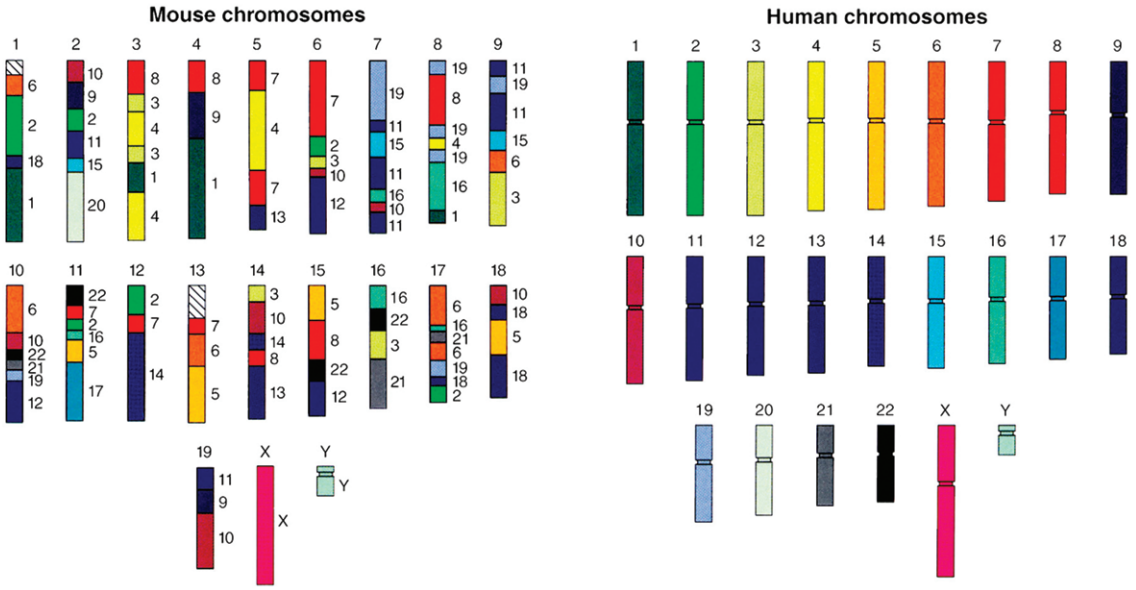
CHAPTER 5



Genome Sequences and Evolution

CHAPTER OUTLINE

5.1 Introduction

Since the first complete organismal genomes were sequenced in 1995, both the speed and range of sequencing have greatly improved. The first genomes to be sequenced were small bacterial genomes of less than 2 megabase (Mb) in size. By 2002, the human genome of about 3,200 Mb had been sequenced. Genomes have now been sequenced from a wide range of organisms, including bacteria, archaeans, yeasts, and other unicellular eukaryotes, plants, and animals, including worms, flies, and mammals.

Perhaps the single most important piece of information provided by a genome sequence is the number of genes. (See the chapter titled *The Content of the Genome* for a discussion about the difficulties of defining a gene; for our purposes, the term *gene* refers to a DNA sequence transcribed to a functional RNA molecule.) *Mycoplasma genitalium,* a free-living parasitic bacterium, has the smallest known genome of any organism, with about only 470 genes. The genomes of free-living bacteria have from 1,700 to 7,500 genes. Archaean genomes have a smaller range of 1,500 to 2,700 genes. The smallest unicellular eukaryotic genomes have about 5,300 genes. Nematode worms and fruit flies have roughly 21,700 and 17,000 genes, respectively. Surprisingly, the number rises only to 20,000 to 25,000 for mammalian genomes.

FIGURE 5.1 summarizes the minimum number of genes found in six groups of organisms. A cell requires a minimum of about 500 genes, a free-living cell requires about 1,500 genes, a eukaryotic cell requires more than 5,000 genes, a multicellular organism requires more than 10,000 genes, and an organism with a nervous system requires more than 13,000 genes. Many species have more than the minimum number of genes required, so the number of genes can vary widely, even among closely related species.

Within prokaryotes and unicellular eukaryotes, most genes are unique. Within multicellular eukaryotic genomes, however, some genes are arranged into families of related members. Of course, some genes are unique (meaning the family has only one member), but many belong to families with 10 or more members. The number of different families may be a better indication of the overall complexity of the organism than the number of genes.

Some of the most insightful information comes from comparing genome sequences. The growing number of complete genome sequences has provided valuable opportunities to study genome structure and organization. As genome sequences of related species become available, there are opportunities to compare not only individual gene differences but also large-scale genomic differences in aspects such as gene distribution, the proportions of nonrepetitive and repetitive DNA and their functional potentials, and the number of copies of repetitive sequences. By making these comparisons, we can gain insight into the historical genetic events that have shaped the genomes of individual species and of the adaptive and nonadaptive forces at work following these events. For example, with the sequences now available for both the human and chimpanzee genomes, it is possible to begin to address some of the questions about what makes humans unique.

FIGURE 5.1 The minimum gene number required for any type of organism increases with its complexity.

(a) Photo of intracellular bacterium courtesy of Gregory P. Henderson and Grant J. Jensen, California Institute of Technology.
(b) Courtesy of Rocky Mountain Laboratories, NIAID, NIH.
(c) Courtesy of Eishi Noguchi, Drexel University College of Medicine.
(d) Courtesy of Carolyn B. Marks and David H. Hall, Albert Einstein College of Medicine, Bronx, NY.
(e) Courtesy of Keith Weller/USDA.
(f) © Photodisc.

The availability of the genome sequences of genetic "model organisms" (e.g., *Escherichia coli*, yeast, *Drosophila*, *Arabidopsis*, and humans) in the late 1990s and early 2000s allowed comparisons between major taxonomic groups such as prokaryote versus eukaryote, animal versus plant, or vertebrate versus invertebrate. More recently, data from multiple genomes within lower-level taxonomic groups (classes down to genera) have allowed closer examination of genome evolution. Such comparisons have the advantage of highlighting changes that have occurred much more recently and are less obscured by additional changes, such as multiple mutations at the same site. In addition, evolutionary events specific to a taxonomic group can be explored. For example, human–chimpanzee comparisons can provide information about primate-specific genome evolution, particularly when compared with an **outgroup** (a species that is less closely related, but close enough to show substantial similarity) such as the mouse. One recent milestone in this field of **comparative**

genomics is the completion of genome sequences of nearly 30 species of the genus *Drosophila*. These types of fine-scale comparisons will continue as more genomes from the same species become available.

What questions can be addressed by comparative genomics? First, the evolution of individual genes can be explored by comparing genes descended from a common ancestor. To some extent, the evolution of a genome is a result of the evolution of a collection of individual genes, so comparisons of homologous sequences within and between genomes can help to answer questions about the adaptive (i.e., naturally selected) and nonadaptive changes that occur to these sequences. The forces that shape coding sequences are usually quite different from those that affect noncoding regions (e.g., introns, untranslated regions, or regulatory regions) of the same gene: Coding and regulatory regions more directly influence phenotype (though in different ways), making selection a more important aspect of their evolution than for noncoding regions. Second, researchers can also explore the mechanisms that result in changes in the structure of the genome, such as gene duplication, expansion and contraction of repetitive arrays, transposition, and polyploidization.

5.2 Prokaryotic Gene Numbers Range Over an Order of Magnitude

KEY CONCEPT

- The minimum number of genes for a parasitic prokaryote is about 500; for a free-living nonparasitic prokaryote, it is about 1,500.

Large-scale efforts have now led to the sequencing of many genomes. The range of known genome sizes (as summarized in **TABLE 5.1**) extends from the 0.6×10^6 base pairs (bp) of a mycoplasma to the 3.3×10^9 bp of the human genome, and includes several important model organisms, such as yeasts, the fruit fly, and a nematode worm. Many plant genomes are much larger; the genome of bread wheat (*Triticum aestivum* L.) is 17 gigabases (Gb; five times the size of the human genome), though it should be noted that the species is hexaploid.

The sequences of the genomes of prokaryotes show that most of the DNA (typically 85% to 90%) encodes RNA or polypeptide. **FIGURE 5.2** shows that the range of prokaryotic genome sizes is an order of magnitude and that the genome size is proportional to the number of genes. The typical gene averages just under 1,000 bp in length.

All of the prokaryotes with genome sizes below 1.5 Mb are parasites—they can live within a eukaryotic host that provides them with small molecules. Their genome sizes suggest the minimum number of functions required for a cellular organism. All classes of genes are reduced in number compared to prokaryotes with larger genomes, but the most significant reduction is in loci that encode enzymes involved with metabolic functions (which are largely provided by the host cell) and with regulation of gene expression. *Mycoplasma genitalium* has the smallest genome, with about 470 genes.

Archaeans have biological properties that are intermediate between those of other prokaryotes and those of eukaryotes, but their genome sizes and gene numbers fall in the same range as those of bacteria. Their genome sizes vary from 1.5 to 3 Mb, corresponding to 1,500 to 2,700 genes. *Methanococcus jannaschii* is a methane-producing species that lives under high pressure and temperature. Its total gene number is similar to that of *Haemophilus influenzae*, but fewer of its genes can be identified on the basis of comparison with genes known in other organisms. Its apparatus for gene expression resembles that of eukaryotes more than that of prokaryotes, but its apparatus for cell division better resembles that of prokaryotes.

The genomes of archaea and the smallest free-living bacteria suggest the minimum number of genes required to make a cell able to function independently in its environment. The smallest archaeal genome has approximately 1,500 genes. The free-living nonparasitic bacterium with the smallest known genome is the thermophile *Aquifex aeolicus*, with a 1.5-Mb genome and 1,512 genes. A "typical" Gram-negative bacterium, *H. influenzae*, has 1,743 genes, the average size of which is about 900 bp. So, we can conclude that about 1,500 genes are required by an exclusively free-living organism.

Prokaryotic genome sizes extend over about an order of magnitude, from 0.6 Mb to less than 8 Mb. As expected, the larger genomes have more genes. The prokaryotes with the largest genomes, *Sinorhizobium meliloti* and *Mesorhizobium loti*, are nitrogen-fixing bacteria that live on plant roots. Their genome sizes (about 7 Mb) and total gene numbers (more than 7,500) are similar to those of yeasts.

The size of the genome of *E. coli* is in the middle of the range for prokaryotes. The common laboratory strain has 4,288 genes, with an average length of about 950 bp and an average separation between genes of 118 bp. There can be quite significant differences between strains, however. The known extremes in genome size among strains of *E. coli* are from 4.6 Mb with 4,249 genes to 5.5 Mb with 5,361 genes.

We still do not know the functions of all of these genes; functions have been identified for more than 80% of the genes. In most of these genomes, about 60% of the genes can be identified on the basis of homology with known genes in other species. These genes fall approximately equally into classes whose products function in metabolism, cell structure or transport of components, and gene expression and its regulation. In virtually every genome, 20% of the genes have not yet been ascribed any function. Many of these genes can be found in related organisms, which implies that they have a conserved function.

There has been some emphasis on sequencing the genomes of pathogenic bacteria, given their medical significance. An important insight into the nature of pathogenicity has been provided by the demonstration that **pathogenicity islands** are a characteristic feature of their genomes. These

TABLE 5.1 Genome sizes and gene numbers are known from complete sequences for several organisms. Lethal loci are estimated from genetic data.

Species	Genome Size (Mb)	Genes	Lethal Loci
Mycoplasma genitalium	0.58	470	~300
Rickettsia prowazekii	1.11	834	
Haemophilus influenzae	1.83	1,743	
Methanococcus jannaschi	1.66	1,738	
Bacillus subtilis	4.2	4,100	
Escherichia coli	4.6	4,288	1,800
Saccharomyces cerevisiae	13.5	6,043	1,090
Schizosaccharomyces pombe	12.5	4,929	
Arabidopsis thaliana	119	25,498	
Oryza sativa	466	~30,000	
Drosophila melanogaster	165	13,601	3,100
Caenorhabditis elegans	97	18,424	
Homo sapiens	3,200	~20,000	

are large regions (from 10 to 200 kb) that are present in the genomes of pathogenic species but absent from the genomes of nonpathogenic variants of the same or related species. Their GC content often differs from that of the rest of the genome, and it is likely that these regions are spread among bacteria by a process of **horizontal transfer**. For example, the bacterium that causes anthrax (*Bacillus anthracis*) has two large plasmids (extrachromosomal DNA molecules), one of which has a pathogenicity island that includes the gene encoding the anthrax toxin.

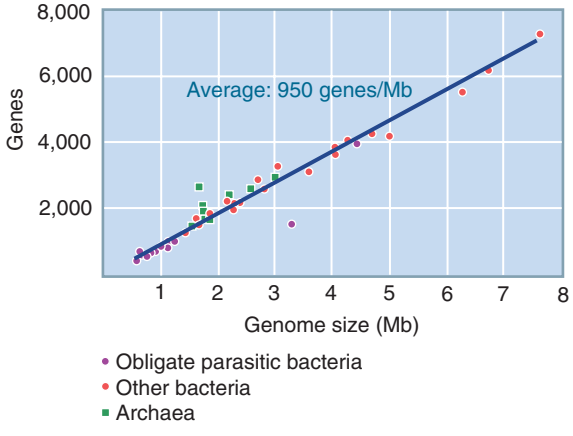


FIGURE 5.2 The number of genes in bacterial and archaeal genomes is proportional to genome size.

5.3 Total Gene Number Is Known for Several Eukaryotes

KEY CONCEPT

- There are 6,000 genes in yeast; 21,700 in a nematode worm; 17,000 in a fly; 25,000 in the small plant *Arabidopsis;* and probably 20,000 to 25,000 in mammals.

As we look at eukaryotic genomes, the relationship between genome size and gene number is weaker than that of prokaryotes. The genomes of unicellular eukaryotes fall in the same size range as the largest bacterial genomes. Multicellular eukaryotes have more genes, but the number does not correlate well with genome size, as can be seen in **FIGURE 5.3**.

The most extensive data for unicellular eukaryotes are available from the sequences of the genomes of the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*.

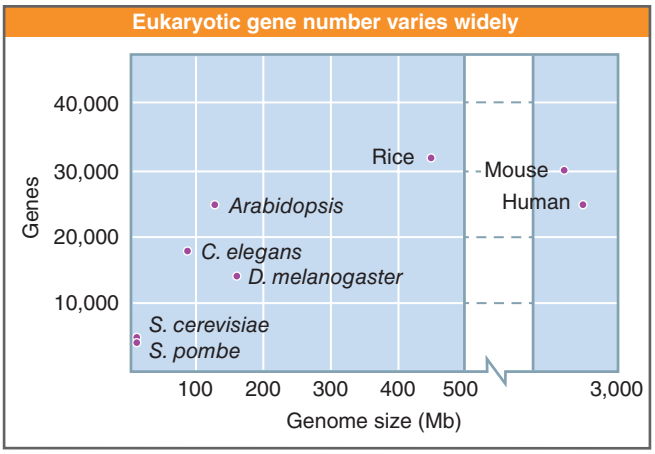


FIGURE 5.3 The number of genes in a eukaryote varies from 6,000 to 32,000 but does not correlate with the genome size or the complexity of the organism.

FIGURE 5.4 summarizes the most important features. The yeast genomes of 13.5 Mb and 12.5 Mb have roughly 6,000 and 5,000 genes, respectively. The average open reading frame (ORF) is about 1.4 kb, so that about 70% of the genome is occupied by coding regions. The major difference between them is that only 5% of *S. cerevisiae* genes have introns, compared to 43% in *S. pombe*. The density of genes is high; organization is generally similar, although the spaces between genes are a bit shorter in *S. cerevisiae*. About half of the genes identified by the sequence were either known previously or related to known genes. The remaining genes were previously unknown, which gives some indication of the number of new types of genes that can be discovered by sequence analysis.

The identification of long reading frames on the basis of sequence is quite accurate. However, ORFs encoding fewer than 100 amino acids cannot be identified solely by sequence because of the high occurrence of false positives. Analysis of gene expression suggests that only about 300 of 600 such ORFs in *S. cerevisiae* are likely to be functional genes.

A powerful way to validate gene structure is to compare sequences in closely related species: If a gene is functional, it is likely to be conserved. Comparisons between the sequences of four closely related yeast species suggest that 503 of the genes originally identified in *S. cerevisiae* do not have orthologs in the other species and therefore should not be considered functional genes. This reduces the total estimated gene number for *S. cerevisiae* to 5,726.

The genome of *Caenorhabditis elegans* varies between regions rich in genes and regions in which genes are more sparsely distributed. The total sequence contains about 21,700 genes. Only about 42% of the genes have suspected orthologs outside Nematoda.

The fruit fly genome is larger than the nematode worm genome, but there are fewer genes in the various species for which complete genome information is available (ranging from estimates of 14,400 in *Drosophila melanogaster* to 17,300 in *Drosophila persimilis*). The number of different transcripts is somewhat larger as the result of alternative splicing. We do not understand why *C. elegans*—arguably, a similarly complex organism—has 30% more genes than the fly, but it might be because *C. elegans* has a larger average number of genes per gene family than does *D. melanogaster*, so the numbers of *unique* genes of the two species are more similar. A comparison of 12 *Drosophila* genomes reveals that there can be a fairly large range of gene number (about 20%) among closely related species. In some cases, there are several thousand genes that are species-specific. This forcefully emphasizes the lack of an exact relationship between gene number and complexity of the organism.

The plant *Arabidopsis thaliana* has a genome size intermediate between those of the worm and the fly, but has a larger gene number (about 25,000) than either. This again shows the lack of a clear relationship between complexity and gene number and also emphasizes a special quality of plants, which can have more genes (due to ancestral duplications) than animal cells (except for vertebrates; see the section *Genome Duplication Has Played a Role in Plant and Vertebrate Evolution* later in this chapter). A majority of the *Arabidopsis* genome is found in duplicated segments, suggesting that there was an ancient doubling of the genome (to result in a tetraploid). Only 35% of *Arabidopsis* genes are present as single copies.

The genome of rice (*Oryza sativa*) is about 43 times larger than that of *Arabidopsis*, but the number of genes is only about 25% larger, estimated at 32,000. Repetitive DNA occupies 42% to 45% of the genome. More than 80% of the genes found in *Arabidopsis* are also found in rice. Of these common genes, about 8,000 are found in *Arabidopsis* and rice but not in any of the bacterial or animal genomes that have been sequenced. This is probably the set of genes that encodes plant-specific functions, such as photosynthesis.

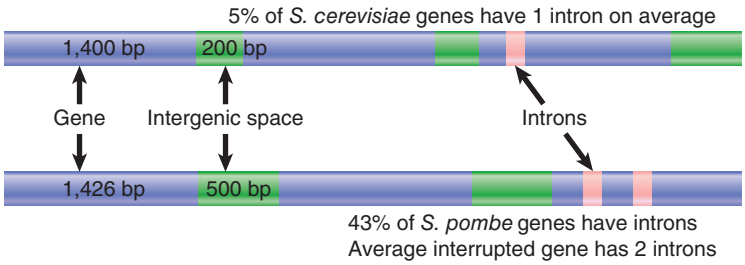


FIGURE 5.4 The *S. cerevisiae* genome of 13.5 Mb has 6,000 genes, almost all uninterrupted. The *S. pombe* genome of 12.5 Mb has 5,000 genes, almost half having introns. Gene sizes and spacing are fairly similar.

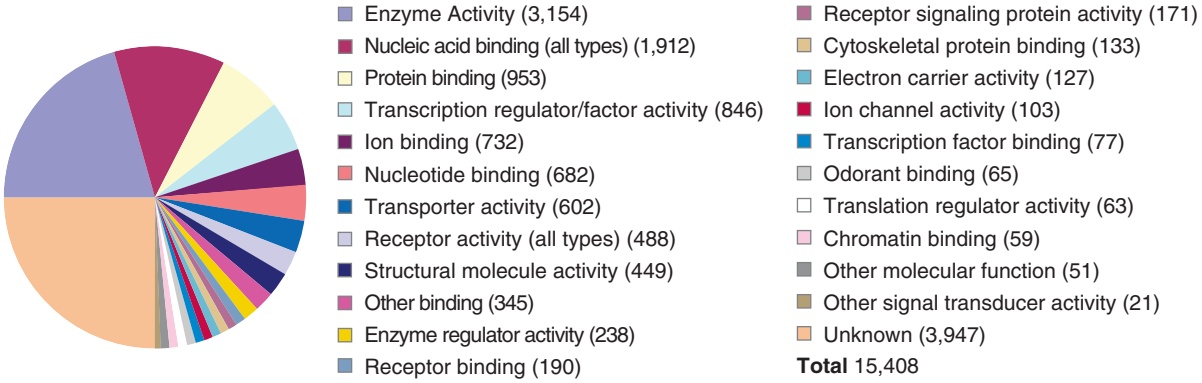


FIGURE 5.5 Functions of *Drosophila* genes based on comparative genomics of 12 species. The functions of about a quarter of the genes of *Drosophila* are unknown.

Data from: *Drosophila* 12 Genomes Consortium, 2007. "Evolution of genes and genomes on the *Drosophila* phylogeny," *Nature* 450: 203–218.

From 12 sequenced *Drosophila* genomes, we can form an impression of how many genes are devoted to each type of function. (In 2016, there are 15 additional complete *Drosophila* species genome sequences available, but these have not yet been fully analyzed.) **FIGURE 5.5** breaks down the functions into different categories. Among the genes that are identified, we find more than 3,000 enzymes, about 900 transcription factors, and about 700 transporters and ion channels. About a quarter of the genes encode products of unknown function.

Eukaryotic polypeptide sizes are greater than those of prokaryotes. The archaean *M. jannaschii* and bacterium *E. coli* have average polypeptide lengths of 287 and 317 amino acids, respectively, whereas *S. cerevisiae* and *C. elegans* have average polypeptide lengths of 484 and 442 amino acids, respectively. Large polypeptides (with more than 500 amino acids) are rare in prokaryotes but comprise a significant component (about one-third) in eukaryotes. The increase in length is due to the addition of extra domains, with each domain typically constituting 100 to 300 amino acids. However, the increase in polypeptide size is responsible for only a very small part of the increase in genome size.

Another insight into gene number is obtained by counting the number of expressed protein-coding genes. If we relied upon the estimates of the number of different messenger RNA (mRNA) species that can be counted in a cell, we would conclude that the average vertebrate cell expresses roughly 10,000 to 20,000 genes. The existence of significant overlaps between the mRNA populations in different cell types would suggest that the total expressed gene number for the organism should be within the same order of magnitude. The estimate for the total human gene number of about 20,000 (see the section *The Human Genome Has Fewer Genes Than Originally Expected* later in this chapter) would imply that a significant proportion of the total gene number is actually expressed in any particular cell.

