

15

Genomics

Chapter-at-a-Glance

The Challenge of Sequencing Entire Genomes

15.1 Genomics

New DNA sequencing technology can be used to determine the complete nucleotide sequence of an entire genome.

15.2 The Human Genome

The draft sequence for the human genome was published in February 2001. Perhaps the greatest surprises were how few genes there are in our genome, and that nearly half of our genome is composed of transposons containing no human genes.

15.3 Comparing Genomes

Researchers are producing a wealth of genome sequences. Many dozens of prokaryotes have had their genomes sequenced, and an increasing number of eukaryote genomes are being reported.

Putting Genomic Information to Work

15.4 Gene Microarrays

Hundreds of thousands of gene sequences can be placed on a glass chip and examined for differences from the reference human sequence.

15.5 Proteomics: The Next Frontier

Protein arrays, much like DNA microarrays, are now being developed to study all the proteins an organism possesses.

15.6 The Ethics of Genetic Testing

Genetic testing holds many potential benefits for diagnosing and preventing disease, but it also leads to many ethical questions regarding its uses.



The DNA strand you see here in this model contains within its twisting spiral the information specifying part of a protein's amino acid sequence. Humans contain about 20,000 to 25,000 protein-encoding genes, far fewer than scientists had expected. Over 98% of our DNA does not code for protein sequences—indeed, nearly half of human DNA seems to have been donated by mobile genetic elements called transposons. Biologists call the totality of the DNA in a cell its “genome.” The human genome, composed of some 6 billion nucleotides of DNA, has been sequenced, and the genome sequences of other animals and plants are rapidly being completed. This treasure trove of gene data is revolutionizing the study of evolution, because the sequence changes and gene rearrangements that occur during evolution are there for all to see. Telltale single nucleotide differences serve as convenient markers for gene disorders. Many of the thousands of nucleotide differences that make you unique will soon be detectable by routine gene microarray screening, unleashing a host of ethical and privacy issues.

15.1 Genomics

Recent years have seen an explosion of interest in comparing the entire DNA content of different organisms, a new field of biology called **genomics**. While initial successes focused on organisms with relatively small numbers of genes, researchers have recently completed the sequencing of several large eukaryotic genomes, including our own.

The full complement of genetic information of an organism—all of its genes and other DNA—is called its **genome**. The first genome to be sequenced was a very simple one: a small bacterial virus called ϕ -X174. Frederick Sanger, inventor of the first practical way to sequence DNA, obtained the sequence of this 5,375-nucleotide genome in 1977. This was followed by the sequencing of dozens of prokaryotic genomes. The advent of automated DNA sequencing machines in recent years has made the DNA sequencing of much larger eukaryotic genomes practical.

Sequencing DNA

DNA sequencing is a process that allows scientists to read each nucleotide in a strand of DNA. In sequencing DNA, DNA is cut into fragments using restriction enzymes. A DNA fragment of unknown sequence is then amplified, so there are thousands of copies of the fragment. The DNA fragments are then mixed with copies of DNA polymerase, copies of a primer (recall from chapter 12 that DNA polymerase can only add nucleotides onto an existing strand of nucleotides), a supply of the four nucleotide bases, and a supply of four different chain-terminating chemical tags. The chemical tags act as one of the four nucleotide bases in DNA synthesis, undergoing complementary base pairing. First, heat is applied to denature the double-stranded DNA fragments. The solution is then allowed to cool, allowing the primer (the lighter blue box in figure 15.1 ①) to bind to a single strand of the DNA, and

synthesis of the complementary strand proceeds. Whenever a chemical tag is added instead of a nucleotide base, the synthesis stops, as shown in the figure. For example, the terminating red “T” was added after three normal nucleotides and synthesis stopped. Because of the relatively low concentration of the chemical tags compared with the nucleotides, a tag that binds to G on the DNA fragment, for example, will not necessarily be added to the first G site. Thus, the mixture will contain a series of double-stranded DNA fragments of different lengths, corresponding to the different distances the polymerase traveled from the primer before a chain-terminating tag was incorporated (six fragments are shown in ①).

The series of fragments are then separated according to size by gel electrophoresis. The fragments become arrayed like the rungs of a ladder, each rung being one base longer than the one below it. Compare the lengths of the fragments in ① and their positions on the gel in ②. The shortest fragment is one nucleotide long (G), and that is also the lowest rung on the gel. In automated DNA sequencing, fluorescently colored chemical tags are used to label the fragments, one color corresponding to each nucleotide. Computers read off the colors on the gel to determine the DNA sequence and display this sequence as a series of colored peaks (③ and ④). What made the attempt to sequence large eukaryotic genomes practical was the development in the mid-1990s of automated sequencers that perform electrophoresis of DNA fragments in capillary tubes instead of the traditional gel slabs. These systems can handle about 1,000 samples a day, with only 15 minutes of human attention. A research institute with several hundred such instruments can sequence about 100 Mbp (million base pairs) every day.

15.1 Powerful automated DNA sequencing technology has begun to reveal the DNA sequences of entire genomes.

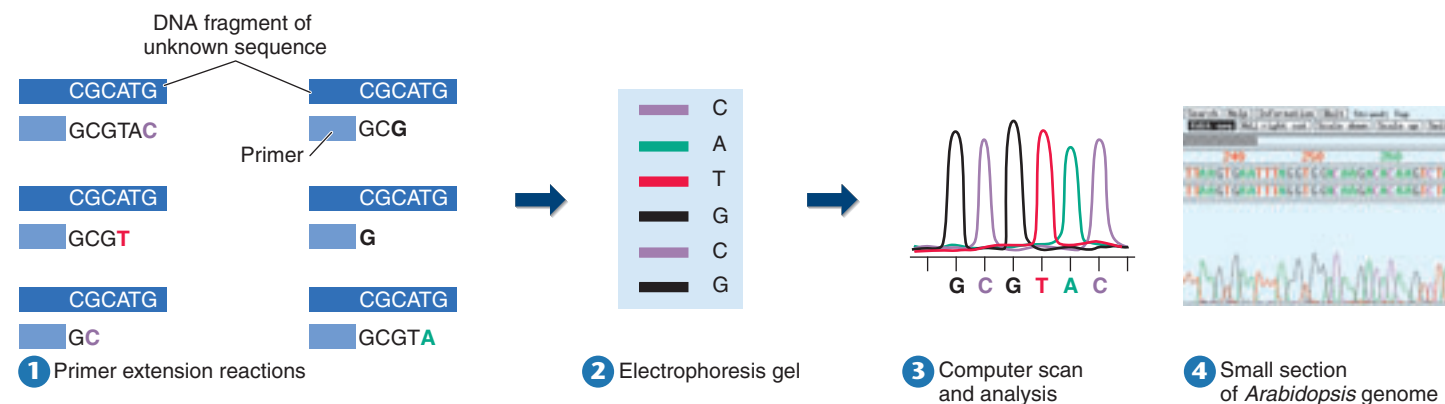


Figure 15.1 How to sequence DNA.

① DNA is sequenced by adding complementary bases to a single-stranded fragment. DNA synthesis stops when a chemical tag is inserted instead of a nucleotide, resulting in different sizes of DNA fragments. ② The DNA fragments of varying lengths are separated by gel electrophoresis, the smaller fragments migrating farther down the gel. (The boldface letters indicate the chemical tags added in step ① that stopped the replication process.) ③ Computers scan the gel, from smallest to largest fragments, and display the DNA sequence as a series of colored peaks. ④ Data from an automated DNA-sequencing run show the nucleotide sequence for a small section of the *Arabidopsis* (plant) genome.

15.2 The Human Genome

On June 26, 2000, geneticists announced that the entire human genome had been sequenced. This effort presented no small challenge, as the human genome is huge—more than 3 billion base pairs, which is the largest genome sequenced to date. To get an idea of the magnitude of the task, consider that if all 3.2 billion base pairs were written down on the pages of this book, the book would be 500,000 pages long, and it would take you about 60 years, working eight hours a day, every day, at five bases a second, to read it all.

Geography of the Genome

The preliminary report of the human genome sequence, published in 2001, estimated the number of protein-encoding genes to be 30,000. The final report, published in 2004, lowered that estimate to 20,000 to 25,000 protein-encoding genes. As we will discuss later, this is scarcely more than in nematodes at 21,000 genes, not quite double the number in *Drosophila* at 13,000 genes, and but a quarter of the number that had been anticipated by scientists based on the number of unique messenger RNA (mRNA) molecules.

