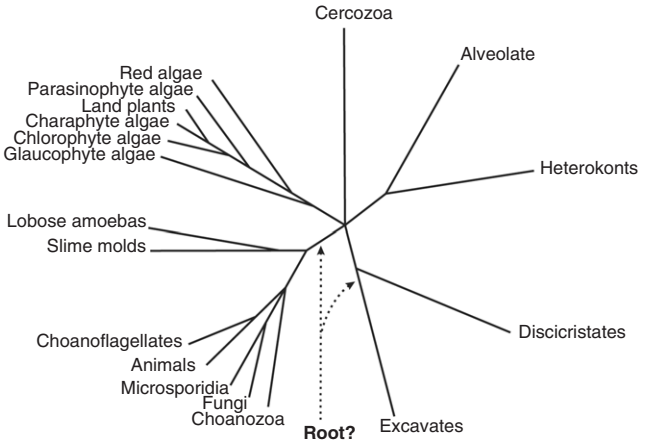

FIGURE 9.17 A consensus phylogeny of eukaryotes based on both molecular and ultrastructural data. The vast majority of characterized eukaryotes, with the notable exception of major subgroups of amoebae, can now be assigned to one of eight major groups. Opisthokonts ("basal flagellum") have a single basal flagellum on reproductive cells and flat mitochondrial cristae (most eukaryotes have tubular ones). Eukaryotic photosynthesis originated in plants; theirs are the only plastids with just two outer membranes. Heterokonts ("different flagellae") have a unique flagellum decorated with hollow tripartite hairs (stramenopiles) and, usually, a second plain one. Cercozoans are amoebae with filose pseudopodia, often living within tests (hard outer shells), some very elaborate (foraminiferans). Amoebozoa are mostly naked amoebae (lacking tests), often with lobose pseudopodia for at least part of their life cycle. Alveolates have systems of cortical alveoli directly beneath their plasma membranes. Discicristates have discoid mitochondrial cristae and, in some cases, a deep (excavated) ventral feeding groove. Amitochondrial excavates lack substantial molecular phylogenetic support, but most have an excavated ventral feeding groove, and all lack mitochondria. Adapted from Baldauf (2003), reproduced by permission (© American Association for the Advancement of Science).

clock hypothesis—that the rate of molecular sequence divergence is often constant for a particular gene over multiple lineages—promised to shed new light on the entire question. After initial controversy about the universality of the clock, tests were devised to reject genes that violated rate-constancy. In theory, phylogenetic trees based on clocklike genes would have branch lengths proportional to elapsed time. If properly calibrated against at least one well-established divergence time, the differences among gene sequences could be used to extrapolate the timings of all phylogenetic events in the tree. Of course, things are never so easy, and one of the main shortcomings of the plan was the relatively large variance of estimates based on too few sites/genes (Benton and Ayala, 2003; Hedges and Kumar, 2003). Not surprisingly, estimates inferred using many genes or proteins are more robust (Hedges and Kumar, 2003). It is also important to consider differences

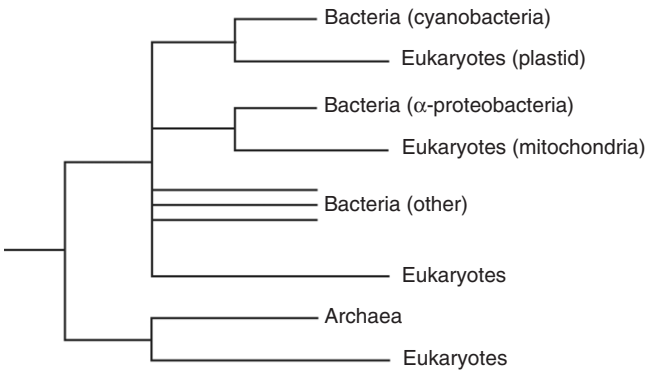

FIGURE 9.18 Eukaryotes consistently evolve faster than prokaryotes. This tree indicates the four general locations where eukaryotic protein sequences typically cluster in the evolutionary tree of prokaryotes. The rate of evolution on each lineage (branch) is indicated diagrammatically by relative branch length (long branch signifies fast rate; branch length is not meant to be proportional to time or actual rate). Adapted from Hedges *et al.* (2001), reproduced by permission of the author.

among major taxa; for example, Hedges *et al.* (2001) found that protein products of eukaryotic genes evolve, on the average, 1.2 to 1.6 times faster than their prokaryotic homologs.

A related question in comparative genomics involves the rate of neutral evolution (essentially equal to mutation rate) (Kimura, 1983). Does it vary from lineage to lineage and among genomic regions? These questions have been intensively studied in mammals. Early studies (Wu and Li, 1985; Li and Tanimura, 1987) reported that the neutral rate for rodents was up to several times greater than those for primates and artiodactyls. Other studies also reported significant differences among lineages or regions (Britten, 1986; Wolfe *et al.,* 1989; Mouchiroud *et al.,* 1995; Matassi *et al.,* 1999). Generation time differences (Li and Tanimura, 1987) and differences in DNA repair mechanisms (Britten, 1986) have been suggested as factors to account for this variation. However, the numbers of genes used in these studies were small and estimation errors may have been made because of incorrect fossil dates or inappropriate outgroups (Easteal *et al.,* 1995). A large-scale study using more than 5000 genes and representatives of a broad range of placental mammal groups concluded that there is little significant variation in neutral evolution rate among lineages, and that the rate is approximately 2.2×10^{-9} per base pair per year (Kumar and Subramanian, 2002). This improves prospects for accurate phylogenetic timing using neutral (4-fold degenerate) DNA sites. Ellegren *et al.* (2003) presented a detailed account of the current debate on rates of mutation and neutral evolution in mammalian genomes.