Eukaryotic genes are transcribed individually, with each gene producing a **monocistronic mRNA**. There is only one general exception to this rule: In the genome of *C. elegans,* about 15% of the genes are organized into units transcribed to polycistronic mRNAs, which are associated with the use of *trans*-splicing to allow expression of the downstream genes in these units (see the *RNA Splicing and Processing* chapter).

5.4 How Many Different Types of Genes Are There?

KEY CONCEPTS

- The sum of the number of unique genes and the number of gene families is an estimate of the number of types of genes.
- The minimum size of the proteome can be estimated from the number of types of genes.

Some genes are unique; others belong to families in which the other members are related (but not usually identical). The proportion of unique genes declines, and the proportion of genes in families increases, with increasing genome size. Some genes are present in more than one copy or are related to one another, so the number of different types of genes is less than the total number of genes. We can divide the total number of genes into sets that have related members, as defined by comparing their exons. (A gene family arises by repeated duplication of an ancestral gene followed by accumulation of changes in sequence among the copies. Most often the members of a family are similar but not identical.) The number of types of genes is calculated by adding the number of unique genes (for which there is no other related gene at all) to the numbers of families that have two or more members.

FIGURE 5.6 compares the total number of genes with the number of distinct families in each of six genomes. In bacteria, most genes are unique, so the number of distinct families is close to the total gene number. The situation is different even in the unicellular eukaryote *S. cerevisiae*, for which there is a significant proportion of repeated genes. The most striking effect is that the number of genes increases

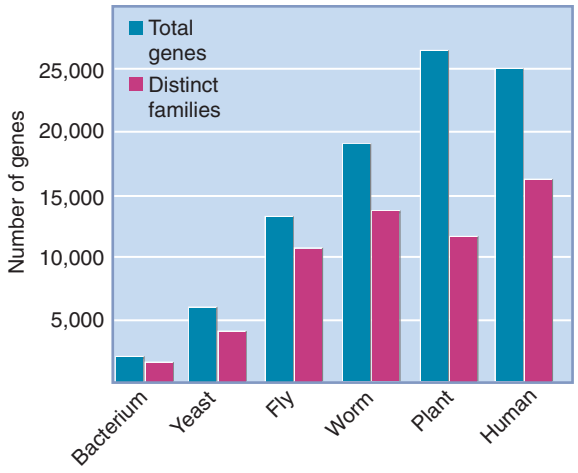


FIGURE 5.6 Many genes are duplicated, and as a result the number of different gene families is much smaller than the total number of genes. This histogram compares the total number of genes with the number of distinct gene families.

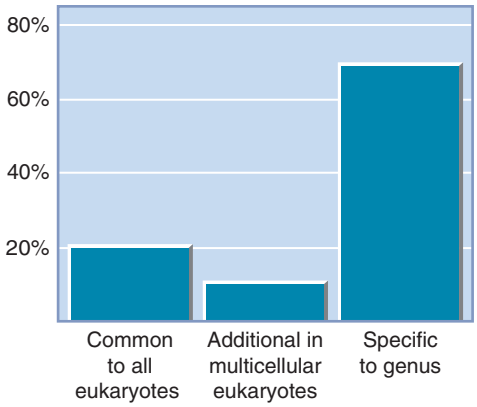


FIGURE 5.7 The fruit fly genome can be divided into genes that are (probably) present in all eukaryotes, additional genes that are (probably) present in all multicellular eukaryotes, and genes that are more specific to subgroups of species that include flies.

quite sharply in the multicellular eukaryotes, but the number of gene families does not change much.

TABLE 5.2 shows that the proportion of unique genes drops sharply with increasing genome size. When there are gene families, the number of members in a family is small in bacteria and unicellular eukaryotes, but is large in multicellular eukaryotes. Much of the extra genome size of *Arabidopsis* is due to families with more than four members.

If every gene is expressed, the total number of genes will account for the total number of polypeptides required by the organism (the proteome). However, there are two factors that can cause the proteome to be different from the total gene number. First, genes can be duplicated, and, as a result, some of them encode the same polypeptide (although it might be expressed at a different time or in a different type of cell) and others might encode related polypeptides that also play the same role at different times or in different cell types. Second, the proteome can be larger than the number of genes because some genes can produce more than one polypeptide by alternative splicing or other means.

What is the core proteome—the basic number of the different types of polypeptides in the organism? Although difficult to estimate because of the possibility of alternative splicing, a minimum estimate is provided by the number of gene families, ranging from 1,400 in bacteria, to about 4,000 in yeast, to 11,000 for the fly, to 14,000 for the worm.

What is the distribution of the proteome by type of protein? The 6,000 proteins of the yeast proteome include 5,000 soluble proteins and 1,000 transmembrane proteins. About half of the proteins are cytoplasmic, a quarter are in the nucleus, and the remainder are split between the mitochondrion and the endoplasmic reticulum (ER)/Golgi system.

How many genes are common to all organisms (or to groups such as bacteria or multicellular eukaryotes), and how many are specific to lower-level taxonomic groups? **FIGURE 5.7** shows the comparison of fly genes to those of the worm (another multicellular eukaryote) and yeast (a unicellular eukaryote). Genes that encode corresponding

TABLE 5.2 The proportion of genes that are present in multiple copies increases with genome size in multicellular eukaryotes.

	Unique Genes	Families with Two to Four Members	Families with More Than Four Members
H. influenzae	89%	10%	1%
S. cerevisiae	72%	19%	9%
D. melanogaster	72%	14%	14%
C. elegans	55%	20%	26%
A. thaliana	35%	24%	41%

polypeptides in different species are called **orthologous genes**, or **orthologs** (see the chapter titled *The Interrupted Gene*). Operationally, we usually consider that two genes in different organisms are orthologs if their sequences are similar over more than 80% of the length. By this criterion, about 20% of the fly genes have orthologs in both yeast and the worm. These genes are probably required by all eukaryotes. The proportion increases to 30% when the fly and worm are compared, probably representing the addition of gene functions that are common to multicellular eukaryotes. This still leaves a major proportion of genes as encoding proteins that are required specifically by either flies or worms, respectively.

A minimum estimate of the size of an organismal proteome can be deduced from the number and structures of genes, and a cellular or organismal proteome size can also be directly measured by analyzing the total polypeptide content of a cell or organism. Using such approaches, researchers have identified some proteins that were not suspected on the basis of genome analysis; this has led to the identification of new genes. Researchers use several methods for large-scale analysis of proteins. They can use mass spectrometry for separating and identifying proteins in a mixture obtained directly from cells or tissues. Hybrid proteins bearing tags can be obtained by expression of cDNAs made by ligating the sequences of ORFs to appropriate expression vectors that incorporate the sequences for affinity tags. This allows array analysis to be used to analyze the products. These methods also can be effective in comparing the proteins of two tissues—for example, a tissue from a healthy individual and one from a patient with a disease—to pinpoint the differences.

After we know the total number of proteins, we can ask how they interact. By definition, proteins in structural multiprotein assemblies must form stable interactions with one another. Also, proteins in signaling pathways interact with one another transiently. In both cases, such interactions can be detected in test systems where essentially a readout system magnifies the effect of the interaction. Such assays cannot detect all interactions; for example, if one enzyme in a metabolic pathway releases a soluble metabolite that then interacts with the next enzyme, the proteins might not interact directly.

As a practical matter, assays of pairwise interactions can give us an indication of the minimum number of independent structures or pathways. An analysis of the ability of all 6,000 predicted yeast proteins to interact in pair-wise combinations shows that about 1,000 proteins can bind to at least one other protein. Direct analyses of complex formation have identified 1,440 different proteins in 232 multiprotein complexes. This is the beginning of an analysis that will lead to defining the number of functional assemblies or pathways. A comparable analysis of 8,100 human proteins identified 2,800 interactions, but this is more difficult to interpret in the context of the larger proteome.

In addition to functional genes, there are also copies of genes that have become nonfunctional (identified as such by mutations in their protein-coding sequences). These are called pseudogenes. The number of pseudogenes can be large. In the mouse and human genomes, the number of pseudogenes is about 10% of the number of (potentially) functional genes (see the chapter titled *The Content of the Genome*). Some of these pseudogenes may serve the function of acting as targets for regulatory microRNAs; see the *Regulatory RNA* chapter.

5.5 The Human Genome Has Fewer Genes Than Originally Expected

KEY CONCEPTS

- Only 1% of the human genome consists of exons.
- The exons comprise about 5% of each gene, so genes (exons plus introns) comprise about 25% of the genome.
- The human genome has about 20,000 genes.
- Roughly 60% of human genes are alternatively spliced.
- Up to 80% of the alternative splices change protein sequence, so the human proteome has 50,000 to 60,000 members.

The human genome was the first vertebrate genome to be sequenced. This massive task has revealed a wealth of information about the genetic makeup of our species and about the evolution of genomes in general. Our understanding is deepened further by the ability to compare the human genome sequence with other sequenced vertebrate genomes.

Mammal genomes generally fall into a narrow size range, averaging about 3×10^9 bp (see the section *Pseudogenes Are Nonfunctional Gene Copies* later in this chapter). The mouse genome is about 14% smaller than the human genome, probably because it has had a higher rate of deletion. The genomes contain similar gene families and genes, with most genes having an ortholog in the other genome but with differences in the number of members of a family, especially in those cases for which the functions are specific to the species (see the chapter titled *The Content of the Genome*). Originally estimated to have about 30,000 genes, the mouse genome is now estimated to have more protein-coding genes than the human genome does, about 25,000. **FIGURE 5.8** plots the distribution of the mouse genes. The 25,000 protein-coding genes are accompanied by about 3,000 genes representing RNAs that do not encode proteins; these are generally small (aside from the ribosomal RNAs). Almost half of these genes encode transfer RNAs. In addition to the functional genes, about 1,200 pseudogenes have been identified.

The haploid human genome contains 22 autosomes plus the X and Y chromosomes. The chromosomes range in size from 45 to 279 Mb, making a total genome size of 3,235 Mb (about 3.2×10^9 bp). On the basis of chromosome structure, the genome can be divided into regions of euchromatin (containing many functional genes) and heterochromatin, with a much lower density of functional genes (see the *Chromosomes* chapter). The euchromatin comprises the majority of the genome, about 2.9×10^9 bp. The identified genome

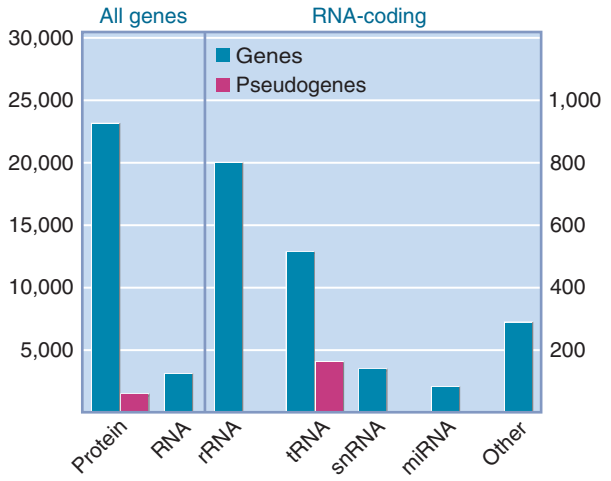


FIGURE 5.8 The mouse genome has about 25,000 protein-coding genes, which include about 1,200 pseudogenes. There are about 3,000 RNA-coding genes.

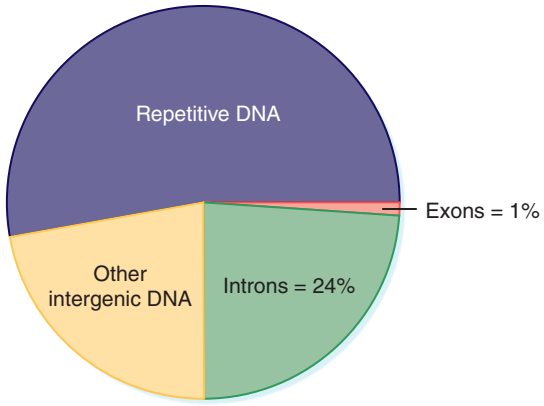


FIGURE 5.9 Genes occupy 25% of the human genome, but protein-coding sequences are only a small part of this fraction.

sequence represents more than 90% of the euchromatin. In addition to providing information on the genetic content of the genome, the sequence also identifies features that may be of structural importance.

FIGURE 5.9 shows that a very small proportion (about 1%) of the human genome is accounted for by the exons that actually encode polypeptides. The introns that constitute the remaining sequences of protein-coding genes bring the total of DNA involved with producing proteins to about 25%. As shown in **FIGURE 5.10**, the average human gene is 27 kb long with nine exons that include a total coding sequence of 1,340 bp. Therefore, the average coding sequence is only 5% of the length of an average protein-coding gene.

Two independent sequencing efforts for the human genome produced estimates of 30,000 and 40,000 genes, respectively. One measure of the accuracy of the analyses is whether they identify the same genes. The surprising answer is that the overlap between the two sets of genes is only about 50%, as summarized in **FIGURE 5.11**. An earlier analysis of the human gene set based on RNA transcripts had identified about 11,000 genes, almost all of which are present in both the large human gene sets, and which account for the major part of the overlap between them. So there is no question about the authenticity of half of each human gene set, but we have yet to establish the relationship between the other half of each set. The discrepancies illustrate the pitfalls of large-scale sequence analysis! As the sequence is analyzed further (and as other genomes are sequenced with which it can be compared), the number of actual genes has declined, and is now estimated to be about 20,000.

By any measure, the total human gene number is much smaller than was originally estimated—most estimates before the genome was sequenced were about 100,000. This represents a relatively small increase over the gene number of fruit flies and nematode worms (recent work suggests as many as 17,000 and 21,700, respectively), not to mention the plants *Arabidopsis* (25,000) and rice (32,000). However, we should not be particularly surprised by the notion that it does not take a great number of additional genes to make a more complex organism. The difference in DNA sequences between the human and chimpanzee genomes is extremely small (there is 98.5% similarity), so it is clear that the functions and interactions between a similar set of genes can produce different results. The functions of specific groups of genes can be especially important because detailed comparisons of orthologous genes in humans and chimpanzees suggest that there has been rapid evolution of certain classes of genes, including some involved in early development, olfaction, and hearing—all functions that are relatively specialized in these species.

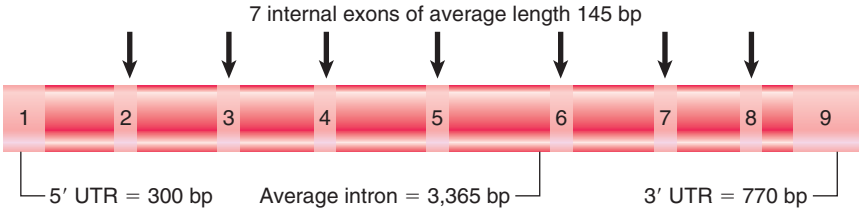


FIGURE 5.10 The average human gene is 27 kb long and has 9 exons usually comprising 2 longer exons at each end and 7 internal exons. The UTRs in the terminal exons are the untranslated (noncoding) regions at each end of the gene. (This is based on the average. Some genes are extremely long, which makes the median length 14 kb with 7 exons.)

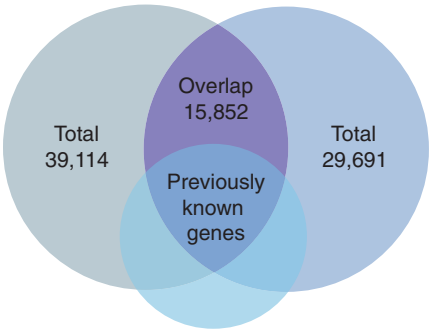


FIGURE 5.11 The two sets of genes identified in the human genome overlap only partially, as shown in the two large upper circles. However, they include almost all previously known genes, as shown by the overlap with the smaller, lower circle.

The number of protein-coding genes is less than the number of potential polypeptides because of mechanisms such as alternative splicing, alternate promoter selection, and alternate poly(A) site selection that can result in several polypeptides from the same gene (see the *RNA Splicing and Processing* chapter). The extent of alternative splicing is greater in humans than in flies or worms; it affects more than 60% of the genes (perhaps more than 90%), so the increase in size of the human proteome relative to that of the other eukaryotes might be larger than the increase in the number of genes. A sample of genes from two chromosomes suggests that the proportion of the alternative splices that actually result in changes in the polypeptide sequence is about 80%. If this occurs genome-wide, the size of the proteome could be 50,000 to 60,000 members.

However, in terms of the diversity of the number of gene families, the discrepancy between humans and the other eukaryotes might not be so great. Many of the human genes belong to gene families. An analysis of more than 20,000 genes identified 3,500 unique genes and 10,300 gene pairs. As can be seen from Figure 5.6, this extrapolates to a number of gene families only slightly larger than that of worms or flies.

5.6 How Are Genes and Other Sequences Distributed in the Genome?

KEY CONCEPTS

- Repeated sequences (present in more than one copy) account for more than 50% of the human genome.
- The great bulk of repeated sequences consists of copies of nonfunctional transposons.
- There are many duplications of large chromosome regions.

Are genes uniformly distributed in the genome? Some chromosomes are relatively "gene poor" and have more than 25% of their sequences as "deserts"—regions longer than 500 kb where there are no ORFs. Even the most gene-rich chromosomes have more than 10% of their sequences as deserts. So overall, about 20% of the human genome consists of deserts that have no protein-coding genes.

Repetitive sequences account for approximately 50% of the human genome, as seen in **FIGURE 5.12**. The repetitive sequences fall into five classes:

- Transposons (either active or inactive) account for the majority of repetitive sequences (45% of the genome). All transposons are found in multiple copies.
- Processed pseudogenes, about 3,000 in all, account for about 0.1% of total DNA. (These are sequences that arise by insertion of a reverse transcribed DNA copy of an mRNA sequence into the genome; see the section *Pseudogenes Are Nonfunctional Gene Copies* later in this chapter.)
- Simple sequence repeats (highly repetitive DNA such as CA repeats) account for about 3% of the genome.
- Segmental duplications (blocks of 10 to 300 kb that have been duplicated into a new region) account for about 5% of the genome. For a small percentage of cases, these duplications are found on the same chromosome; in the other cases, the duplicates are on different chromosomes.

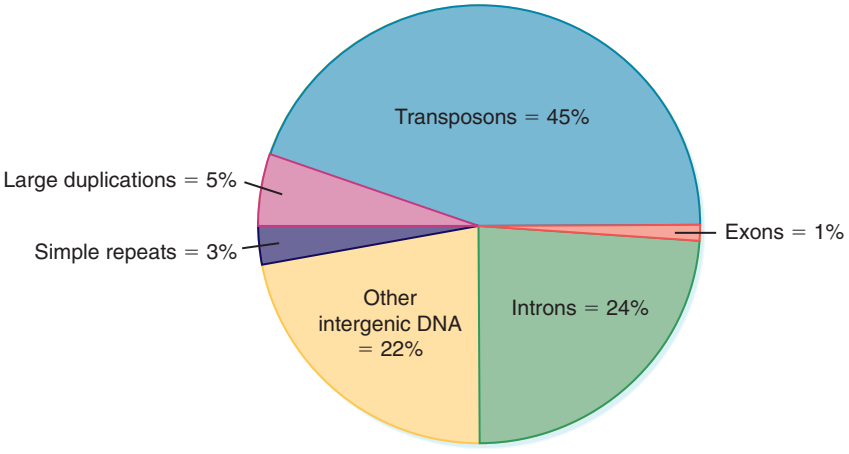


FIGURE 5.12 The largest component of the human genome consists of transposons. Other repetitive sequences include large duplications and simple repeats.

- Tandem repeats form blocks of one type of sequence. These are especially found at centromeres and telomeres.

The sequence of the human genome emphasizes the importance of transposons. Many transposons have the capacity to replicate themselves and insert into new locations. They can function exclusively as DNA elements or can have an active form that is RNA (see the chapter titled *Transposable Elements and Retroviruses*). Most of the transposons in the human genome are nonfunctional; very few are currently active. However, the high proportion of the genome occupied by these elements indicates that they have played an active role in shaping the genome. One interesting feature is that some currently functional genes originated as transposons and evolved into their present condition after losing the ability to transpose. At least 50 genes appear to have originated in this manner.

Segmental duplication at its simplest involves the tandem duplication of some region within a chromosome (typically because of an aberrant recombination event at meiosis; see the *Clusters and Repeats* chapter). However, in many cases the duplicated regions are on different chromosomes, implying that either there was originally a tandem duplication followed by a translocation of one copy to a new site or that the duplication arose by some different mechanism altogether. The extreme case of a segmental duplication is when an entire genome is duplicated, in which case the diploid genome initially becomes tetraploid. As the duplicated copies evolve differences from one another, the genome can gradually become effectively a diploid again, although homologies between the diverged copies leave evidence of the event. This is especially common in plant genomes. The present state of analysis of the human genome identifies many individual duplicated regions, and there is evidence for a whole-genome duplication in the vertebrate lineage (see the section *Genome Duplication Has Played a Role in Plant and Vertebrate Evolution* later in this chapter).

One curious feature of the human genome is the presence of sequences that do not appear to have coding functions but that nonetheless show an evolutionary conservation higher than the background level. As detected by comparison with other genomes (e.g., the mouse genome), these represent about 5% of the total genome. Are these sequences associated with protein-coding sequences in some functional way? Their density on chromosome 18 is the same as elsewhere in the genome, although chromosome 18 has a significantly lower concentration of protein-coding genes. This suggests indirectly that their function is not connected with structure or expression of protein-coding genes.

5.7 The Y Chromosome Has Several Male-Specific Genes

KEY CONCEPTS

- The Y chromosome has about 60 genes that are expressed specifically in the testis.
- The male-specific genes are present in multiple copies in repeated chromosomal segments.
- Gene conversion between multiple copies allows the active genes to be maintained during evolution.

The sequence of the human genome has significantly extended our understanding of the role of the sex chromosomes. It is generally thought that the X and Y chromosomes have descended from a common, very ancient autosome pair. Their evolution has involved a process in which the X chromosome has retained most of the original genes, whereas the Y chromosome has lost most of them.

The X chromosome is like the autosomes insofar as females have two copies and crossing over can take place between them. The density of genes on the X chromosome is comparable to the density of genes on other chromosomes.

The Y chromosome is much smaller than the X chromosome and has many fewer genes. Its unique role results from the fact that only males have the Y chromosome, of which there is only one copy, so Y-linked loci are effectively haploid instead of diploid like all other human genes.

For many years, the Y chromosome was thought to carry almost no genes except for one or a few genes that determine maleness. The large majority of the Y chromosome (more than 95% of its sequence) does not undergo crossing over with the X chromosome, which led to the view that it could not contain active genes because there would be no means to prevent the accumulation of deleterious mutations. This region is flanked by short **pseudoautosomal regions** that frequently exchange with the X chromosome during male meiosis. It was originally called the nonrecombining region but now has been renamed the **male-specific region**.