How can human cells contain four times as many kinds of mRNA as there are genes? Recall from chapter 13 that in a typical human gene, the sequence of DNA nucleotides that specifies a protein is broken into many bits called exons, scattered among much longer segments of nontranslated DNA called introns. Imagine this paragraph was a human gene; all the occurrences of the letter “e” could be considered exons, while the rest would be noncoding introns. Look at figure 15.2, which breaks up the human genome into different types of DNA that will be discussed later; you see that introns make up 24% of the human genome.

When a cell uses a human gene to make a protein, it first manufactures mRNA copies of the gene, then splices the exons together. Now here’s the turn of events researchers had not anticipated: The transcripts of human genes are often spliced together in different ways, called alternative splicing. As we discussed in chapter 13, each exon is actually a module; one exon may code for one part of a protein, another for a different part of a protein. When the exon transcripts are mixed in different ways, very different protein shapes can be built.

With alternative mRNA splicing, it is easy to see how 25,000 genes can encode four times as many proteins. The added complexity of human proteins occurs because the gene parts are put together in new ways. Great music is made from simple tunes in much the same way.

In addition to the fragmenting of genes by the scattering of exons throughout the genome, there is another interesting “organizational” aspect of the genome. Genes are not distributed evenly over the genome. The small chromosome number 19 is packed densely with genes, transcription factors, and other functional elements. The much larger chromosome numbers 4 and 8, by contrast, have few genes. On most chromosomes, vast stretches of seemingly barren DNA fill the chromosomes between scattered clusters rich in genes.

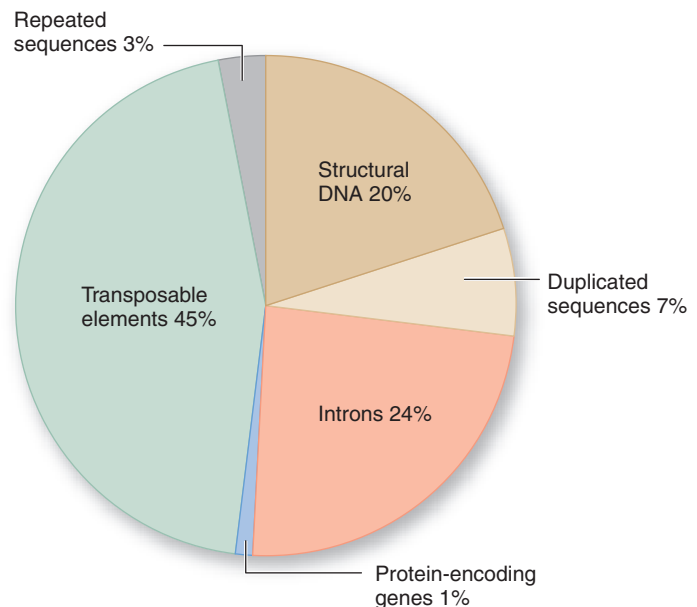


Figure 15.2 The human genome.

Very little of the human genome is devoted to protein-encoding genes. Surprisingly, a large portion of it appears to be composed of transposable elements.

DNA That Codes for Proteins

Four different classes of protein-encoding genes are found in the human genome, differing largely in gene copy number.

Single-copy genes. Many eukaryotic genes exist as single copies at a particular location on a chromosome.

Mutations in these genes produce recessive Mendelian inheritance of those traits. Silent copies inactivated by mutation, called *pseudogenes*, are as common as protein-encoding genes.

Segmental duplications. Human chromosomes contain many segmental duplications, whole blocks of genes that have been copied over from one chromosome to another. Blocks of similar genes in the same order are found throughout the genome. Chromosome 19 seems to have been the biggest borrower, with blocks of genes shared with 16 other chromosomes.

Multigene families. Many genes exist as parts of multigene families, groups of related but distinctly different genes that often occur together in a cluster. Multigene families contain from three to several dozen genes. Although they differ from each other, the genes of a multigene family are clearly related in their sequences, making it likely that they arose from a single ancestral sequence.

Tandem clusters. These groups of repeated genes consist of DNA sequences that are repeated many thousands of times, one copy following another in tandem array. By transcribing all of the copies in these tandem clusters simultaneously, a cell can rapidly obtain large amounts of the product they encode. For example, the genes encoding rRNA are present in clusters of several hundred copies.

Noncoding DNA

One of the most notable characteristics of the human genome is the startling amount of noncoding DNA it possesses. Only 1% to 1.5% of the human genome is coding DNA, devoted to genes encoding proteins. Each of your cells has about six feet of DNA stuffed into it, but of that, less than one inch is devoted to genes! Nearly 99% of the DNA in your cells seems to have little or nothing to do with the instructions that make you who you are. Table 15.1 provides an overview of the types of sequences found in the human genome. True genes are scattered about the human genome in clumps among the much larger amounts of noncoding DNA, like isolated hamlets in a desert.

There are four major types of noncoding human DNA:

Noncoding DNA within genes. As we discussed on the previous page, a human gene is made up of numerous fragments of protein-encoding information (exons) embedded within a much larger matrix of noncoding DNA (introns). Together, introns make up about 24% of the human genome, and exons about 1%.

Structural DNA. Some regions of the chromosomes remain highly condensed, tightly coiled, and untranscribed throughout the cell cycle. Called constitutive heterochromatin, these portions—about 20% of the DNA—tend to be localized around the centromere, or located near the telomeres, or ends, of the chromosome.

Repeated sequences. Scattered about chromosomes are simple sequence repeats (SSRs). An SSR is a two- or three-nucleotide sequence like CA or CCG, repeated like a broken record thousands and thousands of times. SSRs make up about 3% of the human genome. An additional 7% is devoted to other sorts of duplicated sequences. Repetitive sequences with excess C and G tend to be found in the neighborhood of genes, while A- and T-rich repeats dominate the nongene deserts. The light bands on chromosome karyotypes now have an explanation—they are regions rich in GC and genes (see figure 11.24). Dark bands signal neighborhoods rich in AT and thin on genes. Chromosome 19, dense with genes, has few dark bands.

Transposable elements. Fully 45% of the human genome consists of mobile bits of DNA called transposable elements. Discovered by Barbara McClintock in 1950 (she won the Nobel Prize in Physiology or Medicine for her discovery in 1983), transposable elements are bits of DNA that are able to jump from one location on a chromosome to another—tiny molecular versions of Mexican jumping beans. They work like a “cut-and-paste” or “copy-and-paste” word processing function.

Human chromosomes contain five sorts of transposable elements. Fully 20% of the genome consists of long interspersed nuclear elements (LINEs). An ancient and very successful element, LINEs are about 6 kb (6,000 DNA bases) long, and contain all the equipment needed for transposition, including genes for a DNA-loop-nicking enzyme and a reverse transcriptase.

Nested within the genome’s LINEs are over half a million copies of a parasitic element called *Alu*, composing 10% of the human genome. *Alu* is only about 300 bases long, and has no transposition machinery of its own; like a flea on a dog, *Alu* moves with the LINE it resides within. Just as a flea sometimes jumps to a different dog, so *Alu* sometimes uses the enzymes of its LINE to move to a new chromosome location. Often jumping right into genes, *Alu* transpositions cause many harmful mutations.

Three other types of transposable elements are also present in the human genome. Eight percent of the genome is devoted to long terminal repeats (LTRs). LTRs and LINEs are a type of transposon called “retrotransposons” because they involve an RNA intermediate. Three percent is devoted to DNA transposons, which copy themselves as DNA rather than RNA. And, some 4% is devoted to dead transposons, elements that have lost the signals for replication and so can no longer jump.

15.2 The entire 3.2-billion-base-pair human genome has been sequenced. Gene sequences vary greatly in copy number, some occurring many thousands of times, others only once. Only about 1% to 1.5% of the human genome is devoted to protein-encoding genes. Much of the rest is composed of transposable elements.

TABLE 15.1 TYPES OF DNA SEQUENCES FOUND IN THE HUMAN GENOME

Class	Frequency	Description
Protein-encoding genes	1%–1.5%	Translated exons, within some 25,000 genes scattered about the chromosomes
Introns	24%	Noncoding DNA comprising the great majority of most genes
Structural DNA	20%	Constitutive heterochromatin, localized near centromeres and telomeres
Repeated sequences	3%	Simple sequence repeats (SSRs) of a few nucleotides repeated thousands and thousands of times
Duplicated sequences	7%	Duplicated sequences, other than the SSRs
Transposable elements	45%	20% long interspersed nuclear elements (LINEs), active transposons 15% other transposable elements, including long terminal repeats (LTRs) 10% the parasite sequence (<i>Alu</i>), present in half a million copies

The Y Chromosome— Men Really Are Different

Our view of the differences between the sexes has recently undergone a radical revision. How do males and females differ? Seen through a biologist's eyes, the most basic difference between males and females is that all females have two copies of the so-called X chromosome. The X chromosome is about the same size as the other 44 human chromosomes, which also occur in pairs, and like them is packed with some 1,000 genes.