CONCLUDING REMARKS AND FUTURE PROSPECTS

Complete Genome Sequencing

Table 9.1 lists 18 eukaryotes as "completely" sequenced, including two plants, nine animals, four fungi, and three protists. Because of decreasing costs (now at a few cents per base for large-scale shotgun sequencing), this list is expected to grow rapidly, perhaps reaching 100 in only four or five years. Of course, many factors have influenced the choice of organisms to be sequenced. Traditional model organisms have been favored, as have organisms with small genomes. From this admittedly biased sample, a picture of phylogenetics and gene relationships is nevertheless coming to light. One of the surprises in the picture is how much all the sampled kingdoms have in common. Figure 9.19 shows the relative number of shared and unique protein domains among 14 of these species (Chothia *et al.,* 2003). The overwhelming fraction of domains in each organism is shared by all

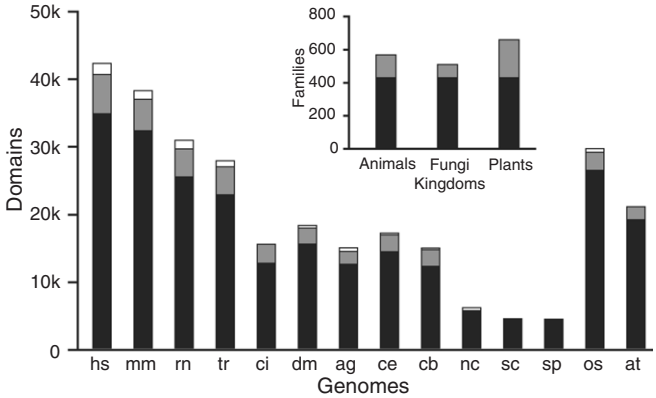

FIGURE 9.19 Contribution of common families to the protein repertoires of various eukaryotes. All or part of about 50% of eukaryote sequences are homologous to domains in proteins of known structure. The numbers of domains that belong to the 429 families common to all 14 eukaryotes studied are shown in black. Additional contributions of families common to the genomes in only one kingdom are shown in gray. For the animal genomes—human (*Homo sapiens,* hs), mouse (*Mus musculus,* mm), rat (*Rattus norvegicus,* rn), pufferfish (*Takifugu rubripes,* tr), sea squirt (*Ciona intestinalis,* ci), fruit fly (*Drosophila melanogaster,* dm), mosquito (*Anopheles gambiae,* ag), and nematodes (*Caenorhabditis elegans,* ce, and *C. briggsae,* cb)—there are 136 additional common families. For the three fungi—bread mold (*Neurospora crassa,* nc), budding yeast (*Saccharomyces cerevisiae,* sc), and fission yeast (*Schizosaccharomyces pombe,* sp)—there are 75 additional common families. For the two plants—rice (*Oryza sativa,* os) and thale cress (*Arabidopsis thaliana,* at)—there are 229 additional common families. Adapted from Chothia *et al.* (2003), reproduced by permission (© American Association for the Advancement of Science).

eukaryotes, a smaller fraction is kingdom-specific, and a mere sliver is unique to the organism, even though, because of the sparse sampling, some organisms on the list represent a broad group such as dicot plants or fishes. An emerging theme of the entire eukaryote sequencing effort is the common basis of eukaryotic life. Although ancestral characters, such as mitochondria, are occasionally lost, and duplicated protein genes acquire novel functions, there is still a great common core of protein domains that constitute the machinery of eukaryotic life, whether animal or plant, fungus or protist, unicellular or multicellular.

The NHGRI maintains a list of "high priority sequencing targets." Large NHGRI-sponsored sequencing centers that have excess capacity resulting from completion of human, mouse, and rat sequencing may use the released capacity for the sequencing of these organisms (Olson and Varki, 2003). As of May 22, 2002, this particular list included the chimpanzee *P. troglodytes,* the chicken, the honeybee, the sea urchin *S. purpuratus,* a protozoan *T. thermophila,* and 15 fungi (Check, 2002). Plants are excluded from this list, because those are handled in the United States by the National Science Foundation and the United States Department of Agriculture. The NHGRI's priority list is being reviewed and updated three times a year and will no doubt soon be expanded to include other model eukaryotic organisms.

In terms of economic impact and potential direct value to a large portion of the world's population, comparative plant genomics, especially with respect to cereal plants, is of immense importance. The factors of population growth (conservatively predicted at 2 billion additional mouths to feed over the next 50 years), the requirement to better manage and preserve world ecosystems, and the as-yet not fully understood effects of global climate change present a formidable challenge to agriculture. Although cereal production uses about half of available farmland and accounts, directly or indirectly, for ⅔ of all human caloric intake, per capita cereal production has been declining (Dyson, 1999). Extracting more food value from cereal crops is key to feeding a hungry world for the foreseeable future and this requires a deeper understanding of the organisms. Comparative plant genomics has been instrumental in building a better understanding of the complete spectrum of cereal plant characteristics and the genetic bases of the differences among crops. Because cereal genomes tend to be large with an abundance of repetitive sequences (common hexaploid wheat *T. aestivum* has a 17 Gb haploid genome size), the availability of high-throughput sequencing methods was essential to this program.

In April 2001 a meeting was held at the International Maize and Wheat Improvement Center (CIMMYT) in Mexico that outlined five prongs of cereal research: (1) alleviating abiotic stress, (2) alleviating biotic stress, (3) adding value to cereals, (4) improving the yield potential of some cereals, especially by modifying photosynthesis, and (5) coordinating the development of comprehensive,

freely available genomic tools and databases for improving cereals. An international Cereals Comparative Genomics Initiative (CCGI) program was sponsored by the Rockefeller Foundation and USAID to help advance these goals.

After basic research, the next steps toward crop enhancement using results from comparative genomics would be the actual production, testing, and distribution of genetically modified crop seeds. This raises complex political and ethical issues. The process is not unprecedented, however. Commercially produced transgenic Bt (insecticidal) corn, cotton, and soybeans and Roundup-Ready (antiherbicidal) corn and soybeans are now common in some countries, including the United States. Calgene's transgenic Flavr Savr tomato was approved for use in the United States in 1994, but was not a commercial success because of problems with characteristics of the crop related to its growth and harvesting (Nicholl, 2002). In 1999, on the noncommercial, humanitarian front, Ingo Potrykus and collaborators developed a method of inserting beta-carotene, a precursor of vitamin A, into rice endosperm where it does not normally occur (Ye *et al.,* 2000). Such rice was dubbed "Golden Rice" and had great potential of improving the nutrition of hundreds of millions of people, including hundreds of thousands of children who go blind from vitamin A deficiency. Even though it was essentially publicly funded and altruistically motivated, this project has encountered many obstacles of a bureaucratic, legal, and political nature (Potrykus, 2001) and is progressing slowly.