Detailed sequencing of the Y chromosome shows that the male-specific region contains three types of sequences, as illustrated in **FIGURE 5.13**:

- The *X-transposed sequences* consist of a total of 3.4 Mb comprising some large blocks that result from a transposition from band q21 in the X chromosome about 3 or 4 million years ago. This is specific to the human lineage. These sequences do not recombine with the X chromosome and have become largely inactive. They now contain only two functional genes.
- The *X-degenerate segments* of the Y chromosome are sequences that have a common origin with the X chromosome (going back to the common autosome from which both X and Y have descended) and contain genes or pseudogenes related to X-linked genes. There are 14 functional genes and 13 pseudogenes. Thus far, the functional genes have defied the trend for genes to be eliminated from chromosomal regions that cannot recombine at meiosis.
- The *ampliconic segments* have a total length of 10.2 Mb and are internally repeated on the Y chromosome. There are eight large palindromic blocks. They include nine protein-coding gene families, with copy numbers per family ranging from 2 to 35. The name *amplicon* reflects the fact that the sequences have been internally amplified on the Y chromosome.

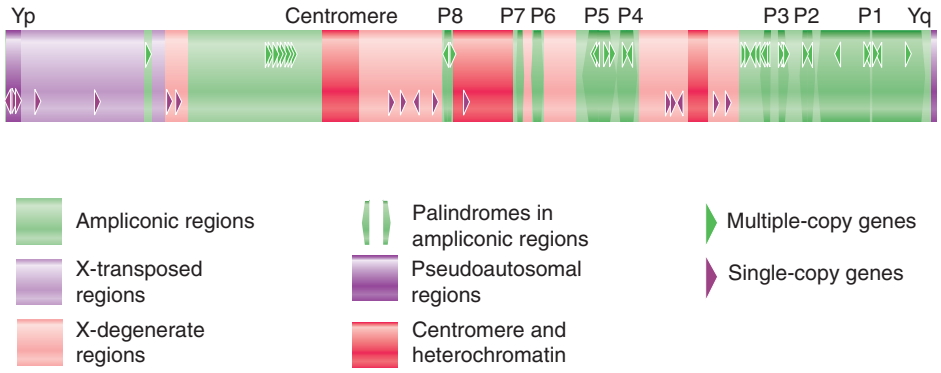


FIGURE 5.13 The Y chromosome consists of X-transposed regions, X-degenerate regions, and amplicons. The X-transposed and X-degenerate regions have 2 and 14 single-copy genes, respectively. The amplicons have 8 large palindromes (P1–P8), which contain 9 gene families. Each family contains at least 2 copies.

Totaling the genes in these three regions, the Y chromosome contains 156 transcription units, of which half represent protein-coding genes and half represent pseudogenes.

The presence of the functional genes is explained by the fact that the existence of closely related gene copies in the ampliconic segments allows gene conversion between multiple copies of a gene to be used to regenerate functional copies. The most common needs for multiple copies of a gene are quantitative (to provide more protein product) or qualitative (to encode proteins with slightly different properties or that are expressed at different times or in different tissues). However, in this case the essential function is evolutionary. In effect, the existence of multiple copies allows recombination within the Y chromosome itself to substitute for the evolutionary diversity that is usually provided by recombination between allelic chromosomes.

Most of the protein-coding genes in the ampliconic segments are expressed specifically in testes and are likely to be involved in male development. If there are roughly 60 such genes out of a total human gene set of about 20,000, the genetic difference between male and female humans is only about 0.3%.

5.8 How Many Genes Are Essential?

KEY CONCEPTS

- Not all genes are essential. In yeast and flies, individual deletions of less than 50% of the genes have detectable effects.
- When two or more genes are redundant, a mutation in any one of them might not have detectable effects.
- We do not fully understand the persistence of genes that are apparently dispensable in the genome.

The force of natural selection ensures that functional genes are retained in the genome. Mutations occur at random, and a common mutational effect in an ORF will be to damage the protein product. An organism with a damaging mutation will be at a disadvantage in competition and ultimately the mutation might be eliminated from a population. However, the frequency of a disadvantageous allele in the population is balanced between the generation of new copies of the allele by mutation and the elimination of the allele by selection. Reversing this argument, whenever we see an intact, expressed ORF in the genome, researchers assume that its product plays a useful role in the organism. Natural selection must have prevented mutations from accumulating in the gene. The ultimate fate of a gene that ceases to be functional is to accumulate mutations until it is no longer recognizable.

The maintenance of a gene implies that it does not confer a selective disadvantage to the organism. However, in the course of evolution, even a small relative advantage can be the subject of natural selection, and a phenotypic defect might not necessarily be immediately detectable as the result of a mutation. Also, in diploid organisms, a new recessive mutation can be "hidden" in heterozygous form for many generations. However, researchers would like to know how many genes are actually essential, meaning that their absence is lethal to the organism. In the case of diploid organisms, it means, of course, that the homozygous null mutation is lethal.

We might assume that the proportion of essential genes will decline with an increase in genome size, given that larger genomes can have multiple related copies of particular gene functions. So far this expectation has not been borne out by the data.

One approach to the issue of gene number is to determine the number of essential genes by mutational analysis. If we saturate some specified region of the chromosome with mutations that are lethal, the mutations should map into a number of complementation groups that correspond to the number of lethal loci in that region. By extrapolating to the genome as a whole, we can estimate the total essential gene number.

In the organism with the smallest known genome (*M. genitalium*), random insertions have detectable effects in only about two-thirds of the genes. Similarly, fewer than half of the genes of *E. coli* appear to be essential. The proportion is even lower in the yeast *S. cerevisiae*. When insertions were introduced at random into the genome in one early analysis, only 12% were

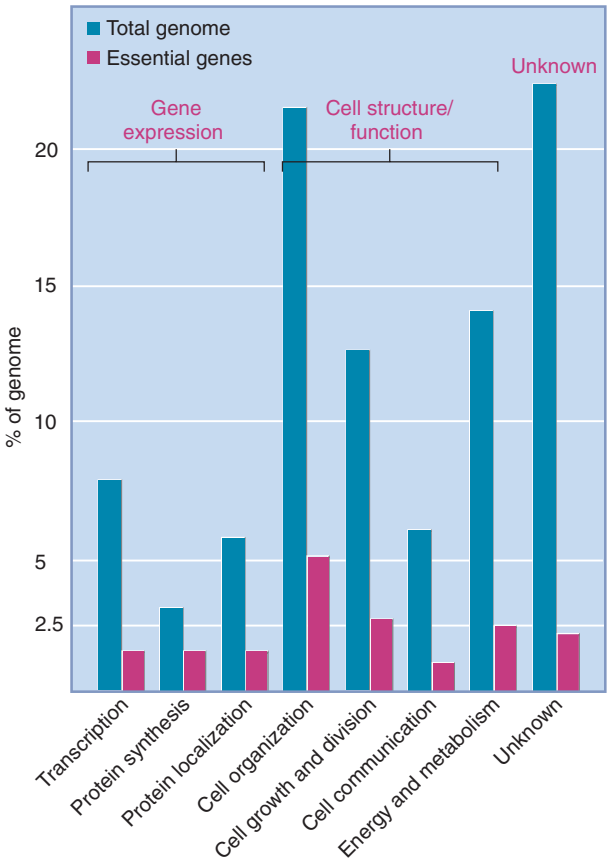


FIGURE 5.14 Essential yeast genes are found in all classes. Blue bars show the total proportion of each class of genes, and pink bars show those that are essential.

lethal and another 14% impeded growth. The majority (70%) of the insertions had no effect. A more systematic survey based on completely deleting each of 5,916 genes (more than 96% of the identified genes) shows that only 18.7% are essential for growth on a rich medium (i.e., when nutrients are fully provided). **FIGURE 5.14** shows that these include genes in all categories. The only notable concentration of defects is in genes encoding products involved in protein synthesis, for which about 50% are essential. Of course, this approach underestimates the number of genes that are essential for the yeast to live in the wild when it is not so well provided with nutrients.

FIGURE 5.15 summarizes the results of a systematic analysis of the effects of loss of gene function in the nematode worm *C. elegans*. The sequences of individual genes were predicted from the genome sequence, and by targeting an inhibitory RNA against these sequences (see the *Regulatory RNA* chapter) a large collection of worms was made in which one predicted gene was prevented from functioning in each worm. Detectable effects on the phenotype were only observed for 10% of these knockdowns, suggesting that most genes do not play essential roles.

There is a greater proportion of essential genes (21%) among those worm genes that have counterparts in other eukaryotes, suggesting that highly conserved genes tend to have more basic functions. There is also an increased proportion of essential genes among those that are present in only one copy per haploid genome, compared with those for which there are multiple copies of related or identical genes. This suggests that many of the multiple genes might be relatively recent duplications that can substitute for one another's functions.

Extensive analyses of essential gene number in a multicellular eukaryote have been made in *Drosophila* through attempts to correlate visible aspects of chromosome structure with the number of functional genetic units. The notion that this might be possible originated from the presence of bands in the polytene chromosomes of *D. melanogaster*. (These chromosomes are found at certain developmental stages and represent an unusually extended physical form in which a series of bands [more formally called chromomeres] are evident; see the *Chromosomes* chapter.) From the time of the early concept that the bands might represent a linear order of genes, there has been an attempt to correlate the organization of genes with the organization of bands. There are about 5,000 bands in the *D. melanogaster* haploid set; they vary in size over an order of magnitude, but on average there are about 20 kb of DNA per band.

The basic approach is to saturate a chromosomal region with mutations. Usually the mutations are simply collected as lethals without analyzing the cause of the lethality. *Any mutation that is lethal is taken to identify a locus that is essential for the organism. Sometimes mutations cause visible deleterious effects short of lethality, in which case we also define them as essential loci.* When the mutations are placed into complementation groups, the number can be compared with the number of bands in the region, or individual complementation groups might even be assigned to individual bands. The purpose of

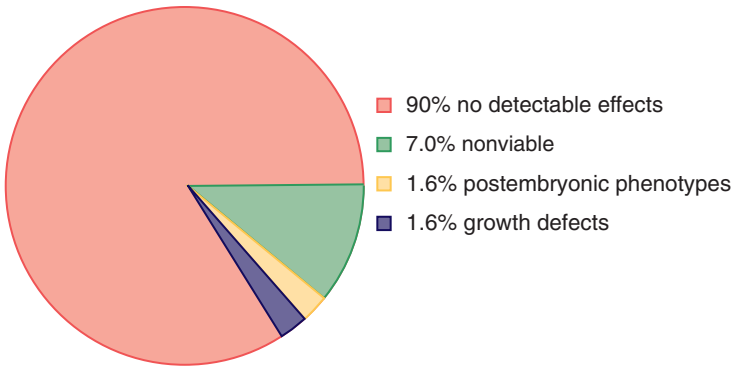


FIGURE 5.15 A systematic analysis of loss of function for 86% of worm genes shows that only 10% have detectable effects on the phenotype.

these experiments has been to determine whether there is a consistent relationship between bands and genes. For example, does every band contain a single gene?

Totaling the analyses that have been carried out since the 1970s, the number of essential complementation groups is about 70% of the number of bands. It is an open question as to whether there is any functional significance to this relationship. Regardless of the cause, the equivalence gives us a reasonable estimate for the essential gene number of around 3,600. By any measure, the number of essential loci in *Drosophila* is significantly less than the total number of genes.

If the proportion of essential human genes is similar to that of other eukaryotes, we would predict a range of 4,000 to 8,000 genes in which mutations would be lethal or produce evidently damaging effects. As of 2015, nearly 8,000 human genes in which mutations cause evident defects have been identified. This might actually exceed the upper range of the predicted total, especially in view of the fact that many lethal genes are likely to act so early in development that we never see their effects. This sort of bias might also explain the results in **TABLE 5.3**, which show that the majority of known genetic defects are due to point mutations (where there is more likely to be at least some residual function of the gene).

TABLE 5.3 Most known genetic defects in human genes are due to point mutations. The majority directly affect the protein sequence. The remainder is due to insertions, deletions, or rearrangements of varying sizes.

Type of Defect	Proportion of Genetic Defects Caused
Missense/nonsense	58%
Splicing	10%
Regulatory	< 1%
Small deletions	16%
Small insertions	6%
Large deletions	5%
Large rearrangements	2%

How do we explain the persistence of genes whose deletion appears to have no effect? The most likely explanation is that the organism has alternative ways of fulfilling the same function. The simplest possibility is that there is **redundancy**, with some genes present in multiple copies. This is certainly true in some cases, in which multiple related genes must be knocked out in order to produce an effect. In a slightly more complex scenario, an organism might have two separate biochemical pathways capable of providing some activity. Inactivation of either pathway by itself would not be damaging, but the simultaneous occurrence of mutations in genes from both pathways would be deleterious.

Such situations can be tested by combining mutations. In this approach, deletions in two genes, neither of which is lethal by itself, are introduced into the same strain. If the double mutant dies, the strain is called a **synthetic lethal**. This technique has been used to great effect with yeast, for which the isolation of double mutants can be automated. The procedure is called **synthetic genetic array analysis (SGA)**. **FIGURE 5.16** summarizes the results of an analysis in which an SGA screen was made for each of 132 viable deletions by testing whether it could survive in combination with any one of 4,700 viable deletions. Every one of the tested genes had at least one partner with which the combination was lethal, and most of the tested genes had many such partners; the median is 25 partners and the greatest number is shown by one tested gene that had 146 lethal partners. A small proportion (about 10%) of the interacting mutant pairs encode polypeptides that interact physically.

This result goes some way toward explaining the apparent lack of effect of so many deletions. Natural selection will act against these deletions when they are found in lethal pair-wise combinations. To some degree, the organism is protected against the damaging effects of mutations by built-in redundancy. There is, however, a price in the form of accumulating the "genetic load" of mutations that are not deleterious in themselves but that might cause serious problems when combined with other such mutations in future

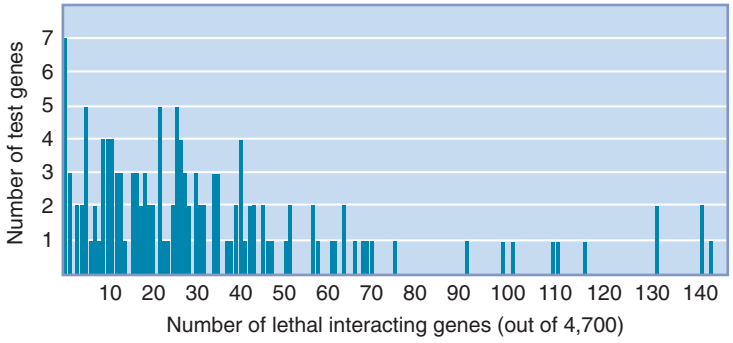


FIGURE 5.16 All 132 mutant test genes have some combinations that are lethal when they are combined with each of 4,700 nonlethal mutations. This chart shows how many lethal interacting genes there are for each test gene.

generations. Presumably, the loss of the individual genes in such circumstances produces a sufficient disadvantage to maintain the functional gene during the course of evolution.

5.9 About 10,000 Genes Are Expressed at Widely Differing Levels in a Eukaryotic Cell

KEY CONCEPTS

- In any particular cell, most genes are expressed at a low level.
- Only a small number of genes, whose products are specialized for the cell type, are highly expressed.
- mRNAs expressed at low levels overlap extensively when different cell types are compared.
- The abundantly expressed mRNAs are usually specific for the cell type.
- About 10,000 expressed genes might be common to most cell types of a multicellular eukaryote.

The proportion of DNA containing protein-coding genes being expressed in a specific cell at a specific time can be determined by the amount of the DNA that can hybridize with the mRNAs isolated from that cell. Such a saturation analysis conducted for many cell types at various times typically identifies about 1% of the DNA being expressed as mRNA. From this researchers can calculate the number of protein-coding genes, as long as they know the average length of an mRNA. For a unicellular eukaryote such as yeast, the total number of expressed protein-coding genes is about 4,000. For somatic tissues of multicellular eukaryotes, including both plants and vertebrates, the number is usually 10,000 to 15,000. (The only consistent exception to this type of value is presented by mammalian brain cells, for which much larger numbers of genes appear to be expressed, although the exact number is not certain.)

Researchers can use kinetic analysis of the reassociation of an RNA population to determine its sequence complexity. This type of analysis typically identifies three components in a eukaryotic cell. Just as with a DNA reassociation curve, a single component hybridizes over about 2 decades of Rot values (RNA concentration × time), and a reaction extending over a greater range must be resolved by computer curve-fitting into individual components. Again, this represents what is really a continuous spectrum of sequences.

FIGURE 5.17 shows an example of an excess mRNA × cDNA reaction that generates three components:

- The first component has the same characteristics as a control reaction of ovalbumin mRNA with its DNA copy. This suggests that the first component is in fact just ovalbumin mRNA (which indeed is about half of the mRNA mass in oviduct tissue).
- The next component provides 15% of the reaction, with a total length of 15 kb. This corresponds to 7 to 8 mRNA species with an average length of 2,000 bases.
- The last component provides 35% of the reaction, which corresponds to a length of 26 Mb. This corresponds to about 13,000 mRNA species with an average length of 2,000 bases.

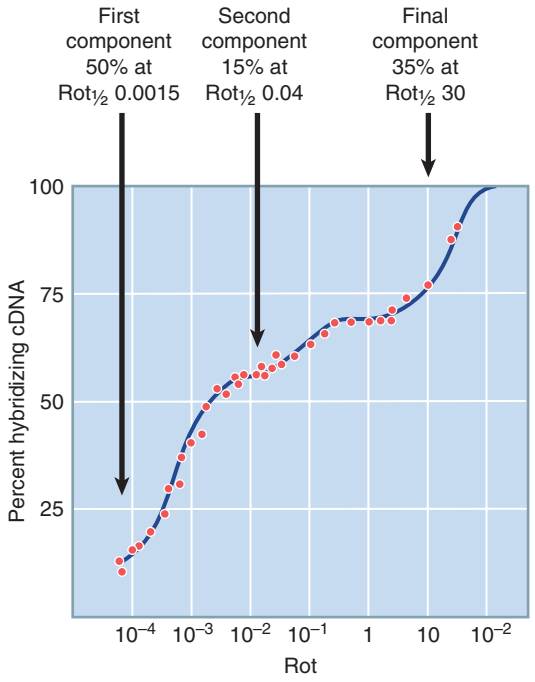


FIGURE 5.17 Hybridization between excess mRNA and cDNA identifies several components in chick oviduct cells, each characterized by the $Rot_{½}$ of reaction.

From this analysis, we can see that about half of the mass of mRNA in the cell represents a single mRNA, about 15% of the mass is provided by a mere seven to eight mRNAs, and about 35% of the mass is divided into the large number of 13,000 mRNA types. It is therefore obvious that the mRNAs comprising each component must be present in very different amounts.

The average number of molecules of each mRNA per cell is called its **abundance**. Researchers can calculate it quite simply if the total mass of a specific mRNA type in the cell is known. In the example of chick oviduct cells shown in Figure 5.17, the total mRNA can be accounted for as 100,000 copies of the first component (ovalbumin mRNA), 4,000 copies of each of 7 or 8 other mRNAs in the second component, and only about 5 copies of each of the 13,000 remaining mRNAs that constitute the last component.

We can divide the mRNA population into two general classes, according to their abundance:

- The oviduct is an extreme case, with so much of the mRNA represented by only one type, but most cells do contain a small number of RNAs present in many copies each. This **abundant mRNA** component typically consists of fewer than 100 different mRNAs present in 1,000 to 10,000 copies per cell. It often corresponds to a major part of the mass, approaching 50% of the total mRNA.
- About half of the mass of the mRNA consists of a large number of sequences, of the order of 10,000, each represented by only a small number of copies in the mRNA—say, fewer than 10. This is the **scarce mRNA** (or **complex mRNA**) class. It is this class that drives a saturation reaction.

Many somatic tissues of multicellular eukaryotes have an expressed gene number in the range of 10,000 to 20,000. How much overlap is there between the genes expressed in different tissues? For example, the expressed gene number of chick liver is between 11,000 and 17,000, compared with the value for oviduct of 13,000 to 15,000. How much do these two sets of genes overlap? How many are specific for each tissue? These questions are usually addressed by analyzing the transcriptome—the set of sequences represented in RNA.

We see immediately that there are likely to be substantial differences among the genes expressed in the abundant class. Ovalbumin, for example, is synthesized only in the oviduct and not at all in the liver. This means that 50% of the mass of mRNA in the oviduct is specific to that tissue.

However, the abundant mRNAs represent only a small proportion of the number of expressed genes. In terms of the total number of genes of the organism, and of the number of changes in transcription that must be made between different cell types, we need to know the extent of overlap between the genes represented in the scarce mRNA classes of different cell phenotypes.

Comparisons between different tissues show that, for example, about 75% of the sequences expressed in liver and oviduct are the same. In other words, about 12,000 genes are expressed in both liver and oviduct, 5,000 additional genes are expressed only in liver, and 3,000 additional genes are expressed only in oviduct.

The scarce mRNAs overlap extensively. Between mouse liver and kidney, about 90% of the scarce mRNAs are identical, leaving a difference between the tissues of only 1,000 to 2,000 expressed genes. The general result obtained in several comparisons of this sort is that only about 10% of the mRNA sequences of a cell are unique to it. The majority of mRNAs are common to many—perhaps even all—cell types.

This suggests that the common set of expressed gene functions, numbering perhaps about 10,000 in mammals, comprise functions that are needed in all cell types. Sometimes, this type of function is referred to as a housekeeping gene or **constitutive gene**. It contrasts with the activities represented by specialized functions (such as ovalbumin or globin) needed only for particular cell phenotypes. These are sometimes called **luxury genes**.

5.10 Expressed Gene Number Can Be Measured En Masse

KEY CONCEPTS

- DNA microarray technology allows a snapshot to be taken of the expression of the entire genome in a yeast cell.
- About 75% (approximately 4,500 genes) of the yeast genome is expressed under normal growth conditions.
- DNA microarray technology allows for detailed comparisons of related animal cells to determine (for example) the differences in expression between a normal cell and a cancer cell.

Recent technology allows more systematic and accurate estimates of the number of expressed protein-coding genes. One approach (serial analysis of gene expression, or SAGE) allows a unique sequence tag to be used to identify each mRNA. The technology then allows the abundance of each tag to be measured. This approach identifies 4,665 expressed genes in *S. cerevisiae* growing under normal conditions, with abundances varying from 0.3 to fewer than 200 transcripts/cell. This means that about 75% of the total gene number (about 6,000) is expressed under these conditions. **FIGURE 5.18** summarizes the number of different mRNAs that is found at each different abundance level.