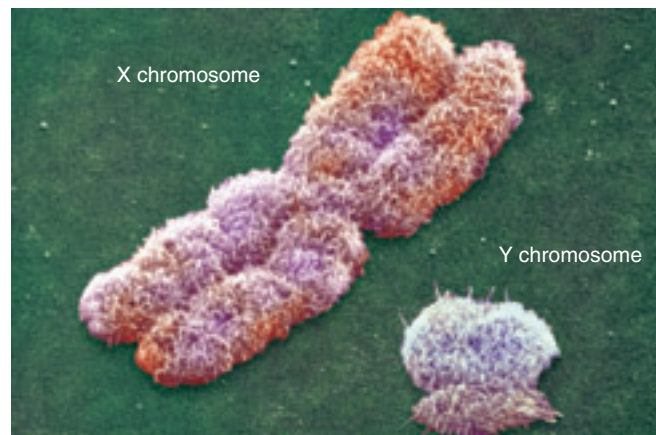
Biologists surmise that the reason there are two copies of the X and other chromosomes is to allow for the repair of the inevitable damage that occurs over time to individual genes because of wear and tear, chemical damage, and mistakes in copying. Because this sort of damage is passed on to offspring, it tends to accumulate over time. For this reason, genes must be edited every so often to repair the accumulated mutations.

How can a cell detect and edit out a mutation involving only one or a few nucleotides in one strand of DNA? How does it know which of the two DNA strands is the “correct” version and which is the altered one? This neat trick is achieved by every cell having two nearly identical copies of each chromosome. By comparing the two versions with each other, a cell can identify the “typos” and fix them.

Here is how it works: When a cell detects a chromosomal DNA duplex with a difference between its two DNA strands, that duplex is “repaired” by the rather Draconian expedient of chopping out the entire region, on both strands of the DNA molecule. No effort is made by the cell to determine which strand is correct—both are discarded. The gap that this creates is filled by copying off the sequence present at that region on the other chromosome. All this editing happens when the two versions of the chromosome are paired closely together in the early stages of meiosis.

So what are we to make of males? Males, by contrast to females, have only one copy of this X chromosome, not two. The other chromosome of the pair in males is called the Y chromosome and is much smaller than the X (the X and Y chromosomes of a male are shown in replicated form in the photo above). Biologists thought until very recently that the Y chromosome had only a few active genes. Because there is no other Y to serve as a pairing partner in meiosis, most of its genes had been thought to have decayed, the victims of accumulated mutations, leaving the Y chromosome a genetic wasteland with only a very few active genes surviving on it.

We now know this view to have been way too simple. In June of 2003, researchers reported the full gene sequence of the human Y chromosome, and it was nothing like biologists had expected. The human Y chromosome contains not one or two active genes, but 78!



Taking all these genes into account, geneticists conclude that men and women differ by 1% to 2% of their genomes—which is the same as the difference between a man and a male chimpanzee (or a woman and a female chimpanzee). So we are going to have to reexamine the basis of the differences between the sexes. A lot more of it may be built into the genes than we had supposed.

The Y chromosome is much smaller than the X and can only pair up with the X at the tips. Thus there can be no close pairing between X and Y during meiosis, the sort of pairing that allows the proofreading and editing just discussed. We can now see that there is a very good reason evolution has acted to prevent the close pairing of X and Y—those 78 Y chromosome genes. Because close pairing allows the exchange of large segments as well as small ones, any association of X and Y would lead to gene swapping, and the male-determining genes of the Y chromosome would sneak into the X, making everybody male.

One mystery remains. If the Y chromosome cannot pair with the X chromosome, how does it make do without copy-editing to prevent the accumulation of mutations? Why hasn't mutation-driven loss of genes long ago driven males to extinction? The answer to this question is right there for us to see in the Y chromosome sequence, and an elegant answer it is. Most of the 78 active genes on the Y chromosome lie within eight vast palindromes, which are regions of the DNA sequence that repeat the same sequence twice, running in opposite directions, like the sentence “Madam, I'm Adam,” or Napoleon's quip “Able was I ere I saw Elba.”

A palindrome has a very neat property: It can bend back on itself, forming a hairpin loop in which the two strands are aligned with nearly identical DNA sequences. This is the same sort of situation—alignment of nearly identical stretches of chromosomes—which permits the copy-edit of the X chromosome during meiosis. Thus in the Y chromosome, mutations can be “corrected” by conversion to the undamaged sequence preserved on the other arm of the palindrome. Damage does not accumulate, and males persist.

15.3 Comparing Genomes

Comparing genomes (entire DNA sequences) of different species provides a powerful new tool for biologists. Over 100 different prokaryotic genomes have been completed, as well as over 30 different eukaryotic genomes (some are shown in table 15.2), with dozens of other eukaryotes in the process of being sequenced. As the draft (preliminary) sequences of these genomes become available, our knowledge of the tree of life will become far clearer. The genomic sequences that are already completed give exciting clues of what is to come. Eight findings stand out as particularly important.

Finding One: More Complex Organisms Tend to Have More Genes

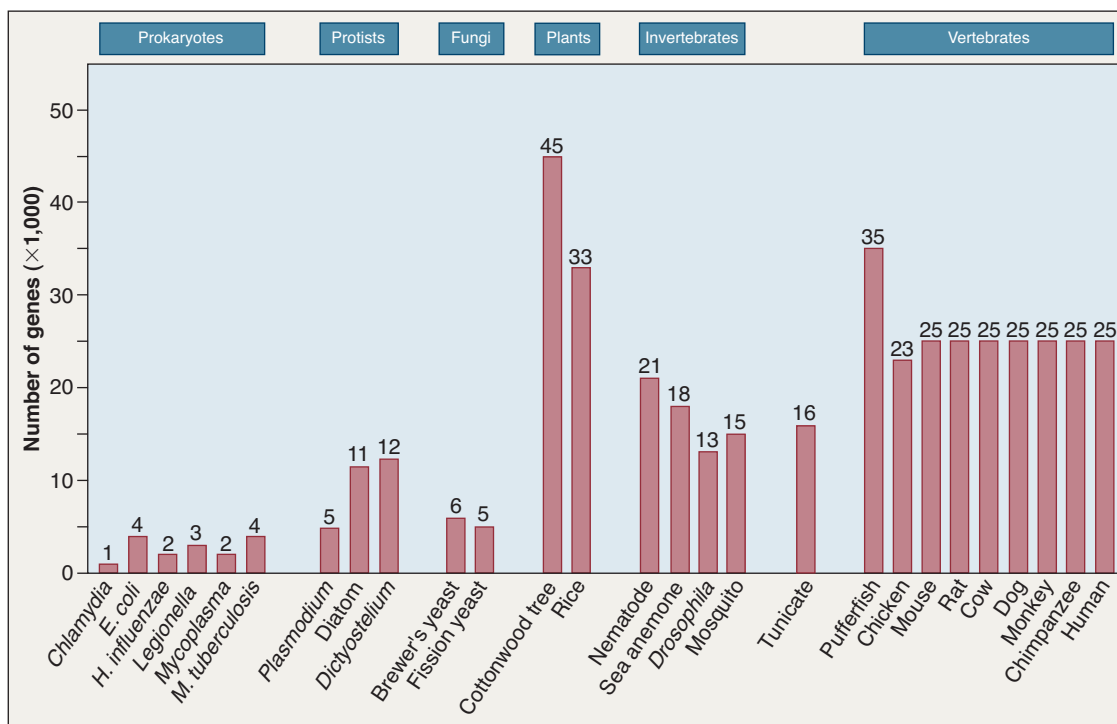
Comparing the number of genes in the genomes of organisms suggests that in a very rough way more complex organisms have more genes. You can see in figure 15.3 that insects (*Drosophila* and mosquito) have twice as many genes as single-celled organisms like bacteria and yeasts, and about half as many as mammals.

Finding Two: All Mammals Have the Same Size Genomes

Perhaps the most surprising finding of the human genome project was the small number of genes required to encode a human being—only about 20,000 to 25,000. Comparing genomes of different organisms, another surprising finding has emerged: All mammals are very much alike, with the same number of genes. As shown in figure 15.3, cows, rats, mice, dogs, monkeys, chimpanzees, and humans each have 20,000 to 25,000 genes. Genome sequences of the cat, rabbit, orangutan, elephant, opossum and numerous other mammals are being completed, and all are expected to possess genomes of this same size.