However, basic research and high-throughput sequencing continue at an increasingly rapid and economical pace. Sequencing efforts are in progress for various amphibians and fishes, several more insects, turkey, sea urchin, cow, dog, horse, kangaroo, pig, numerous fungi, algae, oat, coffee, soybean, cotton, barley, banana, corn, and protist parasites such as *Giardia* and *Leishmania*. Sequencing of the honeybee (*Apis mellifera*), jointly funded by the NHGRI and the U.S. Department of Agriculture (USDA), was recently completed in draft form. Draft sequences have also recently been completed for the chicken (*Gallus domesticus*) by the Washington University Genome Sequencing Center, and for the chimpanzee (*Pan troglodytes*) by the Broad Institute and Washington University Genome Sequencing Center. Updates of past, present, and future projects are available from the Genomes OnLine Database (GOLD) (www.genomesonline.org).

Genomic studies of eukaryotes have been biased toward vertebrates (mammals in particular), economically important plants, traditional model organisms, and pathogens. The range of eukaryote diversity is very broad, however, and knowledge of it is increasing every year. For example, recent studies show an unexpected abundance and diversity of the so-called picoeukaryotes (eukaryotes less than a few micrometers in size) (Moon-van der Staay *et al.,* 2001). The existence of very small eukaryotes, similar in size to, and even smaller than, many bacteria, has been known for some time. The currently smallest known autotrophic eukaryote

is *Ostreococcus tauri* (Courties *et al.,* 1998). This planktonic organism isolated from the Mediterranean is less than 1 μm in size and has a 10.2 Mb genome consisting of 14 linear chromosomes, plus several mitochondria and one chloroplast. It now appears that such miniature eukaryotes may be a more significant part of the planktonic community than formerly appreciated (Lopez-Garcia *et al.,* 2001) and may be represented in at least five of the eight major divisions of eukaryotes shown in Figure 9.17. Such eukaryotes tend to have small genomes and may provide an inexpensive way of exploring eukaryote genome diversity (Baldauf, 2003).

Partial-Genome Comparisons

On April 14, 2003, the International Human Genome Sequencing Consortium announced (final) "completion" of the stated goals of the Human Genome Project, to the specified degree of coverage, accuracy (at least 99.99%), and annotation. This was accomplished approximately two and a half years ahead of the projected completion date of 2005, and at a cost of about 10% less than the projected $3 billion. In addition, a great deal of extra data were produced, including more than 3 million SNPs, and cDNAs for more than 70% of human and mouse genes. All the data have been quickly deposited in public databases, with no restriction on usage. But is a sequencing project ever really completed? The real value of sequence data lies in its annotation. This aspect of the job is never-ending, but pointing the way is the NHGRI Encyclopedia of DNA Elements (ENCODE) initiative, which seeks to develop efficient, comprehensive, high-throughput technologies for the identification and verification of all types of sequence-based functional elements, particularly those other than coding sequences, such as non-protein-coding genes, transcriptional regulatory elements, and determinants of chromosome structure and function. Development of these technologies will allow not only comprehensive annotation of the immediate target, the human genome, but will lead to efficient analysis of homologous regions in a range of mammals and other vertebrates for broad comparative purposes.

A similar comparative approach can be taken using partial genomic sequences. As a complement to whole-genome sequencing efforts, researchers in the NIH Intramural Sequencing Center Comparative Sequencing Program sequenced segments of genomic DNA from 12 vertebrate species, all orthologous to a segment of about 1.8 Mb on human Chromosome 7, containing 10 genes, including the gene mutated in cystic fibrosis (Thomas *et al.,* 2003). From these sequences they identified more than 1000 "multispecies conserved sequences" (MCSs), most of which did not involve known coding sequences. These MCSs were found to be overrepresented in regions immediately upstream of transcription start sites and

in introns, and most do not correspond to known regulatory elements. This work thus provides a rich supply of candidates for future functional studies.

The Tree of Life

One of the principal ongoing projects of modern biology is the construction of a comprehensive tree of life showing evolutionary relationships. Early attempts at finding the relationships among taxa were based on morphological characters and had many successes, as in sorting out the relationships among major groups such as birds, mammals, reptiles, amphibians, and fishes. Some finer divisions of the tree of life remained controversial, such as the relation among mammalian orders, and the deepest divisions among major groups of organisms remained murky. Introduction of molecular methods eventually resolved some questions, such as placement of the cetaceans, and revised the view of the deepest divisions of the tree of life, but the use of only a few genes proved yet unable to resolve, for example, basic questions about the radiation of mammals.

A recent important project umbrella relating to comparative genomics is the National Science Foundation's Tree of Life Initiative. The NSF has solicited proposals for establishing a phylogenetic framework, Darwin's "Great Tree of Life," for the approximately 1.7 million described species of organisms. Many phylogenetic studies have been done by small teams of researchers to elucidate relationships within taxa of interest using molecular or morphological methods, but the Tree of Life project is aimed at funding larger multidisciplinary teams that are able to apply as much evidence as possible to create a definitive large-scale structure to which smaller specialized studies will eventually contribute. Teams of investigators also will be supported for projects in data acquisition, analysis, algorithm development, and dissemination in computational phylogenetics and phyloinformatics. Because it summarizes biological diversity, such a great tree would be useful in many fields, such as tracking the origin and spread of emerging diseases and their vectors, bioprospecting for pharmaceutical and agrochemical products, targeting biological control of invasive species, and evaluating risk factors for species conservation and ecosystem restoration. This project is timely in that, on one hand, a great flood of new molecular characters in the form of sequences is now available, whereas on the other hand, there appears to be a major extinction event induced by human activity, the result of which is that information about species, many of them as-yet unknown, is being lost. The first round of funding has supported methodological studies as well as taxon-specific studies on roundworms, spiders, and birds. In support of such large-scale phylogenetic efforts, it has recently been shown that current methods for constructing phylogenetic trees can be scaled up to infer very large phylogenies (Tamura *et al.,* 2004).