One powerful technology uses chips that contain **microarrays**, which are arrays of many tiny DNA oligonucleotide samples. Their construction is made possible by knowledge of the sequence of the entire genome. In the case of *S. cerevisiae*, each of 6,181 ORFs is represented on the microarray by twenty 25-mer oligonucleotides that perfectly match the sequence of the mRNA and 20 mismatched oligonucleotides that differ at one base position. The expression level of any gene is calculated by subtracting the average signal of a mismatch from its perfect match partner. The entire yeast genome can be represented on four chips. This technology is sensitive enough to detect transcripts of 5,460 genes (about 90% of the genome) and shows that many genes are expressed at low levels, with abundances of 0.1 to 0.2 transcript/cell. (An abundance of less than 1 transcript/cell means that not all cells have a copy of the transcript at any given moment.)

The technology allows not only measurement of levels of gene expression but also detection of differences in expression in mutant cells compared to wild-type cells growing under different conditions, and so on. The results of comparing two states are expressed in the form of a grid, in which each square represents a particular gene and the relative change in expression is indicated by color. These data can be converted to a **heat map** showing wild-type versus mutant expression of genes under different conditions. **FIGURE 5.19**

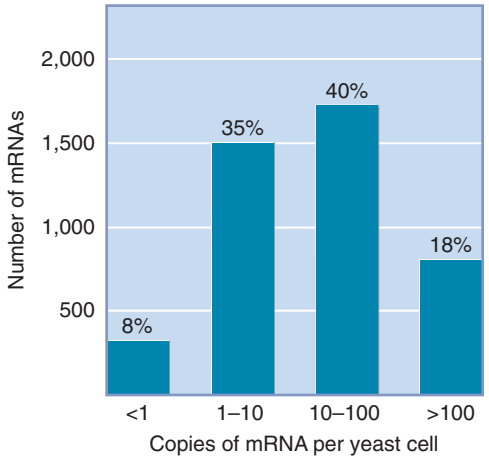


FIGURE 5.18 The abundances of yeast mRNAs vary from less than 1 per cell (meaning that not every cell has a copy of the mRNA) to more than 100 per cell (encoding the more abundant proteins).

Image courtesy of Rachel E. Ellsworth, Clinical Breast Care Project, Windber Research Institute.

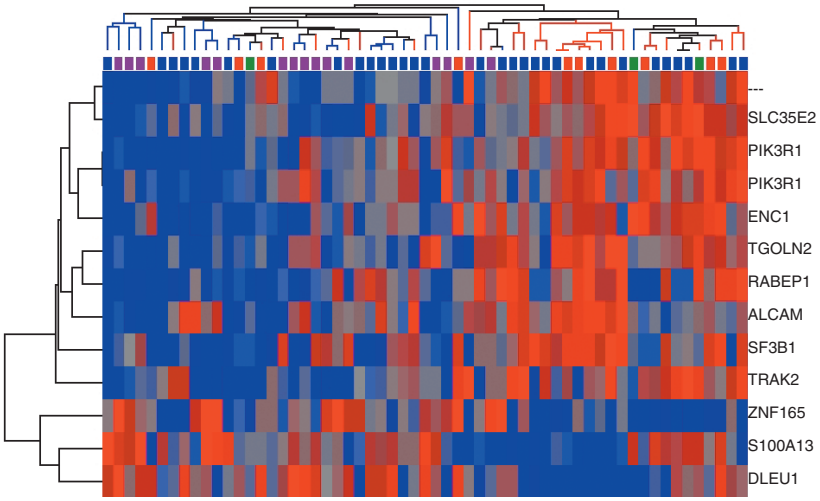


FIGURE 5.19 "Heat map" of 59 invasive breast tumors from women who breastfed for at least 6 months (red lines above map) or who never breastfed (blue lines). Different tumor subtypes are denoted by the blue, green, red, and purple bars above the map. In the map, the expression of a number of genes (listed at the right) in the tumor is compared to their expression in normal breast tissue: red = higher expression, blue = lower expression, gray = equal expression.

Image courtesy of Rachel E. Ellsworth, Clinical Breast Care Project, Windber Research Institute.

shows the difference in expression of a number of genes between normal human breast tissue and cancerous breast tumors. The heat map compares women who breastfed with those who did not, and overall shows that for many genes women who breastfed had increased gene expression.

The extension of this and newer technologies (e.g., deep RNA sequencing; see the chapter titled *The Content of the Genome*) to animal cells will allow the general descriptions based on RNA hybridization analysis to be replaced by exact descriptions of the genes that are expressed, and the abundances of their products, in any particular cell type. A gene expression map of *D. melanogaster* detects transcriptional activity in some stage of the life cycle in almost all (93%) of predicted genes and shows that 40% have alternatively spliced forms.

5.11 DNA Sequences Evolve by Mutation and a Sorting Mechanism

KEY CONCEPTS

- The probability of a mutation is influenced by the likelihood that the particular error will occur and the likelihood that it will be repaired.
- In small populations, the frequency of a mutation will change randomly and new mutations are likely to be eliminated by chance.
- The frequency of a neutral mutation largely depends on genetic drift, the strength of which depends on the size of the population.
- The frequency of a mutation that affects phenotype will be influenced by negative or positive selection.

Biological evolution is based on two sets of processes: the generation of genetic variation and the sorting of that variation in subsequent generations. Variation among chromosomes can be generated by recombination (see the chapter titled *Homologous and Site-Specific Recombination*); variation among sexually reproducing organisms results from the combined processes of meiosis and fertilization. Ultimately, however, variation among DNA sequences is a result of mutation.

Mutation occurs when DNA is altered by replication error or chemical changes to nucleotides, or when electromagnetic radiation breaks or forms chemical bonds, and the damage remains unrepaired at the time of the next DNA replication event (see the chapter titled *Repair Systems*). Regardless of the cause, the initial damage can be considered an "error." In principle, a base can mutate to any of the other three standard bases, though the three possible mutations are not equally likely due to biases incurred by the mechanisms of damage (see the section *There May Be Biases in Mutation, Gene Conversion, and Codon Usage* later in this chapter) and differences in the likelihood of repair of the damage.

For example, if mutation from one base to any of the other three is equally probable, *transversion mutations* (from a pyrimidine to a purine, or vice versa) would be twice as frequent as *transition mutations* (from one pyrimidine to another, or one purine to another; see the *Genes Are DNA and Encode RNAs and Polypeptides* chapter). However, the observation is usually the opposite: Transitions occur roughly twice as frequently as transversions. This might be because (1) spontaneous transitional errors occur more frequently than transversional errors; (2) transversional errors are more likely to be detected and corrected by DNA repair mechanisms; or (3) both of these are true. Given that transversional errors result in distortion of the

	A	T	C	G
A	–			
T		–		
C			–	
G				–

FIGURE 5.20 A simple model of mutational change in which α is the probability of a transition and β is the probability of a transversion.

Reproduced from MEGA (Molecular Evolutionary Genetics Analysis) by S. Kumar, K. Tamura, and J. Dudley. Used with permission of Masatoshi Nei, Pennsylvania State University.

DNA duplex as either pyrimidines or purines are paired together, and that base-pair geometry is used as a fidelity mechanism (see the *DNA Replication* and *Repair Systems* chapters), it is less likely for a DNA polymerase to make a transversional error. The distortion also makes it easier for transversional errors to be detected by postreplication repair mechanisms. As shown in **FIGURE 5.20**, a basic model of mutation would be that the probabilities of transitions are equal (α), as are those of transversions (β), and that α > β. More complex models could have different probabilities for the individual substitution mutations, and could be tailored to individual taxonomic groups from actual data on mutation rates in those groups.

If a mutation occurs in the coding region of a protein-coding gene, it can be characterized by its effect on the polypeptide product of the gene. A substitution mutation that does not change the amino acid sequence of the polypeptide product is a **synonymous mutation**; this is a specific type of **silent mutation**. (Silent mutations include those that occur in noncoding regions.) A **nonsynonymous mutation** in a coding region does alter the amino acid sequence of the polypeptide product, resulting in either a missense codon (for a different amino acid) or a nonsense (termination) codon. The effect of the mutation on the phenotype of the organism will influence the fate of the mutation in subsequent generations.

Mutations in genes other than those encoding polypeptides and mutations in noncoding sequences can, of course, also be subject to selection. In noncoding regions, a mutational change can alter the regulation of a gene by directly changing a regulatory sequence or by changing the secondary structure of the DNA in such a way that some aspect of the gene's expression (such as transcription rate, RNA processing, or mRNA structure influencing translation rate) is affected. However, many changes in noncoding regions might be selectively **neutral mutations**, having no effect on the phenotype of the organism.

If a mutation is selectively neutral or near neutral, its fate is predictable only in terms of probability. The random changes in the frequency of a mutational variant in a population are called **genetic drift**; this is a type of "sampling error" in which, by chance, the offspring genotypes of a particular set of parents do not precisely match those predicted by Mendelian inheritance. In a very large population, the random effects of genetic drift tend to average out, so there is little change in the frequency of each variant. However, in a small population, these random changes can be quite significant and genetic drift can have a major effect on the genetic variation of the population. **FIGURE 5.21** shows a simulation comparing the random changes in allele frequency for seven populations of 10 individuals each with those of seven populations of 100 individuals each. Each population begins with two alleles, each with a frequency of 0.5. After 50 generations, most of the small populations have lost one or the other allele ($p = 1$ means only one allele is left and $p = 0$ means only the other allele is left), whereas the large populations have retained both alleles (though their allele frequencies have randomly drifted from the original 0.5).

Genetic drift is a random process. The eventual fate of a particular variant is not strictly predictable, but the current frequency of the variant is a measure of the probability that it will eventually be *fixed* (replacing all other variants) in the population. In other words, a new mutation (with a low frequency in a population) is very likely to be lost from the population by chance. However, if by chance it becomes more frequent, it has a greater probability of being retained in the population. Over the long term, a variant might either be lost from the population or fixed, but in the short term there might be randomly fluctuating variation for a particular locus, especially in smaller populations where **fixation** or loss occurs more quickly.

On the other hand, if a new mutation is not selectively neutral and does affect phenotype, natural selection will play a role in its increase or decrease in frequency in the population. The speed of its frequency change will partly depend on how much of an advantage or disadvantage the mutation confers to the organisms that carry it. It will also depend on whether it is dominant or recessive; in general, because dominant mutations are "exposed" to natural selection when they first appear, they are affected by selection more rapidly.

Mutations are random with regard to their effects, and thus the common result of a nonneutral mutation is for the phenotype to be negatively affected, so selection often acts primarily to eliminate new mutations (though this might be somewhat delayed in the likely event that the mutation is recessive). This is called **negative** (or **purifying**) **selection** (see the chapter titled *The Interrupted Gene*). The overall result of negative selection is for there to be little variation within a population as new variants are generally eliminated. More rarely, a new mutation might be subject to **positive selection** (see the chapter titled *The Interrupted Gene*) if it happens to confer an advantageous phenotype. This type of selection will also tend to reduce variation within a population, as the new mutation eventually replaces the original sequence, but can result in greater variation *between* populations, provided they are isolated from one another, as different mutations occur in these different populations.

The question of how much observed genetic variation in a population or species (or the lack of such variation) is due to selection and how much is due to genetic drift is a long-standing one in population genetics. In the next section, we

FIGURE 5.21 The fixation or loss of alleles by random genetic drift occurs more rapidly in populations of 10 **(a)** than in populations of 100. **(b)** *p* is the frequency of one of two alleles at a locus in the population.

Data courtesy of Kent E. Holsinger, University of Connecticut (http://darwin.eeb.uconn.edu).

look at some ways that selection on DNA sequences might be detected by testing for significant differences from the expectations of evolution of neutral mutations.

5.12 Selection Can Be Detected by Measuring Sequence Variation

KEY CONCEPTS

- The ratio of nonsynonymous to synonymous substitutions in the evolutionary history of a gene is a measure of positive or negative selection.
- Low heterozygosity of a gene might indicate recent selective events.
- Comparing the rates of substitution among related species can indicate whether selection on the gene has occurred.
- Most functional genetic variation in the human species affects gene regulation and not variation in proteins.

Many methods have been used over the years for analyzing selection on DNA sequences. With the development of DNA sequencing techniques in the 1970s (see the chapter titled *Methods in Molecular Biology and Genetic Engineering*), the automation of sequencing in the 1990s, and the development of high-throughput sequencing in the 21st century, large numbers of partial or complete genome sequences are becoming available. Coupled with the polymerase chain reaction (PCR), which amplifies specific genomic regions, DNA sequence analysis has become a valuable tool in many applications, including the study of selection on genetic variants.

There is now an abundance of DNA sequence data from a wide range of organisms in various publicly available databases. Homologous gene sequences have been obtained from many species as well as from different individuals of the same species. This allows for determination of genetic changes among species with common ancestry as compared to changes within a species. These comparisons have led to the observation that some species (e.g., *D. melanogaster*) have high levels of DNA sequence polymorphism among individuals, most likely as a result of neutral mutations and random genetic drift within populations. (Other species, such as humans, have moderate levels of polymorphism, and without further investigation, the relative roles of genetic drift and selection in keeping these levels low is not immediately clear. This is one use for techniques to detect selection on sequences.) By conducting both interspecific and intraspecific DNA sequence analysis, the level of divergence due to species differences can be determined.

Some neutral mutations are synonymous mutations, but not all synonymous mutations are neutral. Although at first this might seem unlikely, the concentrations of individual tRNAs that specify a particular amino acid in a cell are not equal. Some cognate transfer RNAs (tRNAs) (different tRNAs that carry the same amino acid) are more abundant than others, and a specific codon might lack sufficient tRNAs, whereas a different codon for the same amino acid might have a sufficient number. In the case of a codon that requires a rare tRNA in that organism, ribosomal frameshifting or other alterations in translation may occur (see the chapter titled *Using the Genetic Code*). It also might be that a particular codon is necessary to maintain mRNA structure. Alternatively, there might be a nonsynonymous mutation to an amino acid with the same general characteristics, with little or no effect on the folding and activity of the polypeptide. In either case neutral sequence changes have little effect on the organism. However, a nonsynonymous mutation might result in an amino acid with different properties, such as a change from a polar to a nonpolar amino acid, or from a hydrophobic amino acid to a hydrophilic one in a protein embedded in a phospholipid bilayer. Such changes are likely to have functional effects that are deleterious to the role of the polypeptide and thus to the organism. Depending on the location of the amino acid in the polypeptide, such a change might cause only a slight disruption of protein folding and activity. Only in rare cases is an amino acid change advantageous; in this case the mutational change might become subjected to positive selection and ultimately lead to fixation of this variant in the population.

One common approach for determining selection is to use codon-based sequence information to study the evolutionary history of a gene. Researchers can do this by counting the number of synonymous (K_s) and nonsynonymous (K_a) amino acid substitutions in orthologous genes (see the chapter titled *The Interrupted Gene*) and determining the K_a/K_s ratio. This ratio is indicative of the selective constraints on the gene. A K_a/K_s ratio of 1 is expected for those genes that evolve neutrally, with amino acid sequence changes being neither favored nor disfavored. In this case, the changes that occur do not usually affect the activity of the polypeptide, and this serves as a suitable control. A K_a/K_s ratio <1 is most commonly observed and indicates negative selection, where amino acid replacements are disfavored because they affect the activity of the polypeptide. Thus, there is selective pressure to retain the original functional amino acid at these sites in order to maintain proper protein function.

Positive selection is indicated when the K_a/K_s ratio is >1, but is rarely observed. This means that the amino acid changes are advantageous and might become fixed in the population. One example of this is the antigenic proteins of some pathogens, such as viral coat proteins, which are under strong selection pressure to evade the immune response of the host. A second example is some reproductive proteins that are under *sexual selection* (selection on traits found in one sex). As a third example, the K_a/K_s ratios for the peptide-binding regions of mammalian MHC genes, the products of which function in immunological self-recognition by displaying both "self" and "nonself" antigens, are typically in the range of 2 to 10, indicating strong selection for new variants. This is expected because these proteins represent the cellular uniqueness of individual organisms.

The detection of a positive K_a/K_s ratio might be rare in part because the average value must be greater than one over a length of sequence. If a single substitution in a gene is being positively selected, but flanking regions are under negative selection, the average ratio across the sequence might actually be negative. In contrast, the K_a/K_s ratios for histone genes are typically much less than one, suggesting strong negative selection on these genes. Histones are DNA-binding proteins that make up the basic structure of chromatin (see the chapter titled *Chromatin*) and alterations to their structures are likely to result in deleterious effects on chromosome integrity and gene expression.

In addition to the difficulty of detecting strong selection on a single substitution variant when K_a/K_s is averaged over a stretch of DNA, mutational hotspots can also affect this measure. There have been reports of unusually highly mutable regions of some protein-coding genes that encode a high proportion of polar amino acids; such a bias might influence the interpretation of the K_a/K_s ratio because a higher point mutation rate might be incorrectly interpreted as a higher substitution rate. The lesson seems to be that although codon-based methods of detecting selection can be useful, their limitations must be taken into account.

Researchers can use intraspecific DNA sequence analysis to detect positive selection by comparing the nucleotide sequence between two alleles or two individuals of the same species. Nucleotide sequences are expected to evolve neutrally at a rate proportional to the mutation rate; variation in this rate at specific nucleotides affects the *heterozygosity* of a population (the proportion of heterozygotes for a particular locus). If a variant sequence is favored, the variant will increase in frequency and eventually become fixed in the population, and the site will show a reduction in nucleotide heterozygosity. Closely linked neutral variants can also become fixed, a phenomenon termed **genetic hitchhiking**.

These regions are characterized by having a lower level of DNA sequence polymorphism. (However, it is important to remember that reduced polymorphism can have other causes, such as negative selection or genetic drift.)

In practice it is more reliable to carry out both interspecific and intraspecific DNA sequence comparisons to detect deviations from neutral evolutionary expectations. By including sequence information from at least one closely related species, species-specific DNA polymorphisms can be distinguished from ancestral polymorphisms, and more accurate information regarding the link between the polymorphisms and between species differences can be obtained. With this combined analysis, the degree of nonsynonymous changes between species can be determined. If evolution is primarily neutral, the ratio of nonsynonymous to synonymous changes *within* species is expected to be the same as the ratio *between* species. An excess of nonsynonymous changes might be evidence for positive selection on these amino acids, whereas a lower ratio might indicate that negative selection is conserving sequences.

One example is the comparison of 12 sequences of the *Adh* gene in *D. melanogaster* to each other and to *Adh* sequences from *Drosophila simulans* and *Drosophila yakuba*, as shown in **TABLE 5.4**. A simple contingency chi-square test on these data shows that there are significantly more fixed nonsynonymous changes between species than similar polymorphisms in *D. melanogaster*. The high proportion of nonsynonymous differences among species suggests positive selection on *Adh* variants in these species, as does the lower proportion of such differences in one species, given that nonneutral variation would not be expected to persist for very long within a species.

Relative rate tests can also be used to detect the signature of selection. This involves (at a minimum) three related species: two that are closely related and one outgroup representative. The substitution rate is compared between the close relatives, and each is compared to the outgroup species to see if the substitution rates are similar. This removes the dependence of the analysis on time, as long as the phylogenetic relationships between the species are certain. If the rate of substitutions between related species compared to the rate between these and the outgroup species is different, this might be an indication of selection on the sequence. For example, the protein lysozyme, which functions to digest bacterial cell walls and is a general antibiotic in many species, has evolved to be active at low pH in ruminating mammals, where it functions to digest dead bacteria in the gut. **FIGURE 5.22** shows that the number of amino acid (i.e., nonsynonymous) substitutions for lysozyme in the cow/deer (ruminant) lineage is higher than that of the nonruminant pig outgroup.

TABLE 5.4 Nonsynonymous and synonymous variation in the *Adh* locus in *Drosophila melanogaster* ("polymorphic") and between *D. melanogaster*, *D. simulans*, and *D. yakuba* ("fixed").

	Nonsynonymous	Synonymous
Fixed	7	17
Polymorphic	2	42

Data from J. H. McDonald and M. Kreitman, *Nature* 351 (1991): 652–654.

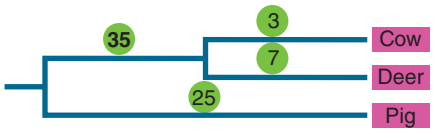


FIGURE 5.22 A higher number of nonsynonymous substitutions in lysozyme sequences in the cow/deer lineage as compared to the pig lineage is a result of adaptation of the protein for digestion in ruminant stomachs.

Data from: N. H. Barton, et al. 2007. *Evolution*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. Original figure appeared in Gillespie J. H. 1994. *The Causes of Molecular Evolution*. Oxford University Press.

This method must take into account that some genes accumulate nucleotide or amino acid substitutions more rapidly (these are said to be *fast-clock*; see the next section *A Constant Rate of Sequence Divergence Is a Molecular Clock*) in some species than in others, possibly due to differences in metabolic rate, generation time, DNA replication time, or DNA repair efficiency. To deal with this difference, additional related species need to be examined in order to identify and eliminate fast-clock effects. The reliability of this approach is improved if larger numbers of distantly related species are included. However, it is difficult to make accurate comparisons between taxonomic groups due to the inherent rate differences. As more work in this area has been done, corrections to adjust for differences in substitution rates have been developed.

Another method for detecting selection utilizes estimates of polymorphism at specific genetic loci. For example, sequence analysis of the *Teosinte branched 1* (*tb1*) locus, an important gene in domesticated maize, has been used to characterize the nucleotide substitution rate in domesticated and wild maize (teosinte) varieties, with an estimate of 2.9×10^{-8} to 3.3×10^{-8} base substitutions per year. For a neutrally evolving gene, the ratio of a measure of nucleotide diversity (p) in domesticated maize to p in wild teosinte is about 0.75, but it is less than 0.1 in the *tb1* region. The interpretation is that strong selection in domesticated maize has severely reduced variation for this gene.

As genome-wide data on nucleotide diversity become available, regions of low diversity can indicate recent selection. Millions of single nucleotide polymorphisms (SNPs) are being characterized in humans, nonhuman animals, and plants, as well as in other species. One approach that has been applied to the human genome is to look for an association between an allele's frequency and its **linkage disequilibrium** with other genetic markers surrounding it. (Linkage disequilibrium is a measure of an association between an allele at one locus and an allele at a different locus.) When a new mutation occurs on one chromosome, it initially has high linkage

disequilibrium with alleles at other polymorphic loci on the same chromosome. In a large population, a neutral allele is expected to rise to fixation slowly, so recombination and mutation will break up associations between loci and linkage disequilibrium will decrease. On the other hand, an allele under positive selection will rise to fixation more quickly and linkage disequilibrium will be maintained. By sampling SNPs across the genome, researchers can establish a general background level of linkage disequilibrium that accounts for local variations in rates of recombination, and any significantly higher measures of linkage disequilibrium can be detected. **FIGURE 5.23** shows the slowly decreasing linkage disequilibrium (measured by the increasing fraction of recombinant chromosomes) with increasing chromosomal distance from a variant of the *G6PD* locus that confers resistance to malaria in African human populations. This pattern suggests that this allele has been under strong recent selection—carrying along with it linked alleles at other loci—and that recombination has not yet had time to break up these interlocus associations.