Figure 15.3
Comparing genome size.

All mammals have the same size genome, 20,000 to 25,000 protein-encoding nuclear genes. The unexpectedly larger sizes of the plant and pufferfish genomes are thought to reflect whole-genome duplications rather than increased complexity.



Finding Three: A Large Number of Genes Are New to Science

As biologists begin to examine the treasure trove of information provided by genomic sequences, another big surprise is that in each of the completely sequenced genomes so far there are large numbers of unfamiliar protein-encoding genes. The genome of the prokaryote *Aeropyrum pernix*, for example, contains more than 1,500 genes—57% of its total genome—not found in any other organism. Some 4,000 genes found in the genome of *Mycobacterium tuberculosis*, one of the best-studied bacteria, fall into the same category. Eukaryote genomes also contain many unexpected genes. Human chromosome 6 contains 1,557 protein-encoding genes, only 772 of which have been previously described. Similarly, human chromosome 7 contains 1,150 protein-encoding genes, only 605 of which have been previously described. The same pattern is seen in every genome examined so far—many genes are new to science. While some of these newly described genes resemble other genes whose functions we know, we don't know what these particular proteins are doing in the organism; other new genes that look conventional in every way have never been seen before and have no known function. Despite centuries of examination by biologists, it seems that organisms still have a lot of unknown equipment.

Finding Four: Large Differences in Genome Sizes Sometimes Arise Through Duplication of Chromosomes or Entire Genomes

Even a casual look at figure 15.3 reveals three glaring exceptions to the general finding that more complex organisms have more genes—cottonwood trees, rice, and pufferfish all have far more genes than their organismal complexity would suggest. A pufferfish has 40% more genes than a human! Do you really think this little fish is that much more complex than you are? Something else must be going on. When the genomes of

TABLE 15.2 SOME EUKARYOTIC GENOMES

Organism	Estimated Genome Size (Mbp)	Number of Genes ($\times 1,000$)	Nature of Genome
Vertebrates			
 <i>Homo sapiens</i> (human)	3,200	20–25	The first large genome to be sequenced; the number of transcribable genes is far less than expected; much of the genome is occupied by repeated DNA sequences.
 <i>Pan troglodytes</i> (chimpanzee)	2,800	20–25	There are few base substitutions between chimp and human genomes, less than 2%, but many small sequences of DNA have been lost as the two species diverged, often with significant effects.
 <i>Mus musculus</i> (mouse)	2,500	25	Roughly 80% of mouse genes have a functional equivalent in the human genome; importantly, large portions of the noncoding DNA of mouse and human have been conserved; overall, rodent genomes (mouse and rat) appear to be evolving more than twice as fast as primate genomes (humans and chimpanzees).
 <i>Gallus gallus</i> (chicken)	1,000	20–23	One-third the size of the human genome; genetic variation among domestic chickens seems much higher than in humans.
 <i>Fugu rubripes</i> (pufferfish)	365	35	The <i>Fugu</i> genome is only one-ninth the size of the human genome, yet it contains 10,000 more genes.
Invertebrates			
 <i>Caenorhabditis elegans</i> (nematode)	97	21	The fact that every cell of <i>C. elegans</i> has been identified makes its genome a particularly powerful tool in developmental biology.
 <i>Drosophila melanogaster</i> (fruit fly)	137	13	<i>Drosophila</i> telomere regions lack the simple repeated segments that are characteristic of most eukaryotic telomeres. About one-third of the genome consists of gene-poor centric heterochromatin.
 <i>Anopheles gambiae</i> (mosquito)	278	15	The extent of similarity between <i>Anopheles</i> and <i>Drosophila</i> is approximately equal to that between human and pufferfish.
 <i>Nematostella vectensis</i> (sea anemone)	450	18	The genome of this cnidarian is much more like vertebrate genomes than nematode or insect genomes that appear to have become streamlined by evolution.
Plants			
 <i>Oryza sativa</i> (rice)	430	33–50	The rice genome contains only 13% as much DNA as the human genome, but roughly twice as many genes; like the human genome, it is rich in repetitive DNA.
 <i>Populus trichocarpa</i> (cottonwood tree)	500	45	This fast-growing tree is widely used by the timber and paper industries. Its genome, fifty times smaller than the pine genome, is one-third heterochromatin.
Fungi			
 <i>Saccharomyces cerevisiae</i> (brewer's yeast)	13	6	<i>S. cerevisiae</i> was the first eukaryotic cell to have its genome fully sequenced.
Protists			
 <i>Plasmodium falciparum</i> (malaria parasite)	23	5	The <i>Plasmodium</i> genome has an unusually high proportion of adenine and thymine. Scarcely 5,000 genes contain the bare essentials of the eukaryotic cell.

these three organisms are examined, the reason for their high gene number becomes apparent. In each instance, the great increase in gene number has been the result of whole-genome duplication, a process called *polyploidy*, not the addition of new genes.

We can look at the pufferfish more closely to see how this happens. The last common ancestor of pufferfish and humans was a primitive bony fish that lived some 230 million years ago. The descendants of this long-extinct fish evolved into two distinct lineages, the ray-finned fishes (including the pufferfish) and the lobe-finned fishes (including the ancestors of humans). Sometime early in the ray-finned fish lineage, the entire genome duplicated! When the positions of more than 6,000 pufferfish genes are compared with the positions of corresponding genes in the human genome, one chromosomal region in humans matches *two* in pufferfish, across the entire genome. Illustrated in figure 15.4 (with duplicated genes highlighted in color), this pattern is a clear reflection of the whole-genome duplication among ray-finned fishes. Many similar duplications arose during the evolution of the plants.

Finding Five: Key Genes Tend to Be Conserved

Since the advent of gene sequencing machines made it possible to compare DNA sequences in different organisms, it has become clear that certain sequences are widespread among the kingdoms of life. Multiple copies of a 180-nucleotide sequence, the so-called *HOX* gene, occur in the genomes of all animals, both invertebrates and vertebrates. *HOX* genes play a key role in guiding development of the body plan, determining the number and orientation of body segments. Changes in four of these *HOX* genes have been shown to account in large part for the major differences in the bodies of crustaceans and insects. Related *HOX*-like gene sequences have been found in plants and yeasts, and even in prokaryotes. Apparently this sequence evolved very early in the history of life, and was of such importance that it has been conserved in animals, fungi, and plants virtually unchanged for hundreds of millions of years.

How common are such highly conserved genes? As more entire genomes become sequenced, it will become possible to address this question. Already it is clear that the number will be large. Only about one-third of the genes in cottonwood trees and rice appear to be in some sense “plant” genes, not found in any animal or fungal genome sequenced so far. These include the many thousands of genes involved in photosynthesis and photosynthetic anatomy. Two-thirds of the plant genomes are devoted to genes similar to those found in animal and fungi genomes, particularly genes involved in basic intermediary metabolism, in genome replication and repair, and in protein synthesis.

Finding Six: Rates of Evolution Vary Greatly

Comparison of rodent (mouse and rat) and primate (human and chimpanzee) genomes reveals that since mice and human lineages last shared a common ancestor about 75

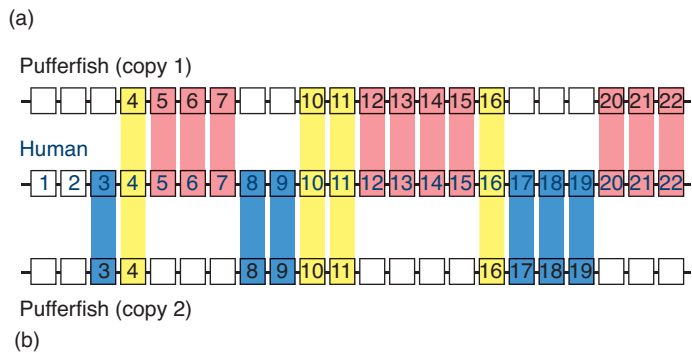


Figure 15.4 Detecting whole-genome duplications.

The genome of the pufferfish (a) has in its ancestry undergone a whole-genome duplication. In (b), human and pufferfish genes are shown. The middle section shows human genes numbered 1 through 22 that are also found in the same position in pufferfish. Some of them, in yellow, have related genes found in two different chromosomal locations in pufferfish (labeled as copy 1 and copy 2). Others have relatives only on copy 1 (red) or copy 2 (blue); the other copy having been lost or diversified.

million years ago, rodent DNA has mutated about twice as fast as primate DNA. This is a fascinating observation in search of an explanation. The difference in generation time between mice and humans (the average time between two successive generations or the time needed for offspring to become parents) could account for some of this difference, as mice have much shorter generation times and would have had more opportunities to mix and match genomic components during meiosis.