The Charter of Genomics

The availability of sequence data on a large scale has already begun transforming the modern view of biological patterns and processes at the taxon, population, individual, and biochemical levels. Although the Human Genome Project was the quintessential "big science" project, data from it are having the effect of democratizing and globalizing science. Sequence data are now freely available everywhere and require only inexpensive consumer-grade computing equipment to process. Laboratories throughout the world can now add this genomic view to their biological investigations with little additional investment in terms of equipment. Potential applications of these kinds of knowledge to problems in medicine and agriculture range from safer use of pharmaceuticals (pharmacogenomics) to developing effective countermeasures to devastating crop diseases such as rice blast. Of course, knowledge is always a two-edged sword and many questions relating to privacy, safety, and other ethical concerns will arise. These applications and concerns have been laid out by the United States National Human Genome Research Institute, which defined its continuing mission and its vision for the future of the genomics community in terms of three broad areas (Collins *et al.,* 2003).

The first area deals with the application of genomics to biology, which is concerned with elucidating the structure and function of genomes. Goals consistent with this area are: (1) comprehensive identification of the structural and functional components encoded in the human genome, (2) elucidation of the organization of genetic networks and protein pathways in establishing phenotypes, (3) development of a detailed understanding of heritable variation in the human genome, and (4) understanding evolutionary variation among species and the underlying mechanisms.

The second area deals with the application of genomics to human health, which is focused on translating genome-based knowledge into health benefits. The goals of this area are to develop (1) robust strategies for identifying genetic contributions to disease and drug response, (2) strategies to identify gene variants that contribute to good health and resistance to disease, (3) genome-based approaches to prediction of disease susceptibility, drug response, and detection of illness, (4) genome-based approaches to molecular taxonomy of disease states, and (5) new understanding of genes and pathways to develop new therapeutic approaches to disease.

The third area is the application of genomics to society, which is focused on promoting the use of genomics to maximize benefits and minimize harm to society. The goals of this area are to (1) develop policy options for the uses of genomics in both medical and nonmedical settings, (2) understand the relationships between genomics, race, and ethnicity, and the consequences of uncovering these relationships, (3) understand the consequences of deciphering the genomic contributions to human traits and behaviors, and (4) assess how to define the

ethical boundaries for uses of genomic information. Unlike the original set of goals for the NHGRI, these are quite broad and open-ended and can be viewed as a comprehensive charter for genomics in the coming century.

Although its roots can be traced back to the earliest chromosomal work, comparative genomics involving (nearly) complete genome sequencing is a science still in its infancy. Fast-growing and full of potential, its maturation is likely to influence an increasingly broad array of biological disciplines. Already, widespread implications can be envisioned for evolutionary biology, medicine, and agriculture; in some cases, these have already become reality. The large-scale comparison, and perhaps even manipulation, of genomes is a complex undertaking involving numerous empirical, analytical, and ethical issues. Undoubtedly, both important challenges and exciting discoveries lie ahead for genome biology.

REFERENCES

Aach J, Bulyk ML, Church GM, *et al.* 2001. Computational comparison of two draft sequences of the human genome. *Nature* 409: 856–859.

Adams MD, Celniker SE, Holt RA, *et al.* 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185–2195.

Adams MD, Dubnick M, Kerlavage AR, *et al.* 1992. Sequence identification of 2,375 human brain genes. *Nature* 355: 632– 634.

Adams MD, Kelley JM, Gocayne JD, *et al.* 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651– 1656.

Antequera F, Bird A. 1993. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci USA* 90: 11995–11999.

Aparicio S, Chapman J, Stupka E, *et al.* 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297: 1301–1310.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.

Archidiacono N, Storlazzi CT, Spalluto C, *et al.* 1998. Evolution of chromosome Y in primates. *Chromosoma* 107: 241–246.

Aristotle. 1953. *Generation of Animals*. Cambridge, MA: Harvard University Press; W. Heinemann Ltd.

Avarello R, Pedicini A, Caiulo A, *et al.* 1992. Evidence for an ancestral alphoid domain on the long arm of human chromosome 2. *Hum Gene* 89: 247–249.

Baldauf SL. 2003. The deep roots of eukaryotes. *Science* 300: 1703–1706.

Band MR, Larson JH, Rebeiz M, *et al.* 2000. An ordered comparative map of the cattle and human genomes. *Genome Res* 10: 1359–1368.

Bennett JW. 1997. White Paper: genomics for filamentous fungi. *Fungal Genet Biol* 21: 3–7.

Bennett MD, Leitch IJ, Price HJ, Johnston JS. 2003. Comparisons with *Caenorhabditis* (~ 100 Mb) and *Drosophila* (~ 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~ 157 Mb and thus ~ 25% larger than the Arabidopsis Genome Initiative estimate of ~ 125 Mb. *Ann Bot* 91: 547–557.

Benton MJ, Ayala FJ. 2003. Dating the tree of life. *Science* 300: 1698–1700.

Berget SM, Moore C, Sharp PA. 1977. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA* 74: 3171–3175.

Bernardi G. 1995. The human genome: organization and evolutionary history. *Annu Rev Genet* 29: 445–476.

Bernardi G. 2001. Misunderstandings about isochores. Part 1. *Gene* 276: 3–13.

Bernardi G, Mouchiroud D, Gautier C. 1988. Compositional patterns in vertebrate genomes: conservation and change in evolution. *J Mol Evol* 28: 7–18.

Bernardi G, Olofsson B, Filipski J, *et al.* 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228: 953–958.

Bird AP. 1986. CpG-rich islands and the function of DNA methylation. *Nature* 321: 209–213.