The availability of multiple complete human genome sequences and the ability to rapidly resequence specific regions of the genome in many individuals allows large-scale measurement of genetic variation in the human species. As described earlier, a lack of genetic variation in a stretch of DNA can indicate negative selection on that sequence, implying that the sequence is functional. If the analysis includes individuals from many populations, we can determine whether individual variations are unique, shared by other members of a specific population, or found globally. Surprisingly, such studies show that the majority of *functional* variations in the human genome are *not* nonsynonymous changes in coding sequences, but are found in noncoding sequences such as introns or intergenic regions! In other words, protein variations account for only a small percentage of functional differences among humans. Presumably, the large percentage of functional variation in noncoding regions reflects differences in regulatory regions (see the chapters in Part III, *Gene Regulation*). Also, most of these variations are found in most or all sampled populations and are not limited to one or a few populations. Clearly, despite many apparent differences among individual humans, there is genetic unity to the human species, and most of the differences are not with the proteins being produced in cells, but *when* and *where* they are being produced.

The 1000 Genomes Project began in 2008 with the initial goal of sequencing at least 1,000 individual anonymous human genomes to assess comprehensive human genetic variation. During the first 2 years of the project, sequencing progressed at a rate that was the equivalent of two genomes per day using reduced-cost, next-generation sequencing techniques. The sequence data are available in free-access public databases. By late 2015, more than 2,500 human genomes had been sequenced.

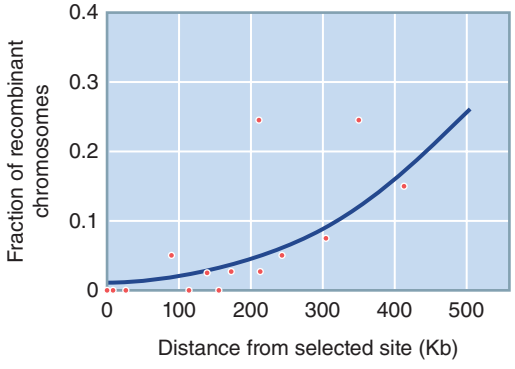


FIGURE 5.23 The fraction of recombinants between an allele of *G6PD* and alleles at nearby loci on a human chromosome remains low, suggesting that the allele has rapidly increased in frequency by positive selection. The allele confers resistance to malaria.

Data from: E. T. Wang, et al. 2006. *Proc Natl Acad Sci USA* 103:135–140.

5.13 A Constant Rate of Sequence Divergence Is a Molecular Clock

KEY CONCEPTS

- The sequences of orthologous genes in different species vary at nonsynonymous sites (where mutations have caused amino acid substitutions) and synonymous sites (where mutation has not affected the amino acid sequence).
- Synonymous substitutions accumulate about 10 times faster than nonsynonymous substitutions.
- The evolutionary divergence between two DNA sequences is measured by the corrected percentage of positions at which the corresponding nucleotides differ.
- Substitutions can accumulate at a more or less constant rate after genes separate, so that the divergence between any pair of globin sequences is proportional to the time since they shared common ancestry.

Most changes in gene sequences occur by mutations that accumulate slowly over time. Point mutations and small insertions and deletions occur by chance, probably with more or less equal probability in all regions of the genome. The exceptions to this are *hotspots*, where mutations occur much more frequently. Recall from the section *DNA Sequences Evolve by Mutation and a Sorting Mechanism* earlier in this chapter that most nonsynonymous mutations are deleterious and will be eliminated by negative selection, whereas the rare advantageous substitution will spread through the population and eventually replace the original sequence (fixation). Neutral variants are expected to be lost or fixed in the population due to random genetic drift. What proportion of mutational changes in a protein-coding gene sequence is selectively neutral is a historically contentious issue.

The rate at which substitutions accumulate is a characteristic of each gene, presumably depending at least in part on its functional flexibility with regard to change. Within a species, a gene evolves by mutation followed by fixation within the single population. Recall that when we study the genetic variation of a species, we see only the variants that

have been maintained, whether by selection or genetic drift. When multiple variants are present they might be stable, or they might in fact be transient because they are in the process of being fixed (or lost).

When a single species separates into two new species, each of the resulting species constitutes an independent evolutionary lineage. By comparing orthologous genes in two species, we see the differences that have accumulated between them since the time when their ancestors ceased to interbreed. Some genes are highly conserved, showing little or no change from species to species. This indicates that most changes are deleterious and therefore eliminated.

The difference between two genes is expressed as their **divergence**, the percentage of positions at which the nucleotides are different, corrected for the possibility of convergent mutations (the same mutation at the same site in two separate lineages) and true revertants. There is usually a difference in the rate of evolution among the three codon positions within genes, because mutations at the third base position often are synonymous, as are some at the first position.

In addition to the coding sequence, a gene contains untranslated regions. Here again, most mutations are potentially neutral, apart from their effects on either secondary structure or (usually rather short) regulatory signals.

Although synonymous mutations are expected to be neutral with regard to the polypeptide, they could affect gene expression via the sequence change in RNA (see the section *DNA Sequences Evolve by Mutation and a Sorting Mechanism* earlier in this chapter). Another possibility is that a change in synonymous codons calls for a different tRNA to respond, influencing the efficiency of translation. Species generally show a **codon bias**; when there are multiple codons for the amino acid, one codon is found in protein-coding genes in a high percentage, whereas the remaining codons are found in low percentages. There is a corresponding percentage difference in the tRNA types that recognize these codons. Consequently, a change from a common to a rare synonymous codon can reduce the rate of translation due to a lower concentration of appropriate tRNAs. (Alternatively, there might be a nonadaptive explanation for codon bias; see the section *There Might Be Biases in Mutation, Gene Conversion, and Codon Usage* later in this chapter.)

Researchers can measure the divergence of proteins (representing nonsynonymous changes in their genes) over time by comparing species for which there is paleontological evidence for the time of their divergence. Such data provide two general observations. First, different proteins evolve at different rates. For example, fibrinopeptides evolve quickly, cytochrome *c* evolves slowly, and hemoglobin evolves at an intermediate rate. Second, for some proteins (including the three just mentioned), the rate of evolution is approximately constant over millions of years. In other words, for a given type of protein, the divergence between any pair of sequences is (more or less) proportional to the time since they shared a common ancestor. This provides a **molecular clock** that measures the accumulation of substitutions at an approximately constant rate during the evolution of a particular protein-coding gene.

There can also be molecular clocks for paralogous proteins diverging within a species lineage. To take the example of the human β- and δ-globin chains (see the section *Globin Clusters Arise by Duplication and Divergence* later in this chapter and the *Clusters and Repeats* chapter), there are 10 differences in 146 amino acids, a divergence of 6.9%. The DNA sequence has 31 changes in 441 nucleotides (7%). However, the nonsynonymous and synonymous changes are distributed very differently. There are 11 changes in the 330 nonsynonymous sites (3.3%), but 20 changes in only 111 synonymous sites (18%). This gives corrected rates of divergence of 3.7% in the nonsynonymous sites and 32% in the synonymous sites, an order of magnitude in difference.

The striking difference in the divergence of nonsynonymous and synonymous sites demonstrates the existence of much greater constraints on nucleotide changes that alter polypeptide sequences compared to those that do not. Many fewer amino acid changes are neutral.

Suppose that we take the rate of synonymous substitutions to indicate the underlying rate of mutational fixation (assuming there is no selection at all at the synonymous sites). Then, over the period since the β and δ genes diverged, there should have been changes at 32% of the 330 nonsynonymous sites, for a total of 105. All but 11 of them have been eliminated, which means that about 90% of the mutations were not retained.

The rate of divergence can be measured as the percent difference per million years or as its reciprocal, the **unit evolutionary period (UEP)**—the time in millions of years that it takes for 1% divergence to accumulate. After the rate of the molecular clock has been established by pairwise comparisons between species (remembering the practical difficulties in establishing the actual time since the existence of the common ancestor), it can be applied to paralogous genes within a species. From their divergence, we can calculate how much time has passed since the duplication that generated them.

By comparing the sequences of orthologous genes in different species, the rate of divergence at both nonsynonymous and synonymous sites can be determined, as plotted in **FIGURE 5.24**.

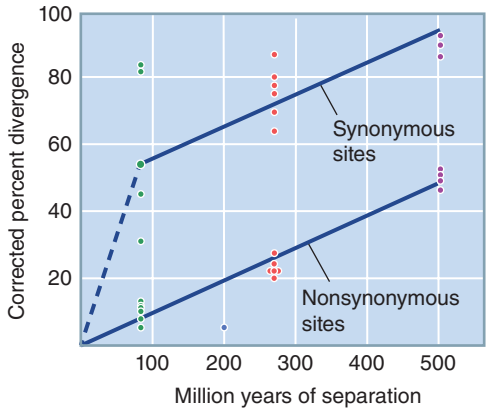


FIGURE 5.24 Divergence of DNA sequences depends on evolutionary separation. Each point on the graph represents a pairwise comparison.

In pairwise comparisons, there is an average divergence of 10% in the nonsynonymous sites of either the α- or β-globin genes of mammal lineages that have been separated since the mammalian radiation occurred roughly 85 million years ago. This corresponds to a nonsynonymous divergence rate of 0.12% per million years.

The rate is approximately constant when the comparison is extended to genes that diverged in the more distant past. For example, the average nonsynonymous divergence between orthologous mammalian and chicken globin genes is 23%. Relative to a common ancestor at roughly 270 million years ago, this gives a rate of 0.09% per million years.

Going farther back, we can compare the α- with the β-globin genes within a species. They have been diverging since the original duplication event about 500 million years ago (see **FIGURE 5.25**). They have an average nonsynonymous divergence of about 50%, which gives a rate of 0.1% per million years.

The summary of these data in Figure 5.24 shows that nonsynonymous divergence in the globin genes has an average rate of about 0.096% per million years (for a UEP of 10.4). Considering the uncertainties in estimating the times at which the species diverged, the results lend good support to the idea that there is a constant molecular clock.

The data on synonymous site divergence are much less clear. In every case, it is evident that the synonymous site divergence is much greater than the nonsynonymous site divergence, by a factor that varies from 2 to 10. However, the range of synonymous site divergences in pairwise comparisons is too great to establish a molecular clock, so we must base temporal comparisons on the nonsynonymous sites.

From Figure 5.24, it is clear that the rate of evolution at synonymous sites is only approximately constant over time. If we assume that there must be zero divergence at zero years of separation, we see that the rate of synonymous site divergence is much greater for the first approximately 100 million years of separation. One interpretation is that roughly half of the synonymous sites are rapidly (within 100 million years) saturated by mutations; this half behaves as neutral sites. The other half accumulates mutations more slowly, at a rate approximately the same as that of the nonsynonymous sites; this half represents sites that are synonymous with regard to the polypeptide but that are under selective constraint for some other reason.

Now we can reverse the calculation of divergence rates to estimate the times since paralogous genes were duplicated. The difference between the human β and α genes is 3.7% for nonsynonymous sites. At a UEP of 10.4, these genes must have diverged 10.4 × 3.7 = about 40 million years ago—about the time of the separation of the major primate lineages: New World monkeys, Old World monkeys, and great apes (including humans). All of these taxonomic groups have both β and δ genes, which suggests that the gene divergence began just before this point in evolution.

Proceeding further back, the divergence between the nonsynonymous sites of γ and ε genes is 10%, which corresponds to a duplication event about 100 million years ago. The separation between embryonic and fetal globin genes therefore might have just preceded or accompanied the mammalian radiation.

An evolutionary tree for the human globin genes is presented in **FIGURE 5.26**. Paralogous groups that evolved before the mammalian radiation—such as the separation of

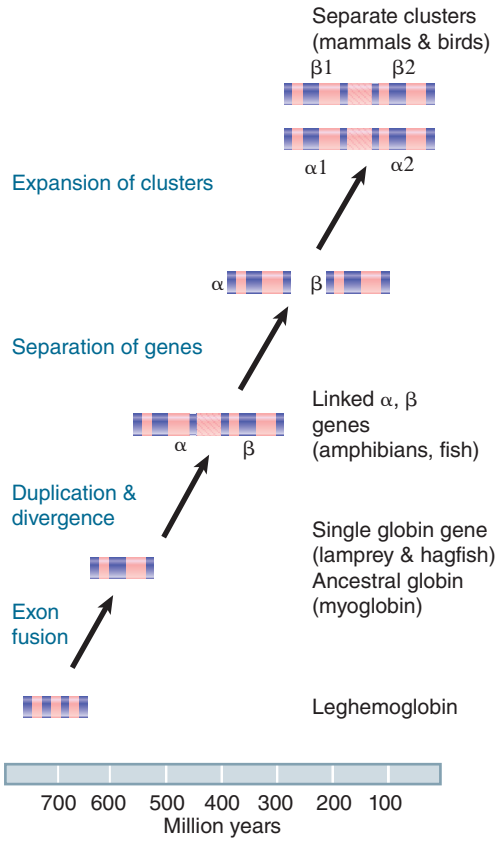


FIGURE 5.25 All globin genes have evolved by a series of duplications, transpositions, and mutations from a single ancestral gene.

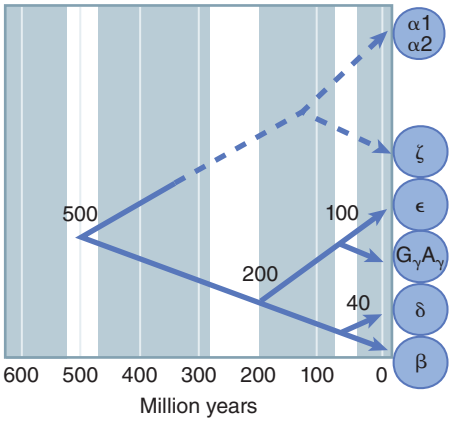


FIGURE 5.26 Nonsynonymous site divergences between pairs of β-globin genes allow the history of the human cluster to be reconstructed. This tree accounts for the separation of classes of globin genes.

β/δ from γ—should be found in all mammals. Paralogous groups that evolved afterward—such as the separation of β- and δ-globin genes—should be found in individual lineages of mammals.

In each species, there have been comparatively recent changes in the structures of the clusters. We know this because we see differences in gene number (one adult β-globin gene in humans, two in the mouse) or in type (most often concerning whether there are separate embryonic and fetal genes).

When sufficient data have been collected on the sequences of a particular gene or gene family, the analysis can be reversed and comparisons between orthologous genes can be used to assess taxonomic relationships. If a molecular clock has been established, the time to common ancestry between the previously analyzed species and a species newly introduced to the analysis can be estimated.

5.14 The Rate of Neutral Substitution Can Be Measured from Divergence of Repeated Sequences

KEY CONCEPT

- The rate of substitution per year at neutral sites is greater in the mouse genome than in the human genome, probably because of a higher mutation rate.

We can make the best estimate of the rate of substitution at neutral sites by examining sequences that do not encode polypeptide. (We use the term *neutral* here rather than *synonymous* because there is no coding potential.) An informative comparison can be made by comparing the members of a common repetitive family in the human and mouse genomes.

The principle of the analysis is summarized in **FIGURE 5.27**. We begin with a family of related sequences that have evolved by duplication and substitution from an original ancestral sequence. We assume that the ancestral sequence can be deduced by taking the base that is most common at each position. Then we can calculate the divergence of each individual family member as the proportion of bases that differ from the deduced ancestral sequence. In this example, individual members vary from 0.13 to 0.18 divergence and the average is 0.16.

One family used for this analysis in the human and mouse genomes derives from a sequence that is thought to have ceased to be functional at about the time of the common ancestor between humans and rodents (the LINEs family; see the *Transposable Elements and Retroviruses* chapter). This means that it has been diverging under limited selective pressure for the same length of time in both species. Its average divergence in humans is about 0.17 substitutions per site, corresponding to a rate of 2.2×10^{-9} substitutions per base per year over the 75 million years since the separation. However, in the mouse genome, neutral substitutions have occurred at twice this rate, corresponding to 0.34 substitutions per site in the family, or a rate of 4.5×10^{-9}. Note, however, that if we calculated the rate per generation instead of per year, it would be greater in humans than in the mouse (2.2×10^{-8} as opposed to 10^{-9}).

These figures probably underestimate the rate of substitution in the mouse; at the time of divergence, the rates in both lineages would have been the same and the difference must have evolved since then. The current rate of neutral substitution per year in the mouse is probably two to three times greater than the historical average. At first glance, these rates would seem to reflect the balance between the occurrence of mutations (which can be higher in species with higher metabolic rates, like the mouse) and the loss of them due to genetic drift, which is largely a function of population size, because genetic drift is a type of "sampling error" where allele frequencies fluctuate more widely in smaller populations. In addition to eliminating neutral alleles more quickly, smaller population sizes also allow faster fixation and loss of neutral alleles. Rodent species tend to have short generation times

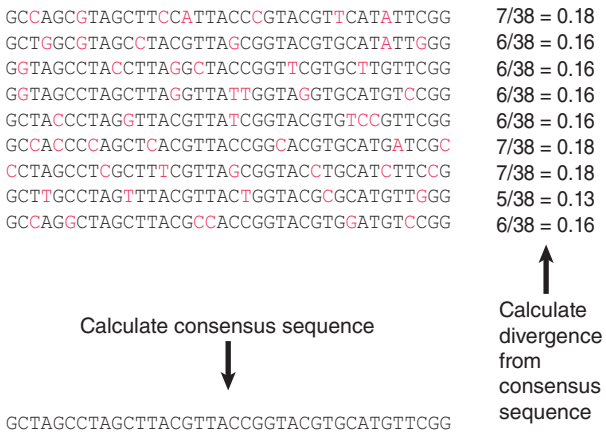


FIGURE 5.27 An ancestral consensus sequence for a family is calculated by taking the most common base at each position. The divergence of each existing current member of the family is calculated as the proportion of bases at which it differs from the ancestral sequence.

(allowing more opportunities for substitutions per year), but species with short generation times also tend to have larger population sizes, so the effects of more substitutions per year but less fixation of neutral alleles would cancel each other out. The higher substitution rate in mice is probably due primarily to a higher mutation rate.

Comparing the mouse and human genomes allows us to assess whether syntenic (homologous) regions show signs of conservation or have differed at the rate predicted from accumulation of neutral substitutions. The proportion of sites that show signs of selection is about 5%. This is much higher than the proportion found in exons (about 1%). This observation implies that the genome includes many more stretches whose sequence is important for functions other than encoding RNA. Known regulatory elements are likely to comprise only a small part of this proportion. This number also suggests that most (i.e., the rest) of the genome sequences do not have any function that depends on the exact sequence.

5.15 How Did Interrupted Genes Evolve?

KEY CONCEPTS

- An interesting evolutionary question is whether genes originated with introns or were originally uninterrupted.
- Interrupted genes that correspond either to proteins or to independently functioning noncoding RNAs probably originated in an interrupted form (the "introns early" hypothesis).
- The interruption allowed base order to better satisfy the potential for stem–loop extrusion from duplex DNA, perhaps to facilitate recombination repair of errors.
- A special class of introns is mobile and can insert themselves into genes.

The structure of many eukaryotic genes suggests a concept of the eukaryotic genome as a sea of mostly unique DNA sequences in which exon "islands" separated by intron "shallows" are strung out in individual gene "archipelagoes." What was the original form of genes?

- The **"introns early" hypothesis** is the proposal that introns have always been an integral part of the gene. Genes originated as interrupted structures, and those now without introns have lost them in the course of evolution.
- The **"introns late" hypothesis** is the proposal that the ancestral protein-coding sequences were uninterrupted and that introns were subsequently inserted into them.

In simple terms, can the difference between eukaryotic and prokaryotic gene organizations be accounted for by the acquisition of introns in the eukaryotes or by the loss of introns in the prokaryotes?

One point in favor of the "introns early" model is that the mosaic structure of genes suggests an ancient combinatorial approach to the construction of genes to encode novel proteins; this is a hypothesis known as **exon shuffling**. Suppose that an early cell had a number of separate protein-coding sequences; it is likely to have evolved by reshuffling different polypeptide units to construct new proteins. Although we recognize the advantages of this mechanism for gene evolution, that does not necessarily mean that it was the primary reason for the *initial* evolution of the mosaic structure. Introns might have greatly assisted, but might not have been critical for, the recombination of protein-coding gene segments. Thus, a disproof of the combinatorial hypothesis would neither disprove the "introns early" hypothesis nor support the "introns late" hypothesis.

If a protein-coding unit (now known as an exon) must be a continuous series of codons, every such reshuffling event would require a precise recombination of DNA to place separate protein-coding units in sequence and in the same reading frame (a one-third probability in any one random joining event). However, if this combination does not produce a functional protein, the cell might be damaged because the original sequence of protein-coding units might have been lost.

The cell might survive, though, if some of the experimental recombination occurs in RNA transcripts, leaving the DNA intact. If a translocation event could place two protein-coding units in the same transcription unit, various RNA splicing "experiments" to combine the two proteins into a single polypeptide chain could be explored. If some combinations are not successful, the original protein-coding units remain available for further trials. In addition, this scenario does not require the two protein-coding units to be recombined precisely into a continuous coding sequence. There is evidence supporting this scenario: Different genes have related exons, as if each gene had been assembled by a process of exon shuffling (see the chapter titled *The Interrupted Gene*).

FIGURE 5.28 illustrates the result of a translocation of a random sequence that includes an exon into a gene. In some organisms, exons are very small compared to introns, so it is likely that the exon will insert within an intron and be flanked by functional 5′ and 3′ splice sites. Splice sites are recognized in sequential pairs, so the splicing mechanism should recognize the 5′ splice site of the original intron and the 3′ splice site of the introduced exon, instead of the 3′ splice site of the original intron. Similarly, the 5′ splice site of the new exon and the 3′ splice site of the original intron might be recognized as a pair, so the new exon will remain between the original two exons in the mature RNA transcript. As long as the new exon is in the same reading frame as the original exons (a one-third probability at each end), a new, longer polypeptide will be produced. Exon shuffling events could have been responsible for generating new combinations of exons during evolution.