The insects are the most species-rich and morphologically diverse animal group on earth. Two insect genomes have been sequenced. *Drosophila melanogaster* has been a laboratory model for genetic studies for much of the last century, and is arguably the best-understood gene system in biology. *Anopheles gambiae*, the malaria mosquito, has an enormous impact on the world’s health, and the genomes of both were sequenced in 2002. The fruit fly *Drosophila* and the mosquito *Anopheles* are separated by approximately 250 million years of evolution, and appear to have evolved more rapidly over that interval than vertebrates, but keep in mind that they also have much shorter life cycles as well. The extent of similarity between these two insects is equivalent to that between humans and pufferfish, which diverged 450 million years ago.

What Makes Us Human?

For more than a century since Charles Darwin published *The Descent of Man* in 1871, his claim that chimpanzees (*Pan troglodytes*) are our closest relatives has received strong scientific support. The biochemical and anatomical evidence that will be outlined in chapter 27 is striking. The family tree you see to the right illustrates the conclusion biologists have come to—that among the apes, chimps are our closest nonhuman cousins.

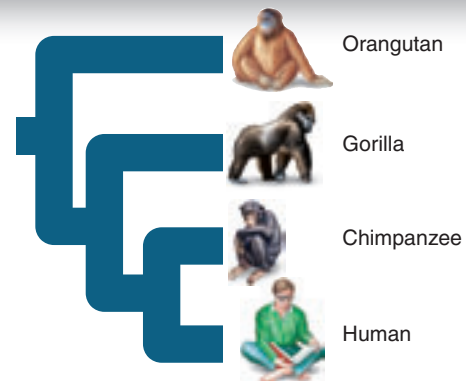
Perhaps one of the most interesting questions in biology is “How different are we?” Just what makes us human and not chimpanzee? Surely what sets us apart is our brain—about two million years ago, hominid brains began to enlarge until today your brain is three times as large as a chimpanzee’s. But the brains of humans and chimps are not anatomically very different; so while brain size has mattered (else why the increase), the difference we seek must be more subtle, perhaps reflecting how individual neurons are constructed or distributed. It appears that to answer this question clearly, we are going to have to look to the genomes—the human genome surely should have retained the imprint of our brain’s recent evolution.

So, 134 years after Darwin’s claim, with the publication in 2005 of the complete sequence of the chimpanzee genome, human and chimpanzee genomes finally can be directly compared. Detailed sequence comparisons of the two genomes indicate that just 1.4% of the DNA is different at the level of single letters of genetic code. This corresponds to 35 million single nucleotide changes.

How significant is this level of difference? Humans differ from one another by some 10 million single nucleotide changes (called *single nucleotide polymorphisms*, or *SNPs*) scattered throughout the human genome. So it seems that, in a very rough sense, you are four times as different from a chimpanzee as you are from another person. That is very alike indeed.

There are differences between human and chimpanzee genomes other than nucleotide substitutions, and these differences have a great impact. When the DNA sequences of chimp chromosomes are lined up nucleotide-to-nucleotide with their human counterparts, there are five million gaps, places where one or the other sequence is missing small stretches of DNA. Most of the gaps are less than 30 nucleotides long. They add another 3% to the total genome difference between the species. Gaps within gene sequences appear to be one of the major evolutionary mechanisms shaping primate species.

By extending the comparison of sequence gaps to other great apes (gorillas and orangutans), it is possible to infer whether a given sequence was added in one lineage or deleted in the other. Deletions and insertions seem to have occurred at similar frequencies in both lineages, with deletions far more common than insertions. Apparently humans and chimps have independently tended to lose bits



of their genomes in the 6 million years since they began to diverge from each other. The ancestor of humans and chimpanzees seems to have had a larger genome that was pared down differently in humans and chimps as the two species evolved, their genomes drifting apart.

The gaps seem to affect gene expression. Investigators used microarray chips, discussed later in this chapter, to compare patterns of gene transcription activity in human and chimpanzee tissues. While the same array of protein-encoding genes are present in chimp and human brain and liver tissues, about 20% of the genes showed significant variation in their expression in chimps and humans. These findings indicate that there are thousands of genes that are expressed differently in humans and chimpanzees. It would seem that much of the difference between the two species lies in which genes are transcribed, and when. Perhaps the best explanation of why a chimpanzee develops into a chimpanzee and not a human is that the genes are expressed at different times and possibly in different tissues.

Now we may ask: Of the many genes that are different between humans and our closest relatives, which are most responsible for making us “human”? In an attempt to answer this more pointed question, researchers in 2006 looked to see which human genome regions had evolved particularly rapidly in recent times. Comparing chimpanzee, mouse, and rat genomes with the human genome, they found 49 regions with a sequence that has changed very little among these other mammals, but has diverged very rapidly since our last common ancestor with chimpanzees. All but two lie outside protein-encoding sequences, and most lie near genes that have brain-development functions, orchestrating how our brain develops.

An even more direct approach to answering this question is being made by sequencing the genome of another human species, the now-extinct Neanderthals (you will encounter them in chapter 27). DNA obtained from inside a leg bone of a Neanderthal male who died about 38,000 years ago is being used to sequence the Neanderthal genome. Preliminary results were published in 2006, with the complete genome anticipated in late 2008. This analysis should let us see which of the rapidly evolving genome differences between us and chimpanzees have occurred since the human-Neanderthal split 0.5 million years ago. It is here that we must look to discover which evolutionary changes have made us what we are.

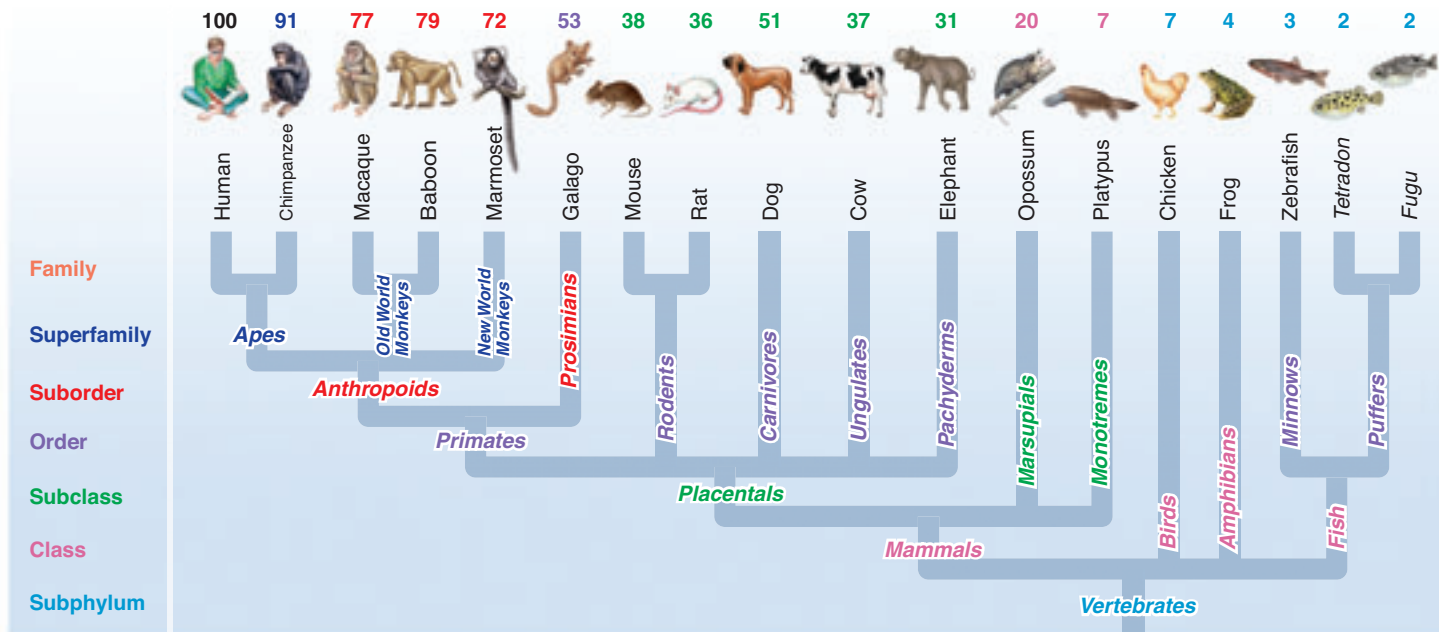


Figure 15.5 Genomic similarity reflects evolutionary relatedness.