Blumenthal T, Gleason KS. 2003. *Caenorhabditis elegans* operons: form and function. *Nat Rev Genet* 4: 112–120.

Boguski MS. 1995. The turning point in genome research. *Trends Biochem Sci* 20: 295–296.

Bolshakov VN, Topalis P, Blass C, *et al.* 2002. A comparative genomic analysis of two distant diptera, the fruit fly, *Drosophila melanogaster*, and the malaria mosquito, *Anopheles gambiae*. *Genome Res* 12: 57–66.

Brenner S, Elgar G, Sandford R, *et al.* 1993. Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* 366: 265–268.

Brett D, Hanke J, Lehmann G, *et al.* 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* 474: 83–86.

Brickley P. 2002. A scrap over sequences, take two. *The Scientist* 16 (May 13): 55.

Bridges CB. 1935. Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. *J Hered* 26: 60–64.

Britten RJ. 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* 231: 1393–1398.

Britten RJ. 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc Natl Acad Sci USA* 99: 13633–13635.

Britten RJ, Kohne DE. 1968. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* 161: 529–540.

Britton-Davidian J, Catalan J, da Graca Ramalhinho M, *et al.* 2000. Rapid chromosomal evolution in island mice. *Nature* 403: 158.

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

Carlton JM, Angiuoli SV, Suh BB, *et al.* 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 419: 512–519.

Caspersson T, Zech L, Johansson C. 1970. Differential binding of alkylating fluorochromes in human chromosomes. *Exp Cell Res* 60: 315–319.

Celniker SE, Wheeler DA, Kronmiller B, *et al.*, 2002. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol* 3: research0079.1–0079.14.

Chargaff E. 1980. In praise of smallness: how we can return to small science? *Perspect Biol Med* 23: 370–385.

Check E. 2002. Priorities for genome sequencing leave macaques out in the cold. *Nature* 417: 473–474.

Chervitz SA, Aravind L, Sherlock G, *et al.* 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282: 2022–2028.

Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. *Science* 300: 1701–1703.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 12: 1–8.

Chowdhary BP, Raudsepp T, Fronicke L, and Scherthan H. 1998. Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH. *Genome Res* 8: 577–589.

Christophides GK, Zdobnov E, Barillas-Mury C, *et al.* 2002. Immunity-related genes and gene families in *Anopheles gambiae*. *Science* 298: 159–165.

Claverie JM. 2001. What if there are only 30,000 human genes? *Science* 291: 1255–1257.

Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. *Nature* 422: 835–847.

Collins FS, Morgan M, Patrinos A. 2003. The Human Genome Project: lessons from large-scale biology. *Science* 300: 286–290.

Courties C, Perasso R, Chrétiennot-Dinet M-J, *et al.* 1998. Phylogenetic analysis and genome size of *Ostreococcus tauri* (Chlorophyta, Prasinophyceae). *J Phycol* 34: 844–849.

Cyranoski D. 2002. Almost human. *Nature* 418: 910–912.

Deamer DW, Branton D. 2002. Characterization of nucleic acids by nanopore analysis. *Acc Chem Res* 35: 817–825.

Decottignies A, Sanchez-Perez I, Nurse P. 2003. *Schizosaccharomyces pombe* essential genes: a pilot study. *Genome Res* 13: 399–406.

Dehal P, Satou Y, Campbell RK, *et al.* 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 298: 2157–2167.

Dimopoulos G, Muller HM, Levashina EA, Kafatos FC. 2001. Innate immune defense against malaria infection in the mosquito. *Curr Opin Immunol* 13: 79–88.

Doolittle RF. 1986. *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*. Mill Valley, CA: University Science Books.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. *Science* 284: 2124–2129.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284: 601–603.

Dujon B. 1996. The yeast genome project: what did we learn? *Trends Genet* 12: 263–270.

Duret L, Mouchiroud D, Gautier C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol* 40: 308–317.

Dyson T. 1999. World food trends and prospects to 2025. *Proc Natl Acad Sci USA* 96: 5929–5936.

Easteal S, Collet C, Betty D. 1995. *The Mammalian Molecular Clock*. New York: R.G. Landes.

Ebersberger I, Metzler D, Schwarz C, Paabo S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* 70: 1490–1497.

Edwalds-Gilbert G, Veraldi KL, Milcarek C. 1997. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res* 25: 2547–2561.

Eichler EE, Sankoff D. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* 301: 793–797.

Ellegren H, Smith NG, Webster MT. 2003. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* 13: 562–568.

Enard W, Khaitovich P, Klose J, *et al.* 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* 296: 340–343.

Estop AM, Garver JJ, Egozcue J, *et al.* 1983. Complex chromosome homologies between the rhesus monkey (*Macaca mulatta*) and man. *Cytogenet Cell Genet* 35: 46–50.

Farr CJ, Goodfellow PN. 1992. Hidden messages in genetic maps. *Science* 258: 49.

Fickett JW, Wasserman WW. 2000. Discovery and modeling of transcriptional regulatory regions. *Curr Opin Biotechnol* 11: 19–24.

Ford CE, Hamerton JL. 1956. Chromosomes of man. *Nature* 178: 1020–1023.

Galagan JE, Calvo SE, Borkovich KA, *et al.* 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422: 859–868.

Gale MD, Devos KM. 1998. Plant comparative genetics after 10 years. *Science* 282: 656–659.

Galison PL. 1997. *Image and Logic: A Material Culture of Microphysics*. Chicago: University of Chicago Press.

Gardner MJ, Hall N, Fung E, *et al.* 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419: 498–511.

Gellin J, Echard G, Benne F, Gillois M. 1981. Pig gene mapping: PKM2-MPI-NP synteny. *Cytogenet Cell Genet* 30: 59–62.

George AJ. 2002. Is the number of genes we possess limited by the presence of an adaptive immune system? *Trends Immunol* 23: 351–355.

Gilbert W, Bodmer WF. 1986. Two cheers for human gene sequencing. *The Scientist* 1 (Oct. 20): 11.