Given that it is difficult to envision (1) the assembly of long chains of amino acids by some template-independent process and (2) that such assembled chains would be able to self-replicate, it is widely believed that the most successful early self-replicating molecules were nucleic acids—probably RNA. Indeed, RNA molecules can act both as coding

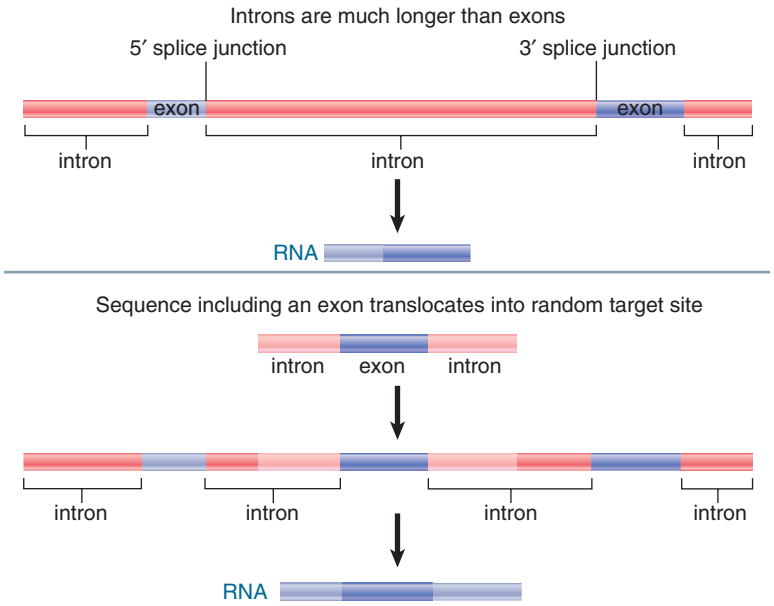


FIGURE 5.28 An exon surrounded by flanking sequences that is translocated into an intron can be spliced into the RNA product.

templates and as catalysts (i.e., *ribozymes*; see the chapter titled *Catalytic RNA*). It was probably by virtue of their catalytic activities that prototypic molecules in the early "RNA world" were able to self-replicate; the templating property would have emerged later.

Many functions mediated by nucleic acid could have competed for genome space in the RNA world. As suggested elsewhere in this text (see the chapter titled *The Interrupted Gene*), these functions can be seen as exerting pressures: AG pressure (the pressure for purine-enrichment in exons); GC pressure (the genome-wide pressure for a distinctive balance between the proportions of the two sets of Watson–Crick pairing bases); single-strand parity pressure (the genome-wide pressure for parity between A and T, and between G and C, in single-stranded nucleic acids); and, probably related to the latter, fold pressure (the genome-wide pressure for single-stranded nucleic acid, whether in free form or extruded from duplex forms, to adopt secondary and higher-order stem–loop structures). For present purposes, the functions served by these pressures need not concern us. The fact that the pressures are so widely spread among organisms suggests important roles in the economy of life (survival and reproduction), rather than mere neutrality.

To these pressures competing for genome space would have been added pressures for increased catalytic activities, ribozyme pressure being supplemented or superseded by protein pressure (the pressure to encode a sequence of amino acids with potential enzymatic activity) after a translation system had evolved. Mutation that happened to generate protein-coding potential would have been favored, but would also be competing against preexisting nucleic acid level pressures. In other words, exons might have been latecomers to an evolving molecular system. Given the redundancy of the genetic code, especially at the third base positions of codons, accommodations could have been explored in the course of evolution so that a protein-encoding region would, to a degree, have been subject to selection by nucleic acid pressures *within itself*. Thus, coding sequences could be selected for both their protein-coding potential and their effects on DNA structure.

Constellations of exons that were *slowly* evolving under negative selection (see the chapter titled *The Interrupted Gene*) would have been able to adapt to accommodate nucleic acid pressures. Exon sequences that could accommodate both protein and nucleic acid pressures would have been conserved. However, those evolving more *rapidly* under positive selection would not have been able to afford this luxury. Thus, some nucleic acid level pressures (e.g., fold pressure) would have been diverted to neighboring introns, resulting in the conservation of the latter.

Some RNA transcripts perform functions by virtue of their secondary and higher-order structures, not by acting as templates for translation. These RNAs, which often interact with proteins, include *Xist* that is involved in X-chromosome inactivation (see the *Epigenetics II* chapter) and the tRNAs and ribosomal RNAs (rRNAs) that facilitate the translation of mRNAs. Generally, these single-stranded RNAs have the same sequence of bases as one strand (the RNA-synonymous strand) of the corresponding DNA.

It is important to note that because these RNAs have structures that serve their distinctive functions (often cytoplasmic), it does not follow that the *same* structures will serve the (nuclear) functions of the corresponding DNAs equally well. Thus, we should not be surprised that, even though there is no ultimate protein product, RNA genes are interrupted and the transcripts are spliced to generate mature RNA products. Similarly, there are sometimes introns in the 5′ and 3′ untranslated regions of pre-mRNAs that must be spliced out.

Therefore, information for the overtly functional parts of genes can be seen as having had to intrude into genomes

that were already adapted to numerous preexisting pressures operating at the nucleic acid level. A reconfiguration of pressures usually could not have occurred if the genic function-encoding parts existed as contiguous sequences. The outcome was that DNA segments corresponding to the genic function-encoding parts were often interrupted by other DNA segments catering to the basic needs of the genome. A further fortuitous outcome would have been a facilitation of the intermixing of functional parts to allow the evolutionary testing of new combinations.

Apart from these pressures on genome space, there are selection pressures acting at the organismal level. For example, birds tend to have shorter introns than mammals, which has led to the controversial hypothesis that there has been selection pressure for compaction of the genome because of the metabolic demands of flight. For many microorganisms (such as bacteria and yeast), evolutionary success can be equated with the ability to rapidly replicate DNA. Smaller genomes can be more rapidly replicated than larger ones, so it might be the pressure for compaction of genomes that led to uninterrupted genes in most microorganisms. Long protein-encoding sequences had to accommodate numerous genomic pressures in addition to protein pressure.

There is evidence that introns have been lost from some members of gene families. See the chapter titled *The Interrupted Gene* for examples from the insulin and actin gene families. In the case of the actin gene family, it is sometimes not clear whether the presence of an intron in a member of the family indicates the ancestral state or an insertion event. Overall, current evidence suggests that genes originally had sequences now called introns but can evolve with both the loss and gain of introns.

Organelle genomes show the evolutionary connections between prokaryotes and eukaryotes. There are many general similarities between mitochondria or chloroplasts and certain bacteria because those organelles originated by endosymbiosis, in which a bacterial cell lived within the cytoplasm of a eukaryotic prototype. Although there are similarities to bacterial genetic processes—such as protein and RNA synthesis—some organelle genes possess introns and therefore resemble eukaryotic nuclear genes. Introns are found in several chloroplast genes, including some that are homologous to *E. coli* genes. This suggests that the endosymbiotic event occurred before introns were lost from the prokaryotic lineage.

Mitochondrial genome comparisons are particularly striking. The genes of yeast and mammalian mitochondria encode virtually identical proteins in spite of a considerable difference in gene organization. Vertebrate mitochondrial genomes are very small and extremely compact, whereas yeast mitochondrial genomes are larger and have some complex interrupted genes. Which is the ancestral form? Yeast mitochondrial introns (and certain other introns) can be mobile—they are independent sequences that can splice out of the RNA and insert DNA copies elsewhere—which suggests that they might have arisen by insertions into the genome (see the *Catalytic RNA* chapter). Even though most evidence supports "introns early," there is reason to believe that, in addition to the introduction of mobile elements, ongoing accommodations to various extrinsic and intrinsic (genomic) pressures might result, from time to time, in the emergence of new introns ("introns late").

As for the role of introns, it is easy to dismiss intronic characteristics such as an enhanced potential to extrude stem–loop structures as an adaptation to assist accurate splicing. An analogy has been drawn between the transmission of genic messages and the transmission of electronic messages, in which a message sequence is normally interrupted by error-correcting codes. Although there is no evidence that similar types of code operate in genomes, it is possible that fold pressure arose to aid in the detection and correction of sequence errors by recombination repair. So important would be the latter that in many circumstances fold pressure might trump protein pressure (see the *Repair Systems* chapter).

5.16 Why Are Some Genomes So Large?

KEY CONCEPTS

- There is no clear correlation between genome size and genetic complexity.
- There is an increase in the minimum genome size associated with organisms of increasing complexity.
- There are wide variations in the genome sizes of organisms within many taxonomic groups.

The total amount of DNA in the (haploid) genome is a characteristic of each living species known as its **C-value**. There is enormous variation in the range of C-values, from less than 10^6 base pairs (bp) for a mycoplasma to more than 10^{11} bp for some plants and amphibians.

FIGURE 5.29 summarizes the range of C-values found in different taxonomic groups. There is an increase in the *minimum* genome size found in each group as the complexity

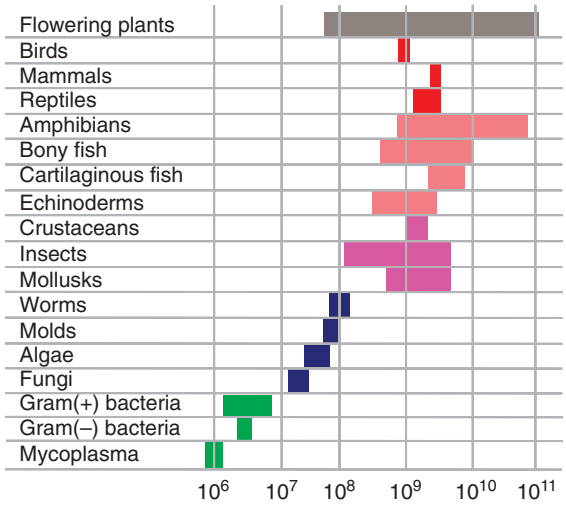


FIGURE 5.29 DNA content of the haploid genome increases with morphological complexity of lower eukaryotes, but varies extensively within some groups of animals and plants. The range of DNA values within each group is indicated by the shaded area.

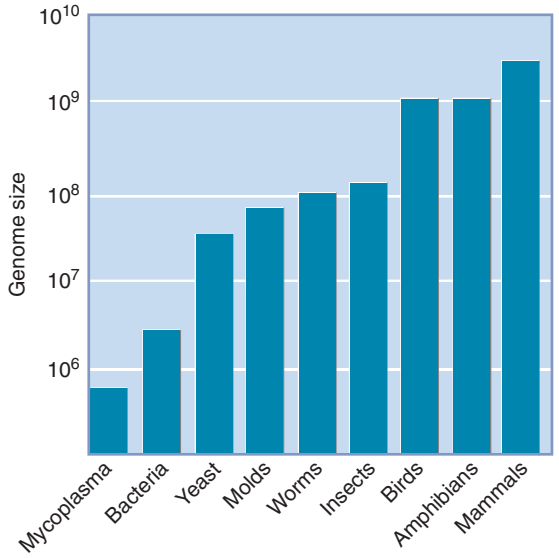


FIGURE 5.30 The minimum genome size found in each taxonomic group increases from prokaryotes to mammals.

increases. Although C-values are greater in the multicellular eukaryotes, we do see some wide variations in the genome sizes within some groups.

Plotting the minimum amount of DNA required for a member of each group suggests in **FIGURE 5.30** that an increase in genome size is required for increased complexity in prokaryotes, fungi, and invertebrate animals.

Mycoplasma are the smallest prokaryotes and have genomes only about three times the size of a large bacteriophage and smaller than those of some megaviruses. More typical bacterial genome sizes start at about 2×10^6 bp. Unicellular eukaryotes (whose lifestyles can resemble those of prokaryotes) also get by with genomes that are small, although their genomes are larger than those of most bacteria. However, being eukaryotic does not imply a vast increase in genome size, per se; a yeast can have a genome size of about 1.3×10^7 bp, which is only about twice the size of an average bacterial genome.

A further twofold increase in genome size is adequate to support the slime mold *Dictyostelium discoideum,* which is able to live in either unicellular or multicellular modes. Another increase in complexity is necessary to produce the first fully multicellular organisms; the nematode worm *C. elegans* has a DNA content of 8×10^7 bp.

We also can see the steady increase in genome size with complexity in the listing in **TABLE 5.5** of some of the most commonly studied organisms. It is necessary for insects, birds, amphibians, and mammals to have larger genomes than those of unicellular eukaryotes. However, after this point there is no clear relationship between genome size and morphological complexity of the organism.

We know that eukaryotic genes are much larger than the sequences needed to encode polypeptides because exons might comprise only a small part of the total length of a gene. This explains why there is much more DNA than is needed to provide reading frames for all the proteins of the organism. Large parts of an interrupted gene might not encode amino

TABLE 5.5 The genome sizes of some commonly studied organisms.

Phylum	Species	Genome
Algae	*Pyrenomas salina*	6.6×10^5
Mycoplasma	*M. pneumoniae*	1.0×10^6
Bacterium	*E. coli*	4.2×10^6
Yeast	*S. cerevisiae*	1.3×10^7
Slime mold	*D. discoideum*	5.4×10^7
Nematode	*C. elegans*	8.0×10^7
Insect	*D. melanogaster*	1.8×10^8
Bird	*G. domesticus*	1.2×10^9
Amphibian	*X. laevis*	3.1×10^9
Mammal	*H. sapiens*	3.3×10^9

acids. In addition, in multicellular organisms there can be significant lengths of DNA between genes, some of which function in gene regulation. So it might not be possible to deduce anything about the number of genes or the complexity of the organism from the overall size of the genome.

The **C-value paradox** refers to the lack of correlation between genome size and genetic and morphological complexity (e.g., the number of different cell types). There are some extremely curious observations about relative genome size, such as that the toad *Xenopus* and humans have genomes of essentially the same size. In some taxonomic groups there are large variations in DNA content between organisms that do not vary much in complexity, as seen in Figure 5.29. (This is especially marked in insects, amphibians, and plants, but does not occur in birds, reptiles, and mammals, which all show little variation within the group—an approximately 23-fold range of genome sizes.) A cricket has a genome 11 times the size of that of a fruit fly. In amphibians, the smallest genomes are less than 10^9 bp, whereas the largest are about 10^{11} bp. There is unlikely to be a large difference in the number of genes needed for the development of these amphibians. Some fish species have about the same number of genes as mammals have, but other fish genomes (such as that of the pufferfish fugu) are more compact, with smaller introns and shorter intergenic spaces. Still others are tetraploid. The extent to which this variation is selectively neutral or subject to natural selection is not yet fully understood.

In mammals, additional complexity is also a consequence of the alternative splicing of genes that allows two or more protein variants to be produced from the same gene

(see the chapter titled *RNA Splicing and Processing*). With such mechanisms, increased complexity need not be accompanied by an increased number of genes.

5.17 Morphological Complexity Evolves by Adding New Gene Functions

KEY CONCEPTS

- In general, comparisons of eukaryotes to prokaryotes, multicellular to unicellular eukaryotes, and vertebrate to invertebrate animals show a positive correlation between gene number and morphological complexity as additional genes are needed with generally increased complexity.
- Most of the genes that are unique to vertebrates are involved with the immune or nervous systems.

Comparison of the human genome sequence with sequences found in other species is revealing about the process of evolution. **FIGURE 5.31** shows an analysis of human genes according to the breadth of their distribution among all cellular organisms. Beginning with the most generally distributed (upper-right corner of the figure), about 21% of genes are common to eukaryotes and prokaryotes. These tend to encode proteins that are essential for all living forms—typically basic metabolism, replication, transcription, and translation. Moving clockwise, another approximately 32% of genes are found in eukaryotes in general—for example, they can be found in yeast. These tend to encode proteins involved in functions that are general to eukaryotic cells but not to bacteria—for example, they might be concerned with the activities of organelles or cytoskeletal components. Another approximately 24% of genes are generally found in animals. These include genes necessary for multicellularity and for development of different tissue types. Approximately 22% of genes are unique to vertebrate animals. These mostly encode proteins of the immune and nervous systems; they encode very few enzymes, consistent with the idea that enzymes have ancient origins, and that metabolic pathways originated early in evolution. Therefore, we see that the evolution of more complex morphology and specialization requires the addition of groups of genes representing the necessary new functions.

One way to define essential proteins is to identify the proteins present in all proteomes. Comparing the human proteome in more detail with the proteomes of other organisms, 46% of the yeast proteome, 43% of the worm proteome, and 61% of the fruit fly proteome are represented in the human proteome. A key group of about 1,300 proteins is present in all four proteomes. The common proteins are basic "housekeeping" proteins required for essential functions, falling into the types summarized in **FIGURE 5.32**. The main functions are transcription and translation (35%), metabolism (22%), transport (12%), DNA replication and

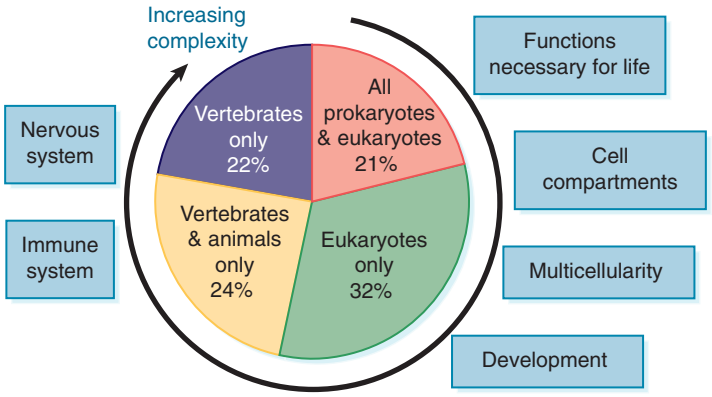


FIGURE 5.31 Human genes can be classified according to how widely their homologs are distributed in other species.

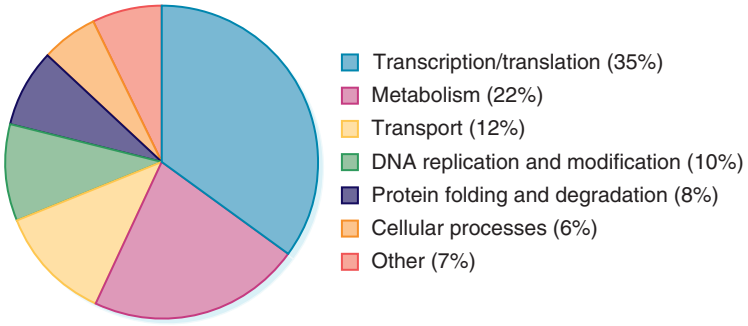


FIGURE 5.32 Common eukaryotic proteins are involved with essential cellular functions.

modification (10%), protein folding and degradation (8%), and cellular processes (6%), with the remaining 7% dedicated to various other functions.

One of the striking features of the human proteome is that it has many unique proteins compared with those of other eukaryotes but has relatively few unique protein domains (portions of proteins having a specific function). Most protein domains appear to be common to the animal kingdom. However, there are many unique protein architectures, defined as unique combinations of domains. **FIGURE 5.33** shows that the greatest proportion of unique proteins consists of transmembrane and extracellular proteins. In yeast, the majority of architectures are associated with intracellular proteins. There are about twice as many intracellular architectures in fruit flies (or nematode worms), but there is a strikingly higher proportion of transmembrane and extracellular proteins, as might be expected from the additional functions required for the interactions between the cells of a multicellular organism. The additions in intracellular architectures required in a vertebrate (typified by the human genome) are relatively small, but there is, again, a higher proportion of transmembrane and extracellular architectures.

It has long been known that the genetic difference between humans and chimpanzees (our nearest relative) is very small, with 98.5% identity between genomes. The sequence of the chimpanzee genome now allows us to investigate the 1.5% of differences in more detail to see whether features responsible for "humanity" can be identified. (Genome sequences for the nonhuman primates orangutan and gorilla as well as the Paleolithic human species of Neanderthals and Denisovans are also now available for comparison.) The comparison shows 35×10^6 nucleotide substitutions (1.2% sequence difference overall), 5×10^6 deletions or insertions (making about 1.5% of the euchromatic sequence specific to each species), and many chromosomal rearrangements. Homologous proteins are usually very similar: 29% are identical, and in most cases there are only one or two amino acid differences between the species in the protein. In fact, nucleotide substitutions occur less often in genes encoding polypeptides than are likely to be involved in specifically human traits, suggesting that protein evolution is not a major factor in human–chimpanzee differences. This leaves larger-scale changes in gene structure and/or changes in gene regulation as the major candidates. Some 25% of nucleotide substitutions occur in CpG dinucleotides (among which are many potential regulator sites).

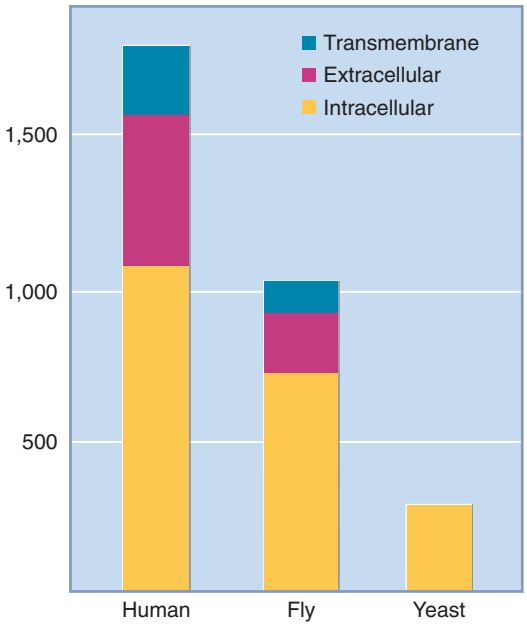


FIGURE 5.33 Increasing complexity in eukaryotes is accompanied by accumulation of new proteins for transmembrane and extracellular functions.

5.18 Gene Duplication Contributes to Genome Evolution

KEY CONCEPT

- Duplicated genes can diverge to generate different genes, or one copy might become an inactive pseudogene.

Exons act as modules for building genes that are tried out in the course of evolution in various combinations (see the chapter titled *The Interrupted Gene*). At one extreme, an individual exon from one gene might be copied and used in another gene. At the other extreme, an entire gene, including both exons and introns, might be duplicated. In such a case, mutations can accumulate in one copy without elimination by natural selection as long as the other copy is under selection to remain functional. The selectively neutral copy might then evolve to a new function, become expressed at a different time or in a different cell type from the first copy, or become a nonfunctional pseudogene.

FIGURE 5.34 summarizes the present view of the rates at which these processes occur. There is about a 1% probability

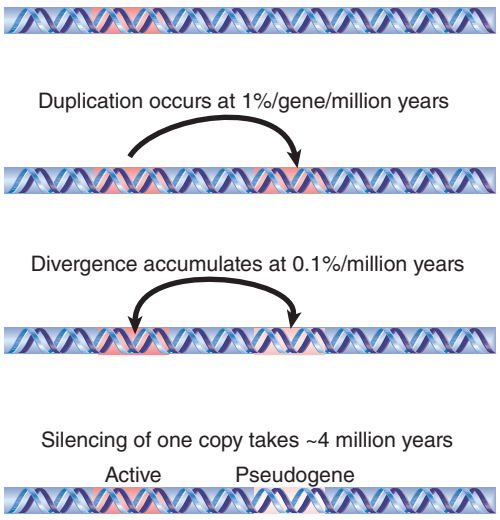


FIGURE 5.34 After a globin gene has been duplicated, differences can accumulate between the copies. The genes can acquire different functions or one of the copies may become a nonfunctional pseudogene.

that a particular gene will be included in a duplication in a period of 1 million years. After the gene has duplicated, differences evolve as the result of the occurrence of different mutations in each copy. These accumulate at a rate of about 0.1% per million years (see the section *A Constant Rate of Sequence Divergence Is a Molecular Clock* earlier in this chapter).