The number above each organism is the percent of the nucleotides in selected regions of that organism’s genome that match those of the same regions in the human genome. The more distantly-related animals to humans, found toward the right side of the figure, also have fewer similarities with the human genome, compared to the more closely-related animals.

Finding Seven: Darwin Was Right—Genomes of Relatives Are More Alike

A key challenge of modern biology is to find a way to link the evolution of DNA sequences, which we are now able to study in great detail, with the evolution of the form and structure of complex organisms. Comparing genomes and portions of genomes of different species provides a powerful new tool to explore these relationships.

Comparison of genome sequences has already provided solid confirmation of a key prediction of Darwin’s theory of evolution, which is that close relatives would be expected to have fewer genetic differences than more distant ones. You will explore Darwin’s arguments closely in chapter 17, but here you need only note that his theory is one of increasing divergence over time, with all living things related back in time, part of a “family tree” of life. It is just this sort of family tree that we see revealed in the genome sequences now being completed. As Darwin’s theory predicts, the closer the relatives, the less the genomic difference we see (figure 15.5). All the other genomes in our order (primates) are more like the human genome than are any of those of another order, such as rodents (mouse and rat).

In general, as you proceed in figure 15.5 through the taxonomic categories from very close relatives (in the same family as humans) to very distant ones (different orders in the same class as humans), you can see clearly that genomic similarity decreases as taxonomic distance increases—just as Darwin’s theory predicts. This analysis is explored in more depth in chapter 17 in “A Closer Look,” pages 314–315.

Finding Eight: Noncoding DNA Is Not “Junk”

An unexpected finding in comparing mouse and human genomes lies in the similarities between the “junk” DNA, mostly transposons, in the two species. Large portions of the noncoding DNA (that is, sequences that do not appear to encode proteins) have been conserved. This DNA has not changed over the millions of years since humans and mice diverged from a common ancestor, and it has independently ended up in comparable regions of the genome.

It’s beginning to look like this noncoding “junk” DNA may have more of a function than was previously assumed. The possibility that it is rich in regulatory RNA sequences, such as described in chapter 13, is being actively investigated. In studies of two species of fruit flies, researchers found that 40% to 70% of noncoding DNA evolves much more slowly than gene-encoding DNA. This implies that these noncoding sequences are being maintained by natural selection.

Similarly, when researchers collected RNA transcripts made by a variety of different mouse tissues, as many as 4,280 RNA transcripts could not be matched to any known mouse protein-encoding gene. This suggests that a large part of the transcribed genome consists of non-protein-encoding genes—that is, of transcripts that function as regulatory RNA.

15.3 Genomes are more than instruction books for building and maintaining an organism; they contain vast amounts of information on the history of life. The growing number of fully sequenced genomes in all kingdoms is leading to a revolution in comparative evolutionary biology.

15.4 Gene Microarrays

Microarrays

A **gene microarray** is a glass square smaller than a postage stamp, covered with hundreds of thousands of different single strands of DNA rising from the surface like blades of grass. At each position on the glass plate, a particular DNA sequence of a hundred or more nucleotides is assembled, forming the microarray. The gene microarray chip you see in figure 15.6, called a *GeneChip* by its manufacturer, contains all known human gene sequences and can be purchased for as little as \$200.

How could you use such a microarray chip to delve into a person's genes? All you would have to do is to obtain a little of the person's DNA, say from a blood sample, and denature it to form single-stranded DNA. You would then flush fluid containing the person's denatured DNA over the chip surface. Wherever the DNA has a gene matching one of the microarray strands, it will stick to it in a way a computer can detect.

Gene microarrays can also be used to determine patterns of gene expression. To do this, mRNA isolated from the cells being studied is reverse transcribed, using fluorescently labeled nucleotides, to make complementary DNA (cDNA; see section 14.4). Because the cDNA contains fluorescently labeled nucleotides, it is easily recognized by a computer. When this labeled cDNA is mixed with a gene microarray representing many thousands of genes, spots light up on the computer screen corresponding to those genes being transcribed in the cells.

Similarly, two different sources of DNA can be compared, such as DNA from two different individuals, to determine their levels of genetic similarities. In this case, the DNA from the two sources is labeled with different-colored fluorescent labels, typically one labeled with a green fluorescent dye and the other with a red fluorescent dye. Spots that fluoresce are places where the samples of DNA bind to DNA on the microarray; the spots are reddish where one source binds and greenish where the other source binds. Where the two sources have similar DNA sequences, they bind to the same spot on the microarray, and it shows up as yellow spots. The more yellow spots there are, the more similar the source DNAs are.

Researchers are busily comparing the “reference sequence” of the human genome with the DNA of individual people, and noting any differences they detect. In this way, they are finding SNPs (single nucleotide polymorphisms), or spot differences in the identity of particular nucleotides, which record every way in which a particular individual differs from the reference sequence. Some SNPs are associated with disorders like cystic fibrosis or sickle-cell anemia. Others may give you red hair or elevated cholesterol in your blood. The human genome tells us that SNPs can be expected to occur at a frequency of about 5 per 1,000 nucleotides, scattered about randomly over the chromosomes. Each of us can be expected to differ from the standard “type sequence” by thousands of nucleotide SNPs.



Figure 15.6 Humanity on a chip.

Microarrays such as this Affymetrix GeneChip now include all known human genes.

Gene Microarrays Raise Critical Issues of Personal Privacy

Humans are thought to contain some 10 million different SNPs, all of which could reside on a small library of gene microarrays. When your DNA is flushed over a SNP microarray, the sequences that light up will instantly reveal your SNP profile, the genetic characteristics that make up who you are. Genes that might affect your health, your behavior, your future potential—all are there to be read. Your SNP profile will reflect all of this variation: a table of contents of your chromosomes, a molecular window to who you are.

When millions of such SNP profiles have been gathered over the coming years, computers will be able to identify other individuals with profiles like yours, and, by examining health records, standard personality tests, and the like, correlate parts of your profile with particular traits. Even behavioral characteristics involving many genes, which until now have been thought too complex to ever analyze, cannot resist a determined assault by a computer comparing SNP profiles.