Gionti M, Ristoratore F, Di Gregorio A, *et al.* 1998. Cihox5, a new *Ciona intestinalis* Hox-related gene, is involved in regionalization of the spinal cord. *Dev Genes Evol* 207: 515–523.

Goff SA, Ricke D, Lan TH, *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92–100.

Goffeau A, Barrell BG, Bussey H, *et al.,* 1996. Life with 6000 genes. *Science* 274: 546, 563–567.

Goujon P. 2001. *From Biotechnology to Genomes*. River Edge, NJ: World Scientific Pub. Co.

Graves JA. 1996. Mammals that break the rules: genetics of marsupials and monotremes. *Annu Rev Genet* 30: 233–260.

Gregory TR. 2001. *Animal Genome Size Database*. www.genomesize.com.

Gruskin KD, Smith TF. 1987. Molecular genetics and computer analyses. *Comput Appl Biosci* 3: 167–170.

Gu X, Wang Y, Gu J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet* 31: 205–209.

Gupta RS. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* 62: 1435–1491.

Hahn MW, Wray GA. 2002. The g-value paradox. *Evol Dev* 4: 73–75.

Hanke J, Brett D, Zastrow I, *et al.* 1999. Alternative splicing of human genes: more the rule than the exception? *Trends Genet* 15: 389–390.

Hartl DL. 2000. Molecular melodies in high and low C. *Nat Rev Genet* 1: 145–149.

Hawksworth D. 1991. The fungal dimension of biodiversity: magnitude, significance and conservation. *Mycol Res* 95: 641–655.

Hedges SB. 2002. The origin and evolution of model organisms. *Nat Rev Genet* 3: 838–849.

Hedges SB, Kumar S. 2002. Vertebrate genomes compared. *Science* 297: 1283–1285.

Hedges SB, Kumar S. 2003. Genomic clocks and evolutionary timescales. *Trends Genet* 19: 200–206.

Hedges SB, Chen H, Kumar S, *et al.* 2001. A genomic timescale for the origin of eukaryotes. *BMC Evol Biol* 1: 4.1–4.10.

Hill CA, Fox AN, Pitts RJ, *et al.* 2002. G protein-coupled receptors in *Anopheles gambiae*. *Science* 298: 176–178.

Holt RA, Subramanian GM, Halpern A, *et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129–149.

Houck ML, Kumamoto AT, Gallagher DS, Benirschke K. 2001. Comparative cytogenetics of the African elephant (*Loxodonta africana*) and Asiatic elephant (*Elephas maximus*). *Cytogenet Cell Genet* 93: 249–252.

Huang L, Guan RJ, Pardee AB. 1999. Evolution of transcriptional control from prokaryotic beginnings to eukaryotic complexities. *Crit Rev Eukaryot Gene Expr* 9: 175–182.

Huang Q, Fu YX, Boerwinkle E. 2003. Comparison of strategies for selecting single nucleotide polymorphisms for case/control association studies. *Hum Genet* 113: 253–257.

Hubbard T, Barker D, Birney E, *et al.* 2002. The Ensembl genome database project. *Nucleic Acids Res* 30: 38–41.

Hungerford DA, Chandra HS, Snyder RL. 1967. Somatic chromosomes of a black rhinoceros (*Diceros bicornis* Gray 1821). *Am Nat* 101: 357–358.

Hynes M. 2003. The *Neurospora crassa* genome opens up the world of filamentous fungi. *Genome Biol* 4: 271.1–271.4.

Iannuzzi L, Di Meo GP, Perucatti A, Bardaro T. 1998. ZOO-FISH and R-banding reveal extensive conservation of human chromosome regions in euchromatic regions of river buffalo chromosomes. *Cytogenet Cell Genet* 82: 210–214.

Ijdo J, Baldini A, Ward DC, *et al.* 1991. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc Natl Acad Sci USA* 88: 9051–9055.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.

Ioannidis JP. 2003. Genetic associations: false or true? *Trends Mol Med* 9: 135–138.

Jiang Z, Melville JS, Cao H, *et al.* 2002. Measuring conservation of contiguous sets of autosomal markers on bovine and porcine genomes in relation to the map of the human genome. *Genome* 45: 769–776.

John B, Miklos GLG. 1988. *The Eukaryote Genome in Development and Evolution*. London: Allen & Unwin.

Katinka MD, Duprat S, Cornillot E, *et al.* 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414: 450–453.

Kellis M, Patterson N, Endrizzi M, *et al.* 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241–254.

Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Korf I, Flicek P, Duan D, Brent MR. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* 17 (Suppl. 1): S140–S148.

Krishnan G. 1975. Nature of tunicin and its interaction with other chemical components of the tunic of the ascidian, *Polyclinum madrasensis* Sebastian. *Indian J Exp Biol* 13: 172–176.

Kruglyak L, Nickerson DA. 2001. Variation is the spice of life. *Nat Genet* 27: 234–236.

Krumlauf R, Marzluf GA. 1980. Genome organization and characterization of the repetitive and inverted repeat DNA sequences in *Neurospora crassa*. *J Biol Chem* 255: 1138–1145.

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. *Nature* 392: 917–920.

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci USA* 99: 803–808.

Kumar S, Gadagkar SR, Filipski A, Gu X. 2001. Determination of the number of conserved chromosomal segments between species. *Genetics* 157: 1387–1395.

Kumar S, Mitnik C, Valente G, Floyd-Smith G. 2000. Expansion and molecular evolution of the interferon-induced 2′-5′ oligoadenylate synthetase gene family. *Mol Biol Evol* 17: 738–750.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* 424: 147–151.

Levy HP, Schultz RA, Cohen MM. 1992. Comparative gene mapping in the species *Muntiacus muntjac*. *Cytogenet Cell Genet* 61: 276–281.

Li W-H. 2002. Are isochore sequences homogeneous? *Gene* 300: 129–139.

Li W-H, Tanimura M. 1987. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* 326: 93–96.

Link AJ, Olson MV. 1991. Physical map of the *Saccharomyces cerevisiae* genome at 110–kilobase resolution. *Genetics* 127: 681–698.