Unless the gene encodes a product that is required in high concentration in the cell, the organism is not likely to need to retain two identical copies of the gene. As differences evolve between the duplicated genes, one of two types of event is likely to occur:

- Both of the gene copies remain necessary. This can happen either because the differences between them generate proteins with different functions, or because they are expressed specifically at different times or in different cell types.
- If this does not happen, one of the genes is likely to become a pseudogene because it will by chance gain a deleterious mutation and there will be no purifying selection to eliminate this copy, so by genetic drift the mutant version might increase in frequency and fix in the species. Typically, this takes about 4 million years for globin genes; in general, the time to fixation of a neutral mutant depends on the generation time and the effective population size, with genetic drift being a stronger force in smaller populations. In such a situation, it is purely a matter of chance which of the two copies becomes nonfunctional. (This can contribute to incompatibility between different individuals, and ultimately to speciation, if different copies become nonfunctional in different populations.)

Analysis of the human genome sequence shows that about 5% of the genome comprises duplications of identifiable segments ranging in length from 10 to 300 kb. These duplications have arisen relatively recently; that is, there has not been sufficient time for divergence between them for their homology to become obscured. They include a proportional share (about 6%) of the expressed exons, which shows that the duplications are occurring more or less without regard to genetic content. The genes in these duplications might be especially interesting because of the implication that they have evolved recently and therefore could be important for recent evolutionary developments (such as the separation of the human lineage from that of other primates).

5.19 Globin Clusters Arise by Duplication and Divergence

KEY CONCEPTS

- All globin genes are descended by duplication and mutation from an ancestral gene that had three exons.
- The ancestral gene gave rise to myoglobin, leghemoglobin, and α- and β-globins.
- The α- and β-globin genes separated in the period of early vertebrate evolution, after which duplications generated the individual clusters of separate α- and β-like genes.
- When a gene has been inactivated by mutation, it can accumulate further mutations and become a pseudogene (ψ), which is homologous to the functional gene(s) but has no functional role (or at least has lost its original function).

The most common type of gene duplication generates a second copy of the gene close to the first copy. In some cases, the copies remain associated and further duplication can generate a cluster of related genes. The best characterized example of a gene cluster is that of the globin genes, which constitute an ancient gene family fulfilling a function that is central to animals: the transport of oxygen.

The major constituent of the vertebrate red blood cell is the globin tetramer, which is associated with its heme (iron-binding) group in the form of hemoglobin. Functional globin genes in all species have the same general structure: They are divided into three exons. Researchers conclude that all globin genes have evolved from a single ancestral gene, and by tracing the history of individual globin genes within and between species we can learn about the mechanisms involved in the evolution of gene families.

In red blood cells of adult mammals, the globin tetramer consists of two identical α chains and two identical β chains. Embryonic red blood cells contain hemoglobin tetramers that are different from the adult form. Each tetramer contains two identical α-like chains and two identical β-like chains, each of which is related to the adult polypeptide and is later replaced by it in the adult form of the protein. This is an example of developmental control, in which different genes are successively switched on and off to provide alternative products that fulfill the same function at different times.

The division of globin chains into α-like and β-like reflects the organization of the genes. Each type of globin is encoded by genes organized into a single cluster. The structures of the two clusters in the primate genome are illustrated in **FIGURE 5.35**. Pseudogenes are indicated by the symbol ψ.

Stretching over 50 kb, the β cluster contains 5 functional genes (ε, two γ, δ, and β) and one nonfunctional pseudogene (ψβ). The two γ genes differ in their coding sequence in only one amino acid: The G variant has glycine at position 136, whereas the A variant has alanine.

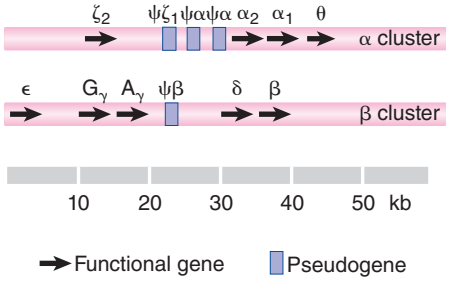


FIGURE 5.35 Each of the α-like and β-like globin gene families is organized into a single cluster, which includes functional genes and pseudogenes (ψ).

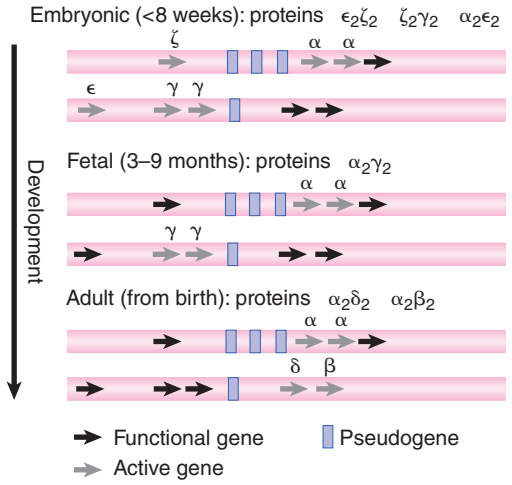


FIGURE 5.36 Different hemoglobin genes are expressed during embryonic, fetal, and adult periods of human development.

The more compact α cluster extends over 28 kb and includes one functional ζ gene, one nonfunctional ζ pseudogene, two α genes, two nonfunctional α pseudogenes, and the θ gene of unknown function. The two α genes encode the same protein. Two (or more) identical genes present on the same chromosome are described as **nonallelic genes**.

The details of the relationship between embryonic and adult hemoglobins vary with the species. The human pathway has three stages: embryonic, fetal, and adult. The distinction between embryonic and adult is common to mammals, but the number of preadult stages varies. In humans, ξ and α are the two α-like chains. The β-like chains are γ, δ, and β. **FIGURE 5.36** shows how the chains are expressed at different stages of development. There is also tissue-specific expression associated with the developmental expression: Embryonic hemoglobin genes are expressed in the yolk sac, fetal genes are expressed in the liver, and adult genes are expressed in bone marrow.

In the human pathway, ζ is the first α-like chain to be expressed, but it is soon replaced by α. In the β-pathway, ε and γ are expressed first, with δ and β replacing them later. In adults, the $\alpha_2\beta_2$ form provides 97% of the hemoglobin, $\alpha_2\delta_2$ provides about 2%, and about 1% is provided by persistence of the fetal form $\alpha_2\gamma_2$.

What is the significance of the differences between embryonic and adult globins? The embryonic and fetal forms have a higher affinity for oxygen, which is necessary to obtain oxygen from the mother's blood. This helps to explain why there is no direct equivalent (although there is temporal expression of globins) in, for example, the chicken, for which the embryonic stages occur outside the mother's body (i.e., within the egg).

Functional genes are defined by their transcription to RNA and ultimately (for protein-coding genes) by the polypeptides they encode. Pseudogenes are defined as having lost their ability to produce functional versions of polypeptides they originally encoded. The reasons for their inactivity vary: The deficiencies might be in transcription, translation, or both. A similar general organization is found in all vertebrate globin gene clusters, but details of the types, numbers, and order of genes all vary, as illustrated in **FIGURE 5.37**. Each cluster contains both embryonic and adult genes. The total lengths of the clusters vary widely. The longest known cluster is found in the goat genome, where a basic cluster of four genes has been duplicated twice. The distribution of functional genes and pseudogenes differs in each case, illustrating the random nature of the evolution of one copy of a duplicated gene to a pseudogene.

The characterization of these gene clusters makes an important general point. There can be more members of a gene family, both functional and nonfunctional, than we would suspect on the basis of protein analysis. The extra functional genes might represent duplicates that encode identical polypeptides, or they might be related to—but different from—known proteins (and presumably expressed only briefly or in low amounts).

With regard to the question of how much DNA is needed to encode a particular function, we see that encoding the β-like globins requires a range of 20 to 120 kb in different mammals. This is much greater than we would expect just from scrutinizing the known β-globin proteins or even from considering the individual genes. However, clusters of this type are not common; most genes are found as individual loci.

From the organization of globin genes in a variety of species, we should be able to trace the evolution of present globin gene clusters from a single ancestral globin gene. Our present view of the evolutionary history was pictured in Figure 5.25.

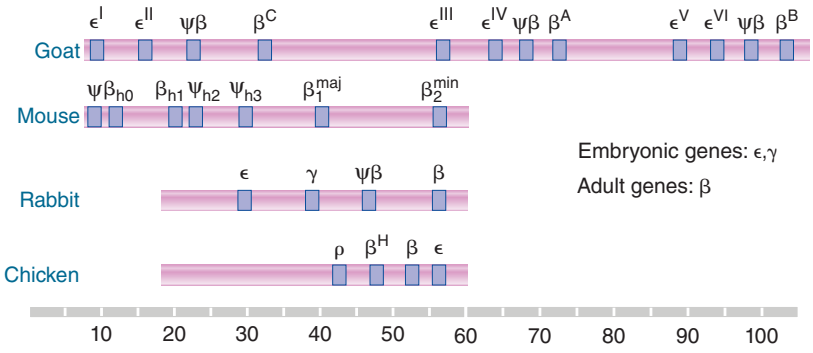


FIGURE 5.37 Clusters of β-globin genes and pseudogenes are found in vertebrates. Seven mouse genes include two early embryonic genes, one late embryonic gene, two adult genes, and two pseudogenes. Rabbits and chickens each have four genes.

The leghemoglobin gene of plants, which is related to the globin genes, might provide some clues about the ancestral form, though of course the modern leghemoglobin gene has evolved for just as long as the animal globin genes. (Leghemoglobin is an oxygen carrier found in the nitrogen-fixing root nodules of legumes.) The furthest back that we can trace a true globin gene is to the sequence of the single chain of mammalian myoglobin, which diverged from the globin lineage about 800 million years ago in the ancestors of vertebrates. The myoglobin gene has the same organization as globin genes, so we can take the three-exon structure to represent that of their common ancestor.

Some members of the class *Chondrichthyes* (cartilaginous fish) have only a single type of globin chain, so they must have diverged from the lineage of other vertebrates before the ancestral globin gene was duplicated to give rise to the α and β variants. This appears to have occurred about 500 million years ago, during the evolution of the *Osteichthyes* (bony fish).

The next stage of globin evolution is represented by the state of the globin genes in the amphibian *Xenopus laevis,* which has two globin clusters. However, each cluster contains both α and β genes, of both larval and adult types. Therefore, the cluster must have evolved by duplication of a linked α–β pair, followed by divergence between the individual copies. Later, the entire cluster was duplicated.

The amphibians separated from the reptilian/mammalian/avian line about 350 million years ago, so the separation of the α- and β-globin genes must have resulted from a transposition in the reptilian/mammalian/avian forerunner after this time. This probably occurred in the period of early tetrapod evolution. There are separate clusters for α- and β-globins in both birds and mammals; therefore the α and β genes must have been physically separated before the mammals and birds diverged from their common ancestor, an event estimated to have occurred about 270 million years ago. Evolutionary changes have taken place within the separate α and β clusters in more recent times, as we saw from the description of the divergence of the individual genes in the section *A Constant Rate of Sequence Divergence Is a Molecular Clock* earlier in this chapter.

5.20 Pseudogenes Have Lost Their Original Functions

KEY CONCEPTS

- Processed pseudogenes result from reverse transcription and integration of mRNA transcripts.
- Nonprocessed pseudogenes result from incomplete duplication or second-copy mutation of functional genes.
- Some pseudogenes might gain functions different from those of their parent genes, such as regulation of gene expression, and take on different names.

As discussed earlier in this chapter, pseudogenes are copies of functional genes that have altered or missing regions such that they presumably do not produce polypeptide products with the original function; they can be nonfunctional or have altered function, and the RNA products might serve regulatory functions. For example, as compared to their functional counterparts, many pseudogenes have frameshift or nonsense mutations that disable their protein-coding functionality. There are two types of pseudogenes characterized by their modes of origin.

Processed pseudogenes result from the reverse transcription of mature mRNA transcripts into cDNA copies, followed by their integration into the genome. This might occur at a time when active reverse transcriptase is present in the cell, such as during active retroviral infection or retroposon activity (see the *Transposable Elements and Retroviruses* chapter). The transcript has undergone processing (see the *RNA Splicing and Processing* chapter), so a processed pseudogene usually lacks the regulatory regions necessary for normal expression. Although it initially contains the coding sequence of a functional polypeptide, it is nonfunctional as soon as it is formed. Such pseudogenes also lack introns and may contain the remnant of the mRNA's poly(A) tail (see the *RNA Splicing and Processing* chapter) as well as the flanking direct repeats characteristic of insertion of retroelements (see the *Transposable Elements and Retroviruses* chapter).

The second type, **nonprocessed pseudogenes**, arises from inactivating mutations in one copy of a multiple-copy or single-copy gene or from incomplete duplication of a functional gene. Often, these are formed by mechanisms that result in tandem duplications. An example of a β-globin pseudogene is shown in **FIGURE 5.38**. If a gene is duplicated in its entirety with intact regulatory regions, there can be two functional copies for a time, but inactivating mutations in one copy would not necessarily be subject to negative selection. Thus, gene families are ripe for the origin of nonprocessed pseudogenes, as evidenced by the existence of several pseudogenes in the globin gene family (see the section *Globin Clusters Arise by Duplication and Divergence* earlier in this chapter). Alternatively, an incomplete duplication of a functional gene, resulting in a copy missing regulatory regions and/or coding sequence, would be "dead on arrival" as an instant pseudogene.

There are approximately 20,000 pseudogenes in the human genome. Ribosomal protein (RP) pseudogenes comprise a large family of pseudogenes, with approximately 2,000 copies. These are processed pseudogenes; presumably the high copy number is a function of the high expression rate of the approximately 80 copies of functional RP genes. Their insertion into the genome is apparently mediated by the L1 retrotransposon (see the *Transposable Elements and Retroviruses* chapter). RP genes are highly conserved among species, so it is possible to identify RP pseudogene orthologs in species with a long history of separate evolution and for which whole genome sequences are available. For example, as shown in **TABLE 5.6**, more than two-thirds of human RP pseudogenes are also found in the chimpanzee genome, whereas less than a dozen are shared between humans and rodents. This suggests that most RP pseudogenes are of more recent origin in both primates and rodents, and that most ancestral RP pseudogenes have been lost by deletion or mutational decay beyond recognition.

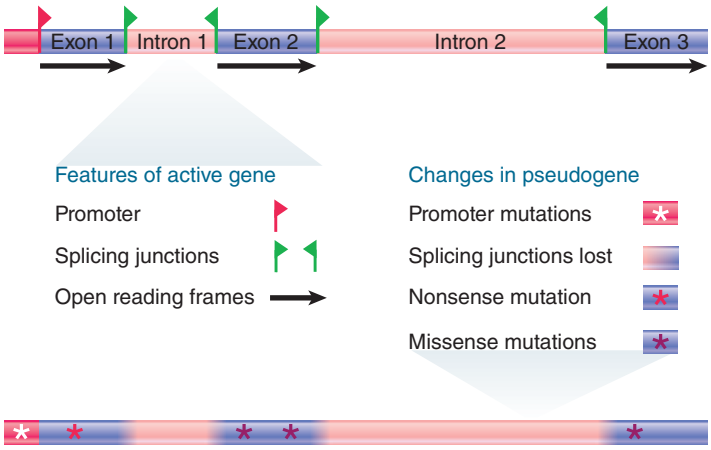


FIGURE 5.38 Many changes have occurred in a β-globin gene since it became a pseudogene.

Interestingly, the rate of evolution of RP pseudogenes is slower than that of the neutral rate (as determined by the rate of substitution in ancient repeats across the genome), suggesting negative selection and implying a functional role for RP pseudogenes. Although pseudogenes are nonfunctional when initially formed, there are clear examples of former pseudogenes (originally identified as pseudogenes because of sequence differences with their functional counterparts that would presumably render them nonfunctional) becoming *neofunctionalized* (taking on a new function) or *subfunctionalized* (taking on a subfunction or complementary function of the parent gene). When functional again, they would be subject to selection and thus evolve more slowly than expected under a neutral model.

How might a pseudogene gain a new function? One possibility is that translation, but not transcription, of the pseudogene has been disabled. The pseudogene encodes an RNA transcript that is no longer translatable but can affect expression or regulation of the still-functional "parent" gene. In the mouse, the processed pseudogene *Makorin1-p1* stabilizes transcripts of the functional *Makorin1* gene. Several endogenous siRNAs (see the *Regulatory RNA* chapter) are encoded by pseudogenes. A second possibility is that a processed pseudogene might be inserted in a location that provides them with new regulatory regions, such as transcription factor binding sites, which allow them to be expressed in a tissue-specific manner unlike that of the parent gene.

TABLE 5.6 Most human *RP* pseudogenes are of recent origin; many are shared with the chimpanzee but absent from rodents.

Human–chimpanzee	1282
Human–mouse	6
Human–rat	11
Mouse–rat	494

Data from S. Balasubramanian, et al., *Genome Biol.* 20 (2009): R2.

5.21 Genome Duplication Has Played a Role in Plant and Vertebrate Evolution

KEY CONCEPTS

- Genome duplication occurs when polyploidization increases the chromosome number by a multiple of two.
- Genome duplication events can be obscured by the evolution and/or loss of duplicates as well as by chromosome rearrangements.
- Genome duplication has been detected in the evolutionary history of many flowering plants and of vertebrate animals.

As discussed in the section *Gene Duplication Contributes to Genome Evolution* earlier in this chapter, genomes can evolve by duplication and divergence of individual genes or of chromosomal segments carrying blocks of genes. However, it appears that some of the major metazoan lineages have had genome duplications in their evolutionary histories. Genome duplication is accomplished by **polyploidization**, as when a tetraploid ($4n$) variety arises from a diploid ($2n$) ancestral lineage.

There are two major mechanisms of polyploidization. **Autopolyploidy** occurs when a species endogenously gives rise to a polyploid variety; this usually involves fertilization by unreduced gametes. **Allopolyploidy** is a result of hybridization between two reproductively compatible species such that diploid sets of chromosomes from both parental species

are retained in the hybrid offspring. As with autopolyploids, the process generally involves the accidental production of unreduced gametes. In both cases, new tetraploids are usually reproductively isolated from the diploid parental species because backcrossed hybrids are triploid and sterile, as some chromosomes are without homologs during meiosis.

Following the successful establishment of a polyploidy species, many mutations can be essentially neutral. As with gene duplications, nonsynonymous substitutions are "covered" by the redundant functional copy of the same gene. In the case of a genome duplication, the deletion of a gene or chromosomal segment or the loss of a chromosome pair might have little phenotypic effect. In addition to the loss of chromosomal segments, chromosomal rearrangements such as inversions and translocations will shuffle the locations and orders of blocks of genes. Over a long period of time, such events can obscure ancestral polyploidization. However, there might still be evidence of polyploidization in the presence of redundant chromosomes or chromosomal segments within a genome.

One successful approach to detecting ancient polyploidization is to compare many pairs of paralogous (duplicated) genes within a species and establish an age distribution of gene duplication events. Many events of approximately the same age can be taken as evidence of polyploidization. As seen in **FIGURE 5.39**, genome duplication events will appear as peaks above the general pattern of random events of gene duplication and copy loss. This approach, along with an analysis of chromosomal locations of gene duplications, suggests that the evolutionary histories of the unicellular yeast *S. cerevisiae* and many flowering plants include one or more genome duplication events. The genetic model of the land plant *Arabidopsis thaliana*, for example, has a history of two, or possibly three, polyploidization events.

Because polyploidization is more common in plants than in animals, it is not surprising that most detected examples of genome duplication are in plant species. However, genome duplication appears to have played an important role in vertebrate evolution, specifically in ray-finned fishes. As evidence, the zebrafish genome contains seven *Hox* clusters as compared to four clusters in tetrapod genomes, suggesting that there was a tetraploidization event followed by secondary loss of one cluster. The analysis of other fish genomes suggests that this event occurred before the diversification of this taxonomic group. The presence of four *Hox* clusters in tetrapods (and at least four in other vertebrates), together with the observation of other shared gene duplications as compared to invertebrate animal genomes, itself suggests that there might have been two major polyploidization events prior to the evolution of vertebrates. In reference to "two rounds of polyploidization," this has been termed the **2R hypothesis**.

This hypothesis leads to the prediction that many vertebrate genes, like the *Hox* clusters, will be found in four times the copy number as compared to their orthologs in invertebrate species. The subsequent observation that less than 5% of vertebrate genes show this 4:1 ratio seems weak support for the hypothesis at best. However, it is to be expected that after nearly 500 million years of evolution, many of the additional copies of genes would have been deleted, evolved significantly to take on new functions, or become pseudogenes and decayed beyond recognition. Stronger support, however, comes from analyses that take into account the map position of duplications that date to the time of the common ancestor of vertebrates. The ancient gene duplications that do show the 4:1 pattern tend to be found in clusters, even after a half-billion years of chromosomal rearrangements. The vertebrates evidently began their evolutionary history as octoploids. The 2R hypothesis is tempting as an explanation for the burst of morphological complexity that accompanied the evolution of vertebrates, although as yet there is little evidence of a direct correlation between the genomic and morphological changes in this taxonomic group.

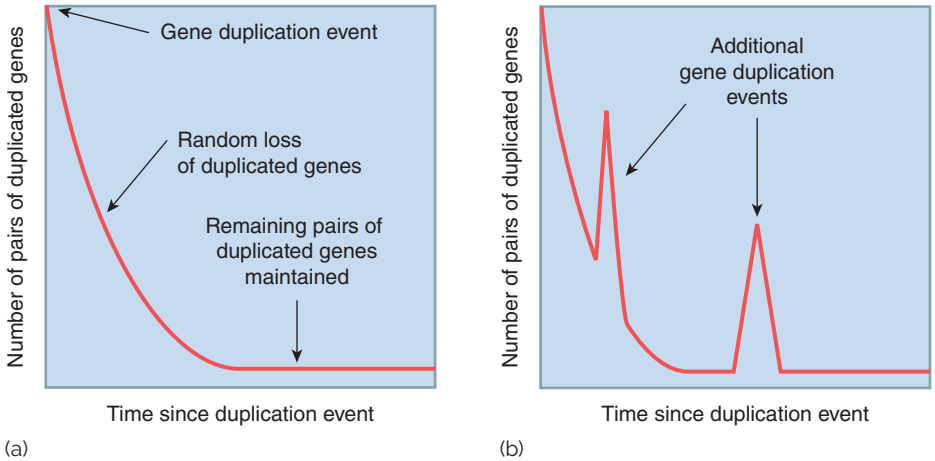


FIGURE 5.39 **(a)** A constant rate of gene duplication and loss shows an exponentially decreasing age distribution of duplicated gene pairs. **(b)** A genome duplication event shows a secondary peak in the age distribution as many genes are duplicated at the same time.