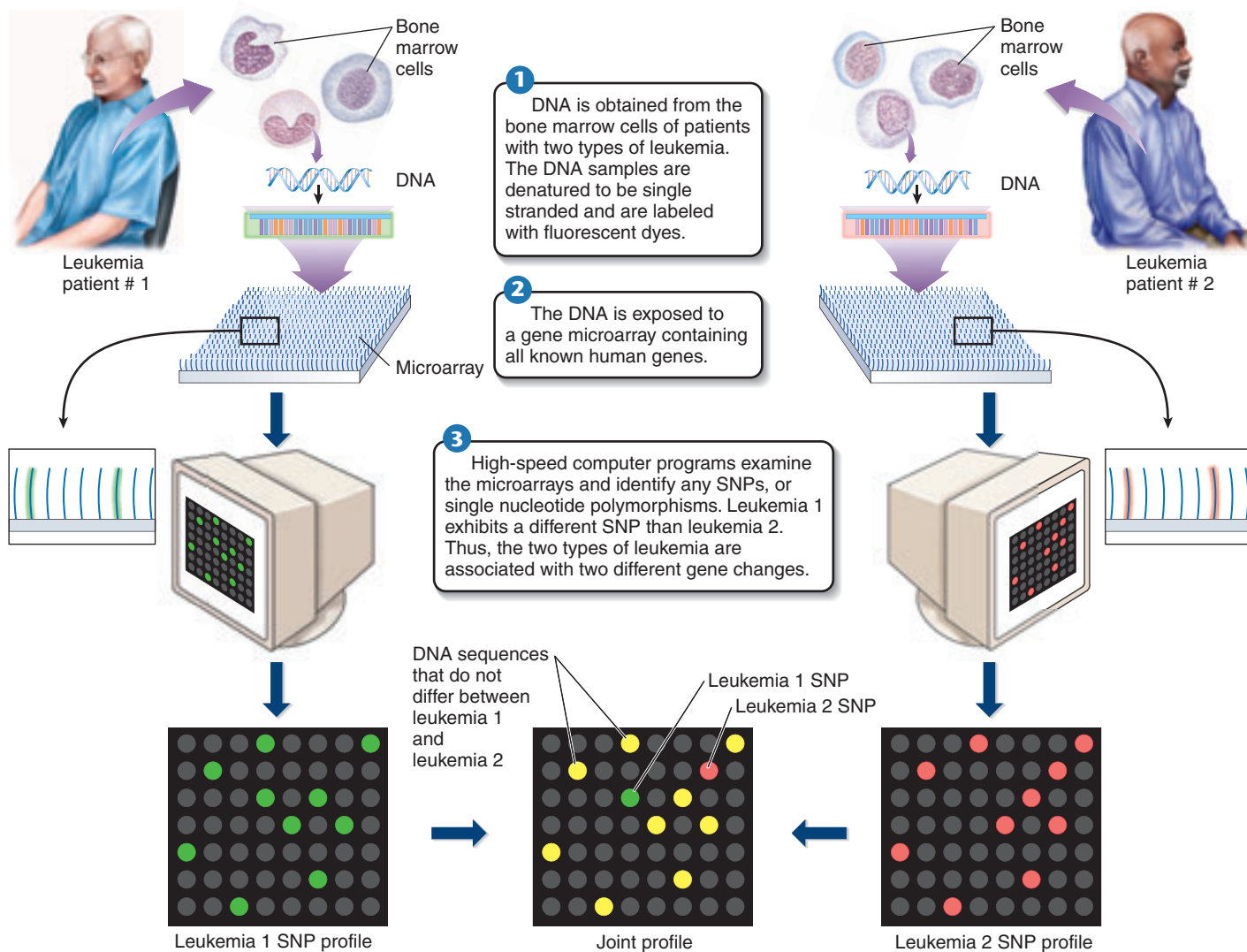


Figure 15.7 Gene microarrays can help in identifying precise subtypes of cancer.

How Microarray Gene Chips Can Be Used to Screen for Cancer

One of the biggest decisions facing an oncologist (cancer doctor) treating a tumor is to select the proper treatment. Most cancer cells look alike, although the tumors may in fact be caused by quite different types of cancer. If the oncologist could clearly identify the cancer, very targeted therapies might be possible. Unable to tell the difference for sure, however, oncologists take no chances. Tumors are treated with therapy that attacks all cancers, usually with severe side effects.

Researchers Todd Golub and Eric Lander took a vital step toward treating cancer, using gene microarrays to sniff out the differences between different subtypes of a deadly cancer of the immune system. Their method is illustrated in figure 15.7. Golub and Lander obtained bone marrow cells from patients with two types of leukemia (cancer of white blood cells) **1** and exposed DNA from each to microarrays containing 6,817 human genes **2**. The two types of leukemia each showed gene changes

from normal, but, importantly, the changes were different in each case, as shown in the “joint profile” where a computer compares the two profiles. Each had their own characteristic SNP profile **3**. Screening with the microarray thus provided a rapid and reliable way to determine which form of leukemia a patient possessed, greatly facilitating treatment.

Researchers have announced plans to compile a database of hundreds of thousands of SNPs over the next two years. Screening SNPs and comparing them with known SNP databases will soon allow doctors to screen each of us for copies of genes leading to genetic disorders.

15.4 A gene microarray is a discrete collection of gene fragments on a stamp-sized chip that can be used to screen for the presence of particular gene variants. Microarrays allow rapid screening of gene profiles, a tool that promises to have a revolutionary impact on medicine and society.

15.5 Proteomics: The Next Frontier

Ideally, a researcher would like to be able to examine a nucleotide sequence and know what sort of functional protein the gene specifies. However, efforts to calculate what shape a protein will assume from knowledge of its amino acid sequence have proven difficult, even with the aid of computers. However, by determining the actual structure of proteins produced by sequenced genes, researchers are beginning to get a clearer picture of how a gene sequence relates to protein shape.

Knowing that certain sequences tend to produce certain protein shapes such as twisting helices or kinks, powerful computer programs are now having considerable success in screening gene sequences within genomes for these particular telltale sequences, allowing increased success in predicting the structure of a protein from the nucleotide sequence of the gene encoding it. This fast-growing area of genomics, which involves the organization, storage, and indexing of sequence information, is one aspect of what is now loosely called **bioinformatics**.

With the sequencing of the human genome now essentially complete, researchers have begun an even more challenging task: the cataloging and analysis of every protein in the human body, an endeavor called **proteomics**. Each gene's nucleotide sequence specifies an amino acid sequence that folds in a cer-

tain way, producing a protein whose shape gives it a particular function. Only by understanding the protein shapes that genes produce can we really understand a genome.

Protein arrays, just like DNA microarrays, are now being developed to study all the proteins an organism possesses, its **proteome**. These arrays are screened using antibodies to specific proteins. Antibodies are fluorescently labeled so they can be detected, and the patterns on the protein array can then be determined by computer analysis. Technological advances are under way that will allow many proteins to be characterized on a mass scale in much less time than it took to uncover the structure of individual proteins in the past.

Fortunately, while there may be as many as a million different proteins, most are just variations on a handful of themes. The same shared structural motifs—barrels, helices, molecular zippers—are found in the proteins of plants, insects, and humans. The maximum number of distinct motifs has been estimated as fewer than 5,000. About 1,000 of these motifs have already been cataloged. Both public and privately financed efforts are now under way to detail the shapes of all the common motifs.

15.5 Like genomics, proteomics is a new approach that will enable analysis of proteins and comparisons at the protein level.

15.6 The Ethics of Genetic Testing

A greatly expanded use of genetic testing becomes potentially possible using gene microarrays. If a doctor knows that a patient carries a gene known to cause a particular disease, the doctor and patient can be better prepared to watch for symptoms and administer preventative treatments when available. The great increase in the power of genetic screening that gene microarray screening offers has the potential to alleviate great suffering and save lives. But there are ethical questions associated with this new technology, and we will need to carefully examine the ethical, legal, and social implications of this screening.

Among the chief ethical concerns are issues regarding confidentiality. If gene microarray screening reveals that a person carries a gene associated with risk of a disease, who should have access to this information? What if the screening reveals genes associated with undesirable personality traits? While doctor/patient confidentiality laws have always protected this type of information, such genetic risk factors affect not only the patient but also the patient's family. Do family members have a right to know that a person in the family carries a genetic mutation that could also affect them?

The potential for discrimination is also a very real concern. Information regarding a genetic risk factor for disease could lead to discrimination by employers and insurance companies. A person who is more likely to get sick may be seen as a risk to potential employers. Insurance companies may elect to restrict coverage on a person who carries a mutated gene. Is this right?

Other issues such as patient responsibility and treatments are also in question. Is it ethical to test for diseases that may not be treatable? Patients may have a right to know if they carry risk factors, but if no treatment is available, is it ethical to inform them? For some patients, a diagnosis may allow them time to prepare but for others, it may cause great anguish and distress. Who decides when testing should be performed?

A key aspect that may not be fully apparent to many patients is the concept of probability and risk in genetics. Just because a person carries a mutated gene does not mean that the person will definitely develop the disease, and the absence of a mutated gene does not give the "all clear" that a person will not develop the disease. The *BRCA1* and *BRCA2* mutations that have been identified for breast cancer account for 15% of all breast cancers and are associated with a lifetime risk of between 55% and 85%. What this means is that while these two mutations indicate a very high probability that the patient will develop breast cancer during her life, it is not 100%. Also, negative tests for either of these two genes does not mean that a woman won't develop breast cancer, it just means that she doesn't have a higher probability of developing the disease than the general population. This aspect of probability has to be communicated to the patient.

15.6 Genetic testing offers hope for predictive and preventative medicine, but it also opens the door for abuses and ethical issues.

Do Vertebrate Genomes Evolve as Darwin Predicted?

The genomic comparisons presented in figure 15.5 establish clearly that the genomes of close relatives are more similar than those of more distant ones, just as Darwin's theory predicts. Such comparisons are relative, however, as the analysis in the figure says nothing about how rapidly a particular taxonomic category evolves. Taxonomic categories are listed on the left side of figure 15.5, with orders being more distantly related than families, and classes even more distantly related. Do the genomes of vertebrates (animals with a backbone) in different classes from one another evolve more slowly than the genomes of vertebrates in different orders? And if we don't know about such evolutionary rates, how can we be sure how much the differences in genomic relatedness analyzed in figure 15.5 truly reflect the ongoing accumulation of evolutionary change? Might they instead be the result of some unsuspected peculiarity in how we define the taxonomic categories?

Fortunately, a more direct analysis is possible. The evolutionary history of the vertebrates is quite well known from fossils, and because many of these fossils have been independently dated using tools such as radioisotope dating (see page 47), it is possible to recast the analysis of figure 15.5 in terms of concrete intervals of time, and test directly whether or not vertebrate genomes accumulate more differences over longer periods of time as Darwin's theory predicts.