Lo N, Tokuda G, Watanabe H, *et al*. 2000. Evidence from multiple gene sequences indicates that termites evolved from wood-feeding cockroaches. *Curr Biol* 10: 801–804.

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C, Moreira D. 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* 409: 603–607.

Macer D. 1991. Whose genome project? *Bioethics* 5: 183–211.

Mannhaupt G, Montrone C, Haase D, *et al.* 2003. What's in the genome of a filamentous fungus? Analysis of the *Neurospora* genome sequence. *Nucleic Acids Res* 31: 1944–1954.

March R. 2000. Pharmacogenomics: the genomics of drug response. *Yeast* 17: 16–21.

Margulis L. 1996. Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc Natl Acad Sci USA* 93: 1071–1076.

Martin SL, Blackmon BP, Rajagopalan R, *et al.* 2002. MagnaportheDB: a federated solution for integrating physical and genetic map data with BAC end derived sequences for the rice blast fungus *Magnaporthe grisea*. *Nucleic Acids Res* 30: 121–124.

Matassi G, Sharp PM, Gautier C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr Biol* 9: 786–791.

McCarthy JJ, Hilfiker R. 2000. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat Biotechnol* 18: 505–508.

Messing J. 2001. Do plants have more genes than humans? *Trends Plant Sci* 6: 195–196.

Mironov AA, Fickett JW, Gelfand MS. 1999. Frequent alternative splicing of human genes. *Genome Res* 9: 1288–1293.

Moon-van der Staay SY, De Wachter R, Vaulot D. 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* 409: 607–610.

Mouchiroud D, Gautier C, Bernardi G. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. *J Mol Evol* 40: 107–113.

Mounsey A, Bauer P, Hope IA. 2002. Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. *Genome Res* 12: 770–775.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420: 520–562.

Mullis K, Faloona F, Scharf S, *et al.* 1986. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* 51 (Pt. 1): 263–273.

Mural RJ, Adams MD, Myers EW, *et al.* 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* 296: 1661–1671.

Murphy WJ, Stanyon R, O'Brien SJ. 2001. Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol* 2: reviews0005.1–0005.8.

Murphy WJ, Sun S, Chen ZQ, *et al.* 1999. Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Res* 9: 1223–1230.

Nadeau JH. 1989. Maps of linkage and synteny homologies between mouse and man. *Trends Genet* 5: 82–86.

Nakashima K, Yamada L, Satou Y, *et al*. 2004. The evolutionary origin of animal cellulose synthase. *Dev Genes Evol* 214: 81–88.

Nash WG, O'Brien SJ. 1982. Conserved regions of homologous G-banded chromosomes between orders in mammalian evolution: carnivores and primates. *Proc Natl Acad Sci USA* 79: 6631–6635.

Nash WG, Wienberg J, Ferguson-Smith MA, *et al.* 1998. Comparative genomics: tracking chromosome evolution in the family ursidae using reciprocal chromosome painting. *Cytogenet Cell Genet* 83: 182–192.

Nei M. 1969. Gene duplication and nucleotide substitution in evolution. *Nature* 221: 40–42.

Nicholl DST. 2002. *An Introduction to Genetic Engineering.* 2nd edition. Cambridge: Cambridge University Press.

Niimura Y, Gojobori T. 2002. *In silico* chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. *Proc Natl Acad Sci USA* 99: 797–802.

Nilsen TW. 1989. Trans-splicing in nematodes. *Exp Parasitol* 69: 413–416.

Normile D. 2001. Gene expression differs in human and chimp brains. *Science* 292: 44–45.

O'Brien SJ, Nash WG. 1982. Genetic mapping in mammals: chromosome map of domestic cat. *Science* 216: 257–265.

O'Brien SJ, Stanyon R. 1999. Phylogenomics: ancestral primate viewed. *Nature* 402: 365–366.

O'Brien SJ, Eisenberg JF, Miyamoto M, *et al.* 1999. Genome maps 10. Comparative genomics. Mammalian radiations. Wall chart. *Science* 286: 463–478.

O'Brien SJ, Menotti-Raymond M, Murphy WJ, *et al.* 1999. The promise of comparative genomics in mammals. *Science* 286: 458–462, 479–481.

Ohno S. 1970. *Evolution by Gene Duplication*. Berlin: Springer-Verlag.
Ohno S. 1972. So much "junk" DNA in our genome. In: Smith HH ed. *Evolution of Genetic Systems*. New York: Gordon and Breach, 366–370.
Okubo K, Hori N, Matoba R, *et al.* 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat Genet* 2: 173–179.
Oliver SG, van der Aart QJ, Agostoni-Carbone ML, *et al.* 1992. The complete DNA sequence of yeast chromosome III. *Nature* 357: 38–46.
Olson M, Hood L, Cantor C, Botstein D. 1989. A common language for physical mapping of the human genome. *Science* 245: 1434–1435.
Olson MV, Varki A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat Rev Genet* 4: 20–28.
Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* 284: 604–607.
Pal C, Hurst LD. 2000. The evolution of gene number: are heritable and non-heritable errors equally important? *Heredity* 84: 393–400.
Pardue ML, Gall JG. 1970. Chromosomal localization of mouse satellite DNA. *Science* 168: 1356–1358.
Pearson H. 2003. Geneticists play the numbers game in vain. *Nature* 423: 576.
Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet* 17: 23–28.
Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* 384: 346–349.
Potrykus I. 2001. Golden rice and beyond. *Plant Physiol* 125: 1157–1161.
Pruitt KD, Maglott DR. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29: 137–140.
Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428: 493–521.
Raudsepp T, Fronicke L, Scherthan H, *et al.* 1996. Zoo-FISH delineates conserved chromosomal segments in horse and man. *Chromosome Res* 4: 218–225.
Richard F, Messaoudi C, Bonnet-Garnier A, *et al.* 2003. Highly conserved chromosomes in an Asian squirrel (*Menetes berdmorei*, Rodentia: Sciuridae) as demonstrated by ZOO-FISH with human probes. *Chromosome Res* 11: 597–603.
Rivera MC, Jain R, Moore JE, Lake JA. 1998. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* 95: 6239–6244.
Rizzon C, Marais G, Gouy M, Biemont C. 2002. Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome. *Genome Res* 12: 400–407.
Roberts L. 1987. Agencies vie over Human Genome Project. *Science* 237: 486–488.
Roest Crollius H, Jaillon O, Bernot A, *et al.* 2000a. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat Genet* 25: 235–238.
Roest Crollius H, Jaillon O, Bernot A, *et al.* 2002. Genome-wide comparisons between human and *Tetraodon*. *Ernst Schering Res Found Workshop*: 11–29.
Roest Crollius H, Jaillon O, Dasilva C, *et al.* 2000b. Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res* 10: 939–949.
Russo E. 2001. Behind the sequence. *The Scientist* 15 (Mar. 5): 1.
Saccone S, de Sario A, Wiegant J, *et al.* 1993. Correlations between isochores and chromosomal bands in the human genome. *Proc Natl Acad Sci USA* 90: 11929–11933.
Sachidanandam R, Weissman D, Schmidt SC, *et al.* 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928–933.
Sanger F, Air GM, Barrell BG, *et al.* 1977. Nucleotide-sequence of bacteriophage Phi X174 DNA. *Nature* 265: 687–695.
Scherthan H, Cremer T, Arnason U, *et al.* 1994. Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nat Genet* 6: 342–347.