Data from: Blanc, G. and Wolfe, K. H. 2004. *Plant Cell* 16:1667–1678.

5.22 What Is the Role of Transposable Elements in Genome Evolution?

KEY CONCEPT

- Transposable elements tend to increase in copy number when introduced to a genome but are kept in check by negative selection and transposition regulation mechanisms.

Transposable elements (TEs) are mobile genetic elements that can be integrated into the genome at multiple sites and (for some elements) also excised from an integration site. (See the chapter titled *Transposable Elements and Retroviruses* for an extensive discussion of the types and mechanisms of TEs.) The insertion of a TE at a new site in the genome is called **transposition**. One type of TE, the retrotransposon, transposes via an RNA intermediate; a new copy of the element is created by transcription, followed by reverse transcription to DNA and subsequent integration at a new site.

Most TEs integrate at sequences that are random (at least with respect to their functions). As such, they are a major source of the problems associated with insertion mutations: frameshifts if inserted into coding regions and altered gene expression if inserted into regulatory regions. The number of copies of a particular TE in a species' genome therefore depends on several factors: the rate of integration of the TE, its rate of excision (if any), selection on individuals with phenotypes altered by TE integration, and regulation of transposition.

TEs effectively act as intracellular parasites and, like other parasites, might need to strike an evolutionary balance between their own proliferation and the detrimental effects on the "host" organism. Studies on *Drosophila* TEs confirm that the mutational integration of TEs generally has deleterious, sometimes lethal, phenotypic effects. This suggests that negative selection plays an important role in the regulation of transposition; individuals with high levels of transposition are less likely to survive and reproduce. However, we might expect that both TEs and their hosts might evolve mechanisms to limit transposition, and in fact both are observed. In one example of TE self-regulation, the *Drosophila* P element encodes a transposition repressor protein that is active in somatic tissue (see the *Transposable Elements and Retroviruses* chapter). In addition, there are two major cellular mechanisms for transposition regulation:

- In an RNA interference-like mechanism (see the *Regulatory RNA* chapter) involving piRNAs, the RNA intermediates of retrotransposons can be selectively degraded.
- In mammals, plants, and fungi, a DNA methyltransferase methylates cytosines within TEs, resulting in transcriptional silencing (see the *Epigenetics I* chapter).

In any case, it is rare for TE proliferation to continue unchecked but rather to be limited by negative selection and/or regulation of transposition. However, following introduction of a TE to a genome, the copy number can increase to many thousands or millions before some equilibrium is achieved, particularly if TEs are integrated into introns or intergenic DNA where phenotypic effects will be absent or minimal. As a result, genomes might contain a high proportion of moderately or highly repetitive sequences (see the chapter titled *The Content of the Genome*).

5.23 There Can Be Biases in Mutation, Gene Conversion, and Codon Usage

KEY CONCEPTS

- Mutational bias can account for a high AT content in organismal genomes.
- Gene conversion bias, which tends to increase GC content, can act in partial opposition to the mutational bias.
- Codon bias might be a result of adaptive mechanisms that favor particular sequences, and of gene conversion bias.

As discussed in the section *DNA Sequences Evolve by Mutation and a Sorting Mechanism* earlier in this chapter, the probability of a particular mutation is a function of the probability that a particular replication error or DNA-damaging event will occur and the probability that the error will be detected and repaired before the next DNA replication. To the extent that there is bias in these two events, there is bias in the types of mutations that occur (for example, a bias for transition mutations over transversion mutations despite the greater number of possible transversions).

Observations of the distributions of types of mutations over a taxonomically wide range of species (including prokaryotes and unicellular and multicellular eukaryotes), assessed by direct observation of mutational variants or by comparing sequence differences in pseudogenes, show a consistent pattern of a bias toward a high AT genomic content. The reasons for this are complex, and different mechanisms might be more or less important in different taxonomic groups, but there are two likely mechanisms. First, the common mutational source of spontaneous deamination of cytosine to uracil, or of 5-methylcytosine to thymine, promotes the transition mutation of C-G to T-A. Uracil in DNA is more likely to be repaired than thymine (see the *Genes Are DNA and Encode RNAs and Polypeptides* chapter), so methylated cytosines (often found in CG doublets) are not only mutation hotspots but specifically biased toward producing a T-A pair. Second, oxidation of guanine to 8-oxoguanine can result in a C-G to A-T transversion because 8-oxoguanine pairs more stably with adenine than with cytosine.

Despite this *mutational bias*, in analyses in which the expected equilibrium base composition is predicted from the observed rates of specific types of mutations, the observed AT content is generally lower than expected. This suggests that some mechanism or mechanisms are working to counteract the mutational bias toward A-T. One possibility is that this is adaptive; a highly biased base composition limits the mutational possibilities and consequently limits evolutionary

potential. However, as discussed next, there might be a nonadaptive explanation.

A second possible source of bias in genomic base composition is **gene conversion**, which occurs when heteroduplex DNA containing mismatched base pairs, often resulting from the resolution of a Holliday junction during recombination or double-strand break repair, is repaired using the mutated strand as a template (see the *Clusters and Repeats* chapter and the *Homologous and Site-Specific Recombination* chapter). Interestingly, observations of gene conversion events in animals and fungi show a clear bias toward G-C, though the mechanism is unclear. In support of this observation, chromosomal regions of high recombinational activity show more mutations to G-C, and regions with low recombinational activity tend to be A-T rich. The observed rates of gene conversion per site tend to be of the same order of magnitude or higher than mutation rates; thus gene conversion bias alone might account for the lower than expected AT content being driven higher by mutational bias. **Gene conversion bias** might also be partly responsible for another universally observed bias in genome composition, codon bias (see the section *A Constant Rate of Sequence Divergence Is a Molecular Clock* earlier in this chapter).

Due to the degeneracy of the genetic code, most of the amino acids found in polypeptides are represented by more than one codon in a genetic message. However, the alternate codons are not generally found in equal frequencies in genes; particularly in highly expressed genes, one codon of the two, four, or six that call for a particular amino acid is often used at a much higher frequency than the others. One explanation for this bias is that a particular codon might be more efficient at recruiting an abundant tRNA type, such that the rate or accuracy of translation is greater with higher usage of that codon. There might be additional adaptive consequences of particular exon sequences: Some might contribute to splicing efficiency, form secondary structures that affect mRNA stability, or be less subject to frameshift mutations than others (e.g., mononucleotide repeats that promote slippage). However, biased gene conversion remains a (nonadaptive) possibility, as well. Intriguingly, the synonymous site for most codons is the 3′ end, and high-usage codons in eukaryotes almost always end in G or C, as is consistent with the hypothesis that biased gene conversion drives codon bias. Clearly, the causes of codon bias are complex and might involve both adaptive and nonadaptive mechanisms.

Summary

- Genomes that have been sequenced include those of many bacteria and archaea, yeasts, nematode worms, fruit flies, mice, many plants, humans, and other species. The minimum number of genes required for a living cell (though a parasite) is about 470. The minimum number required for a free-living cell is about 1,500. A typical Gram-negative bacterium has about 1,500 genes. Genomes of strains of *E. coli* have gene numbers varying from 4,300 to 5,400. The average bacterial gene is about 1,000 bp long and is separated from the next gene by a space of about 100 bp. The yeasts *S. pombe* and *S. cerevisiae* have 5,000 and 6,000 genes, respectively.
- Although the fruit fly *D. melanogaster* has a larger genome than the nematode worm *C. elegans,* the fly has fewer genes (17,000) than the worm (21,700). The plant *Arabidopsis* has 25,000 genes, and the lack of a clear relationship between genome size and gene number is shown by the fact that the rice genome is 4 times larger but contains only 28% more genes (about 32,000). Mammals have 20,000 to 25,000 genes, many fewer than had been originally expected. The complexity of development of an organism can depend on the nature of the interactions between genes as well as their total number. In each organismal genome that has been sequenced, only about 50% of the genes have defined functions. Analysis of lethal genes suggests that only a minority of genes is essential in each organism.
- The sequences comprising a eukaryotic genome can be classified in three groups: nonrepetitive sequences are unique; moderately repetitive sequences are dispersed and repeated a small number of times in the form of related, but not identical, copies; and highly repetitive sequences are short and usually repeated as tandem arrays. The proportions of the types of sequence are characteristic for each genome, although larger genomes tend to have a smaller proportion of nonrepetitive DNA. Almost 50% of the human genome consists of repetitive sequences, the majority corresponding to transposon sequences. Most structural genes are located in nonrepetitive DNA. The complexity of nonrepetitive DNA is a better reflection of the complexity of the organism than the total genome complexity.
- Genes are expressed at widely varying levels. There might be 10^5 copies of mRNA for an abundant gene whose protein is the principal product of the cell, 10^3 copies of each mRNA for fewer than 10 moderately abundant transcripts, and fewer than 10 copies of each mRNA for more than 10,000 scarcely expressed genes. Overlaps between the mRNA populations of cells of different phenotypes are extensive; the majority of mRNAs are present in most cells.
- New variation in a genome is introduced by mutation. Although mutation is random with respect to function, the types of mutations that actually occur are biased by the probabilities of various changes to DNA and of types of DNA repair. This variation is sorted by random genetic drift (if variation is selectively neutral and/or populations are small) and negative or positive selection (if the variation affects phenotype).
- The past influence of selection on a gene sequence can be detected by comparing homologous sequences among and within species. The K_a/K_s ratio compares nonsynonymous with synonymous changes; either

an excess or a deficiency of nonsynonymous mutations might indicate positive or negative selection, respectively. Comparing the rates of evolution or the amount of variation for a locus among different species can also be used to assess past selection on DNA sequences. Applying these techniques to human genome sequences reveals that most functional variation is in noncoding (presumably regulatory) regions.

- Synonymous substitutions accumulate more rapidly than nonsynonymous substitutions (which affect the amino acid sequence). Researchers can sometimes use the rate of divergence at nonsynonymous sites to establish a molecular clock, which can be calibrated in percent divergence per million years. The clock can then be used to calculate the time of divergence between any two members of the family.
- Certain genes share only some of their exons with other genes, suggesting that they have been assembled by addition of exons representing functional "modular units" of the protein. Such modular exons may have been incorporated into a variety of different proteins. The hypothesis that genes have been assembled by accumulation of exons implies that introns were present in the genes of protoeukaryotes. Some of the relationships between orthologous genes can be explained by loss of introns from the primordial genes, with different introns being lost in different lines of descent.
- The proportions of repetitive and nonrepetitive DNA are characteristic for each genome, although larger genomes tend to have a smaller proportion of unique sequence DNA. The amount of nonrepetitive DNA is a better reflection of the complexity of the organism than the total genome size; the greatest amount of nonrepetitive DNA in genomes is about 2×10^9 bp.
- About 5,000 genes are common to prokaryotes and eukaryotes (though individual species might not carry all of these genes) and most are likely to be involved in basic functions. A further 8,000 genes are found in multicellular organisms. Another 5,000 genes are found in animals, and an additional 5,000 (largely involved with the immune and nervous systems) are found in vertebrates.
- An evolving set of genes might remain together in a cluster or might be dispersed to new locations by chromosomal rearrangement. Researchers can sometimes use the organization of existing clusters to infer the series of events that has occurred. These events act with regard to sequence rather than function and therefore include pseudogenes as well as functional genes. Pseudogenes that arise by gene duplication and inactivation are nonprocessed, whereas those that arise via an RNA intermediate are processed. Pseudogenes can become secondarily functional due to gain of function mutations or via their untranslatable RNA products.
- In some taxonomic groups, genome duplication (or polyploidization) can provide raw material for subsequent genome evolution. This process has shaped many flowering plant genomes and appears to have been a factor in early vertebrate evolution.
- Copies of transposable elements can propagate within genomes and sometimes result in a large proportion of repetitive sequences in genomes. The number of copies of an element is kept in check by selection, self-regulation, and host regulatory mechanisms.
- There are several sources of bias affecting the base composition of a genome. Mutational bias tends to result in higher AT content, whereas gene conversion bias acts to lower it somewhat. The universally observed codon biases of protein-coding sequences in genomes can be influenced by selection as well as gene conversion bias.

References

5.1 Introduction

Review

Lynch, M. (2007). *The Origins of Genome Architecture.* Sunderland, MA: Sinauer Associates Inc.

5.2 Prokaryotic Gene Numbers Range Over an Order of Magnitude

Reviews

Bentley, S. D., and Parkhill, J. (2004). Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.* 38, 771–792.

Hacker, J., and Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbio.* 54, 641–679.

Research

Blattner, F. R., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474.

Deckert, G., et al. (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus. Nature* 392, 353–358.

Galibert, F., et al. (2001). The composite genome of the legume symbiont *Sinorhizobium meliloti. Science* 293, 668–672.

5.3 Total Gene Number Is Known for Several Eukaryotes

Research

Adams, M. D., et al. (2000). The genome sequence of *D. melanogaster. Science* 287, 2185–2195.

Arabidopsis Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* 408, 796–815.

C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans:* a platform for investigating biology. *Science* 282, 2012–2022.

Duffy, A., and Grof, P. (2001). Psychiatric diagnoses in the context of genetic studies of bipolar disorder. *Bipolar Disord* 3, 270–275.

Dujon, B., et al. (1994). Complete DNA sequence of yeast chromosome XI. *Nature* 369, 371–378.

Goff, S. A., et al. (2002). A draft sequence of the rice genome(*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–114.

Johnston, M., et al. (1994). Complete nucleotide sequence of *S. cerevisiae* chromosome VIII. *Science* 265, 2077–2082.
Kellis, M., et al. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254.
Oliver, S. G., et al. (1992). The complete DNA sequence of yeast chromosome III. *Nature* 357, 38–46.
Wilson, R., et al. (1994). 22 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans. Nature* 368, 32–38.
Wood, V., et al. (2002). The genome sequence of *S. pombe. Nature* 415, 871–880.

5.4 How Many Different Types of Genes Are There?

Reference

Rual, J. F., et al. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437, 1173–1178.

Reviews

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* 422, 198–207.
Hanash, S. (2003). Disease proteomics. *Nature* 422, 226–232.
Phizicky, E., et al. (2003). Protein analysis on a proteomic scale. *Nature* 422, 208–215.
Sali, A., et al. (2003). From words to literature in structural proteomics. *Nature* 422, 216–225.

Research

Agarwal, S., et al. (2002). Subcellular localization of the yeast proteome. *Genes. Dev.* 16, 707–719.
Arabidopsis Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* 408, 796–815.
Gavin, A. C., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147.
Ho, Y., et al. (2002). Systematic identification of protein complexes in *S. cerevisiae* by mass spectrometry. *Nature* 415, 180–183.
Rubin, G. M., et al. (2000). Comparative genomics of the eukaryotes. *Science* 287, 2204–2215.
Uetz, P., et al. (2000). A comprehensive analysis of protein–protein interactions in *S. cerevisiae. Nature* 403, 623–630.
Venter, J. C., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1350.

5.5 The Human Genome Has Fewer Genes Than Originally Expected

Research

Clark, A. G., et al. (2003). Inferring nonneutral evolution from human–chimp–mouse orthologous gene trios. *Science* 302, 1960–1963.
Hogenesch, J. B., et al. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* 106, 413–415.
International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.
Mouse Genome Sequencing Consortium, et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562.
Venter, J. C., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1350.

5.6 How Are Genes and Other Sequences Distributed in the Genome?

Reference

Nusbaum, C., et al. (2005). DNA sequence and analysis of human chromosome 18. *Nature* 437, 551–555.

5.7 The Y Chromosome Has Several Male-Specific Genes

Research

Skaletsky, H., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423, 825–837.

5.8 How Many Genes Are Essential?

Research

Giaever, G., et al. (2002). Functional profiling of the *S. cerevisiae* genome. *Nature* 418, 387–391.
Goebl, M. G., and Petes, T. D. (1986). Most of the yeast genomic sequences are not essential for cell growth and division. *Cell* 46, 983–992.
Hutchison, C. A., et al. (1999). Global transposon mutagenesis and a minimal mycoplasma genome. *Science* 286, 2165–2169.
Kamath, R. S., et al. (2003). Systematic functional analysis of the *C. elegans* genome using RNAi. *Nature* 421, 231–237.
Tong, A. H., et al. (2004). Global mapping of the yeast genetic interaction network. *Science* 303, 808–813.

5.9 About 10,000 Genes Are Expressed at Widely Differing Levels in a Eukaryotic Cell

Research

Hastie, N. B., and Bishop, J. O. (1976). The expression of three abundance classes of mRNA in mouse tissues. *Cell* 9, 761–774.

5.10 Expressed Gene Number Can Be Measured En Masse

Reviews

Mikos, G. L. G., and Rubin, G. M. (1996). The role of the genome project in determining gene function: insights from model organisms. *Cell* 86, 521–529.
Young, R. A. (2000). Biomedical discovery with DNA arrays. *Cell* 102, 9–15.

Research

Holstege, F. C. P., et al. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95, 717–728.
Hughes, T. R., et al. (2000). Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126.
Stolc, V., et al. (2004). A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* 306, 655–660.
Velculescu, V. E., et al. (1997). Characterization of the yeast transcriptosome. *Cell* 88, 243–251.

5.12 Selection Can Be Detected by Measuring Sequence Variation

Research

Clark, R. M., et al. (2004). Pattern of diversity in the genomic region near the maize domestication gene *tb1. Proc. Natl. Acad. Sci. USA* 101, 700–707.
Clark, R. M., et al. (2005). Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. *Mol. Biol. Evol.* 22, 2304–2312.

Geetha, V., et al. (1999). Comparing protein sequence-based and predicted secondary structure-based methods for identification of remote homologs. *Protein Eng.* 12, 527–534.

McDonald, J. H., and Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351, 652–654.

Robinson, M., et al. (1998). Sensitivity of the relative-rate test to taxonomic sampling. *Mol. Biol. Evol.* 15, 1091–1098.

Wang, E. T., et al. (2006). Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci. USA* 103, 135–140.

5.13 A Constant Rate of Sequence Divergence Is a Molecular Clock

Research

Dickerson, R. E. (1971). The structure of cytochrome *c* and the rates of molecular evolution. *J. Mol. Evol.* 1, 26–45.

5.14 The Rate of Neutral Substitution Can Be Measured from Divergence of Repeated Sequences

Research

Waterston, R. H., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562.

5.15 How Did Interrupted Genes Evolve?

Review

Belshaw, R., and Bensasson, D. (2005). The rise and fall of introns. *Heredity* 96, 208–213.

Joyce, G. F., and Orgel, L. E. (2006). Progress toward understanding the origin of the RNA world. In: *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*, 3rd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Research

Barrette, I. H., et al. (2001). Introns resolve the conflict between base order-dependent stemloop potential and the encoding of RNA or protein: further evidence from overlapping genes. *Gene.* 270, 181–189. (See http://post.queensu.ca/~forsdyke/introns1.htm.)

Coulombe-Huntington, J., and Majewski, J. (2007). Characterization of intron-loss events in mammals. *Genome Research* 17, 23–32.

Forsdyke, D. R. (1981). Are introns in-series error detecting sequences? *J. Theoret. Biol.* 93, 861–866.

Forsdyke, D. R. (1995). A stem-loop "kissing" model for the initiation of recombination and the origin of introns. *Mol. Biol. Evol.* 12, 949–958.

Hughes, A. L., and Friedman, R. (2008). Genome size reduction in the chicken has involved massive loss of ancestral protein-coding genes. *Mol. Biol. Evol.* 25, 2681–2688.

Raible, F., et al. (2005). Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* 310, 1325–1326.

Roy, S. W., and Gilbert, W. (2006). Complex early genes. *Proc. Natl. Acad. Sci. USA* 102, 1986–1991.

5.16 Why Are Some Genomes So Large?

Review

Gall, J. G. (1981). Chromosome structure and the C-value paradox. *J. Cell. Biol.* 91, 3s–14s.

Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* 76, 65–101.

5.17 Morphological Complexity Evolves by Adding New Gene Functions

Reference

Chimpanzee Sequencing and Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69–87.

Research

Giaever, G., et al. (2002). Functional profiling of the *S. cerevisiae* genome. *Nature* 418, 387–391.

Goebl, M. G., and Petes, T. D. (1986). Most of the yeast genomic sequences are not essential for cell growth and division. *Cell* 46, 983–992.

Hutchison, C. A., et al. (1999). Global transposon mutagenesis and a minimal mycoplasma genome. *Science* 286, 2165–2169.

Kamath, R. S., et al. (2003). Systematic functional analysis of the *C. elegans* genome using RNAi. *Nature* 421, 231–237.

Tong, A. H., et al. (2004). Global mapping of the yeast genetic interaction network. *Science* 303, 808–813.

5.18 Gene Duplication Contributes to Genome Evolution

Research

Bailey, J. A., et al. (2002). Recent segmental duplications in the human genome. *Science* 297, 1003–1007.

5.19 Globin Clusters Arise by Duplication and Divergence

Review

Hardison, R. (1998). Hemoglobins from bacteria to man: evolution of different patterns of gene expression. *J. Exp. Biol.* 201, 1099–1117.

5.20 Pseudogenes Have Lost Their Original Functions

Research

Balasubramanian, S., et al. (2009). Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome. Biol.* 10, R2.

Esnault, C., et al. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* 24, 363–367.

Kaneko, S., et al. (2006). Origin and evolution of processed pseudogenes that stabilize functional Makorin1 mRNAs in mice, primates and other mammals. *Genetics* 172, 2421–2429.

Review

Balakirev, E. S., and Ayala, F. J. (2003). Pseudogenes: are they "junk" or functional DNA? *Ann. Rev. Genet.* 37, 123–151.

5.21 Genome Duplication Has Played a Role in Plant and Vertebrate Evolution

Research

Abbasi, A. A. (2008). Are we degenerate tetraploids? More genomes, new facts. *Biol. Direct.* 3, 50.

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678.

Dehal, P., and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS. Biol.* 3, e314.

Review

Furlong, R. F., and Holland, P. W. (2002). Were vertebrates octoploid? *Phil. Trans. R. Soc. Lond. B.* 357, 531–544.

Kasahara, M. (2007). The 2R hypothesis: an update. *Curr. Opin. Immunol.* 19, 547–552.

5.22 What Is The Role of Transposable Elements in Genome Evolution?

Research

Shen, S., et al. (2011). Widespread establishment and regulatory impact of Alu exons in human genes. *Proc. Natl. Acad. Sci. USA* 108, 2837–2842.

5.23 There May Be Biases in Mutation, Gene Conversion, and Codon Usage

Research

Rocha, E. P. C. (2004). Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome. Res.* 14, 2279–2286.