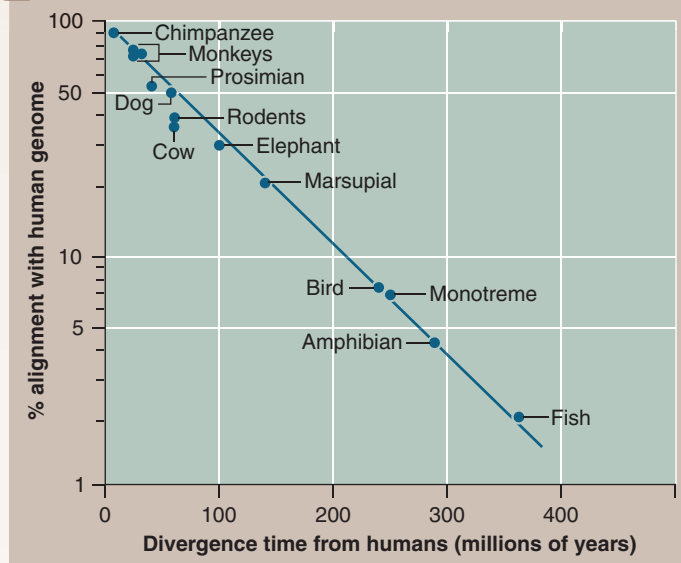
For each of the 17 nonhuman vertebrates in figure 15.5, the graph to the upper right plots genomic similarity—the percent (%) alignment with humans of 44 ENCODE regions (that is, how alike the DNA sequences of these regions are to that of the human genome)—plotted against that vertebrate's divergence time (that is, how many millions of years have elapsed since that vertebrate and humans shared a common ancestor in the fossil record). Thus the last common ancestor shared by birds and humans was an early reptile called a dicynodont that lived some 240 million years ago, and since then the genomes have changed so much that only 7% of their ENCODE sequences are still the same.

1. Applying Concepts

a. Variable. In the graph, which is the dependent variable?

b. Comparing Categories. Of the 17 kinds of vertebrates included in the study, which has the genome most similar to that of humans? Which has the least similar? Which of these two vertebrates diverged from humans more recently (that is, which had a more recent common ancestor in the fossil record)?

Genome Similarity and Divergence Time



2. Interpreting Data

a. What general statement can be made regarding the relationship between the % alignment with the human genome and taxonomic relatedness such as seen in figure 15.5—referring back to figure 15.5 and comparing it to this graph, are members of the same family found clustered together in the graph? Are vertebrates in the same order as humans more or less divergent than vertebrates in different orders? Are vertebrates in the same class as humans more or less divergent than those in different classes?

b. Four mammal genomes (cow, mouse, rat, and dog) all diverged from humans at about the same time 60 million years ago. Do all four genomes appear to have evolved at the same rate [Hint: Compare the vertical distances of points to the blue line]?

3. Making Inferences

a. In general, what is the relationship seen between similarity of a vertebrate genome to the human one (% alignment) and the time since that vertebrate and humans shared a common ancestor (divergence time)?

b. What is the significance of the fact that the relationship in the graph is a straight line when plotted on a log scale?

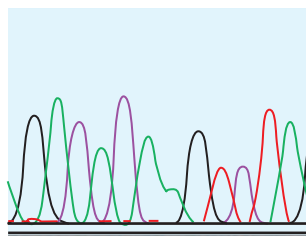
4. Drawing Conclusions Does this survey support the contention that evolution as seen in the fossil record is reflected at the DNA level? Do vertebrate genomes accumulate more differences over longer periods of time?

5. Further Analysis This analysis involves only 30 Mb (megabase or million bases), just 1% of the 3,000-Mb vertebrate genome. Do you think examining a larger portion of the genome would change the outcome of the analysis? Explain. If so, how would you recommend choosing which additional portions of the genome should be examined?

The Challenge of Sequencing Entire Genomes

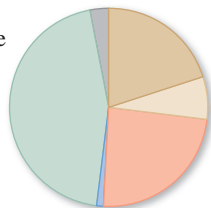
15.1 Genomics

- In sequencing DNA, the genome is cut into fragments using restriction enzymes. The fragments are amplified and the copies are mixed with the primer, DNA polymerase, a supply of the four nucleotides, and a smaller supply of chemically tagged nucleotides that terminate replication. The double-stranded DNA fragments are separated into single strands, and new DNA strands form on the template strands. When a chemically tagged base is added to the growing DNA strand, DNA synthesis stops. The mixture will contain double-stranded DNA fragments of different lengths. The DNA mixture is separated by fragment size using gel electrophoresis.
- By examining the fragments from shortest to longest and identifying the chemically tagged nucleotide at the end of each fragment, the nucleotide sequence of the DNA fragment can be determined. A computer scan of a fragment of DNA is shown here from **figure 15.1** where each peak represents a different fragment-terminating nucleotide. The DNA fragments are then linked together to reveal the sequence of the entire genome.



15.2 The Human Genome

- The human genome contains about 20,000 to 25,000 genes, not much more than other organisms and far less than what was expected based on the number of unique mRNA molecules present in our cells.
- The difference in the number of genes and the number of mRNA molecules is most likely due to alternative splicing in genes. Noncoding intron regions of the DNA represent 24% of the human genome, as indicated here by the orange area in the pie chart from **figure 15.2**. The coding regions of a gene are called exons and are scattered among the introns.
- Genes are not evenly distributed among the chromosomes in the human genome. Some chromosomes are loaded with genes, while others are sparsely populated. Genes also cluster in certain areas within a chromosome.
- Genes are organized in different ways in the genome. Many genes appear only once and are called single-copy genes. Some genes are contained within blocks of genes, called segmental duplications, that have been copied from one chromosome to another. Some genes are grouped with related genes in multigene families. Other genes are repeated many thousands of times, in repeating DNA sequences called tandem clusters.
- Nearly 99% of the human genome contains noncoding segments of DNA (**table 15.1**) that include introns, structural DNA around the centromere regions, groups of short repeating sequences, and transposable elements. Transposable elements called LINEs, for long interspersed nuclear elements, contain the genes needed for transposition. Other transposons, such as *Alu*, cannot undergo transposition on their own and are found within LINE regions.



15.3 Comparing Genomes

- Comparing the genomes of different species reveals relationships between species and is revolutionizing the study of comparative evolutionary biology (**table 15.2**). Eight main concepts have been

identified. The first finding is that more complex organisms tend to have more genes (**figure 15.3**).

- Findings two and three are that all mammals have approximately the same size genome and that a large number of genes being identified are new to science. Some with unknown functions.
- Finding four states that large differences in genomes sometimes arise through duplication of chromosomes, a process called polyploidy. Finding five is that genes controlling key processes tend to be conserved between even distantly related organisms.
- Sixth, the rate of evolution of genomes varies greatly. The differences in generation time may account for some variation.
- Finding seven states that the genomes of more closely related organisms are more alike than those of more distantly related organisms (**figure 15.5**).
- Finding eight is that the noncoding “junk” DNA may serve an important regulatory function and is conserved between organisms which suggests that it is maintained through natural selection.

Putting Genomic Information to Work

15.4 Gene Microarrays

- A microarray is a collection of hundreds of thousands of known DNA sequences displayed on a small glass plate, like the *GeneChip* shown here from **figure 15.6**.
- When a DNA sample is denatured to single strands and washed over the microarray, complementary sequences will bind to the DNA on the microarray. Computers detect the double-stranded DNA and analysis reveals which genes are present in the sample DNA.
- The level of gene expression can also be detected using microarrays. Fluorescently-tagged cDNAs are produced from a cell’s mRNAs and are applied to a microarray. Complementary binding of the cDNAs to the microarray will result in a pattern that can be analyzed by a computer. The DNA of two different organisms or individuals can be compared using a similar procedure.
- Researchers are identifying sequences that have small differences in nucleotide sequences in the human genome. These small variations, called single nucleotide polymorphisms (SNPs) are associated with genetic disorders or result in variations in genetic traits.
- Researchers are identifying sequences associated with diseases and subtypes of cancers. These “reference sequences” can be used to identify individuals that are more susceptible to certain diseases or aid in their treatments (**figure 15.7**).



15.5 Proteomics: The Next Frontier

- Sequencing the human genome has advanced the study of proteins, an area called proteomics. Researchers are now identifying relationships between gene sequences and protein structure and function. Certain sequences tend to produce similar protein shapes and this information will help scientists more accurately predict protein shapes based on DNA sequences.

15.6 The Ethics of Genetic Testing

- Genetic testing can help identify not only people who are affected by a genetic disorder, but also those who could be affected. This opens possibilities for predictive and preventative medicine but also leads to questions about its use and how to protect the information. Many ethical questions related to genetic testing are being discussed.