Schwartz S, Kent WJ, Smit A, *et al.* 2003. Human-mouse alignments with BLASTZ. *Genome Res* 13: 103–107.

Seabright M. 1971. A rapid banding technique for human chromosomes. *Lancet* 2: 971–972.

Sebat J, Lakshmi B, Troge J, *et al.* 2004. Large-scale copy number polymorphism in the human genome. *Science* 305: 525–528.

Selker EU. 1990. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu Rev Genet* 24: 579–613.

Seltman H, Roeder K, Devlin B. 2003. Evolutionary-based association analysis using haplotype data. *Genet Epidemiol* 25: 48–58.

Sibley CG, Ahlquist JE. 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J Mol Evol* 20: 2–15.

Sibley CG, Ahlquist JE. 1987. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *J Mol Evol* 26: 99–121.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, *et al.* 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423: 825–837.

Spieth J, Brooke G, Kuersten S, *et al.* 1993. Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell* 73: 521–532.

Stanyon R, Yang F, Cavagna P, *et al.* 1999. Reciprocal chromosome painting shows that genomic rearrangement between rat and mouse proceeds ten times faster than between humans and cats. *Cytogenet Cell Genet* 84: 150–155.

Stephens JC. 1999. Single-nucleotide polymorphisms, haplotypes, and their relevance to pharmacogenetics. *Mol Diagn* 4: 309–317.

Strauss EC, Kobori JA, Siu G, Hood LE. 1986. Specific-primer-directed DNA sequencing. *Anal Biochem* 154: 353–360.

Sturtevant AH. 1913. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J Exp Zool* 14: 43–59.

Suedmeyer WK, Houck ML, Kreeger J. 2003. Klinefelter syndrome (39 XXY) in an adult Siberian tiger (*Panthera tigris altaica*). *J Zoo Wildl Med* 34: 96–99.

Sulston JE, Schierenberg E, White JG, Thomson JN. 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev Biol* 100: 64–119.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci USA* 101: 11030–11035.

Tavare S, Marshall CR, Will O, *et al.* 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* 416: 726–729.

Taylor JG, Choi EH, Foster CB, Chanock SJ. 2001. Using genetic variation to study human disease. *Trends Mol Med* 7: 507–512.

Thiessen KM, Lalley PA. 1986. New gene assignments and syntenic groups in the baboon (*Papio papio*). *Cytogenet Cell Genet* 42: 19–23.

Thiessen KM, Lalley PA. 1987. Gene assignments and syntenic groups in the sacred baboon (*Papio hamadryas*). *Cytogenet Cell Genet* 44: 82–88.

Thomas CA. 1971. The genetic organization of chromosomes. *Annu Rev Genet* 5: 237–256.

Thomas JW, Touchman JW, Blakesley RW, *et al.* 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424: 788–793.

Tjio JH, Levan A. 1956. The chromosome number of Man. *Hereditas* 42: U1–6.

Venter JC, Adams MD, Myers EW, *et al.* 2001. The sequence of the human genome. *Science* 291: 1304–1351.

Volti R. 2001. *Society and Technological Change*. New York: Worth Publishers.

Wade N. 2002. Thrown aside, genome pioneer plots a rebound. *New York Times* April 30.

Wakefield MJ, Graves JA. 1996. Comparative maps of vertebrates. *Mamm Genome* 7: 715–716.

Wiehe T, Gebauer-Jung S, Mitchell-Olds T, Guigo R. 2001. SGP-1: prediction and validation of homologous genes based on sequence alignments. *Genome Res* 11: 1574–1583.

Wienberg J, Stanyon R. 1995. Chromosome painting in mammals as an approach to comparative genomics. *Curr Opin Genet Dev* 5: 792–797.

Wildman DE, Uddin M, Liu G, *et al.* 2003. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus *Homo*. *Proc Natl Acad Sci USA* 100: 7181–7188.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2: 333–341.

Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* 337: 283–285.

Wong GK, Wang J, Tao L, *et al.* 2002. Compositional gradients in Gramineae genes. *Genome Res* 12: 851–856.

Wood VR, Gwilliam MA, Rajandream M, *et al.* 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* 415: 871–880.

Wu CI, Li WH. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci USA* 82: 1741–1745.

Ye X, Al-Babili S, Kloti A, *et al.* 2000. Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science* 287: 303–305.

Yu J, Hu S, Wang J, *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79–92.

Yunis JJ, Prakash O. 1982. The origin of man: a chromosomal pictorial legacy. *Science* 215: 1525–1530.

Zdobnov EM, von Mering C, Letunic I, *et al.* 2002. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* 298: 149–159.

Zhang J. 2000. Protein-length distributions for the three domains of life. *Trends Genet* 16: 107–109.

Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17: 821–828.